A Spectral Approach to Network Design and Experimental Design

by

Hong Zhou

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Nikhil Bansal
Professor
Dept. of Math and CS, Eindhoven University of Technology

Supervisor: Lap Chi Lau
Associate Professor
Chertion School of Computer Science, University of Waterloo

Internal Member: Eric Blais
Associate Professor
Chertion School of Computer Science, University of Waterloo

Yaoliang Yu
Assistant Professor
Chertion School of Computer Science, University of Waterloo

Internal-External Member: Chaitanya Swamy
Professor
Dept. of Combinatorics and Optimization, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the 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.

Statement of Contributions

The main results of this thesis are based on the following papers that I have coauthored.

1. [98]: *A Spectral Approach to Network Design.* Joint work with Lap Chi Lau. A preliminary version appears in Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC '20, page 826–839, 2020.

2. [97]: *A Local Search Framework for Experimental Design.* Joint work with Lap Chi Lau. To appear in Proceedings of the 32nd Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '21, 2021.

3. [38]: *Network Design for s-t Effective Resistance.* Joint work with: Pak Hay Chan, Lap Chi Lau, Aaron Schild, Sam Chiu-Wai Wong. ArXiv preprint arXiv:1904.03219, 2019.

Abstract

Over the last decade, the spectral sparsification technique has become a powerful tool in designing fast graph algorithms for various problems with numerous applications. In this thesis, we extend this spectral approach, and show that it is also very powerful in designing approximation algorithms for classical network design and experimental design problems.

The central piece in this thesis is a problem called spectral rounding, which is inspired by spectral sparsification and studied in an earlier work on experimental design. In this problem, we are given vectors $v_1, \ldots, v_m$ each with a non-negative cost, and a fractional solution $x \in [0,1]^m$. The task is to find an integral solution $z \in \{0,1\}^m$ such that the spectrum of the integral solution is similar to the one of the fractional solution, i.e. $\sum_i z(i) \cdot v_i v_i^\top \approx \sum_i x(i) \cdot v_i v_i^\top$, and the integral cost is approximately equal to the fractional cost.

We observe that the spectral rounding problem underlies a large family of network design and experimental design problems. With this perspective, we bring new insights into these well-studied problems. For network design, we show that the spectral rounding technique provides a novel and general approach to significantly extend the scope of problems that can be solved efficiently. For experimental design, we show that the spectral rounding technique provides a unified and elegant framework that matches and improves all known existing algorithmic results.

There are two key techniques that we will use in this thesis. The first one is regret minimization, which is well-known to the online optimization community and has been used for spectral sparsification. We use it to control the spectrum of the integral solution in the spectral rounding problem. The second key technique is concentration inequalities for analyzing adaptive random sampling processes, which enable us to satisfy spectral and linear constraints simultaneously with high probability.

Acknowledgements

The journey of finishing this thesis has been long and colorful. I would like to take this opportunity to thank those who have helped and supported me along this journey.

First and foremost, I would like to convey my sincere gratitude and thanks to my supervisor Lap Chi Lau. I am deeply indebted to him for kindly providing me the opportunity to start this journey and constantly supporting me through it. I am also indebted to him for his valuable and thorough comments on my thesis. This thesis would not have been possible without his inspiring questions and his patience and encouragements. His clear-thinking, persistence in research, and his writing and presenting style shaped the way that I pursue my research.

I am also grateful to Nikhil Bansal, Eric Blais, Chaitanya Swamy, and Yaoliang Yu for reading my thesis, serving on my examining committee, and providing many insightful comments.

I was fortunate to have great collaborators: Pak Hay Chan, Aaron Schild, and Sam Chiu-wai Wong. I also greatly appreciate Tsz Chiu Kwok and Akshay Ramachandran for insightful and helpful discussions that improved this thesis.

I must also thank my friend Weiwei Wu who inspired my interests in theoretical computer science before I started this journey.

Thank you to my friends and fellow students in Waterloo: Vedat Levi Alev, Abhinav Bommireddi, Pak Hay (Alan) Chan, Hicham El-Zein, Nathaniel Harms, Amit Levi, Vijay Menon, Akshay Ramachandran, Anurag Murty Naredla, Kam Chuen (Alex) Tung, and also to faculty members and other students in the Algorithms and Complexity group and Combinatorial and Optimization Department, who created a friendly and welcoming environment around me. In particular, I would like to thank my long-term officemates Vedat, Alan, and Akshay for their supports and the enjoyable conversations throughout the years.

I also like to extend my thanks to friends and members in Lincoln Road Chapel Mandarin Congregation for making me feel at home in Waterloo.

I was fortunate to be born to my wonderful parents Xingwei Zhou and Pingfang Luo. I cannot thank them enough for their unconditional and never ending love and support. I

am also fortunate to have a wonderful gift from God, my beloved wife Melody Zhao. I am grateful to have her sweet love and full support in my life.

Last but definitely not the least, I thank the gracious and merciful God for providing me peace and strength throughout this colorful journey.

Dedication

Dedicated to my grandparents.

Table of Contents

Chapter 1

Introduction

In the past decade, the linear algebraic perspective to solving graph problems has become a powerful tool in designing fast graph algorithms [132, 44, 21, 7, 9, 128]. In this thesis, we extend this spectral approach and find new connections and interesting results in network design and experimental design.

1.1 The Central Problem: Spectral Rounding

The following spectral rounding problem is the central problem in this thesis. The version we stated in the abstract is equivalent to the following one by a simple reduction (see, e.g., Section 6.1.3).

Question 1.1.1 (Spectral Rounding)**.** *Suppose we are given vectors* $v_1, \ldots, v_m \in \mathbb{R}^d$ *and* $x \in [0,1]^m$ *such that* $\sum_{i=1}^m x(i) \cdot v_i v_i^\top = I_d$*, where* I_d *is the d-dimensional identity matrix. Given a non-negative "cost" vector* $c \in \mathbb{R}_+^m$*, find* $z \in \{0,1\}^m$ *such that*

$$\sum_{i=1}^{m} z(i) \cdot v_i v_i^\top \approx I_d \qquad \text{and} \qquad \langle c, z \rangle \approx \langle c, x \rangle.$$

This problem is similar to the spectral sparsification problem introduced by Spielman and Teng [133]. In spectral sparsification, the goal is to find a *sparse* non-negative vector

$y \in \mathbb{R}^m_+$ to approximate the spectral properties of a given fractional vector x. Spectral rounding is different in that we want to find an *integral* vector $z \in \{0,1\}^m$ to approximate the spectral properties of x and preserve the cost simultaneously.

To approximate the spectral properties of a fractional vector, we consider two different settings.

- One-sided spectral rounding: Find $z \in \{0,1\}^m$ such that
$$\sum_{i=1}^m z(i) \cdot v_i v_i^\top \gtrsim I_d \qquad \text{and} \qquad \langle c, z\rangle \approx \langle c, x\rangle.$$
- Two-sided spectral rounding: Find $z \in \{0,1\}^m$ such that
$$\sum_{i=1}^m z(i) \cdot v_i v_i^\top \approx I_d \qquad \text{and} \qquad \langle c, z\rangle \approx \langle c, x\rangle.$$

The one-sided spectral rounding was formulated and studied by Allen-Zhu, Li, Singh, and Wang [6] when the cost vector $c = \mathit{1}_m$. We extend their results to incorporate general costs. The following is our main algorithmic result for the one-sided spectral rounding.

Theorem (Informal)**.** *Suppose we are given vectors* $v_1, \ldots, v_m \in \mathbb{R}^d$ *and* $x \in [0,1]^m$ *such that* $\sum_{i=1}^m x(i) \cdot v_i v_i^\top = I_d$. *For any given non-negative vector* $c \in \mathbb{R}^m_+$, *if* $\langle c, x\rangle$ *is large enough, then one-sided spectral rounding is always possible. In particular, there is a polynomial time randomized algorithm that returns a solution* $z \in \{0,1\}^m$ *with high probability such that*
$$\sum_{i=1}^m z(i) \cdot v_i v_i^\top \succsim I_d \qquad \textit{and} \qquad \langle c, z\rangle \approx \langle c, x\rangle.$$

We design an iterative randomized rounding algorithm to prove the above theorem. Initially, each vector v_i is selected at random with probability $x(i)$ independently. Then, in each iteration, the algorithm adaptively samples a vector v_i to remove from the current solution set S, and samples a vector v_i to add into S. The sampling probabilities of v_i and v_j are based on their contributions to the spectrum of the current solution and the corresponding fractional values $x(i)$ and $x(j)$. This adaptive randomized sampling approach is the main theme in this thesis, which allows us to improve the spectrum and preserve linear constraints simultaneously.

When all the vectors are short and the cost constraint is ignored, a recent result of Kyng, Luh, and Song [91] proves that two-sided spectral rounding is always possible. We extend their result to show that the two-sided spectral rounding with short vectors and general costs is always possible. We remark that these are existential results, as the proofs use the nonconstructive interlacing polynomial method.

1.2 Applications to Network Design

Network design is a central topic in combinatorial optimization, approximation algorithms and operations research. The general setting of network design is to find a minimum cost subgraph satisfying certain requirements. The most well-studied problem is the survivable network design problem [71, 1, 73, 67], where the requirement is to have at least a specified number f_{uv} of edge-disjoint paths between every pair of vertices u, v. A seminal work of Jain [79] introduced the iterative rounding method for linear programming to design a 2-approximation algorithm for the survivable network design problem, and this method has been extended to various more general settings [62, 66, 43, 93, 94, 56, 64, 96, 16]. There are also other linear programming based algorithms such as randomized rounding [135, 68, 32, 12, 75] to obtain important algorithmic results for network design. It is widely recognized that linear programming is the most general and powerful approach in designing approximation algorithms for network design problems.

Using spectral rounding, we provide a completely different approach to this well-studied topic. The spectral requirement in the spectral rounding problem not only captures pairwise edge connectivity requirements in survivable network design, but also allows us to have a control over many other useful and interesting properties, e.g., algebraic connectivity, graph expansion, pairwise effective resistance, etc. Before our work, there was no good approximation algorithm for these properties, even individually. The spectral approach provides us a powerful tool to tackle a generalized network design problem that incorporates connectivity constraints, effective resistance constraints, algebraic connectivity constraints, and some other unstructured linear constraints simultaneously. The following is our main result for the generalized network design problem, which significantly extends the scope of useful properties that a network designer could control simultaneously to design better networks.

Theorem (Informal). *For any ε, there is a convex programming based polynomial time randomized algorithm to return an integral solution z of the generalized network design problem that simultaneously satisfies all the connectivity constraints, the effective resistance constraints, the algebraic connectivity constraint and the capacity constraints exactly with high probability. The cost of the integral solution z is*

$$\langle \mathsf{c}, \mathsf{z} \rangle \leqslant (1 + O(\varepsilon)) \cdot \mathsf{opt} + O\left(\frac{n \left\|\mathsf{c}\right\|_\infty}{\varepsilon}\right)$$

with high probability, where opt is the cost of an optimal (fractional) solution, n is the number of vertices in the graph and $\left\|\mathsf{c}\right\|_\infty$ is the maximum cost of an edge. Furthermore, unstructured linear constraints can be satisfied approximately with high probability.

Besides the generalized network design problem, we also show that spectral rounding can be applied to spectral network design problems with spectral objective functions, e.g., maximizing algebraic connectivity [69], minimizing total effective resistance [70], etc. Finally, we mention that the spectral rounding techniques are useful for graph problems other than network design. For example, we use it to design new algorithms for additive spectral sparsification, a new notion of sparsification recently introduced by by Bansal, Svensson and Trevisan [20].

1.3 Applications to Experimental Design

Experimental design is a classical topic in statistics [60, 13, 121, 74]. Recently, it has found applications in various areas, e.g., machine learning [10, 29, 114, 34], signal processing [80, 37, 40, 41], numerical linear algebra [50, 51, 28, 15], etc.

In experimental design problems, we are given vectors $\mathsf{u}_1, \ldots, \mathsf{u}_n \in \mathbb{R}^d$ and a budget $b \geqslant d$, the goal is to choose a (multi-)subset S of b vectors so that S is a representative of all the n vectors. There are different objective functions to measure the quality of the representative set S. The most popular and well-studied objective functions are related to spectral properties of those vectors in S:

- D-design: Maximizing $\left(\det\left(\sum_{i\in S} \mathsf{v}_i \mathsf{v}_i^\top\right)\right)^{\frac{1}{d}}$.
- A-design: Minimizing $\operatorname{tr}\left(\left(\sum_{i\in S} \mathsf{v}_i \mathsf{v}_i^\top\right)^{-1}\right)$.
- E-design: Maximizing $\lambda_{\min}\left(\sum_{i\in S} \mathsf{v}_i \mathsf{v}_i^\top\right)$.

All three experimental design problems are NP-hard [33, 142] and also APX-hard [136, 118, 33]. Despite the long history and the wide interest, strong approximation algorithms for these problems have been obtained only very recently [6, 131, 118]. These state-of-the-art algorithms use completely different techniques for each different experimental design problem.

We show that the spectral rounding technique leads to an elegant framework to design and analyze both rounding algorithms and combinatorial algorithms for experimental design problems. This framework provides a unifying approach to match and improve all known results in D/A/E-design and to obtain new results in previously unknown settings.

Theorem (Informal). *For rounding algorithms, there is a unified randomized local-search framework that matches and improves all known rounding algorithms for D/A/E-design. Furthermore, the framework works in the more general setting to approximately satisfy multiple knapsack constraints.*

For combinatorial algorithms, a similar framework provides a new analysis of the classical Fedorov's exchange method. The new analysis shows this simple local search algorithm works well as long as there exists an almost optimal solution with good condition number.

1.4 Beyond Spectral Rounding

In [38], together with our coauthors, we proposed to incorporate effective resistance metric into network design, as an interpolation of shortest path distance and edge-connectivity between vertices. Incorporating effective resistances can also allow one to control some natural quantities about random walks on the resulting subgraph, such as the commute time between vertices [39] and the cover time [112, 52]. We note that effective resistances have interesting connections to many other graph problems, including spectral sparsification [132], maximum flow computation [44, 109, 120], asymmetric traveling salesman problem [9], and random spanning tree generation [115, 128]. We believe that it is a useful property to be incorporated into network design. We also would like to remark that our work in [38] inspired the subsequent work on spectral rounding and applications to network design in this thesis.

In the last part of this thesis, we present our results for the s-t effective resistance network design problem, which was proposed in [38]. In this problem, we are given an input graph, two designated vertices s and t, and a non-negative integer k. The goal is to find a subgraph with at most k edges to minimize the effective resistance between s and t. The following is our main result.

Theorem (Informal)**.** *The s-t effective resistance network design problem is* NP*-hard, and there exists a randomized $O(1)$-approximation algorithm for this problem.*

It is worth pointing out that the spectral rounding technique leads to a $(1+\varepsilon)$-approximation algorithm when $k \geqslant \Omega(n/\varepsilon^2)$, where n is the number of vertices in the graph. Nevertheless, the constant approximation algorithm in the above theorem outperforms the spectral rounding based algorithm significantly when $k \ll n$.

1.5 Organization

- In Chapter 2, we introduce concepts and present preliminary results that will be used in this thesis, e.g., linear algebra, convex optimization, graph theory, electrical networks, etc. Furthermore, we survey some particularly important and relevant topics in this chapter, which include spectral sparsification and the interlacing polynomial method.

- In Chapter 3, we survey some basic concentration inequalities and prove a new concentration inequality for self-adjusting random processes, which is a key tool in this thesis.

- In Chapter 4, we provide a comprehensive review of the regret minimization framework, which is another key technique in this thesis. We first review two equivalent algorithmic frameworks, the Follow-The-Regularizer-Leader algorithm and the mirror descent method for regret minimization. Then, we derive a generic regret bound, which slightly generalizes the known bounds. Finally, we present a new randomized spectral sparsification algorithm with the regret minimization framework, which illustrates the key theme in this thesis.

- In Chapter 5, we formally formulate the spectral rounding problem, and present an iterative randomized rounding algorithm for the one-sided spectral rounding and a non-constructive proof for the two-sided spectral rounding.

- In Chapter 6, we show the applications of spectral rounding techniques to various graph problems, including generalized survivable network design, spectral network design, and additive spectral sparsification.

- In Chapter 7, we present a refined analysis of spectral rounding, and provide a unifying algorithmic framework for experimental design problems.

- In Chapter 8, we present both algorithmic and hardness results of the s-t effective resistance network design problem.

- Finally, in Chapter 9, we conclude the thesis and discuss future directions.

Chapter 2

Preliminaries

We write $\mathbb{R}$ and $\mathbb{R}_+$ as the sets of real numbers and non-negative real numbers, and $\mathbb{Z}$ and $\mathbb{Z}_+$ as the sets of integers and non-negative integers. We also write $\mathbb{C}$ as the set of complex numbers. Given a positive integer $d \geqslant 1$, we denote $[d] := \{1, \ldots, d\}$ as the set of integers from 1 to d, and denote $2^{[d]} := \{S \subseteq [d]\}$ as the power set of $[d]$. We use $\mathbb{P}[\cdot]$ to denote the probability of a random event, and $\mathbb{E}[\cdot]$ to denote the expectation of a random variable.

2.1 Linear Algebra

Throughout this thesis, we use italic sans-serif fonts for vectors and matrices, e.g., x, A, and all the vectors and matrices only have real entries.

2.1.1 Vectors

Let $\mathbb{R}^d$ denote the d-dimensional Euclidean space. Given a finite set S, let $\mathbb{R}^S$ denote the $|S|$-dimensional Euclidean space where the Cartesian coordinates are indexed by S. We write $\mathsf{1}_d$ as the d-dimensional all-one vector or simply $\mathsf{1}$ when the dimension is clear from the context. We denote $\{\mathsf{e}_1, \ldots, \mathsf{e}_d \in \mathbb{R}^d\}$ as the standard basis of the d-dimensional

Euclidean space. Given a vector $x \in \mathbb{R}^d$, we write $x(i)$ as the i-th entry of vector x, and write $x(S) := \sum_{i \in S} x(i)$ for any subset $S \subseteq [d]$. For $p \geqslant 1$, we denote

$$\|x\|_p := \left(\sum_{i=1}^{d} |x(i)|^p \right)^{\frac{1}{p}}$$

as the *ℓ_p-norm* of x. For example, $\|x\|_2$ is the Euclidean norm, and $\|x\|_\infty = \max_i |x(i)|$ is the maximum norm.

A vector $v \in \mathbb{R}^d$ is a column vector, and its transpose is denoted by $v^\top$. Given two vectors $x, y \in \mathbb{R}^d$, the *inner product* is defined as $\langle x, y \rangle := \sum_{i=1}^{d} x(i) \cdot y(i)$. The Cauchy-Schwarz inequality says that

$$\langle x, y \rangle \leqslant \|x\|_2 \cdot \|y\|_2 .$$

2.1.2 Matrices and Eigenvalues

We denote the $d \times d$ identity matrix by I_d or simply I when the dimension is clear from the context. The *inverse* of a square matrix $M \in \mathbb{R}^{d \times d}$, denoted by M^{-1}, is a square matrix in $\mathbb{R}^{d \times d}$ that satisfies $MM^{-1} = I_d$. We write $M(i,j)$ as the (i,j)-th entry of a matrix M.

Given a matrix $M \in \mathbb{R}^{d \times d}$, a nonzero vector $v \in \mathbb{C}^d$ is an *eigenvector* of M if there exists $\lambda \in \mathbb{C}$ such that $Mv = \lambda v$, where the scalar λ is known as the *eigenvalue* associated with v. Given an eigenvalue λ of a matrix $M \in \mathbb{R}^{d \times d}$, the subspace defined by $E := \{v \mid (M - \lambda I_d)v = 0\}$ is called the *eigenspace* of M associated with λ, and the dimension of E is called the *geometric multiplicity* of λ. The following is a simple but useful fact about the eigenvalues of the product of two matrices.

Lemma 2.1.1. *Let $A \in \mathbb{R}^{d_1 \times d_2}$ and $B \in \mathbb{R}^{d_2 \times d_1}$ for some integers $d_1, d_2 \geqslant 1$. Then, the matrices AB and BA have the same set of nonzero eigenvalues.*

Proof. For any nonzero eigenvalue λ of the matrix AB associated with an eigenvector $v \in \mathbb{R}^{d_1}$, we can verify that $(BA)Bv = B(AB)v = \lambda Bv$. Thus, $Bv \in \mathbb{R}^{d_2}$ is an eigenvector of BA with eigenvalue λ. The same argument applies to the nonzero eigenvalues of matrix

BA. Therefore, we can find a one-to-one mapping between the nonzero eigenvalues of AB and BA. □

Remark. *The above proof can be extended to show that the eigenvalue λ of AB and BA have the same multiplicity, but this simpler version is enough for our applications.*

A matrix $M \in \mathbb{R}^{d \times d}$ is *symmetric* if $M = M^\top$, where $M^\top$ is the transpose of M. We denote the set of all d-dimensional symmetric matrices by $\mathbb{S}^d$. It is a fundamental result that any $d \times d$ real symmetric matrix has d real eigenvalues $\lambda_1 \leqslant \ldots \leqslant \lambda_d$ and an orthonormal basis of eigenvectors, which is known as the *Spectral Theorem* for real symmetric matrices.

Theorem 2.1.2 (Spectral Theorem, see, e.g., [77])**.** *Let $M \in \mathbb{S}^d$ be a real symmetric matrix. Then, all the eigenvalues $\lambda_1, \ldots, \lambda_d$ of M are real. Furthermore, M has a set of orthonormal real eigenvectors $v_1, \ldots, v_d \in \mathbb{R}^d$ such that*

$$M = \sum_{i=1}^{d} \lambda_i \cdot v_i v_i^\top \quad \textit{and} \quad \langle v_i, v_j \rangle = 0 \textit{ for all } i \neq j \quad \textit{and} \quad \|v_i\|_2 = 1 \textit{ for all } i \in [d].$$

Let $M \in \mathbb{S}^d$ be a real symmetric matrix. We refer to $M = \sum_{i=1}^d \lambda_i \cdot v_i v_i^\top$ defined in Theorem 2.1.2 as the *eigendecomposition* of M. We write $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ as the maximum and the minimum eigenvalue of the real symmetric matrix M. The following variational characterization of eigenvalues is well-known.

Theorem 2.1.3 (see, e.g., Theorem 4.2.2 in [77])**.** *Let $M \in \mathbb{S}^d$ be a real symmetric matrix with eigendecomposition $M = \sum_{i=1}^d \lambda_i \cdot v_i v_i^\top$, where $\lambda_1 \leqslant \ldots \leqslant \lambda_d$. Then,*

$$\lambda_k = v_k^\top M v_k = \sup_{\substack{x \in \operatorname{span}\{v_1, \ldots, v_k\} \\ \text{and } \|x\|_2 = 1}} x^\top M x = \inf_{\substack{x \in \operatorname{span}\{v_k, \ldots, v_d\} \\ \text{and } \|x\|_2 = 1}} x^\top M x, \quad \textit{for all } k \in [d].$$

The following theorem shows that a factorization, which generalizes eigendecomposition of a real symmetric matrix, always exists for a general rectangular matrix.

Theorem 2.1.4 (Singular Value Decomposition (SVD), see, e.g., [77])**.** *Let $M \in \mathbb{R}^{d_1 \times d_2}$ be a rectangular matrix with rank $r \leqslant d := \min\{d_1, d_2\}$. There exist a diagonal matrix*

$\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_r)$ *where* $\sigma_1, \dots, \sigma_r > 0$, *two matrices* $U \in \mathbb{R}^{d_1 \times r}, V \in \mathbb{R}^{d_2 \times r}$ *both with orthonormal columns such that* $M = U\Sigma V^\top$. *Furthermore,* $\sigma_1, \dots, \sigma_r$ *are uniquely determined by the positive square root of the positive eigenvalues of* $MM^\top$ *(the same as* $M^\top M$*).*

Based on the SVD decomposition, we can define a generalized inversion of a general rectangular matrix.

Definition 2.1.5 (Moore–Penrose Inverse)**.** *Let* $M = U\Sigma V^\top$ *be the SVD decomposition of the matrix* $M \in \mathbb{R}^{d_1 \times d_2}$ *with rank* r*. The* Moore–Penrose inverse *(or simply* pseudoinverse*) of* M *is defined as* $M^\dagger := V\Sigma^{-1}U^\top$.

It is easy to verify that the pseudoinverse of matrix M satisfies $MM^\dagger M = M$ and $M^\dagger M M^\dagger = M^\dagger$. When $M \in \mathbb{S}^d$ is a symmetric square rank-r matrix with eigendecomposition $M = \sum_{i=1}^r \lambda_i \cdot v_i v_i^\top$, then the pseudoinverse of M can be written as $M^\dagger = \sum_{i=1}^r \frac{1}{\lambda_i} \cdot v_i v_i^\top$

2.1.3 Positive Semidefinite/Definite Matrices

Definition 2.1.6 (Positive Semidefinite Matrix)**.** *A real symmetric matrix* $M \in \mathbb{S}^d$ *is called a* positive semidefinite *(PSD) matrix, denoted as* $M \succcurlyeq 0$, *if any of the following equivalent conditions holds*

- *All the eigenvalues of* M *are non-negative.*
- *The quadratic form* $x^\top M x \geqslant 0$ *for any vector* $x \in \mathbb{R}^d$.
- *There exists a matrix* $U \in \mathbb{R}^{n \times d}$ *for some* $n \geqslant 1$ *such that* $M = U^\top U$.

The square root of a PSD matrix $M \succcurlyeq 0$ is defined as $M^{\frac{1}{2}} := \sum_i \sqrt{\lambda_i} \cdot v_i v_i^\top$. We denote $M \succ 0$ as a *positive definite* (PD) matrix M, which is real symmetric and all eigenvalues are positive. We use $A \succcurlyeq B$ to denote $A - B \succcurlyeq 0$ and $A \succ B$ to denote $A - B \succ 0$ for matrices A and B. We write $\mathbb{S}_+^d$ as the set of all d-dimensional PSD matrices, and $\mathbb{S}_{++}^d$ as the set of all d-dimensional PD matrices. The following Lemma characterizes the semidefiniteness of a blocked matrix by its Schur complement.

Lemma 2.1.7 (see, e.g., [3, 77])**.** *Let $M \in \mathbb{S}^d$ be a blocked symmetric matrix*

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}.$$

Then, $M \succcurlyeq 0$ if and only if $A \succcurlyeq 0$, $(I - AA^\dagger)B = 0$, and $C - B^\top A^\dagger B \succcurlyeq 0$. The matrix $C - B^\top A^\dagger B$ is called the generalized Schur complement *of A in M.*

The following is a useful fact that we will use in multiple occasions.

Claim 2.1.8. *Given $X \in \mathbb{S}^d_{++}$ and $Y \in \mathbb{R}^{d \times d_1}$, then $Y^\top X^{-1} Y \prec I_{d_1}$ if and only if $YY^\top \prec X$. Similarly, $Y^\top X^{-1} Y \preccurlyeq I_{d_1}$ is equivalent to $YY^\top \preccurlyeq X$.*

Proof. Since $Y^\top X^{-1} Y = Y^\top X^{-\frac{1}{2}} X^{-\frac{1}{2}} Y$, it follows from Lemma 2.1.1 that

$$\lambda_{\max}(X^{-\frac{1}{2}} YY^\top X^{-\frac{1}{2}}) = \lambda_{\max}(Y^\top X^{-1} Y).$$

Since $A \prec I$ is equivalent to $\lambda_{\max}(A) < 1$ for symmetric matrix A, the condition $Y^\top X^{-1} Y \prec I_{d_1}$ is equivalent to $X^{-\frac{1}{2}} YY^\top X^{-\frac{1}{2}} \prec I_d$, which is further equivalent to $YY^\top \prec X$. The second part of the claim follows by the same argument. □

Given two real matrices A and B of the same size, the *Frobenius inner product* of A, B is denoted as $\langle A, B \rangle := \sum_{i,j} A(i,j) \cdot B(i,j)$. The following is a standard fact.

Fact 2.1.9. *Let $A \succcurlyeq 0$ and $B \succcurlyeq C \succcurlyeq 0$, then it holds that $\langle A, B \rangle \geqslant 0$ and $\langle A, B \rangle \geqslant \langle A, C \rangle$.*

We write $\|M\|_{\mathrm{F}} := \sqrt{\langle M, M \rangle}$ as the *Frobenius norm* of a matrix M, and write $\|M\|_{\mathrm{op}} := \max_{\|x\|_2 = 1} \|Mx\|_2$ as the *operator norm* of a matrix M. For symmetric matrices, the operator norm is just the largest absolute value of its eigenvalues. For positive semidefinite matrices, the operator norm is just its largest eigenvalue.

2.1.4 Trace and Determinant

The *trace* of a matrix $M \in \mathbb{R}^{d \times d}$, denoted by $\operatorname{tr}(M)$, is defined as the sum of the diagonal entries of M. We can check by definition that trace satisfies *cyclic property*, i.e. $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ for A and B with appropriate sizes. We also note that $\langle A, B \rangle = \operatorname{tr}(A^\top B)$ for A and B of the same size. A PSD matrix $M \succcurlyeq 0$ with $\operatorname{tr}(M) = 1$ is called a *density matrix*. We denote $\Delta^d := \{M \succcurlyeq 0 \mid \sum_{i=1}^d \lambda_i(M) = 1\}$ as the set of all d-dimensional density matrices.

The *determinant* of a matrix $M \in \mathbb{R}^{d \times d}$, denoted by $\det(M)$, is defined as $\det(M) := \sum_{\sigma \in S_d} \operatorname{sgn}(\sigma) \prod_{i=1}^d M(i, \sigma(i))$, where S_d is the set of all permutations on $[d]$ and $\operatorname{sgn}(\sigma)$ is the signature of σ. It is well-known that both trace and determinant are related to the eigenvalues of a matrix. In particular,

$$\operatorname{tr}(M) = \sum_{i=1}^d \lambda_i(M) \qquad \text{and} \qquad \det(M) = \prod_{i=1}^d \lambda_i(M),$$

where $\lambda_i(M)$ denotes the i-th eigenvalue of M.

The following is a simple claim that bounds the trace of the square root of a density matrix, which will be invoked multiple times in this thesis.

Claim 2.1.10. *For any $d \times d$ matrix $A \succcurlyeq 0$ satisfying $\operatorname{tr}(A) = 1$, $\operatorname{tr}\left(A^{\frac{1}{2}}\right) \leqslant \sqrt{d}$.*

Proof. Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of A. It holds that

$$\operatorname{tr}\left(A^{\frac{1}{2}}\right) = \sum_{i=1}^d \sqrt{\lambda_i} \leqslant \sqrt{d} \cdot \sqrt{\sum_{i=1}^d \lambda_i} = \sqrt{d},$$

where the inequality follows by Cauchy-Schwartz, and the last equality follows by the assumption $\operatorname{tr}(A) = 1$. □

The following two lemmas describe the change of the determinant function under rank-one and rank-two updates. The first one is the well-known matrix determinant lemma.

Lemma 2.1.11 (Matrix Determinant Lemma, see, e.g., [77]). *For any invertible matrix X and any vector $v \in \mathbb{R}^d$,*

$$\det(X \pm vv^\top) = \det(X)(1 \pm \langle vv^\top, X^{-1} \rangle).$$

The following determinant lower bound under a rank-two update is a simple consequence and was implicitly contained in [108]. We provide a proof for completeness.

Lemma 2.1.12. *Given a matrix $A \succ 0$ and two vectors $u, v \in \mathbb{R}^d$, if $\langle uu^\top, A^{-1} \rangle \leqslant 1$, then*

$$\det(A - uu^\top + vv^\top) \geqslant \det(A) \left(1 - \langle uu^\top, A^{-1} \rangle\right) \left(1 + \langle vv^\top, A^{-1} \rangle\right).$$

Proof. We first consider the case when $\langle uu^\top, A^{-1} \rangle < 1$. This is equivalent to $uu^\top \prec A$ by Claim 2.1.8. Thus $A - uu^\top \succ 0$. Applying Lemma 2.1.11 twice,

$$\begin{aligned} \det(A - uu^\top + vv^\top) &= \det(A) \cdot \left(1 - \langle uu^\top, A^{-1} \rangle\right) \cdot \left(1 + \langle vv^\top, (A - uu^\top)^{-1} \rangle\right) \\ &\geqslant \det(A) \left(1 - \langle uu^\top, A^{-1} \rangle\right) \left(1 + \langle vv^\top, A^{-1} \rangle\right), \end{aligned}$$

where the last inequality holds as $0 \prec A - uu^\top \preccurlyeq A$.

In the case when $\langle uu^\top, A^{-1} \rangle = 1$, the RHS of the lemma is zero. Claim 2.1.8 implies $A - uu^\top \succcurlyeq 0$, thus it follows that the LHS is non-negative. □

2.1.5 Matrix Inversion with Perturbations

The following lemmas describe the change of matrix inversion under rank-one and general perturbations.

Lemma 2.1.13 (Sherman-Morrison Formula [129])**.** *Suppose $A \in \mathbb{R}^{d \times d}$ is an invertible matrix, and $u, v \in \mathbb{R}^d$. Then, $A + uv^\top$ is invertible if and only if $1 + v^\top A^{-1} u \neq 0$, and under this case*

$$\left(A + uv^\top\right)^{-1} = A^{-1} - \frac{A^{-1} uv^\top A^{-1}}{1 + v^\top A^{-1} u}.$$

The following is a corollary of Sherman-Morrison formula, which follows by restricting to the eigenspaces associated with nonzero eigenvalues of the matrix A.

Corollary 2.1.14. *Suppose we are given a symmetric matrix $A \in \mathbb{S}^d$, a vector $v \in \mathbb{R}^d$ that lives in the eigenspaces associated with nonzero eigenvalues of A, and a number $c \in \mathbb{R}$. If $1 + c \cdot v^\top A^\dagger v \neq 0$, then*

$$\left(A + c \cdot vv^\top\right)^\dagger = A^\dagger - \frac{c \cdot A^\dagger vv^\top A^\dagger}{1 + c \cdot v^\top A^\dagger v}.$$

Lemma 2.1.15 (Woodbury Matrix Identity [143])**.** *Suppose $A \in \mathbb{R}^{d\times d}$ and $C \in \mathbb{R}^{k\times k}$ are two invertible matrices, and $U \in \mathbb{R}^{d\times k}$, $V \in \mathbb{R}^{k\times d}$. Then, $A + UCV$ is invertible if and only if $C^{-1} + VA^{-1}U$ is also invertible, and under this case*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

2.2 Convex Analysis and Convex Optimization

In this section, we review some basic notions and facts in convex analysis and convex optimization that we will see throughout this thesis. Most of the contents follow the exposition of two textbooks, by Boyd and Vandenberghe [31] and by Hiriart-Urruty and Lemaréchal [76]. We include some short proofs for completeness.

2.2.1 Differentiation in Euclidean Space

Given a point $x \in \mathbb{R}^d$, the *ε-neighbourhood* around x is defined as $\{y \in \mathbb{R}^d \mid \|x - y\|_2 < \varepsilon\}$ for $\varepsilon > 0$. Given a set $\mathcal{S} \subseteq \mathbb{R}^d$ and a point $x \in \mathcal{S}$, we say x is in the *interior* of $\mathcal{S}$ if there exists an $\varepsilon > 0$ such that the ε-neighbourhood around x is contained in $\mathcal{S}$. We denote $\mathrm{int}(\mathcal{S})$ as the set of all interior points of $\mathcal{S}$. A set $\mathcal{S} \subseteq \mathbb{R}^d$ is *open* if and only if $\mathcal{S} = \mathrm{int}(\mathcal{S})$.

Let $\boldsymbol{f} : \mathcal{D} \to \mathbb{R}^m$ be a vector-valued function defined on an open domain $\mathcal{D} \subseteq \mathbb{R}^d$. We say $\boldsymbol{f}$ is *continuous* at $x \in \mathcal{D}$ if

$$\lim_{h\to 0} \|\boldsymbol{f}(x + h) - \boldsymbol{f}(x)\|_2 = 0.$$

We denote $\mathrm{L}(\mathbb{R}^d, \mathbb{R}^m)$ as the set of all linear maps from $\mathbb{R}^d$ to $\mathbb{R}^m$. Every linear map in $\mathrm{L}(\mathbb{R}^d, \mathbb{R}^m)$ can be uniquely represented by a matrix in $\mathbb{R}^{m\times d}$.

We say $\boldsymbol{f}$ is *differentiable* at $x \in \mathcal{D}$ if there exists a (necessarily unique) linear map $J \in \mathrm{L}(\mathbb{R}^d, \mathbb{R}^m)$ such that

$$\lim_{h\to 0} \frac{\|\boldsymbol{f}(x + h) - \boldsymbol{f}(x) - J(h)\|_2}{\|h\|_2} = 0. \tag{2.1}$$

Note that the Euclidean norms in the denominator and numerator are defined in different dimensions, and we can replace the Euclidean norms by any other norms in finite dimensional spaces. The unique linear map J is called the *differential* of $\boldsymbol{f}$ at x, which is also known as the *Jacobian operator* of $\boldsymbol{f}$ at x. We denote the differential of $\boldsymbol{f}$ at x by $\boldsymbol{f}'(x)$, or by $\frac{\mathrm{d}}{\mathrm{d}x}\boldsymbol{f}(x)$ when we want to specify the varying variable. Note that $\boldsymbol{f}'(x)$ is a linear map from $\mathbb{R}^d$ to $\mathbb{R}^m$, which can be uniquely represented by a *Jacobian matrix* in $\mathbb{R}^{m\times d}$. Thus, $\boldsymbol{f}'$ can be treated as a function that maps from $\mathbb{R}^d$ to $\mathbb{R}^{m\times d}$. We say the differential $\boldsymbol{f}'(x)$ exists in $\mathcal{S} \subseteq \mathcal{D}$ if $\boldsymbol{f}'(x)$ exists and there exists some $\varepsilon > 0$ such that the ε-neighbourhood around x is contained in $\mathcal{S}$.

We say $\boldsymbol{f}$ is *continuously differentiable* at $x \in \mathcal{D}$ if $\boldsymbol{f}$ is differentiable at x and the function defined by the differentials $\boldsymbol{f}' : \mathbb{R}^d \to \mathrm{L}(\mathbb{R}^d, \mathbb{R}^m)$ is continuous at x.

The above definitions for vector-valued functions can be naturally generalized to matrix-valued functions with matrix domains, where the inner product is replaced by Frobenius inner product, and the Euclidean norm is replaced by Frobenius norm.

For a real function $f : \mathcal{D} \to \mathbb{R}$ defined on a domain $\mathcal{D} \subseteq \mathbb{R}^d$, the differential of f at x, i.e. $f'(x)$, can be uniquely represented by a column vector in $\mathbb{R}^d$. We refer to this vector as the *gradient* of f at x, and denote it by $\nabla f(x)$. Note that $f' : \mathbb{R}^d \to \mathrm{L}(\mathbb{R}^d, \mathbb{R})$ is a function that maps a vector $x \in \mathcal{D}$ to the differential of f at x.

Let $x \in \mathcal{D}$ be some point where f is finite. The *directional derivative* of f at x with respect to some direction $\boldsymbol{d} \in \mathbb{R}^d$ is defined as

$$f'(x; \boldsymbol{d}) := \lim_{\lambda \to 0} \frac{f(x + \lambda \boldsymbol{d}) - f(x)}{\lambda}.$$

As a special case, we denote $\partial_j f(x) := f'(x; \mathsf{e}_j)$ as the *partial derivative* of f with respect to the variable at the j-th coordinate. We say the directional derivative $f'(x; \boldsymbol{d})$ exists in $\mathcal{S} \subseteq \mathcal{D}$, if there exists $\varepsilon > 0$ such that an ε-neighbourhood of x along the direction $\boldsymbol{d}$ is contained in $\mathcal{S}$.

If f is differentiable at x, then all directional derivatives and partial derivatives of f at x exist. The gradient can be written as a column vector $\nabla f(x) = (\partial_1 f(x), \ldots, \partial_d f(x))^\top$, and $f'(x; \boldsymbol{d}) = \langle \nabla f(x), \boldsymbol{d} \rangle$.

It is a well-known fact in analysis that, if all the partial derivatives exist and are continuous, then the function is continuously differentiable.

Theorem 2.2.1 (see, e.g., [127])**.** *Let f be a function that maps $\mathcal{D}$ in $\mathbb{R}^d$ to $\mathbb{R}$. Then, f is continuously differentiable at $\mathsf{x} \in \mathcal{D}$ if and only if the partial derivative $\partial_j f$ exists and is continuous at x for all $j \in [d]$. Furthermore, the statement still holds if we replace the partial derivatives $\{\partial_j f(\mathsf{x})\}_{j\in[d]}$ by a collection of directional derivatives $\{f'(\mathsf{x}, \mathsf{d}_i)\}_{i\in[d]}$, where $\mathsf{d}_1, \dots, \mathsf{d}_d$ form a basis of $\mathbb{R}^d$.*

We will need the gradients of the following functions in this thesis. We provide proofs in Appendix A.1 for completeness.

Fact 2.2.2. *Let $f : \mathbb{S}^d_{++} \to \mathbb{R}$ be defined as $f(X) = \log\det(X)$. Then, f is differentiable at any $X \succ 0$ with $\nabla f(X) = X^{-1}$.*

Fact 2.2.3. *Let $f : \mathbb{S}^d_{++} \to \mathbb{R}$ be defined as $f(X) = \operatorname{tr}(X^{-1})$. Then, f is differentiable at any $X \succ 0$ with $\nabla f(X) = -X^{-2}$.*

Fact 2.2.4. *Let $f : \mathbb{S}^d_{+} \to \mathbb{R}$ be defined as $f(X) = \operatorname{tr}(X^{\frac{1}{2}})$. Then, f is differentiable at any $X \succ 0$ with $\nabla f(X) = \frac{1}{2}X^{-\frac{1}{2}}$.*

For a positive definite matrix $X \in \mathbb{S}^d_{++}$ with eigendecomposition $X = U\Lambda U^\top$ where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$, we define $\log X := U \operatorname{diag}(\log\lambda_1, \dots, \log\lambda_d) U^\top$.

Fact 2.2.5. *Let $f : \mathbb{S}^d_{+} \to \mathbb{R}$ be defined as $f(X) = \langle X, \log X - I_d\rangle$. Then, f is differentiable at any $X \succ 0$ with $\nabla f(X) = \log X$.*

A real function f is *twice differentiable* at x if f is continuously differentiable at x and the first order differential $f' : \mathbb{R}^d \to \mathrm{L}(\mathbb{R}^d, \mathbb{R})$ is differentiable at x. By the definition of differentiability, there exists a unique linear map $H \in \mathrm{L}(\mathbb{R}^d, \mathrm{L}(\mathbb{R}^d, \mathbb{R}))$ that satisfies (2.1) with $\boldsymbol{f} = f'$ and $J = H$. We call this unique linear map H the *Hessian* of f at x. When f is twice differentiable at x, the Hessian of f at x can be represented by a symmetric matrix in $\mathbb{S}^d$ with the second order partial derivative $\partial^2_{i,j} f(x)$ at the (i,j)-th entry. We call this matrix the *Hessian matrix* of f at x and denote it by $\nabla^2 f(x)$.

A real function $f : \mathcal{D} \to \mathbb{R}$ is a *continuous (differentiable, continuously differentiable, twice differentiable) function* if f is continuous (differentiable, continuously differentiable, twice differentiable) on the whole domain $\mathcal{D}$.

2.2.2 Convex Set and Convex Function

A set $\mathcal{C}$ is *convex* if, for any two points $x, y \in \mathcal{C}$, the whole line segment between x and y, i.e. $\lambda x + (1-\lambda)y \in \mathcal{C}$ for all $0 \leqslant \lambda \leqslant 1$, is contained in $\mathcal{C}$.

A function $f : \mathcal{D} \to \mathbb{R}$ is a *convex function*, if the domain $\mathcal{D}$ is convex and for any $x, y \in \mathcal{D}$ and $\lambda \in [0,1]$ the following inequality holds

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y). \tag{2.2}$$

Note that it suffices to check the mid-point $\lambda = \frac{1}{2}$ to establish convexity for continuous function f. The well-known *Jensen's inequality* follows from the convexity of a function immediately.

Lemma 2.2.6 (Jensen's Inequality, see, e.g., [31])**.** *Let $f : \mathcal{D} \to \mathbb{R}$ be a convex function, $x_1, \ldots, x_n \in \mathcal{D}$ be n points in the domain of f. For any $\lambda_1, \ldots, \lambda_n \geqslant 0$ with $\sum_{i=1}^n \lambda_i = 1$, it holds that,*

$$f\left(\sum_{i=1}^n \lambda_i x_i\right) \leqslant \sum_{i=1}^n \lambda_i f(x_i).$$

There is a geometric view of the convex functions through the notion of epigraph. Let $f : \mathcal{D} \to \mathbb{R}$ be a function defined on the domain $\mathcal{D} \subseteq \mathbb{R}^d$. The *epigraph* of a function f is defined as $\operatorname{epi}(f) := \{(x, u) \mid x \in \mathcal{D}, u \geqslant f(x)\} \subseteq \mathbb{R}^{d+1}$. The following simple fact follows from definition directly, but insightfully connects convex functions with convex sets.

Lemma 2.2.7 (see, e.g., [31])**.** *A function f is convex if and only if the epigraph* epi(f) *is a convex set.*

A function $f : \mathcal{D} \to \mathbb{R}$ is a *strictly convex function*, if $\mathcal{D}$ is convex and (2.2) holds with strict inequality

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y) \quad \text{for all } x \neq y \in \mathcal{D} \text{ and } 0 < \lambda < 1.$$

There is an equivalent way to define convexity (or strict convexity) by restricting to a single variable function.

Lemma 2.2.8 (see, e.g., [31])**.** *A function $f : \mathcal{D} \to \mathbb{R}$ is a (strictly) convex function if and only if for any $x, y \in \mathcal{D}$, the single variable function $g_{x,y}(t) := f(x + t(y - x))$ is (strictly) convex on $[0, 1]$.*

The following is a necessary and sufficient condition known as the *first order condition* for a differentiable function f being (strictly) convex.

Lemma 2.2.9 (First Order Condition, see, e.g., [31])**.** *A differentiable function $f : \mathcal{D} \to \mathbb{R}$ with a convex domain $\mathcal{D}$ is convex if and only if $f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle$ for all $x, y \in \mathcal{D}$. The function f is strictly convex if and only if $f(y) > f(x) + \langle \nabla f(x), y - x\rangle$ for all $x \neq y \in \mathcal{D}$.*

When f is not fully differentiable but the directional differential derivative exists at some point with respect to some direction, we still have a similar necessary condition.

Lemma 2.2.10 (see, e.g., [31])**.** *Let $f : \mathcal{D} \to \mathbb{R}$ be a convex function. If the directional derivative of f at $x \in \mathcal{D}$ with respect to a direction $d \in \mathbb{R}^d$ exists, then for any t such that $y = x + td \in \mathcal{D}$ it holds that $f(y) \geqslant f(x) + tf'(x; d)$.*

The following is a necessary and sufficient condition known as the *second order condition* for a twice differentiable function f being convex.

Lemma 2.2.11 (Second Order Condition, see, e.g., [31])**.** *Let $f : \mathcal{D} \to \mathbb{R}$ be a twice differentiable function on a convex domain $\mathcal{D}$. If the Hessian matrix $\nabla^2 f(x) \succcurlyeq 0$ for all $x \in \mathcal{D}$, then f is convex. Conversely, when the domain $\mathcal{D}$ is open, f is convex implies that the Hessian matrix $\nabla^2 f(x) \succcurlyeq 0$ for all $x \in \mathcal{D}$.*

Remark. *In the necessity of the second order condition, the openness assumption of the domain $\mathcal{D}$ is important. Consider a two variable function $f(x,y) = x^2 - y^2$ with domain $\mathcal{D} = \mathbb{R} \times \{0\}$. The Hessian matrix $\nabla^2 f(x,y) = \left(\begin{smallmatrix} 2 & 0 \\ 0 & -2 \end{smallmatrix}\right)$ is nowhere positive semidefinite, but the function f is convex on the domain $\mathcal{D}$.*

The following is a sufficient condition for f being strictly convex. However, we remark that this condition is not a necessary condition for strict convexity.

Lemma 2.2.12 (see, e.g., [31]). *Given a function $f : \mathcal{D} \to \mathbb{R}$, if the domain $\mathcal{D}$ is convex and the Hessian matrix $\nabla^2 f(\mathsf{x}) \succ 0$ for all $\mathsf{x} \in \mathcal{D}$, then f is strictly convex.*

A function $f : \mathcal{D} \to \mathbb{R}$ is *concave/strictly concave* if $-f$ is convex/strictly convex. The following fact says that taking pointwise infimum over a family of concave functions preserves concavity.

Lemma 2.2.13 (see, e.g., [31]). *Let $f : \mathcal{D}_{\mathcal{X}} \times \mathcal{D}_{\mathcal{Y}} \to \mathbb{R}$ be a function such that $f(\mathsf{x}, \mathsf{y})$ is concave in x for any given $\mathsf{y} \in \mathcal{D}_{\mathcal{Y}}$, where the domain $\mathcal{D}_{\mathcal{Y}}$ could be an infinite set. Then, the function $g : \mathcal{D}_{\mathcal{X}} \to \mathbb{R}$ defined by $g(\mathsf{x}) = \inf_{\mathsf{y} \in \mathcal{D}_{\mathcal{Y}}} f(\mathsf{x}, \mathsf{y})$ is concave.*

The following lemma is a simple corollary.

Lemma 2.2.14. *Let $\mathsf{M} \in \mathbb{S}^d$ be a real symmetric matrix. The function of the smallest eigenvalue $\lambda_{\min}(\mathsf{M})$ is a concave function on $\mathbb{S}^d$, and the function of the largest eigenvalue $\lambda_{\max}(\mathsf{M})$ is a convex function on $\mathbb{S}^d$.*

Proof. By Theorem 2.1.3, we can write the smallest eigenvalue of $\mathsf{M} \in \mathbb{S}^d$ as

$$\lambda_{\min}(\mathsf{M}) = \inf_{x \in \mathbb{R}^d \text{ and } \|\mathsf{x}\|_2 = 1} \langle \mathsf{x}\mathsf{x}^\top, \mathsf{M} \rangle,$$

which is the point-wise infimum of a family of linear functions in M. By Lemma 2.2.13, $\lambda_{\min}(\mathsf{M})$ is concave in M. The convexity of $\lambda_{\max}(\mathsf{M})$ follows from similar argument. □

Examples of Convex/Concave Functions

The following are two well-known concave functions. We provide proofs in Appendix A.2 for completeness.

Fact 2.2.15. *The function $f(X) = \log\det(X)$ is concave on $\mathbb{S}^d_{++}$.*

Fact 2.2.16. *The function $f(X) = \det(X)^{\frac{1}{d}}$ is concave on $\mathbb{S}^d_{+}$.*

Then, we show convexity/concavity of some trace related functions. We follow an idea from the lecture notes of Lee [99], which lifts the convexity/concavity of univariate functions to trace related functions with a matrix domain.

Given an interval $I \subseteq \mathbb{R}$, define a set of symmetric matrices

$$\mathbb{S}^d_I := \{X \in \mathbb{S}^d \mid \text{all eigenvalues of } X \text{ belong to } I.\}$$

By Lemma 2.2.14, $\lambda_{\max}$ is convex and $\lambda_{\min}$ is concave. Thus, we can verify that $\mathbb{S}^d_I$ is a convex set.

Then, we lift a function $g : I \to \mathbb{R}$ to a function that maps $\mathbb{S}^d_I$ to $\mathbb{R}$ as follows. For any matrix $X \in \mathbb{S}^d_I$ with eigendecomposition $X = U\Lambda U^\top$ where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$, we define

$$g(X) := U \begin{pmatrix} g(\lambda_1) & & \\ & \ddots & \\ & & g(\lambda_d) \end{pmatrix} U^\top.$$

Lemma 2.2.17. *For any interval $I \subseteq \mathbb{R}$, if a function $g : I \to \mathbb{R}$ is (strictly) convex on I, then the function $X \mapsto \operatorname{tr}(g(X))$ is (strictly) convex on $\mathbb{S}^d_I$.*

Proof. We prove the convexity, and the strict convexity follows by the same argument.

For any $X, Y \in \mathbb{S}^d_I$, we are going to show

$$\operatorname{tr}\Big(f(\lambda X + (1-\lambda)Y)\Big) \leqslant \lambda \operatorname{tr}\Big(f(X)\Big) + (1-\lambda)\operatorname{tr}\Big(f(Y)\Big). \tag{2.3}$$

Let $\{u_i\}$ be the orthonormal basis of the eigenvectors of $\lambda X + (1-\lambda)Y$. The i-th eigenvalue of $\lambda X + (1-\lambda)Y$ is equal to $\langle u_i, (\lambda X + (1-\lambda)Y)u_i\rangle$ by Theorem 2.1.3. Thus,

$$\begin{aligned}
\operatorname{tr}\Big(f(\lambda X + (1-\lambda)Y)\Big) &= \sum_{i=1}^{d} f(\langle u_i, (\lambda X + (1-\lambda)Y)u_i\rangle) \\
&= \sum_{i=1}^{d} f(\lambda\langle u_i, Xu_i\rangle + (1-\lambda)\langle u_i, Yu_i\rangle) \\
&\leqslant \sum_{i=1}^{d} \Big(\lambda f(\langle u_i, Xu_i\rangle) + (1-\lambda) f(\langle u_i, Yu_i\rangle)\Big),
\end{aligned}$$

where the last inequality follows by the convexity of $f : I \to \mathbb{R}$.

Let $\{v_j\}$ be the orthonormal basis of the eigenvectors of X. For any u_i, we can write it as $u_i = (\sum_{j=1}^{d} v_j v_j^\top)u_i = \sum_{j=1}^{d}\langle v_j, u_i\rangle v_j$. This implies that $\langle u_i, Xu_i\rangle = \sum_{j=1}^{d}\langle v_j, u_i\rangle^2\langle v_j, Xv_j\rangle$ since $\{v_j\}$ are eigenvectors of X.

Using the fact that u_i is a unit vector and $\{v_j\}$ is an orthonormal basis, it follows that

$$\sum_{j=1}^{d}\langle v_j, u_i\rangle^2 = u_i^\top\left(\sum_{j=1}^{d} v_j v_j^\top\right)u_i = \|u_i\|_2^2 = 1. \tag{2.4}$$

Since $f : I \to \mathbb{R}$ is a convex function, we apply Jensen's inequality Lemma 2.2.6 to show

$$f(\langle u_i, Xu_i\rangle) = f\Big(\sum\nolimits_{j=1}^{d}\langle v_j, u_i\rangle^2\langle v_j, Xv_j\rangle\Big) \leqslant \sum_{j=1}^{d}\langle v_j, u_i\rangle^2 f(\langle v_j, Xv_j\rangle).$$

This implies that

$$\sum_{i=1}^{d}\lambda f(\langle u_i, Xu_i\rangle) \leqslant \lambda\sum_{j=1}^{d}\Big(f(\langle v_j, Xv_j\rangle)\cdot\sum_{i=1}^{d}\langle v_j, u_i\rangle^2\Big) = \lambda\operatorname{tr}\Big(f(X)\Big),$$

where we used $\sum_{i=1}^{d}\langle v_j, u_i\rangle^2 = \|v_j\|_2^2 = 1$ (follows similarly as (2.4)). Applying the same argument, we can upper bound the term $(1-\lambda)\sum_{i=1}^{d} f(\langle u_i, Yu_i\rangle)$ by $(1-\lambda)\operatorname{tr}\big(f(Y)\big)$, thus the inequality (2.3) follows. □

We can use the second order condition Lemma 2.2.12 to check that $x \mapsto x^{-1}$ is strictly convex on $(0, +\infty)$, $x \mapsto -2\sqrt{x}$ is strictly convex on $[0, +\infty)$, and $x \log x - x$ is strictly convex on $[0, +\infty)$. Therefore, the following facts follow by Lemma 2.2.17.

Fact 2.2.18. *The function $f(X) = \operatorname{tr}(X^{-1})$ is strictly convex on $\mathbb{S}^d_{++}$.*

Fact 2.2.19. *The function $f(X) = -2\operatorname{tr}(X^{\frac{1}{2}})$ is strictly convex on $\mathbb{S}^d_{+}$.*

Fact 2.2.20. *The function $f(X) = \langle X, \log X - I_d \rangle = \operatorname{tr}(X \log X - X)$ is strictly convex on $\mathbb{S}^d_{+}$.*

2.2.3 Optimality Conditions in Convex Optimization

The following is a general formulation of an optimization problem.

$$\begin{aligned} \mathsf{opt}_P := \underset{x \in \mathcal{D}}{\text{minimize}} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leqslant 0, \quad i \in \{1, \dots, m\}, \\ & h_j(x) = 0, \quad j \in \{1, \dots, p\}, \end{aligned} \tag{2.5}$$

where $f_0, f_1, \dots, f_m, h_1, \dots, h_p$ are functions defined over domain $\mathcal{D} \subseteq \mathbb{R}^d$. We denote the set of all feasible solutions by

$$\mathcal{X} := \{x \in \mathcal{D} : f_i(x) \leqslant 0, \forall i = 1, \dots, m \ \text{ and } \ h_j(x) = 0, \forall j = 1, \dots, p\}.$$

The following simple fact gives a necessary condition for a solution $x^* \in \mathbb{R}^d$ being optimal based on directional derivatives.

Lemma 2.2.21. *Given the general optimization program (2.5) with a feasible set $\mathcal{X}$, let $x^* \in \mathcal{X}$ be an optimal solution. If there exists some direction $d \in \mathbb{R}^d$ such that the directional derivative $f_0'(x^*; d)$ exists in $\mathcal{X}$, then the directional derivative $f_0'(x^*; d) = 0$.*

The lemma holds since if $f_0'(x^*; d) \neq 0$ then moving x^* to $x^* - \varepsilon f_0'(x^*) d$ would decrease the value of the objective function. Note that we do not assume convexity or differentiability in general.

We say (2.5) is a *convex program* for a convex optimization problem if the objective function f_0 and inequality constraints $f_1, \ldots, f_m$ are convex functions, and equality constraints $h_1, \ldots, h_p$ are affine functions in the form of $\langle \mathsf{a}_j, \mathsf{x} \rangle - b_j$ for all $j = 1, \ldots, p$. In the following, we assume (2.5) is a convex program with some special cases being explicitly mentioned.

The following is a necessary and sufficient condition for a point in the domain being an optimal solution.

Lemma 2.2.22 (see, e.g., [31]). *Suppose f_0 is a differentiable convex objective function, then $\mathsf{x} \in \mathcal{X}$ is an optimal solution to the convex optimization problem* (2.5) *if and only if*

$$\langle \nabla f_0(\mathsf{x}), \mathsf{y} - \mathsf{x} \rangle \geqslant 0 \qquad \text{for all } \mathsf{y} \in \mathcal{X}. \tag{2.6}$$

Proof. Since f_0 is a differentiable convex function, by the first order condition Lemma 2.2.9, for any $\mathsf{x}, \mathsf{y} \in \mathcal{X}$, it holds that

$$f_0(\mathsf{y}) \geqslant f_0(\mathsf{x}) + \langle \nabla f_0(\mathsf{x}), \mathsf{y} - \mathsf{x} \rangle.$$

Thus, if $\mathsf{x} \in \mathcal{X}$ satisfies (2.6), then for any $\mathsf{y} \in \mathcal{X}$ it holds that $f_0(\mathsf{y}) \geqslant f_0(\mathsf{x})$, which shows x is an optimal solution to (2.5).

Conversely, if x is an optimal solution but (2.6) does not hold, then there exist some $\mathsf{y} \in \mathcal{X}$ such that $\langle \nabla f_0(\mathsf{x}), \mathsf{y} - \mathsf{x} \rangle < 0$. Let $g(t) = f_0(\mathsf{z}_t)$, where $\mathsf{z}_t = \mathsf{x} + t(\mathsf{y} - \mathsf{x})$ for $t \in [0, 1]$. Note that $g'(t) \mid_{t=0} = \langle \nabla f_0(\mathsf{x}), \mathsf{y} - \mathsf{x} \rangle < 0$. Thus, for small enough $0 \leqslant t \leqslant 1$, it holds that $g(t) < g(0)$, which is equivalent to $f_0(\mathsf{z}_t) < f_0(\mathsf{x})$. Since both $\mathsf{x}, \mathsf{y} \in \mathcal{X}$, z_t is also in $\mathcal{X}$ by the convexity of $\mathcal{X}$. Thus, we have a feasible solution z_t with strictly smaller objective value than that of x, a contradiction. □

Corollary 2.2.23. *Suppose f_0 is convex and differentiable at $\mathsf{x} \in \mathcal{X}$ with $\nabla f_0(\mathsf{x}) = 0$, then x is an optimal solution to* (2.5).

Corollary 2.2.24. *Suppose f_0 is a differentiable convex function and the feasible set $\mathcal{X}$ is open. Then, $\mathsf{x} \in \mathcal{X}$ is an optimal solution to the convex optimization problem* (2.5) *if and only if $\nabla f_0(\mathsf{x}) = 0$.*

Proof. Suppose $x \in \mathcal{X}$ is an optimal solution and $\nabla f_0(x) \neq 0$. Since $\mathcal{X}$ is open, the point $y = x - \varepsilon \nabla f_0(x)$ is still in $\mathcal{X}$ for small enough $\varepsilon \geqslant 0$. However, $\langle \nabla f_0(x), y - x \rangle = -\varepsilon \left\| \nabla f_0(x) \right\|_2^2 < 0$, which implies x is not optimal by Lemma 2.2.22, a contradiction. The other direction of the corollary follows from Corollary 2.2.23. □

2.2.3.1 Lagrangian Duality

Given the convex optimization problem in (2.5), which is referred by the *primal problem*, we introduce a dual variable $\lambda_i \geqslant 0$ for each inequality constraint $f_i(x) \leqslant 0$ for all $i = 1, \dots, m$. We also introduce a dual variable $\nu_j \in \mathbb{R}$ for each affine constraint $h_j(x) = 0$ for all $j = 1, \dots, p$. Define the *Lagrangian function* $L : \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ as

$$L(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x).$$

The *Lagrange dual function*, or simply *dual function*, $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is defined as

$$g(\lambda, \nu) := \inf_{x \in \mathcal{D}} L(x, \lambda, \nu).$$

The *dual program* of the convex program in (2.5) is defined as

$$\begin{aligned} \mathsf{opt}_D := \ & \underset{\lambda \in \mathbb{R}^m, \nu \in \mathbb{R}^p}{\text{maximize}} & & g(\lambda, \nu) \\ & \text{subject to} & & \lambda \geqslant 0. \end{aligned} \tag{2.7}$$

Notice that the following *weak duality* property always holds regardless the problem is convex or not:

$$\mathsf{opt}_D = \max_{\lambda \geqslant 0, \nu \in \mathbb{R}^p} \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \leqslant \inf_{x \in \mathcal{X}} f_0(x) = \mathsf{opt}_P,$$

where we recall that $\mathcal{X}$ is the feasible set of the primal problem (2.5). The inequality holds as, for any fixed $\lambda \geqslant 0$ and ν, we have $\lambda_i f_i(x) \leqslant 0, \forall i \in [m]$ and $\nu_j h_j(x) = 0, \forall j \in [p]$ for all feasible $x \in \mathcal{X}$, which implies that $L(x, \lambda, \nu) \leqslant f_0(x)$ for $x \in \mathcal{X} \subseteq \mathcal{D}$.

Weak duality shows that the *duality gap*, i.e. opt_P–opt_D, is nonnegative. We refer to the property that the primal and dual problem have zero duality gap by *strong duality*, which does not necessarily hold in general. It usually holds for convex programs, but requires some additional conditions on the constraints.

Before we formally state one of the conditions, we first introduce two related notions. Given a set $\mathcal{C}$, the *affine hull* of $\mathcal{C}$ is defined by

$$\operatorname{aff}(\mathcal{C}) := \left\{ \sum_{i=1}^{k} \lambda_i \mathsf{x}_i \;\middle|\; k \geqslant 1; \mathsf{x}_1, \ldots, \mathsf{x}_k \in \mathcal{C}; \sum_{i=1}^{k} \lambda_i = 1 \right\}.$$

Given a set $\mathcal{C}$, the *relative interior* of $\mathcal{C}$ is defined by

$$\operatorname{relint}(\mathcal{C}) := \{\mathsf{x} \in \mathcal{C} \mid \exists r > 0, B(\mathsf{x}, r) \cap \operatorname{aff}(\mathcal{C}) \subseteq \mathcal{C}\},$$

where $B(\mathsf{x}, r) := \{\mathsf{y} \mid \|\mathsf{y} - \mathsf{x}\|_2 \leqslant r\}$ is a ball around x with radius r. The following is a commonly used condition to guarantee the strong duality for convex programs.

Definition 2.2.25 (Slater's Condition)**.**

$$\exists \mathsf{x} \in \operatorname{relint}(\mathcal{D}) \quad \textit{such that} \quad f_i(\mathsf{x}) < 0, \quad \forall i = 1, \ldots, m \quad \textit{and} \quad h_j(\mathsf{x}) = 0, \quad j = 1, \ldots, p.$$

We state the following theorem without proof, and refer the readers to Section 5.3.2 of [31] for a proof.

Theorem 2.2.26 (see, e.g., [31])**.** *If the convex program in* (2.5) *satisfies Slater's condition, then strong duality holds and the dual optimal is attained with zero duality gap.*

2.2.3.2 Karush–Kuhn–Tucker Conditions

The following is the well-known *Karush-Kuhn-Tucker (KKT) conditions* in optimization.

$$\begin{array}{ll}
\text{(Primal feasibility)} & f_i(\mathsf{x}) \leqslant 0 \;\; \forall i = 1, \ldots m \;\; \text{and} \;\; h_j(\mathsf{x}) = 0 \;\; \forall j = 1, \ldots, p, \\
\text{(Dual feasibility)} & \lambda(i) \geqslant 0 \quad \forall i = 1, \ldots, m, \\
\text{(Complementary slackness)} & \lambda(i) f_i(\mathsf{x}) = 0 \quad \forall i = 1, \ldots, m, \\
\text{(Lagrangian optimality)} & \nabla f_0(\mathsf{x}) + \sum_{i=1}^{m} \lambda(i) \nabla f_i(\mathsf{x}) + \sum_{j=1}^{p} \nu(j) \nabla h_j(\mathsf{x}) = 0,
\end{array}$$

The following proposition says that KKT conditions are both necessary and sufficient for a convex program to attain strong duality. Note that convexity is not required for the necessity.

Proposition 2.2.27. *Suppose $f_0, f_1, \ldots, f_m$ are all convex and differentiable functions over an open domain $\mathcal{D}$, then there exist $x^* \in \mathcal{D}, \lambda^* \in \mathbb{R}^m, \nu^* \in \mathbb{R}^p$ that satisfy KKT conditions if and only if there exists an primal optimal solution x^* to (2.5) and a dual optimal solution (λ^*, ν^*) to (2.7) that attain zero duality gap.*

Proof. We first prove the necessity. If x^* is an optimal solution of the primal problem and (λ^*, ν^*) is an optimal solution of the dual problem, then the primal feasibility and dual feasibility conditions are automatically satisfied. By the zero duality gap assumption, it follows that

$$\begin{aligned} f_0(x^*) = g(\lambda^*, \nu^*) &= \inf_{x \in \mathcal{D}} L(x, \lambda^*, \nu^*) \\ &= \inf_{x \in \mathcal{D}} \left\{ f_0(x) + \sum_{i=1}^m \lambda^*(i) f_i(x) + \sum_{j=1}^p \nu^*(j) h_j(x) \right\} \\ &\leqslant f_0(x^*) + \sum_{i=1}^m \lambda^*(i) f_i(x^*) + \sum_{j=1}^p \nu^*(j) h_j(x^*) \qquad (2.8) \\ &\leqslant f_0(x^*), \end{aligned}$$

where the last inequality follows as $\lambda^*(i) f_i(x^*) \leqslant 0$ and $h_j(x^*) = 0$ by the feasibility of x^* and (λ^*, ν^*). Thus, the two inequalities should hold with equality.

In particular, the last inequality is an equality implies that $\lambda^*(i) f_i(x^*) = 0$ for all $i \in [m]$ due to the feasibility of x^* and λ^*, thus the complementary slackness condition holds. The second last inequality being an equality implies that x^* minimizes the function $L(x, \lambda^*, \nu^*)$ over the open domain $\mathcal{D}$. The minimizer of $L(x, \lambda^*, \nu^*)$ over an open domain, x^*, should satisfy

$$\nabla_x L(x^*, \lambda^*, \nu^*) = \nabla f_0(x) + \sum_{i=1}^m \lambda^*(i) \nabla f_i(x^*) + \sum_{j=1}^p \nu^*(j) \nabla h_j(x^*) = 0.$$

Thus, x^*, λ^*, ν^* also satisfy the Lagrangian optimality condition.

Then, we prove the sufficiency. Assume that there exist $x^* \in \mathcal{D}, \lambda^* \in \mathbb{R}^m, \nu^* \in \mathbb{R}^p$ that satisfy KKT conditions. The primal and dual feasibility conditions ensure that x^* is feasible to the primal problem and (λ^*, ν^*) is feasible to the dual problem. Since $\lambda^* \geqslant 0$, $f_0, f_1, \ldots, f_m$ are convex and $h_1, \ldots, h_p$ are affine, the Lagrangian function $L(x, \lambda^*, \nu^*)$ is convex in x. The Lagrangian optimality condition says that $\nabla_x L(x^*, \lambda^*, \nu^*) = 0$, thus x^* is a minimizer of $L(x, \lambda^*, \nu^*)$ over the convex domain $\mathcal{D}$ by Corollary 2.2.23. Thus, it follows that

$$\begin{aligned} g(\lambda^*, \nu^*) &= \inf_{x \in \mathcal{D}} L(x, \lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*) \\ &= f_0(x^*) + \sum_{i=1}^m \lambda^*(i) f_i(x^*) + \sum_{j=1}^p \nu^*(j) h_j(x^*) = f_0(x^*), \end{aligned}$$

where the last equality follows by primal feasibility that $h_j(x^*) = 0$ for all $j \in [p]$ and complementary slackness condition that $\lambda^*(i) f_i(x^*) = 0$ for all $i \in [m]$.

Hence, x^* and (λ^*, ν^*) attain zero duality gap, and they are primal optimal and dual optimal solutions separately. □

If the primal convex program (2.5) satisfies Slater's condition, then strong duality holds, and thus Proposition 2.2.27 implies the following theorem.

Theorem 2.2.28. *Suppose $f_0, f_1, \ldots, f_m$ are all convex and differentiable functions over an open domain $\mathcal{D}$, and the primal convex program (2.5) satisfies Slater's condition. Then, $x^* \in \mathcal{D}$ is an primal optimal solution of (2.5) if and only if there exist $\lambda^* \in \mathbb{R}^m, \nu^* \in \mathbb{R}^p$, together with x^*, satisfy KKT conditions.*

When the convex optimization problem contains only affine equality constraints, the Slater's condition is satisfied, and the following fact is a direct consequence of Theorem 2.2.28.

Corollary 2.2.29. *Suppose f_0 is a differentiable convex objective function and there are only affine constraints of the form $Ax = b$ with $A \in \mathbb{R}^{p \times d}$ and $b \in \mathbb{R}^p$ in problem (2.5).*

Then, x is an optimal solution to the primal problem (2.5) if and only if there exists $\nu \in \mathbb{R}^p$ such that

$$\nabla f_0(\mathsf{x}) + \mathsf{A}^\top \nu = 0 \quad \textit{and} \quad \mathsf{A}\mathsf{x} = \mathsf{b}.$$

For those problems with nondifferentiable objective or constraint functions, Proposition 2.2.27 and Theorem 2.2.28 cannot be applied. The subdifferential theory for convex functions [126, 76] can deal with those nondifferentiable functions. However, in this thesis we do not need to use the full strength of the advanced background. The following special treatment of Proposition 2.2.27 suffices for our application in Chapter 8.

Proposition 2.2.30. *Suppose $f_0, f_1, \ldots, f_m$ are functions over a domain $\mathcal{D}$. If there exist a primal optimal solution x^* to (2.5) and a dual optimal solution (λ^*, ν^*) to (2.7) that attain zero duality gap, then x^*, λ^* and ν^* satisfy the primal/dual feasibility conditions and the complementary slackness condition. Furthermore, for some direction $\mathsf{d} \in \mathbb{R}^d$, if the directional derivative $f_i'(\mathsf{x}^*; d)$ exits in $\mathcal{D}$ for all $i = 0, 1, \ldots, m$, then the following condition holds.*

$$f_0'(\mathsf{x}^*; \mathsf{d}) + \sum_{i=1}^{m} \lambda^*(i) f_i'(\mathsf{x}^*; \mathsf{d}) + \sum_{j=1}^{p} \nu^* h_j'(\mathsf{x}^*; \mathsf{d}) = 0.$$

Proof. The necessity of the primal/dual feasibility and complementary slackness conditions follows from exactly the same argument as in the proof of Proposition 2.2.27. The only difference is the implications of the inequality (2.8) being a equality. Recall that this means x^* is a minimizer of the function $L(\mathsf{x}, \lambda^*, \nu^*)$ over the domain $\mathcal{D}$. By the assumption that $f_0'(\mathsf{x}^*; \mathsf{d}), f_1'(\mathsf{x}^*; \mathsf{d}), \ldots, f_m'(\mathsf{x}^*; \mathsf{d})$ exist in $\mathcal{D}$, we know that $L'(\mathsf{x}^*, \lambda^*, \nu^*)$ also exists in $\mathcal{D}$ (with $L(\mathsf{x}, \lambda^*, \nu^*)$ being a function in x). The proposition follows by applying Lemma 2.2.21. □

2.3 Graphs and Laplacian Matrices

Let $G = (V, E)$ be an undirected graph with edge weight $\mathsf{w}(e) \geqslant 0$ on each edge $e \in E$. The number of vertices and the number of edges are denoted by $n := |V|$ and $m := |E|$.

For a subset of edges $F \subseteq E$, the total weight of edges in F is $\mathsf{w}(F) := \sum_{e \in F} \mathsf{w}(e)$. For a subset of vertices $S \subseteq V$, the set of edges with one endpoint in S and one endpoint in $V - S$ is denoted by $\delta(S)$. For a vertex v, the set of edges incident on a vertex v is $\delta(v) := \delta(\{v\})$, and the weighted *degree* of v is $\deg(v) := \mathsf{w}(\delta(v))$. The *expansion* of a set S and *expansion* of a graph G are defined as

$$\phi(S) := \frac{|\delta(S)|}{|S|} \qquad \text{and} \qquad \phi(G) := \min_{0 \leqslant |S| \leqslant \frac{n}{2}} \phi(S).$$

The symbol ϕ is usually used for graph conductance. However, we will not use graph conductance in this thesis, thus we save ϕ for expansion.

The *adjacency matrix* $\mathsf{A} \in \mathbb{R}^{n \times n}$ of the graph is defined as $\mathsf{A}(u,v) = \mathsf{w}(uv)$ for all $uv \in E$, and $\mathsf{A}(u,v) = 0$ otherwise. The *Laplacian matrix* $\mathsf{L} \in \mathbb{R}^{n \times n}$ of the graph is defined as $\mathsf{L} = \mathsf{D} - \mathsf{A}$ where $\mathsf{D} \in \mathbb{R}^{n \times n}$ is the diagonal degree matrix with $\mathsf{D}(u,u) = \deg(u)$ for all $u \in V$. Similarly, the *signless Laplacian matrix* $\mathsf{L}^+ \in \mathbb{R}^{n \times n}$ is defined as $\mathsf{L}^+ = \mathsf{D} + \mathsf{A}$. For each edge $e = uv \in E$, let $\mathsf{b}_e := \mathcal{X}_u - \mathcal{X}_v$ where $\chi_u \in \mathbb{R}^n$ is the vector with one in the u-th entry and zero otherwise. The Laplacian matrix of G with respect to edge weight w can also be written as

$$\mathsf{L}_\mathsf{w} = \sum_{e \in E} \mathsf{w}(e) \cdot \mathsf{b}_e \mathsf{b}_e^\top = \mathsf{B} \mathsf{W} \mathsf{B}^\top,$$

where $\mathsf{W} = \operatorname{diag}(w)$ and $\mathsf{B} \in \mathbb{R}^{V \times E}$ is the n-by-m matrix with b_e being the e-th column.

Let $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n$ be the eigenvalues of L with corresponding orthonormal eigenvectors $\mathsf{v}_1, \mathsf{v}_2, \ldots, \mathsf{v}_n$ so that $\mathsf{L} = \sum_{i=1}^n \lambda_i \cdot \mathsf{v}_i \mathsf{v}_i^\top$. It is well-known that the Laplacian matrix is positive semidefinite. It is also well-known that $\lambda_1 = 0$ with $\mathsf{v}_1 = \frac{1}{\sqrt{n}} \mathit{1}$ as the corresponding unit eigenvector. The second smallest eigenvalue λ_2 is known as *algebraic connectivity*, and $\lambda_2 > 0$ if and only if G is connected. The pseudoinverse of the Laplacian matrix L of a connected graph is defined as

$$\mathsf{L}^\dagger := \sum_{i=2}^n \frac{1}{\lambda_i} \cdot \mathsf{v}_i \mathsf{v}_i^\top,$$

which maps every vector b orthogonal to v_1 to a vector y such that $\mathsf{L}\mathsf{y} = \mathsf{b}$. We write $\mathsf{L}^{\frac{\dagger}{2}}$ as the square root of $\mathsf{L}^\dagger$.

The following fact is useful for eigenvalue maximization and imposing eigenvalue lower bounds. The proof is similar to the one of Lemma 2.2.14, with an additional observation that L_w's first eigenvector $v_1 = \frac{1}{\sqrt{n}}\mathbb{1}$ for any $w \geqslant 0$.

Lemma 2.3.1 (see, e.g., [69]). *$\lambda_2(L_w)$ is a concave function with respect to w for $w \geqslant 0$.*

2.4 Electrical Flow and Effective Resistance

In this section, we introduce the notions of electric network, electrical flow and effective resistance. An undirected graph $G = (V, E)$ with edge weights $w \in \mathbb{R}^E$ can be interpreted as an *electric network*, where each edge $e \in E$ is treated as a resistor with resistance $r_e = \frac{1}{w(e)}$.

To define an electrical flow on graph G, we start with defining a unit s-t flow on an undirected graph $G = (V, E)$. We first fix an orientation of the edges of G arbitrarily. Let $B \in \mathbb{R}^{V \times E}$ be the matrix defined in Section 2.3 that is consistent with this orientation, i.e. for an edge $e = uv \in E$ oriented from u to v, $B(u, e) = 1$, $B(v, e) = -1$, and zero otherwise in the e-th column of B. A unit *s-t flow* $f : E \to \mathbb{R}$ is an m-dimensional vector that satisfies flow conservation constraints:

$$Bf = b_{st} \quad \text{or equivalently} \quad \sum_{e=uv\ :\ u \in \delta^+(v)} f(e) - \sum_{e=uv\ :\ u \in \delta^-(v)} f(e) = \begin{cases} 1 & v = s \\ -1 & v = t \\ 0 & \text{otherwise,} \end{cases}$$

where $b_{st} := \chi_s - \chi_t$, and $\delta^+(v)$ and $\delta^-(v)$ are the set of outgoing and incoming neighbours of v with respect to the fixed orientation. Note that positive $f(e)$ indicates the flow on e is in the same direction as the orientation of e, while negative $f(e)$ indicates the opposite.

The unit *s-t electrical flow* is a unit s-t flow f that also satisfies the *Ohm's law*: There exists a potential vector $\varphi \in \mathbb{R}^V$ such that for all $e = uv \in E$ oriented from u to v,

$$f(e) = w(e) \cdot (\varphi(u) - \varphi(v)).$$

Given a potential vector φ, we will use the orientation where all edges are pointing from the high potential endpoint to the low potential endpoint, so that $\mathsf{f}(e)$ is nonnegative for all $e \in E$ in the rest of this thesis. Notice that the constraints imposed by Ohm's law can be written as a linear system in f and φ

$$\mathsf{f} = \mathsf{W}\mathsf{B}^\top \varphi, \tag{2.9}$$

where $\mathsf{W} \in \mathbb{R}^{E \times E}$ is a diagonal matrix with $\mathsf{w}(e)$ in the (e, e)-th entry, and $\mathsf{B} \in \mathbb{R}^{V \times E}$ is defined in Section 2.3 with signs of the e-th column defined according to the orientation of edge e. Combining the flow conservation constraint and the Ohm's law, we can check that the potential vector $\varphi \in \mathbb{R}^V$ of the unit s-t electrical flow is a solution to the linear system

$$\mathsf{B}\mathsf{W}\mathsf{B}^\top \varphi = \mathsf{b}_{st} \quad \Longrightarrow \quad \mathsf{L}_\mathsf{w} \cdot \varphi = \mathsf{b}_{st}.$$

Note that if G is connected $\varphi = \mathsf{L}_\mathsf{w}^\dagger \mathsf{b}_{st}$ is a solution, and any solution is given by $\varphi + c \cdot \mathit{1}$ for $c \in \mathbb{R}$. Thus, the electrical flow f satisfying (2.9) is uniquely defined. Furthermore, the *effective resistance* between s and t in G can be uniquely defined as

$$\mathrm{Reff}_G(s, t) := \varphi(s) - \varphi(t), \tag{2.10}$$

which is the potential difference between s and t when one unit of electrical flow is sent from s to t. When s and t are disconnected in the underlying graph, we can not send electrical flow from s to t. In this case, $\mathrm{Reff}_G(s, t)$ is defined to be $+\infty$. The s-t effective resistance can be interpreted as the resistance of the whole graph G as a big resistor when an electrical flow is sent from s to t.

We describe some notational conventions for effective resistance in this thesis. The subscript G in $\mathrm{Reff}_G(s, t)$ is dropped when G is clear from the context. We write $\mathrm{Reff}_\mathsf{w}(s, t)$ to emphasize that the underlying graph has edge weights w. We also need to frequently refer to the s-t effective resistance as a function of w. In this case we write

$$\mathrm{Reff}_{st}(\mathsf{w}) := \mathrm{Reff}_\mathsf{w}(s, t).$$

2.4.1 Formulas for Effective Resistance

For series and parallel electrical circuits, effective resistances are easy to compute.

Fact 2.4.1 (resistance of series and parallel circuits)**.** *Let s, t be two designated vertices.*

- *If s and t are connected by a series of k edges, each with resistance $r_1, ..., r_k$, then the s-t effective resistance is* $\mathrm{Reff}(s,t) = r_1 + \cdots + r_k$.
- *If s and t are connected by k parallel edges, each with resistance $r_1, ..., r_k$, then the s-t effective resistance is* $\mathrm{Reff}(s,t) = \left(\frac{1}{r_1} + \cdots + \frac{1}{r_k}\right)^{-1}$.

For general graphs, one can write the effective resistance in terms of the Laplacian matrix. When G is connected and φ is the potential vector satisfies the Ohm's law, $\varphi = L_w^\dagger b_{st} + c \cdot 1$ by the discussion above the definition (2.10). Thus, we can write

$$\mathrm{Reff}_w(s,t) = \varphi(s) - \varphi(t) = b_{st}^\top L_w^\dagger b_{st}.$$

In general, the above expression also applies to s-t connected graphs. Given edge weights $w \in \mathbb{R}_+^E$, we say w is s-t *connected* if s and t are connected in the subgraph spanned by those edges in the support of w, i.e. $\{e \in E \mid w(e) > 0\}$. The set of all s-t connected edge weights can be formally defined by

$$\mathcal{D}_{st} := \{w \in \mathbb{R}_+^E \mid w(\delta(S)) > 0, \quad \text{for all } S \subseteq V \text{ with } s \in S \text{ and } t \notin S\}. \tag{2.11}$$

Note that $\mathcal{D}_{st}$ is the intersection of some halfspaces, which is a convex set.

Then, we observe that $\mathrm{Reff}_{st}(w)$ coincides with $b_{st}^\top L_w^\dagger b_{st}$ over those $w \in \mathcal{D}_{st}$.

Fact 2.4.2. *$w \in \mathcal{D}_{st}$ if and only if b_{st} lives in the range of L_w, i.e. $(I - L_w L_w^\dagger) b_{st} = 0$. Furthermore,* $\mathrm{Reff}_{st}(w) = b_{st}^\top L_w^\dagger b_{st}$ *for any $w \in \mathcal{D}_{st}$, and* $\mathrm{Reff}_{st}(w) = +\infty$ *otherwise.*

Proof. When $w \in \mathcal{D}_{st}$, s and t are in the same connected component $S \subseteq V$. The matrix L_w is in a block form with blocks corresponding to S and its complement $\bar{S}$ (could be empty), i.e. $L_w = \begin{pmatrix} L_S & 0 \\ 0 & L_{\bar{S}} \end{pmatrix}$. Let b_{st} be also in a block form with $b_{st} = \begin{pmatrix} b_{st}^S \\ 0 \end{pmatrix}$ where $b_{st}^S \in \mathbb{R}^S$. Then,

$$L_w L_w^\dagger b_{st} = \begin{pmatrix} L_S L_S^\dagger & 0 \\ 0 & L_{\bar{S}} L_{\bar{S}}^\dagger \end{pmatrix} \begin{pmatrix} b_{st}^S \\ 0 \end{pmatrix} = \begin{pmatrix} L_S L_S^\dagger b_{st}^S & 0 \\ 0 & 0 \end{pmatrix} = b_{st},$$

where the last equality follows as S is a connected component, the null space of L_S is spanned by $\mathit{1}_S$, and b_{st}^S is orthogonal to $\mathit{1}_S$. This also implies $\mathsf{b}_{st}^\top \mathsf{L}_\mathsf{w}^\dagger \mathsf{b}_{st} = (\mathsf{b}_{st}^S)^\top \mathsf{L}_S^\dagger \mathsf{b}_{st}^S$. Since restricting to the component S does not change effective resistance, thus we have $\mathrm{Reff}_{st}(\mathsf{w}) = \mathsf{b}_{st}^\top \mathsf{L}_\mathsf{w}^\dagger \mathsf{b}_{st}$.

When $\mathsf{w} \notin \mathcal{D}_{st}$, we assume s is in a connected component S and $t \in \bar{S}$. Clearly, $\mathrm{Reff}_{st}(\mathsf{w}) = +\infty$ by definition. We show that b_{st} is not in the range of L_w. Note that b_{st} is in the following block form $\mathsf{b}_{st} = \begin{pmatrix} \chi_s^S \\ -\chi_t^{\bar{S}} \end{pmatrix}$, where $\chi_s^S \in \{0,1\}^S$ is the indicator vector of s restricted to S and $\chi_t^{\bar{S}}$ is similar. Thus

$$\mathsf{L}_\mathsf{w} \mathsf{L}_\mathsf{w}^\dagger \mathsf{b}_{st} = \begin{pmatrix} \mathsf{L}_S \mathsf{L}_S^\dagger & 0 \\ 0 & \mathsf{L}_{\bar{S}} \mathsf{L}_{\bar{S}}^\dagger \end{pmatrix} \begin{pmatrix} \chi_s^S \\ -\chi_t^{\bar{S}} \end{pmatrix} = \begin{pmatrix} \mathsf{L}_S \mathsf{L}_S^\dagger \chi_s^S & 0 \\ 0 & -\mathsf{L}_{\bar{S}} \mathsf{L}_{\bar{S}}^\dagger \chi_t^{\bar{S}} \end{pmatrix} \neq \mathsf{b}_{st},$$

where the last inequality follows as, $\mathit{1}_S$ spans the null space of L_S for connected S and χ_s^S is not orthogonal to $\mathit{1}_S$, which implies that $\mathsf{L}_S \mathsf{L}_S^\dagger \chi_s^S \neq \chi_s^S$ after the projection onto the range of L_S. □

2.4.2 Thomson's Principle and Rayleigh's Monotonicity Principle

The effective resistance can also be characterized by the energy of a flow. The *energy* of an s-t flow f is defined as

$$\mathcal{E}(\mathsf{f}) := \sum_{e \in E} \frac{\mathsf{f}(e)^2}{\mathsf{w}(e)} = \sum_{e \in E} r_e \cdot \mathsf{f}(e)^2.$$

Thomson's principle [84] states that the unit s-t electrical flow is the unique unit s-t flow that minimizes the energy.

Theorem 2.4.3 (Thomson's principle [84]). *Let f^* be the unit electrical s-t flow in an s-t connected graph G. Then*

$$\mathrm{Reff}_G(s,t) = \min_{\mathsf{f}} \{\mathcal{E}(\mathsf{f}) \mid \mathsf{f} \text{ is a unit } s\text{-}t \text{ flow in } G\} = \mathcal{E}(\mathsf{f}^*).$$

Proof. If there is any edge with weight $w(e) = 0$, the edge has resistance ∞ and any energy minimizer f sends 0 unit of flow on e (otherwise the energy is unbounded). Thus, we can assume that $w > 0$ by removing all zero-weight edges without loss of generality. We rewrite the energy minimization problem in the following form (the rescaling does not affect the optimizer).

$$\begin{aligned} \underset{f \in \mathbb{R}^E}{\text{minimize}} \quad & \frac{1}{2} \cdot f^\top \operatorname{diag}(w)^{-1} f \\ \text{subject to} \quad & Bf = b_{st}. \end{aligned}$$

This is an optimization problem with a quadratic objective function and linear equality constraints. The optimizer of the convex program is unique due to the strict convexity of the objective function. By Corollary 2.2.29, a flow f is the unique optimizer if and only if there exists $\varphi \in \mathbb{R}^V$ such that

$$Bf = b_{st} \qquad \text{and} \qquad \nabla_f \left(\frac{1}{2} \cdot f^\top \operatorname{diag}(w)^{-1} f \right) - B^\top \varphi = \operatorname{diag}(w)^{-1} f - B^\top \varphi = 0.$$

The second equality implies $f(e) = w(e) \cdot (\varphi(u) - \varphi(v))$ for each edge $e = uv \in E$ oriented from u to v according to B, and the first equality guarantees f is a unit s-t flow. The unique optimizer f satisfies Ohm's law with respect to potential vector φ, and hence is equal to the unit s-t electrical flow f^*.

To see that the energy is equal to the effective resistance, note that the flow value on edge $e = uv$ in the unit s-t electrical flow satisfies $f^*(e) = w(e) \cdot (\varphi(u) - \varphi(v)) = w(e) \cdot b_e^\top L_w^\dagger b_{st}$ and thus

$$\begin{aligned} \mathcal{E}(f^*) = \sum_{e \in E} w(e) \cdot (b_e^\top L_w^\dagger b_{st})^2 &= b_{st}^\top L_w^\dagger \left(\sum_{e \in E} w(e) \cdot b_e b_e^\top \right) L_w^\dagger b_{st} \\ &= b_{st}^\top L_w^\dagger L_w L_w^\dagger b_{st} = b_{st}^\top L_w^\dagger b_{st} = \operatorname{Reff}_G(s,t), \end{aligned}$$

where the second equality follows as $(b_e^\top L_w^\dagger b_{st})^2 = b_{st}^\top L_w^\dagger b_e b_e^\top L_w^\dagger b_{st}$. □

A corollary of Thomson's principle is the following intuitive fact known as the Rayleigh's monotonicity principle.

Theorem 2.4.4 (Rayleigh's monotonicity principle)**.** *The s-t effective resistance cannot increase if the resistance of an edge is decreased.*

Proof. Let $w \in \mathbb{R}_+^E$ be the original edge weights. We increase the edge weight of the e-th entry (or equivalently decrease the resistance) to make a new vector w' for an arbitrary edge $e \in E$. Let f be the unit electrical flow with respect to w. Then, it follows that

$$\mathrm{Reff}_w(s,t) = \sum_{e\in E} \frac{f(e)^2}{w(e)} \geqslant \sum_{e\in E} \frac{f(e)^2}{w'(e)} \geqslant \inf_{f':\text{ unit } s\text{-}t \text{ flow}} \sum_{e\in E} \frac{f'(e)^2}{w'(e)} = \mathrm{Reff}_{w'}(s,t),$$

where the first inequality follows as $w' \geqslant w$, and the last equality follows by Thomson's principle Theorem 2.4.3. □

2.4.3 Convexity and Differentiability of Effective Resistance

We will also use the convexity of effective resistance to write convex programming relaxations for problems with effective resistance objective function (or constraints).

In [70], Ghosh, Boyd, and Saberi showed that s-t effective resistance is convex with respect to the edge weights w over the domain where the support of w spans the whole graph. We can extend the domain to $\mathbb{R}_+^E$ by a continuity argument. In the following, we slightly extend the proof of [70] (or the one in Example 3.4 of [31]) without invoking the continuity argument.

Lemma 2.4.5. *The s-t effective resistance on a graph $G = (V, E; w)$ is a convex function with respect to the edge weights w on the domain $\mathbb{R}_+^E$.*

Proof. For any given $w, w' \in \mathbb{R}_+^E$ and $\lambda \in [0,1]$, if w is s-t disconnected, then the inequality $\mathrm{Reff}_{st}(\lambda w + (1-\lambda)w') \leqslant \lambda\,\mathrm{Reff}_{st}(w) + (1-\lambda)\,\mathrm{Reff}_{st}(w')$ always holds, as $\mathrm{Reff}_{st}(w) = +\infty$ by Fact 2.4.2. Thus, it suffices to consider the restricted domain $\mathcal{D}_{st}$. We claim that the following is an equivalent characterization of the domain $\mathcal{D}_{st}$.

$$\mathcal{D}_{st} = \{w \in \mathbb{R}_+^E \mid (I - L_w L_w^\dagger) b_{st} = 0\} = \{w \in \mathbb{R}_+^E \mid (I - L_w L_w^\dagger) b_{st} = 0, L_w \succcurlyeq 0\}.$$

The first equality follows from Fact 2.4.2. For the second equality, note that the condition $L_w \succcurlyeq 0$ is redundant given $w \in \mathbb{R}^E_+$, but it will be useful in the following characterization.

To prove $\mathrm{Reff}_{st}(w)$ is convex on the domain $w \in \mathcal{D}_{st}$, according to Lemma 2.2.7, it suffices to show that the following epigraph of the s-t effective resistance is convex:

$$\begin{aligned}\mathrm{epi}(\mathrm{Reff}_{st}) &= \{(w,u) \in \mathbb{R}^{m+1} \mid w \in \mathcal{D}_{st}, b_{st}^\top L_w^\dagger b_{st} \leqslant u\} \\ &= \{(w,u) \in \mathbb{R}^{m+1} \mid w \in \mathbb{R}^E_+, (I - L_w L_w^\dagger) b_{st} = 0, L_w \succcurlyeq 0, b_{st}^\top L_w^\dagger b_{st} \leqslant u\},\end{aligned}$$

where the second equality follows by the equivalent characterization of $\mathcal{D}_{st}$. By the Schur complement characterization of semidefiniteness in Lemma 2.1.7, it follows that

$$\mathrm{epi}(\mathrm{Reff}_{st}) = \left\{(w,u) \in \mathbb{R}^{m+1} \middle| w \in \mathbb{R}^E_+, \begin{pmatrix} L_w & b_{st} \\ b_{st}^\top & u \end{pmatrix} \succcurlyeq 0 \right\}.$$

For any $(w_1, u_1), (w_2, u_2) \in \mathrm{epi}(\mathrm{Reff}_{st})$ and $\lambda \in [0,1]$, we can check that the convex combination $(w,u) = \lambda(w_1, u_1) + (1-\lambda)(w_2, u_2)$ satisfies

$$w = \lambda w_1 + (1-\lambda) w_2 \geqslant 0 \quad \text{and} \quad \begin{pmatrix} L_w & b_{st} \\ b_{st}^\top & u \end{pmatrix} = \lambda \begin{pmatrix} L_{w_1} & b_{st} \\ b_{st}^\top & u_1 \end{pmatrix} + (1-\lambda) \begin{pmatrix} L_{w_2} & b_{st} \\ b_{st}^\top & u_2 \end{pmatrix} \succcurlyeq 0,$$

where both inequalities follow by the fact that $(w_1, u_1), (w_2, u_2) \in \mathrm{epi}(\mathrm{Reff}_{st})$. □

Finally, we consider the differentiability of $\mathrm{Reff}_{st}(w)$. It is not difficult to verify that if $w > 0$, then $\mathrm{Reff}_{st}(w)$ is differentiable at w, We also have discussed that $\mathrm{Reff}_{st}(w)$ has finite value when $w \in \mathcal{D}_{st}$. Can we say $\mathrm{Reff}_{st}(w)$ is differentiable over the whole domain $\mathcal{D}_{st}$? The answer is no, because there are many points $w \in \mathbb{R}^E_+$ contain zero entries and lay on the boundary of $\mathcal{D}_{st}$. $\mathrm{Reff}_{st}(w)$ is undefined when w contains negative edge weights, thus $\mathrm{Reff}_{st}(w)$ is differentiable at w if and only if $w \in \mathbb{R}^E_{++}$. Nevertheless, we still can show that partial derivatives of $\mathrm{Reff}_{st}(w)$ exist at $w \in \mathcal{D}_{st}$ with respect to those coordinates with $w(e) > 0$.

Lemma 2.4.6. *Given any $w \in \mathcal{D}_{st}$, if $w(e) > 0$, then $\partial_e \mathrm{Reff}_{st}(w)$ exists in $\mathcal{D}_{st}$ and $\partial_e \mathrm{Reff}_{st}(w) = -(b_{st}^\top L_w^\dagger b_e)^2$.*

Proof. By the definition of directional derivatives in Section 2.2.1,

$$\partial_e \operatorname{Reff}_{st}(w) = \lim_{\lambda \to 0} \frac{\operatorname{Reff}_{st}(w + \lambda \chi_e) - \operatorname{Reff}_{st}(w)}{\lambda}.$$

We consider the term $\operatorname{Reff}_{st}(w + \lambda\chi_e)$. For small enough λ (could be negative), $w + \lambda\chi_e$ is still s-t connected. By Fact 2.4.2,

$$\operatorname{Reff}_{st}(w) = b_{st}^\top L_w^\dagger b_{st} \quad \text{and} \quad \operatorname{Reff}_{st}(w + \lambda\chi_e) = b_{st}^\top L_{w+\lambda\chi_e}^\dagger b_{st} = b_{st}^\top \left(L_w + \lambda b_e b_e^\top\right)^\dagger b_{st}.$$

Since $w(e) > 0$, b_e lives in the range of L_w. For small enough λ, $1 + \lambda b_e^\top L_w^\dagger b_e \neq 0$. Thus, by Corollary 2.1.14, for small enough λ we have

$$\left(L_w + \lambda b_e b_e^\top\right)^\dagger = L_w^\dagger - \frac{\lambda}{1 + \lambda b_e^\top L_w^\dagger b_e} \cdot L_w^\dagger b_e b_e^\top L_w^\dagger,$$

which implies

$$\operatorname{Reff}_{st}(w + \lambda\chi_e) = \operatorname{Reff}_{st}(w) - \frac{\lambda (b_{st}^\top L_w^\dagger b_e)^2}{1 + \lambda b_e^\top L_w^\dagger b_e}.$$

Thus,

$$\partial_e \operatorname{Reff}_{st}(w) = \lim_{\lambda \to 0} -\frac{(b_{st}^\top L_w^\dagger b_e)^2}{1 + \lambda b_e^\top L_w^\dagger b_e} = -(b_{st}^\top L_w^\dagger b_e)^2.$$

□

2.5 Spectral Sparsification

In the study of combinatorial graph algorithms, a natural yet powerful idea is to use a sparse object as the representative of the input graph to help solve the original problem more efficiently [57]. Various notions of sparse representatives have been proposed and studied in the literature. Among them, the *cut-sparsifier* that approximately preserves values of all cuts [82, 83] is one of the most well-studied notions. In an influential work, Benczúr and Karger [23] showed that every graph admits a cut-sparsifier of size $O(\frac{n \log n}{\varepsilon^2})$ that preserves values of all cuts up to a $(1 \pm \varepsilon)$ factor, and furthermore the cut-sparsifier can be constructed in nearly linear time. Later, Spielman and Teng [133] introduced a more general notion, *spectral sparsification*, which is defined as follows.

Definition 2.5.1 (Spectral Sparsifier)**.** *Given an undirected graph $G = (V, E_G; \mathsf{w})$ with edge weights $\mathsf{w} \in \mathbb{R}_+^{E_G}$, an undirected graph $H = (V, E_H; \widetilde{\mathsf{w}})$ on the same vertex set V with edge weights $\widetilde{\mathsf{w}} \in \mathbb{R}_+^{E_H}$ is a $(1+\varepsilon)$-*spectral sparsifier *of G for some $\varepsilon \geqslant 0$ if the following holds*

$$(1+\varepsilon)^{-1} \cdot \mathsf{x}^\top \mathsf{L}_G \mathsf{x} \leqslant \mathsf{x}^\top \mathsf{L}_H \mathsf{x} \leqslant (1+\varepsilon) \cdot \mathsf{x}^\top \mathsf{L}_G \mathsf{x}, \qquad \textit{for all } \mathsf{x} \in \mathbb{R}^V,$$

where L_G and L_H are Laplacian matrices of the edge-weighted graph G and H.

When we restrict the constraints to those vectors $\mathsf{x} \in \{0,1\}^V$, then the above definition is exactly the same as that for cut sparsification. This shows spectral sparsification is a strictly stronger notion than cut sparsification.

Spielman and Teng [133] gave the first algorithm to construct a $(1+\varepsilon)$-spectral sparsifier of size $O(\frac{n \log^c n}{\varepsilon^2})$ (for some large constant c) in nearly linear time, and the constructed spectral sparsifier is a reweighted subgraph of the input. Spectral sparsification was a key ingredient in Spielman and Teng's first nearly linear time Laplacian linear system solver [134]. Since then, spectral sparsifers have found many applications in designing fast algorithms, e.g., faster solver for Laplacian linear systems and generalizations [89, 88, 90], computing maximum flow [44, 109, 120], fast random spanning trees sampling [115, 54, 128], measuring edge centrality [103], etc.

Besides numerous applications, spectral sparsification has became an interesting research topic in itself. The techniques that developed to attack spectral sparsification led to a solution to the famous Kadison-Singer problem [110, 111]. In the remaining of this section, we briefly survey several important developments about constructing spectral sparsifiers.

We start with reducing the problem into a nicer form by a preprocessing step. Let $\mathsf{L}_G = \sum_{e \in E_G} \mathsf{w}(e) \cdot \mathsf{b}_e \mathsf{b}_e^\top$ be the Laplacian matrix of the input graph. Without loss of generality, we assume the graph G is connected, so that L_G is of rank $n-1$ with a null space spanned by the all-one vector $\mathit{1}$. Let $\mathsf{L}_G = \sum_{i=2}^n \lambda_i \cdot \mathsf{u}_i \mathsf{u}_i^\top = \mathsf{U} \Lambda \mathsf{U}^\top$ be the eigendecomposition of L_G, where $\Lambda \in \mathbb{R}^{(n-1)\times(n-1)}$ is a diagonal matrix with $\Lambda(i,i) = \lambda_i$, and $\mathsf{U} \in \mathbb{R}^{n \times (n-1)}$ contains the i-th eigenvecotr u_i at the i-th column. We consider the following linear transformation

$$\mathsf{v}_e := \mathsf{U}^\top \mathsf{L}_G^{\frac{\dagger}{2}} \mathsf{b}_e \in \mathbb{R}^{n-1} \qquad \text{for all } e \in E_G. \tag{2.12}$$

Note that the squared vector length $\|v_e\|_2^2 = b_e^\top L_G^\dagger b_e$ is exactly the effective resistance between the two endpoints of edge e in the input graph, and

$$\sum_{e \in E_G} w(e) \cdot v_e v_e^\top = U^\top L_G^{\frac{\dagger}{2}} \left(\sum_{e \in E_G} w(e) \cdot b_e b_e^\top \right) L_G^{\frac{\dagger}{2}} U = U^\top L_G^{\frac{\dagger}{2}} L_G L_G^{\frac{\dagger}{2}} U = U^\top U = I_{n-1},$$

as the columns of U are orthonormal. Thus, we have transformed the construction of spectral sparsifier into the following nicer form, where we only need to ensure that the maximum and minimum eigenvalue of the solution are close to one. After solving the following problem, we can easily turn the solution into a solution of the original problem by reversing the preprocessing step.

Problem 2.5.2. *Suppose we are given vectors $v_1, \ldots, v_m \in \mathbb{R}^n$ and weights $w \in \mathbb{R}_+^m$ such that $\sum_{i=1}^m w(i) \cdot v_i v_i^\top = I_n$. For some given $\varepsilon > 0$, find a reweighting $\widetilde{w} \in \mathbb{R}_+^m$ such that*

$$(1+\varepsilon)^{-1} I_n \preccurlyeq \sum_{i=1}^m \widetilde{w}(i) \cdot v_i v_i^\top \preccurlyeq (1+\varepsilon) I_n \qquad \textit{and} \qquad |\{i \in [m] : \widetilde{w}(i) \neq 0\}| \textit{ is small.}$$

2.5.1 Effective Resistance Sampling

Spielman and Srivastava proposed a randomized algorithm in [132] to construct a $(1+\varepsilon)$-spectral sparsifier of size $O(\frac{n \log n}{\varepsilon^2})$ in nearly linear time, which matches and extends Benczúr and Karger's result for cut sparsification [23]. The algorithm is a very simple and elegant sampling algorithm.

```
Effective Resistance Sampling

1. Initialization: $Z_0 \leftarrow 0$ and $\tau \leftarrow O(\frac{n \log n}{\varepsilon^2})$.
2. For $t \leftarrow 1$ to $\tau$ do
    (a) Independently sample a vector $i_t = i \in [m]$ with probability
            $\mathbb{P}[i_t = i] \propto w(i) \cdot \|v_i\|_2^2$.
    (b) Let $\Delta_t := \frac{n}{\tau \|v_{i_t}\|_2^2} v_{i_t} v_{i_t}^\top$, and update $Z_t \leftarrow Z_{t-1} + \Delta_t$.
```

The term $w(i) \cdot \|v_i\|_2^2$ in the sampling probability is known as the *leverage score* of vector v_i. In the graph setting, it is the effective resistance of an edge when the input graph is unweighted. The size of each update Δ_t is bounded by $\left\|\frac{n}{\tau\|v_{i_t}\|_2^2} \cdot v_{i_t} v_{i_t}^\top\right\|_{\mathrm{op}} = \frac{n}{\tau}$, and the expected update of each iteration is $\mathbb{E}\left[\frac{n}{\tau\|v_{i_t}\|_2^2} \cdot v_{i_t} v_{i_t}^\top\right] = \frac{1}{\tau} I_d$, where we need to use the fact that $\sum_{i=1}^m w(i) \cdot \|v_i\|_2^2 = n$. By matrix concentration inequalities (e.g., matrix chernoff bound Theorem 3.4.1, matrix Bernstein's inequality Theorem 3.4.3, see Section 3.4 for more details), after $\tau = O(\frac{n \log n}{\varepsilon^2})$ iterations the resulting subgraph is a $(1+\varepsilon)$-spectral sparsifier with high probability. Note that with the independent effective resistance sampling approach, the $\log n$ factor cannot be improved, since $\Omega(n \log n)$ iterations are required to guarantee that no isolated vertex exists when the input is an unweighted complete graph.

2.5.2 Barrier Function Methods

Batson, Spielman, and Srivastava designed a new deterministic algorithm to construct a $(1+\varepsilon)$-spectral sparsifier of size $O(\frac{n}{\varepsilon^2})$ in polynomial time [21]. We will refer to the algorithm by BSS in this thesis. This result is quite remarkable in several aspects. It beats the independent random sampling, and the dependence of ε in the bound $O(\frac{n}{\varepsilon^2})$ is tight up to a constant factor. Furthermore, the barrier function idea led to a solution to Kadison-Singer problem, which we will explain more in Section 2.6.

To illustrate the idea of the BSS algorithm, we first consider a naive plan as follows. Through an iterative process, we maintain an upper barrier u_t and a lower barrier l_t to ensure the whole spectrum of the current solution is located within the range of the barriers $[l_t, u_t]$. In each iteration, we select a vector v_{i_t} from the input and add an update $\Delta_t = c_t \cdot v_{i_t} v_{i_t}^\top$ with appropriately chosen reweighting $c_t \geqslant 0$. Then, we increase a nontrivial amount in both barriers such that the gap between the barriers $u_t - l_t$ is approximately preserved and the spectrum of the new solution is still contained in $[l_t, u_t]$. After certain number of iterations, when l_t is large enough (in the order of $\Omega(\frac{u_t - l_t}{\varepsilon})$), the condition number of the solution is bounded by $\frac{u_t}{l_t} = 1 + \frac{u_t - l_t}{l_t} = 1 + O(\varepsilon)$.

However, as the maximum and the minimum eigenvalue are very sensitive to the solution update, it is not easy to control how much we should shift the barriers algorithmically.

Thus, instead of controlling the maximum and the minimum eigenvalue directly, Batson, Spielman, and Srivastava introduced two barrier potential functions which are more robust to the solution updates

$$\Phi^u(Z) = \operatorname{tr}\left((uI - Z)^{-1}\right) \qquad \text{and} \qquad \Phi^l(Z) = \operatorname{tr}\left((Z - lI)^{-1}\right), \tag{2.13}$$

where Z is the current solution. Intuitively, when $lI \prec Z \prec uI$, the barrier potential functions measure how far the eigenvalues of Z are from the barriers. In particular, if any of the eigenvalues of Z approaches a barrier, the corresponding potential function will blow up dramatically.

Thus, the key task of the algorithm is to add an appropriately reweighted vector v such that $\Phi^{u_t+\delta_u}(Z_t + vv^\top) \leqslant \Phi^{u_t}(Z_t)$ and $\Phi^{l_t+\delta_l}(Z_t + vv^\top) \leqslant \Phi^{l_t}(Z_t)$ (with significant shifts of δ_u, δ_l where $\delta_u \approx \delta_l$). With an averaging argument, they showed that there always exists a vector from the input that can be reweighted to satisfy all the desired properties. With appropriately chosen parameters, it only takes $O(\frac{n}{\varepsilon^2})$ iterations to return a $(1+\varepsilon)$-spectral sparsifier. Moreover, they showed that this is the optimal dependence in ε by using Ramanujan graph as a tight example.

2.5.3 Regret Minimization Reformulation

Allen-Zhu, Liao, and Orecchia [7] proposed an algorithm for spectral sparsification based on a general regret minimization framework (which is well-known to the online optimization community). There are several advantages of their algorithm. For example, there is no need to explicitly maintain the shift of the barriers, and the algorithm analysis can be done in a more principled way within a general mirror descent framework.

They recovered the optimal approximation guarantee of BSS using the same potential function. By introducing a generalized potential function in the regret minimization framework, they managed to improve the running time to almost-quadratic time (a naive implementation of the BSS algorithm takes $\Omega(n^4)$ time).

As a key tool in this thesis, we will have a comprehensive review of the regret minimization framework in Chapter 4.

2.5.4 Potential Function Guided Adaptive Sampling

The barrier potential functions in the BSS algorithm provide a nice way to control the eigenvalues of a solution, which is the key ingredient in achieving the optimal approximation guarantee. However, the choice of the parameters looks a bit mysterious. In particular, the existence of a good vector and the appropriate reweighting in each iteration in the BSS algorithm is guaranteed by an averaging argument. The relation between the reweighting and the shift of the barriers is not explicit. Also, we may get the impression that there are very few good vectors and we need to pick a good vector very carefully in each iteration.

To design a fast algorithm for spectral sparsification, Lee and Sun [101] combined the random sampling approach in [132] with the potential function approach in [21]. The algorithm provides an explicit description of how we shall reweight the edges. Moreover, their algorithm shows that we have much more flexibility in choosing vectors and assigning reweightings in each iteration. In particular, they proved that choosing a random vector following an approximate probability distribution would work, and there is no need to maintain that the potential value is non-increasing.

Lee and Sun first combined the two potential functions in the BSS algorithm into a single one

$$\Phi^{u,l}(Z) = \mathrm{tr}\left((uI_n - Z)^{-1}\right) + \mathrm{tr}\left((Z - lI_n)^{-1}\right),$$

and they introduced a sampling procedure based on this single potential function.

```
Potential Function Guided Adaptive Sampling

1. Initialization: $Z_1 \leftarrow 0$, $l_1 \leftarrow -\frac{d}{\varepsilon}$, $u_1 \leftarrow \frac{d}{\varepsilon}$, and $\tau \leftarrow O(\frac{n}{\varepsilon^2})$.
2. For $t \leftarrow 1$ to $\tau$ do
    (a) Let $W_t \leftarrow \Phi^{u_t,l_t}(Z_t)$ be the current potential value.
    (b) Sample a vector $i_t = i \in [m]$ with probability
        $p_i = \frac{w(i)}{W_t} \cdot \left( \langle v_i v_i^\top, (u_t I_n - Z_t)^{-1} \rangle + \langle v_i v_i^\top, (Z_t - l_t I_n)^{-1} \rangle \right).$
```

(c) Let $\Delta_t \leftarrow \frac{w(i_t)}{p_{i_t}} v_{i_t} v_{i_t}^\top$, and update $Z_{t+1} \leftarrow Z_t + \Delta_t$.

(d) $u_{t+1} \leftarrow u_t + \frac{1}{1-W_t}$ and $l_{t+1} \leftarrow l_t + \frac{1}{1+W_t}$.

We make some brief remarks about the algorithm description. First, the expected change of solution in each iteration is $\mathbb{E}[\Delta_t] = \sum_{i=1}^m w(i) \cdot v_i v_i^\top = I_n$. Second, the two terms $\langle v_i v_i^\top, (u_t I_n - Z_t)^{-1} \rangle$ and $\langle v_i v_i^\top, (Z_t - l_t I_n)^{-1} \rangle$ in the sampling probability can be treated as "relative effective resistance". Finally, the above description of the algorithm is not exactly the same as the one in [101], however it captures all the essential elements.

When the potential value W_t is not too large, the change of the potential value (before shifting the barriers) is bounded by

$$\Phi^{u_t,l_t}(Z_t + \Delta_t) \leqslant \Phi^{u_t,l_t}(Z_t) + \frac{\langle \Delta_t, (u_t I_n - Z_t)^{-2} \rangle}{1 - W_t} - \frac{\langle \Delta_t, (Z_t - l_t I_n)^{-2} \rangle}{1 + W_t}.$$

Hence, given the current solution Z_t, the expected change of the potential value can be bounded by

$$\begin{aligned} \mathbb{E}[\Phi^{u_t,l_t}(Z_t + \Delta_t)] - \Phi^{u_t,l_t}(Z_t) &\leqslant \frac{\mathrm{tr}((u_t I_n - Z_t)^{-2})}{1 - W_t} - \frac{\mathrm{tr}((Z_t - l_t I_n)^{-2})}{1 + W_t} \\ &= -\frac{1}{1 - W_t} \cdot \frac{\partial}{\partial u_t} \Phi_{u_t,l_t}(Z_t) - \frac{1}{1 + W_t} \cdot \frac{\partial}{\partial l_t} \Phi_{u_t,l_t}(Z_t). \end{aligned}$$

Notice that if we increase u_t by $\frac{1}{1-W_t}$ and l_t by $\frac{1}{1+W_t}$, then the potential value approximately increases by

$$\frac{1}{1 - W_t} \cdot \frac{\partial}{\partial u_t} \Phi^{u_t,l_t}(Z_t) + \frac{1}{1 + W_t} \cdot \frac{\partial}{\partial l_t} \Phi^{u_t,l_t}(Z_t).$$

This means that we can shift the upper barrier $u_{t+1} \approx u_t + \frac{1}{1-W_t}$ and the lower barrier $l_{t+1} \approx l_t + \frac{1}{1+W_t}$ to maintain that the potential value is not increased in expectation. Therefore, if we start with small enough potential value, then we can maintain that the potential value is small and the gap between $u_t - l_t$ is approximately preserved. If we select a vector deterministically to preserve small potential value in each iteration, then a similar argument as in BSS would go through and the algorithm will return a linear-sized sparsifier.

One of the main advantages of this randomized approach is that we can do batch update for the sampling probabilities. Together with a different but related potential function suggested in [7], Lee and Sun managed to improve the running time of the algorithm to almost-linear time. In a followup work, Lee and Sun [100] finally improved the construction time of spectral sparsifier to nearly-linear time. The key ingredient of the followup work is a completely new potential function, and a SDP-based one-sided sparsifier construction. We are not going to discuss more technical details of this work, as it is beyond our focus in this thesis.

Going back to the potential based adaptive sampling process, the potential value W_t is a random variable. Thus, analyzing the final value of u_τ and l_τ requires more work. However, with a more careful analysis, one can show that u_τ and l_τ are highly concentrated around their expectations, and the expectations give the desired condition number. We will discuss more details of the analysis of adaptive sampling based algorithm in Section 4.3, and provide an algorithm unifying the regret minimization algorithms in [7] and the adaptive sampling algorithm in [101].

Finally, we remark that our iterative randomized rounding algorithm for spectral rounding was inspired by the adaptive sampling idea in [101].

2.6 Interlacing Polynomials

In Chapter 5, we will apply a result from discrepancy theory [91] to the two-sided spectral rounding problem. The result in [91] is based on the method of interlacing polynomials. For completeness, we review some background of interlacing polynomials in this section.

To motivate the techniques of interlacing polynomials for the problems in this thesis, we consider the connections between spectral sparsification and Weaver's conjecture.

Conjecture 2.6.1 (Weaver's Conjecture KS_2 [141])**.** *Given vectors* $\mathsf{u}_1, \ldots, \mathsf{u}_m \in \mathbb{R}^d$ *such that* $\sum_{i=1}^m \mathsf{u}_i \mathsf{u}_i^\top = \mathsf{I}_d$ *and* $\|\mathsf{u}_i\|_2^2 \leqslant \varepsilon$ *for some* $\varepsilon \in (0,1)$ *for all* $i \in [m]$*, there exists a subset*

$S \subseteq [m]$ and a constant $\varepsilon' > 0$ (independent of m and d) such that

$$\left(\frac{1}{2} - \varepsilon'\right) \cdot I_d \preccurlyeq \sum_{i \in S} u_i u_i^\top \preccurlyeq \left(\frac{1}{2} + \varepsilon'\right) \cdot I_d.$$

Comparing with spectral sparsification (Problem 2.5.2), Weaver's conjecture considers a setting where we are not allowed to reweight the vectors, i.e. the output vectors are unreweighted. Another difference is the short vector assumption $\|u_i\|_2^2 \leqslant \varepsilon$ for all $i \in [m]$. To see the assumption is necessary, suppose there is a vector with $\|u_i\|_2^2 \geqslant 1 - \delta$, then the partition contains i would have maximum eigenvalue $\geqslant 1 - \delta$ and the other partition that does not contain i would have minimum eigenvalue $< \delta$.

Weaver [141] showed that the above discrepancy theoretical statement is equivalent to the Kadison-Singer problem, a famous problem in functional analysis with its origin from quantum physics [81]. In 2013, Marcus, Spielman and Srivastava [111] resolved the problem affirmatively using interlacing polynomial techniques and a barrier function argument.

2.6.1 Solution to Kadison-Singer Problem

To solve the Kadison-Singer problem, Marcus, Spielman and Srivastava proved a probabilistic statement summarized in the following theorem.

Theorem 2.6.2 (Theorem 1.4 in [111]). *Let $v_1, \ldots, v_m \in \mathbb{R}^d$ be independent random vectors with finite support such that*

$$\mathbb{E}\left[\sum_{i=1}^{m} v_i v_i^\top\right] = I_d \qquad \textit{and} \qquad \mathbb{E}\big[\, \|v_i\|_2^2 \,\big] \leqslant \varepsilon \;\; \textit{for all } i \in [m],$$

then

$$\mathbb{P}\left[\left\|\sum_{i=1}^{m} v_i v_i^\top\right\|_{\mathrm{op}} \leqslant (1 + \sqrt{\varepsilon})^2\right] > 0.$$

Weaver's conjecture can be derived from Theorem 2.6.2 using the following reduction. Let $u_1, \ldots, u_m \in \mathbb{R}^d$ be the input vectors satisfying the conditions in Conjecture 2.6.1.

We define independent random vectors $v_1, \ldots, v_m \in \mathbb{R}^{2d}$ such that v_i is chosen uniformly at random from the set $\{\binom{\sqrt{2}u_i}{0}, \binom{0}{\sqrt{2}u_i}\}$. We can verify that $\mathbb{E}[\sum_{i=1}^m v_i v_i^\top] = I_{2d}$ and $\|u_i\|_2^2 \leqslant 2\varepsilon$ for all $i \in [m]$. Thus, Theorem 2.6.2 implies that there exists an outcome of v_i's, or a subset $S \subseteq [m]$, such that

$$\left\| \sum_{i=1}^m v_i v_i^\top \right\|_{\mathrm{op}} = \left\| \begin{pmatrix} 2\sum_{i\in S} u_i u_i^\top & \\ & 2\sum_{j\notin S} u_j u_j^\top \end{pmatrix} \right\|_{\mathrm{op}} \leqslant (1+\sqrt{2\varepsilon})^2.$$

Since $\sum_{i=1}^m u_i u_i^\top = I_d$, this implies $(\frac{1}{2} - O(\sqrt{\varepsilon})) \cdot I_d \preccurlyeq \sum_{i\in S} u_i u_i^\top \preccurlyeq (\frac{1}{2} + O(\sqrt{\varepsilon})) \cdot I_d$, which confirms Conjecture 2.6.1.

In the following, without giving rigorous proofs, we describe some key ideas and techniques used in proving Theorem 2.6.2. To establish Theorem 2.6.2, Marcus, Spielman and Srivastava considered the expected characteristic polynomial of the random matrix $\sum_{i=1}^m v_i v_i^\top$, i.e. $\mathbb{E}\left[\chi\left[\sum_{i=1}^m v_i v_i^\top\right](x)\right]$, where $\chi[M](x) := \det(xI_d - M)$ is the *characteristic polynomial* of matrix M.

The plan of proving Theorem 2.6.2 in [111] consists of two steps:

- Step 1: Show that there exists an outcome of v_i's with the largest root of the characteristic polynomial (i.e. $\lambda_{\max}(\sum_{i=1}^m v_i v_i^\top)$) upper bounded by the largest root of the expected characteristic polynomial

$$\text{max-root}\left(\det\left(x \cdot I_d - \sum_{i=1}^m v_i v_i^\top\right)\right) \leqslant \text{max-root}\left(\mathbb{E}\left[\chi\left[\sum_{i=1}^m v_i v_i^\top\right](x)\right]\right).$$

- Step 2: Bound the largest root of the expected characteristic polynomial by

$$\text{max-root}\left(\mathbb{E}\left[\chi\left[\sum_{i=1}^m v_i v_i^\top\right](x)\right]\right) \leqslant (1+\sqrt{\varepsilon})^2.$$

Interlacing Argument for Step 1

Marcus, Spielman and Srivastava [111] gave an equivalent way to describe the expected characteristic polynomial, which is crucial in both Step 1 and Step 2.

Theorem 2.6.3 (Theorem 4.1 in [111]). *Let $v_1, \ldots, v_m \in \mathbb{R}^d$ be independent random vectors with finite support, and $A_i = \mathbb{E}[v_i v_i^\top]$ for $i = 1, \ldots, m$. Then, it holds that*

$$\mathbb{E}\left[\chi\left[\sum_{i=1}^m v_i v_i^\top\right](x)\right] = \mu[A_1, \ldots, A_m](x),$$

where

$$\mu[A_1, \ldots, A_m](x) := \left(\prod_{i=1}^m \left(1 - \partial_{x_i}\right)\right) \det\left(xI + \sum_{i=1}^m x_i A_i\right)\Big|_{x_1=0,\ldots,x_m=0}$$

is referred as the mixed characteristic polynomial *of $A_1, \ldots, A_m$.*

To get some intuition about the theorem, we consider a simpler setting, where there is a deterministic invertible matrix B and a random vector v such that $\mathbb{E}[vv^\top] = A$. It holds that

$$\mathbb{E}\big[\det(B - vv^\top)\big] = \mathbb{E}\Big[\det(B)\Big(1 - \operatorname{tr}\left(B^{-1}vv^\top\right)\Big)\Big] = \det(B)\Big(1 - \operatorname{tr}\left(B^{-1}A\right)\Big),$$

where we used Lemma 2.1.11 for the first equality, and the last equality follows by linearity of expectation and trace. Then, we consider a univariate polynomial in x,

$$\det(B + xA) = \det(B)\det(I + xB^{-1}A) = \det(B) \cdot \prod_{i=1}^d (1 + x\lambda_i),$$

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $B^{-1}A$. We apply the operator $(1 - \partial_x)$ to the above polynomial, and then set $x = 0$, it follows that

$$(1 - \partial_x)\det(B + xA)\mid_{x=0} = \det(B)\left(1 - \sum_{i=1}^d \lambda_i\right) = \det(B)\Big(1 - \operatorname{tr}\left(B^{-1}A\right)\Big).$$

Thus, we have established that, for invertible B, it holds that

$$\mathbb{E}\big[\det(B - vv^\top)\big] = (1 - \partial_x)\det(B + x\mathbb{E}[vv^\top])\mid_{x=0}.$$

The key observation here is that, when A is rank-one, $\det(B + xA)$ is a affine-linear function in x. By a continuity argument, we can show the conclusion still holds for singular B. Theorem 2.6.3 follows by applying this argument repeatedly together with the independence of v_i's. The following is a simple observation based on a similar argument.

Lemma 2.6.4. *Let $A_i = \mathbb{E}[v_i v_i^\top]$ for all $i \in [m]$. The mixed characteristic polynomial $\mu[A_1, A_2, \ldots, A_m](x)$ is a convex combination of $\{\mu[w_1 w_1^\top, A_2, \ldots A_m](x)\}_{w_1}$, where $w_1 \in$* $\operatorname{support}(v_1)$.

The mixed characteristic polynomial characterization is important as it connects the expected characteristic polynomial with a key notion called stable polynomials.

Definition 2.6.5 ((Real) Stable Polynomials)**.** *A multivariate polynomial $p(x_1, \ldots, x_m)$ is* stable *if $p(x_1, \ldots, x_m)$ has no root in the region $\{(x_1, \ldots, x_m) : \mathrm{Im}(z_i) > 0$ for all $i \in [m]\}$. $p(x_1, \ldots, x_m)$ is* real stable *if p is stable and all the coefficients of p are real.*

The definition directly implies a univariate polynomial is real stable if and only if it is real-rooted. The following is an important example of real stable polynomial.

Lemma 2.6.6 (Proposition 2.4 in [26])**.** *The polynomial $p(x_1, \ldots, x_m) = \det(\sum_{i=1}^m x_i A_i)$ is real stable for $A_1, \ldots, A_m \succcurlyeq 0$.*

Many operations can preserve real stability. Here are two related examples.

Lemma 2.6.7 (see, e.g., [139])**.** *Let $p(x_1, \ldots, x_m)$ be a real stable polynomial with m variables.*

- *(Restrictions): For any $a \in \mathbb{R}$, $p(a, x_2, \ldots, x_m)$ is a real stable polynomial with $m-1$ variables.*
- *(Differentiation): For any $t \in \mathbb{R}$, $(1 + t\partial_{x_1})p(x_1, \ldots, x_m)$ is a real stable polynomial with m variables.*

Using the above facts, the mixed characteristic polynomial $\mu[A_1, \ldots, A_m](x)$ is real stable (or real-rooted) for positive semidefinite matrices $A_1, \ldots, A_m$. Note that the statement works for any choices of PSD matrices $A_1, \ldots, A_m$. Therefore, fixing the outcome of $v_1, \ldots, v_k$ being $w_1, \ldots, w_k$ for some $k \in [m]$, the mixed characteristic polynomial $\mu[w_1 w_1^\top, \ldots, w_k w_k^\top, A_{k+1}, \ldots, A_m](x)$ is also real-rooted. Together with Lemma 2.6.4, the real-rooted polynomial $\mu[A_1,\ldots,A_m](x)$ is a convex combination of real-rooted polynomials $\{\mu[w_1w_1^\top,A_2,\ldots,A_m](x)\}_{w_1}$, where w_1's come from the support of v_1. Therefore, the following lemma implies that there exists an outcome of v_1, say w, such that the largest root of $\mu[ww^\top,A_2,\ldots,A_m](x)$ is upper bounded by $\mu[A_1,\ldots,A_m](x)$.

Lemma 2.6.8 (see, e.g., [61, 49, 45])**.** *Let $p_1(x),\ldots,p_k(x)$ be univariate real-rooted polynomials of the same degree with positive leading coefficients. For $\lambda_1,\ldots,\lambda_k \geqslant 0$ and $\sum_{i=1}^k \lambda_i = 1$, if the convex combination $p = \sum_{i=1}^k \lambda_i p_i$ is also real-rooted, then there exists an $i \in [k]$ such that the largest root of p_i is at most the largest root of p.*

Apply the same argument inductively, we can show that there exists outcome v_i's such that the largest root of $\mu[v_1v_1^\top,\ldots,v_mv_m^\top](x) = \det(xI - \sum_{i=1}^m v_iv_i^\top)$ is at most the largest root of $\mu[A_1,\ldots,A_m](x)$. This finishes the high level description of the proof of Step 1. Marcus, Spielman and Srivastava introduced a new notion "interlacing family of polynomials" to formally prove Step 1, see [111] for more details. Notice that Step 1 is not constructive, as we don't know how to evaluate the largest root of a mixed characteristic polynomial efficiently in general. In each level, we can only guarantee that there exists a polynomial in the next level having a smaller root, but cannot efficiently identify it.

Multivariate Barrier Argument for Step 2

Marcus, Spielman and Srivastava [111] generalized the barrier potential function argument in [21] to bound the largest root of the mixed characteristic polynomial $\mu[A_1,\ldots,A_m](x)$.

We start with defining a sequence of polynomials with $m+1$ variables $x, x_1, \ldots, x_m$.

$$P_k(x,x_1,\ldots,x_m) := \left(\prod_{i=1}^{k}\left(1-\partial_{x_i}\right)\right)\det\left(xI+\sum_{i=1}^{m}x_iA_i\right), \qquad \text{for all } k=0,1,\ldots,m.$$

Note that $P_m(x,0,\ldots,0)$ is exactly the mixed characteristic polynomial $\mu[A_1,\ldots,A_m](x)$.

The plan is to inductively prove that some $z_k = (z_0^k, z_1^k, \ldots, z_m^k) \in \mathbb{R}^{m+1}$ is *above* the roots of $P_k(x,x_1,\ldots,x_m)$, i.e. $P_k(z) > 0$ for any $z \geqslant z_k$. We denote the set of all points above the roots of p by $\mathbf{Ab}_p$. The final goal is to prove that $z_m = (z^*, 0, \ldots, 0)$

is above the roots of $P_m(x, x_1, \ldots, x_m)$, which is enough to show the largest root of the $\mu[A_1, \ldots, A_m](x)$ is upper bounded by z^*. We remark that this plan is slightly different from the original presentation in [111]. We follow the presentation in [91] in order to make the later comparison more clear.

The candidate choice of z_k is

$$z_k = (t + \delta, \underbrace{0, \ldots, 0}_{k \text{ zeros}}, \underbrace{-\delta, \ldots, -\delta}_{(m-k) \text{ of } -\delta\text{'s}}) \qquad \text{for some } t > 0 \text{ and } \delta > 0.$$

For the base case, it is easy to verify that for any $y = (y, y_1, \ldots, y_m) \geqslant 0$,

$$P_0(z_0 + y) = \det\left((t + \delta + y)I + \sum_{i=1}^{m}(y_i - \delta)A_i\right) = \det\left(\sum_{i=1}^{m}(t + y + y_i)A_i\right) \geqslant \det(tI) > 0,$$

where we used $\sum_{i=1}^m A_i = \sum_{i=1}^m \mathbb{E}[v_i v_i^\top] = I$ and all $A_1, \ldots, A_m \succcurlyeq 0$. Thus, the base case holds. To proceed by induction, we need to quantify the impact of the operator $(1 - \partial_{x_k})$ on the upper barrier z_k, i.e. how much do we need to shift the upper barrier. This was done by introducing a notion of multivariate barrier function. Let $p(x_1, \ldots, x_m)$ be a real stable polynomial, and $z \in \mathbf{Ab}_p$. Define the barrier function of p at the direction i at z as

$$\Phi_p^i(z) := \partial_{x_i} \log p(z) = \frac{\partial_{x_i} p(z)}{p(z)}. \tag{2.14}$$

The key property of this multivariate barrier function (see [111] for a proof) is that, when the polynomial p is real stable and $z \in \mathbf{Ab}_p$, then the univariate function

$$\Phi_p^i(z + te_j) : t \to \mathbb{R} \text{ is non-negative, non-increasing, and convex for } t \geqslant 0.$$

Let's consider a simple example to get some intuition. Let $A \in \mathbb{S}^d$ be a symmetric matrix with eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_d$, and let $p(x) = \det(xI - A) = \prod_{i=1}^d (x - \lambda_i)$ be the univariate characteristic polynomial of A. Then, the barrier function $\Phi_p(x) = \sum_{i=1}^d \frac{1}{x - \lambda_i}$ is exactly the barrier potential function used in [21]. This function is non-negative, non-increasing and convex for $x > \lambda_1$ the largest eigenvalue of A. In particular, the value of the barrier function blows up when x is approaching the boundary of $\mathbf{Ab}_p$. This intuition carries over to the multivariate barrier function. If we can keep the value of $\Phi_p^i(z)$ small, then z is far away from the boundary of $\mathbf{Ab}_p$ in the i-th direction, we have more flexibility to modify the polynomial p.

Lemma 2.6.9 (Lemma 5.9 and 5.10 in [111])**.** *Suppose $p(x_1, \ldots, x_m)$ is real stable, $\mathsf{z} \in \mathbf{Ab}_p$, and $\delta > 0$ satisfies $\Phi_p^j(\mathsf{z}) \leqslant 1 - \delta^{-1}$ for some $j \in [m]$. Then, for all $i \in [m]$, it holds that $\Phi^i_{p-\partial_{x_j}p}(\mathsf{z} + \delta \mathsf{e}_j) \leqslant \Phi_p^i(\mathsf{z})$. Furthermore, $\mathsf{z} + \delta \mathsf{e}_j$ is above the roots of $(1 - \partial_{x_j})p$.*

The above lemma in [111] says that, if the barrier function of p at the direction j at z is small enough (measured in terms of δ), i.e. z is far away from the boundary of $\mathbf{Ab}_p$ in the j-th direction, then moving the upper barrier z by δ in the j-th direction suffices to maintain a valid upper barrier and also to guarantee the barrier functions of $(1 - \partial_{x_j})p$ are not increased in all directions. The proof of this lemma crucially relies on the monotonicity and convexity of the multivariate barrier function, which are guaranteed by the real stability of polynomials $P_0, \ldots, P_m$.

Thus, it remains to set the parameters t and δ such that $t + \delta$ is as small as possible and the induction proof based on Lemma 2.6.9 can go through. By setting $t = \varepsilon + \sqrt{\varepsilon}$ and $\delta = 1 + \sqrt{\varepsilon}$, we can make sure that $\Phi^i_{P_0}(\mathsf{z}_0) \leqslant 1 - \delta^{-1}$ for all $i \in [m]$. Thus, applying Lemma 2.6.9 repeatedly, we can finally conclude that $t + \delta = (1 + \sqrt{\varepsilon})^2$ is larger than any root of the mixed characteristic polynomial $\mu[\mathsf{A}_1, \ldots, \mathsf{A}_m](x)$, and establish Step 2.

2.6.2 Kyng, Luh and Song's Extension

Kyng, Luh and Song proved the following theorem in [91].

Theorem 2.6.10 (Theorem 1.4 in [91])**.** *Let $\mathsf{v}_1, \ldots, \mathsf{v}_m \in \mathbb{R}^d$, and $\xi_1, \ldots, \xi_m$ be independent random scalar variables with finite support. There exists a choice of outcomes $\epsilon_1, \ldots, \epsilon_m$ in the support of $\xi_1, \ldots, \xi_m$ such that*

$$\left\| \sum_{i=1}^m \mathbb{E}\left[\xi_i\right] \cdot \mathsf{v}_i \mathsf{v}_i^\top - \sum_{i=1}^m \epsilon_i \cdot \mathsf{v}_i \mathsf{v}_i^\top \right\|_{\mathrm{op}} \leqslant 4 \left\| \sum_{i=1}^m \mathbf{Var}[\xi_i] (\mathsf{v}_i \mathsf{v}_i^\top)^2 \right\|_{\mathrm{op}}^{\frac{1}{2}}.$$

The above theorem does not directly generalize Theorem 2.6.2 since the rank-one random matrices involved in Theorem 2.6.10 always have a rank-one expectation, while the rank-one random matrices can have arbitrary PSD expectation in Theorem 2.6.2. However, the above theorem is sufficient to prove Weaver's Conjecture. Let $\mathsf{v}_1, \ldots, \mathsf{v}_m \in \mathbb{R}^d$ satisfy the conditions in Conjecture 2.6.1, i.e. $\sum_{i=1}^m \mathsf{v}_i \mathsf{v}_i^\top = I$ and $\|\mathsf{v}_i\|_2^2 \leqslant \varepsilon$. Let ξ_i be independent uniform random variable over $\{\pm 1\}$ for all $i \in [m]$. Then, Theorem 2.6.10 implies there exists $\epsilon_i \in \{\pm 1\}$ for all $i \in [m]$ such that

$$\left\| \sum_{i=1}^m \epsilon_i \cdot \mathsf{v}_i \mathsf{v}_i^\top \right\|_{\mathrm{op}} \leqslant 4 \left\| \sum_{i=1}^m \mathbf{Var}[\xi_i] (\mathsf{v}_i \mathsf{v}_i^\top)^2 \right\|_{\mathrm{op}}^{1/2} \leqslant 4 \left\| \max_{i \in [m]} \|\mathsf{v}_i\|_2^2 \cdot \sum_{i=1}^m \mathbf{Var}[\xi_i] \mathsf{v}_i \mathsf{v}_i^\top \right\|_{\mathrm{op}}^{1/2} \leqslant 4\sqrt{\varepsilon},$$

where the last inequality follows as $\mathbf{Var}[\xi_i] = 1$, $\mathsf{v}_i \mathsf{v}_i^\top \leqslant \varepsilon I$ and $\sum_{i=1}^m \mathsf{v}_i \mathsf{v}_i^\top = I$. This further implies the existence of the partition required by Conjecture 2.6.1, with slightly worse constant than [111].

While Theorem 2.6.2 requires a normalized expectation I_d, Theorem 2.6.10 allows the sum of rank-one random matrices to have expectations in arbitrary forms. This gives us some more flexibility in applications, e.g., the two-sided spectral rounding in Chapter 5. Theorem 2.6.10 also allows us to have a refined control of the deviation from the expectation by incorporating individual $\mathbf{Var}[\xi_i]$ and $\|\mathsf{v}_i\|_2^2$ into the upper bound.

In the following, we denote $\sigma^2 = \left\| \sum_{i=1}^m \mathbf{Var}[\xi_i] (\mathsf{v}_i \mathsf{v}_i^\top)^2 \right\|_{\mathrm{op}}$, and $\mu_i := \mathbb{E}[\xi_i]$ and $\tau_i := \sqrt{\mathbf{Var}[\xi_i]}$ for all $i \in [m]$.

To prove Theorem 2.6.10, Kyng, Luh and Song followed a similar two-step framework as in [110, 111] with several new ideas. The first challenge is that the framework in [110, 111] cannot deal with the largest and smallest eigenvalues simultaneously. A simple and nice idea in [91] to overcome this difficulty is, instead of reasoning the characteristic polynomial of $\sum_{i=1}^m (\xi_i - \mu_i) \mathsf{v}_i \mathsf{v}_i^\top$, they reason about the polynomial

$$\det\left(x^2 I - \left(\sum_{i=1}^m (\xi_i - \mu_i) \mathsf{v}_i \mathsf{v}_i^\top \right)^2 \right) = \det\left(xI - \sum_{i=1}^m (\xi_i - \mu_i) \mathsf{v}_i \mathsf{v}_i^\top \right) \cdot \det\left(xI + \sum_{i=1}^m (\xi_i - \mu_i) \mathsf{v}_i \mathsf{v}_i^\top \right),$$

where the largest root is exactly the operator norm of $\sum_{i=1}^m (\xi_i - \mu_i) \mathsf{v}_i \mathsf{v}_i^\top$. The expectation of the above polynomial has the following form.

Proposition 2.6.11 (Proposition 3.3 in [91]). *Let $\mathsf{v}_1, \ldots, \mathsf{v}_m \in \mathbb{R}^d$, let ξ_i be independent* *random variable with mean μ_i and variance τ_i^2, then*

$$\mathbb{E}\left[\det\left(x^2 I - \left(\sum_{i=1}^{m}(\xi_i - \mu_i)\mathsf{v}_i\mathsf{v}_i^\top\right)^2\right)\right]$$

$$= \left(\prod_{i=1}^{m}\left(1 - \frac{1}{2}\partial_{x_i}^2\right)\right)\det\left(xI + \sum_{i=1}^{m} x_i\tau_i\mathsf{v}_i\mathsf{v}_i^\top\right)^2\Bigg|_{x_1=\cdots=x_m=0}. \qquad (2.15)$$

A key observation here is that $(1-\partial_x^2) = (1+\partial_x)(1-\partial_x)$ is also a real stability preserving operation. Thus, the expected polynomial in (2.15) is real stable. With similar interlacing argument in Step 1 of the proof for Theorem 2.6.2, there exists an outcome ϵ_i in the support of ξ_i for all $i \in [m]$ such that $\left\|\sum_{i=1}^m(\epsilon_i - \mu_i)\mathsf{v}_i\mathsf{v}_i^\top\right\|_{\mathrm{op}}$ is at most the largest root of (2.15).

To upper bound the largest root of (2.15), again we define some intermediate polynomials with $m+1$ variables $x, x_1, \ldots, x_m$.

$$P_k(x, x_1, \ldots, x_m) := \left(\prod_{i=1}^{k}\left(1 - \frac{1}{2}\partial_{x_i}^2\right)\right)\det\left(xI + \sum_{i=1}^{m} x_i\tau_i\mathsf{v}_i\mathsf{v}_i^\top\right)^2, \qquad \text{for all } i \in [m].$$

The goal is to show, for all $k = 0, \ldots, m$,

$$\mathsf{z}_k = (4\sigma, \underbrace{0, \ldots, 0}_{k \text{ zeros}}, -\delta_{k+1}, \ldots, -\delta_m) \qquad (2.16)$$

is above the roots of P_k inductively, where $\delta_i = \frac{2\tau_i}{\sigma}\|\mathsf{v}_i\|_2^2$ for all $i \in [m]$. In particular, if this is true for $k = m$, then we can conclude that the largest root of (2.15) is upper bounded by 4σ.

The induction proof works similarly as Step 2 in the proof of Theorem 2.6.2. Kyng, Luh and Song used the same multivariate barrier functions defined in (2.14) to guide the induction steps. The key difference is that they need to deal with the operator $1 - \frac{1}{2}\partial_{x_i}^2$ instead of $1 - \partial_{x_i}$.

This operator has been studied by Anari and Oveis Gharan [8] for another variant of Theorem 2.6.2, where the random vectors v_i's are associated with an underlying strong Rayleigh distribution. As observed by Anari and Oveis Gharan, informally, the effect of

$(1-\partial_{x_i})$ is up shifting the barrier by $1+\Theta(\delta_i)$, and the effect of $(1+\partial_{x_i})$ is down shifting the barrier by $1-\Theta(\delta_i)$, thus the total effect of $(1-\partial_{x_i}^2)=(1-\partial_{x_i})(1+\partial_{x_i})$ is shifting the barrier by at most $\pm\Theta(\delta_i)$. Note that the multivariate barrier function analysis in [111] cannot tolerant any shift smaller than 1. This shows that the operator $(1-\partial_{x_i}^2)$ is crucial to enable the small shift δ_i of the barriers z_k in (2.16).

Using a similar analysis in [8], Kyng, Luh and Song proved an analog of Lemma 2.6.9.

Lemma 2.6.12 (Lemma 5.3 in [91])**.** *Suppose $p(x_1,\ldots,x_m)$ is real stable and $\mathsf{z}\in\mathbf{Ab}_p$. If $\Phi_p^j(\mathsf{z})<\sqrt{2}$, then $\mathsf{z}\in\mathbf{Ab}_{(1-\frac{1}{2}\partial_{x_j}^2)p}$. Further, if $\delta^{-1}\Phi_p^j(\mathsf{z})+\frac{1}{2}\Phi_p^j(\mathsf{z})^2\leqslant 1$ for some $\delta>0$, then $\Phi^i_{(1-\frac{1}{2}\partial_{x_j}^2)p}(\mathsf{z}+\delta\mathsf{e}_j)\leqslant\Phi_p^i(\mathsf{z})$ for all $i\in[m]$.*

This lemma guarantees that a similar induction as Step 2 in the proof of Theorem 2.6.2 goes through for the choices of upper barriers z_k's in (2.16), which suffices to establish Theorem 2.6.10.

Chapter 3

Concentration Inequalities

In probability theory, *concentration inequalities* aim at analyzing the probability of a random variable deviating from certain value (typically its expectation). In this thesis, we crucially rely on concentration inequalities to analyze several randomized algorithms in Chapter 4, Chapter 5, Chapter 7, and Chapter 8. In this chapter, we introduce those concentration inequalities that will be used.

We start with an elementary yet useful inequality.

Fact (Markov's Inequality)**.** *Let X be a nonnegative random variable. For any $\delta > 0$,*

$$\mathbb{P}\left[X \geqslant \delta\right] \leqslant \frac{\mathbb{E}\left[X\right]}{\delta}.$$

When we take $\delta = c \cdot \mathbb{E}\left[X\right]$ for some $c \geqslant 1$, the probability of X deviating from the expectation by a factor of c can be bounded by $\frac{1}{c}$.

Markov's inequality could be tight in general. However, when X has some nice properties, we can derive stronger, even exponential, tail bounds. A generic approach to derive an exponential tail bound is based on the following observation,

$$\mathbb{P}\left[X \geqslant \delta\right] = \mathbb{P}\left[e^{\lambda X} \geqslant e^{\lambda \delta}\right] \leqslant \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda \delta}},$$

for any $\lambda > 0$, where the inequality follows from Markov's inequality. By optimizing over $\lambda > 0$, we get a tail bound known as Chernoff inequality,

$$\mathbb{P}\left[X \geqslant \delta\right] \leqslant \inf_{\lambda > 0} \left\{ e^{-\lambda\delta} \cdot \mathbb{E}\left[e^{\lambda X}\right] \right\}. \tag{3.1}$$

The term $\mathbb{E}\left[e^{\lambda X}\right]$ can be treated as a function in λ, which is known as *moment generating function* of the random variable X. Bounding the moment generating function is a key step in deriving a good tail bound.

3.1 Sum of Independent Random Variables

One of the most well-studied and well-understood setting in concentration inequalities is X being the sum of independent random variables, i.e. $X = \sum_{i=1}^{n} X_i$, where $X_1, \ldots, X_n$ are independent random variables. In this setting, due to the independence, it suffices to consider the moment generating function of each random variable X_i separately, as the moment generating function of X can be written as

$$\mathbb{E}\left[e^{\lambda X}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right]. \tag{3.2}$$

Different approaches in bounding $\mathbb{E}\left[e^{\lambda X_i}\right]$ provide different (sometimes incomparable) tail bounds. For example, when X_i's are i.i.d. 0-1 random variables, we can bound $\mathbb{E}\left[e^{\lambda X_i}\right]$ by

$$\mathbb{E}\left[e^{\lambda X_i}\right] = (1 - \mathbb{E}[X_i]) + \mathbb{E}[X_i] \cdot e^{\lambda} = 1 + (e^{\lambda} - 1) \cdot \mathbb{E}[X_i] \leqslant \exp((e^{\lambda} - 1) \cdot \mathbb{E}[X_i]), \tag{3.3}$$

where we used $1 + p \leqslant e^{p}$ for any $p \in \mathbb{R}$ for the last inequality. Based on the above bound, we can derive the well-known Chernoff bound.

Theorem 3.1.1 (Chernoff Bound, see, e.g., [27])**.** *Let $X_1, \ldots, X_n \in \{0, 1\}$ be i.i.d. random variables. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X]$, then for any $\delta > 0$,*

$$\mathbb{P}\left[X \geqslant (1+\delta)\mu\right] \leqslant \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \qquad \text{for } \delta > 0, \qquad \text{and}$$

$$\mathbb{P}\left[X \leqslant (1-\delta)\mu\right] \leqslant \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} \qquad \text{for } \delta \in [0, 1].$$

Proof. By independence of X_i's and (3.3), for any $\lambda > 0$ it holds that

$$\mathbb{E}[e^{\lambda X}] \leqslant \exp\left((e^{\lambda} - 1) \cdot \sum_{i=1}^{n} \mathbb{E}[X_i]\right) \leqslant \exp((e^{\lambda} - 1) \cdot \mu).$$

By Chernoff inequality (3.1),

$$\mathbb{P}\left[X \geqslant (1+\delta)\mu\right] \leqslant \inf_{\lambda>0}\{e^{-(1+\delta)\mu} \cdot \mathbb{E}[e^{\lambda X}]\} \leqslant \inf_{\lambda>0} e^{-\lambda(1+\delta)\mu + (e^{\lambda}-1)\mu}.$$

Optimizing λ over the function $(e^{\lambda} - 1 - \lambda(1+\delta))\mu$, we obtain the first bound in the theorem by taking $\lambda = \ln(1+\delta)$. With a similar argument, for any $\delta \in [0,1]$, we can prove the second bound in the theorem. □

In a more general setting, Hoeffding's lemma addresses the case where each X_i is bounded with a slightly more sophisticated argument than (3.3).

Lemma 3.1.2 (Hoeffding's Lemma, see, e.g., [27])**.** *Let X be a real random variable such that $a \leqslant X \leqslant b$, then for any $\lambda > 0$*

$$\mathbb{E}\left[e^{\lambda X}\right] \leqslant e^{\lambda \mathbb{E}[X] + \frac{1}{8}\cdot\lambda^2 (b-a)^2}.$$

Based on Hoeffding's Lemma, the following widely applied Hoeffding bound is just a simple consequence, which can be proved with a similar treatment as in the proof of Theorem 3.1.1.

Theorem 3.1.3 (Hoeffding's Inequality, see, e.g., [27])**.** *Let $X_1, \ldots, X_n \in [0, R]$ be n independent random variables. Let $X = \sum_{i=1}^{n} X_i$, and $\mu = \mathbb{E}[X]$, then for any $\delta > 0$,*

$$\mathbb{P}\left[X \geqslant (1+\delta)\mu\right] \leqslant e^{-\frac{2\delta^2\mu^2}{nR^2}}.$$

As another example, Bernstein's inequality bounds the term $\mathbb{E}[e^{\lambda X_i}]$ using the the variance of X_i explicitly, which produces a different tail bound. Usually, Bernstein's inequality has the assumption $|X_i| \leqslant R$, we learnt the idea of relaxing the lower bound $X_i \geqslant -R$ from [137]. We provide a proof for Bernstein's inequality in a more general setting in the next section (see Theorem 3.2.3).

Theorem 3.1.4 (Bernstein's Inequality, see, e.g., [27]). *Let $X_1, \ldots, X_n$ be n independent random variables such that $\mathbb{E}[X_i] = 0$ and $X_i \leqslant R$ for some $R > 0$, $\forall i \in [n]$. Let $X = \sum_{i=1}^n X_i$ and $\sigma^2 = \sum_{i=1}^n \mathbb{E}[X_i^2]$, then for any $\delta > 0$*

$$\mathbb{P}\left[X \geqslant \delta\right] \leqslant e^{-\frac{\delta^2/2}{\sigma^2+\delta R/3}}.$$

We may notice that Hoeffiding bound and Bernstein's inequality are not comparable, i.e. we cannot say one is strictly better than the other. When the sum of variance $\sigma^2 \approx nR^2$, then Hoeffiding bound is better. When $\sigma^2 \ll nR^2$ and t is not too large, then Bernstein's inequality is better.

3.2 Martingales

Many of the concentration inequalities for sum of independent random variables can be extended to the setting where the random variables are weakly correlated. One particular useful and well-studied setting is about martingales.

Definition 3.2.1 (Martingale). *A sequence of random variables $Y_1, \ldots, Y_t, \ldots$ is a* martingale *with respect to a sequence of random variables $Z_1, \ldots, Z_t, \ldots$ if for all $t > 0$, it holds that*

1. *Y_t is a function of $Z_1, \ldots, Z_{t-1}$;*
2. *$\mathbb{E}\left[|Y_t|\right] < \infty$;*
3. *$\mathbb{E}\left[Y_{t+1} - Y_t | Z_1, \ldots, Z_{t-1}\right] = 0$.*

The sequence $\{X_t = Y_t - Y_{t-1}\}_t$ is known as the difference sequence *of martingale $\{Y_t\}_t$.*

Note that if the difference sequence $\{X_t\}_t$ is formed by independent random variables with finite expectation then $\{Y_t\}_t$ is a martingale. An important observation to generalize the techniques used in concentration inequalities for sum of independent random variables to the martingale setting is that

$$\mathbb{E}\left[e^{\lambda \sum_{t=1}^{\tau} X_t}\right] = \mathbb{E}\left[e^{\lambda \sum_{t=1}^{\tau-1} X_t} \cdot \mathbb{E}[e^{\lambda X_\tau} \mid X_1, \ldots, X_{\tau-1}]\right]. \tag{3.4}$$

Therefore, if we can bound the conditional expectation $\mathbb{E}[e^{\lambda X_t} \mid X_1, \ldots, X_{t-1}]$, then we will still be able to bound the tails using Chernoff inequality (3.1).

Based on this observation, several concentration inequalities for the sum of independent variables have their counterparts in martingale settings. Freedman's inequality is one of the most frequently used martingale inequalities, which generalizes Bernstein's inequality Theorem 3.1.4.

Theorem 3.2.2 (see, e.g., [63, 137])**.** *Let $\{Y_t\}_t$ be a real-valued martingale with respect to $\{Z_t\}_t$, and $\{X_t = Y_t - Y_{t-1}\}_t$ be the difference sequence. Assume that $X_t \leqslant R$ deterministically for all $t \geqslant 1$. Let $W_t := \sum_{j=1}^{t} \mathbb{E}[X_j^2|Z_1, ..., Z_{j-1}]$ for $t \geqslant 1$. Then, for all $\delta \geqslant 0$ and $\sigma^2 > 0$,*

$$\mathbb{P}\left(\exists t \geqslant 1 : Y_t \geqslant \delta \text{ and } W_t \leqslant \sigma^2\right) \leqslant \exp\left(\frac{-\delta^2/2}{\sigma^2 + R\delta/3}\right).$$

Informally, Freedman's inequality says the martingale $\{Y_t\}_t$ deviates from 0 (the expectation) only when the difference sequence accumulates large enough "energy" (i.e. variance) in the random process. Note that, in Theorem 3.2.2, W_t is the sum of conditional variances of X_j's, which is a random variable. To prove the exact statement (i.e. $\exists t \geqslant 1$) in Theorem 3.2.2, we need to invoke some martingale specific stopping time argument. However, in this thesis we mainly care about the deviation of Y_τ at some fixed time step τ. Thus, instead of proving the original Freedman's inequality, we provide a proof for the following simplified version, which is sufficient for our applications.

Theorem 3.2.3. *Let $\{Y_t\}_t$ be a real-valued martingale with respect to $\{Z_t\}_t$, and $\{X_t = Y_t - Y_{t-1}\}_t$ be the difference sequence. Suppose we are given some fixed $R > 0$ and $\sigma_t^2 \geqslant 0$ for $1 \leqslant t \leqslant \tau$. Assume $X_t \leqslant R$ and $\mathbb{E}\left[X_t^2|Z_1, ..., Z_{t-1}\right] \leqslant \sigma_t^2$ deterministically for $1 \leqslant t \leqslant \tau$, and further assume that $\sigma^2 := \sum_{t=1}^{\tau} \sigma_t^2 > 0$. Then, for all $\delta \geqslant 0$,*

$$\mathbb{P}\left[Y_\tau \geqslant \delta\right] \leqslant \exp\left(\frac{-\delta^2/2}{\sigma^2 + R\delta/3}\right).$$

Proof. For any $\lambda > 0$ and $t \in [\tau]$,

$$e^{\lambda X_t} = 1 + \lambda X_t + f_\lambda(X_t) \cdot X_t^2, \quad \text{where} \quad f_\lambda(x) := \begin{cases} \frac{e^{\lambda x} - 1 - \lambda x}{x^2}, & x \neq 0 \\ \frac{\lambda^2}{2}, & x = 0 \end{cases}.$$

We prove that $f_\lambda(x)$ is monotone increasing in x by showing the derivative $f'_\lambda(x) > 0$ for all $x \in \mathbb{R}$. Note that,

$$f'_\lambda(0) = \frac{\lambda^3}{3!} > 0 \qquad \text{and} \qquad f'_\lambda(x) = \frac{x(e^x + 1) - 2(e^x - 1)}{x^3} \qquad \text{for } x \neq 0.$$

Let $g(x) := x(e^x + 1) - 2(e^x - 1)$. To prove $f_\lambda(x)$ is monotone increasing, it suffices to show that $g(x) > 0$ for all $x > 0$ and $g(x) < 0$ for all $x < 0$. Since $g(0) = 0$, it is enough to show that $g'(x) > 0$ for all $x \neq 0$. This follows from

$$g'(0) = 0 \quad \text{and} \quad g''(x) = xe^x \begin{cases} > 0, & \text{when } x > 0 \\ < 0, & \text{when } x < 0 \end{cases}.$$

Thus, we have established the monotonicity of $f'_\lambda(x)$.

In the following, we denote $\mathbb{E}_t[\cdot]$ as $\mathbb{E}[\cdot \mid Z_1, \ldots, Z_{t-1}]$ for simplicity. Since $X_t \leqslant R$ and $\mathbb{E}_t[X_t] = 0$, it holds that

$$\begin{aligned} &e^{\lambda X_t} \leqslant 1 + \lambda X_t + f_\lambda(R) \cdot X_t^2 \\ &\qquad \Longrightarrow \quad \mathbb{E}_t\left[e^{\lambda X_t}\right] \leqslant 1 + f_\lambda(R) \cdot \mathbb{E}_t[X_t^2] \leqslant \exp(f_\lambda(R) \cdot \mathbb{E}_t[X_t^2]) \leqslant \exp(f_\lambda(R) \cdot \sigma_t^2), \end{aligned}$$

where the second last inequality follows by $1 + p \leqslant e^p$ for all $p \in \mathbb{R}$, and the last inequality follows by the assumption $\mathbb{E}_t[X_t^2] \leqslant \sigma_t^2$.

Applying the observation (3.4) repeatedly, we have

$$\mathbb{E}[e^{\lambda Y_\tau}] = \mathbb{E}\Big[e^{\lambda Y_{\tau-1}} \cdot \mathbb{E}_\tau[e^{\lambda X_\tau}]\Big] \leqslant \mathbb{E}[e^{\lambda Y_{\tau-1}}] \cdot e^{f_\lambda(R) \cdot \sigma_\tau^2} \leqslant \cdots \leqslant e^{f_\lambda(R) \cdot \sum_{t=1}^{\tau} \sigma_t^2} = e^{f_\lambda(R) \cdot \sigma^2}.$$

Then, by Chernoff inequality (3.1),

$$\mathbb{P}\left[Y_\tau \geqslant \delta\right] \leqslant \inf_{\lambda > 0} \exp\left(-\lambda\delta + f_\lambda(R) \cdot \sigma^2\right).$$

Optimizing λ over the function $-\lambda\delta + f_\lambda(R) \cdot \sigma^2$, we take $\lambda = \frac{1}{R}\ln(1 + \frac{\delta R}{\sigma^2})$ (notice that $\frac{\delta R}{\sigma^2} \geqslant 0$ by our assumption). It follows that

$$\mathbb{P}\left[Y_\tau \geqslant \delta\right] \leqslant \exp\left(-\frac{\delta}{R}\ln\left(1 + \frac{\delta R}{\sigma^2}\right) + \frac{\sigma^2}{R^2}\left(\frac{\delta R}{\sigma^2} - \ln\left(1 + \frac{\delta R}{\sigma^2}\right)\right)\right) = e^{-\frac{\sigma^2}{R^2}\cdot h\left(\frac{\delta R}{\sigma^2}\right)},$$

where $h(x) := (1+x)\ln(1+x) - x$. We can verify that $h(x) \geqslant f(x) := \frac{x^2}{2(1+x/3)}$ for $x \geqslant 0$ by a similar argument at the beginning of the proof. More specifically, we can check $h''(x) - f''(x) = \frac{1}{1+x} - \frac{27}{(x+3)^3} > 0$ for all $x > 0$. This shows $h'(x) - f'(x)$ is monotone increasing when $x > 0$. Since $h'(0) = f'(0) = 0$ and $h(0) = f(0) = 0$, $h(x) - f(x)$ is also monotone increasing and nonnegative for all $x \geqslant 0$. Therefore, we finish the proof by observing that

$$\mathbb{P}\left[Y_\tau \geqslant \delta\right] \leqslant e^{-\frac{\sigma^2}{R^2}\cdot h\left(\frac{\delta R}{\sigma^2}\right)} = e^{-\frac{\delta^2/2}{\sigma^2 + \delta R/3}}. \qquad \square$$

3.3 Concentration Inequality for Self-adjusting Random Process

Recently, some variants of Freedman's inequality for martingales have been applied to obtain algorithmic discrepancy results [18, 17]. For the applications in this thesis, we prove another variant which applies to non-martingales with a "self-adjusting" property, that if Y_t is (more) positive then $E[Y_{t+1}] - Y_t$ is (more) negative and vice versa. With this self-adjusting property, intuitively Y_t cannot be too far away from zero, and the following theorem provides a quantitative bound that is similar to that in Freedman's inequality. Although the proof of the theorem follows from relatively standard techniques, to the best of our knowledge, we are not aware of similar inequalities in the literature. The theorem will be a key tool in analyzing the algorithm for the spectral rounding problem in Chapter 5.

Theorem 3.3.1. *Let $\{Y_t\}_t$ be a sequence of random variables, and $X_t := Y_t - Y_{t-1}$ be the difference sequence. Suppose that there exist $R, \sigma > 0, \beta_u, \beta_l \geqslant 0$ and $\gamma_1 \in (0, \frac{1}{2}), \gamma_2 > 0$ with $\gamma_1 \leqslant \gamma_2/R$ such that the following properties hold for all $t \geqslant 1$:*

1. (Bounded difference:) $|X_t| \leqslant R$ *with probability one.*

2. (Self adjusting:) $-\gamma_1 Y_{t-1} - \beta_l \leqslant \mathbb{E}[X_t \mid Y_0, ..., Y_{t-1}] \leqslant -\gamma_1 Y_{t-1} + \beta_u$.

3. (Bounded variance:) $\mathbb{E}[X_t^2 \mid Y_0, \ldots, Y_{t-1}] \leqslant \gamma_2 Y_{t-1} + \sigma$.

4. (Initial concentration:) *For any* $a \in [-\frac{1}{R}, \frac{1}{R}]$, *the initial random variable* Y_0 *satisfies* $\mathbb{E}\left[e^{aY_0}\right] \leqslant e^{a^2\sigma/\gamma_1}$.

Then, for any $\eta > 0$ *and any* $t \geqslant 0$, *it holds that*

$$\mathbb{P}\left[Y_t \geqslant \frac{\beta_u}{\gamma_1} + \eta\right] \leqslant \exp\left(-\frac{\eta^2\gamma_1/\gamma_2}{4(\sigma/\gamma_2 + \beta_u/\gamma_1) + 2\eta}\right)$$

and

$$\mathbb{P}\left[Y_t \leqslant -\frac{\beta_l}{\gamma_1} - \eta\right] \leqslant \exp\left(-\frac{\eta^2\gamma_1/\gamma_2}{4\sigma/\gamma_2 + \eta}\right).$$

Proof. The proof is by computing the moment generating function of Y_t and applying Markov's inequality, which is standard in concentration inequalities as we have seen. In the following, we write the conditional expectation as $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|Y_0, ..., Y_{t-1}]$ for simplicity.

Upper Tail: We start with the proof for the upper tail. For any $a \in [0, \gamma_1/\gamma_2]$, the conditional moment generating function of X_t with any given $Y_0, ..., Y_{t-1}$ is

$$\begin{aligned}
\mathbb{E}_t\left[e^{aX_t}\right] = \mathbb{E}_t\left[\sum_{l=0}^{\infty} \frac{a^l X_t^l}{l!}\right] &\leqslant \mathbb{E}_t\left[1 + aX_t + \frac{X_t^2}{R^2}\sum_{l=2}^{\infty}\frac{(aR)^l}{l!}\right] \\
&= 1 + a\mathbb{E}_t[X_t] + \mathbb{E}_t[X_t^2] \cdot \frac{e^{aR} - 1 - aR}{R^2} \\
&\leqslant 1 + a\mathbb{E}_t[X_t] + a^2\mathbb{E}_t[X_t^2] \\
&\leqslant 1 - a\gamma_1 Y_{t-1} + a\beta_u + a^2\gamma_2 Y_{t-1} + a^2\sigma \\
&\leqslant \exp\left(a^2\sigma + a\beta_u - (\gamma_1 - a\gamma_2)aY_{t-1}\right),
\end{aligned}$$

where the first inequality is by the bounded difference property that $|X_t| \leqslant R$ always, the second inequality is by $aR \leqslant 1$ for $a \in [0, \gamma_1/\gamma_2]$ because $\gamma_1 \leqslant \gamma_2/R$ and the inequality

$e^p \leqslant 1 + p + p^2$ for $p \leqslant 1$, the third inequality is by the self-adjusting property and the bounded variance property and $a \geqslant 0$, and the last inequality uses $1 + p \leqslant e^p$ for $p \in \mathbb{R}$. Then we can bound the moment generating function of Y_t as

$$\begin{aligned}
\mathbb{E}_{Y_0,\ldots,Y_t}\left[e^{aY_t}\right] &= \mathbb{E}_{Y_0,\ldots,Y_{t-1}}\left[e^{aY_{t-1}} \cdot \mathbb{E}_t\left[e^{aX_t}\right]\right] \\
&\leqslant \mathbb{E}_{Y_0,\ldots,Y_{t-1}}\left[\exp\left(a^2\sigma + a\beta_u + (1 - (\gamma_1 - a\gamma_2))aY_{t-1}\right)\right] \\
&\leqslant \exp\left(a^2\sigma + a\beta_u\right) \cdot \mathbb{E}_{Y_0,\ldots,Y_{t-1}}\left[\exp\left(a\left(1 - (\gamma_1 - a\gamma_2)\right)Y_{t-1}\right)\right] \\
&= \exp\left(a^2\sigma + a\beta_u\right) \cdot \mathbb{E}_{Y_0,\ldots,Y_{t-1}}\left[\exp\left(f(a) \cdot Y_{t-1}\right)\right],
\end{aligned}$$

where we define $f(a) := a(1 - (\gamma_1 - a\gamma_2))$. Note that, $\gamma_1 - a\gamma_2 \in [0, 1]$ for $a \in [0, \gamma_1/\gamma_2]$, which implies $f(a) \in [0, a]$. Define the sequence $a_{(0)} = a$ and $a_{(i)} = f(a_{(i-1)})$ for $i \geqslant 1$. Apply the same argument inductively, it follows that

$$\begin{aligned}
\mathbb{E}_{Y_0,\ldots,Y_t}\left[e^{aY_t}\right] &\leqslant \exp\left(\sum_{i=0}^{t-1}\left(a_{(i)}^2\sigma + a_{(i)}\beta_u\right)\right) \cdot \mathbb{E}_{Y_0}\left[e^{a_{(t)}Y_0}\right] \\
&\leqslant \exp\left(\sum_{i=0}^{t-1}\left(a_{(i)}^2\sigma + a_{(i)}\beta_u\right) + \frac{a_{(t)}^2\sigma}{\gamma_1}\right),
\end{aligned}$$

where the last inequality follows from the initial concentration property of Y_0 for $a_{(t)} \leqslant a \leqslant \gamma_1/\gamma_2 \leqslant 1/R$. To bound the moment generating function, we use the following claim whose proof follows from the definition of the sequence $\{a_{(i)}\}_i$.

Claim 3.3.2. *The sequence $\{a_{(i)}\}_{i\geqslant 0}$ is decreasing and dominated by the geometric sequence $\{ar^i\}_{i\geqslant 0}$ with common ratio $r := 1 - (\gamma_1 - a\gamma_2)$. The sequence $\{a_{(i)}^2\}_i$ is also decreasing and dominated by the geometric sequence $\{a^2r^{2i}\}_{i\geqslant 0}$ with common ratio r^2. Furthermore, $r^2 < r < 1$ when $a \in [0, \gamma_1/\gamma_2)$.*

Using Claim 3.3.2, when $a \in [0, \gamma_1/\gamma_2)$, we can upper bound the moment generating function by

$$
\begin{aligned}
\mathbb{E}_{Y_0,\dots,Y_t}\left[e^{aY_t}\right] &\leqslant \exp\left(\left(a^2\sigma + a\beta_u\right)\sum_{i=0}^{t-1} r^i + \frac{a^2\sigma r^t}{\gamma_1}\right) \\
&= \exp\left(\left(a^2\sigma + a\beta_u\right)\cdot\frac{1-r^t}{1-r} + \frac{a^2\sigma r^t}{\gamma_1}\right) \\
&= \exp\left(\frac{a^2\sigma + a\beta_u}{\gamma_1 - a\gamma_2}\cdot(1-r^t) + \frac{a^2\sigma r^t}{\gamma_1}\right) \\
&\leqslant \exp\left(\frac{a^2\sigma + a\beta_u}{\gamma_1 - a\gamma_2}\right),
\end{aligned}
$$

where the last inequality uses $a \in [0, \gamma_1/\gamma_2)$. By Markov inequality, for any $a \in [0, \gamma_1/\gamma_2)$ and any $\eta > 0$,

$$
\begin{aligned}
\mathbb{P}\left[Y_t \geqslant \frac{\beta_u}{\gamma_1} + \eta\right] = \mathbb{P}\left[e^{aY_t} \geqslant e^{a\left(\frac{\beta_u}{\gamma_1}+\eta\right)}\right] &\leqslant \mathbb{E}_{Y_0,\dots,Y_t}\left[e^{aY_t}\right]\cdot e^{-a\left(\frac{\beta_u}{\gamma_1}+\eta\right)} \\
&\leqslant \exp\left(\frac{a^2\sigma + a\beta_u}{\gamma_1 - a\gamma_2} - a\left(\frac{\beta_u}{\gamma_1} + \eta\right)\right) \\
&= \exp\left(\frac{a^2(\sigma + \gamma_2\beta_u/\gamma_1)}{\gamma_1 - a\gamma_2} - a\eta\right) \\
&= \exp\left(\frac{a^2(\sigma/\gamma_2 + \beta_u/\gamma_1)}{\gamma_1/\gamma_2 - a} - a\eta\right)
\end{aligned}
$$

To prove the best upper bound, we optimize over a and set

$$
a = \frac{\gamma_1}{\gamma_2}\cdot\left(1 - \sqrt{\frac{\sigma/\gamma_2 + \beta_u/\gamma_1}{(\sigma/\gamma_2 + \beta_u/\gamma_1) + \eta}}\right) = \frac{\gamma_1}{\gamma_2}\cdot\left(1 - \sqrt{\frac{\nu}{\nu+\eta}}\right),
$$

where we use $\nu := \sigma/\gamma_2 + \beta_u/\gamma_1$ as a shorthand. Notice that $a \in [0, \gamma_1/\gamma_2)$ as $\sigma, \gamma_2, \eta > 0$ and $\beta_u \geqslant 0$, so the above probability bound applies. Putting this choice of a back into the exponent on the right hand side, the exponent is

$$
\frac{a^2(\sigma/\gamma_2 + \beta_u/\gamma_1)}{\gamma_1/\gamma_2 - a} - a\eta = \frac{\gamma_1}{\gamma_2}\cdot\underbrace{\left(\frac{(1-\sqrt{\nu/(\nu+\eta)})^2\cdot\nu}{\sqrt{\nu/(\nu+\eta)}} - \left(1 - \sqrt{\frac{\nu}{\nu+\eta}}\right)\cdot\eta\right)}_{(*)}.
$$

Simplifying the second term on the right hand side,

$$\begin{aligned}
(*) &= \left(1 + \frac{\nu}{\nu+\eta} - 2\sqrt{\frac{\nu}{\nu+\eta}}\right) \cdot \sqrt{\nu(\nu+\eta)} - \left(1 - \sqrt{\frac{\nu}{\nu+\eta}}\right) \cdot \eta \\
&= \sqrt{\nu(\nu+\eta)} + \nu\sqrt{\frac{\nu}{\nu+\eta}} - 2\nu - \eta + \eta\sqrt{\frac{\nu}{\nu+\eta}} \\
&= -(2\nu+\eta) + 2\sqrt{\nu(\nu+\eta)} \\
&= -(2\nu+\eta) + \sqrt{(2\nu+\eta)^2 - \eta^2} \\
&= -(2\nu+\eta) + (2\nu+\eta)\sqrt{1 - \frac{\eta^2}{(2\nu+\eta)^2}} \\
&\leqslant -\frac{\eta^2/2}{2\nu+\eta},
\end{aligned}$$

where we used $\sqrt{1-p} \leqslant 1-p/2$ for $p \in [0,1]$ in the last inequality. Therefore, we conclude that

$$\mathbb{P}(Y_t \geqslant \eta) \leqslant \exp\left(\frac{a^2(\sigma/\gamma_2 + \beta_u/\gamma_1)}{\gamma_1/\gamma_2 - a} - a\eta\right) \leqslant \exp\left(-\frac{\eta^2\gamma_1/\gamma_2}{4(\sigma/\gamma_2 + \beta_u/\gamma_1) + 2\eta}\right),$$

which completes the proof for the upper tail.

Lower Tail: The proof for the lower tail is quite similar to that for the upper tail. The main difference is that we work with the moment generating function $\mathbb{E}[e^{-aY_t}]$, instead of $\mathbb{E}[e^{aY_t}]$. For any $a \in [0, \gamma_1/\gamma_2]$, the conditional moment generating function of $-X_t$ is

$$\begin{aligned}
\mathbb{E}_t\left[e^{-aX_t}\right] = \mathbb{E}_t\left[\sum_{l=0}^{\infty} \frac{(-a)^l X_t^l}{l!}\right] &\leqslant \mathbb{E}_t\left[1 - aX_t + \frac{X_t^2}{R^2}\sum_{l=2}^{\infty}\frac{(aR)^l}{l!}\right] \\
&= 1 - a\mathbb{E}_t[X_t] + \mathbb{E}_t[X_t^2] \cdot \frac{e^{aR} - 1 - aR}{R^2} \\
&\leqslant 1 - a\mathbb{E}_t[X_t] + a^2\mathbb{E}_t[X_t^2] \\
&\leqslant 1 + a\gamma_1 Y_{t-1} + a\beta_l + a^2\gamma_2 Y_{t-1} + a^2\sigma \\
&\leqslant \exp\left(a^2\sigma + a\beta_l + (\gamma_1 + a\gamma_2)aY_{t-1}\right),
\end{aligned}$$

where the first inequality is by the bounded difference property $|X_t| \leqslant R$ and $a \geqslant 0$, the second inequality is by $aR \leqslant 1$ for $a \in [0, \gamma_1/\gamma_2]$ because $\gamma_1 \leqslant \gamma_2/R$ and the inequality

$e^p \leqslant 1+p+p^2$ for $p \leqslant 1$, the third inequality is by the self-adjusting property and the bounded variance property and $a \geqslant 0$, and the last inequality is by $1+p \leqslant e^p$ for $p \in \mathbb{R}$. Then we can bound the moment generating function of Y_t as

$$\begin{aligned} \mathbb{E}_{Y_0,\ldots,Y_t}\left[e^{-aY_t}\right] &= \mathbb{E}_{Y_0,\ldots,Y_{t-1}}\left[e^{-aY_{t-1}} \cdot \mathbb{E}_t\left[e^{-aX_t}\right]\right] \\ &\leqslant \mathbb{E}_{Y_0,\ldots,Y_{t-1}}\left[\exp\left(a^2\sigma + a\beta_l - a(1-(\gamma_1+a\gamma_2))Y_{t-1}\right)\right] \\ &\leqslant \exp\left(a^2\sigma + a\beta_l\right) \cdot \mathbb{E}_{Y_0,\ldots,Y_{t-1}}\left[\exp\left(-a\left(1-(\gamma_1+a\gamma_2)\right)Y_{t-1}\right)\right] \\ &= \exp\left(a^2\sigma + a\beta_l\right) \cdot \mathbb{E}_{Y_0,\ldots,Y_{t-1}}\left[\exp\left(-g(a)\cdot Y_{t-1}\right)\right], \end{aligned}$$

where we define $g(a) := a(1-(\gamma_1+a\gamma_2))$. By the given condition $\gamma_1 \in (0, \frac{1}{2})$, it holds that $\gamma_1 + a\gamma_2 \in [0,1]$ for $a \in [0, \gamma_1/\gamma_2]$ which implies $g(a) \in [0,a]$.

Define the sequence $a_{(0)} = a$ and $a_{(i)} = g(a_{(i-1)})$ for $i \geqslant 1$. Apply the same argument inductively, it follows that

$$\begin{aligned} \mathbb{E}_{Y_0,\ldots,Y_t}\left[e^{-aY_t}\right] &\leqslant \exp\left(\sum_{i=0}^{t-1}\left(a_{(i)}^2\sigma + a_{(i)}\beta_l\right)\right) \cdot \mathbb{E}_{Y_0}\left[e^{-a_{(t)}Y_0}\right] \\ &\leqslant \exp\left(\sum_{i=0}^{t-1}\left(a_{(i)}^2\sigma + a_{(i)}\beta_l\right) + \frac{a_{(t)}^2\sigma}{\gamma_1}\right), \end{aligned}$$

where the last inequality follows from the initial concentration property of Y_0 for $a_{(t)} \leqslant a \leqslant \gamma_1/\gamma_2 \leqslant 1/R$. To bound the moment generating function, we use the following claim whose proof follows from the definition of the sequence $\{a_{(i)}\}_i$.

Claim 3.3.3. *The sequence $\{a_{(i)}\}_{i\geqslant 0}$ is decreasing and dominated by the geometric sequence $\{ar^i\}_{i\geqslant 0}$ with common ratio $r := 1-(\gamma_1+a\gamma_2)$. The sequence $\{a_{(i)}^2\}_i$ is also decreasing and dominated by the geometric sequence $\{a^2r^{2i}\}_{i\geqslant 0}$ with common ratio r^2. Furthermore, $r \in (0,1)$ for $a \in [0,\gamma_1/\gamma_2]$ with $\gamma_1 \in (0,\frac{1}{2})$ and $\gamma_2 > 0$.*

Using Claim 3.3.3, when $a \in [0, \gamma_1/\gamma_2]$, we can upper bound the moment generating function by

$$\begin{aligned}\mathbb{E}_{Y_0,\dots,Y_t}\left[e^{-aY_t}\right] &\leqslant \exp\left(\left(a^2\sigma + a\beta_l\right)\sum_{i=0}^{t-1} r^i + \frac{a^2\sigma r^t}{\gamma_1}\right)\\ &= \exp\left(\left(a^2\sigma + a\beta_l\right)\cdot\frac{1-r^t}{1-r} + \frac{a^2\sigma r^t}{\gamma_1}\right)\\ &= \exp\left(\frac{a^2\sigma + a\beta_l}{\gamma_1 + a\gamma_2}\cdot(1-r^t) + \frac{a^2\sigma r^t}{\gamma_1}\right)\\ &\leqslant \exp\left(\frac{a^2\sigma + a\beta_l}{\gamma_1}\cdot(1-r^t) + \frac{a^2\sigma r^t}{\gamma_1}\right)\\ &\leqslant \exp\left(\frac{a^2\sigma + a\beta_l}{\gamma_1}\right),\end{aligned}$$

where we used $a\gamma_2 \geqslant 0$ and $r \in (0,1)$ in the second last inequality.

By Markov inequality, for any $a \in [0, \gamma_1/\gamma_2]$ and any $\eta > 0$,

$$\begin{aligned}\mathbb{P}\left[Y_t \leqslant -\frac{\beta_l}{\gamma_1} - \eta\right] = \mathbb{P}\left[e^{-aY_t} \geqslant e^{a\left(\frac{\beta_l}{\gamma_1}+\eta\right)}\right] &\leqslant \mathbb{E}_{Y_0,\dots,Y_t}\left[e^{-aY_t}\right]\cdot e^{-a\left(\frac{\beta_l}{\gamma_1}+\eta\right)}\\ &\leqslant \exp\left(\frac{a^2\sigma + a\beta_l}{\gamma_1} - a\left(\frac{\beta_l}{\gamma_1}+\eta\right)\right)\\ &= \exp\left(\frac{a^2\sigma}{\gamma_1} - a\eta\right).\end{aligned}$$

When $\eta \leqslant 2\sigma/\gamma_2$, we set $a = (\eta\gamma_1)/(2\sigma) \in [0, \gamma_1/\gamma_2]$, so the above probability bound applies and gives

$$\mathbb{P}\left[Y_t \leqslant -\frac{\beta_l}{\gamma_1} - \tau\right] \leqslant \exp\left(-\frac{\eta^2\gamma_1}{4\sigma}\right) \leqslant \exp\left(-\frac{\eta^2\gamma_1/\gamma_2}{4\sigma/\gamma_2 + \eta}\right).$$

When $\eta > 2\sigma/\gamma_2$, we simply set $a = \gamma_1/\gamma_2$, and the above probability bound gives

$$\mathbb{P}\left[Y_t \leqslant -\frac{\beta_l}{\gamma_1} - \eta\right] \leqslant \exp\left(\frac{\gamma_1}{\gamma_2}\cdot\left(\frac{\sigma}{\gamma_2} - \eta\right)\right) \leqslant \exp\left(-\frac{\eta^2\gamma_1/\gamma_2}{4\sigma/\gamma_2 + \eta}\right),$$

where the last inequality holds by the assumption that $\eta > 2\sigma/\gamma_2$. This finishes the proof for the lower tail and thus the proof of Theorem 3.3.1. □

3.4 Matrix Concentration Inequalities

All the previously mentioned concentration inequalities in Section 3.1 and Section 3.2 have their counterparts in matrix settings. In this section, we briefly survey these matrix concentration inequalities, which were used to analyze the effective resistance sampling algorithm in Section 2.5. Since we will not use these matrix concentration inequalities further in this thesis, we only give a high level overview without getting into technical details, most of the contents can be found in [137, 138].

For the matrix concentration inequalities, the main setting of the studies is as follows. Given a sequence of random matrices X_i in $\mathbb{S}^d$, bound the following tail probability

$$\mathbb{P}\left[\lambda_{\max}\left(\sum\nolimits_i X_i\right) \geqslant \delta\right].$$

Similar to the scalar random variables case, for any $\theta > 0$ we can bound by Markov's inequality that

$$\mathbb{P}\left[\lambda_{\max}\left(\sum\nolimits_i X_i\right) \geqslant \delta\right] \leqslant \mathbb{P}\left[e^{\theta\lambda_{\max}(\sum_i X_i)} \geqslant e^{\theta\delta}\right] \leqslant e^{-\theta\delta} \cdot \mathbb{E}\left[e^{\theta\lambda_{\max}(\sum_i X_i)}\right].$$

Notice that

$$e^{\theta\lambda_{\max}(\sum_i X_i)} \leqslant \sum_{k=1}^{d} e^{\theta\lambda_k(\sum_i X_i)} = \operatorname{tr}\left(e^{\sum_i \theta X_i}\right),$$

for any fixed symmetric matrices X_i's. Thus, we can derive an analog of Chernoff inequality (3.1) as follows

$$\mathbb{P}\left[\lambda_{\max}\left(\sum\nolimits_i X_i\right) \geqslant \delta\right] \leqslant \inf_{\theta>0}\left\{e^{-\theta\delta} \cdot \mathbb{E}\left[\operatorname{tr}\left(e^{\sum_i \theta X_i}\right)\right]\right\}.$$

However, controlling the term $\mathbb{E}\left[\operatorname{tr}\left(e^{\sum_i \theta X_i}\right)\right]$, i.e. the trace of the matrix moment generating function, is very challenging, which requires powerful tools. To demonstrate the idea, we assume X_i's are independent in the following discussions.

Ahlswede and Winter [2] dealt this term with Golden–Thompson inequality (see, e.g., [24] for a proof)

$$\operatorname{tr}\left(e^{A+B}\right) \leqslant \operatorname{tr}\left(e^{A} \cdot e^{B}\right).$$

With Golden-Thompson inequality, the trace of the matrix moment generating function can be bounded as follows

$$\begin{aligned}\mathbb{E}\big[\operatorname{tr}\big(e^{\sum_{i=1}^{n}\theta X_i}\big)\big] \leqslant \mathbb{E}\big[\operatorname{tr}\big(e^{\sum_{i=1}^{n-1}\theta X_i}\cdot e^{\theta X_n}\big)\big] &= \operatorname{tr}\Big(\big(\mathbb{E}e^{\sum_{i=1}^{n-1}\theta X_i}\big)\cdot\big(\mathbb{E}e^{\theta X_n}\big)\Big)\\ &\leqslant \operatorname{tr}\big(\mathbb{E}e^{\sum_{i=1}^{n-1}\theta X_i}\big)\cdot\lambda_{\max}\big(\mathbb{E}e^{\theta X_n}\big),\end{aligned}$$

where the equality follows by independence of X_i's and linearity of trace and expectation. Repeatedly applying Golden-Thompson inequality, we have

$$\mathbb{E}\big[\operatorname{tr}\big(e^{\sum_i \theta X_i}\big)\big] \leqslant \operatorname{tr}(I_d)\cdot\prod\nolimits_i \lambda_{\max}\big(\mathbb{E}e^{\theta X_i}\big) \leqslant d\cdot\exp\Big(\sum\nolimits_i \lambda_{\max}\big(\log\mathbb{E}e^{\theta X_i}\big)\Big). \tag{3.5}$$

In [138], Tropp used another powerful tool Lieb's theorem [104] to deal with the trace of matrix moment generating function, and managed to improve the above bound. Lieb's theorem says, for a fixed matrix $M\in\mathbb{S}^d$, the function $X\mapsto\operatorname{tr}\big[\exp(M+\log X)\big]$ is concave on $\mathbb{S}^d_{++}$. With Lieb's theorem, we can deal with the trace of matrix moment generating function as follows. Let $\mathbb{E}_k[\cdot]$ denotes $\mathbb{E}[\cdot\mid X_1,\ldots X_{k-1}]$. It holds that

$$\begin{aligned}\mathbb{E}_n\big[\operatorname{tr}\big(e^{\sum_{i=1}^{n}\theta X_i}\big)\big] = \mathbb{E}_n\big[\operatorname{tr}\big(e^{\sum_{i=1}^{n-1}\theta X_i+\log e^{\theta X_n}}\big)\big] &\leqslant \operatorname{tr}\big(e^{\sum_{i=1}^{n-1}\theta X_i+\log\mathbb{E}_n[e^{\theta X_n}]}\big)\\ &= \operatorname{tr}\big(e^{\sum_{i=1}^{n-1}\theta X_i+\log\mathbb{E}[e^{\theta X_n}]}\big),\end{aligned}$$

where the inequality follows by Jensen's inequality Lemma 2.2.6 for the concave trace function in X_n (by Lieb's theorem), and the last equality follows by the independence of X_i's. Then, we consider X_{n-1},

$$\begin{aligned}\mathbb{E}_{n-1}\mathbb{E}_n\big[\operatorname{tr}\big(e^{\sum_{i=1}^{n}\theta X_i}\big)\big] &\leqslant \mathbb{E}_{n-1}\big[\operatorname{tr}\big(e^{\sum_{i=1}^{n-1}\theta X_i+\log\mathbb{E}[e^{\theta X_n}]}\big)\big]\\ &= \mathbb{E}_{n-1}\big[\operatorname{tr}\big(e^{\sum_{i=1}^{n-2}\theta X_i+\log\mathbb{E}[e^{\theta X_n}]+\theta X_{n-1}}\big)\big]\\ &\leqslant \operatorname{tr}\big(e^{\sum_{i=1}^{n-2}\theta X_i+\log\mathbb{E}[e^{\theta X_n}]+\log\mathbb{E}[e^{\theta X_{n-1}}]}\big)\end{aligned}$$

where the last inequality follows by the same argument as previous step. Repeat the same argument for $X_{n-2},\ldots,X_1$ one by one, it follows that

$$\mathbb{E}\big[\operatorname{tr}\big(e^{\sum_i\theta X_i}\big)\big] \leqslant \operatorname{tr}\Big(\exp\Big(\sum\nolimits_i\log\mathbb{E}[e^{\theta X_i}]\Big)\Big) \leqslant d\cdot\exp\Big(\lambda_{\max}\Big(\sum\nolimits_i\log\mathbb{E}[e^{\theta X_i}]\Big)\Big). \tag{3.6}$$

It is not hard to see the bound in (3.6) is always better than the one in (3.5). Although the two bounds give the same result in the worst case, Tropp pointed out that (3.5) can be worst than (3.6) by a factor of d in many situations [138].

Using (3.6), Tropp proved the following generalization of Chernoff bound, Hoeffding bound, and Bernstein's inequality for sequences of independent random matrices [138].

Theorem 3.4.1 (Matrix Chernoff Bound [138]). *Let $\{X_i\}_i$ be a sequence of independent random matrices in $\mathbb{S}^d_+$. Assume $\lambda_{\max}(X_i) \leqslant R$ for each i deterministically. Let $\mu_{\min} := \lambda_{\min}(\sum_i \mathbb{E}[X_i])$ and $\mu_{\max} := \lambda_{\max}(\sum_i \mathbb{E}[X_i])$. Then*

$$\mathbb{P}\left[\lambda_{\max}\left(\sum\nolimits_i X_i\right) \geqslant (1+\delta)\mu_{\max}\right] \leqslant d \cdot \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R} \qquad \text{for } \delta > 0, \qquad \text{and}$$

$$\mathbb{P}\left[\lambda_{\min}\left(\sum\nolimits_i X_i\right) \leqslant (1-\delta)\mu_{\min}\right] \leqslant d \cdot \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R} \qquad \text{for } \delta \in [0,1].$$

Theorem 3.4.2 (Matrix Hoeffding Bound [138]). *Let $\{X_i\}_i$ be a sequence of independent random matrices in $\mathbb{S}^d$. Let A_i be a sequence of deterministic matrices in $\mathbb{S}^d$. Assume $\mathbb{E}[X_i] = 0$ and $X_i^2 \preccurlyeq A_i^2$ for each i deterministically. Then, for any $\delta > 0$*

$$\mathbb{P}\left[\lambda_{\max}\left(\sum\nolimits_i X_i\right) \geqslant \delta\right] \leqslant d \cdot e^{-\frac{\delta^2}{8\sigma^2}} \quad \text{where } \sigma^2 := \left\|\sum\nolimits_i A_i^2\right\|_{\text{op}}.$$

Theorem 3.4.3 (Matrix Bernstein's Inequality [138]). *Let $\{X_i\}_i$ be a sequence of independent random matrices in $\mathbb{S}^d$. Assume $\mathbb{E}[X_i] = 0$ and $\lambda_{\max}(X_i^2) \leqslant R$ for each i deterministically. Then, for any $\delta > 0$*

$$\mathbb{P}\left[\lambda_{\max}\left(\sum\nolimits_i X_i\right) \geqslant \delta\right] \leqslant d \cdot e^{-\frac{\delta^2/2}{\sigma^2 + R\delta/3}} \quad \text{where } \sigma^2 := \left\|\sum\nolimits_i \mathbb{E}[X_i^2]\right\|_{\text{op}}.$$

When the sequence of random matrices are not independent but form a matrix martingale, Tropp proved a generalized Freedman's inequality using Lieb's theorem in [137].

A sequence of random matrices $Y_1, \ldots, Y_t, \ldots$ is a *matrix martingale* if

$$\mathbb{E}[Y_{t+1} - Y_t \mid Y_1, \ldots, Y_t] = 0 \qquad \text{and} \qquad \mathbb{E}\left[\, \|Y_t\|_{\text{op}} \right] < +\infty \qquad \text{for all } t \geqslant 1.$$

Theorem 3.4.4 (Matrix Freedman's Inequality [137]). *Let $\{Y_t\}_t$ be a matrix martingale, where each matrix $Y_t \in \mathbb{S}^d$. Let $\{X_t\}_t$ be the difference sequence, and denote $W_t := \sum_{j=1}^{t} \mathbb{E}[X_j^2 \mid Y_1, \ldots, Y_{j-1}]$ for $t \geqslant 1$. Assume $\lambda_{\max}(X_t) \leqslant R$ deterministically for some $R > 0$. Then, for any $\delta \geqslant 0$ and $\sigma^2 > 0$,*

$$\mathbb{P}\left[\exists t \geqslant 1 : \lambda_{\max}(Y_t) \geqslant \delta \text{ and } \|W_t\|_{\mathrm{op}} \leqslant \sigma^2\right] \leqslant d \cdot e^{-\frac{\delta^2/2}{\sigma^2 + R\delta/3}}.$$

Finally, we remark that the matrix Chernoff bound, matrix Bernstein's inequality, and the matrix Freedman's inequality (except matrix Hoeffding bound) can be applied in the analysis of the effective resistance sampling algorithm for spectral sparsification in Section 2.5.

Chapter 4

Regret Minimization Framework

In this chapter, we review the regret minimization framework for spectral sparsification and one-sided spectral rounding [7, 6], and derive several slightly more general statements than those in [7, 6]. Based on the regret minimization framework, we design a new randomized algorithm to construct linear-sized spectral sparsifiers, which unifies the regret minimization based algorithm in [7] and the potential function guided adaptive sampling algorithm in [101]. The randomized adaptive sampling idea used in this new variant will be repeatedly used in later chapters (e.g., Chapter 5 and Chapter 7).

The regret minimization framework is for optimization in an online setting. In each iteration t, the player chooses an action matrix A_t from the set of density matrices $\Delta^d := \{A \in \mathbb{S}^d \mid A \succcurlyeq 0, \operatorname{tr}(A) = 1\}$, which can be understood as a probability distribution over the set of unit vectors. The player then observes a feedback matrix F_t and incurs a loss of $\langle A_t, F_t\rangle$. After τ iterations, the *regret* of the player is defined as

$$R_\tau := \sum_{t=1}^{\tau}\langle A_t, F_t\rangle - \inf_{B\in\Delta^d}\sum_{t=1}^{\tau}\langle B, F_t\rangle = \sum_{t=1}^{\tau}\langle A_t, F_t\rangle - \lambda_{\min}\left(\sum_{t=1}^{\tau} F_t\right),$$

which is the difference between the loss of the player actions and the loss of the best fixed action B, that can be assumed to be a rank one matrix $vv^\top$. The objective of the player is to minimize the regret. A well-known algorithm for regret minimization is Follow-The-

Regularized-Leader (FTRL), which plays the action

$$A_t = \operatorname{argmin}_{A \in \Delta^d} \left\{ w(A) + \alpha \cdot \sum_{i=0}^{t-1} \langle A, F_i \rangle \right\} \quad \text{for all } t \geqslant 1, \tag{4.1}$$

where $w : \mathbb{R}^{d \times d} \to \mathbb{R}$ is a convex differentiable (on $\mathbb{S}^d_{++}$) regularizer and α is a parameter called the learning rate that balances the loss and the regularization. We remark that, similar to [6], our setting allows us to have a nonzero initial feedback matrix F_0 which is given before the game starts. This will give us more flexibility in some applications.

Different choices of regularization give different algorithms for regret minimization. A popular choice is the *entropy regularizer*

$$w(A) = \langle A, \log A - I \rangle.$$

Entropy regularizer gives the well-known matrix multiplicative weight update algorithm (see Remark 4.1.19).

For the purpose of spectral sparsification, however, the ultimate goal is not to control the regret, but instead to control the objective value of the best offline action matrix $\inf_{B \in \Delta^d} \sum_{t=1}^{T} \langle B, F_t \rangle$. For this task, it turns out that the $\ell_{1-\frac{1}{q}}$*-regularizer*

$$w(A) = -\frac{q}{q-1} \operatorname{tr}\left(A^{1-\frac{1}{q}}\right)$$

introduced in [7] is more effective. The $\ell_{\frac{1}{2}}$-regularizer with $q = 2$ is already sufficient to provide an optimal algorithm for spectral sparsification. The $\ell_{1-\frac{1}{q}}$-regularizer with large constant q can be used to improve the running time of the algorithm [7, 101]. As the running time is not our main concern, we will be focusing on $\ell_{\frac{1}{2}}$-regularizer throughout this thesis.

Organization

In this chapter, we will review the FTRL algorithm for regret minimization and show how to apply the framework to construct spectral sparsifiers.

We first review the mirror descent method, an equivalent description of the FTRL algorithm in Section 4.1. Then we derive a generic regret bound with general feedback matrices in Section 4.2. Finally, with the machinery from regret minimization, we present an alternative randomized sampling algorithm for spectral sparsification that unifies known algorithms in [7] and [101] in Section 4.3.

4.1 FTRL Algorithm and Mirror Descent Method

As we have mentioned, there is an equivalent description of the FTRL algorithm using the mirror descent method framework. In this section, we formally prove the equivalence. We start with introducing an important notion. Given a (strictly) convex function $w : \mathcal{D} \to \mathbb{R}$ which is differentiable on the interior of the domain $\mathrm{int}(\mathcal{D})$, the *Bregman divergence* with respect to w is defined as

$$D_w(x, y) := w(x) - w(y) - \langle \nabla w(y), x - y \rangle, \tag{4.2}$$

where $x \in \mathcal{D}$ and $y \in \mathrm{int}(\mathcal{D})$. For example, for the entropy regularizer $w(X) = \langle X, \log X - I \rangle$ for $X \succcurlyeq 0$, with the gradient given in Fact 2.2.5, the Bregman divergence can be written as

$$\begin{aligned} D_w(X, Y) &= \langle X, \log X - I \rangle - \langle Y, \log Y - I \rangle - \langle \log Y, X - Y \rangle \\ &= \mathrm{tr}(Y - X) + \langle X, \log X - \log Y \rangle \end{aligned}$$

for $X \succcurlyeq 0$ and $Y \succ 0$. For the $\ell_{\frac{1}{2}}$-regularizer $w(X) = -2\,\mathrm{tr}(X^{\frac{1}{2}})$ for $X \succcurlyeq 0$, with the gradient given in Fact 2.2.4, the Bregman divergence can be written as

$$\begin{aligned} D_w(X, Y) &= -2\,\mathrm{tr}(X^{\frac{1}{2}}) + 2\,\mathrm{tr}(Y^{\frac{1}{2}}) + \langle Y^{-\frac{1}{2}}, X - Y \rangle \\ &= \langle Y^{-\frac{1}{2}}, X \rangle + \mathrm{tr}(Y^{\frac{1}{2}}) - 2\,\mathrm{tr}(X^{\frac{1}{2}}) \end{aligned} \tag{4.3}$$

for $X \succcurlyeq 0$ and $Y \succ 0$. More properties of Bregman divergence will be discussed in Section 4.1.1.

Then, we formally describe the mirror descent method for regret minimization.

Mirror Descent Method for Regret Minimization

1. $\mathsf{A}_1 \leftarrow \operatorname{argmin}_{\mathsf{A} \in \Delta^d}\{w(\mathsf{A}) + \alpha\langle \mathsf{A}, \mathsf{F}_0\rangle\}$.

2. For $t \leftarrow 2, 3, \ldots$ do the following

$$\widetilde{\mathsf{A}}_t \leftarrow \operatorname*{argmin}_{\mathsf{A} \succcurlyeq 0}\big\{D_w(\mathsf{A}, \mathsf{A}_{t-1}) + \alpha\langle \mathsf{A}, \mathsf{F}_{t-1}\rangle\big\}, \tag{4.4}$$

$$\mathsf{A}_t \leftarrow \operatorname*{argmin}_{\mathsf{A} \in \Delta^d}\{D_w(\mathsf{A}, \widetilde{\mathsf{A}}_t)\}. \tag{4.5}$$

We remark that we set the initial action matrix to ensure that the mirror descent method is equivalent to the FTRL algorithm (see Section 4.1.4 for more details).

Informally, Bregman divergence measures the "distance" between two points with respect to the convex function w. The mirror descent method first returns a PSD matrix $\widetilde{\mathsf{A}}_t$ that has a small loss with respect to the new feedback matrix F_{t-1}, but also does not deviate too much from the previous action matrix A_{t-1}. Then, it projects $\widetilde{\mathsf{A}}_t$ back to the space of action matrices (density matrices).

This subsection is organized as follows. In Section 4.1.1, we first review some properties of Bregman divergence, which will be useful in the later analysis. Then, we discuss some desired properties of the regularizers in Section 4.1.2 and use these properties to show the mirror descent method is well-defined in Section 4.1.3. Finally, in Section 4.1.4, we formally prove the equivalence between the FTRL algorithm and mirror descent method for those regularizers with the desired properties.

4.1.1 Bregman Divergence

Since Bregman divergence is a key notion in the mirror descent method, we first review several classical properties of Bregman divergence which will be useful in further analysis.

Lemma 4.1.1. *Let $w : \mathcal{D} \to \mathbb{R}$ be a (strictly) convex function that is differentiable on* $\mathrm{int}(\mathcal{D})$*. Then, the Bregman divergence associated with w satisfies the following properties.*

- *Given $y \in \mathrm{int}(\mathcal{D})$, $x \mapsto D_w(x, y)$ is a (strictly) convex function with gradient $\nabla w(x) - \nabla w(y)$ for $x \in \mathrm{int}(\mathcal{D})$.*
- *Non-negativity: Given any $x \in \mathcal{D}$ and any $y \in \mathrm{int}(\mathcal{D})$, it holds that $D_w(x, y) \geqslant 0$.*
- *Three-point-equality: Given any $x, y \in \mathrm{int}(\mathcal{D})$ and any $z \in \mathcal{D}$, it holds that $D_w(z, x) + D_w(x, y) - D_w(z, y) = \langle \nabla w(x) - \nabla w(y), x - z\rangle$.*

Proof. The (strict) convexity of the function $D_w(\cdot, y)$ follows from the (strict) convexity of w and the fact that $D_w(x, y)$ is equal to $w(x)$ plus a linear function in x (see (4.2)). The gradient of $D_w(\cdot, y)$ follows directly from the definition.

Given $x \in \mathcal{D}$ and $y \in \mathrm{int}(\mathcal{D})$, w is differentiable at y. By the first order condition Lemma 2.2.9, $w(x) \geqslant w(y) + \langle \nabla w(y), x - y\rangle$. Rearranging the terms, the nonnegativity follows.

The three-point-equality can be checked from the definition.

$$\begin{aligned} D_w(z, x) + D_w(x, y) &= w(z) - w(x) - \langle \nabla w(x), z - x\rangle + w(x) - w(y) - \langle \nabla w(y), x - y\rangle \\ &= D_w(z, y) + \langle \nabla w(y), z - x\rangle - \langle \nabla w(x), z - x\rangle \\ &= D_w(z, y) + \langle \nabla w(x) - \nabla w(y), x - z\rangle. \end{aligned}$$ □

Given a point $x \in \mathcal{D}$ where f is differentiable at x and a closed convex set $\mathcal{C} \subseteq \mathcal{D}$, the *Bregman projection* of x onto $\mathcal{C}$ with respect to w is defined as

$$x^* = \mathrm{argmin}_{z \in \mathcal{C}} D_w(z, x). \tag{4.6}$$

We recall a simple fact about strictly convex functions.

Fact 4.1.2. *Let f be a strictly convex function on a convex domain $\mathcal{D}$. If there exists a minimizer of f over $\mathcal{D}$, then it is unique.*

Proof. Assume there exists two distinct minimizer $x \neq y \in \mathcal{D}$ with $f(x) = f(y) = \min_{z\in\mathcal{D}} f(z)$. We consider the function value of the mid-point between x and y, i.e. $\frac{x+y}{2}$. By strict convexity of f, it holds that

$$f\left(\frac{x+y}{2}\right) < \frac{f(x)+f(y)}{2} = \min_{z\in\mathcal{D}} f(z).$$

Since $\mathcal{D}$ is convex, we find a point $\frac{x+y}{2} \in \mathcal{D}$ with strictly smaller value than the minimizer over $\mathcal{D}$, contradiction. □

Therefore, when w is strictly convex ($D_w(\cdot, x)$ is also strictly convex by Lemma 4.1.1), the Bregman projection in (4.6) is uniquely defined since $\mathcal{C}$ is a closed set. Bregman projection generalizes the notion of orthogonal projection. To see this, we can take $w(x) = \|x\|_2^2$ and verify that $D_w(x, y) = \|x - y\|_2^2$.

Finally, we show that the Bregman divergence satisfies a generalized Pythagorean theorem.

Theorem 4.1.3 (Generalized Pythagorean Theorem, see, e.g., [7, 35])**.** *Suppose $w : \mathcal{D} \to \mathbb{R}$ is a strictly convex function that is differentiable on $\mathcal{D}$, and $\mathcal{C} \subseteq \mathcal{D}$ is a closed convex set. Let $x \in \mathcal{D}$ and $x^* = \operatorname{argmin}_{z\in\mathcal{C}} D_w(z, x)$ be the Bregman projection of x onto $\mathcal{C}$ with respect to w. Then, for any $y \in \mathcal{C}$, it holds that*

$$D_w(y, x) \geqslant D_w(y, x^*) + D_w(x^*, x).$$

This further implies $D_w(y, x) \geqslant D_w(y, x^*)$ due to the nonnegativity of Bregman divergence.

Proof. By the three-point-equality in Lemma 4.1.1,

$$D_w(y, x^*) + D_w(x^*, x) - D_w(y, x) = \langle \nabla w(x^*) - \nabla w(x), x^* - y\rangle.$$

Thus, it suffices to show $\langle \nabla w(x^*) - \nabla w(x), x^* - y\rangle \leqslant 0$ for any $y \in \mathcal{C}$.

Let $f(z) = D_w(z, x)$ for the given x, which is a strictly convex function by Lemma 4.1.1. Since $x^* = \operatorname{argmin}_{z\in\mathcal{C}} D_w(z, x) = \operatorname{argmin}_{z\in\mathcal{C}} f(z)$ is the minimizer of the convex function f over the convex set $\mathcal{C}$, by Lemma 2.2.22, it implies

$$\langle \nabla f(x^*), y - x^*\rangle \geqslant 0 \quad \forall y \in \mathcal{C} \qquad \Longrightarrow \qquad \langle \nabla w(x^*) - \nabla w(x), y - x^*\rangle \geqslant 0 \quad \forall y \in \mathcal{C},$$

where we used $\nabla D_w(x^*, x) = \nabla w(x^*) - \nabla w(x)$ for a fixed x by Lemma 4.1.1. □

Remark 4.1.4. *Note that, in the above theorem, we assume that w is differentiable over the whole domain. However, some regularizers are not differentiable on boundary points. Thus, the term $D_w(y, x^*)$ is not defined for this type of regularizers if x^* lays on the boundary. In Section 4.1.2, we will introduce some restrictions on the regularizers to ensure that it does not happen.*

4.1.2 Choices for Regularizers

Consider the definition of Bregman divergence $D_w(x, y)$ in (4.2), if y is on the boundary of $\mathcal{D}$ (which is not differentiable), then $D_w(x, y)$ is undefined. Thus, in the mirror descent steps (4.4) and (4.5), we need to guarantee that $A_{t-1} \succ 0$ and $\widetilde{A}_t \succ 0$. Furthermore, the domain of computing $\widetilde{A}_t$ is unbounded, thus we also need to make sure that the minimum in step (4.4) is attained. Finally, a not crucial but naturally desirable property is that the minimizers in both (4.4) and (4.5) are uniquely defined. To summarize, we would like to choose regularizer w such that

1. The minimizers of (4.4) and (4.5) stay away from the boundary, i.e. $\widetilde{A}_t, A_t \succ 0$.
2. The minimizer $\widetilde{A}_t$ of (4.4) is attained.
3. The minimizers of (4.4) and (4.5) are uniquely defined.

The first and the third point depend on the properties of the function $D_w(\cdot, y)$ for a given $y \in \operatorname{int}(\mathcal{D})$. After throwing away those terms that do not depend on the first variable x in the Bregman divergence, minimizing $D_w(\cdot, y)$ is equivalent to minimizing $w(\cdot) - \langle \nabla w(y), \cdot \rangle$. Thus, the objective functions in the two steps (4.4) and (4.5) essentially have the same form, i.e. $w(x) - \langle c, x \rangle$.

For the first point, one option is to choose $w(x)$ to be a barrier function, which blows up when x approaches the boundary. However, this requirement is too restrictive. For example, both the entropy regularizer and $\ell_{1-\frac{1}{q}}$-regularizer do not meet this requirement.

Potentially, this might also exclude many good solutions. Another option is to require that, when the solution approaching the boundary from some direction, the directional derivative of w in that direction blows up. As we will show later, this requirement can effectively prevent the minimizer from staying on the boundary.

As for the second point, the property does not solely depend on the regularizer w. It also depends on the feedback matrix F_{t-1}. Thus, we need to make further assumption on F_{t-1} to guarantee the minimizer of (4.4) is attained.

If w is a strictly convex function, we can guarantee the uniqueness in the last point easily by Fact 4.1.2.

We will formally prove the mirror descent steps (4.4) and (4.5) are well-defined in Section 4.1.3. In the remaining of this subsection, we formally introduce the desired properties of the regularizers, and analyze those properties for further use.

Desired Properties of the Regularizers and Implications

In the following discussions of this subsection, we do not restrict the domain to $\mathbb{S}^d_+$, instead we consider a general closed domain $\mathcal{D} \subseteq \mathbb{R}^n$ in a finite dimensional Euclidean space $\mathbb{R}^n$.

Definition 4.1.5 (Legendre Function)**.** *A convex function $w : \mathcal{D} \to \mathbb{R}$ is a* Legendre function *(or a* convex function of Legendre type*), if w satisfies the following conditions.*

1. *(Interior Differentiability:) w is differentiable on $\operatorname{int}(\mathcal{D})$, where $\operatorname{int}(\mathcal{D}) \neq \emptyset$.*
2. *(Interior Strict Convexity:) w is strictly convex on $\operatorname{int}(\mathcal{D})$.*
3. *(Boundary Barrier:) For any $\mathsf{x} \in \mathcal{D} \setminus \operatorname{int}(\mathcal{D})$ on the boundary and any $\mathsf{y} \in \operatorname{int}(\mathcal{D})$ in the interior, it holds that*

$$\lim_{t \downarrow 0} \langle \nabla w(\mathsf{x} + t(\mathsf{y} - \mathsf{x})), \mathsf{y} - \mathsf{x} \rangle = -\infty,$$

where $t \downarrow 0$ denotes t approaching 0 from the side larger than 0.

Remark. *Legendre functions have many nice properties, especially in terms of their convex conjugates. In this thesis, we will not get into the details of conjugate duality theory. We refer the interested readers to the text of Rockafellar (Chapter 26 in [126]) for a nice treatment of this topic.*

We first consider a useful fact about strictly convex functions.

Fact 4.1.6. *Let f be a strictly convex function on a convex domain $\mathcal{D}$. If f is differentiable at two distinct points $x \neq y \in \mathcal{D}$, then $\nabla f(x) \neq \nabla f(x)$.*

Proof. By the first order condition Lemma 2.2.9, for $x \neq y \in \mathcal{D}$, it follows that

$$f(x) > f(y) + \langle \nabla f(y), x - y \rangle \qquad \text{and} \qquad f(y) > f(x) + \langle \nabla f(x), y - x \rangle.$$

If $\nabla f(x) = \nabla f(y)$, then adding the two inequality up gives $f(x) + f(y) > f(x) + f(y)$, contradiction. □

A direct consequence of the Fact 4.1.6 is that ∇f is a one-to-one mapping from the differentiable points of f to their gradients.

Corollary 4.1.7. *Let f be a strictly convex function on a convex domain $\mathcal{D}$, and let $\mathcal{S} \subset \mathcal{D}$ be the set of differentiable points of f. Then, there exists a one-to-one mapping from $\mathcal{S}$ to $\nabla f(\mathcal{S})$, i.e. we can define an inverse map $(\nabla f)^{-1}$ such that $(\nabla f)^{-1}(\nabla f(x)) = x$ for any $x \in \mathcal{S}$, and $\nabla f((\nabla f)^{-1}(y)) = y$ for any $y \in \nabla f(\mathcal{S})$.*

Next, we show that if the regularizer w satisfies the interior differentibility and boundary barrier conditions in Definition 4.1.5, then the minimizer of the function $w(x) - \langle c, x \rangle$ (if exist) should not stay on the boundary. Note that the function $w(x) - \langle c, x \rangle$ is closely related to the objective function in (4.4) and (4.5), and we do not need strict convexity for this lemma.

Lemma 4.1.8. *Let w be a convex regularizer satisfying the interior differentibility and boundary barrier conditions in Definition 4.1.5. Let c be an arbitrary vector in $\mathbb{R}^n$, and $\mathcal{C} \subseteq \mathbb{R}^n$ be a convex set such that $\mathcal{C} \cap \operatorname{int}(\mathcal{D}) \neq \emptyset$. If the minimizer $x^* = \operatorname{argmin}_{x \in \mathcal{D} \cap \mathcal{C}} \{w(x) - \langle c, x \rangle\}$ exists, then $x^* \in \operatorname{int}(\mathcal{D}) \cap \mathcal{C}$.*

Proof. For the sake of contradiction, suppose there exists a feasible minimizer $x^* \notin \operatorname{int}(\mathcal{D}) \cap \mathcal{C}$ (thus x^* lays on the boundary $\mathcal{D} \setminus \operatorname{int}(\mathcal{D})$). Fix any $y \in \operatorname{int}(\mathcal{D}) \cap \mathcal{C}$, we consider the function $g : [0, 1] \to \mathbb{R}$ defined by

$$g(t) = w(x^* + t(y - x^*)) - \langle c, x^* + t(y - x^*)\rangle.$$

Note that $x^* + t(y - x^*) \in \operatorname{int}(\mathcal{D}) \cap \mathcal{C}$ for $t \in (0, 1]$ by convexity of $\mathcal{D}$ and $\mathcal{C}$. Since w satisfies the interior differentiability condition in Definition 4.1.5, g is differentiable on $(0, 1]$ with derivative

$$g'(t) = \langle \nabla w(x^* + t(y - x^*)), y - x^*\rangle - \langle c, y - x^*\rangle.$$

Since $\langle c, y - x^*\rangle$ does not depend on t, the boundary barrier condition in Definition 4.1.5 implies $\lim_{t \downarrow 0} g'(t) = -\infty$. By the convexity of g (induced by the convexity of w) and the first order condition Lemma 2.2.9, we have $g(0) \geqslant g(t) - tg'(t)$. Thus, for small enough $t > 0$, it holds that $g(0) > g(t)$. This implies $w(x^*) - \langle c, x^*\rangle > w(z_t) - \langle c, z_t\rangle$, where $z_t = x^* + t(y - x^*) \in \operatorname{int}(\mathcal{D}) \cap \mathcal{C}$ for some small enough $t > 0$. This contradicts to the assumption that x^* is a global minimizer on $\mathcal{D} \cap \mathcal{C}$. □

Then, we show that if w is a Legendre function (Definition 4.1.5) subject to some restrictions on c, then the minimizer of $w(x) - \langle c, x\rangle$ behaves nicely.

Lemma 4.1.9. *Suppose we are given a Legendre function w and a vector $c \in \mathbb{R}^d$. There is a unique minimizer $x^* = \operatorname{argmin}_{x \in \mathcal{D}}\{w(x) - \langle c, x\rangle\}$ if and only if $c \in \nabla w(\operatorname{int}(\mathcal{D}))$. Furthermore, the unique minimizer $x^* = (\nabla w)^{-1}(c) \in \operatorname{int}(\mathcal{D})$.*

Proof. As w is a Legendre function (Definition 4.1.5), we can apply Lemma 4.1.8 with $\mathcal{C} = \mathbb{R}^n$ and conclude that if x^* exists then it must be in $\operatorname{int}(\mathcal{D})$. Therefore, instead of the whole domain $\mathcal{D}$, we can equivalently optimize $f_0(x) = w(x) - \langle c, x\rangle$ over the interior $\operatorname{int}(\mathcal{D})$, i.e.

$$\min_{x \in \mathcal{D}} f_0(x) = \min_{x \in \operatorname{int}(\mathcal{D})} f_0(x).$$

Since w is strictly convex on $\text{int}(\mathcal{D})$, if x^* exists, then it is necessarily unique by Fact 4.1.2. Thus, it suffices to show that the minimum of $\min_{x \in \text{int}(\mathcal{D})} f_0(x)$ is attained if and only if $c \in \nabla w(\text{int}(\mathcal{D}))$.

As $\text{int}(\mathcal{D})$ is open and w is differentiable over $\text{int}(\mathcal{D})$ by the interior diferentiability condition in Definition 4.1.5, $x \in \text{int}(\mathcal{D})$ is an optimal solution if and only if $\nabla f_0(x) = \nabla w(x) - c = 0$ (by Corollary 2.2.24).

Thus, if the minimizer x^* exists, then $c = \nabla w(x^*) \in \nabla w(\text{int}(\mathcal{D}))$. On the other hand, as w is strictly convex on $\text{int}(\mathcal{D})$, we can apply Corollary 4.1.7 to show ∇w is invertible on $\text{int}(\mathcal{D})$. Therefore, if $c \in \nabla w(\text{int}(\mathcal{D}))$ then there exists a unique $x^* \in \text{int}(\mathcal{D})$ such that $\nabla w(x^*) = c$, i.e. $x^* = (\nabla w)^{-1}(c)$, which attains the the global minimum of f_0 on $\mathcal{D}$. □

Remark. *Given a convex function $w : \mathcal{D} \to \mathbb{R}$, we can define* Fenchel conjugate *of w as $w^*(c) := -\inf_{x \in \mathcal{D}}\{w(x) - \langle c, x\rangle\}$. The relationship between w and its conjugate in Lemma 4.1.9 can be generalized to general convex function settings in terms of subdifferentials. To keep the exposition simple, we do not introduce the new notions and refer interested readers to [126, 76] for more details.*

It is easy to verify whether a univariate function is Legendre or not. For a function with matrix domain, it is less trivial to verify it. However, there is a nice way to lift a univariate Legendre function to a function with matrix domain that are unitary-invariant and symmetric with respect to the eigenvalues of the input matrix (see [22] and [102] for more details). In particular, by lifting the univariate Legendre functions $x \log x - x$ and $-\frac{q}{q-1}x^{1-\frac{1}{q}}$, one can show that the entropy regularizer and the general $\ell_{1-\frac{1}{q}}$-regularizer (for $q > 1$) are both Legendre functions. We include an elementary but not so insightful proof for $\ell_{\frac{1}{2}}$-regularizer in Appendix A.3.

Lemma 4.1.10 (see, e.g., [22] and [102])**.** *The entropy regularizer $w(X) = \langle X, \log X - I\rangle$ and $\ell_{1-\frac{1}{q}}$-regularizer $w(X) = -\frac{q}{q-1}\operatorname{tr}(X^{1-\frac{1}{q}})$ (for $q > 1$) are Legendre functions.*

4.1.3 Mirror Descent Method

In this subsection, we formally prove that the two steps (4.4) and (4.5) in mirror descent method are well-defined when the regularizer w is a Legendre function.

The first step (4.4) is an easy consequence of Lemma 4.1.9.

Lemma 4.1.11. *Suppose the regularizer $w : \mathbb{S}_+^d \to \mathbb{R}$ is a Legendre function. For $t \geqslant 2$ in step (4.4), if $A_{t-1} \succ 0$ and $\nabla w(A_{t-1}) - \alpha F_{t-1} \in \nabla w(\mathbb{S}_{++}^d)$, then the minimizer in (4.4) can be uniquely determined as $\widetilde{A}_t = (\nabla w)^{-1}(\nabla w(A_{t-1}) - \alpha F_{t-1}) \succ 0$.*

Proof. By the definition of Bregman divergence (4.2),

$$\begin{aligned} D_w(A, A_{t-1}) &= w(A) - w(A_{t-1}) - \langle \nabla w(A_{t-1}), A - A_{t-1} \rangle \\ &= w(A) - \langle \nabla w(A_{t-1}), A \rangle - \big(w(A_{t-1}) - \langle \nabla w(A_{t-1}), A_{t-1} \rangle \big). \end{aligned}$$

As the second part does not depend on A, the optimization in (4.4) is equivalent to

$$\operatorname{argmin}_{A \succcurlyeq 0} \{ D_w(A, A_{t-1}) + \alpha \langle A, F_{t-1} \rangle \} = \operatorname{argmin}_{A \succcurlyeq 0} \{ w(A) - \langle \nabla w(A_{t-1}) - \alpha F_{t-1}, A \rangle \}.$$

Since $\nabla w(A_{t-1}) - \alpha F_{t-1} \in \nabla w(\mathbb{S}_{++}^d)$ and w is a Legendre function, we apply Lemma 4.1.9 with $c = \nabla w(A_{t-1}) - \alpha F_{t-1}$ and conclude that $\widetilde{A}_t = (\nabla w)^{-1}(\nabla w(A_{t-1}) - \alpha F_{t-1}) \succ 0$ is the unique minimizer for (4.4). □

Before we deal with the second step (4.5) of mirror descent method, we prove the following lemma which is also useful for the analysis of FTRL algorithm.

Lemma 4.1.12. *Suppose the regularizer $w : \mathbb{S}_+^d \to \mathbb{R}$ is a Legendre function and is continuous on the set of density matrices Δ^d. For any symmetric matrix $C \in \mathbb{S}^d$,*

$$\min_{X \in \Delta^d} \{ w(X) - \langle C, X \rangle \} \tag{4.7}$$

has a unique minimizer defined as

$$X^* = (\nabla w)^{-1}(C + c \cdot I_d) \succ 0,$$

where $c \in \mathbb{R}$ is a unique scalar such that $C + c \cdot I_d \in (\nabla w)^{-1}(\mathbb{S}_{++}^d)$ and $\operatorname{tr}(X^*) = 1$.

Proof. It is well-known that the infimum of a continuous function over a closed and bounded set is attained (see, e.g., [127]). As $w(\cdot) - \langle C, \cdot\rangle$ is continuous on Δ^d by the assumption on w and Δ^d is closed and bounded, there exists at least one $X^* \in \Delta^d$ that attains the minimum. It remains to show the uniqueness and derive the expression for the minimizer X^*.

As w is a Legendre function, we can apply Lemma 4.1.8 with $\mathcal{D} = \mathbb{S}_+^d$ and $\mathcal{C} = \{X \in \mathbb{S}_+^d \mid \operatorname{tr}(X) = 1\}$, and conclude that the minimizer X^* of (4.7) must stay in the interior of $\mathbb{S}_+^d$, i.e. $X^* \succ 0$. Thus, the optimization in (4.7) is equivalent to

$$\min_{X \in \Delta^d} \{w(X) - \langle C, X\rangle\} = \min_{X \succ 0:\ \operatorname{tr}(X)=1} \{w(X) - \langle C, X\rangle\}. \tag{4.8}$$

As w is strictly convex over on $\mathbb{S}_{++}^d$, the uniqueness of the minimizer X^* follows from Fact 4.1.2.

To derive the expression for X^*, we note that the objective function is differentiable over the open domain $\mathbb{S}_{++}^d$ and there is only one linear constraint. By Corollary 2.2.29, X^* is an optimal solution if and only if there exists a $c \in \mathbb{R}$ such that

$$\nabla w(X^*) - C - c \cdot I_d = 0 \qquad \text{and} \qquad \operatorname{tr}(X^*) = 1.$$

As ∇w is invertible on $\mathbb{S}_{++}^d$ by the strict convexity of w (Corollary 4.1.7), it holds that $X^* = (\nabla w)^{-1}(C + c \cdot I_d) \succ 0$. The uniqueness of c follows from the uniqueness of the minimizer X^*. □

In the second step of mirror descent method, given an $\widetilde{A}_t \succ 0$, we would like to project it onto the closed and bounded set Δ^d according to the "distance" defined by the Bregman divergence.

Lemma 4.1.13. *Suppose the regularizer $w : \mathbb{S}_+^d \to \mathbb{R}$ is a Legendre function and is continuous on the set of density matrices Δ^d. For $t \geqslant 2$ in step (4.5), if $\widetilde{A}_t \succ 0$, then the minimizer in (4.5) can be uniquely determined as*

$$A_t = (\nabla w)^{-1}(\nabla w(\widetilde{A}_t) + r_t \cdot I_d) \succ 0,$$

where r_t is the unique value such that $\nabla w(\widetilde{A}_t) + r_t \cdot I_d \in \nabla w(\mathbb{S}_{++}^d)$ and $\operatorname{tr}(A_t) = 1$.

Proof. Similar to the proof of Lemma 4.1.11, the optimization in (4.5) is equivalent to

$$\operatorname{argmin}_{A \in \Delta^d} \{D_w(A, \widetilde{A}_t)\} = \operatorname{argmin}_{A \in \Delta^d} \{w(A) - \langle \nabla w(\widetilde{A}_t), A \rangle\}.$$

Thus, the lemma follows by applying Lemma 4.1.12 with $C = \nabla w(\widetilde{A}_t)$. □

Remark. *We make the continuous assumption of w to guarantee that the optimal of the convex program (4.8) is attained. We may want to use strong duality of (4.8) to achieve the same goal. However, the strong duality that follows from the Slater's condition only guarantees that the dual optimal is attained. Consider a simple example $\inf_{x>0} x^{-1}$, the optimal is not attained.*

Nevertheless, we remark that the continuous assumption of w can be replaced by the assumption that w is a closed proper convex function, which is part of the definition of Legendre function in [126]. In general, based on the Bregman divergence induced by a Legendre function, the projection of an interior point onto a closed convex set (not necessarily bounded) is unique and staying in the interior (of the general domain). As Lemma 4.1.13 is already enough for our application, we decide to keep the exposition simple and not to prove the general statement and refer the readers to Bauschke and Borwein's work [22] for more details.

From the proof Lemma 4.1.13, we see that if the regularizer w is a Legendre function, then the Bregman projection of any $X \in \mathbb{S}^d_{++}$ onto Δ^d with respect to w, i.e. $X^* = \operatorname{argmin}_{A \in \Delta^d} D_w(A, X)$, is in $\mathbb{S}^d_{++}$. Thus, w is differentiable at X^*. This resolves the issue mentioned in Remark 4.1.4. Thus, we can avoid the non-differentiable issue and derive the following corollary of the generalized Pythagorean theorem (Theorem 4.1.3), which will be used later.

Corollary 4.1.14. *Suppose $w : \mathbb{S}^d_+ \to \mathbb{R}$ is a Legendre function. Let $X \in \mathbb{S}^d_{++}$ and $X^* = \operatorname{argmin}_{Z \in \Delta^d} D_w(Z, X)$ be the Bregman projection of X onto Δ^d with respect to w. Then, for any $Y \in \Delta^d$, it holds that $D_w(Y, X) \geqslant D_w(Y, X^*)$.*

Finally, we notice that if the initial action matrix A_1 is a positive definite matrix, then applying Lemma 4.1.11 and Lemma 4.1.13 repeatedly shows the mirror descent method is well-defined for each subsequent iteration $t \geqslant 2$.

4.1.4 Equivalence of FTRL Algorithm and Mirror Descent Method

Now, we are ready to formally prove the equivalence of FTRL algorithm and mirror descent method. We first derive the expression for the action matrices of the FTRL algorithm without any assumptions on the feedback matrices.

Lemma 4.1.15. *Suppose the regularizer $w : \mathbb{S}_+^d \to \mathbb{R}$ is a Legendre function and is continuous on the set of density matrix Δ^d. The FTRL algorithm returns a unique minimizer of* (4.1)

$$A_t = (\nabla w)^{-1}\Big(-\alpha \sum_{i=0}^{t-1} F_i + l_t \cdot I_d \Big) \succ 0 \qquad \textit{for all } t \geqslant 0,$$

where $l_t \in \mathbb{R}$ is the unique scalar such that $-\alpha \sum_{i=0}^{t-1} F_i + l_t \cdot I_d \in \nabla w(\mathbb{S}_{++}^d)$ and $\operatorname{tr}(A_t) = 1$.

Proof. The lemma follows by applying Lemma 4.1.12 with $C = -\alpha \sum_{i=1}^{t-1} F_i$. □

With some additional assumptions on the feedback matrices, we show the equivalence of the FTRL algorithm and mirror descent method.

Proposition 4.1.16. *Suppose the regularizer $w : \mathbb{S}_+^d \to \mathbb{R}$ is a Legendre function and is continuous on the set of density matrix Δ^d. If the feedback matrices F_t's satisfy $\nabla w(A_t) - \alpha F_t \in \nabla w(\mathbb{S}_{++}^d)$ for all $t \geqslant 1$, then both the mirror descent method and FTRL algorithm play the following action matrix*

$$A_t = (\nabla w)^{-1}\Big(-\alpha \sum_{i=0}^{t-1} F_i + l_t \cdot I_d \Big) \succ 0 \qquad \textit{for all } t \geqslant 1$$

where $l_t \in \mathbb{R}$ is the unique scalar such that $-\alpha \sum_{i=0}^{t-1} F_i + l_t \cdot I_d \in \nabla w(\mathbb{S}_{++}^d)$ and $\operatorname{tr}(A_t) = 1$.

Proof. By Lemma 4.1.15, the FTRL algorithm will play the action matrices claimed in the proposition, even without the assumption on F_t's. In the following, we prove by induction the claim that mirror descent method also plays the same sequence of action matrices.

Take $t = 1$ as the base case, the claim follows by the initial setting of the mirror descent method. Furthermore, Lemma 4.1.15 guarantees $A_1 \succ 0$. Thus, we can apply

Lemma 4.1.11 and Lemma 4.1.13 inductively. Assume the claim is true for the first $t-1$ iterations. We consider the t-th iteration. By the induction hypothesis, $\mathsf{A}_{t-1} = (\nabla w)^{-1}\big(-\alpha \sum_{i=0}^{t-2} \mathsf{F}_i + l_{t-1} \cdot \mathsf{I}_d\big) \succ 0$. Together with the assumption $\nabla w(\mathsf{A}_{t-1}) - \alpha \mathsf{F}_{t-1} \in \nabla w(\mathbb{S}^d_{++})$, we can apply Lemma 4.1.11 and find a unique minimizer of (4.4)

$$\widetilde{\mathsf{A}}_t = (\nabla w)^{-1}(\nabla w(\mathsf{A}_{t-1}) - \alpha \mathsf{F}_{t-1}) = (\nabla w)^{-1}\Bigg(-\alpha \sum_{i=0}^{t-1} \mathsf{F}_i + l_{t-1} \cdot \mathsf{I}_d\Bigg) \succ 0,$$

where the last equality follows by the induction hypothesis on A_{t-1}. Then, we apply Lemma 4.1.13 and conclude that

$$\mathsf{A}_t = (\nabla w)^{-1}(\nabla w(\widetilde{\mathsf{A}}_t) + r_t \cdot \mathsf{I}_d) = (\nabla w)^{-1}\Bigg(-\alpha \sum_{i=0}^{t-1} \mathsf{F}_i + l_{t-1} \cdot \mathsf{I}_d + r_t \cdot \mathsf{I}_d\Bigg) \succ 0,$$

where $r_t \in \mathbb{R}$ is the unique scalar such that $-\alpha \sum_{i=0}^{t-1} \mathsf{F}_i + (l_{t-1} + r_t) \cdot \mathsf{I}_d \in \nabla w(\mathbb{S}^d_{++})$ and $\operatorname{tr}(\mathsf{A}_t) = 1$. Due to the uniqueness of l_{t-1} and l_t (for the FTRL algorithm), we must have $l_t = l_{t-1} + r_t$. □

Remark. *Notice that we do not make any assumption on the initial feedback matrix F_0. This will give us more flexibility in applications (e.g., in Chapter 5). This is exactly the reason that we set the initial action matrix for mirror descent method from iteration $t = 1$ instead of $t = 0$.*

Remark. *As mentioned by Allen-Zhu, Liao, and Orrichia in [7], for the equivalence in Proposition 4.1.16, the assumption on the feedback matrices, i.e. $\nabla w(\mathsf{A}_t) - \alpha \mathsf{F}_t \in \nabla w(\mathbb{S}^d_{++})$ for all $t \geqslant 1$, can be removed if we use a one-step mirror descent update that combines (4.4) and (4.5). More specifically, using $\mathsf{A}_t = \operatorname{argmin}_{\mathsf{A} \in \Delta^d}\{D_w(\mathsf{A}, \mathsf{A}_{t-1}) + \alpha\langle \mathsf{F}_{t-1}, \mathsf{A}\rangle\}$ in the mirror descent method is exactly equivalent to the FTRL algorithm. However, we focus on the two-step description as it is crucial in the analysis of the regret bound.*

After proving the equivalence of the FTRL algorithm and the mirror descent method for general regularizers that are Legendre functions, we consider the special cases of $\ell_{\frac{1}{2}}$-regularizer and entropy regularizer.

Corollary 4.1.17. *For the $\ell_{\frac{1}{2}}$-regularizer $w(A) = -2\operatorname{tr}(A^{\frac{1}{2}})$, if the feedback matrices F_t's satisfy $A_t^{-\frac{1}{2}} + \alpha F_t \succ 0$ for all $t \geqslant 1$, then both FTRL algorithm and mirror descent method play the following action matrix*

$$A_t = \left(\alpha \sum_{i=0}^{t-1} F_i - l_t \cdot I_d\right)^{-2} \qquad \text{for all } t \geqslant 1, \tag{4.9}$$

where l_t is the unique scalar such that $A_t \succ 0$ and $\operatorname{tr}(A_t) = 1$.

Proof. For $A \succ 0$, $\nabla w(A) = -A^{-\frac{1}{2}}$ by Fact 2.2.4. Thus, $(\nabla w)^{-1}(A) = A^{-2}$ for $A \succ 0$ and $\nabla w(\mathbb{S}^d_{++}) = \{X \prec 0\}$. The condition $\nabla w(A_t) - \alpha F_t \in \nabla w(\mathbb{S}^d_{++})$ is equivalent to $A_t^{-\frac{1}{2}} + \alpha F_t \succ 0$. We have checked that $\ell_{\frac{1}{2}}$-regularizer is a Legendre function in Lemma 4.1.10, and it is obvious that $\ell_{\frac{1}{2}}$-regularizer is continuous on Δ^d. Therefore, the corollary follows from Proposition 4.1.16 directly. □

Corollary 4.1.18. *For the entropy regularizer $w(A) = \langle A, \log A - I_d\rangle$, both FTRL algorithm and mirror descent method play the following action matrix*

$$A_t = \exp\left(l_t \cdot I_d - \alpha \sum_{i=0}^{t-1} F_i\right) \qquad \text{for all } t \geqslant 1, \tag{4.10}$$

where l_t is the unique scalar such that $A_t \succ 0$ and $\operatorname{tr}(A_t) = 1$.

Proof. For $A \succ 0$, $\nabla w(A) = \log A$ by Fact 2.2.5. Thus, $(\nabla w)^{-1}(A) = e^A$ for $A \succ 0$ and $\nabla w(\mathbb{S}^d_{++}) = \mathbb{S}^d$. Therefore, we do not need to impose any restriction on the feedback matrices F_t's. Since entropy regularizer is a Legendre function (Lemma 4.1.10) and the entropy regularizer is continuous on Δ^d, the corollary follows from Proposition 4.1.16 directly. □

Remark 4.1.19. *Note that the scalar l_t in (4.10) is used to normalize the matrix to a density matrix, thus A_t can be rewritten as*

$$A_t = \frac{\exp\left(-\alpha \sum_{i=0}^{t-1} F_i\right)}{\operatorname{tr}\left(\exp\left(-\alpha \sum_{i=0}^{t-1} F_i\right)\right)}.$$

This is exactly the action matrix used by the matrix multiplicative update method (see, e.g., Arora, Hazan, and Kale's survey [11] for more details).

4.2 A Generic Regret Bound with General Feedback Matrices

In this section, we first derive a regret bound (Lemma 4.2.1) with respect to a general regularizer that is a Legendre function. Then, we start to focus on the $\ell_{\frac{1}{2}}$-regularizer and derive a generic regret bound with general feedback matrices (Theorem 4.2.6). This generic bound will be used later in Section 6.3. With this generic bound, we derive some corollaries with feedback matrices of specific forms (e.g., rank-one matrices and rank-two matrices). These corollaries will be used in Section 4.3 in this chapter and in Chapter 5.

We start with a regret bound for general regularizers.

Lemma 4.2.1. *Suppose we are given a regularizer $w : \mathbb{S}^d_+ \to \mathbb{R}$ which is a Legendre function. Let $F_0, \dots, F_\tau$ be the feedback matrices. We run the mirror descent method on these feedback matrices. Let $\widetilde{A}_2, \dots, \widetilde{A}_{\tau+1}$ be the intermediate matrices defined in (4.4), and $A_1, \dots, A_\tau$ be the action matrices defined in (4.5). We further assume the feedback matrices and action matrices satisfy $\nabla w(A_t) - \alpha F_t \in \nabla w(\mathbb{S}^d_{++})$ for all $t \geqslant 1$. Then, for any $U \in \Delta^d$, the regret with respect to U can be bounded by*

$$R_\tau(U) = \sum_{t=1}^{\tau} \langle F_t, A_t - U \rangle \leqslant \frac{1}{\alpha} \cdot \left(D_w(U, A_1) + \sum_{t=1}^{\tau} D_w(A_t, \widetilde{A}_{t+1}) \right).$$

Proof. For any given $t \geqslant 1$, $\nabla w(A_t) - \alpha F_t \in \nabla w(\mathbb{S}^d_{++})$ by the assumption. Thus, $\widetilde{A}_{t+1} = (\nabla w)^{-1}(\nabla w(A_t) - \alpha F_t) \succ 0$ by Lemma 4.1.11, which implies

$$\nabla w(A_t) - \nabla w(\widetilde{A}_{t+1}) = \alpha F_t.$$

Therefore, the regret (rescaled by a factor α) at iteration $t \geqslant 1$ is given by

$$\begin{aligned} \langle \alpha F_t, A_t - U \rangle &= \langle \nabla w(A_t) - \nabla w(\widetilde{A}_{t+1}), A_t - U \rangle \\ &= D_w(U, A_t) + D_w(A_t, \widetilde{A}_{t+1}) - D_w(U, \widetilde{A}_{t+1}) \\ &\leqslant D_w(U, A_t) - D_w(U, A_{t+1}) + D_w(A_t, \widetilde{A}_{t+1}), \end{aligned}$$

where the second inequality follows from the three-point-equality in Lemma 4.1.1, and the last inequality follows by the corollary of the generalized Pythagorean theorem in Corollary 4.1.14.

With a telescoping sum over all $t \geqslant 1$, it holds that

$$\begin{aligned}\sum_{t=1}^{\tau} \langle \alpha F_t, A_t - U \rangle &\leqslant D_w(U, A_1) - D_w(U, A_{\tau+1}) + \sum_{t=1}^{\tau} D_w(A_t, \widetilde{A}_{t+1}) \\ &\leqslant D_w(U, A_1) + \sum_{t=1}^{\tau} D_w(A_t, \widetilde{A}_{t+1}),\end{aligned}$$

where the last inequality follows as $D_w(U, A_{\tau+1}) \geqslant 0$ by the non-negativity of Bregman divergence in Lemma 4.1.1. □

Remark. *We make some remark about the assumption $\nabla w(A_t) - \alpha F_t \in \nabla w(\mathbb{S}^d_{++})$. As we mentioned before, without this assumption, the FTRL algorithm is still well-defined.*

However, the following example shows that this assumption is important anyway in order to get a good regret bound. Consider the case where $F_0 = 0$, then $A_1 = \frac{1}{d} I_d$ for both entropy and $\ell_{\frac{1}{2}}$-regularizers. When $\tau = 1$ and U is the rank-one projection onto the top eigenspace of F_1 and $F_1 \prec 0$, then

$$\langle F_1, A_1 - U \rangle = \langle F_1, A_1 \rangle + \|F_1\|_{\mathrm{op}} = \frac{1}{d} \operatorname{tr}(F_1) + \|F_1\|_{\mathrm{op}} .$$

The error term $\|F_1\|_{\mathrm{op}}$ is much larger than the loss $\langle F_1, A_1 \rangle$ when F_1 is rank-one.

Nevertheless, it is still interesting to see whether we can bypass the two-step mirror descent analysis and directly analyze the FTRL algorithm.

In the remaining of this section, we specialize to the $\ell_{\frac{1}{2}}$-regularizer.

Proposition 4.2.2. *Let $F_0, \ldots, F_\tau$ be the feedback matrices. We run the mirror descent method on these feedback matrices. Let $\widetilde{A}_2, \ldots, \widetilde{A}_{\tau+1}$ be the intermediate matrices defined in (4.4), and $A_1, \ldots, A_\tau$ be the action matrices defined in (4.5). We further assume the* *feedback matrices and action matrices satisfy* $A_t^{-\frac{1}{2}} + \alpha F_t \succ 0$ *for all* $t \geqslant 1$*. Then, for any* $U \in \Delta^d$*, the regret with respect to* U *can be bounded by*

$$R_\tau(U) = \sum_{t=1}^{\tau} \langle F_t, A_t - U \rangle \leqslant \sum_{t=1}^{\tau} \langle F_t, A_t \rangle + \frac{1}{\alpha} \sum_{t=1}^{\tau} \left(\operatorname{tr}\left(\widetilde{A}_{t+1}^{\frac{1}{2}} \right) - \operatorname{tr}\left(A_t^{\frac{1}{2}} \right) \right) + \frac{D_w(U, A_1)}{\alpha}.$$

The regret bound further implies

$$\lambda_{\min}\left(\sum_{t=0}^{\tau} F_t \right) \geqslant -\frac{1}{\alpha} \sum_{t=1}^{\tau} \left(\operatorname{tr}\left(\widetilde{A}_{t+1}^{\frac{1}{2}} \right) - \operatorname{tr}\left(A_t^{\frac{1}{2}} \right) \right) - \frac{2\sqrt{d}}{\alpha} + \lambda_{\min}(F_0). \tag{4.11}$$

Proof. Let $w(A) = -2\operatorname{tr}(A^{\frac{1}{2}})$ be the $\ell_{\frac{1}{2}}$-regularizer. Then, $\nabla w(A) = -A^{-\frac{1}{2}}$ by Fact 2.2.4, and $\nabla w(\mathbb{S}_{++}^d) = \{X \prec 0\}$. Thus, $A_t^{-\frac{1}{2}} + \alpha F_t \succ 0$ is equivalent to $\nabla w(A_t) - \alpha F_t \in \nabla w(\mathbb{S}_{++}^d)$. We can apply Lemma 4.2.1 and show that for any $U \in \Delta^d$,

$$\sum_{t=1}^{\tau} \langle F_t, A_t - U \rangle \leqslant \frac{1}{\alpha} \cdot \left(D_w(U, A_1) + \sum_{t=1}^{\tau} D_w(A_t, \widetilde{A}_{t+1}) \right).$$

Then, we consider each Bregman divergence term in the summation

$$\begin{aligned} D_w(A_t, \widetilde{A}_{t+1}) &= \langle \widetilde{A}_{t+1}^{-\frac{1}{2}}, A_t \rangle + \operatorname{tr}\left(\widetilde{A}_{t+1}^{\frac{1}{2}} \right) - 2\operatorname{tr}\left(A_t^{\frac{1}{2}} \right) \\ &= \langle A_t^{-\frac{1}{2}} + \alpha F_t, A_t \rangle + \operatorname{tr}\left(\widetilde{A}_{t+1}^{\frac{1}{2}} \right) - 2\operatorname{tr}\left(A_t^{\frac{1}{2}} \right) \\ &= \langle \alpha F_t, A_t \rangle + \operatorname{tr}\left(\widetilde{A}_{t+1}^{\frac{1}{2}} \right) - \operatorname{tr}\left(A_t^{\frac{1}{2}} \right), \end{aligned}$$

where we used (4.3) for the first equality. To prove the second last equality, we notice that

$$\widetilde{A}_{t+1} = (\nabla w)^{-1}(\nabla w(A_t) - \alpha F_t) = (-A_t^{-\frac{1}{2}} - \alpha F_t)^{-2} = (A_t^{-\frac{1}{2}} + \alpha F_t)^{-2}, \tag{4.12}$$

where the first equality follows by Lemma 4.1.11, and the second equality follows as $\nabla w(X) = -X^{-\frac{1}{2}}$ and $(\nabla w)^{-1}(X) = X^{-2}$ by Fact 2.2.4.

Therefore, with $\ell_{\frac{1}{2}}$-regularizer, the regret with respect to $U \in \Delta^d$ can be bounded by

$$\sum_{t=1}^{\tau} \langle F_t, A_t - U \rangle \leqslant \sum_{t=1}^{\tau} \langle F_t, A_t \rangle + \frac{1}{\alpha} \sum_{t=1}^{\tau} \left(\operatorname{tr}\left(\widetilde{A}_{t+1}^{\frac{1}{2}} \right) - \operatorname{tr}\left(A_t^{\frac{1}{2}} \right) \right) + \frac{D_w(U, A_1)}{\alpha}. \tag{4.13}$$

Then, we move on to prove the lower bound for the minimum eigenvalue of $\sum_{t=0}^{\tau} F_t$. Let U be a rank-one projection on the minimum eigenspace of $\sum_{t=0}^{\tau} F_t$. We will show that

$$D_w(U, A_1) \leqslant \langle \alpha F_0, U\rangle + 2\sqrt{d} - \alpha\lambda_{\min}(F_0). \tag{4.14}$$

Note that

$$\begin{aligned} D_w(U, A_1) &= \langle A_1^{-\frac{1}{2}}, U\rangle + \mathrm{tr}(A_1^{\frac{1}{2}}) - 2\,\mathrm{tr}(U^{\frac{1}{2}}) && \text{(by (4.3))} \\ &\leqslant \mathrm{tr}\left(A_1^{\frac{1}{2}}\right) + \langle A_1^{-\frac{1}{2}}, U\rangle && \text{(since } U \succcurlyeq 0) \\ &= \mathrm{tr}\left(A_1^{\frac{1}{2}}\right) + \langle \alpha F_0 - l_1 I, U\rangle && \text{(by (4.9) for } A_1) \\ &\leqslant \sqrt{d} + \langle \alpha F_0, U\rangle - l_1, \end{aligned}$$

where the last inequality follows by the fact that both A_1, U are density matrices and Claim 2.1.10. It remains to lower bound l_1. Since $F_0 \succcurlyeq \lambda_{\min}(F_0)\cdot I_d$, $\mathrm{tr}(A_1) = 1$ and $A_1 \succ 0$, it holds that

$$1 = \mathrm{tr}(A_1) \leqslant \frac{\mathrm{tr}(I_d)}{(\alpha\lambda_{\min}(F_0) - l_1)^2} = \frac{d}{(\alpha\lambda_{\min}(F_0) - l_1)^2} \quad \Longrightarrow \quad l_1 \geqslant \alpha\lambda_{\min}(F_0) - \sqrt{d}. \tag{4.15}$$

Thus, we established (4.14). The lemma follows from (4.13) by plugging in the upper bound (4.14) of $D_w(U, A_1)$. □

Proposition 4.2.2 shows that the term

$$\mathrm{tr}\left(\widetilde{A}_{t+1}^{\frac{1}{2}}\right) - \mathrm{tr}\left(A_t^{\frac{1}{2}}\right) \tag{4.16}$$

is crucial in controlling the minimum eigenvalue of $\sum_{t=0}^{\tau} F_t$. This term was handled with respect to feedback matrices F_t's in some special forms in the literature. For example, Allen-Zhu, Liao and Oreicchia [7] considered the cases where F_t's are rank-one matrices, positive semidefinite matrices or negative semidefinite matrices. Later, Allen-Zhu, Li, Singh and Wang [6] considered the cases where F_t's are rank-two matrices with both positive and negative eigenvalues.

In the remaining of this section, we show that the techniques in [7, 6] can be easily generalized to handle feedback matrices in general form. More specifically, we consider feedback matrices of the form

$$F_t = P_t P_t^\top - N_t N_t^\top \qquad \text{and satisfying} \qquad A_t^{-\frac{1}{2}} + \alpha F_t \succ 0,$$

where $P_t \in \mathbb{R}^{d \times d_1}$ and $N_t \in \mathbb{R}^{d \times d_2}$ for some $d_1, d_2 \geqslant 0$. Note that, this form is general enough to capture all symmetric feedback matrices.

To bound (4.16), we start with considering the matrix $\widetilde{A}^{\frac{1}{2}}_{t+1}$. Since $A_t^{-\frac{1}{2}} + \alpha F_t \succ 0$ by assumption and $A_t \succ 0$, both $A_t^{-\frac{1}{2}}$ and $A_t^{-\frac{1}{2}} + \alpha F_t$ are invertible. By Woodbury matrix identity Lemma 2.1.15,

$$\begin{aligned}
\widetilde{A}^{\frac{1}{2}}_{t+1} &= \left(A_t^{-\frac{1}{2}} + \alpha P_t P_t^\top - \alpha N_t N_t^\top\right)^{-1} = \left(A_t^{-\frac{1}{2}} + \alpha \begin{pmatrix} P_t & N_t \end{pmatrix} \begin{pmatrix} I_{d_1} & \\ & -I_{d_2} \end{pmatrix} \begin{pmatrix} P_t^\top \\ N_t^\top \end{pmatrix}\right)^{-1} \\
&= A_t^{\frac{1}{2}} - \alpha A_t^{\frac{1}{2}} \begin{pmatrix} P_t & N_t \end{pmatrix} \left(\begin{pmatrix} I_{d_1} & \\ & -I_{d_2} \end{pmatrix} + \alpha \begin{pmatrix} P_t^\top \\ N_t^\top \end{pmatrix} A_t^{\frac{1}{2}} \begin{pmatrix} P_t & N_t \end{pmatrix} \right)^{-1} \begin{pmatrix} P_t^\top \\ N_t^\top \end{pmatrix} A_t^{\frac{1}{2}}.
\end{aligned} \tag{4.17}$$

To spectrally control this matrix, we need the following technical lemma, which is a generalization of Claim 2.10 in [6].

Lemma 4.2.3. *Let* $E = \begin{pmatrix} I_{d_1} & \\ & -I_{d_2} \end{pmatrix}$, *and suppose* $X \in \mathbb{R}^{d_1 \times d_1}$, $Y \in \mathbb{R}^{d_2 \times d_2}$ *and* $Z \in \mathbb{R}^{d_1 \times d_2}$. *If* $\begin{pmatrix} X & Z \\ Z^\top & Y \end{pmatrix} \succcurlyeq 0$ *and* $2Y \prec I_{d_2}$, *then we have*

$$\left(E + \begin{pmatrix} X & Z \\ Z^\top & Y \end{pmatrix}\right)^{-1} \succcurlyeq \left(E + \begin{pmatrix} 2X & \\ & 2Y \end{pmatrix}\right)^{-1}.$$

We note that the matrix $E + \begin{pmatrix} X & Z \\ Z^\top & Y \end{pmatrix}$ may not be positive definite, otherwise the lemma is easy to prove. We defer the proof to the end of this section.

With the above technical lemma, we are ready to control the spectrum of $\widetilde{A}^{\frac{1}{2}}_{t+1}$ and $\operatorname{tr}(\widetilde{A}^{\frac{1}{2}}_{t+1}) - \operatorname{tr}(A_t^{\frac{1}{2}})$.

Lemma 4.2.4. *Suppose we are given* $P \in \mathbb{R}^{d \times d_1}$, $N \in \mathbb{R}^{d \times d_2}$, *and* $A \in \mathbb{S}^d_{++}$. *If* $2N^\top A^{-1} N \prec I_{d_2}$, *then it holds that* $A + PP^\top - NN^\top \succ 0$ *and*

$$\left(A + PP^\top - NN^\top\right)^{-1} \preccurlyeq A^{-1} - A^{-1} P (I_{d_1} + 2P^\top A P)^{-1} P^\top A^{-1} + A^{-1} N (I_{d_2} - 2N^\top A N)^{-1} N^\top A^{-1}.$$

Proof. By the assumption $2N^\top A^{-1}N \prec I_{d_2}$, apply Claim 2.1.8 with $X = A$ and $Y = N$, it follows that $A + PP^\top - NN^\top \succ 0$. We can apply Woodbury matrix identity Lemma 2.1.15 and derive

$$\begin{aligned}&\left(A + PP^\top - NN^\top\right)^{-1} - A^{-1}\\ &\qquad = -A^{-1}\begin{pmatrix} P & N\end{pmatrix}\left(\begin{pmatrix} I_{d_1} & \\ & -I_{d_2}\end{pmatrix} + \begin{pmatrix} P^\top A^{-1}P & P^\top A^{-1}N \\ N^\top A^{-1}P & N^\top A^{-1}N\end{pmatrix}\right)^{-1}\begin{pmatrix} P^\top \\ N^\top\end{pmatrix} A^{-1}. \quad (4.18)\end{aligned}$$

For $A \succ 0$, we can verify that

$$\begin{pmatrix} P^\top A^{-1}P & P^\top A^{-1}N \\ N^\top A^{-1}P & N^\top A^{-1}N\end{pmatrix} = \begin{pmatrix} P^\top \\ N^\top\end{pmatrix} A^{-1}\begin{pmatrix} P & N\end{pmatrix} \succcurlyeq 0.$$

Together with the assumption $2N^\top A^{-1}N \prec I_{d_2}$, we can apply Lemma 4.2.3 with $X = P^\top A^{-1}P$, $Y = N^\top A^{-1}N$ and $Z = P^\top A^{-1}N$ to conclude that

$$\left(\begin{pmatrix} I_{d_1} & \\ & -I_{d_2}\end{pmatrix} + \begin{pmatrix} P^\top A^{-1}P & P^\top A^{-1}N \\ N^\top A^{-1}P & N^\top A^{-1}N\end{pmatrix}\right)^{-1} \succcurlyeq \begin{pmatrix} I_{d_1} + 2P^\top A^{-1}P & \\ & 2N^\top A^{-1}N - I_{d_2}\end{pmatrix}^{-1}.$$

The lemma follows by applying the above inequality to (4.18) and rearranging the terms.

□

We mention a direct consequence of Lemma 4.2.4 which is useful in applications in Chapter 7. By specializing Lemma 4.2.4 into rank-two updates and using Fact 2.1.9, we have the following lemma (which was implicitly contained in the proof of Lemma 2.5 in [6]).

Lemma 4.2.5. *Let $A \in \mathbb{R}^{d\times d} \succ 0$ and $v, u \in \mathbb{R}^d$. If $2\langle vv^\top, A^{-1}\rangle < 1$, then it holds for any $X \succcurlyeq 0$ that*

$$\langle X, \left(A - vv^\top + uu^\top\right)^{-1}\rangle \leqslant \langle X, A^{-1}\rangle + \frac{\langle X, A^{-1}vv^\top A^{-1}\rangle}{1 - 2\langle vv^\top, A^{-1}\rangle} - \frac{\langle X, A^{-1}uu^\top A^{-1}\rangle}{1 + 2\langle uu^\top, A^{-1}\rangle}.$$

Finally, we are ready to present the main generic regret bound in this section.

Theorem 4.2.6. *Suppose the action matrix* $A_t \in \mathbb{R}^{d\times d}$ *is of the form of* (4.9) *for some* $\alpha > 0$*. Suppose the initial feedback matrix* $F_0 \in \mathbb{S}^d$ *is a symmetric matrix, and for all* $t \geqslant 1$*, each feedback matrix* F_t *is of the form* $P_tP_t^\top - N_tN_t^\top$ *for some* $P_t \in \mathbb{R}^{d\times d_1}$*,* $N_t \in \mathbb{R}^{d\times d_2}$ *(*$d_1, d_2 \geqslant 0$*) such that* $\alpha \left\| A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}} \right\|_{\mathrm{op}} < \frac{1}{2}$*. Then, for any density matrix* $U \in \Delta^d$*, the regret with respect to* U *can be bounded by*

$$R_\tau(U) \leqslant \sum_{t=1}^{\tau} \left(\frac{2\alpha\langle N_tN_t^\top, A_t\rangle \cdot \left\| A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}} \right\|_{\mathrm{op}}}{1 - 2\alpha \left\| A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}} \right\|_{\mathrm{op}}} \right.$$
$$\left. + \frac{2\alpha\langle P_tP_t^\top, A_t\rangle \cdot \left\| A_t^{\frac{1}{4}} P_t P_t^\top A_t^{\frac{1}{4}} \right\|_{\mathrm{op}}}{1 + 2\alpha \left\| A_t^{\frac{1}{4}} P_t P_t^\top A_t^{\frac{1}{4}} \right\|_{\mathrm{op}}} \right) + \frac{D_w(A_1, U)}{\alpha}.$$

The above regret bound implies that

$$\lambda_{\min}\left(\sum_{t=0}^{\tau} F_t\right) \geqslant \sum_{t=1}^{\tau}\left(\frac{\langle P_tP_t^\top, A_t\rangle}{1 + 2\alpha \left\| A_t^{\frac{1}{4}} P_t P_t^\top A_t^{\frac{1}{4}} \right\|_{\mathrm{op}}} - \frac{\langle N_tN_t^\top, A_t\rangle}{1 - 2\alpha \left\| A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}} \right\|_{\mathrm{op}}} \right) - \frac{2\sqrt{d}}{\alpha} + \lambda_{\min}(F_0).$$

Proof. Recall that (4.12) says $\widetilde{A}_{t+1} = \left(A_t^{-\frac{1}{2}} + \alpha F_t\right)^{-2} = \left(A_t^{-\frac{1}{2}} + \alpha P_tP_t^\top - \alpha N_tN_t^\top\right)^{-2}$. By Lemma 2.1.1, the matrices $A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}}$ and $N_t^\top A_t^{\frac{1}{2}} N_t$ have the same nonzero eigenvalues. Thus, the assumption $\alpha \left\| A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}} \right\|_{\mathrm{op}} < \frac{1}{2}$ implies $\alpha N_t^\top A_t^{\frac{1}{2}} N_t \prec I_d$, which further implies $A_t^{-\frac{1}{2}} - \alpha N_tN_t^\top \succ 0$ by Claim 2.1.8. Thus, $\widetilde{A}_{t+1} \succ 0$ is well-defined.

Applying Lemma 4.2.4 with $A = A_t^{-\frac{1}{2}}$, $P = \sqrt{\alpha}P_t$, $N = \sqrt{\alpha}N_t$, and then take trace on both sides, we have

$$\operatorname{tr}\left(\widetilde{A}_{t+1}^{\frac{1}{2}}\right) - \operatorname{tr}\left(A_t^{\frac{1}{2}}\right) \leqslant \alpha\langle N_t^\top A_t N_t, (I_{d_2} - 2\alpha N_t^\top A_t^{\frac{1}{2}} N_t)^{-1}\rangle - \alpha\langle P_t^\top A_t P_t, (I_{d_1} + 2\alpha P_t^\top A_t^{\frac{1}{2}} P_t)^{-1}\rangle$$

Note that

$$I_{d_2} - 2\alpha N_t^\top A_t^{\frac{1}{2}} N_t \succcurlyeq \left(1 - 2\alpha \left\| N_t^\top A_t^{\frac{1}{2}} N_t \right\|_{\mathrm{op}}\right) \cdot I_{d_2} \quad \text{and} \quad \left\| N_t^\top A_t^{\frac{1}{2}} N_t \right\|_{\mathrm{op}} = \left\| A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}} \right\|_{\mathrm{op}},$$
$$I_{d_1} + 2\alpha P_t^\top A_t^{\frac{1}{2}} P_t \preccurlyeq \left(1 + 2\alpha \left\| P_t^\top A_t^{\frac{1}{2}} P_t \right\|_{\mathrm{op}}\right) \cdot I_{d_1} \quad \text{and} \quad \left\| P_t^\top A_t^{\frac{1}{2}} P_t \right\|_{\mathrm{op}} = \left\| A_t^{\frac{1}{4}} P_t P_t^\top A_t^{\frac{1}{4}} \right\|_{\mathrm{op}}.$$

Thus,

$$\operatorname{tr}\left(\widetilde{A}_{t+1}^{\frac{1}{2}}\right) - \operatorname{tr}\left(A_t^{\frac{1}{2}}\right) \leqslant \frac{\alpha\langle N_t N_t^\top, A_t\rangle}{1 - 2\alpha\left\|A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}}\right\|_{\mathrm{op}}} - \frac{\alpha\langle P_t P_t^\top, A_t\rangle}{1 + 2\alpha\left\|A_t^{\frac{1}{4}} P_t P_t^\top A_t^{\frac{1}{4}}\right\|_{\mathrm{op}}},$$

where we used the assumption $\alpha\left\|A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}}\right\|_{\mathrm{op}} < \frac{1}{2}$. The theorem follows by plugging the above bound into Proposition 4.2.2 and rearranging the terms. □

Remark. *We remark that, if the feedback matrices are either positive or negative semidefinite, then the constant factor 2 in the denominators can be removed by applying Woodbury matrix identity directly instead of applying Lemma 4.2.4.*

Recovering Results in Special Settings

When we specialize into rank-two feedback matrices, we have the following theorem, which recovers the bound in [6], and will be used in Chapter 5 and Chapter 7.

Theorem 4.2.7. *Suppose the action matrix $A_t \in \mathbb{R}^{d\times d}$ is of the form of (4.9) for some $\alpha > 0$. Suppose the initial feedback matrix $F_0 \in \mathbb{S}^d$ is a symmetric matrix, and for all $t \geqslant 1$, each feedback matrix F_t is of the form $v_{j_t} v_{j_t}^\top - v_{i_t} v_{i_t}^\top$ for some $v_{j_t}, v_{i_t} \in \mathbb{R}^d$ such that $\alpha\langle v_{i_t} v_{i_t}^\top, A_t^{\frac{1}{2}}\rangle < \frac{1}{2}$. Then the minimum eigenvalue of $\sum_{t=0}^{\tau} F_t$ is bounded by*

$$\lambda_{\min}\left(\sum_{t=0}^{\tau} F_t\right) \geqslant \sum_{t=1}^{\tau}\left(\frac{\langle v_{j_t} v_{j_t}^\top, A_t\rangle}{1 + 2\alpha\langle v_{j_t} v_{j_t}^\top, A_t^{\frac{1}{2}}\rangle} - \frac{\langle v_{i_t} v_{i_t}^\top, A_t\rangle}{1 - 2\alpha\langle v_{i_t} v_{i_t}^\top, A_t^{\frac{1}{2}}\rangle}\right) - \frac{2\sqrt{d}}{\alpha} + \lambda_{\min}(F_0).$$

When we specialize into rank-one positive/negative semidefinite matrices, we have the following theorem, which essentially recovers the bound in [7, 5], and will be used in Section 4.3.

Theorem 4.2.8. *Suppose we are given vectors $v_{i_t} \in \mathbb{R}^d$ for $t \geqslant 1$ and a learning rate parameter $\alpha > 0$. Let the initial feedback matrix $F_0 = 0$, and the action matrix $X_t \in \mathbb{R}^{d\times d}$*

be of the form of (4.9) *with respect to feedback matrix* $F_t = v_{i_t} v_{i_t}^\top$ *for* $t \geqslant 1$. *Then the minimum eigenvalue of* $\sum_{t=1}^{\tau} v_{i_t} v_{i_t}^\top$ *is bounded by*

$$\lambda_{\min}\left(\sum_{t=1}^{\tau} v_{i_t} v_{i_t}^\top\right) \geqslant \sum_{t=1}^{\tau} \frac{\langle v_{i_t} v_{i_t}^\top, X_t\rangle}{1 + 2\alpha\langle v_{i_t} v_{i_t}^\top, X_t^{\frac{1}{2}}\rangle} - \frac{2\sqrt{d}}{\alpha}.$$

Let the action matrix $Y_t \in \mathbb{R}^{d\times d}$ *be of the form of* (4.9) *with respect to feedback matrix* $F_t = -v_{i_t} v_{i_t}^\top$ *for* $t \geqslant 1$. *Suppose the vectors* v_{i_t}*'s further satisfy* $\alpha\langle v_{i_t} v_{i_t}^\top, Y_t^{\frac{1}{2}}\rangle < \frac{1}{2}$ *for all* $t \geqslant 1$, *then the maximum eigenvalue of* $\sum_{t=1}^{\tau} v_{i_t} v_{i_t}^\top$ *is bounded by*

$$\lambda_{\max}\left(\sum_{t=1}^{\tau} v_{i_t} v_{i_t}^\top\right) \leqslant \sum_{t=1}^{\tau} \frac{\langle v_{i_t} v_{i_t}^\top, Y_t\rangle}{1 - 2\alpha\langle v_{i_t} v_{i_t}^\top, Y_t^{\frac{1}{2}}\rangle} + \frac{2\sqrt{d}}{\alpha}.$$

When Allen-Zhu, Li, Singh, and Wang applied Theorem 4.2.7 in [6] and Theorem 4.2.8 in [5], a key technical issue is to bound the term $\langle Z_t, A_t\rangle$ and $\langle Z_t, A_t^{\frac{1}{2}}\rangle$, where $Z_t := \sum_{i=0}^{t-1} F_i$ is the partial solution at time t. Since Z_t and A_t have the same eigenbasis due to the closed-form (4.9), we can control the two terms by the following lemma (Claim 2.11 in [6]), which will be used in Chapter 5 and Chapter 7. We include the proof here for completeness.

Lemma 4.2.9 (Claim 2.11 in [6])**.** *Let* $Z \succcurlyeq 0$ *be an* $d \times d$ *positive semidefinite matrix and* $A = (\alpha Z - lI)^{-2}$ *for some* $\alpha > 0$ *where* l *is the unique constant such that* $A \succ 0$ *and* $\mathrm{tr}(A) = 1$. *Then*

$$\langle Z, A\rangle \leqslant \frac{\sqrt{d}}{\alpha} + \lambda_{\min}(Z) \qquad \textit{and} \qquad \alpha\langle Z, A^{\frac{1}{2}}\rangle \leqslant d + \alpha\sqrt{d}\cdot\lambda_{\min}(Z).$$

Proof. Since A and Z have the same eigenbasis, without loss of generality, we can assume both A and Z are diagonal matrices to bound $\langle Z, A\rangle$. Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of Z, it holds that

$$\langle Z, A\rangle = \sum_{i=1}^{d} \frac{\lambda_i}{(\alpha\lambda_i - l)^2} = \frac{1}{\alpha}\sum_{i=1}^{d} \frac{\alpha\lambda_i - l}{(\alpha\lambda_i - l)^2} + \sum_{i=1}^{d} \frac{l/\alpha}{(\alpha\lambda_i - l)^2} = \frac{\mathrm{tr}(A^{\frac{1}{2}})}{\alpha} + \frac{l}{\alpha} \leqslant \frac{\sqrt{d}}{\alpha} + \lambda_{\min}(Z),$$

where the last equality holds as $\mathrm{tr}(A) = 1$ and $A \succ 0$ implies $\alpha\lambda_i - l > 0$ for all $i \in [d]$, and the last inequality holds as $\mathrm{tr}(A^{\frac{1}{2}}) \leqslant \sqrt{d}$ (Claim 2.1.10 and $\mathrm{tr}(A) = 1$) and $l < \alpha\lambda_{\min}(Z)$ since $\alpha\lambda_{\min}(Z) - l > 0$ ($A \succ 0$).

By a similar argument, for the second inequality in the lemma, it holds that

$$\alpha\langle Z, A^{\frac{1}{2}}\rangle = \sum_{i=1}^{d} \frac{\alpha\lambda_i}{\alpha\lambda_i - l} = d + \sum_{i=1}^{d} \frac{l}{\alpha\lambda_i - l} \leqslant d + l \cdot \sqrt{d} \leqslant d + \alpha\sqrt{d} \cdot \lambda_{\min}(Z). \qquad \square$$

4.2.1 Deferred Proof of the Technical Lemma

Proof of Lemma 4.2.3. We first show that both $E + \left(\begin{smallmatrix} X & Z \\ Z^\top & Y \end{smallmatrix}\right)$ and $E + \left(\begin{smallmatrix} 2X & \\ & 2Y \end{smallmatrix}\right)$ are indeed invertible. By the assumption $\left(\begin{smallmatrix} X & Z \\ Z^\top & Y \end{smallmatrix}\right) \succcurlyeq 0$, we have $X \succcurlyeq 0$, which implies $I_{d_1} + 2X \succ 0$. By the assumption $2Y \prec I_{d_2}$, we also have $-I_{d_2} + 2Y \prec 0$. This verifies that $E + \left(\begin{smallmatrix} 2X & \\ & 2Y \end{smallmatrix}\right)$ is invertible. Then, we consider the matrix $E + \left(\begin{smallmatrix} X & Z \\ Z^\top & Y \end{smallmatrix}\right)$. For the sake of contradiction, we assume there is a non-zero vector $\left(\begin{smallmatrix} v_1 \\ v_2 \end{smallmatrix}\right)$ such that

$$\left(E + \begin{pmatrix} X & Z \\ Z^\top & Y \end{pmatrix}\right) \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0 \qquad \Longrightarrow \qquad \begin{cases} (I_{d_1} + X)v_1 + Zv_2 = 0 \\ Z^\top v_1 + (Y - I_{d_2})v_2 = 0 \end{cases}.$$

It further implies that $0 = v_2^\top Z^\top v_1 + v_2^\top (Y - I_{d_2})v_2 = -v_1^\top (I_{d_1} + X)v_1 + v_2^\top (Y - I_{d_2})v_2$. This contradicts with $I_{d_1} + X \succ 0$ and $-I_{d_2} + Y \prec 0$ (which follows from our assumption $X, Y \succcurlyeq 0$ and $2Y \prec I_{d_2}$). Thus, $E + \left(\begin{smallmatrix} X & Z \\ Z^\top & Y \end{smallmatrix}\right)$ is also invertible.

We write the positive semidefinite matrix $\left(\begin{smallmatrix} X & Z \\ Z^\top & Y \end{smallmatrix}\right)$ as

$$\begin{pmatrix} X & Z \\ Z^\top & Y \end{pmatrix} = QQ^\top = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} \begin{pmatrix} Q_1^\top & Q_2^\top \end{pmatrix},$$

where $Q \in \mathbb{R}^{(d_1+d_2)\times r}$, $Q_1 \in \mathbb{R}^{d_1 \times r}$, $Q_2 \in \mathbb{R}^{d_2 \times r}$ and r is the rank of the matrix $\left(\begin{smallmatrix} X & Z \\ Z^\top & Y \end{smallmatrix}\right)$. Then, notice that we also have

$$\widetilde{Q}\widetilde{Q}^\top := \begin{pmatrix} Q_1 \\ -Q_2 \end{pmatrix} \begin{pmatrix} Q_1^\top & -Q_2^\top \end{pmatrix} = \begin{pmatrix} X & -Z \\ -Z^\top & Y \end{pmatrix} \succcurlyeq 0.$$

Now, we consider the difference between $\left(E+\begin{pmatrix} X & Z \\ Z^\top & Y \end{pmatrix}\right)^{-1}$ and $\left(E+\begin{pmatrix} 2X & \\ & 2Y \end{pmatrix}\right)^{-1}$

$$
\begin{aligned}
&\left(E+\begin{pmatrix} X & Z \\ Z^\top & Y \end{pmatrix}\right)^{-1} - \left(E+\begin{pmatrix} 2X & \\ & 2Y \end{pmatrix}\right)^{-1} \\
=&\left(E+\begin{pmatrix} 2X & \\ & 2Y \end{pmatrix} - \widetilde{Q}\widetilde{Q}^\top\right)^{-1} - \left(E+\begin{pmatrix} 2X & \\ & 2Y \end{pmatrix}\right)^{-1} \\
=&\left(E+\begin{pmatrix} 2X & \\ & 2Y \end{pmatrix}\right)^{-1}\widetilde{Q}\left(I_r - \widetilde{Q}^\top\left(E+\begin{pmatrix} 2X & \\ & 2Y \end{pmatrix}\right)^{-1}\widetilde{Q}\right)^{-1}\widetilde{Q}^\top\left(E+\begin{pmatrix} 2X & \\ & 2Y \end{pmatrix}\right)^{-1}
\end{aligned}
$$

where the last equality follows by Woodbury matrix identity Lemma 2.1.15, note that we have already verified that both $E+\begin{pmatrix} X & Z \\ Z^\top & Y \end{pmatrix}$ and $E+\begin{pmatrix} 2X & \\ & 2Y \end{pmatrix}$ are invertible.

For any given $A \in \mathbb{R}^{(d_1+d_2)\times r}$ and $B \in \mathbb{S}^r$, if $B \succcurlyeq 0$ then $ABA^\top \succcurlyeq 0$, as we can verify that $x^\top ABA^\top x = y^\top By \geqslant 0$ for any $x \in \mathbb{R}^{d_1+d_2}$ and $y = A^\top x$. Therefore, to prove the lemma, it suffices to show

$$
I_r - \widetilde{Q}^\top\left(E+\begin{pmatrix} 2X & \\ & 2Y \end{pmatrix}\right)^{-1}\widetilde{Q} \succ 0.
$$

By the condition $2Q_2Q_2^\top = 2Y \prec I_{d_2}$, it holds that

$$
\begin{aligned}
\widetilde{Q}^\top\left(E+\begin{pmatrix} 2X & \\ & 2Y \end{pmatrix}\right)^{-1}\widetilde{Q} &= \begin{pmatrix} Q_1^\top & -Q_2^\top \end{pmatrix}\begin{pmatrix} (I_{d_1}+2X)^{-1} & \\ & (2Y-I_{d_2})^{-1} \end{pmatrix}\begin{pmatrix} Q_1 \\ -Q_2 \end{pmatrix} \\
&= Q_1^\top(I_{d_1}+2Q_1Q_1^\top)^{-1}Q_1 + Q_2^\top(2Q_2Q_2^\top - I_{d_2})^{-1}Q_2 \\
&\prec Q_1^\top(I_{d_1}+2Q_1Q_1^\top)^{-1}Q_1.
\end{aligned}
$$

It remains to show that $Q_1^\top(I_{d_1}+2Q_1Q_1^\top)^{-1}Q_1 \prec I_r$. Apply Woodbury matrix identity Lemma 2.1.15,

$$
\begin{aligned}
Q_1^\top(I_{d_1}+2Q_1Q_1^\top)^{-1}Q_1 &= Q_1^\top(I_{d_1} - 2Q_1(I_r+2Q_1^\top Q_1)^{-1}Q_1^\top)Q_1 \\
&= Q_1^\top Q_1 - 2Q_1^\top Q_1(I_r+2Q_1^\top Q_1)^{-1}Q_1^\top Q_1.
\end{aligned}
$$

Let $Q_1 = U\Lambda V^\top$ be the singular value decomposition of Q_1, where $\Lambda \in \mathbb{R}^{r\times r}$ is a diagonal matrix, $U \in \mathbb{R}^{d_1\times r}$ and $V \in \mathbb{R}^{r\times r}$ are column orthonormal matrices. Thus, it holds that

$$\begin{aligned} Q_1^\top (I_{d_1} + 2Q_1Q_1^\top)^{-1}Q_1 &= V\Lambda^2V^\top - 2V\Lambda^2V^\top(I_r + 2V\Lambda^2V^\top)^{-1}V\Lambda^2V^\top \\ &= V\Lambda^2V^\top - 2V\Lambda^2(I_r + 2\Lambda^2)^{-1}\Lambda^2V^\top \\ &= V\underbrace{\left(\Lambda^2 - 2\Lambda^2(I_r + 2\Lambda^2)^{-1}\Lambda^2\right)}_{M}V^\top, \end{aligned}$$

where the second equality holds as $V \in \mathbb{R}^{r\times r}$ is a column orthonormal matrix, which implies $V^\top V = VV^\top = I_r$ and $(I_r + 2V\Lambda^2V^\top)^{-1} = V(I_r + 2\Lambda^2)^{-1}V^\top$.

Note that matrix M is a r-dimensional diagonal matrix with diagonal $M(i,i) = \frac{\Lambda(i,i)^2}{1+2\Lambda(i,i)^2}$ at the (i,i)-th entry. Since the function $\frac{x}{1+2x} < 1$ for any $x \geqslant 0$, it follows that $M \prec I_r$ and $Q_1^\top(I_{d_1} + 2Q_1Q_1^\top)^{-1}Q_1 \prec I_r$ as desired. □

4.3 Regret Minimization Based Randomized Sampling for Spectral Sparsification

In this section, we demonstrate how to apply the regret minimization framework to design a randomized sampling algorithm for spectral sparsification. This randomized approach is the main theme in this thesis.

We first recall the general version of the spectral sparsification problem.

Problem 2.5.2. *Suppose we are given vectors $v_1, \ldots, v_m \in \mathbb{R}^n$ and weights $w \in \mathbb{R}^m_+$ such that $\sum_{i=1}^m w(i) \cdot v_iv_i^\top = I_n$. For some given $\varepsilon > 0$, find a reweighting $\widetilde{w} \in \mathbb{R}^m_+$ such that*

$$(1+\varepsilon)^{-1}I_n \preccurlyeq \sum_{i=1}^m \widetilde{w}(i)\cdot v_iv_i^\top \preccurlyeq (1+\varepsilon)I_n \qquad \textit{and} \qquad |\{i\in[m] : \widetilde{w}(i) \neq 0\}| \textit{ is small.}$$

To apply the regret minimization framework on spectral sparsifiaction, the sparsifier construction algorithm will act in the role of adversary in the online game setting, and it plays against two sequences of action matrices ($\{X_t\}_t$ for controlling the minimum eigenvalue and $\{Y_t\}_t$ for controlling the maximum eigenvalue). In the t-th iteration, the algorithm picks a vector v_{i_t} from the input. With some appropriate reweighting $c_t \geqslant 0$, the algorithm uses $F_t = c_t \cdot v_{i_t} v_{i_t}^\top$ as a feedback matrix for the action matrix X_t and uses $-F_t$ as a feedback matrix for the action matrix Y_t. Both $\{X_t\}_t$ and $\{Y_t\}_t$ use the FTRL algorithm/mirror descent method with the $\ell_{\frac{1}{2}}$-regularizer. For the sequence of $\{X_t\}_t$, it follows from Theorem 4.2.8 that

$$\lambda_{\min}\left(\sum_{t=1}^{\tau} F_t\right) \geqslant \sum_{t=1}^{\tau} \frac{\langle F_t, X_t\rangle}{1 + 2\alpha\langle F_t, X_t^{\frac{1}{2}}\rangle} - \frac{2\sqrt{n}}{\alpha}.$$

For the sequence of $\{Y_t\}_t$, Theorem 4.2.8 implies

$$\lambda_{\min}\left(\sum_{t=1}^{\tau} -F_t\right) \geqslant \sum_{t=1}^{\tau} \frac{\langle -F_t, Y_t\rangle}{1 + 2\alpha\langle -F_t, Y_t^{\frac{1}{2}}\rangle} - \frac{2\sqrt{n}}{\alpha}.$$

Note that $\lambda_{\min}(-A) = -\lambda_{\max}(A)$ for $A \in \mathbb{S}^d$. Thus, the above lower bound is equivalent to

$$\lambda_{\max}\left(\sum_{t=1}^{\tau} F_t\right) \leqslant \sum_{t=1}^{\tau} \frac{\langle F_t, Y_t\rangle}{1 - 2\alpha\langle F_t, Y_t^{\frac{1}{2}}\rangle} + \frac{2\sqrt{n}}{\alpha}.$$

Therefore, we can control both the maximum/minimum eigenvalues of the final solution.

The deterministic algorithm in [7] essentially follows the above plan but deviates slightly by using different reweightings for the two sequences $\{X_t\}_t$ and $\{Y_t\}_t$. In particular, it picks a vector v_{i_t} in each iteration, and maintains two sequences of action matrices X_t and Y_t such that

$$X_t := \left(\alpha \sum_{l=1}^{t-1} \frac{v_{i_l} v_{i_l}^\top}{\langle v_{i_l} v_{i_l}^\top, X_l\rangle^{\frac{1}{2}}} - l_t I_n\right)^{-2} \quad \text{and} \quad Y_t := \left(u_t I_n - \alpha \sum_{l=1}^{t-1} \frac{v_{i_l} v_{i_l}^\top}{\langle v_{i_l} v_{i_l}^\top, Y_l\rangle^{\frac{1}{2}}}\right)^{-2}.$$

The two sequences of action matrices are connected by the restriction that the selected vector v_{i_t} in the t-th iteration should satisfy $\langle v_{i_t} v_{i_t}^\top, X_t\rangle \geqslant \langle v_{i_t} v_{i_t}^\top, Y_t\rangle$. When $\langle v_{i_t} v_{i_t}^\top, X_t\rangle$ is large, the vector v_{i_t} has large contribution to the lower eigenspace of the matrix $\sum_{l=1}^{t-1} \frac{v_{i_t} v_{i_t}^\top}{\langle v_{i_t} v_{i_t}, X_l\rangle^{\frac{1}{2}}}$. We have similar explanation for large $\langle v_{i_t} v_{i_t}^\top, Y_t\rangle$. Thus, intuitively, this restriction says we want to find a vector v_{i_t} that can move lower eigenvalues more than the higher eigenvalues.

The existence of such a vector v_{i_t} follows by an averaging argument as both X_t and Y_t are density matrices. Finally, the algorithm returns $\sum_{t=1}^{\tau} \frac{v_{i_t} v_{i_t}^\top}{\langle v_{i_t} v_{i_t}^\top, X_t \rangle^{\frac{1}{2}}}$ after at most $\tau = O(n/\varepsilon^2)$ iterations.

Connections to BSS Algorithm

We make some observations about the connections between the barrier function based BSS algorithm and the regret minimization framework based algorithm. We consider the action matrix X_t for controlling the minimum eigenvalue as an example. As the action matrix $X_t \succ 0$, l_t naturally serves as a lower bound of the minimum eigenvalue of the current solution. Then, we recall a key quantity in controlling the eigenvalues via the regret bound, i.e. (4.16),

$$\operatorname{tr}\left(\widetilde{X}_{t+1}^{\frac{1}{2}}\right) - \operatorname{tr}\left(X_t^{\frac{1}{2}}\right) = \operatorname{tr}\left(\left(\alpha F_t + \alpha \sum_{l=1}^{t-1} F_l - l_t I_n\right)^{-1}\right) - \operatorname{tr}\left(\left(\alpha \sum_{l=1}^{t-1} F_l - l_t I_n\right)^{-1}\right),$$

where F_t is the feedback matrix for the sequence of X_t. The function $\operatorname{tr}(X_t^{\frac{1}{2}})$ has very similar form as the lower barrier potential function in BSS algorithm, and $\operatorname{tr}(\widetilde{X}_{t+1}^{\frac{1}{2}})$ corresponds to the value of potential function after adding the solution update F_t but without shifting the current lower barrier. The difference in the two algorithms is that BSS maintains the two barriers and the potential values explicitly, but the regret minimization based framework adjusts the barriers implicitly and adaptively by requiring $\operatorname{tr}(X_t) = 1$.

Based on the above observation, we translate the potential function guided adaptive sampling algorithm in [101] (see Section 2.5 for a more technical discussion about the algorithm in [101]) into the regret minimization framework.

4.3.1 Unifying Regret Minimization and Potential Function Sampling

```
Regret Minimization Based Randomized Sampling Algorithm
Input: vectors $\mathsf{v}_1, \ldots, \mathsf{v}_m \in \mathbb{R}^n$ and weights $\mathsf{w} \in \mathbb{R}^m_+$ such that $\sum_{i=1}^m \mathsf{w}(i) \cdot \mathsf{v}_i \mathsf{v}_i^\top = \mathsf{I}_n$.

1. Initialization: $\mathsf{Z}_1 \leftarrow 0$, $\alpha \leftarrow \varepsilon$ and $\tau = n/\varepsilon^2$.
2. For $t \leftarrow 1$ to $\tau$ do
    (a) Compute $\mathsf{X}_t := (\alpha \mathsf{Z}_t - l_t \mathsf{I}_n)^{-2}$ and $\mathsf{Y}_t := (u_t \mathsf{I}_n - \alpha \mathsf{Z}_t)^{-2}$, where $l_t$ is the unique
        scalar such that $\mathsf{X}_t \succ 0$ and $\mathrm{tr}(\mathsf{X}_t) = 1$, and $u_t$ is the unique scalar such that
        $\mathsf{Y}_t \succ 0$ and $\mathrm{tr}(\mathsf{Y}_t) = 1$.
    (b) Set $W_t \leftarrow \mathrm{tr}\left(\mathsf{X}_t^{\frac{1}{2}}\right) + \mathrm{tr}\left(\mathsf{Y}_t^{\frac{1}{2}}\right)$.
    (c) Sample a vector $i_t = i$ with probability
            $p_i := \frac{\mathsf{w}(i)}{W_t} \cdot \left( \langle \mathsf{v}_i \mathsf{v}_i^\top, \mathsf{X}_t^{\frac{1}{2}} \rangle + \langle \mathsf{v}_i \mathsf{v}_i^\top, \mathsf{Y}_t^{\frac{1}{2}} \rangle \right).$
    (d) Set
            $\Delta_t \leftarrow \frac{\mathsf{w}(i)}{p_{i_t} \sqrt{n}} \cdot \mathsf{v}_{i_t} \mathsf{v}_{i_t}^\top.$
    (e) $\mathsf{Z}_{t+1} \leftarrow \mathsf{Z}_t + \Delta_t$ and $t \leftarrow t + 1$.
3. Return $\mathsf{Z}_\tau$ as the solution.
```

Remark 4.3.1. *We make some remark on the implementation of Step 2(a), which was handled in [7] with a binary search. The key here is to find the unique scalar l_t and u_t. As the calculations of l_t and u_t are similar, we only consider l_t here. First, we observe that $l_t < \alpha\lambda_{\min}(\mathsf{Z}_t)$ as $\mathsf{X}_t \succ 0$. Then, with a similar argument as in (4.15), we can show that $l_t \geqslant \alpha\lambda_{\min}(\mathsf{Z}_t) - \sqrt{n}$, which implies that l_t lays in an interval of length $\sqrt{n}$. Thus, we can compute l_t up to $\pm\delta$ precision within $O(\log \frac{\sqrt{n}}{\delta})$ binary search iterations, which suffices for all the applications in this thesis. Therefore, we ignore the numerical issue and assume we can find the exact value for l_t and u_t throughout the thesis.*

Notice that we have a $\sqrt{n}$ factor in the denominator of the reweighting for the solution update Δ_t. The following simple observation explains the reason of using the $\sqrt{n}$ factor.

Claim 4.3.2. *For all $t \geqslant 1$, it holds that $W_t \leqslant 2\sqrt{n}$ and $\langle \Delta_t, X_t^{\frac{1}{2}} \rangle, \langle \Delta_t, Y_t^{\frac{1}{2}} \rangle < 2$.*

Proof. Since X_t and Y_t are density matrices, $\operatorname{tr}(X_t^{\frac{1}{2}}), \operatorname{tr}(Y_t^{\frac{1}{2}}) \leqslant \sqrt{n}$ by Claim 2.1.10. Thus, $W_t \leqslant 2\sqrt{n}$. Then, we consider $\langle \Delta_t, Y_t^{\frac{1}{2}} \rangle$.

$$\langle \Delta_t, Y_t^{\frac{1}{2}} \rangle = \frac{W_t}{\sqrt{n}} \cdot \frac{\langle v_{i_t} v_{i_t}^\top, Y_t^{\frac{1}{2}} \rangle}{\langle v_{i_t} v_{i_t}^\top, X_t^{\frac{1}{2}} \rangle + \langle v_{i_t} v_{i_t}^\top, Y_t^{\frac{1}{2}} \rangle} < 2,$$

where we used $W_t \leqslant 2\sqrt{n}$ that we have just proved, and $X_t \succ 0$. Similarly, we have $\langle \Delta_t, X_t^{\frac{1}{2}} \rangle < 2$. □

From the above proof we see that the motivation of introducing the $\sqrt{n}$ factor in Δ_t is to cancel the term W_t (i.e. the potential value) so that $\langle \Delta_t, X_t^{\frac{1}{2}} \rangle$ and $\langle \Delta_t, Y_t^{\frac{1}{2}} \rangle$ can be kept small. Therefore, when $\alpha < \frac{1}{4}$, we have $2\alpha\langle \Delta_t, Y_t^{\frac{1}{2}} \rangle < 1$ for all $t \geqslant 1$. Thus, we can apply Theorem 4.2.8 with feedback matrix $F_t = \Delta_t$ for action matrix X_t and feedback matrix $F_t = -\Delta_t$ for action matrix Y_t for $t \geqslant 1$, which gives

$$\begin{aligned}
\lambda_{\min}\left(\sum_{t=1}^{\tau} \Delta_t\right) &\geqslant \sum_{t=1}^{\tau} \frac{\langle \Delta_t, X_t \rangle}{1 + 2\alpha\langle \Delta_t, X_t^{\frac{1}{2}} \rangle} - \frac{2\sqrt{n}}{\alpha} \quad \text{and} \\
\lambda_{\max}\left(\sum_{t=1}^{\tau} \Delta_t\right) &\leqslant \sum_{t=1}^{\tau} \frac{\langle \Delta_t, Y_t \rangle}{1 - 2\alpha\langle \Delta_t, Y_t^{\frac{1}{2}} \rangle} + \frac{2\sqrt{n}}{\alpha}.
\end{aligned} \tag{4.19}$$

We denote the change of the lower bound on the minimum eigenvalue and the upper bound on the maximum eigenvalue in each iteration as

$$\Gamma_t^- := \frac{\langle \Delta_t, X_t \rangle}{1 + 2\alpha\langle \Delta_t, X_t^{\frac{1}{2}} \rangle} \qquad \text{and} \qquad \Gamma_t^+ := \frac{\langle \Delta_t, Y_t \rangle}{1 - 2\alpha\langle \Delta_t, Y_t^{\frac{1}{2}} \rangle}.$$

We consider the conditional expectation of Γ_t^- and Γ_t^+. We denote $\mathbb{E}_t[\Gamma_t^+]$ as the conditional expectation of Γ_t^+ given $\Delta_1, \ldots, \Delta_{t-1}$ and X_t, and write $\mathbb{E}_t[\Gamma_t^-]$ similarly.

Lemma 4.3.3. *Let $\alpha < \frac{1}{4}$. For each $1 \leqslant t \leqslant \tau$, it holds that*

$$\frac{1}{\sqrt{n}(1+4\alpha)} \leqslant \mathbb{E}_t\left[\Gamma_t^-\right] \leqslant \frac{1}{\sqrt{n}} \quad \textit{and} \quad \frac{1}{\sqrt{n}} \leqslant \mathbb{E}_t[\Gamma_t^+] \leqslant \frac{1}{\sqrt{n}(1-4\alpha)}.$$

Proof. The expected change of the spectral lower bound is

$$\mathbb{E}_t[\Gamma_t^-] = \sum_{i=1}^m p_i \cdot \frac{\frac{w(i)}{p_i\sqrt{n}} \cdot \langle v_i v_i^\top, X_t\rangle}{1 + 2\alpha \cdot \langle \Delta_t, X_t^{\frac{1}{2}}\rangle} = \frac{1}{\sqrt{n}} \cdot \sum_{i=1}^m \frac{w(i) \cdot \langle v_i v_i^\top, X_t\rangle}{1 + 2\alpha \cdot \langle \Delta_t, X_t^{\frac{1}{2}}\rangle}$$

Note that the denominator $1 + 2\alpha\langle \Delta_t, X_t^{\frac{1}{2}}\rangle < 1 + 4\alpha$ by Claim 4.3.2. Together with $\sum_{i=1}^m w(i) \cdot v_i v_i^\top = I_n$ and $\mathrm{tr}(X_t) = 1$, it holds that

$$\mathbb{E}_t[\Gamma_t^-] \begin{cases} \geqslant \dfrac{1}{(1+4\alpha)\sqrt{n}} \cdot \displaystyle\sum_{i=1}^m w(i) \cdot \langle v_i v_i^\top, X_t\rangle = \dfrac{1}{(1+4\alpha)\sqrt{n}} \\ \leqslant \dfrac{1}{\sqrt{n}} \cdot \displaystyle\sum_{i=1}^m w(i) \cdot \langle v_i v_i^\top, X_t\rangle = \dfrac{1}{\sqrt{n}} \end{cases}$$

as desired. The bounds on $\mathbb{E}_t[\Gamma_t^+]$ follows by similar calculations.

$$\mathbb{E}_t[\Gamma_t^+] = \sum_{i=1}^m p_i \cdot \frac{\frac{w(i)}{p_i\sqrt{n}} \cdot \langle v_i v_i^\top, Y_t\rangle}{1 - 2\alpha \cdot \langle \Delta_t, Y_t^{\frac{1}{2}}\rangle} = \frac{1}{\sqrt{n}} \cdot \sum_{i=1}^m \frac{w(i) \cdot \langle v_i v_i^\top, X_t\rangle}{1 - 2\alpha \cdot \langle \Delta_t, Y_t^{\frac{1}{2}}\rangle}$$

Now, the denominator $1 - 2\alpha \cdot \langle \Delta_t, Y_t^{\frac{1}{2}}\rangle > 1 - 4\alpha > 0$ by Claim 4.3.2 and the condition $\alpha < \frac{1}{4}$. Together with $\sum_{i=1}^m w(i) \cdot v_i v_i^\top = I_n$ and $\mathrm{tr}(Y_t) = 1$, it holds that

$$\mathbb{E}_t[\Gamma_t^+] \begin{cases} \geqslant \dfrac{1}{\sqrt{n}} \cdot \displaystyle\sum_{i=1}^m w(i) \cdot \langle v_i v_i^\top, Y_t\rangle = \dfrac{1}{\sqrt{n}} \\ \leqslant \dfrac{1}{(1-4\alpha)\sqrt{n}} \cdot \displaystyle\sum_{i=1}^m w(i) \cdot \langle v_i v_i^\top, Y_t\rangle = \dfrac{1}{(1-4\alpha)\sqrt{n}}. \end{cases}$$

□

In expectation, the spectrum of the final solution lays in the range of $\frac{\sqrt{n}}{\alpha^2} \pm \Theta(\frac{\sqrt{n}}{\alpha})$ after n/α^2 iterations (according to (4.19)). This implies a $1 \pm \Theta(\alpha)$ condition number of the final solution, as desired.

With Freedman's inequality Theorem 3.2.3, we show that the behavior of the expectation is indeed typical, i.e. the sum of Γ_t^- and Γ_t^+ are highly concentrated around the expectation.

Lemma 4.3.4. *For $\alpha < \frac{1}{5}$ and any $\delta > 0$, it follows that*

$$\mathbb{P}\left[\sum_{t=1}^{\tau} \Gamma_t^+ \geqslant \sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t^+] + \delta\right] \leqslant \exp\left(-\Omega\Big(\frac{\delta^2}{\tau/\sqrt{n} + \delta}\Big)\right) \quad \textit{and}$$
$$\mathbb{P}\left[\sum_{t=1}^{\tau} \Gamma_t^- \leqslant \sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t^-] - \delta\right] \leqslant \exp\left(-\Omega\Big(\frac{\delta^2}{\tau/\sqrt{n} + \delta}\Big)\right).$$

Proof. We first deal with the sum of Γ_t^+. We define a martingale as follows.

$$X_t := \Gamma_t^+ - \mathbb{E}_t[\Gamma_t^+] \qquad \text{and} \qquad Y_t := \sum_{l=1}^{t} X_l, \qquad \forall 1 \leqslant t \leqslant \tau.$$

Then, we upper bound Γ_t^+ deterministically.

$$\begin{aligned}\Gamma_t^+ &= \frac{\langle \Delta_t, \mathsf{Y}_t\rangle}{1 - 2\alpha\langle \Delta_t, \mathsf{Y}_t^{\frac{1}{2}}\rangle} = \frac{\langle \mathsf{v}_{i_t}\mathsf{v}_{i_t}^\top, \mathsf{Y}_t\rangle}{\sqrt{n}p_{i_t}/\mathsf{w}(i_t) - 2\alpha\langle \mathsf{v}_{i_t}\mathsf{v}_{i_t}^\top, \mathsf{Y}_t^{\frac{1}{2}}\rangle} \\ &= \frac{\langle \mathsf{v}_{i_t}\mathsf{v}_{i_t}^\top, \mathsf{Y}_t\rangle}{(\frac{\sqrt{n}}{W_t} - 2\alpha)\langle \mathsf{v}_{i_t}\mathsf{v}_{i_t}^\top, \mathsf{Y}_t^{\frac{1}{2}}\rangle + \frac{\sqrt{n}}{W_t}\langle \mathsf{v}_{i_t}\mathsf{v}_{i_t}^\top, \mathsf{X}_t^{\frac{1}{2}}\rangle} \leqslant \frac{\langle \mathsf{v}_{i_t}\mathsf{v}_{i_t}^\top, \mathsf{Y}_t\rangle}{(\frac{\sqrt{n}}{W_t} - 2\alpha)\langle \mathsf{v}_{i_t}\mathsf{v}_{i_t}^\top, \mathsf{Y}_t^{\frac{1}{2}}\rangle} \leqslant \frac{2}{1-4\alpha},\end{aligned}$$

where we used $\alpha < \frac{1}{5}$, $W_t \leqslant 2\sqrt{n}$ and $\langle \mathsf{v}_{i_t}\mathsf{v}_{i_t}^\top, \mathsf{Y}_t\rangle \leqslant \langle \mathsf{v}_{i_t}\mathsf{v}_{i_t}^\top, \mathsf{Y}_t^{\frac{1}{2}}\rangle$ (since $0 \prec \mathsf{Y}_t \prec \mathsf{I}_n$) in the last inequality.

Since $\Gamma_t^+ \geqslant 0$, we can also bound the second moment of Γ_t^+ by

$$\mathbb{E}_t[(\Gamma_t^+)^2] \leqslant \frac{2}{1-4\alpha} \cdot \mathbb{E}_t[\Gamma_t^+] = O\Big(\frac{1}{\sqrt{n}}\Big),$$

where the last inequality follows by Lemma 4.3.3. Thus, we have

$$X_t \leqslant \frac{2}{1-4\alpha} = O(1) \qquad \text{and} \qquad \mathbb{E}_t[X_t^2] \leqslant \mathbb{E}_t[(\Gamma_t^+)^2] = O\Big(\frac{1}{\sqrt{n}}\Big) \quad \forall t \in [\tau].$$

Apply Freedman's inequality Theorem 3.2.3 with $R = O(1)$, $\sigma_t^2 = O(\frac{1}{\sqrt{n}})$ for all $t \in [\tau]$, and $\sigma^2 = O(\frac{\tau}{\sqrt{n}})$, we have

$$\mathbb{P}\left[Y_\tau \geqslant \delta\right] \leqslant \exp\left(-\Omega\Big(\frac{\delta^2}{\tau/\sqrt{n}+\delta}\Big)\right).$$

The first probability bound in the lemma (about the deviation of sum of Γ_t^+) follows by observing $Y_\tau \geqslant \delta$ is equivalent to $\sum_{t=1}^{\tau} \Gamma_t^+ \geqslant \sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t^+] + \delta$. The second probability bound in the lemma about the deviation of sum of Γ_t^- follows from essentially the same proof. □

Finally, we are ready to show the Regret Minimization Based Randomized Sampling Algorithm returns a $(1 + O(\varepsilon))$-spectral sparsifier with high probability.

Theorem 4.3.5. *For $\tau = n/\varepsilon^2$ and $\alpha = \varepsilon$, the Regret Minimization Based Randomized Algorithm returns a $(1 + O(\varepsilon))$ spectral sparsifier of size at most n/ε^2 with probability at least* $1 - \exp(-\Omega(\sqrt{n}))$.

Proof. By Lemma 4.3.3,

$$\sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t^+] \leqslant \frac{\tau}{(1-4\alpha)\sqrt{n}}.$$

Together with (4.19), the event

$$\lambda_{\max}(Z_\tau) \geqslant \frac{\tau}{(1-4\alpha)\sqrt{n}} + \frac{3\sqrt{n}}{\alpha} \quad \Longrightarrow \quad \sum_{t=1}^{\tau} \Gamma_t^+ + \frac{2\sqrt{n}}{\alpha} \geqslant \lambda_{\max}(Z_\tau) \geqslant \sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t^+] + \frac{3\sqrt{n}}{\alpha}.$$

Take $\tau = n/\varepsilon^2$ and $\alpha = \varepsilon$, we have

$$\begin{aligned}
\mathbb{P}\left[\lambda_{\max}(Z_\tau) \geqslant \frac{\sqrt{n}}{(1-4\varepsilon)\varepsilon^2} + \frac{3\sqrt{n}}{\varepsilon}\right] &\leqslant \mathbb{P}\left[\sum_{t=1}^{\tau} \Gamma_t^+ \geqslant \sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t^+] + \frac{\sqrt{n}}{\varepsilon}\right] \\
&\leqslant \exp\left(-\Omega\Big(\frac{n/\varepsilon^2}{\sqrt{n}/\varepsilon^2 + \sqrt{n}/\varepsilon}\Big)\right) \\
&\leqslant \exp(-\Omega(\sqrt{n})).
\end{aligned}$$

where the second last inequality follows by applying Lemma 4.3.4 with $\delta = \sqrt{n}/\varepsilon$. Similarly, we can bound the probability of having small minimum eigenvalue,

$$\mathbb{P}\left[\lambda_{\min}(Z_\tau) \leqslant \frac{\sqrt{n}}{(1+4\varepsilon)\varepsilon^2} - \frac{3\sqrt{n}}{\varepsilon}\right] \leqslant \exp(-\Omega(\sqrt{n})).$$

Therefore, with probability at least $1 - \exp(-\Omega(\sqrt{n}))$, the condition number of Z_τ is bounded by

$$\frac{\lambda_{\max}(Z_\tau)}{\lambda_{\min}(Z_\tau)} \leqslant \frac{\frac{\sqrt{n}}{(1-4\varepsilon)\varepsilon^2} + \frac{3\sqrt{n}}{\varepsilon}}{\frac{\sqrt{n}}{(1+4\varepsilon)\varepsilon^2} - \frac{3\sqrt{n}}{\varepsilon}} = 1 + O(\varepsilon). \qquad \square$$

Discussions

We make some remarks about the Regret Minimization Based Randomized Sampling Algorithm which unifies the regret minimization framework and barrier potential function guided adaptive sampling algorithm.

The algorithm in [7] uses different reweightings for the two sequences of action matrices. The choices of the reweightings are not very intuitive and the analysis of the algorithm is more involved. The reweightings in the Regret Minimization Based Randomized Sampling Algorithm were intuitively designed such that the expected update is proportional to the identity matrix. The analysis of the algorithm is also very straight forward. One disadvantage of the Regret Minimization Based Randomized Sampling Algorithm is that it does not seem easy to derandomize the algorithm to give a deterministic algorithm.

As in BSS algorithm, the potential function guided adaptive sampling algorithm in [101] maintains the barriers explicitly, and there are more parameters that need to be kept track of, e.g., the initial potential value, the step size of the solution update, the upper/lower barrier shifts, etc. In contrast, using the machinery from regret minimization, the algorithm we proposed in this section is more principled with only one parameter α to adjust. We can perform the algorithm analysis within a single framework. In particular, we do not need to worry about how to shift the upper/lower barriers in the analysis, as they are handled by the regret minimization framework (with the requirement that each action matrix is a density matrix). Furthermore, the potential value of each iteration can be easily bounded as in Claim 4.3.2, thus the analysis of the eigenvalue bounds also becomes simpler.

Finally, we remark that the most important advantage of using the random sampling idea (as in [101] and this section) is that we can find solutions that satisfy many different properties simultaneously due to the concentration properties of the algorithm. For example, the algorithm can return a sparsifier that maintains the cost of the input graph. We will demonstrate this idea with more concrete examples in Chapter 5, Chapter 6, and Chapter 7.

Chapter 5

Spectral Rounding

In this chapter, we study the spectral rounding problem, which is the central problem in this thesis. We recall the problem statement in Chapter 1.

Question 1.1.1 (Spectral Rounding)**.** *Suppose we are given vectors $v_1, \ldots, v_m \in \mathbb{R}^d$ and $x \in [0,1]^m$ such that $\sum_{i=1}^m x(i) \cdot v_i v_i^\top = I_d$, where I_d is the d-dimensional identity matrix. Given a non-negative "cost" vector $c \in \mathbb{R}^m_+$, find $z \in \{0,1\}^m$ such that*

$$\sum_{i=1}^m z(i) \cdot v_i v_i^\top \approx I_d \qquad \textit{and} \qquad \langle c, z \rangle \approx \langle c, x \rangle.$$

This problem is similar to the spectral sparsification problem introduced by Spielman and Teng [133] (see Section 2.5). In spectral sparsification, the goal is to find a sparse non-negative vector $y \in \mathbb{R}^m_+$ to approximate the spectral properties of a given fractional vector x. Spectral rounding is different in that we want to find an integral vector $z \in \{0,1\}^m$ to approximate the spectral properties of x and preserve the cost simultaneously.

To approximate the spectral properties of a fractional vector, we will consider two different settings.

- One-sided spectral rounding: Find $z \in \{0,1\}^m$ such that

$$\sum_{i=1}^m z(i) \cdot v_i v_i^\top \gtrsim I_d \qquad \text{and} \qquad \langle c, z \rangle \approx \langle c, x \rangle.$$

- Two-sided spectral rounding: Find $z \in \{0,1\}^m$ such that

$$\sum_{i=1}^{m} z(i) \cdot v_i v_i^\top \approx I_d \qquad \text{and} \qquad \langle c, z \rangle \approx \langle c, x \rangle.$$

Organization

We present the previous work about both settings of spectral rounding and state our main results in Section 5.1. Then, we present our one-sided spectral rounding algorithm in Section 5.2. We present our non-constructive two-sided spectral rounding bound in Section 5.3. Finally, we gave some tight examples about one-sided spectral rounding in Section 5.4.

5.1 Previous Work and Our Contributions

The most relevant works for spectral rounding are from spectral sparsification and discrepancy theory. There are two previous theorems that imply non-trivial results for spectral rounding.

One-Sided Spectral Rounding

Allen-Zhu, Li, Singh, and Wang [6] formulated and proved the following spectral rounding theorem, using a regret minimization framework for spectral sparsification [7] (see Chapter 4).

Theorem 5.1.1 (Theorem 2.1 in [6]). *Let $v_1, v_2, \ldots, v_m \in \mathbb{R}^d$, $x \in [0,1]^m$ and $k = \sum_{i=1}^m x(i)$. Suppose $\sum_{i=1}^m x(i) \cdot v_i v_i^\top = I_d$ and $k \geqslant 5d/\varepsilon^2$ for some $\varepsilon \in (0, \frac{1}{3}]$. Then there is a polynomial time algorithm to return a subset $S \subseteq [m]$ with*

$$|S| \leqslant k \quad \text{and} \quad \sum_{i \in S} v_i v_i^\top \succcurlyeq (1 - 3\varepsilon) \cdot I_d.$$

Theorem 5.1.1 can be understood as a one-sided spectral rounding result, where the fractional solution x is rounded to a zero-one solution while the budget constraint is satisfied and the spectral lower bound is approximately satisfied. Through a general reduction, this theorem implies near-optimal approximation algorithms for a large class of experimental design problems (see Section 7.2).

Two-Sided Spectral Rounding

The techniques in spectral sparsification have been extended greatly to prove discrepancy theorems in spectral settings [111, 9, 91]. The following recent result by Kyng, Luh, and Song [91] provides the most refined formulation in the discrepancy setting, using the method of interlacing polynomials and the barrier arguments developed in [110, 111, 9] (see Section 2.6 for a more detailed survey).

Theorem 2.6.10 (Theorem 1.4 in [91])**.** *Let $v_1, ..., v_m \in \mathbb{R}^d$, and $\xi_1, ..., \xi_m$ be independent random scalar variables with finite support. There exists a choice of outcomes $\epsilon_1, ..., \epsilon_m$ in the support of $\xi_1, ..., \xi_m$ such that*

$$\left\| \sum_{i=1}^{m} \mathbb{E}\left[\xi_i\right] \cdot v_i v_i^\top - \sum_{i=1}^{m} \epsilon_i \cdot v_i v_i^\top \right\|_{\mathrm{op}} \leqslant 4 \left\| \sum_{i=1}^{m} \mathbf{Var}[\xi_i](v_i v_i^\top)^2 \right\|_{\mathrm{op}}^{\frac{1}{2}}.$$

We note that Theorem 2.6.10 implies the following two-sided spectral rounding result, which is very similar to Corollary 1.7 in [91] but with a weaker assumption, where we only need $\left\|\sum_{i=1}^m x(i) \cdot v_i v_i^\top\right\|_{\mathrm{op}} \leqslant 1$ instead of $\left\|\sum_{i=1}^m v_i v_i^\top\right\|_{\mathrm{op}} \leqslant 1$ as in [91]. The proof will be presented in Section 5.3 in a more general setting.

Corollary 5.1.2. *Let $v_1, ..., v_m \in \mathbb{R}^d$ and $x \in [0,1]^m$. Suppose $\sum_{i=1}^m x(i) \cdot v_i v_i^\top = I_d$ and $\|v_i\|_2 \leqslant \varepsilon$ for all $i \in [m]$. Then there exists a subset $S \subseteq [m]$ satisfying*

$$(1 - O(\varepsilon)) \cdot I_d \preccurlyeq \sum_{i \in S} v_i v_i^\top \preccurlyeq (1 + O(\varepsilon)) \cdot I_d.$$

Comparing to Theorem 5.1.1, the advantage of Corollary 5.1.2 is that it provides a two-sided spectral approximation. On the other hand, Corollary 5.1.2 requires the assumption that all vectors are short, and it has no guarantee on the size of S. Also, it is important to point out that the proof of Corollary 5.1.2 does not provide a polynomial time algorithm to find such a subset.

5.1.1 Our Contributions

We extend the previous results on spectral rounding to incorporate non-negative linear constraints, which can satisfy the requirements for network design problems (see Section 6.1).

Our main result for spectral rounding considers one-sided spectral rounding.

Theorem 5.1.3. *Suppose we are given* $\mathsf{v}_1, ..., \mathsf{v}_m \in \mathbb{R}^d$ *and* $\mathsf{x} \in [0,1]^m$ *such that* $\sum_{i=1}^m \mathsf{x}(i) \cdot \mathsf{v}_i \mathsf{v}_i^\top = \mathsf{I}_d$*. For any* $\varepsilon \in (0, \frac{1}{4})$*, there is a polynomial time randomized algorithm that returns a solution* $\mathsf{z} \in \{0,1\}^m$ *such that*

$$\sum_{i=1}^m \mathsf{z}(i) \cdot \mathsf{v}_i \mathsf{v}_i^\top \succcurlyeq \mathsf{I}_d$$

with probability at least $1 - \exp(-\Omega(d))$*. Furthermore, for any* $\mathsf{c} \in \mathbb{R}_+^m$*, the solution* z *satisfies the upper bound*

$$\langle \mathsf{c}, \mathsf{z} \rangle \leqslant (1 + 6\varepsilon)\langle \mathsf{c}, \mathsf{x} \rangle + \frac{15d \|\mathsf{c}\|_\infty}{\varepsilon}$$

with probability at least $1 - \exp(-\Omega(d))$*, and the solution* z *satisfies the lower bound*

$$\langle \mathsf{c}, \mathsf{z} \rangle \geqslant \langle \mathsf{c}, \mathsf{x} \rangle - \delta d \|\mathsf{c}\|_\infty$$

with probability at least $1 - \exp(-\Omega(\min\{\varepsilon\delta, \varepsilon\delta^2\} \cdot d))$ *for* $\delta > 0$.

The main advantage of Theorem 5.1.3 over Theorem 5.1.1 is that we can prove $\langle \mathsf{c}, \mathsf{z} \rangle$ is not too far from $\langle \mathsf{c}, \mathsf{x} \rangle$ for an arbitrary vector $\mathsf{c} \in \mathbb{R}_+^m$ with high probability. Note that the guarantee on linear constraints can be applied to up to exponentially many constraints. This allows us to incorporate multiple linear constraints in applications to network design (see Chapter 6) and experimental design (see Chapter 7).

We note that there are examples showing that the additive error term $O(d \|c\|_\infty / \varepsilon)$ in Theorem 5.1.3 is tight up to a constant factor (see Section 5.4).

For two-sided spectral rounding, we show that Corollary 5.1.2 can be extended to incorporate one given non-negative linear constraint.

Theorem 5.1.4. *Let $v_1, ..., v_m \in \mathbb{R}^d$, $x \in [0,1]^m$ and $c \in \mathbb{R}^m_+$. Suppose $\sum_{i=1}^m x(i) \cdot v_i v_i^\top = I_d$, $\|v_i\| \leqslant \varepsilon < \frac{1}{8}$ for all $i \in [m]$ and $\|c\|_\infty \leqslant \varepsilon^2 \langle c, x \rangle$. Then there exists $z \subseteq \{0,1\}^m$ such that*

$$(1-8\varepsilon) \cdot I_d \preccurlyeq \sum_{i=1}^m z(i) \cdot v_i v_i^\top \preccurlyeq (1+8\varepsilon) \cdot I_d \quad \textit{and} \quad (1-8\varepsilon)\langle c, x \rangle \leqslant \langle c, z \rangle \leqslant (1+8\varepsilon)\langle c, x \rangle.$$

Note that the linear constraint c in Theorem 5.1.4 is required to be given as part of the input, while it is not required so in Theorem 5.1.3. Theorem 5.1.4 is useful in bounding the integrality gap for convex programs for network design problems, showing strong approximation results when the assumptions are satisfied (see Section 6.1.4). Also, we will show in Section 6.3 that it can be used in the study of additive unweighted spectral sparsification [20], proving an optimal existential result.

Technical Overview for the One-Sided Spectral Rounding Algorithm

Our algorithms for one-sided spectral rounding is based on the regret minimization framework developed in [7, 6] for spectral sparsification and experimental design. Let us first review the previous work. To prove Theorem 5.1.1, Allen-Zhu, Li, Singh, and Wang [6] analyzed a local search algorithm where they start from an arbitrary subset S_0 of k vectors, and in each iteration $t \geqslant 1$ they find a pair of vectors $i \in S_{t-1}$ and $j \notin S_{t-1}$ so that roughly speaking $\lambda_{\min}(\sum_{l \in S_{t-1}-i+j} v_l v_l^\top) > \lambda_{\min}(\sum_{l \in S_{t-1}} v_l v_l^\top)$, and then they set $S_t = S_{t-1} - i_t + j_t$. Using the framework of regret minimization, with the $\ell_{\frac{1}{2}}$-regularizer introduced in [7], they proved that the task of finding a pair to improve the minimum eigenvalue can be reduced to finding a pair $i_t \in S_{t-1}$ and $j_t \notin S_{t-1}$ so that

$$\frac{\langle v_{j_t} v_{j_t}^\top, A_t \rangle}{1 + 2\alpha \langle v_{j_t} v_{j_t}^\top, A_t^{\frac{1}{2}} \rangle} - \frac{\langle v_{i_t} v_{i_t}^\top, A_t \rangle}{1 - 2\alpha \langle v_{i_t} v_{i_t}^\top, A_t^{\frac{1}{2}} \rangle} \geqslant \Delta > 0, \tag{5.1}$$

where A_t is the action matrix defined in (4.9) based on the current solution S_{t-1}. Using a delicate argument, they proved that if $i_t \in S_{t-1}$ (subject to the restriction that $2\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle < 1$) is chosen to minimize $\langle v_i v_i^\top, A_t\rangle / \big(1 - 2\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle\big)$ and $j_t \notin S_{t-1}$ is chosen to maximize $\langle v_j v_j^\top, A_t\rangle / \big(1 + 2\alpha\langle v_j v_j^\top, A_t^{\frac{1}{2}}\rangle\big)$, then this pair i_t, j_t satisfies the above inequality with $\Delta = \frac{\varepsilon}{k}$ as long as $\lambda_{\min}(\sum_{l \in S_{t-1}} v_l v_l^\top) \leqslant 1 - 3\varepsilon$. This implies, by the regret minimization framework, that the local search algorithm will succeed to find a solution S_τ with $\lambda_{\min}(\sum_{l \in S_\tau} v_l v_l^\top) \geqslant 1 - 3\varepsilon$ within $\tau \leqslant \frac{k}{\varepsilon}$ iterations. See Chapter 4 for more details about the regret minimization framework.

To incorporate non-negative linear constraints, our idea is to turn the deterministic local search algorithm into an iterative randomized rounding algorithm. In this randomized rounding algorithm, we first construct an initial solution S_0 by adding each i into S_0 with probability $x(i)$ independently. This will ensure that $c(S_0) \approx \langle c, x\rangle$ with high probability. In each iteration $t \geqslant 1$, based on the current solution S_{t-1}, we construct a probability distribution to sample a vector v_{i_t} to be removed from S_{t-1}, and a probability distribution to sample a vector v_{j_t} to be added to S_{t-1}. To maintain $c(S_t) \approx \langle c, x\rangle$, the basic idea is to remove a vector v_i with probability proportional to $1 - x(i)$ and add a vector v_j with probability proportional to $x(j)$, but doing so may not satisfy the spectral lower bound with good probability. Instead, we prove that if we recompute the sampling probability so that a vector v_i is removed with probability proportional to $(1 - x(i)) \cdot (1 - 2\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle)$ and a vector v_j is added with probability proportional to $x(j) \cdot (1 + 2\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle)$, then (5.1) is satisfied with expected progress $\mathbb{E}[\Delta] \geqslant \frac{\varepsilon}{k}$ as long as $\lambda_{\min}(\sum_{l \in S_{t-1}} v_l v_l^\top) \leqslant 1 - 2\varepsilon$. Informally, a vector pointing to a direction that is not well covered by the current solution is more likely to be added and less likely to be removed, to ensure that the spectral lower bound will be satisfied. However, this changes the expectation on the linear constraint, but we can bound the error by the additive term $O(\frac{d\|c\|_\infty}{\varepsilon})$. Note that there are examples showing that this additive error is unavoidable if our goal is to satisfy the spectral lower bound exactly (see Section 5.4), so our analysis is tight up to a constant factor. Compared to the deterministic approach in [6], this randomized approach uses the fractional solution x more crucially in the rounding procedure, and we note that it can be used to give a simpler proof of the deterministic local search algorithm in [6] (see Remark 5.2.5).

The advantage of the randomized approach is that we can prove that the random variables are concentrated around their expected values, so that we can handle multiple non-negative linear constraints simultaneously. Since the sampling probabilities change over time based on the previous samples, the random variables that we consider are not a sum of independent random variables and thus Chernoff type bounds cannot be applied. For the spectral lower bound, we will define a martingale and use Freedman's inequality to prove that the total progress we make in (5.1) is concentrated around its expected value. For the non-negative linear constraints, we show that they satisfy an interesting "self-adjusting" property, such that if $c(S_t) - \langle c, x \rangle$ is (more) positive then $\mathbb{E}\left[c(S_{t+1})\right] - c(S_t)$ is (more) negative and vice versa, so intuitively $c(S_t) \approx \langle c, x \rangle$ with high probability for any t. This sequence of random variables is not a martingale and so Freedman's inequality cannot be applied. Instead, we prove a new concentration inequality (see Section 3.3) for this self-adjusting process that provides a quantitative bound similar to that in Freedman's inequality. We note that the iterative randomized rounding algorithm does not even need to know the linear constraint c in advance in order to return a solution S with $c(S) \approx \langle c, x \rangle$. This property is quite similar to that of a recent rounding algorithm by Bansal [16] combining iterative rounding and randomized rounding as we will discuss in Section 6.1.5.

We remark that our approach to turn a deterministic algorithm into a randomized algorithm is inspired by the fast algorithm for spectral sparsification by Lee and Sun [101], where they turned the deterministic algorithm by Batson, Spielman and Srivastava [21] into a randomized algorithm that recomputes the sampling probabilities in different phases (see Section 2.5 for more details). In their algorithm, the advantage of the randomized algorithm is to sample many vectors in parallel instead of carefully choosing one vector at a time as in [21]. In our algorithm, the advantage of the randomized algorithm is to approximately preserves many linear constraints simultaneously using arguments about expectation and concentration, while it is not clear how to modify the proofs in the deterministic local search algorithm in [6] to prove that there is always a pair of vectors v_i, v_j which makes enough progress in (5.1) and at the same time $c(j) - c(i)$ is small, even if there is only one constraint c and it is given in advance. We believe that this probabilistic approach will be useful in designing algorithms using the regret minimization framework.

5.2 Constructive One-Sided Spectral Rounding

In this section, we first present our iterative randomized rounding algorithm for one-sided spectral rounding and state the main technical result, a bicriteria approximation theorem, in Section 5.2.1. Then, we prove the bicriteria approximation theorem by analyzing the probability of approximately meeting the spectral requirement in Section 5.2.2 and analyzing the probability of approximately satisfying the linear constraint in Section 5.2.3. Finally, we prove our main result for one-sided spectral rounding Theorem 5.1.3 in Section 5.2.4.

5.2.1 Iterative Randomized Rounding Algorithm

We modify the deterministic local search algorithm in [6] to an iterative randomized rounding algorithm so as to approximately satisfy arbitrary non-negative linear constraints. In this randomized algorithm, we first construct an initial solution S_0 by adding each vector v_i into S_0 with probability $\mathsf{x}(i)$ independently. In each iteration $t \geqslant 1$, based on the current solution S_{t-1}, we construct a probability distribution to sample a vector v_{i_t} to be removed from S_{t-1}, and a probability distribution to sample a vector v_{j_t} to be added to S_{t-1}. The basic idea is that a vector v_i is removed with probability proportional to $1 - \mathsf{x}(i)$ and a vector v_j is added with probability proportional to $\mathsf{x}(j)$, but the probability is also adjusted based on the vector's contribution to the minimum eigenvalue of the current solution. We remark that it is possible that no vector is removed and/or no vector is added in an iteration. The algorithm stops when the minimum eigenvalue of the current solution is at least $1 - 2\varepsilon$. The following is the formal description of the algorithm.

Iterative Randomized Swapping Algorithm

```
Input: $v_1, ..., v_m \in \mathbb{R}^d$ and $x \in [0,1]^m$ with $\sum_{i=1}^m x(i) \cdot v_i v_i^\top = I_d$, and an error parameter $\gamma \in (0, \frac{1}{2})$.

Output: a subset $S \subseteq [m]$ such that $\sum_{i \in S} v_i v_i^\top \succcurlyeq (1-2\gamma) I_d$ and $c(S) \approx \langle c, x \rangle$ for any $c \in \mathbb{R}_+^m$ with high probability.

1. Initialization: $S_0 \leftarrow \emptyset$, $\alpha \leftarrow \frac{\sqrt{d}}{\gamma}$, $k \leftarrow m + 2\alpha\sqrt{d}$.
2. Add $i$ into $S_0$ independently with probability $x(i)$ for each $i \in [m]$.
3. Let $Z_1 \leftarrow \sum_{i \in S_0} v_i v_i^\top$ and $t \leftarrow 1$.
4. While $\lambda_{\min}(Z_t) < 1 - 2\gamma$ do
    (a) Compute the action matrix $A_t \leftarrow (\alpha Z_t - l_t I_d)^{-2}$, where $l_t \in \mathbb{R}$ is the unique value such that $A_t \succ 0$ and $\text{tr}(A_t) = 1$.[1]
    (b) Define $S'_{t-1} := \{i \in S_{t-1} : 2\alpha \langle v_i v_i^\top, A_t^{\frac{1}{2}} \rangle < \frac{1}{2}\}$.
    (c) Sample $i_t$ from the following probability distribution:
            $\mathbb{P}[i_t = i] = \frac{1}{k}(1 - x(i))(1 - 2\alpha \langle v_i v_i^\top, A_t^{\frac{1}{2}} \rangle)$ for $i \in S'_{t-1}$,
        and $\mathbb{P}[i_t = \emptyset] = 1 - \sum_{i \in S'_{t-1}} \mathbb{P}[i_t = i]$.
    (d) Sample $j_t$ from the following probability distribution:
            $\mathbb{P}[j_t = j] = \frac{x(j)}{k}(1 + 2\alpha \langle v_j v_j^\top, A_t^{\frac{1}{2}} \rangle)$ for $j \in [m] \backslash S_{t-1}$,
        and $\mathbb{P}[j_t = \emptyset] = 1 - \sum_{j \in [m] \backslash S_{t-1}} \mathbb{P}[j_t = j]$.
    (e) Set $S_t \leftarrow S_{t-1} \cup \{j_t\} \backslash \{i_t\}$, $Z_{t+1} \leftarrow \sum_{i \in S_t} v_i v_i^\top$, and $t \leftarrow t + 1$.
5. Return $S = S_{t-1}$ as the solution.
```

Before we state the main result of this algorithm, we first check that the algorithm is well-defined.

Claim 5.2.1. *The probability distributions in each iteration of the iterative randomized swapping algorithm are well-defined.*

Proof. To verify that the probability distribution for sampling i_t is well-defined, we need to show that $\mathbb{P}[i_t = i] \geqslant 0$ for $i \in S'_{t-1}$ and $\sum_{i \in S'_{t-1}} \mathbb{P}[i_t = i] \leqslant 1$. Since $A_t \succ 0$, $x(i) \in [0,1]$ and $2\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle \leqslant \frac{1}{2}$ for $i \in S'_{t-1}$, it follows that for $i \in S'_{t-1}$ we have

$$0 \leqslant \mathbb{P}[i_t = i] = \frac{1}{k}(1 - x(i))(1 - 2\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle) \leqslant \frac{1}{k},$$

and this implies that $\sum_{i \in S'_{t-1}} \mathbb{P}[i_t = i] \leqslant \frac{|S'_{t-1}|}{k} \leqslant \frac{m}{k} < 1$ by the definition of k.

Next we verify that the probability distribution for sampling j_t is well-defined. It is clear that $\mathbb{P}[j_t = j] \geqslant 0$ as $A_t \succ 0$ and $x(j) \in [0,1]$. We claim that

$$\sum_{j \in [m] \setminus S_{t-1}} \mathbb{P}[j_t = j] \leqslant \sum_{j \in [m]} \mathbb{P}(j_t = j) \leqslant 1$$

as

$$\begin{aligned}\sum_{j \in [m]} \mathbb{P}[j_t = j] = \frac{1}{k}\sum_{j=1}^{m} x(j) \cdot (1 + 2\alpha\langle v_j v_j^\top, A_t^{\frac{1}{2}}\rangle) &= \frac{1}{k}\left(\sum_{j=1}^{m} x(j) + 2\alpha\,\mathrm{tr}(A_t^{\frac{1}{2}})\right) \\ &\leqslant \frac{1}{k}\left(m + 2\alpha\sqrt{d}\right) = 1,\end{aligned}$$

where the second equality is by the assumption that $\sum_{j=1}^{m} x(j) \cdot v_j v_j^\top = I_d$, the last equality is by the definition of k, and the inequality uses that $x(j) \in [0,1]$, and the bound that $\mathrm{tr}(A_t^{\frac{1}{2}}) \leqslant \sqrt{d}$ from Claim 2.1.10. □

The following is the main technical result for one-sided spectral rounding.

[1] Step 4(a) can be implemented with a binary search step as mentioned in Remark 4.3.1.

Theorem 5.2.2. *Suppose we are given $v_1, ..., v_m \in \mathbb{R}^d$, $x \in [0,1]^m$ such that $\sum_{i=1}^m x(i) \cdot v_i v_i^\top = I_d$. For any $\gamma \in (0, \frac{1}{2})$, the iterative randomized swapping algorithm returns a subset $S \subseteq [m]$ satisfying*

$$\sum_{i \in S} v_i v_i^\top \succcurlyeq (1-2\gamma) \cdot I_d$$

within $\frac{qk}{\gamma}$ iterations with probability at least $1 - \exp\left(-\Omega(q\sqrt{d})\right)$ for $q \geqslant 2$. Furthermore, for any $c \in \mathbb{R}_+^m$ and any $\delta_1 \in [0,1]$, $\delta_2 \in [0,1]$ and $\delta_3 > 0$, the probability that the returned solution S satisfies the cost upper bound is

$$\mathbb{P}\left[c(S) \leqslant (1+\delta_1)\langle c, x\rangle + \frac{15d\,\|c\|_\infty}{\gamma}\right] \geqslant 1 - \exp\left[-\Omega\left(\frac{\delta_1 d}{\gamma}\right)\right],$$

and the probability that the returned solution S satisfies the cost lower bound is

$$\mathbb{P}\left[c(S) \geqslant (1-\delta_2)\langle c, x\rangle - \delta_3 d\,\|c\|_\infty\right] \geqslant 1 - \exp\Big[-\Omega\big(\min\{\delta_2\delta_3, \gamma\delta_3^2\} \cdot d\big)\Big].$$

Remark. *If we set $\delta_1 = \delta_2 = \gamma$ and $\delta_3 = \frac{1}{\gamma}$, then Theorem 5.2.2 states that the returned solution S satisfies*

$$(1-\gamma)\langle c, x\rangle - \frac{d\,\|c\|_\infty}{\gamma} \leqslant c(S) \leqslant (1+\gamma)\langle c, x\rangle + \frac{15d\,\|c\|_\infty}{\gamma}$$

with probability at least $1 - \exp(-\Omega(d))$ for any $c \in \mathbb{R}_+^m$. We introduce $\delta_1, \delta_2, \delta_3$ to have a more refined control of the failure probability of the lower bound, and this will be relevant in showing that linear covering constraints can be almost satisfied.

As a corollary of Theorem 5.2.2, we can also satisfy the linear constraint $c(S) \leqslant \langle c, x\rangle$ exactly when $\langle c, x\rangle$ is large by sacrificing a little bit in the spectral lower bound.

Corollary 5.2.3. *Let $v_1, \ldots, v_m \in \mathbb{R}^d$ and $x \in [0,1]^m$. Let $c \in \mathbb{R}_+^m$ and $C = \langle c, x\rangle$. Suppose $\sum_{i=1}^m x(i) \cdot v_i v_i^\top = I_d$ and $C \geqslant \frac{15d\|c\|_\infty}{\gamma^2}$ for some $\gamma \in (0, \frac{1}{2})$. Then, there is a randomized polynomial time algorithm that returns an integral solution $z \in \{0,1\}^m$ such that $\langle c, z\rangle \leqslant C$ and $\sum_{i=1}^m z(i) \cdot v_i v_i^\top \succcurlyeq (1-4\gamma) I_d$ with probability at least $1 - \exp(-\Omega(d))$.*

Proof. The idea is to scale down x then apply Theorem 5.2.2. We let $\eta = 1 - 2\gamma$ and set $y := \eta x$ and $u_i := \frac{v_i}{\sqrt{\eta}}$, which implies

$$\langle c, y \rangle = \eta \langle c, x \rangle = \eta C \qquad \text{and} \qquad \sum_{i=1}^{m} y(i) \cdot u_i u_i^\top = \sum_{i=1}^{m} x(i) \cdot v_i v_i^\top = I_d.$$

We apply Theorem 5.2.2 on $u_1, \ldots, u_m$ and y, c with $\delta_1 = \gamma, q = \sqrt{d}$ to obtain $z \in \{0,1\}^m$ so that

$$\sum_{i=1}^{m} z(i) \cdot u_i u_i^\top \succcurlyeq (1 - 2\gamma) I_d \qquad \Longrightarrow \qquad \sum_{i=1}^{m} z(i) \cdot v_i v_i^\top \succcurlyeq \eta(1 - 2\gamma) I_d \succcurlyeq (1 - 4\gamma) I_d.$$

and

$$\langle c, z \rangle \leqslant (1 + \gamma) \langle c, y \rangle + \frac{15d \left\| c \right\|_\infty}{\gamma} \leqslant (1 + \gamma)(1 - 2\gamma)C + \gamma C < C,$$

where we use the assumptions that $\frac{15d\|c\|_\infty}{\gamma^2} \leqslant C$. The failure probability is at most $\exp(-\Omega(d))$. □

5.2.2 Analysis of the Minimum Eigenvalue

In this subsection, we first show that the minimum eigenvalue of the current solution Z_t reaches the target $1 - 2\gamma$ within polynomial time with high probability in Section 5.2.2.1. This establishes the first part of Theorem 5.2.2.

Then we show that, if we continue the randomized swapping process after reaching the minimum eigenvalue target, the minimum eigenvalue of Z_t can be maintained at a relatively high level for a period of time with high probability in Section 5.2.2.2. We do not need the bounded minimum eigenvalue property for the results in this chapter. However, it is a key property in the analysis of the improved rounding algorithms for D/A-design (see Section 7.3).

5.2.2.1 Reaching the Minimum Eigenvalue Target

We first prove that, during the execution of the iterative randomized swapping algorithm, the probability of the minimum eigenvalue of Z_t is less than $1 - 2\gamma$ for all the first $\tau = \frac{qk}{\gamma}$

iterations is at most $\exp(-\Omega(q\sqrt{d}))$ for $q \geqslant 2$.

We will bound the minimum eigenvalue of the solution using the regret minimization framework (see Chapter 4). The initial feedback matrix is $F_0 = Z_1$, which is constructed randomly using x. In each iteration $t \geqslant 1$, after computing the action matrix A_t, the algorithm responds with the feedback matrix $F_t = v_{j_t} v_{j_t}^\top - v_{i_t} v_{i_t}^\top$. Note that $Z_{\tau+1} = \sum_{t=0}^{\tau} F_t$. Define

$$\Delta_t^+ := \frac{\langle v_{j_t} v_{j_t}^\top, A_t \rangle}{1 + 2\alpha \langle v_{j_t} v_{j_t}^\top, A_t^{\frac{1}{2}} \rangle}, \quad \Delta_t^- := \frac{\langle v_{i_t} v_{i_t}^\top, A_t \rangle}{1 - 2\alpha \langle v_{i_t} v_{i_t}^\top, A_t^{\frac{1}{2}} \rangle} \quad \text{and} \quad \Delta_t := \Delta_t^+ - \Delta_t^-. \tag{5.2}$$

Note that $2\alpha \langle v_{i_t} v_{i_t}^\top, A_t^{\frac{1}{2}} \rangle < \frac{1}{2} < 1$ for $t \geqslant 1$ by the definition of S'_{t-1}, and so Δ_t^- is well-defined for $t \geqslant 1$. We also note that $F_0 \not\succcurlyeq 0$, thus Theorem 4.2.7 implies that

$$\lambda_{\min}\left(Z_{\tau+1}\right) = \lambda_{\min}\left(\sum_{t=0}^{\tau} F_t\right) \geqslant \sum_{t=1}^{\tau} \Delta_t - \frac{2\sqrt{d}}{\alpha} + \lambda_{\min}(F_0) \geqslant \sum_{t=1}^{\tau} \Delta_t - 2\gamma. \tag{5.3}$$

To lower bound the minimum eigenvalue, we will prove that $\sum_{t=1}^{\tau} \Delta_t \geqslant 1$ with high probability. In the following, we bound the expected value of $\sum_{t=1}^{\tau} \Delta_t$, and then use Freeman's martingale inequality to bound the probability that $\sum_{t=1}^{\tau} \Delta_t$ deviates significantly from its expected value. In the remaining of this section, we write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid S_{t-1}]$ as the expectation conditional on the set S_{t-1}.

Lemma 5.2.4. *Let $\tau > \tau' \geqslant 0$ be two time steps in the iterative randomized swapping algorithm. Let $\lambda := \max_{\tau' < t \leqslant \tau} \lambda_{\min}(Z_t)$. Then*

$$\sum_{t=\tau'+1}^{\tau} \mathbb{E}_t[\Delta_t] \geqslant \sum_{t=\tau'+1}^{\tau} \frac{1}{k}\Big(1 - \frac{\sqrt{d}}{\alpha} - \lambda_{\min}(Z_t)\Big) \geqslant \frac{\tau - \tau'}{k}\Big(1 - \frac{\sqrt{d}}{\alpha} - \lambda\Big).$$

Proof. We first consider the expected gain of adding the vector j_t. By the definition of the probability distribution of j_t,

$$\begin{aligned}\mathbb{E}_t[\Delta_t^+] &= \frac{1}{k}\sum_{j\in[m]\setminus S_{t-1}} x(j)\cdot(1+2\alpha\langle v_jv_j^\top, A_t^{\frac{1}{2}}\rangle)\cdot\frac{\langle v_jv_j^\top, A_t\rangle}{1+2\alpha\langle v_jv_j^\top, A_t^{\frac{1}{2}}\rangle}\\ &= \frac{1}{k}\sum_{j\in[m]\setminus S_{t-1}} x(j)\cdot\langle v_jv_j^\top, A_t\rangle \\ &= \frac{1}{k}\Big(1-\sum_{j\in S_{t-1}} x(j)\langle v_jv_j^\top, A_t\rangle\Big),\end{aligned}\tag{5.4}$$

where the last equality is by $\sum_{j=1}^m x(j)\cdot v_jv_j^\top = I_d$ and $\mathrm{tr}(A_t)=1$ by the definition of A_t.

Then we consider the expected loss of removing the vector i_t. By the definition of the probability distribution of i_t,

$$\begin{aligned}\mathbb{E}_t[\Delta_t^-] &= \sum_{i\in S'_{t-1}}\frac{1}{k}(1-x(i))(1-2\alpha\langle v_iv_i^\top, A_t^{\frac{1}{2}}\rangle)\cdot\frac{\langle v_iv_i^\top, A_t\rangle}{1-2\alpha\langle v_iv_i^\top, A_t^{\frac{1}{2}}\rangle}\\ &= \frac{1}{k}\sum_{i\in S'_{t-1}}(1-x(i))\langle v_iv_i^\top, A_t\rangle\\ &\leqslant \frac{1}{k}\sum_{i\in S_{t-1}}(1-x_i)\langle v_iv_i^\top, A_t\rangle\\ &\leqslant \frac{1}{k}\Big(\lambda_{\min}(Z_t)+\frac{\sqrt{d}}{\alpha}-\sum_{i\in S_{t-1}}x(i)\cdot\langle v_iv_i^\top, A_t\rangle\Big),\end{aligned}\tag{5.5}$$

where the first inequality is because $x(i)\in[0,1]$ and $\langle v_iv_i^\top, A_t\rangle \geqslant 0$ as $A_t\succ 0$, and the last inequality follows from Lemma 4.2.9 that $\langle Z_t, A_t\rangle \leqslant \frac{\sqrt{d}}{\alpha}+\lambda_{\min}(Z_t)$.

The lemma follows by combining (5.4) and (5.5) and summing over t and using $\lambda = \max_t \lambda_{\min}(Z_t)$. □

Remark 5.2.5. *If we sample i_t with probability*

$$\mathbb{P}\left[i_t=i\right] = \frac{(1-x(i))(1-2\alpha\langle v_iv_i^\top, A_t^{\frac{1}{2}}\rangle)}{\sum_{j\in S'_{t-1}}(1-x(j))(1-2\alpha\langle v_jv_j^\top, A_t^{\frac{1}{2}}\rangle)}\quad \text{for } i\in S'_{t-1},$$

so that $\sum_{i \in S'_{t-1}} \mathbb{P}[i_t = i] = 1$ and define probability of sampling j_t likewise. We can maintain the size of S_t unchanged. Thus, we can start with a solution with $l := \sum_{i=1}^m x(i) + O(\frac{d}{\varepsilon})$ vectors and guarantee that the solution at each iteration still has exactly l vectors. A similar statement about the expected progress as in Lemma 5.2.4 can be proved. This implies that there exists a good pair $i_t \in S_{t-1}$ and $j_t \notin S_{t-1}$, which gives a solution set of size l satisfying the spectral lower bound approximately. Together with a preprocessing step as in Section 5.2.4, this gives a simpler proof of the deterministic algorithm of [6]. We use this idea to design a combinatorial algorithm for E-design without solving a convex program. When the input vectors to the combinatorial algorithm are normalized by the optimal fractional solution, the algorithm recovers the result in [6]. See Section 7.4.3 for more details.

Lemma 5.2.6. *Let $\tau > \tau' \geqslant 0$ be two time steps in the iterative randomized swapping algorithm*[2]*. Let $\lambda := \max_{\tau' < t \leqslant \tau} \lambda_{\min}(Z_t)$. Then, for any $\eta > 0$,*

$$\mathbb{P}\left[\sum_{t=\tau'+1}^{\tau} \Delta_t \leqslant \sum_{t=\tau'+1}^{\tau} \mathbb{E}_t[\Delta_t] - \eta\right] \leqslant \exp\left(-\frac{\eta^2 \alpha k/2}{(\tau - \tau')(1 + \lambda + \sqrt{d}/\alpha) + \eta k/3}\right).$$

Proof. We define the following sequences of random variables where

$$X_t := \mathbb{E}_t[\Delta_t] - \Delta_t, \quad Y_t := \sum_{l=\tau'+1}^{t} X_l, \quad \text{for all } t \geq \tau' + 1, \qquad \text{and} \qquad Y_{\tau'} = 0.$$

Observe that $\{Y_t\}_t$ is a martingale with respect to $\{S_t\}_t$. We use Freedman's inequality to bound $\mathbb{P}[Y_\tau \geqslant \eta]$. To apply Freedman's inequality, we need to upper bound X_t and $\mathbb{E}_t[X_t^2]$. Note that

$$0 \leqslant \Delta_t^+ = \frac{\langle v_{j_t} v_{j_t}^\top, A_t \rangle}{1 + 2\alpha \langle v_{j_t} v_{j_t}^\top, A_t^{\frac{1}{2}} \rangle} \leqslant \frac{\langle v_{j_t} v_{j_t}^\top, A_t \rangle}{2\alpha \langle v_{j_t} v_{j_t}^\top, A_t^{\frac{1}{2}} \rangle} \leqslant \frac{1}{2\alpha},$$

where the last inequality holds as $0 \prec A_t \preccurlyeq I$. Also,

$$0 \leqslant \Delta_t^- = \frac{\langle v_{i_t} v_{i_t}^\top, A_t \rangle}{1 - 2\alpha \langle v_{i_t} v_{i_t}^\top, A_t^{\frac{1}{2}} \rangle} \leqslant \frac{\langle v_{i_t} v_{i_t}^\top, A_t^{\frac{1}{2}} \rangle}{1 - 2\alpha \langle v_{i_t} v_{i_t}^\top, A_t^{\frac{1}{2}} \rangle} \leqslant \frac{1}{2\alpha},$$

[2] We introduce an arbitrary starting time step τ' for flexibility in later applications.

where the second last inequality is by $0 \prec A_t \preccurlyeq I$, and the first and last inequality are because i_t is chosen from the set $S'_{t-1} := \{i \mid 4\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle < 1\}$. (We remark that this upper bound on Δ_t^- is exactly the reason for the definition of S'_{t-1}.) As these lower and upper bounds on Δ_t^+ and Δ_t^- hold with probability one, we have the deterministic upper bound $X_t \leqslant R := \frac{1}{\alpha}$ as

$$X_t = \mathbb{E}_t[\Delta_t] - \Delta_t \leqslant \mathbb{E}_t[\Delta_t^+] + \Delta_t^- \leqslant \frac{1}{\alpha} = R.$$

Next, we upper bound

$$\begin{aligned}\mathbb{E}_t[X_t^2] \leqslant \mathbb{E}_t[\Delta_t^2] \leqslant R \cdot \mathbb{E}_t[|\Delta_t|] &\leqslant \frac{1}{\alpha}\Big(\mathbb{E}_t[\Delta_t^+] + \mathbb{E}_t[\Delta_t^-]\Big) \\ &\leqslant \frac{1}{\alpha k}\Big(1 + \lambda + \frac{\sqrt{d}}{\alpha}\Big),\end{aligned}$$

where the last inequality follows from (5.4) and (5.5) that $\mathbb{E}_t[\Delta_t^+] \leqslant \frac{1}{k}$ and $\mathbb{E}_t[\Delta_t^-] \leqslant \frac{\lambda+\sqrt{d}/\alpha}{k}$. Applying Freedman's inequality Theorem 3.2.3 with $R = \frac{1}{\alpha}$, $\sigma_t^2 = \frac{1+\lambda+\sqrt{d}/\alpha}{\alpha k}$ for all $\tau' < t \leqslant \tau$, and $\sigma^2 = \frac{(\tau-\tau')(1+\lambda+\sqrt{d}/\alpha)}{\alpha k}$, it follows that[3]

$$\mathbb{P}(Y_\tau \geqslant \eta) \leqslant \exp\left(-\frac{\eta^2/2}{\sigma^2 + R\eta/3}\right) = \exp\left(-\frac{\eta^2 \alpha k/2}{(\tau-\tau')(1+\lambda+\sqrt{d}/\alpha) + \eta k/3}\right).$$

The lemma follows as $Y_\tau \geqslant \eta$ is equivalent to $\sum_{t=1}^{\tau} \Delta_{\tau'+1} \leqslant \sum_{t=\tau'+1}^{\tau} \mathbb{E}_t[\Delta_t] - \eta$. □

We are ready to prove that the algorithm terminates in a polynomial number of iterations with high probability.

Theorem 5.2.7. *Let $\tau \geqslant 0$ be the first time such that the solution set S_τ of the iterative randomized swapping algorithm satisfies $\sum_{i \in S_\tau} v_i v_i^\top \succcurlyeq (1-2\gamma) \cdot I_d$. Then, the probability that $\tau \leqslant \frac{qk}{\gamma}$ for $q \geqslant 2$ is at most $\exp(-\Omega(q\sqrt{d}))$.*

Proof. Let $\tau = \frac{qk}{\gamma}$. Suppose $\lambda = \max_{0<t\leqslant\tau+1} \lambda_{\min}(Z_t) < 1-2\gamma$. Then, since $\alpha = \frac{\sqrt{d}}{\gamma}$, apply Lemma 5.2.4 with $\tau' = 0$ gives that

$$\sum_{t=1}^{\tau} \mathbb{E}_t[\Delta_t] \geqslant \frac{\tau}{k}\Big(1 - \frac{\sqrt{d}}{\alpha} - \lambda\Big) = \frac{q}{\gamma}(1-\gamma-\lambda) > q,$$

[3] Recall that the sequence $\{Y_t\}$ starts from $Y_{\tau'}$, instead of Y_0.

and the regret minimization bound in (5.3) implies that

$$1 - 2\gamma > \lambda_{\min}(Z_{\tau+1}) \geqslant \left(\sum_{t=1}^{\tau} \Delta_t\right) - 2\gamma \quad \Longrightarrow \quad \sum_{t=1}^{\tau} \Delta_t < 1.$$

Therefore,

$$\begin{aligned}\mathbb{P}\left[\bigcap_{t=1}^{\tau+1}\Big(\lambda_{\min}(Z_t) < 1 - 2\gamma\Big)\right] &\leqslant \mathbb{P}\left[\sum_{t=1}^{\tau}\Delta_t < \sum_{t=1}^{\tau}\mathbb{E}_t[\Delta_t] - (q-1)\right] \\ &\leqslant \exp\left(-\frac{(q-1)^2\alpha k/2}{(qk/\gamma)(1+(1-2\gamma)+\sqrt{d}/\alpha)+(q-1)k/3}\right) \\ &\leqslant \exp(-\Omega(q\sqrt{d})),\end{aligned}$$

where the second inequality is by Lemma 5.2.6 with $\eta = q - 1$ and $\tau = \frac{qk}{\gamma}, \tau' = 0$ and the last inequality is by the assumption that $q \geqslant 2$ and $\alpha = \frac{\sqrt{d}}{\gamma}$. □

So, for example, the probability that the algorithm does not terminate in $\frac{2k}{\gamma}$ iterations is at most $\exp(-\Omega(\sqrt{d}))$ and the probability that it does not terminate in $\frac{k\sqrt{d}}{\gamma}$ iterations is at most $\exp(-\Omega(d))$.

5.2.2.2 Maintaining the Minimum Eigenvalue

We show that, once the minimum eigenvalue of the current solution Z_t reaches the target $1 - 2\gamma$, the minimum eigenvalue of Z_t can be maintained being at least a constant for a period of time. We do not need this property to solve the one-sided spectral rounding problem, but it will be useful in applications to experimental design (see Section 7.3 for details).

Proposition 5.2.8. *Suppose $0 < \gamma \leqslant \frac{1}{8}$. Assume there is no termination condition in the iterative randomized swapping algorithm, and the minimum eigenvalue hit the target $\lambda_{\min}(Z_{\tau_1}) \geqslant 1 - 2\gamma$ at some time step τ_1, then the probability that $\lambda_{\min}(Z_t) \geqslant \frac{1}{4}$ for all the next τ steps $\tau_1 \leqslant t \leqslant \tau_1 + \tau$ is at least $1 - \tau^2 \cdot e^{-\Omega(\sqrt{d})}$.*

Proof. Consider the bad event that there exists a time $t \in [\tau_1, \tau_1 + \tau]$ with $\lambda_{\min}(Z_t) < \frac{1}{4}$. As the initial solution Z_{τ_1} satisfies $\lambda_{\min}(Z_{\tau_1}) \geqslant 1 - 2\gamma$, there must exist a time period $[t_0, t_1] \subseteq [\tau_1, \tau_1 + \tau)$ such that $\lambda_{\min}(Z_{t_0}) \geqslant 1 - 2\gamma$, $\lambda_{\min}(Z_{t_1+1}) < \frac{1}{4}$, and $\lambda_{\min}(Z_t) \in [\frac{1}{4}, 1 - 2\gamma)$ for all $t \in [t_0 + 1, t_1]$.

We show that the decrease of the minimum eigenvalue from t_0 to t_1 implies that the sum of Δ_t defined in (5.2) has decreased significantly. Let $F_{t_0} = Z_{t_0}$ and $F_t = v_{j_t} v_{j_t}^\top - v_{i_t} v_{i_t}^\top$ for all $t \in [t_0 + 1, t_1]$. Note that $\alpha \langle v_{i_t} v_{i_t}^\top, A_t^{\frac{1}{2}} \rangle \leqslant \frac{1}{4}$ holds for for any $t \geqslant 1$. So, it follows from Theorem 4.2.7 with $\alpha = \frac{\sqrt{d}}{\gamma}$ that

$$\frac{1}{4} > \lambda_{\min}(Z_{t_1+1}) \geqslant \sum_{t=t_0+1}^{t_1} \Delta_t - \frac{2\sqrt{d}}{\alpha} + \lambda_{\min}(Z_{t_0}) \geqslant \sum_{t=t_0+1}^{t_1} \Delta_t + 1 - 4\gamma \implies \sum_{t=t_0+1}^{t_1} \Delta_t < 4\gamma - \frac{3}{4}.$$

On the other hand, Δ_t is expected to be positive when $\lambda_{\min}(Z_t) < 1 - 2\gamma$. The expectation bound in Lemma 5.2.4 with $\tau' = t_0$, $\tau = t_1$, $\lambda < 1 - 2\gamma$ and $\alpha = \sqrt{d}/\gamma$ implies that

$$\sum_{t=t_0+1}^{t_1} \mathbb{E}_t[\Delta_t] \geqslant \frac{(t_1 - t_0)\gamma}{k}.$$

So, the sum has a large deviation from the expectation, i.e.

$$\sum_{t=t_0+1}^{t_1} \Delta_t \leqslant \sum_{t=t_0+1}^{t_1} \mathbb{E}_t[\Delta_t] - \left(\frac{3}{4} - 4\gamma + \frac{(t_1 - t_0)\gamma}{k}\right).$$

We can apply the concentration bound in Lemma 5.2.6 with $\eta = \frac{3}{4} - 4\gamma + \frac{(t_1 - t_0)\gamma}{k}$ and $\lambda < 1 - 2\gamma$ to upper bound this probability by

$$\begin{aligned} &\mathbb{P}\left[\sum_{t=t_0+1}^{t_1} \Delta_t \leqslant \sum_{t=t_0+1}^{t_1} \mathbb{E}_t[\Delta_t] - \left(\frac{3}{4} - 4\gamma + \frac{(t_1 - t_0)\gamma}{k}\right)\right] \\ \leqslant\ & \exp\left(-\frac{\left(\frac{3}{4} - 4\gamma + \frac{(t_1-t_0)\gamma}{k}\right)^2 \alpha k/2}{(t_1 - t_0)(1 + 1 - 2\gamma + \sqrt{d}/\alpha) + \left(\frac{3}{4} - 4\gamma + \frac{(t_1-t_0)\gamma}{k}\right)k/3}\right) \\ \leqslant\ & \exp\left(-\Omega(\sqrt{d})\right), \end{aligned}$$

where the last inequality follows as the denominator is in the order of $\Theta(k + t_1 - t_0)$ for $\alpha = \frac{\sqrt{d}}{\gamma}$ and $0 < \gamma \leqslant \frac{1}{8}$, and the numerator is in the order of $\Omega\Big(1+\frac{(t_1-t_0)\gamma}{k}+\frac{(t_1-t_0)^2\gamma^2}{k^2}\Big)\cdot\frac{k\sqrt{d}}{\gamma} = \Omega(\frac{k}{\gamma} + t_1 - t_0)\cdot\sqrt{d}$ for $\alpha = \frac{\sqrt{d}}{\gamma}$ and $0 < \gamma \leqslant \frac{1}{8}$. The proposition follows by applying the union bound over the at most τ^2 possible pairs of t_0 and t_1 from time τ_1 to $\tau_1 + \tau$. □

5.2.3 Analysis of the Linear Constraint

For an arbitrary non-negative linear constraint $c \in \mathbb{R}^m_+$, the goal in this subsection is to prove that $c(S_t) \approx \langle c, x\rangle$ with high probability for any t by using the concentration inequality for self-adjusting random processes (Theorem 3.3.1) in Section 3.3. We recall that $c(S_t) := \sum_{i\in S_t} c(i)$ is the "cost" of the solution at time t. We first bound the expected change of the cost in an iteration.

Lemma 5.2.9. *Suppose* $\lambda_{\min}(Z_t) < 1$. *Then*

$$\frac{1}{k}\Big(\langle c, x\rangle - c(S_{t-1})\Big) \leqslant \mathbb{E}_t[c(j_t) - c(i_t)] \leqslant \frac{1}{k}\Big(\langle c, x\rangle - c(S_{t-1}) + \frac{14d\,\|c\|_\infty}{\gamma}\Big).$$

Proof. We first bound the conditional expectation of $c(j_t)$. By the probability distribution of j_t,

$$\begin{aligned}\mathbb{E}_t[c(j_t)] &= \frac{1}{k}\sum_{j\in[m]\setminus S_{t-1}} c(j)x(j)(1 + 2\alpha\langle v_j v_j^\top, A_t^{\frac{1}{2}}\rangle)\\ &= \frac{1}{k}\Big(\langle c, x\rangle - \sum_{j\in S_{t-1}} c(j)\cdot x(j) + 2\alpha\sum_{j\in[m]\setminus S_{t-1}} c(j)x(j)\langle v_j v_j^\top, A_t^{\frac{1}{2}}\rangle\Big).\end{aligned}$$

Note that

$$\begin{aligned}0 \leqslant 2\alpha\sum_{j\in[m]\setminus S_{t-1}} c(j)x(j)\langle v_j v_j^\top, A_t^{\frac{1}{2}}\rangle &\leqslant 2\alpha\,\|c\|_\infty\sum_{j=1}^m x(j)\langle v_j v_j^\top, A_t^{\frac{1}{2}}\rangle\\ &= 2\alpha\,\|c\|_\infty\,\mathrm{tr}(A_t^{\frac{1}{2}}) \leqslant \frac{2d\,\|c\|_\infty}{\gamma},\end{aligned}$$

where the equality holds as $\sum_{j=1}^{m} x(j) \cdot v_j v_j^\top = I_d$ and the last inequality is by Claim 2.1.10 and $\alpha = \frac{\sqrt{d}}{\gamma}$. Therefore,

$$\frac{1}{k}\Big(\langle c, x\rangle - \sum_{i \in S_{t-1}} c(i)x(i)\Big) \leqslant \mathbb{E}_t[c(j_t)] \leqslant \frac{1}{k}\Big(\langle c, x\rangle - \sum_{i \in S_{t-1}} c(i)x(i) + \frac{2d\,\|c\|_\infty}{\gamma}\Big). \tag{5.6}$$

Next we bound the expectation of $c(i_t)$. By the probability distribution of i_t,

$$\begin{aligned}
\mathbb{E}_t[c(i_t)] &= \frac{1}{k}\sum_{i \in S'_{t-1}} c(i)(1 - x(i))(1 - 2\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle) \\
&= \frac{1}{k}\Big(\sum_{i \in S'_{t-1}} c(i)(1 - x(i)) - 2\alpha \sum_{i \in S'_{t-1}} c(i)(1 - x(i))\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle\Big) \\
&= \frac{1}{k}\Big(c(S_{t-1}) - \Big(\sum_{i \in S_{t-1}} c(i)x(i)\Big) - \sum_{i \in S_{t-1}\setminus S'_{t-1}} c(i)(1 - x(i)) \\
&\qquad - 2\alpha \sum_{i \in S'_{t-1}} c(i)(1 - x(i))\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle\Big).
\end{aligned}$$

We would like to bound the last two terms of the right hand side. Recall that $S'_{t-1} := \{i \in S_{t-1} \mid 4\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle < 1\}$. This implies that

$$\begin{aligned}
|S_{t-1} \setminus S'_{t-1}| \leqslant \sum_{i \in S_{t-1}\setminus S'_{t-1}} 4\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle &\leqslant 4\alpha \sum_{i \in S_{t-1}} \langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle \\
&\leqslant 4\big(d + \alpha\sqrt{d} \cdot \lambda_{\min}(Z_t)\big) \\
&\leqslant \frac{8d}{\gamma},
\end{aligned}$$

where the second last inequality uses Lemma 4.2.9 and the last inequality is by $\alpha = \frac{\sqrt{d}}{\gamma}$ and the assumption that $\lambda_{\min}(Z_t) < 1$. Since $x \in [0,1]^m$ and $c \geqslant 0$, it follows that the second last term is

$$0 \leqslant \sum_{i \in S_{t-1}\setminus S'_{t-1}} c(i)(1 - x(i)) \leqslant \|c\|_\infty \cdot |S_{t-1}\setminus S'_{t-1}| \leqslant \frac{8d\,\|c\|_\infty}{\gamma}.$$

Similarly, for the last term,

$$\begin{aligned} 0 \leqslant 2\alpha \sum_{i \in S'_{t-1}} c(i)(1-x(i))\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle &\leqslant 2\,\|c\|_\infty \cdot \alpha \sum_{i \in S_{t-1}} \langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle \\ &\leqslant 2\,\|c\|_\infty\,(d + \alpha\sqrt{d}\cdot \lambda_{\min}(Z_t)) \\ &\leqslant \frac{4d\,\|c\|_\infty}{\gamma}. \end{aligned}$$

Plugging back these upper and lower bounds for the last two terms, we obtain

$$\frac{1}{k}\left(c(S_{t-1}) - \left(\sum_{i \in S_{t-1}} c(i)x(i)\right) - \frac{12d\,\|c\|_\infty}{\gamma}\right) \leqslant \mathbb{E}_t[c(i_t)] \leqslant \frac{1}{k}\left(c(S_{t-1}) - \sum_{i \in S_{t-1}} c(i)x(i)\right). \tag{5.7}$$

The lemma follows by combining the bounds for the expectations of $c(i_t)$ and $c(j_t)$ in (5.6) and (5.7). □

To bound the difference between $c(S_t)$ and $\langle c, x\rangle$, we consider the following sequences of random variables where

$$Y_t := c(S_t) - \langle c, x\rangle \text{ for } t \geqslant 0 \quad \text{and} \quad X_t := Y_t - Y_{t-1} = c(j_t) - c(i_t) \text{ for } t \geqslant 1. \tag{5.8}$$

Note that Lemma 5.2.9 shows that the sequence $\{Y_t\}_t$ has the "self-adjusting" property that if Y_t is (more) positive then $\mathbb{E}[Y_{t+1} \mid Y_t] - Y_t$ is (more) negative and vice versa, so intuitively Y_t cannot be too far away from zero. The sequence $\{Y_t\}_t$ is not a martingale, and so we cannot apply Freedman's inequality to prove concentration. Instead, we will use Theorem 3.3.1 to prove that the absolute value of Y_t is small with high probability. To apply Theorem 3.3.1, we need to bound the conditional second moment of X_t and the moment generating function of the initial solution S_0.

Lemma 5.2.10. *Suppose $\lambda_{\min}(Z_t) < 1$. Then*

$$\mathbb{E}_t[(c(j_t) - c(i_t))^2] \leqslant \frac{\|c\|_\infty}{k} \cdot \left(\langle c, x\rangle + c(S_{t-1}) + \frac{2d\,\|c\|_\infty}{\gamma}\right).$$

Proof. Since $c(i) \geqslant 0$ for all $1 \leqslant i \leqslant m$,

$$\begin{aligned}
\mathbb{E}_t[(c(j_t) - c(i_t))^2] &\leqslant \max_{i_t, j_t} |c(j_t) - c(i_t)| \cdot \mathbb{E}_t[|c(j_t) - c(i_t)|] \\
&\leqslant \|c\|_\infty \cdot \mathbb{E}_t[c(j_t) + c(i_t)] \\
&\leqslant \frac{\|c\|_\infty}{k} \Big(\langle c, x \rangle + c(S_{t-1}) + \frac{2d \|c\|_\infty}{\gamma} \Big),
\end{aligned}$$

where the last inequality is by (5.6) and (5.7). □

We use the fact that the initial solution S_0 is generated randomly to bound its moment generating function.

Lemma 5.2.11. *For* $a \in \big[-\|c\|_\infty^{-1}, \|c\|_\infty^{-1} \big]$,

$$\mathbb{E}\left[e^{aY_0}\right] \leqslant e^{a^2 \|c\|_\infty \langle c, x \rangle}.$$

Proof. Let $\mathcal{X}_i$ be the indicator variable where $\mathcal{X}_i = 1$ if $i \in S_0$ and $\mathcal{X}_i = 0$ otherwise. Since the algorithm constructs S_0 by sampling each vector independently with probability $x(i)$, it follows that

$$\mathbb{E}\left[e^{ac(S_0)}\right] = \mathbb{E}\left[e^{a \sum_{i=1}^m \mathcal{X}_i c(i)}\right] = \prod_{i=1}^m \mathbb{E}\left[e^{a \mathcal{X}_i c(i)}\right] = \prod_{i=1}^m \left(1 - x(i) + x(i) e^{ac(i)}\right).$$

Note that $ac(i) \leqslant 1$ as $a \in \big[-\|c\|_\infty^{-1}, \|c\|_\infty^{-1} \big]$ and $c(i) \leqslant \|c\|_\infty$, and thus $e^{ac(i)} \leqslant 1 + ac(i) + a^2 c(i)^2$ as $e^p \leqslant 1 + p + p^2$ for $p \leqslant 1$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[e^{ac(S_0)}\right] \leqslant \prod_{i=1}^m \left(1 + ac(i)x(i) + a^2 c(i)^2 x(i)\right) &\leqslant \exp\left(\sum_{i=1}^m \left(ac(i)x(i) + a^2 \|c\|_\infty c(i)x(i) \right) \right) \\
&= \exp\left((a + a^2 \|c\|_\infty) \langle c, x \rangle \right),
\end{aligned}$$

where the second inequality uses $1 + p \leqslant e^p$ for $p \in \mathbb{R}$ and $c_i \leqslant \|c\|_\infty$ for $1 \leqslant i \leqslant m$. The claim follows as $Y_0 = c(S_0) - \langle c, x \rangle$. □

We are ready to apply the concentration inequality Theorem 3.3.1 to bound the cost.

Theorem 5.2.12. *Suppose the solution set S_t of the iterative randomized swapping algorithm satisfies $\lambda_{\min}(\sum_{i\in S_t} \mathsf{v}_i\mathsf{v}_i^\top) < 1$ for all $0 \leqslant t \leqslant \tau$. Then, for any $\mathsf{c} \in \mathbb{R}_+^m$ and $\delta_1 \in [0,1]$,*

$$\mathbb{P}\left[\mathsf{c}(S_\tau) \leqslant (1+\delta_1)\langle \mathsf{c}, \mathsf{x}\rangle + \frac{15d\left\|\mathsf{c}\right\|_\infty}{\gamma}\right] \geqslant 1 - \exp\left[-\Omega\Big(\frac{\delta_1 d}{\gamma}\Big)\right].$$

Also, for any $\delta_2 \in [0,1]$ and $\delta_3 > 0$,

$$\mathbb{P}\Big[\mathsf{c}(S_\tau) \geqslant (1-\delta_2)\langle \mathsf{c}, \mathsf{x}\rangle - \delta_3 d\left\|\mathsf{c}\right\|_\infty\Big] \geqslant 1 - \exp\Big(-\Omega\left(\min\{\delta_2\delta_3, \gamma\delta_3^2\}\cdot d\right)\Big).$$

Proof. We will apply Theorem 3.3.1 on the sequences $\{X_t\}_t$ and $\{Y_t\}_t$ as defined in (5.8). Firstly, note that $|X_t| \leqslant \left\|\mathsf{c}\right\|_\infty$ by definition for all $t \geqslant 1$. Secondly, as $\mathbb{E}_t[X_t] = \mathbb{E}_t[(\mathsf{c}(j_t) - \mathsf{c}(i_t))]$ and $Y_{t-1} = (\mathsf{c}(S_{t-1}) - \langle \mathsf{c}, \mathsf{x}\rangle)$, Lemma 5.2.9 implies that

$$\mathbb{E}_t[X_t] \leqslant \frac{1}{k}\Big(\langle \mathsf{c}, \mathsf{x}\rangle - \mathsf{c}(S_{t-1}) + \frac{14n\left\|\mathsf{c}\right\|_\infty}{\gamma}\Big) = -\frac{Y_{t-1}}{k} + \frac{14d\left\|\mathsf{c}\right\|_\infty}{k\gamma},$$

and

$$\mathbb{E}_t[X_t] \geqslant \frac{1}{k}\Big(\langle \mathsf{c}, \mathsf{x}\rangle - \mathsf{c}(S_{t-1})\Big) = -\frac{Y_{t-1}}{k}.$$

Thirdly, since $\mathbb{E}_t[X_t^2] = \mathbb{E}_t[(\mathsf{c}(j_t) - \mathsf{c}(i_t))^2]$, Lemma 5.2.10 implies that

$$\begin{aligned}\mathbb{E}_t[X_t^2] &\leqslant \frac{\left\|\mathsf{c}\right\|_\infty}{k}\Big(\langle \mathsf{c}, \mathsf{x}\rangle + \mathsf{c}(S_{t-1}) + \frac{2d\left\|\mathsf{c}\right\|_\infty}{\gamma}\Big)\\ &= \frac{\left\|\mathsf{c}\right\|_\infty}{k}Y_{t-1} + \frac{2\left\|\mathsf{c}\right\|_\infty}{k}\Big(\langle \mathsf{c}, \mathsf{x}\rangle + \frac{d\left\|\mathsf{c}\right\|_\infty}{\gamma}\Big).\end{aligned}$$

Finally, Lemma 5.2.11 states that $\mathbb{E}[e^{aY_0}] \leqslant \exp(a^2\left\|\mathsf{c}\right\|_\infty\langle \mathsf{c}, \mathsf{x}\rangle)$ for $a \in [-\left\|\mathsf{c}\right\|_\infty^{-1}, \left\|\mathsf{c}\right\|_\infty^{-1}]$. By setting

$$R = \left\|\mathsf{c}\right\|_\infty,\ \gamma_1 = \frac{1}{k},\ \gamma_2 = \frac{\left\|\mathsf{c}\right\|_\infty}{k},\ \beta_u = \frac{14d\left\|\mathsf{c}\right\|_\infty}{k\gamma},\ \beta_l = 0,\ \sigma = \frac{2\left\|\mathsf{c}\right\|_\infty}{k}\Big(\langle \mathsf{c}, \mathsf{x}\rangle + \frac{d\left\|\mathsf{c}\right\|_\infty}{\gamma}\Big),$$

we can check that all the conditions of Theorem 3.3.1 are satisfied, including the conditions on the range of parameters (in particular, $\gamma_1 \in (0, \frac{1}{2}), \gamma_2 > 0$ and $\gamma_1 \leqslant \gamma_2 R^{-1}$). Applying

Theorem 3.3.1 with $\eta = \delta_1\langle c, x\rangle + \frac{d\|c\|_\infty}{\gamma}$ for $\delta_1 \in [0,1]$,

$$\begin{aligned}\mathbb{P}\left[c(S_t) \geqslant (1+\delta_1)\langle c, x\rangle + \frac{15d\,\|c\|_\infty}{\gamma}\right] &= \mathbb{P}\left[Y_t \geqslant \frac{\beta_u}{\gamma_1} + \eta\right] \\ &\leqslant \exp\left[-\frac{\eta^2\gamma_1/\gamma_2}{4(\sigma/\gamma_2 + \beta_u/\gamma_1) + 2\eta}\right] \\ &= \exp\left[-\frac{\eta^2/\,\|c\|_\infty}{8\langle c, x\rangle + 64d\,\|c\|_\infty/\gamma + 2\eta}\right] \\ &\leqslant \exp\left[-\Omega\Big(\frac{\delta_1 d}{\gamma}\Big)\right],\end{aligned}$$

where the last inequality is because $\eta = O(\langle c, x\rangle + \frac{d\|c\|_\infty}{\gamma})$ and thus the denominator is $\Theta(\langle c, x\rangle + \frac{d\|c\|_\infty}{\gamma})$, and the numerator is (note that $\delta_1 \in [0,1]$)

$$\frac{\eta^2}{\|c\|_\infty} = \frac{\eta}{\|c\|_\infty}\Big(\delta_1\langle c, x\rangle + \frac{d\,\|c\|_\infty}{\gamma}\Big) \geqslant \frac{d\delta_1}{\gamma}\Big(\langle c, x\rangle + \frac{d\,\|c\|_\infty}{\gamma}\Big).$$

Similarly, for the cost lower bound, we apply Theorem 3.3.1 with $\eta = \delta_2\langle c, x\rangle + \delta_3 d\,\|c\|_\infty$ for $\delta_2 \in [0,1]$ and $\delta_3 > 0$ to obtain

$$\begin{aligned}\mathbb{P}\left[c(S_t) \leqslant (1-\delta_2)\langle c, x\rangle - \delta_3 d\,\|c\|_\infty\right] &= \mathbb{P}\left[Y_t \leqslant -\frac{\beta_l}{\gamma_1} - \eta\right] \\ &\leqslant \exp\left[-\frac{\eta^2\gamma_1/\gamma_2}{4\sigma/\gamma_2 + \eta}\right] \\ &= \exp\left[-\frac{\eta^2/\,\|c\|_\infty}{8(\langle c, x\rangle + d\,\|c\|_\infty/\gamma) + \eta}\right] \\ &\leqslant \exp\left[-\Omega\left(\frac{\delta_3 d(\delta_2\langle c, x\rangle + \delta_3 d\,\|c\|_\infty)}{\langle c, x\rangle + d\,\|c\|_\infty/\gamma + \delta_3 d\,\|c\|_\infty}\right)\right] \\ &\leqslant \exp\left[-\Omega\left(\min\{\delta_2\delta_3, \gamma\delta_3^2\}\cdot d\right)\right],\end{aligned}$$

where the second last inequality is by similar calculations as in the previous case. □

5.2.4 Exact One-Sided Spectral Rounding

Theorem 5.2.2 follows directly from Theorem 5.2.7 and Theorem 5.2.12. This shows that the iterative randomized swapping algorithm will return a solution S with $\sum_{i\in S} v_i v_i^\top \succcurlyeq$

$(1-2\gamma)\cdot I_d$ and $c(S_t) \approx \langle c, x\rangle$ with high probability for any $c \in \mathbb{R}^m_+$.

To prove Theorem 5.1.3 where the goal is to return a solution S with $\sum_{i \in S} v_i v_i^\top \succcurlyeq I_d$, our idea is to scale up the fractional solution x and then apply Theorem 5.2.2. The following is the detailed description of the algorithm.

Exact One-Sided Spectral Rounding

Input: $v_1, \ldots, v_m \in \mathbb{R}^d$ and $x \in [0,1]^m$ with $\sum_{i=1}^m x(i) v_i v_i^\top = I_d$, and an error parameter $\varepsilon \in (0, \frac{1}{4})$.

Output: a subset $S \subseteq [m]$ such that $\sum_{i \in S} v_i v_i^\top \succcurlyeq I_d$ and $c(S) \approx \langle c, x\rangle$ for any $c \in \mathbb{R}^m_+$ with high probability.

1. Define $y(i) := \frac{x(i)}{1-2\varepsilon}$ and $u_i := \sqrt{1-2\varepsilon} \cdot v_i$ for $i \in [m]$. Note that $\sum_{i=1}^m y(i) \cdot u_i u_i^\top = I_d$.
2. Let $S_{\text{big}} := \{i \in [m] : y(i) > 1\}$, $S_{\text{small}} := \{i \in [m] : 0 \leqslant y(i) \leqslant 1\}$, and $Z_{\text{big}} = \sum_{i \in S_{\text{big}}} y(i) \cdot u_i u_i^\top$.
3. Define $w_i := (I_d - Z_{\text{big}})^{-\frac{1}{2}} u_i$ for each $i \in S_{\text{small}}$, so that $\sum_{i \in S_{\text{small}}} y(i) \cdot w_i w_i^\top = I_d$[4].
4. Apply the iterative randomized swapping algorithm with $\gamma = \varepsilon$ and $\{w_i \mid i \in S_{\text{small}}\}$ and $\{y(i) \mid i \in S_{\text{small}}\}$ as input to obtain a solution set $S'_{\text{small}} \subseteq S_{\text{small}}$ with $\sum_{i \in S'_{\text{small}}} w_i w_i^\top \succcurlyeq (1-2\varepsilon) \cdot I_d$.
5. Return $S := S_{\text{big}} \cup S'_{\text{small}}$ as the solution.

Theorem 5.1.3. *Suppose we are given $v_1, \ldots, v_m \in \mathbb{R}^d$ and $x \in [0,1]^m$ such that $\sum_{i=1}^m x(i) \cdot v_i v_i^\top = I_d$. For any $\varepsilon \in (0, \frac{1}{4})$, there is a polynomial time randomized algorithm that returns*

[4] If $I_d - Z_{\text{big}}$ is singular, we first project the vectors to the orthogonal complement of the nullspace before applying the transformation. We can add dummy coordinates to keep the vectors to have the same dimension d for simplicity of the analysis.

a solution $\mathsf{z} \in \{0,1\}^m$ such that

$$\sum_{i=1}^{m} \mathsf{z}(i) \cdot \mathsf{v}_i \mathsf{v}_i^\top \succcurlyeq \mathsf{I}_d$$

with probability at least $1 - \exp(-\Omega(d))$. Furthermore, for any $\mathsf{c} \in \mathbb{R}_+^m$, the solution z satisfies the upper bound

$$\langle \mathsf{c}, \mathsf{z} \rangle \leqslant (1 + 6\varepsilon)\langle \mathsf{c}, \mathsf{x} \rangle + \frac{15d \, \|\mathsf{c}\|_\infty}{\varepsilon}$$

with probability at least $1 - \exp(-\Omega(d))$, and the solution z satisfies the lower bound

$$\langle \mathsf{c}, \mathsf{z} \rangle \geqslant \langle \mathsf{c}, \mathsf{x} \rangle - \delta d \, \|\mathsf{c}\|_\infty$$

with probability at least $1 - \exp\left(-\Omega\left(\min\{\varepsilon\delta, \varepsilon\delta^2\} \cdot d\right)\right)$ for $\delta > 0$.

Proof. We first analyze the spectral lower bound. Applying Theorem 5.2.2 with $\gamma = \varepsilon$ and $q = \sqrt{d}$ to Step 4, we find a solution set $S'_{\text{small}} \subseteq S_{\text{small}}$ with probability at least $1 - \exp(-\Omega(d))$ in polynomial time such that

$$\begin{aligned} \sum_{i \in S'_{\text{small}}} \mathsf{w}_i \mathsf{w}_i^\top \succcurlyeq (1 - 2\varepsilon) \cdot \mathsf{I}_d \implies & \sum_{i \in S'_{\text{small}}} \mathsf{u}_i \mathsf{u}_i^\top \succcurlyeq (1 - 2\varepsilon) \cdot (\mathsf{I}_d - \mathsf{Z}_{\text{big}}) \\ \implies & \sum_{i \in S'_{\text{small}}} \mathsf{v}_i \mathsf{v}_i^\top \succcurlyeq \mathsf{I}_d - \mathsf{Z}_{\text{big}}. \end{aligned}$$

For the vectors in S_{big}, as $\mathsf{x}(i) \in [0,1]$,

$$\sum_{i \in S_{\text{big}}} \mathsf{v}_i \mathsf{v}_i^\top \succcurlyeq \sum_{i \in S_{\text{big}}} \mathsf{x}(i) \cdot \mathsf{v}_i \mathsf{v}_i^\top = \sum_{i \in S_{\text{big}}} \mathsf{y}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top = \mathsf{Z}_{\text{big}}.$$

Therefore, it follows that

$$\sum_{i \in S} \mathsf{v}_i \mathsf{v}_i^\top = \sum_{i \in S'_{\text{small}}} \mathsf{v}_i \mathsf{v}_i^\top + \sum_{i \in S_{\text{big}}} \mathsf{v}_i \mathsf{v}_i^\top \succcurlyeq (\mathsf{I}_d - \mathsf{Z}_{\text{big}}) + \mathsf{Z}_{\text{big}} = \mathsf{I}_d.$$

Next, we prove that $c(S) \approx \langle c, x \rangle$ with high probability for any vector $c \in \mathbb{R}^m_+$. Let $\langle c, x \rangle_{\text{small}} := \sum_{i \in S_{\text{small}}} c(i) \cdot x(i)$ and $\langle c, x \rangle_{\text{big}} := \sum_{i \in S_{\text{big}}} c(i) \cdot x(i)$. For the vectors in S_{big}, as $y(i) > 1$ for $i \in S_{\text{big}}$ and $y(i) = \frac{x(i)}{1-2\varepsilon}$ for all $i \in [m]$, it follows that

$$\langle c, x \rangle_{\text{big}} \leqslant c(S_{\text{big}}) \leqslant \langle c, y \rangle_{\text{big}} = \frac{\langle c, x \rangle_{\text{big}}}{1 - 2\varepsilon}.$$

For the vectors in S_{small}, applying Theorem 5.2.2 with $\delta_1 = \varepsilon$ and $\gamma = \varepsilon$, the returned set S'_{small} in Step 4 satisfies the cost upper bound

$$c(S'_{\text{small}}) \leqslant (1+\varepsilon)\langle c, y \rangle_{\text{small}} + \frac{15d \left\| c \right\|_\infty}{\varepsilon} = \frac{(1+\varepsilon)\langle c, x \rangle_{\text{small}}}{1-2\varepsilon} + \frac{15d \left\| c \right\|_\infty}{\varepsilon}$$

with probability at least $1 - \exp(-\Omega(d))$, which implies that for $\varepsilon \in (0, \frac{1}{4})$,

$$\begin{aligned} c(S) = c(S_{\text{big}}) + c(S'_{\text{small}}) &\leqslant \frac{1+\varepsilon}{1-2\varepsilon}\big(\langle c, x \rangle_{\text{big}} + \langle c, x \rangle_{\text{small}}\big) + \frac{15d \left\| c \right\|_\infty}{\varepsilon} \\ &\leqslant (1+6\varepsilon)\langle c, x \rangle + \frac{15d \left\| c \right\|_\infty}{\varepsilon}. \end{aligned}$$

Similarly, by Theorem 5.2.2 with $\gamma = \varepsilon$, $\delta_2 = \varepsilon$ and $\delta_3 = \delta$ for some $\delta > 0$, the returned set S'_{small} in Step 4 satisfies the cost lower bound

$$c(S'_{\text{small}}) \geqslant (1-\varepsilon)\langle c, y \rangle_{\text{small}} - \delta d \left\| c \right\|_\infty = \frac{1-\varepsilon}{1-2\varepsilon}\langle c, x \rangle_{\text{small}} - \delta d \left\| c \right\|_\infty \geqslant \langle c, x \rangle_{\text{small}} - \delta d \left\| c \right\|_\infty$$

with probability at least $1 - \exp(-\Omega(\min\{\varepsilon\delta, \varepsilon\delta^2\} \cdot d))$, which implies that

$$c(S) = c(S_{\text{big}}) + c(S'_{\text{small}}) \geqslant \langle c, x \rangle_{\text{big}} + \langle c, x \rangle_{\text{small}} - \delta d \left\| c \right\|_\infty = \langle c, x \rangle - \delta d \left\| c \right\|_\infty. \qquad \square$$

5.3 Non-constructive Two-Sided Spectral Rounding

In this section, we show that the two-sided spectral rounding result in Theorem 2.6.10 can be extended to incorporate one non-negative linear constraint that is given as part of the input.

There is a standard reduction used in [48] to construct spectral sparsifiers that satisfy additional linear constraints. Suppose Corollary 5.1.2 were to work for rank two matrices, then we can simply incorporate the linear constraint to the input matrices as $A_i := \begin{pmatrix} v_i v_i^\top & 0 \\ 0 & c(i)/\langle c, x\rangle \end{pmatrix}$ so that $\sum_{i=1}^m x(i) \cdot A_i = I_{d+1}$, and any $z \in \{0,1\}^m$ so that $\sum_{i=1}^m z(i) \cdot A_i \approx I_{d+1}$ would have $\langle c, z\rangle \approx \langle c, x\rangle$. But the rank one assumption is crucial in the proof of Theorem 2.6.10 and it is an open problem to generalize it to work with higher rank matrices.

Our idea is to use the following signing trick, suggested to us by Akshay Ramachandran, to essentially carry out the same reduction using only rank one matrices. We state the results in a more general form, where $\sum_{i=1}^m x(i) \cdot v_i v_i^\top$ is not necessarily equal to the identity matrix, so that we can also apply them to additive spectral sparsifiers in Section 6.3.

Lemma 5.3.1. *Suppose we are given $c \in \mathbb{R}_+^m$, and $v_1, \ldots, v_m \in \mathbb{R}^d$, $x \in [0,1]^m$ such that $\left\|\sum_{i=1}^m x(i) \cdot v_i v_i^\top\right\|_{\text{op}} \leqslant \lambda$ and $\|v_i\|_2 \leqslant l$ for $1 \leqslant i \leqslant m$. There exists a signing $s \in \{\pm 1\}^m$ such that if we let $u_i := \begin{pmatrix} v_i \\ s(i)\sqrt{c(i)\lambda/\langle c, x\rangle} \end{pmatrix} \in \mathbb{R}^{d+1}$ then $\left\|\sum_{i=1}^m x(i) \cdot u_i u_i^\top\right\|_{\text{op}} \leqslant \lambda + l\sqrt{\lambda}$.*

Proof. By the definition of u_i,

$$\sum_{i=1}^m x(i) \cdot u_i u_i^\top = \begin{pmatrix} \sum_{i=1}^m x(i) \cdot v_i v_i^\top & \sum_{i=1}^m s(i)x(i)\sqrt{\frac{c(i)\lambda}{\langle c, x\rangle}} v_i \\ \sum_{i=1}^m s(i)x(i)\sqrt{\frac{c(i)\lambda}{\langle c, x\rangle}} v_i^\top & \sum_{i=1}^m \frac{c(i)x(i)\lambda}{\langle c, x\rangle} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^m x(i) \cdot v_i v_i^\top & 0 \\ 0 & \lambda \end{pmatrix} + \begin{pmatrix} 0 & \sum_{i=1}^m s(i)x(i)\sqrt{\frac{c(i)\lambda}{\langle c, x\rangle}} v_i \\ \sum_{i=1}^m s(i)x(i)\sqrt{\frac{c(i)\lambda}{\langle c, x\rangle}} v_i^\top & 0 \end{pmatrix}$$

The operator norm of the second matrix is bounded by $\left\|\sum_{i=1}^m s(i)x(i)\sqrt{\frac{c(i)\lambda}{\langle c,x\rangle}} v_i\right\|_2$. It follows from triangle inequality that $\left\|\sum_{i=1}^m x(i) \cdot u_i u_i^\top\right\|_{\text{op}} \leqslant \lambda + \left\|\sum_{i=1}^m s(i)x(i)\sqrt{\frac{c(i)\lambda}{\langle c,x\rangle}} v_i\right\|_2$. We show that there is a signing $s \in \{\pm 1\}^m$ such that

$$\left\|\sum_{i=1}^m s(i)x(i)\sqrt{\frac{c(i)\lambda}{\langle c, x\rangle}} \cdot v_i\right\|_2 \leqslant l\sqrt{\lambda},$$

and this will complete the proof. Take a uniform random signing and consider

$$\begin{aligned}
&\mathbb{E}_{s \in \{\pm 1\}^m} \left\| \sum_{i=1}^m s(i) x(i) \sqrt{\frac{c(i)\lambda}{\langle c, x\rangle}} v_i \right\|_2^2 \\
&= \sum_{i=1}^m \mathbb{E}_s \left[s(i)^2 x(i)^2 \left\| v_i \right\|_2^2 \frac{\lambda c(i)}{\langle c, x\rangle} \right] + \sum_{i \neq j} \mathbb{E}_s \left[s(i)s(j)x(i)x(j) \langle v_i, v_j \rangle \frac{\lambda \sqrt{c(i)c(j)}}{\langle c, x\rangle} \right] \\
&= \sum_{i=1}^m x(i)^2 \left\| v_i \right\|_2^2 \frac{\lambda c(i)}{\langle c, x\rangle} \leqslant l^2 \sum_{i=1}^m \frac{\lambda c(i) x(i)}{\langle c, x\rangle} = l^2 \lambda,
\end{aligned}$$

where the last line uses that $s(i)^2 = 1$, $\mathbb{E}[s(i)s(j)] = \mathbb{E}[s(i)] \cdot \mathbb{E}[s(j)] = 0$, and $x(i) \in [0, 1]$, $\left\| v_i \right\|_2 \leqslant l$ in the inequality. This implies that there exists such a signing. □

We apply the signing in Lemma 5.3.1 to incorporate one non-negative linear constraint into the two-sided spectral rounding result of Kyng, Luh and Song [91].

Theorem 5.3.2. *Suppose we are given $c \in \mathbb{R}_+^m$, and $v_1, \ldots, v_m \in \mathbb{R}^d$, $x \in [0, 1]^m$ such that $\left\| \sum_{i=1}^m x(i) \cdot v_i v_i^\top \right\|_{\text{op}} \leqslant \lambda$ and $\left\| v_i \right\|_2 \leqslant l$ for $1 \leqslant i \leqslant m$. Suppose further that $\left\| c \right\|_\infty \leqslant \frac{l^2 \langle c, x\rangle}{\lambda}$ and $l \leqslant \sqrt{\lambda}$. Then there exists $z \in \{0, 1\}^m$ such that*

$$\left\| \sum_{i=1}^m x(i) \cdot v_i v_i^\top - \sum_{i=1}^m z(i) \cdot v_i v_i^\top \right\|_{\text{op}} \leqslant 8l\sqrt{\lambda} \quad \text{and} \quad |\langle c, x\rangle - \langle c, z\rangle| \leqslant \frac{8l}{\sqrt{\lambda}} \langle c, x\rangle$$

Proof. Let $u_i = \begin{pmatrix} v_i \\ s(i)\sqrt{c(i)\lambda / \langle c, x\rangle} \end{pmatrix}$ for $1 \leqslant i \leqslant m$, where $s \in \{\pm 1\}^m$ is the signing promised by Lemma 5.3.1. By the assumption that $\left\| c \right\|_\infty \leqslant \frac{l^2 \langle c, x\rangle}{\lambda}$, it follows that $\left\| u_i \right\|_2^2 = \left\| v_i \right\|_2^2 + \frac{c(i)\lambda}{\langle c, x\rangle} \leqslant 2l^2$. Let ξ_i be a zero-one random variable with probability $x(i)$ being one. Applying Theorem 2.6.10 on $u_1, \ldots, u_m$ and $\xi_1, \ldots, \xi_m$, there exists $z \in \{0, 1\}^m$ such that

$$\begin{aligned}
\left\| \sum_{i=1}^m x(i) \cdot u_i u_i^\top - \sum_{i=1}^m z(i) \cdot u_i u_i^\top \right\|_{\text{op}} &\leqslant 4 \left\| \sum_{i=1}^m \mathbf{Var}[\xi_i] (u_i u_i^\top)^2 \right\|_{\text{op}}^{\frac{1}{2}} \\
&\leqslant 4 \left\| \sum_{i=1}^m x(i) \left\| u_i \right\|_2^2 u_i u_i^\top \right\|_{\text{op}}^{\frac{1}{2}} \\
&\leqslant 4 \sqrt{2l^2 (\lambda + l\sqrt{\lambda})},
\end{aligned}$$

where we use that $\mathbf{Var}[\xi_i] = x(i)(1 - x(i)) \leqslant x(i)$, $\|u_i\|_2^2 \leqslant 2l^2$ and $\left\|\sum_{i=1}^m x(i) \cdot u_i u_i^\top\right\|_{\mathrm{op}} \leqslant \lambda + l\sqrt{\lambda}$ by Lemma 5.3.1. By looking at the top left $d \times d$ block, this implies that

$$\left\|\sum_{i=1}^m x(i) \cdot v_i v_i^\top - \sum_{i=1}^m z(i) \cdot v_i v_i^\top\right\|_{\mathrm{op}} \leqslant 4\sqrt{2l^2(\lambda + l\sqrt{\lambda})} \leqslant 8l\sqrt{\lambda}$$

where we use the assumption that $l \leqslant \sqrt{\lambda}$. By looking at the bottom right entry, we have

$$\left|\sum_{i=1}^m \frac{x(i)c(i)\lambda}{\langle c, x\rangle} - \sum_{i=1}^m \frac{z(i)c(i)\lambda}{\langle c, x\rangle}\right| \leqslant 4\sqrt{2l^2(\lambda + l\sqrt{\lambda})} \leqslant 8l\sqrt{\lambda},$$

which implies $|\langle c, x\rangle - \langle c, z\rangle| \leqslant \frac{8l}{\sqrt{\lambda}}\langle c, x\rangle$. □

This proves Theorem 5.1.4 that incorporates one non-negative linear constraint into Corollary 5.1.2, by plugging $\lambda = 1$ and $l = \varepsilon$ into Theorem 5.3.2.

5.4 Tight Examples

We provide two examples showing the tightness of Theorem 5.1.3.

First, consider the following simple example, which shows the $d\|c\|_\infty$ additive error term is necessary.

Example 5.4.1. *There are $m = 2d$ vectors $v_{11}, v_{12}, ..., v_{d1}, v_{d2} \in \mathbb{R}^d$, a vector $x \in [0, 1]^m$, a vector $c \in \mathbb{R}_+^m$, and a parameter ε. They are defined as follows*

$$\begin{aligned} &x(i1) = 1, \quad v_{i1} = \sqrt{1-\varepsilon} \cdot e_i, \quad c(i1) = 0 \qquad \textit{and} \\ &x(i2) = \varepsilon, \quad v_{i2} = e_i, \quad c(i2) = \|c\|_\infty, \quad \forall i \in \{1, ..., d\}. \end{aligned}$$

Note that $\langle c, x\rangle = \varepsilon d\|c\|_\infty$ and $\sum_{i=1}^d \sum_{j=1,2} x(ij) \cdot v_{ij} v_{ij}^\top = I_d$.

Claim 5.4.2. *For any constant $\alpha > 1$, any $z \in \{0, 1\}^m$ satisfying the spectral lower bound in Example 5.4.1 must have $\langle c, z\rangle \geqslant \alpha\langle c, x\rangle + \Omega(d\|c\|_\infty)$.*

Proof. Note that the only vector $z \in \{0,1\}^m$ that satisfies the spectral lower bound exactly is $z = 1_m$. This implies that $\langle c, z\rangle - \alpha\langle c, x\rangle = d\,\|c\|_\infty - \alpha\varepsilon d\,\|c\|_\infty = (1-\alpha\varepsilon)d\,\|c\|_\infty$. For any $\alpha > 1$, there exists ε such that $\langle c, z\rangle - \alpha\langle c, x\rangle$ is at least say $\frac{d\|c\|_\infty}{2}$. □

Next, we modify an integrality gap example in [118] to show that, even if $c = 1$ and we are allowing integral-solution instead of zero-one solution, the additive error $O(\frac{d\|c\|_\infty}{\varepsilon})$ in Theorem 5.1.3 is best possible.

Example 5.4.3. *The example contains $m = \binom{n}{2}$ vectors $v_1, \ldots, v_m \in \mathbb{R}^{n-1}$, a vector $x \in [0,1]^m$ and a vector $c = 1_m$. Let $U \in \mathbb{R}^{(n-1)\times n}$ be a matrix where the rows form an orthonormal basis of the $(n-1)$-dimensional subspace orthogonal to 1_n. Given some parameter k, we define*

$$v_{ij} = \sqrt{\frac{n-1}{2k}} \cdot U(\chi_i - \chi_j) \qquad \text{and} \qquad x(ij) = \frac{2k}{n(n-1)}, \qquad \forall 1 \leqslant i < j \leqslant n.$$

Note that $\langle c, x\rangle = k$ and $\sum_{i<j} x(ij)\cdot v_{ij}v_{ij}^\top = I_{n-1}$ and x has the smallest $\|x\|_1$ among all vectors satisfying $\sum x(ij)\cdot v_{ij}v_{ij}^\top \succcurlyeq I_{n-1}$.

We will use the following result from [118].

Theorem 5.4.4 (Theorem 7.2 in [118])**.** *Let $G = (V, E)$ be a graph with average degree $d_{\text{avg}} = \frac{2m}{n}$, and let L_G be its unnormalized Laplacian matrix. Then, as long as d_{avg} is large enough, and n is large enough with respect to d_{avg},*

$$\lambda_2(L_G) \leqslant d_{\text{avg}} - \rho\sqrt{d_{\text{avg}}},$$

where $\lambda_2(L_G)$ is the second smallest eigenvalue of L_G, and $\rho > 0$ is an absolute constant. Furthermore, the upper bound for $\lambda_2(L_G)$ still holds for graphs with parallel edges.

Using the above theorem, we can prove the following lemma.

Lemma 5.4.5. *Let $\{v_{ij}\}, c, x$ be defined as in Example 5.4.3. For any $z \in \mathbb{Z}_+^m$, if it satisfies $\sum_{1\leqslant i<j\leqslant n} z(ij)\cdot v_{ij}v_{ij}^\top \succcurlyeq I_{n-1}$, then we have*

$$\langle c, z\rangle \geqslant k + \Omega(\sqrt{kn} + n).$$

Proof. Given any $\mathsf{z} \in \mathbb{Z}_+^m$, let G_{z} be the multi-graph corresponding to z with Laplacian matrix

$$\begin{aligned} \mathsf{L}_{\mathsf{z}} &= \sum_{1 \leqslant i < j \leqslant n} \mathsf{z}(ij)(\chi_i - \chi_j)(\chi_i - \chi_j)^\top \\ \Longrightarrow \qquad \mathsf{U}\mathsf{L}_{\mathsf{z}}\mathsf{U}^\top &= \frac{2k}{n-1}\left(\sum_{1 \leqslant i < j \leqslant n} \mathsf{z}(ij) \cdot \mathsf{v}_{ij}\mathsf{v}_{ij}^\top\right) \succcurlyeq \frac{2k}{n-1}\mathsf{I}_{n-1}, \end{aligned}$$

where the last inequality holds by the assumption on z. Since U is the projection onto the $(n{-}1)$-dimensional subspace that orthogonal to the minimum eigenvector $\mathit{1}_n$, $\lambda_2(\mathsf{L}_{\mathsf{z}}) \geqslant \frac{2k}{n-1}$.

On the other hand, since the average degree of G_{z} is $d_{\text{avg}} = \frac{2\|\mathsf{z}\|_1}{n}$, we apply Theorem 5.4.4 with properly chosen n, for some constant ρ we have

$$\lambda_2(\mathsf{L}_{\mathsf{z}}) \leqslant d_{\text{avg}} - \rho\sqrt{d_{\text{avg}}} \qquad \Longrightarrow \qquad \lambda_2(\mathsf{L}_{\mathsf{z}}) \leqslant \frac{2\|\mathsf{z}\|_1}{n} - \rho\sqrt{\frac{2\|\mathsf{z}\|_1}{n}}.$$

Combining with $\lambda_2(\mathsf{L}_{\mathsf{z}}) \geqslant \frac{2k}{n-1}$, we have

$$\frac{2k}{n-1} \leqslant \frac{2\|\mathsf{z}\|_1}{n} - \rho\sqrt{\frac{2\|\mathsf{z}\|_1}{n}} \quad \Longrightarrow \quad 2k \leqslant 2\|\mathsf{z}\|_1 - \rho\sqrt{2n\|\mathsf{z}\|_1}.$$

For the quadratic inequality $2y^2 - \rho\sqrt{2n}y - 2k \geqslant 0$, we know that the nonnegative solution for y should satisfy

$$y \geqslant \frac{\rho\sqrt{2n} + \sqrt{2\rho^2 n + 16k}}{4}.$$

Therefore, letting $y = \sqrt{\|\mathsf{z}\|_1}$, we have

$$\begin{aligned} \langle \mathsf{c}, \mathsf{z}\rangle = \|\mathsf{z}\|_1 &\geqslant (\rho\sqrt{2n} + \sqrt{2\rho^2 n + 16k})^2/16 \\ &= \rho^2 n/4 + k + \rho\sqrt{4\rho^2 n + 32kn}/8 \\ &\geqslant k + \rho\sqrt{2kn}/2 + \rho^2 n/4 \\ &\geqslant k + \Omega(\sqrt{kn} + n). \end{aligned}$$ □

Suppose we set the parameter $k = qn$ for $q > 16$ in Example 5.4.3. If we apply Theorem 5.1.3 to the vectors $\mathsf{v}_1, \ldots, \mathsf{v}_m \in \mathbb{R}^{n-1}$ and $\mathsf{x} \in [0,1]^m$ defined in Example 5.4.3 with $\varepsilon = \sqrt{n/k} < 1/4$, then there exists a $\mathsf{z} \in \{0,1\}^m$ such that

$$\sum_{1 \leqslant i < j \leqslant n} \mathsf{z}(ij) \cdot \mathsf{v}_{ij} \mathsf{v}_{ij}^\top \succcurlyeq \mathsf{I}_{n-1} \qquad \text{and} \qquad \langle \mathsf{c}, \mathsf{z} \rangle \leqslant (1 + 6\varepsilon)\langle \mathsf{c}, \mathsf{x} \rangle + \frac{15n \left\| \mathsf{c} \right\|_\infty}{\varepsilon} = k + O(\sqrt{kn}),$$

where the last equality uses $\langle \mathsf{c}, \mathsf{x} \rangle = k$. Note that if the additive error term $O(\varepsilon\langle \mathsf{c}, \mathsf{x} \rangle + \frac{n\|\mathsf{c}\|_\infty}{\varepsilon})$ has a better dependency on ε, then we can set ε accordingly such that the cost upper bound will contradict with the lower bound in Lemma 5.4.5. For example, if Theorem 5.1.3 were improved to $\langle \mathsf{c}, \mathsf{z} \rangle \leqslant (1+6\varepsilon)\langle \mathsf{c}, \mathsf{x} \rangle + \frac{15n\|\mathsf{c}\|_\infty}{\sqrt{\varepsilon}}$, then we could set $\varepsilon = (\frac{n}{k})^{\frac{2}{3}}$ which would imply that $\langle \mathsf{c}, \mathsf{z} \rangle \leqslant k + O(k^{\frac{1}{3}} n^{\frac{2}{3}})$, contradicting with the lower bound $\langle \mathsf{c}, \mathsf{z} \rangle \geqslant k + \Omega(\sqrt{kn})$ when k is large enough. This shows Theorem 5.1.3 is tight up to a constant factor in the additive error term $\frac{n\|\mathsf{c}\|_\infty}{\varepsilon}$.

Chapter 6

Applications of Spectral Rounding to Graph Problems

In this chapter, we show the applications of spectral rounding techniques to various graph problems. We first show that spectral rounding significantly extends the scope of the well-studied survivable network design in Section 6.1. Then, we apply spectral rounding to some spectral network design problems in Section 6.2. Finally, we show applications of spectral rounding to additive spectral sparsification in Section 6.3.

6.1 Generalized Survivable Network Design

In this section, we show that the spectral rounding results provide a new approach for the survivable network design problem. The main advantage of this approach is that it significantly extends the scope of useful properties that can be incorporated into survivable network design.

6.1.1 Previous Work on Network Design and Our Main Results

In network design, we are given a graph $G = (V, E)$ where each edge has a cost $\mathsf{c}(e)$, and the objective is to find a minimum cost subgraph that satisfies certain requirements. In survivable network design [71, 79], the requirements are pairwise edge-connectivities, that every pair of vertices u, v should have at least f_{uv} edge-disjoint paths for $u, v \in V$. This captures several classical problems as special cases, including minimum Steiner tree [32], minimum Steiner forest [1, 73], and minimum k-edge-connected subgraph [67]. Jain introduced the iterative rounding method for linear programming to design a 2-approximation algorithm for the survivable network design problem [79]. His proof exploits the nice structures of the connectivity constraints to show that there is always a variable $\mathsf{x}(e)$ with value at least $\frac{1}{2}$ in any extreme point solution to the linear program. His work leads to many subsequent developments in network design [62, 43, 66, 67, 42], and the iterative rounding algorithm is still the only known constant factor approximation algorithm for the survivable network design problem.

Motivated by the need of more realistic models for the design of practical networks, researchers study generalizations of survivable network design problems where we can incorporate additional useful constraints. One well-studied problem is the degree-constrained survivable network design problem, where there is a degree upper bound d_v on each vertex v to control its workload. There is a long line of work on this problem [122, 125, 72, 93, 56, 64, 96] and the iterative rounding method has been extended to incorporate degree constraints into survivable network design successfully. In the general setting [93, 106, 96], there is a polynomial time algorithm to find a subgraph that violates the cost and the degree constraints by a multiplicative factor of at most 2. For interesting special cases such as finding a spanning tree [72, 130] or a Steiner tree [95, 96], there is a polynomial time algorithm that returns a solution that violates the degree constraint by an additive constant.

More generally, one can consider to add linear packing constraints and linear covering constraints into survivable network design [25, 19, 119, 105], but not as much is known about how to approximately satisfy these constraints simultaneously especially when the linear constraints are unstructured.

Another natural constraint is to control the shortest path distance between pairs of vertices, but unfortunately this is proved to be computationally hard [53] to incorporate into network design.

In [38], together with our coauthors, we propose to incorporate the effective resistance metric (see Section 2.4 for a definition) into network design, as an interpolation of shortest path distance and edge-connectivity between vertices. Incorporating effective resistances can also allow one to control some natural quantities about random walks on the resulting subgraph, such as the commute time between vertices [39] and the cover time [112, 52]. We note that effective resistances have interesting connections to many other graph problems, including spectral sparsification [132], maximum flow computation [44], asymmetric traveling salesman problem [9], and random spanning tree generation [115, 128]. We believe that it is a useful property to be incorporated into network design.

There are many other natural spectral constraints that could help in designing better networks, including total effective resistances [70], algebraic connectivity (and graph expansion) [69], and the mixing time of random walks [30]. These constraints are also well-motivated and were studied individually before (without taking other constraints together into consideration, e.g., connectivity requirements), but not much is known about approximation algorithms with nontrivial approximation guarantees for these constraints (see Section 6.2).

Convex Relaxation for Generalized Survivable Network Design

It would be ideal if a network designer can control all of these properties simultaneously to design a good network that suits their need. We can write a convex programming relaxation for this general network design problem incorporating all these constraints.

In the following, the input graph is $G = (V, E)$ with $|V| = n$ and $|E| = m$. The fractional solution is $x \in \mathbb{R}^m$ where the intended solution is to set $x(e) = 1$ if we choose edge e and $\mathsf{x}(e) = 0$ otherwise. The convex program can be written as follows.

$$
\begin{aligned}
&\min_{\mathsf{x}} \langle \mathsf{c}, \mathsf{x} \rangle \\
&\qquad \mathsf{x}(\delta(S)) \geqslant f(S) && \forall S \subseteq V && \text{(connectivity constraints)} \\
&\qquad \mathsf{x}(\delta(v)) \leqslant d_v && \forall v \in V && \text{(degree constraints)} \\
&\qquad \mathsf{A}\mathsf{x} \leqslant \mathsf{a} && \mathsf{A} \in \mathbb{R}_+^{p \times m}, \mathsf{a} \in \mathbb{R}_+^{p} && \text{(linear packing constraints)} \\
&\qquad \mathsf{B}\mathsf{x} \geqslant \mathsf{b} && \mathsf{B} \in \mathbb{R}_+^{q \times m}, \mathsf{b} \in \mathbb{R}_+^{q} && \text{(linear covering constraints)} \\
&\qquad \mathrm{Reff}_{uv}(\mathsf{x}) \leqslant r_{uv} && \forall u, v \in V && \text{(effective resistance constraints)} \\
&\qquad \mathsf{L}_{\mathsf{x}} \succcurlyeq \mathsf{M} && \mathsf{M} \succcurlyeq 0 && \text{(spectral constraints)} \\
&\qquad \lambda_2(\mathsf{L}_{\mathsf{x}}) \geqslant \lambda && && \text{(algebraic connectivity constraint)} \\
&\qquad 0 \leqslant \mathsf{x}(e) \leqslant 1 && \forall e \in E && \text{(capacity constraints)}
\end{aligned}
\tag{CP}
$$

Let us explain the constraints one by one. For the connectivity constraints, we have a connectivity requirement f_{uv} that there are at least f_{uv} edge-disjoint paths between every pair u, v of vertices. For each subset $S \subseteq V$, we let $f(S) := \max_{u,v:u \in S, v \notin S} f_{uv}$ and write a constraint that at least $f(S)$ edges in $\delta(S)$ should be chosen, where $\mathsf{x}(\delta(S))$ denotes $\sum_{e \in \delta(S)} \mathsf{x}(e)$. By Menger's theorem, if an integral solution satisfies all these constraints, then all the connectivity requirements are satisfied. For the degree constraints, each vertex has a degree upper bound d_v and we write a constraint that at most d_v edges in $\delta(v)$ can be chosen, where $\mathsf{x}(\delta(v)) := \sum_{e \in \delta(v)} \mathsf{x}(e)$. For the linear packing and covering constraints, all the entries in $\mathsf{A}, \mathsf{B}, \mathsf{a}, \mathsf{b}$ are nonnegative, and we assume that A, B have at most a polynomial number of rows in n, m. For effective resistance constraints, we have an upper bound r_{uv} on the effective resistance between every pair $u, v \in V$. As in Section 2.4, we write $\mathrm{Reff}_{uv}(\mathsf{x})$ as the effective resistance between u and v in the fractional solution x where each edge e has conductance $\mathsf{x}(e)$. In the spectral and the algebraic connectivity constraints, we write $\mathsf{L}_{\mathsf{x}} := \sum_{e \in E} \mathsf{x}(e) \cdot \mathsf{L}_e$ as the Laplacian matrix of the fractional solution x where L_e is the Laplacian matrix of an edge as defined in Section 2.3. In the spectral constraint, we require that $\mathsf{L}_{\mathsf{x}} \succcurlyeq \mathsf{M}$ for a positive semidefinite matrix M. One could have polynomially many constraints of this form (just as linear packing and covering constraints), but we only write one for simplicity. In the algebraic connectivity constraint, we require the second smallest eigenvalue of the Laplacian matrix of the solution is at least λ, which is related to the graph expansion of the fractional solution as described in Section 2.3.

This convex program can be solved by the ellipsoid method in polynomial time in n and m. There are exponentially many connectivity constraints but we can use a max-flow min-cut algorithm as a polynomial time separation oracle for these constraints (see, e.g., [79]). Other linear constraints can easily be checked efficiently, as we assume there are only polynomially many of them. Next we consider the non-linear constraints. For the effective resistance constraints, it is known (see Lemma 2.4.5) that $\mathrm{Reff}_{uv}(\mathsf{x})$ is a convex function in x. For the algebraic connectivity constraint, it is known (see Lemma 2.3.1) that λ_2 is a concave function in x. For the spectral constraint, the feasible set is a positive semidefinite cone and is convex in x. So the feasible set for these non-linear constraints form a convex set. Also, these non-linear constraints can all be checked in polynomial time using standard numerical computations. Therefore, we can use the ellipsoid algorithm to find an ε-approximate solution to this convex program in polynomial time in n and m with dependency on ε being $\log(\frac{1}{\varepsilon})$.

Statements of Main Results

Our first result for network design is the following approximation algorithm for this general problem. We remark that the degree constraints are not handled in the following result.

Theorem 6.1.1 (Informal, see Theorem 6.1.5 for a formal statement)**.** *Suppose we are given an optimal solution x to the convex program* (CP)*. There is a polynomial time randomized algorithm to return an integral solution z to* (CP) *that simultaneously satisfies all the connectivity constraints, the effective resistance constraints, the spectral constraints, the algebraic connectivity constraint and the capacity constraints exactly with high probability. The objective value of the integral solution z is*

$$\langle \mathsf{c}, \mathsf{z} \rangle \leqslant (1 + O(\varepsilon)) \cdot \mathsf{cp} + O\left(\frac{n \left\| \mathsf{c} \right\|_\infty}{\varepsilon} \right)$$

with high probability, where n is the number of vertices in the graph and $\|\mathsf{c}\|_\infty$ is the maximum cost of an edge. Furthermore, the linear packing constraints and the linear covering constraints are satisfied approximately with high probability.

We remark that the one-sided spectral rounding theorem in [6] (i.e. Theorem 5.1.1) can be modified to prove similar but more restrictive results for network design when the objective function c is the all-one vector and there are no linear covering and packing constraints. This already extends the scope of unweighted network design significantly, but this connection was not made before. For network design, it is desirable to have different costs on edges, and these weighted problems are usually more difficult to solve than the unweighted problems (e.g., minimum k-edge-connected subgraphs [67] vs [79], minimum bounded degree spanning trees [65] vs [72], etc.).

Theorem 6.1.1 provides a $(1 + O(\varepsilon))$-approximation algorithm if $\mathsf{cp} \gtrsim \frac{n\|\mathsf{c}\|_\infty}{\varepsilon^2}$, and a constant factor approximation algorithm if $\mathsf{cp} \gtrsim n\|\mathsf{c}\|_\infty$. We remark that, for survivable network design, the $(1 + O(\varepsilon))$-approximation algorithm does not improve on the 2-approximation algorithm of Jain's result, as Jain's algorithm always returns a solution with cost at most $\mathsf{cp} + 2n\|\mathsf{c}\|_\infty$.

The main advantage of the spectral approach is that it significantly extends the scope of useful properties that can be incorporated into network design, while previously there were no known non-trivial approximation algorithms even for some individual constraints. We demonstrate the use of Theorem 6.1.1 with one concrete setting.

Example 6.1.2. *Suppose the connectivity requirement satisfies $f_{uv} \geqslant k$ for all $u, v \in V$ (e.g., to find a k-edge-connected subgraph). Assume the cost $\mathsf{c}(e)$ of each edge e satisfies $1 \leqslant \mathsf{c}(e) \leqslant O(k)$. Then Theorem 6.1.1 provides a constant factor approximation algorithm for this survivable network design problem. To our knowledge, the only known constant factor approximation algorithm even restricted to this special case is Jain's iterative rounding algorithm. The algorithm in Theorem 6.1.1 provides a completely different spectral algorithm to achieve constant factor approximation in this special case.*

Furthermore, the constant factor approximation algorithm can be achieved while incorporating additional effective resistance constraints (e.g., to upper bound commute times between pairs of vertices), spectral constraints (e.g., to dominate another graph/topology in terms of the number of edges in cuts), algebraic connectivity constraint (e.g., to lower bound graph expansion). Also, additional linear packing and covering constraints can be *satisfied approximately, even when they are unstructured. See Section 6.1.3 for an in-depth discussion.*

Recently, Bansal [16] designed a rounding technique that achieves the guarantees by iterative rounding and randomized rounding simultaneously, and he showed various interesting applications of his techniques. However, he left it as an open question whether there is an $O(1)$-approximation algorithm for survivable network design while satisfying some concentration property of the output. Theorem 6.1.1 provides some progress towards his question (e.g., in the setting in Example 6.1.2), as the guarantees on the linear packing and linear covering constraints satisfy some concentration property as shown in Theorem 6.1.5. We defer to Section 6.1.5 for details.

Our second result for network design is a strong upper bound on the integrality gap of the convex program that incorporates degree constraints as well, assuming the fractional solution x satisfies some additional properties.

Theorem 6.1.3 (Informal)**.** *Suppose we are given a solution x to the convex program* (CP)*. Assume that* $\mathrm{Reff}_{\mathsf{x}}(u,v) \leqslant \varepsilon^2$ *for every* $uv \in E$ *and* $\|\mathsf{c}\|_\infty \leqslant \varepsilon^2\langle \mathsf{c},\mathsf{x}\rangle$ *for some* $\varepsilon \in [0,1]$*. Then, there exists an integral solution z that approximately satisfies all the connectivity constraints, degree constraints, effective resistance constraints, spectral constraints, algebraic connectivity constraints, and capacity constraints with* $\langle \mathsf{c},\mathsf{z}\rangle \leqslant (1+O(\varepsilon))\langle \mathsf{c},\mathsf{x}\rangle$*.*

We remark that Theorem 6.1.3 does not provide a polynomial time algorithm to find such an integral solution, as it is proved using the non-constructive results in discrepancy theory. Also, we note that Theorem 6.1.3 does not handle linear covering and packing constraints. The assumption $\mathrm{Reff}_{\mathsf{x}}(u,v) \leqslant \varepsilon^2$ for every $uv \in E$ may not be satisfied in applications, and we will explain in Section 6.1.4 when it will be satisfied and show that it is not too restrictive.

The organization of remaining of this section is as follows. We begin by explaining how the spectral rounding results can be used to find a solution for this general survivable network design problem in Section 6.1.2. Then we will see the implications of Theorem 5.1.3 to network design in Section 6.1.3 and of Theorem 5.1.4 to network design in Section 6.1.4.

Finally, we discuss how these new results make some progress towards Bansal's question [16] of designing an approximation algorithm for survivable network design with concentration property in Section 6.1.5.

6.1.2 Implications of Spectral Rounding

Suppose we are given an optimal solution x to the convex programming relaxation (CP). To design approximation algorithms, the task is to round this fractional solution x into an integral solution z so that z satisfies all the constraints and $\langle c, z\rangle$ is close to $\langle c, x\rangle$. There are many different types of constraints and it seems difficult to handle them simultaneously. In the spectral approach, the main observation is that if we can find an integral solution z such that $\sum_{e\in E} z(e)L_e \approx \sum_{e\in E} x(e)L_e$ and $\langle c, x\rangle \approx \langle c, z\rangle$, then all the constraints can be (approximately) satisfied simultaneously. We state this observation in the following lemma.

Lemma 6.1.4. *Let $x \in \mathbb{R}^m_+$ be a feasible solution to* (CP). *For $\varepsilon \in [0, \frac{1}{2}]$, any $z \in \mathbb{Z}^m_+$ satisfies*

$$\sum_{e\in E} z(e)\cdot L_e \succcurlyeq (1-\varepsilon)\sum_{e\in E} x(e)\cdot L_e \quad\Longrightarrow\quad \begin{cases} z(\delta(S)) \geqslant (1-\varepsilon)f(S) \text{ for all } S\subseteq V \\ \mathrm{Reff}_z(u,v) \leqslant (1+2\varepsilon)r_{uv} \text{ for all } u,v\in V \\ L_z \succcurlyeq (1-\varepsilon)\cdot M, \\ \lambda_2(L_z) \geqslant (1-\varepsilon)\lambda. \end{cases}$$

For $\varepsilon \in [0,1]$, any $z \in \mathbb{Z}^m_+$ satisfies

$$\sum_{e\in E} z(e)\cdot L_e \preccurlyeq (1+\varepsilon)\sum_{e\in E} x(e)\cdot L_e \quad\Longrightarrow\quad z(\delta(v)) \leqslant (1+\varepsilon)d_v \text{ for all } v\in V.$$

Proof. Let $L_x := \sum_{e\in E} x(e)L_e$ and $L_z := \sum_{e\in E} z(e)L_e$. We start with the connectivity constraints. For any $S\subseteq V$, let $\chi_S \in \mathbb{R}^n$ be the characteristic vector of S with $\chi_S(i) = 1$ if $i\in S$ and zero otherwise. It is well-known that

$$\chi_S^\top L_z \chi_S = \chi_S^\top\left(\sum_{e\in E} z(e)L_e\right)\chi_S = \sum_{e\in E} z(e)\chi_S^\top L_e\chi_S = \sum_{e\in\delta(S)} z(e) = z(\delta(S))$$

and similarly $\chi_S^\top L_x \chi_S = x(\delta(S))$. So, if $L_z \succcurlyeq (1-\varepsilon)L_x$, then for all $S \subseteq V$ we have

$$z(\delta(S)) = \chi_S^\top L_z \chi_S \geqslant (1-\varepsilon)\chi_S^\top L_x \chi_S = (1-\varepsilon)x(\delta(S)) \geqslant (1-\varepsilon)f(S).$$

For the effective resistance constraints, since $L_z \succcurlyeq (1-\varepsilon)L_x$, it implies that $L_z^\dagger \preccurlyeq (1-\varepsilon)^{-1}L_x^\dagger \preccurlyeq (1+2\varepsilon)L_x^\dagger$ for $\varepsilon \in [0, \frac{1}{2}]$, and thus

$$\mathrm{Reff}_z(u,v) = b_{uv}^\top L_z^\dagger b_{uv} \leqslant (1+2\varepsilon) b_{uv}^\top L_x^\dagger b_{uv} = (1+2\varepsilon)\,\mathrm{Reff}_x(u,v) \leqslant (1+2\varepsilon) r_{uv}.$$

The statements about the spectral lower bound and the algebraic connectivity constraint follows directly from the assumption that $L_z \succcurlyeq (1-\varepsilon)L_x$. Finally, for the degree constraints, suppose we are given $L_z \preccurlyeq (1+\varepsilon)L_x$, then it follows that

$$z(\delta(v)) = \chi_v^\top L_z \chi_v \leqslant (1+\varepsilon)\chi_v^\top L_x \chi_v = (1+\varepsilon)x(\delta(v)) \leqslant (1+\varepsilon)d_v. \qquad \square$$

Lemma 6.1.4 says that if z satisfies the spectral lower bound $L_z \succcurlyeq L_x$, then the solution z will simultaneously satisfy all connectivity constraints, effective resistance constraints, spectral constraints, and the algebraic connectivity constraint exactly. Moreover, if z also satisfies the spectral upper bound approximately, then the solution z will approximately satisfy all degree constraints as well.

6.1.3 Applications of One-Sided Spectral Rounding

We apply Theorem 5.1.3 to design approximation algorithms for network design problems that significantly extend the scope of existing techniques.

$$\begin{aligned}
\mathsf{cp} := \min_x \ & \langle c, x\rangle \\
& x(\delta(S)) \geqslant f(S) && \forall S \subseteq V && \text{(connectivity constraints)} \\
& Ax \leqslant a && A \in \mathbb{R}_+^{p\times m}, a \in \mathbb{R}_+^p && \text{(linear packing constraints)} \\
& Bx \geqslant b && B \in \mathbb{R}_+^{q\times m}, b \in \mathbb{R}_+^q && \text{(linear covering constraints)} \\
& \mathrm{Reff}_{uv}(x) \leqslant r_{uv} && \forall u,v \in V && \text{(effective resistance constraints)} \qquad \text{(CP1)} \\
& L_x \succcurlyeq M && M \succcurlyeq 0 && \text{(spectral constraint)} \\
& \lambda_2(L_x) \geqslant \lambda && && \text{(algebraic connectivity constraint)} \\
& 0 \leqslant x(e) \leqslant 1 && \forall e \in E && \text{(capacity constraints)}
\end{aligned}$$

In network design, a zero-one solution corresponds to a subset of edges where each edge is used at most once (satisfying the capacity constraints). The following theorem is a consequence of Theorem 5.1.3.

Theorem 6.1.5. *Suppose we are given an optimal solution x to the convex program (CP1). For any $\varepsilon \in (0, \frac{1}{4})$, there is a polynomial time randomized algorithm to return a zero-one solution $\mathsf{z} \in \{0,1\}^m$ to (CP1) satisfying all the constraints exactly with probability at least $1 - \exp(-\Omega(n))$ except for the linear constraints. The solution z has objective value*

$$\langle \mathsf{c}, \mathsf{z} \rangle \leqslant (1 + 6\varepsilon)\mathsf{cp} + \frac{15n \left\| \mathsf{c} \right\|_\infty}{\varepsilon}$$

with probability at least $1 - \exp(-\Omega(n))$, and satisfies

$$\langle \mathsf{A}(i,:), \mathsf{z} \rangle \leqslant (1 + 6\varepsilon)\mathsf{a}(i) + \frac{15n \left\| \mathsf{A}(i,:) \right\|_\infty}{\varepsilon}$$

with probability at least $1 - \exp(-\Omega(n))$ for each linear packing constraint, where $\mathsf{A}(i,:)$ is the i-th row of A, and satisfies

$$\langle \mathsf{B}(j,:), \mathsf{z} \rangle \geqslant \mathsf{b}(j) - \delta n \left\| \mathsf{B}(j,:) \right\|_\infty,$$

with probability at least $1 - \exp(-\min\{\varepsilon\delta, \varepsilon\delta^2\} \cdot \Omega(n))$ for any $\delta > 0$ for each linear covering constraint, where $\mathsf{B}(j,:)$ is the j-th row of B.

Proof. We assume without loss of generality that the graph $G = (V, E; \mathsf{x})$ formed by the support of the fractional solution x is connected, and so L_x has rank $n - 1$. Then, we apply a transformation similar to (2.12).

Let $\mathsf{L}_\mathsf{x} = \sum_{i=2}^n \lambda_i \cdot \mathsf{u}_i \mathsf{u}_i^\top = \mathsf{U} \Lambda \mathsf{U}^\top$ be the eigendecomposition of L_x, where $\Lambda = \operatorname{diag}(\lambda_2, \ldots, \lambda_n)$ is a diagonal matrix that contains the $n-1$ nonzero eigenvalues of L_x, and the columns of $\mathsf{U} \in \mathbb{R}^{n \times (n-1)}$ are the corresponding eigenvectors. We define

$$\mathsf{v}_e := \mathsf{U}^\top \mathsf{L}_\mathsf{x}^{\frac{\dagger}{2}} \mathsf{b}_e, \qquad \text{for all } e \in E. \tag{6.1}$$

Note that each $v_e \in \mathbb{R}^{n-1}$. Then

$$\sum_{e \in E} x(e) \cdot v_e v_e^\top = U^\top L_x^{\frac{\dagger}{2}} \left(\sum_{e \in E} x(e) b_e b_e^\top \right) L_x^{\frac{\dagger}{2}} U = U^\top L_x^{\frac{\dagger}{2}} L_x L_x^{\frac{\dagger}{2}} U = I_{n-1}.$$

For any $\varepsilon \in (0, \frac{1}{4})$, we apply Theorem 5.1.3 to $x \in [0,1]^E$ and $\{v_e\}_{e \in E}$ to find a zero-one solution $z \in \{0,1\}^E$ such that $\sum_{e \in E} z(e) \cdot v_e v_e^\top \succcurlyeq I_{n-1}$ with probability at least $1 - \exp(-\Omega(n))$.

Since the columns of U are the eigenvectors (u_i's) of L_x which span the $(n-1)$-dimensional subspace orthogonal to $\mathit{1}_n$, it holds that

$$UU^\top = I_n - \frac{1}{n} \mathit{1}\mathit{1}^\top \quad \Longrightarrow \quad \begin{cases} L_x^{\frac{1}{2}} UU^\top L_x^{\frac{1}{2}} = L_x, \\ L_x^{\frac{1}{2}} UU^\top L_x^{\frac{\dagger}{2}} = I_n. \end{cases}$$

Thus, $\sum_{e \in E} z(e) \cdot v_e v_e^\top \succcurlyeq I_{n-1}$ is equivalent to $U^\top L_x^{\frac{\dagger}{2}} \left(\sum_{e \in E} z(e) \cdot b_e b_e^\top \right) L_x^{\frac{\dagger}{2}} U \succcurlyeq I_{n-1}$, which further implies

$$L_x^{\frac{1}{2}} UU^\top L_x^{\frac{\dagger}{2}} \left(\sum_{e \in E} z(e) \cdot b_e b_e^\top \right) L_x^{\frac{\dagger}{2}} UU^\top L_x^{\frac{1}{2}} \succcurlyeq L_x^{\frac{1}{2}} UU^\top L_x^{\frac{1}{2}} \Longrightarrow \sum_{e \in E} z(e) \cdot b_e b_e^\top \succcurlyeq L_x. \tag{6.2}$$

Therefore, the zero-one solution z satisfies all the constraints in (CP1) except for the linear constraints by Lemma 6.1.4.

Theorem 5.1.3 also guarantees that with probability at least $1 - \exp(-\Omega(n))$ the objective value of z is at most

$$\langle c, z \rangle \leqslant (1 + 6\varepsilon) \langle c, x \rangle + \frac{15n \|c\|_\infty}{\varepsilon}.$$

The guarantees for the linear packing constraints follow the same way as for the objective function, and the guarantees for the linear covering constraints follow from the lower bound part of Theorem 5.1.3. □

We demonstrate the use of Theorem 6.1.5 in some concrete settings. The first example shows that Theorem 6.1.5 provides a spectral alternative to Jain's iterative rounding algorithm to achieve $O(1)$-approximation for a fairly general subclass of the survivable network design problem.

Example 6.1.6. *Theorem 6.1.5 is a constant factor approximation algorithm as long as* $n\left\|\mathsf{c}\right\|_\infty = O(\mathsf{cp})$. *Suppose that in our network design problem the average degree is at least* d_{avg} *and the costs on edges are positive integers with* $\left\|\mathsf{c}\right\|_\infty = O(d_{\text{avg}})$ *(e.g., in the minimum* k*-edge-connected subgraph problem every vertex has degree at least* k *and* $1 \leqslant \mathsf{c}(e) \leqslant O(k)$ *for* $e \in E$*, or the solution requires a connected subgraph and* $1 \leqslant c_e \leqslant O(1)$ *for* $e \in E$*, etc). Then* $\mathsf{cp} \geqslant \Omega(d_{\text{avg}} n) \geqslant \Omega(\left\|\mathsf{c}\right\|_\infty n)$ *and Theorem 6.1.5 provides a constant factor approximation algorithm.*

The additive error term $n\left\|\mathsf{c}\right\|_\infty$ is the reason that we could not achieve constant factor approximation in general, but this term is unavoidable in the one-sided spectral rounding setting when we need to satisfy the spectral lower bound exactly. See Section 5.4 for examples showing the limitations. Heuristically, we can compute cp and if $n\left\|\mathsf{c}\right\|_\infty = O(\mathsf{cp})$ then we know Theorem 6.1.5 will provide good approximate solutions.

The second example shows that Theorem 6.1.5 returns good approximate solution to survivable network design while incorporating many other constraints simultaneously.

Example 6.1.7. *Suppose the connectivity requirement is to find a* k*-edge-connected subgraph, or more generally* $f_{uv} \geqslant k$ *for all* $u, v \in V$. *Assume the cost* $\mathsf{c}(e)$ *of each edge* e *is at least one. Then* $\mathsf{cp} \geqslant \Omega(kn)$.

When the cost function satisfies $\left\|\mathsf{c}\right\|_\infty = O(k)$, *then Theorem 6.1.5 implies that there is a polynomial time randomized algorithm to return a simple* k*-edge-connected subgraph satisfying all the constraints in* (CP1) *except for the linear constraints (with some non-trivial guarantees), and the cost of the subgraph is at most a constant factor of the optimal value.*

When the cost function satisfies $\left\|\mathsf{c}\right\|_\infty = O(1)$, *then Theorem 6.1.5 implies that there is a polynomial time randomized algorithm to return a* k*-edge-connected subgraph satisfying* *all the constraints in* (CP1) *except for the linear constraints, and the cost of the subgraph is at most* $1 + O\big(\frac{1}{\sqrt{k}}\big)$ *factor of the optimal value by setting* $\varepsilon = \Theta\big(\frac{1}{\sqrt{k}}\big)$.

The third example shows when the linear packing and covering constraints can be satisfied up to a multiplicative constant factor. See also Section 6.1.5 for a related question asked by Bansal [16].

Example 6.1.8. *For linear covering constraints, suppose they are of the form* $\sum_{e \in F} \mathsf{x}(e) \geqslant b_j$ *for some subset* $F \subseteq E$ *where* $b_j \geqslant n$*, then the returned solution* z *will almost satisfy this constraint as* $\sum_{e \in F} \mathsf{z}(e) \geqslant \mathsf{b}(j) - \delta n \left\| \mathsf{B}(j,:) \right\|_\infty \geqslant (1-\delta)\mathsf{b}(j)$ *for some* $\delta > 0$*. So, these unweighted covering constraints with large right hand side can be incorporated into survivable network design, even though they can be unstructured. By a similar argument, any unweighted packing constraints with large right hand side will be only violated by at most a multiplicative constant factor with high probability. It was not known that Jain's iterative rounding can be adapted to incorporate these linear covering and packing constraints.*

We will present more applications of Theorem 6.1.5 in Section 6.2, where they can be used to design approximation algorithms for network design problems with spectral requirements. These problems were studied in the literature before but not much is known about approximation algorithms with performance guarantees.

6.1.4 Applications of Two-Sided Spectral Rounding

If we can achieve two-sided spectral rounding in network design, then we can also approximately satisfy the degree constraints by Lemma 6.1.4. However, to apply Theorem 5.1.4, we need to satisfy the assumption that the vector lengths are small. It is known that the vector lengths in the spectral rounding setting corresponds to the effective resistance of the edges in the fractional solution x. In the following, we describe when two-sided spectral rounding can be applied, and discuss what are the implications for network design.

$$
\begin{aligned}
\mathsf{cp} := \min_{x}\ & \langle c, x\rangle \\
& x(\delta(S)) \geqslant f(S) && \forall S \subseteq V && \text{(connectivity constraints)} \\
& x(\delta(v)) \leqslant d_v && \forall v \in V && \text{(degree constraints)} \\
& \mathrm{Reff}_{uv}(x) \leqslant r_{uv} && \forall u, v \in V && \text{(effective resistance constraints)} \\
& L_x \succcurlyeq M && M \succcurlyeq 0 && \text{(spectral lower bound)} \\
& \lambda_2(L_x) \geqslant \lambda && && \text{(algebraic connectivity constraint)} \\
& 0 \leqslant x(e) \leqslant 1 && \forall e \in E && \text{(capacity constraints)}
\end{aligned}
\tag{CP2}
$$

Theorem 6.1.9. *Suppose we are given an optimal solution x to the convex program (CP2). For any $\varepsilon \in [0,1]$, if $\mathrm{Reff}_x(u,v) \leqslant \varepsilon^2$ for every $uv \in E$ and $\|c\|_\infty \leqslant \varepsilon^2 \langle c, x\rangle$, then there exists a zero-one solution $z \in \{0,1\}^m$*

$$(1 - O(\varepsilon))L_x \preccurlyeq L_z \preccurlyeq (1 + O(\varepsilon))L_x \quad \text{and} \quad (1 - O(\varepsilon))\langle c, x\rangle \leqslant \langle c, z\rangle \leqslant (1 + O(\varepsilon))\langle c, x\rangle$$

This implies that all the constraints of (CP2) will be approximately satisfied by z (e.g., $z(\delta(S)) \geqslant (1 - O(\varepsilon))f(S)$ for all $S \subseteq V$ and $z(\delta(v)) \leqslant (1 + O(\varepsilon))d_v$ for all $v \in V$) and the objective value of z is at most $(1 + O(\varepsilon))\mathsf{cp}$.

Proof. We apply the same transformation as in (6.1) to obtain vector v_e for each edge $e \in E$ such that $\sum_{e \in E} x(e) \cdot v_e v_e^\top = I_{n-1}$ as in the proof of Theorem 6.1.5. Using the assumption that $\mathrm{Reff}_x(i,j) \leqslant \varepsilon^2$ for every edge $ij \in E$, it follows that

$$\|v_{ij}\|^2 = b_{ij}^\top L_x^{\frac{\dagger}{2}} \left(I_n - \frac{1}{n} \mathit{11}^\top\right) L_x^{\frac{\dagger}{2}} b_{ij} = b_{ij}^\top L_x^\dagger b_{ij} = \mathrm{Reff}_x(i,j) \leqslant \varepsilon^2 \text{ for all } ij \in E,$$

and thus the assumption in Theorem 5.1.4 is satisfied. We can then apply Theorem 5.1.4 on $\{v_e\}_e$ and c to conclude that there exists $z \in \{0,1\}^E$ such that

$$(1 - O(\varepsilon))I_{n-1} \preccurlyeq \sum_{e \in E} z(e) \cdot v_e v_e^\top \preccurlyeq (1 + O(\varepsilon))I_{n-1} \qquad \text{and} \qquad \langle c, z\rangle \leqslant (1 + O(\varepsilon))\langle c, x\rangle.$$

By the definition of v_e's, this implies that

$$(1 - O(\varepsilon))L_x \preccurlyeq L_z = \sum_{e \in E} z(e) \cdot b_e b_e^\top \preccurlyeq (1 + O(\varepsilon))L_x.$$

By Lemma 6.1.4, the zero-one solution z satisfies all the constraints of (CP2) approximately. □

In the following, we compare Theorem 6.1.9 to Theorem 6.1.5.

1. **Approximation guarantees:** When Theorem 6.1.9 applies, it can handle degree constraints as well and basically preserves all properties of the fractional solution (e.g., upper bound and lower bound on every cut). It also gives strong approximation guarantee for the objective value, getting arbitrarily close to the optimal value. However, the constraints are only approximately satisfied, while in Theorem 6.1.5 they are exactly satisfied. Theorem 6.1.9 can only handle one linear constraint, which is used for the objective function, while Theorem 6.1.5 can handle many linear constraints simultaneously with an additive error term.

2. **Assumptions:** Theorem 6.1.5 apply without any assumptions, but Theorem 6.1.9 only applies when $\mathrm{Reff}_x(u,v) \leqslant \varepsilon^2$ for all $uv \in E$ and $\|c\|_\infty \leqslant \varepsilon^2 \langle c, x\rangle$. The assumption about the cost is moderate, as it only requires the maximum cost of an edge is at most ε^2 fraction of the total cost of the solution, which should be satisfied in many applications with small ε. The main restriction is the first assumption about effective resistances, which may not be satisfied in network design applications, and we would like to provide some combinatorial characterizations under which the assumption will hold. Let $\mathrm{Reff}_{\mathrm{diam}} := \max_{u,v} \mathrm{Reff}(u,v)$ be the effective resistance diameter of a graph; note that the maximum is taken over all pairs (not just for edges as required in Theorem 6.1.9). For example, it is known that [39] a d-regular graph with constant expansion has $\mathrm{Reff}_{\mathrm{diam}} \leqslant O(\frac{1}{d})$. So, if the fractional solution x is close to a d-regular expander graph, then Theorem 6.1.9 can be applied with $\varepsilon \geqslant \frac{1}{\sqrt{d}}$. It is proved in [4] that a much milder expansion condition guarantees small effective resistance diameter. For example, in a d-regular graph G, as long as for some $0 < \eta \leqslant \frac{1}{2}$,

$$|\delta(S)| \geqslant \Omega\left((d|S|)^{\frac{1}{2}+\eta}\right) \text{ for all } S \subseteq V \quad \Longrightarrow \quad \mathrm{Reff}_{\mathrm{diam}} \leqslant O\left(\frac{1}{d^{2\eta}}\right).$$

Note that a d-regular graph with constant expansion satisfies the much stronger assumption that $|\delta(S)| \geqslant \Omega(d|S|)$. Informally, the above result only requires $|\delta(S)|$ to be roughly the square root of $d|S|$ to show that the graph has a small effective resistance diameter (e.g., 3-dimensional mesh). So, as long as the fractional solution x is a mild expander as defined in [4], the assumption in Theorem 6.1.9 will be satisfied with small ε. As another example, if the algebraic connectivity $\lambda_2(L_{\mathsf{x}})$ of the fractional solution is at least say $\frac{1}{2\varepsilon^2}$, then we have $\mathrm{Reff}_{\mathrm{diam}} \leqslant \varepsilon^2$ so that Theorem 6.1.9 can be applied. Heuristically, if one could add the constraints that $\mathrm{Reff}_{uv}(\mathsf{x}) \leqslant \varepsilon^2$ for $uv \in E$ so that the convex program (CP2) is still feasible without increasing the objective value too much, then one could then apply Theorem 6.1.9 to bound the integrality gap of the convex program.

3. **Algorithms:** There are polynomial time algorithms to return the solutions guaranteed in Theorem 6.1.5, while the proof of Theorem 6.1.9 is non-constructive. In network design, Theorem 6.1.5 give us approximation algorithms, while Theorem 6.1.9 only gives us integrality gap results for the convex programming relaxation (that there exists a zero-one solution almost satisfying all the constraints with objective value close to the optimal value).

6.1.5 Concentration Property in Survivable Network Design

Recently, Bansal [16] designed a rounding technique that achieves the guarantees by iterative rounding and randomized rounding simultaneously. Suppose there is an iterative rounding algorithm for a problem satisfying some technical assumptions. Bansal's algorithm will satisfy essentially the same guarantees of the iterative rounding algorithm, and simultaneously the following concentration property with $\beta = O(1)$ with respect to linear constraints as if the algorithm does independent randomized rounding.

Definition 6.1.10 (β-concentration)**.** *Let $\beta \geqslant 1$. For a vector valued random variable $X = (X_1, ..., X_m)$, where X_i are possible dependent 0-1 random variables, we say X is β-concentrated around the mean $\mathsf{x} \in \mathbb{R}^m$ where $\mathsf{x}(i) = \mathbb{E}[X_i]$, if for every $\mathsf{a} \in \mathbb{R}^n$ with $M := \max_i |\mathsf{a}(i)|$, $\langle \mathsf{a}, X\rangle$ is well-concentrated and satisfies Bernstein's inequality up to a* *factor of β in the exponent, i.e.*

$$\mathbb{P}\left[\langle a, X\rangle - \langle a, x\rangle \geqslant t\right] \leqslant \exp\left(-\frac{t^2/\beta}{2(\sum_{i=1}^{m} a(i)^2 x(i)(1-x(i)) + Mt/3)}\right).$$

Bansal showed various interesting applications of his techniques, with x being the fractional solution to the linear programming relaxation and X being the zero-one solution output by the approximation algorithm. However, he left it as an open question whether there is an $O(1)$-approximation algorithm for survivable network design (the guarantee achieved by Jain's iterative rounding algorithm) with $O(1)$-concentration property.

Our iterative randomized swapping algorithms satisfy similar but weaker concentration properties. Let $x \in [0,1]^m$ be the fractional solution to the one-sided spectral rounding problem. The algorithm in Theorem 5.1.3 will output a vector-valued random variable $X \in \{0,1\}^m$ such that for any $a \in \mathbb{R}_+^n$ with $M := \max_i a(i)$,

$$\mathbb{E}[\langle a, X\rangle] \leqslant (1+O(\varepsilon))\langle a, x\rangle + O\Big(\frac{nM}{\varepsilon}\Big) \quad \text{and}$$
$$\mathbb{P}[\langle a, X\rangle - \mathbb{E}[\langle a, X\rangle] \geqslant \eta] \leqslant \exp\left[-\Omega\left(\frac{\eta^2}{\sigma^2 + M\eta}\right)\right],$$

where n is the dimension of the problem (i.e. the dimension of the vectors) and $\sigma^2 = O(M(\langle a, x\rangle + nM/\varepsilon))$ is a term related to the variance of the randomized swapping process. In other words, the random variable $\langle a, X\rangle$ is concentrated around the expected value $\mathbb{E}[\langle a, X\rangle]$, but the expected value $\mathbb{E}[\langle a, X\rangle]$ could deviate from $\langle a, x\rangle$ by $O(\varepsilon\langle a, x\rangle + nM/\varepsilon)$ and the concentration property is weaker than the one required in β-concentration, as the upper bound of σ^2 we can obtain is larger than the term $\sum_{i=1}^{m} a(i)^2 x(i)(1-x(i))$ in the β-concentration definition. We note that both Bansal's proof and our proof use Freedman's concentration inequality or its variant. Using Theorem 6.1.5, we made some progress towards Bansal's question.

Corollary 6.1.11. *Let $x \in [0,1]^m$ be an optimal fractional solution to the survivable network design problem (i.e. (CP1) with only connectivity and capacity constraints). Suppose $n\|c\|_\infty = O(\langle c, x\rangle)$. Then there is a randomized polynomial time algorithm to return a solution $z \in \{0,1\}^m$ to the survivable network design problem so that $\langle c, z\rangle \leqslant O(\langle c, x\rangle)$ with* *probability at least* $1-\exp(-\Omega(n))$*. Furthermore, for any* $\mathsf{a} \in \mathbb{R}^m_+$ *and* $\delta \in (0,1)$ *it holds that* $\langle \mathsf{a}, \mathsf{x}\rangle - \delta n \left\|\mathsf{a}\right\|_\infty \leqslant \langle \mathsf{a}, \mathsf{z}\rangle \leqslant O(\langle \mathsf{a}, \mathsf{x}\rangle + n \left\|\mathsf{a}\right\|_\infty)$ *with probability at least* $1-O(\exp(-\Omega(\delta^2 n)))$*.*

We remark that one can add linear constraints a to the convex program in our framework before we apply the rounding, so that we have some control over $\langle \mathsf{a}, \mathsf{x}\rangle$ of the fractional solution x and hence some control over $\langle \mathsf{a}, \mathsf{z}\rangle$ of the zero-one solution z. But it may not be possible to add linear constraints to the relaxation in Bansal's setting, as adding constraints may make the underlying iterative rounding algorithm stops working (e.g., we do not know of an iterative rounding algorithm for the survivable network design problem with additional linear packing or covering constraints). See Example 6.1.8 for a related discussion. Our results suggest that the spectral approach is perhaps more suitable for achieving concentration property for survivable network design.

6.2 Spectral Network Design

There are several previous work on network design problems with spectral requirements. In this section, we will see that these problems are special cases of the general network design problem in Section 6.1, and our results provide improved approximation algorithms for these problems and also generalize these problems to incorporate many additional constraints.

6.2.1 Maximizing Algebraic Connectivity

Ghosh and Boyd [69] study the problem of choosing a subgraph that maximizes the algebraic connectivity (the second smallest eigenvalue of its Laplacian matrix) subject to a cost constraint. The problem is formulated as follows:

$$\begin{aligned} \lambda_{\mathsf{opt}} := \max_{\mathsf{x}\in\mathbb{R}^E} \quad & \lambda_2\left(\sum_{e\in E} \mathsf{x}(e)\cdot \mathsf{b}_e \mathsf{b}_e^\top\right) \\ \text{subject to} \quad & \sum_{e\in E} \mathsf{c}(e)\cdot \mathsf{x}(e) \leqslant C, \\ & \mathsf{x}(e) \in \{0,1\}, \forall e \in E, \end{aligned} \tag{6.3}$$

where $c(e)$ is the cost of edge e for $e \in E$ and C is the given cost budget. As mentioned in [69], the algebraic connectivity is a good measure on the well-connectedness of a graph, as

$$\lambda_2(L_G) \leqslant \min_{S \subseteq V} \frac{n|\delta(S)|}{|S||\bar{S}|} \leqslant 2 \min_{0 \leqslant |S| \leqslant \frac{n}{2}} \frac{|\delta(S)|}{|S|}$$

where the first inequality is proved in [59]. Thus, any graph with large λ_{opt} has no sparse cuts, which also implies that the mixing time of random walks is small.

Ghosh and Boyd show that if the constraint $x(e) \in \{0,1\}$ is relaxed to $x(e) \in [0,1]$, then the relaxation is convex and can be written as a semidefinite program. They proposed a greedy heuristic based on the Fiedler vector for the zero-one cost setting (where $c(e) \in \{0,1\}$ for all e), but they do not provide any approximation guarantee of their heuristic algorithm.

Kolla, Makarychev, Saberi and Teng [86] provide the first algorithm with non-trivial approximation guarantee in the zero-one cost setting. Using subgraph sparsification techniques, they give an algorithm that returns a solution which violates the cost constraint by a factor of at most 8 and having algebraic connectivity at least $\Omega(\lambda_{\mathsf{opt}}^2/\Delta)$ where Δ is the maximum degree of the graph.

We observe that if we project the vectors b_e onto the rank $n-1$ subspace orthogonal to the all-one vector, then the objective function of (6.3) is simply the reciprocal of the objective function of the E-optimal design problem (see Chapter 7). This immediately implies that the result of Allen-Zhu, Li, Singh and Wang [6] (Theorem 5.1.1) can be applied to give a $(1+\varepsilon)$-approximation algorithm for the unweighted problem as long as $C \geqslant 5n/\varepsilon^2$, although this connection was not made before.

The one-sided spectral rounding results in Chapter 5 imply the following approximation result for general non-negative cost functions.

Theorem 6.2.1. *Suppose $C \geqslant \frac{15n\|c\|_\infty}{\varepsilon^2}$ for some $\varepsilon \in (0, \frac{1}{2}]$. There is a polynomial time randomized algorithm which returns a zero-one solution $z \in \{0,1\}^E$ for (6.3) with with probability at least $1 - \exp(-\Omega(n))$ such that*

$$\lambda_2\left(\sum_{e \in E} z(e) \cdot b_e b_e^\top\right) \geqslant (1 - O(\varepsilon))\lambda_{\mathsf{opt}} \quad \text{and} \quad \sum_{e \in E} c(e) \cdot z(e) \leqslant C.$$

Proof. We apply the same transformation as in (6.1) to obtain vector v_e for each edge $e \in E$ such that $\sum_{e \in E} \mathsf{x}(e) \cdot \mathsf{v}_e \mathsf{v}_e^\top = \mathsf{I}_{n-1}$. Then, apply Corollary 5.2.3 with $\gamma = \varepsilon$ on $\{\mathsf{v}_e\}$, x and c to find a $\mathsf{z} \in \{0,1\}^E$ such that

$$\sum_{e \in E} \mathsf{z}(e) \cdot \mathsf{v}_e \mathsf{v}_e^\top \succcurlyeq (1 - 4\varepsilon)\mathsf{I}_{n-1} \qquad \text{and} \qquad \langle \mathsf{c}, \mathsf{z} \rangle \leqslant C$$

with probability at least $1 - \exp(-\Omega(n))$. The theorem follows by the reverse transformation of v_e's in (6.2). □

As shown in Section 6.1, the constraint $\lambda_2(\sum_{e \in E} \mathsf{x}(e) \cdot \mathsf{b}_e \mathsf{b}_e^\top) \geqslant \lambda_{\mathsf{opt}}$ can be incorporated into network design, and so Theorem 6.1.5 implies the following result.

Theorem 6.2.2. *There is a polynomial time randomized algorithm which returns a zero-one solution $z \in \{0,1\}^m$ with probability at least $1 - \exp(-\Omega(n))$ such that*

$$\lambda_2 \left(\sum_{e \in E} \mathsf{z}(e) \cdot \mathsf{b}_e \mathsf{b}_e^\top \right) \geqslant \lambda_{\mathsf{opt}} \quad \text{and} \quad \sum_{e \in E} \mathsf{c}(e) \cdot \mathsf{z}(e) \leqslant (1 + O(\varepsilon))C + O\Big(\frac{n \left\| \mathsf{c} \right\|_\infty}{\varepsilon}\Big).$$

Furthermore, this can be done while incorporating other constraints (e.g., connectivity constraints) as described in Theorem 6.1.5.

6.2.2 Minimizing Total Effective Resistance

Ghosh, Boyd and Saberi [70] study the problem of designing a network that minimizes the total effective resistance. The problem is formulated as follows.

$$\begin{aligned} R_{\mathsf{opt}} := \min_{x \in \mathbb{R}^{|E|}} \quad & \frac{1}{2} \sum_{u,v \in V} \mathrm{Reff}_{uv}(\mathsf{x}) \\ \text{subject to} \quad & \sum_{e \in E} \mathsf{x}(e) \leqslant k, \\ & \mathsf{x}(e) \in \{0,1\}, \forall e \in E. \end{aligned} \tag{6.4}$$

They showed that if the constraint $\mathsf{x}(e) \in \{0,1\}$ is relaxed to $\mathsf{x}(e) \in [0,1]$, then the relaxation is convex and can be written as a semidefinite program. They did not provide any result for the discrete optimization version in (6.4).

Ghosh, Boyd and Saberi [70] also show that the total effective resistance is a useful measure in different problems, e.g., average commute time, power dissipation in a resistor network, Elmore delay in a RC Circuit, total time constant of an averaging network, and euclidean variance. Furthermore, they established a connection between (6.4) and the A-design problem described in Chapter 7. To see this, note that the objective of (6.4) can be written as

$$\frac{1}{2}\sum_{u,v\in V}\mathrm{Reff}_{x}(u,v)=\frac{1}{2}\sum_{u\neq v\in V}b_{uv}^{\top}L_{x}^{\dagger}b_{uv}=\left\langle L_{x}^{\dagger},\frac{1}{2}\sum_{u\neq v\in V}b_{uv}b_{uv}^{\top}\right\rangle$$
$$=\left\langle L_{x}^{\dagger},nI_{n}-1_{n}1_{n}^{\top}\right\rangle=n\,\mathrm{tr}\left(L_{x}^{\dagger}\right),$$

where the last equality follows as $L_x^\dagger$ is orthogonal to 1_n. Hence, minimizing total effective resistance is equivalent to minimizing $\mathrm{tr}(L_x^\dagger)=\mathrm{tr}\left((\sum_{e\in E}x(e)b_eb_e^\top)^\dagger\right)$, which is the same as the A-design objective function after we project the vectors onto the subspace orthogonal to the all-one vector.

With this connection, all the recent algorithms for the A-optimal design can be applied to solve (6.4). For instances, the regret minimization algorithm in [6] gives a $(1+\varepsilon)$-approximation algorithm when $k\geqslant\Omega(n/\varepsilon^2)$, and the proportional volume sampling in [118] achieves $(1+\varepsilon)$-approximation with weaker assumption $k\geqslant\Omega(n/\varepsilon)$.

With a same reduction as in Theorem 6.2.1, we obtain the following approximation result for the more general weighted setting, where every edge has a cost $c(e)$ and we are given a cost budget C as in (6.3).

Theorem 6.2.3. *Suppose $C\geqslant\frac{15n\|c\|_\infty}{\varepsilon^2}$. There is a polynomial time randomized $(1+O(\varepsilon))$-approximation algorithm for the weighted version of (6.4).*

The assumption of C can be improved to $C\geq\Omega(\frac{n\|c\|_\infty}{\varepsilon})$ by a refined analysis of the iterative randomized rounding algorithm (see Corollary 7.1.7).

As shown in Section 6.1, the effective resistance constraints can be incorporated into network design, and so Theorem 6.1.5 implies the following result.

Theorem 6.2.4. *There is a polynomial time randomized algorithm which returns a zero-one solution $\mathsf{z} \in \{0,1\}^m$ with probability at least $1 - \exp(-\Omega(n))$ such that*

$$\frac{1}{2}\sum_{u,v\in V} \mathrm{Reff}_{\mathsf{z}}(u,v) \leqslant R_{\mathsf{opt}} \quad \text{and} \quad \sum_{e\in E} \mathsf{c}(e)\cdot \mathsf{x}(e) \leqslant (1+O(\varepsilon))C + O\Big(\frac{n\,\|\mathsf{c}\|_\infty}{\varepsilon}\Big).$$

Furthermore, this can be done while incorporating other constraints (e.g., connectivity constraints) as described in Theorem 6.1.5.

6.2.3 Network Design for Effective Resistances

Using similar reduction as in Theorem 6.2.1 and Theorem 6.2.3, we can handle network design problems with a budget constraint and the objective function being the sum of multiple pairs of effective resistances. We can obtain similar guarantees as in Theorem 6.2.3.

However, we would like to point out that when the budget C is small, e.g., sublinear in n, the spectral rounding technique cannot provide good approximation guarantee. This is unavoidable as suggested by the tight examples of spectral rounding (see Section 5.4). In Chapter 8, we consider a special case in this regime, and manage to find a constant approximation algorithm without using spectral rounding in special settings.

We leave it as an open problem that whether the spectral techniques can help to design good approximation algorithm in this regime.

6.3 Unweighted Spectral Sparsification

We show that the spectral rounding results can also be applied to the study of unweighted spectral sparsification.

6.3.1 Previous Work

Batson, Spielman, and Srivastava [21] proved that any graph has a $(1 \pm \varepsilon)$-spectral sparsifier with only $O(n/\varepsilon^2)$ edges, by carefully reweighting the edges of the original graph where different edges may have different weights (see Section 2.5 for more details). If we require all the edges to have the same weight, then there are simple examples (e.g., barbell graphs) showing that linear-sized spectral sparsification is not always possible. In a recent paper [20], Bansal, Svensson and Trevisan asked whether there is a non-trivial notion of unweighted spectral sparsification with which linear-sized spectral sparsification is always possible. They study a notion suggested by Oveis Gharan.

Definition 6.3.1 (Additive Unweighted Spectral Sparsifier). *Given a graph $G = (V, E)$ with n vertices, m edges and maximum degree $d_{\max}$, a subgraph $\widetilde{G} = (V, F)$ with $\widetilde{m}$ edges is an additive spectral sparsifier with error $\varepsilon \in [0, 1]$ if*

$$-\varepsilon d_{\max} I_n \preccurlyeq \frac{m}{\widetilde{m}} L_{\widetilde{G}} - L_G \preccurlyeq \varepsilon d_{\max} I_n.$$

Bansal, Svensson and Trevisan [20] prove that sparse additive unweighted spectral sparsification is always possible, and they provide both deterministic and randomized algorithms for constructing these sparsifiers.

Theorem 6.3.2 (Randomized Construction [20]). *Given a graph $G = (V, E)$ with n vertices, m edges, maximum degree $d_{\max}$, and $\varepsilon \in (0, 1)$, there is a polynomial time randomized algorithm that finds a subset of edges $F \subseteq E$ with size $\widetilde{m} = |F| = O(n\varepsilon^{-2} \log(1/\varepsilon)^3)$ such that $\widetilde{G} = (V, F)$ satisfies*

$$-\varepsilon d_{\max} I_n \preccurlyeq \frac{m}{\widetilde{m}} L_{\widetilde{G}} - L_G \preccurlyeq \varepsilon d_{\max} I_n.$$

Theorem 6.3.3 (Deterministic Construction [20]). *Given a graph $G = (V, E)$ with n vertices, m edges, maximum degree $d_{\max}$, and $\varepsilon \in (0, 1)$, there is a polynomial time deterministic algorithm that finds a multi-set F of edges with size $\widetilde{m} = |F| = O(\frac{n}{\varepsilon^2})$ such that $\widetilde{G} = (V, F)$ satisfies*

$$2\frac{m}{\widetilde{m}} D_{\widetilde{G}} - 2D_G - \varepsilon d_{\max} I \preccurlyeq \frac{m}{\widetilde{m}} L_{\widetilde{G}} - L_G \preccurlyeq \varepsilon d_{\max} I,$$

where D_G is the diagonal degree matrix of G and $D_{\widetilde{G}}$ is the diagonal degree matrix of $\widetilde{G}$.

The proof of Theorem 6.3.2 is by Lovász local lemma and the converse of expander mixing lemma by Bilu and Linial. The proof of Theorem 6.3.3 is by the regret minimization framework of Allen-Zhu, Liao and Orecchia [7].

Note that Theorem 6.3.3 has a slightly weaker spectral lower bound guarantee than Theorem 6.3.2. Also, Theorem 6.3.3 can only return a multi-set solution where some edges can be used more than once, and so the sparsifier is integer weighted rather than unweighted where every edge has the same weight.

6.3.2 Nonconstructive Spectral Rounding and Unweighted Spectral Sparsification

We show that the existence of a linear-sized additive unweighted spectral sparsifier follows from the two-sided rounding result in Theorem 5.3.2. The idea is to view the original graph as a fractional solution where every edge e has $\mathsf{x}(e) = \widetilde{m}/m$, and then use Theorem 5.3.2 to round this fractional solution to a zero-one solution while preserving the spectral properties of the original graph. The additional linear constraint in Theorem 5.3.2 allows us to bound the number of edges in the sparsifier.

Theorem 6.3.4. *Suppose we are given a graph $G = (V, E)$ with n vertices, m edges, and maximum degree $d_{\max}$. For any $\varepsilon \in (0, 1]$, there exists a subset of edges $F \subseteq E$ with $|F| \in [(1 - 4\sqrt{2}\varepsilon)\widetilde{m}, (1 + 4\sqrt{2}\varepsilon)\widetilde{m}]$ where $\widetilde{m} = \frac{n}{\varepsilon^2}$ such that*

$$-8\sqrt{2}\varepsilon d_{\max} \mathsf{I}_n \preccurlyeq \mathsf{L}_G - \frac{m}{\widetilde{m}} \sum_{e \in F} \mathsf{b}_e \mathsf{b}_e^\top \preccurlyeq 8\sqrt{2}\varepsilon d_{\max} \mathsf{I}_n.$$

Proof. The plan is to apply Theorem 5.3.2 with $\mathsf{v}_e := \mathsf{b}_e$, $\mathsf{x}(e) := \frac{\widetilde{m}}{m}$ and $c := \mathit{1}_m$. We will first define the parameters λ and l and check that the assumptions $l \leqslant \sqrt{\lambda}$ and $\|\mathsf{c}\|_\infty \frac{\leqslant l^2 \langle \mathsf{c}, \mathsf{x} \rangle}{\lambda}$ in Theorem 5.3.2 are satisfied. Note that

$$\left\| \sum_{e \in E} \mathsf{x}(e) \cdot \mathsf{v}_e \mathsf{v}_e^\top \right\|_{\mathrm{op}} = \frac{\widetilde{m}}{m} \|\mathsf{L}_G\|_{\mathrm{op}} \leqslant \frac{2 d_{\max} \widetilde{m}}{m} \quad \text{and} \quad \|\mathsf{v}_e\| = \sqrt{2} \text{ for all } e \in E.$$

So we define $\lambda := \frac{2d_{\max}\widetilde{m}}{m}$ and $l := \sqrt{2}$. We check that $\lambda = \frac{2d_{\max}\widetilde{m}}{m} = \frac{2dn}{\varepsilon^2 m} \geqslant \frac{2}{\varepsilon^2} \geqslant 2 = l^2$, and $\frac{l^2\langle c,x\rangle}{\lambda} = \frac{2\widetilde{m}}{2d_{\max}\widetilde{m}/m} = \frac{m}{d_{\max}} \geqslant 1 = \|c\|_\infty$. Therefore, we can apply Theorem 5.3.2 to conclude that there exists a subset of edges $F \subseteq E$ (corresponding to the zero-one solution z) such that

$$\left\|\sum_{e\in E} x(e)\cdot v_e v_e^\top - \sum_{e\in F} v_e v_e^\top\right\|_{\text{op}} \leqslant 16\sqrt{\frac{d_{\max}\widetilde{m}}{m}} \quad \text{and}$$

$$\left|\sum_{e\in E} x(e)c(e) - \sum_{e\in F} c(e)\right| \leqslant 8\sqrt{\frac{m}{d_{\max}\widetilde{m}}}\cdot\langle c,x\rangle.$$

Plugging in $x(e) = \frac{\widetilde{m}}{m}$ and $c = \vec{1}$ and $\widetilde{m} = \frac{n}{\varepsilon^2}$, the first statement implies that

$$\left\|L_G - \frac{m}{\widetilde{m}}\sum_{e\in F} v_e v_e^\top\right\|_{\text{op}} = \left\|\sum_{e\in E} v_e v_e^\top - \frac{m}{\widetilde{m}}\sum_{e\in F} v_e v_e^\top\right\|_{\text{op}} \leqslant 16\sqrt{\frac{d_{\max}m}{\widetilde{m}}}$$

$$= 16\sqrt{\frac{\varepsilon^2 d_{\max} m}{n}} \leqslant 8\sqrt{2}\varepsilon d_{\max},$$

where the last inequality uses $m \leqslant \frac{d_{\max}n}{2}$ as the maximum degree is $d_{\max}$. Finally, the second statement implies that

$$\left|\widetilde{m} - |F|\right| \leqslant 8\sqrt{\frac{\varepsilon^2 m}{d_{\max}n}}\cdot\widetilde{m} \leqslant 4\sqrt{2}\varepsilon\widetilde{m}. \qquad \square$$

Note that Theorem 6.3.4 improves Theorem 6.3.2 slightly by removing a factor of $\log^3(1/\varepsilon)$ in the number of edges of the sparsifier. This confirms the existence of unweighted additive spectral sparsifiers with $O(\frac{n}{\varepsilon^2})$ edges, which was not known before. More generally, we can use the same proof with a cost function c with $\|c\|_\infty \leqslant \frac{\|c\|_1}{d_{\max}}$ to obtain a sparsifier with $\widetilde{m} = \frac{n}{\varepsilon^2}$ and

$$\left\|L_G - \frac{m}{\widetilde{m}}\sum_{e\in F} v_e v_e^\top\right\|_{\text{op}} \leqslant 8\sqrt{2}\varepsilon d_{\max} \quad \text{and}$$

$$(1-4\sqrt{2}\varepsilon)\sum_{e\in E} c(e) \leqslant \frac{m}{\widetilde{m}}\sum_{e\in F} c(e) \leqslant (1+4\sqrt{2}\varepsilon)\sum_{e\in E} c(e).$$

We remark that the same reduction in [20] can be used to replace $d_{\max} I$ by $D_G + d_{\text{avg}} I$ where D_G is the diagonal degree matrix of G and d_{avg} is the average degree in G.

The main disadvantage of Theorem 6.3.4 is that it does not provide a polynomial time algorithm to find such a sparsifier. It is a major open problem to make the method of interlacing polynomials used in [110, 111, 91] constructive.

6.3.3 Constructive Spectral Rounding and Unweighted Spectral Sparsification

For the determinstic algorithm, using similar techniques in [5, 6] which proves Lemma 4.2.9, we can strengthen Theorem 6.3.3 by returning a subgraph with no parallel edges.

Theorem 6.3.5. *Given a graph $G = (V, E)$ with n vertices, m edges, maximum degree $d_{\max}$, and $\varepsilon \in (0, \frac{1}{10})$, there is a polynomial time deterministic algorithm that finds a* subset *F of edges with size $\widetilde{m} = |F| = O(\frac{n}{\varepsilon^2})$ such that $\widetilde{G} = (V, F)$ satisfies*

$$2\frac{m}{\widetilde{m}} D_{\widetilde{G}} - 2D_G - O(\varepsilon) d_{\max} I_n \preccurlyeq \frac{m}{\widetilde{m}} L_{\widetilde{G}} - L_G \preccurlyeq O(\varepsilon) d_{\max} I_n.$$

The algorithm is a slight modification of the algorithm in [20], which is a greedy algorithm based on the regret minimization framework. The analysis of the algorithm is also slightly different from the one in [20] that we apply the generic regret bound in Theorem 4.2.6 with feedback matrices in general form (do not need to be positive semidefinite or negative semidefinite). In particular, the feedback matrices are of the following form

$$F_0 = 0 \quad \text{and} \quad F_t = \begin{pmatrix} L_G - mL_e & \\ & L_G^+ - mL_e^+ \end{pmatrix} \quad \text{for some } e \in E \text{ and } t \geqslant 1,$$

where L_G is the Laplacian matrix of the original graph, $L_G^+ := D_G + A_G$ is the signless-Laplacian of the original graph, and L_e and L_e^+ are the Laplacian and signless-Laplacian matrix of a single edge e. Note that we always have $F_t \preccurlyeq 2d_{\max} I_n$, as $L_G \preccurlyeq 2d_{\max} I_n$ and $L_G^+ \preccurlyeq 2d_{\max} I_n$ for a graph G of maximum degree $d_{\max}$.

Greedy Additive Spectral Sparsification

```
Input: An error parameter $\varepsilon \in (0,1)$, and a graph $G = (V, E)$ with $n$ vertices, $m \geqslant \frac{2n}{\varepsilon^2}$
       edges and maximum degree $d_{\max}$.

Output: A subgraph $\widetilde{G}$ of $G$ with $\widetilde{m} = O(\frac{n}{\varepsilon^2})$ edges satisfying

    $$2\frac{m}{\widetilde{m}}\mathsf{D}_{\widetilde{G}} - 2\mathsf{D}_G - O(\varepsilon)d_{\max}\mathsf{I}_n \preccurlyeq \frac{m}{\widetilde{m}}\mathsf{L}_{\widetilde{G}} - \mathsf{L}_G \preccurlyeq O(\varepsilon)d_{\max}\mathsf{I}_n.$$

    1. Initialization: Set $S_0 \leftarrow \emptyset$, $\mathsf{F}_0 \leftarrow 0$, $\tau \leftarrow \frac{n}{\varepsilon^2}$, and $\alpha \leftarrow \frac{\varepsilon}{\sqrt{d_{\max} m}}$.

    2. For $t = 1$ to $\tau$ do

        (a) Compute the action matrix $\mathsf{A}_t = (\alpha \sum_{j=0}^{t-1} \mathsf{F}_j + l_t \mathsf{I}_{2n})^{-2}$, where $l_t \in \mathbb{R}$ is the
            unique value such that $\mathsf{A}_t \succ 0$ and $\mathrm{tr}(\mathsf{A}_t) = 1$[1].

        (b) Select an edge $e_t \in E \backslash S_{t-1}$ such that

            $$\left\langle \mathsf{A}_t, \begin{pmatrix} \mathsf{L}_G - m\mathsf{L}_{e_t} & \\ & \mathsf{L}_G^{+} - m\mathsf{L}_{e_t}^{+} \end{pmatrix} \right\rangle \geqslant -\frac{2\sqrt{n}}{\alpha m} = -\varepsilon d_{\max}.$$

        (c) Set

            $$\mathsf{F}_t \leftarrow \begin{pmatrix} \mathsf{L}_G - m\mathsf{L}_{e_t} & \\ & \mathsf{L}_G^{+} - m\mathsf{L}_{e_t}^{+} \end{pmatrix} \quad \text{and} \quad S_t \leftarrow S_{t-1} \cup \{e_t\}.$$

    3. Return $\widetilde{G} = (V, S_\tau)$ as the solution.
```

Note that we can assume $m \geqslant \frac{2n}{\varepsilon^2} = 2\tau$, as otherwise we can simply return $\widetilde{G} = G$ as our solution. The only difference with the algorithm in [20] is in Step 2(b), where we insist on choosing an edge $e_t \in E \setminus S_{t-1}$ to guarantee that the returned solution is a simple subgraph. If there is no such restriction, then a simple averaging argument in [20] shows

[1] Step 2(a) can be implemented with a binary search step as mentioned in Remark 4.3.1.

that there is an edge $e \in E$ with the inner product in Step 2(b) being non-negative. With this restriction, we will use the closed-form of the action matrix and Lemma 4.2.9 to show that there is still an edge with the inner product in Step 2(b) being not too small. The following lemma is the new ingredient for the proof of Theorem 6.3.5.

Lemma 6.3.6. *For each $1 \leqslant t \leqslant \tau$ and $\alpha = \frac{\varepsilon}{\sqrt{d_{\max} m}}$, there always exists an edge $e \in E \backslash S_{t-1}$ such that*

$$\left\langle A_t, \begin{pmatrix} L_G - mL_e & \\ & L_G^+ - mL_e^+ \end{pmatrix} \right\rangle \geqslant -\frac{2\sqrt{n}}{\alpha m} \geqslant -\varepsilon d_{\max}.$$

Proof. The sum of the inner product over all edges in $E \backslash S_{t-1}$ is

$$\begin{aligned} &\sum_{e \in E \backslash S_{t-1}} \left\langle A_t, \begin{pmatrix} L_G - mL_e & \\ & L_G^+ - mL_e^+ \end{pmatrix} \right\rangle \\ =& \sum_{e \in E} \left\langle A_t, \begin{pmatrix} L_G - mL_e & \\ & L_G^+ - mL_e^+ \end{pmatrix} \right\rangle - \sum_{e \in S_{t-1}} \left\langle A_t, \begin{pmatrix} L_G - mL_e & \\ & L_G^+ - mL_e^+ \end{pmatrix} \right\rangle \\ =& \left\langle A_t, \begin{pmatrix} mL_G - m\sum_{e \in E} L_e & \\ & mL_G^+ - m\sum_{e \in E} L_e^+ \end{pmatrix} \right\rangle - \left\langle A_t, \sum_{e \in S_{t-1}} \begin{pmatrix} L_G - mL_e & \\ & L_G^+ - mL_e^+ \end{pmatrix} \right\rangle \\ =& -\left\langle A_t, \sum_{e \in S_{t-1}} \begin{pmatrix} L_G - mL_e & \\ & L_G^+ - mL_e^+ \end{pmatrix} \right\rangle, \end{aligned}$$

where the last equality follows from $\sum_{e \in E} L_e = L_G$ and $\sum_{e \in E} L_e^+ = L_G^+$. Let

$$Z_t := \sum_{e \in S_{t-1}} \begin{pmatrix} L_G - mL_e & \\ & L_G^+ - mL_e^+ \end{pmatrix},$$

and let the eigenvalues of Z_t be $\lambda_1, ..., \lambda_{2n}$. Note that $\lambda_{\min}(Z_t) \leqslant 0$ as $\operatorname{tr}(L_G) = \operatorname{tr}(L_G^+) = 2m$ and $m \operatorname{tr}(L_e) = m \operatorname{tr}(L_e^+) = 2m$ which imply that $\operatorname{tr}(Z_t) = 0$.

Observe that $A_t = (l_t I_{2n} + \alpha Z_t)^{-2}$ and so A_t and Z_t have the same eigenbasis, and the i-th eigenvalue of A_t is $(l_t + \alpha\lambda_i)^{-2}$. It follows that

$$\begin{aligned} -\langle A_t, Z_t \rangle = \sum_{i=1}^{2n} \frac{-\lambda_i}{(l_t + \alpha\lambda_i)^2} &= \sum_{i=1}^{2n} \frac{l_t/\alpha}{(l_t + \alpha\lambda_i)^2} - \frac{1}{\alpha} \sum_{i=1}^{2n} \frac{l_t + \alpha\lambda_i}{(l_t + \alpha\lambda_i)^2} \\ &= \frac{l_t}{\alpha} - \frac{\operatorname{tr}(A_t^{\frac{1}{2}})}{\alpha} \geqslant -\lambda_{\min}(Z_t) - \frac{\operatorname{tr}(A_t^{\frac{1}{2}})}{\alpha} \geqslant -\frac{\sqrt{n}}{\alpha}, \end{aligned}$$

where the first equality in the second line is because $\operatorname{tr}(A_t) = 1$ and $(l_t + \alpha\lambda_i)^{-1}$ is the i-th eigenvalue of $A_t^{\frac{1}{2}}$, the second last inequality is by $A_t \succ 0$ which implies that $l_t > -\alpha\lambda_{\min}(Z_t)$, and the last inequality is by $\lambda_{\min}(Z_t) \leqslant 0$ and $\operatorname{tr}(A_t^{\frac{1}{2}}) \leqslant \sqrt{n}$ from Claim 2.1.10.

Since $|E \setminus S_{t-1}| = m - t + 1$, an averaging argument shows that there exists an edge $e \in E \backslash S_{t-1}$ such that

$$\left\langle A_t, \begin{pmatrix} L_G - mL_e & \\ & L_G^+ - mL_e^+ \end{pmatrix} \right\rangle \geqslant -\frac{\sqrt{n}}{\alpha(m-t+1)} \geqslant -\frac{2\sqrt{n}}{\alpha m},$$

where the last inequality is because $m - t + 1 \geqslant m - \tau + 1 \geqslant m/2$ by our assumption $\tau = \frac{n}{\varepsilon^2} \leqslant \frac{m}{2}$. Finally, when $\alpha = \frac{\varepsilon}{\sqrt{d_{\max} m}}$,

$$\frac{2\sqrt{n}}{\alpha m} = \frac{2}{\varepsilon}\sqrt{\frac{d_{\max} n}{m}} \leqslant \sqrt{2 d_{\max}} \leqslant \varepsilon d_{\max},$$

where the first inequality is by our assumption $\tau = \frac{n}{\varepsilon^2} \leqslant \frac{m}{2}$, and the second inequality follows from $d_{\max} n \geqslant m \geqslant \frac{2n}{\varepsilon^2}$ which implies $\varepsilon \geqslant \sqrt{\frac{2}{d_{\max}}}$. □

Given Lemma 6.3.6, the rest of the proof is similar to the one in [20], the main difference is that we apply the regret bound in Theorem 4.2.6 instead of the one in [7]. The following lemma bounds several crucial terms in the regret bound in Theorem 4.2.6.

Lemma 6.3.7. *Let* $P_t := \begin{pmatrix} L_G & \\ & L_G^+ \end{pmatrix}^{\frac{1}{2}}$ *and* $N_t := \begin{pmatrix} mL_{e_t} & \\ & mL_{e_t}^+ \end{pmatrix}^{\frac{1}{2}}$ *such that* $P_t P_t^\top - N_t N_t^\top = F_t$. *If* $\alpha = \frac{\varepsilon}{\sqrt{d_{\max} m}}$, *then*

$$\begin{aligned} \langle P_t P_t^\top, A_t \rangle &\leqslant 2 d_{\max}, \qquad \alpha \left\| A_t^{\frac{1}{4}} P_t P_t^\top A_t^{\frac{1}{4}} \right\|_{\mathrm{op}} \leqslant \varepsilon \qquad \textit{and} \\ \langle N_t N_t^\top, A_t \rangle &\leqslant (2+\varepsilon) d_{\max}, \qquad \alpha \left\| A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}} \right\|_{\mathrm{op}} \leqslant 3\varepsilon. \end{aligned}$$

Proof. Since $0 \preccurlyeq L_G, L_G^+ \preccurlyeq 2 d_{\max} I_n$, it follows that $0 \preccurlyeq P_t P_t^\top = \begin{pmatrix} L_G & \\ & L_G^+ \end{pmatrix} \preccurlyeq 2 d_{\max} I_{2n}$. The density matrix A_t has trace one, thus $\langle P_t P_t^\top, A_t \rangle \leqslant 2 d_{\max}$.

Since the feedback matrices F_t have a block diagonal structure, by the closed-form solution of the action matrix in (4.9), A_t also has the same block diagonal structure

$$A_t = \begin{pmatrix} B_t & \\ & C_t \end{pmatrix}, \quad \text{where} \quad 0 \preccurlyeq B_t, C_t \preccurlyeq I_n.$$

Therefore,

$$\left\| A_t^{\frac{1}{4}} P_t P_t^\top A_t^{\frac{1}{4}} \right\|_{\text{op}} = \max \left\{ \left\| B_t^{\frac{1}{4}} L_G B_t^{\frac{1}{4}} \right\|_{\text{op}}, \ \left\| C_t^{\frac{1}{4}} L_G^+ C_t^{\frac{1}{4}} \right\|_{\text{op}} \right\} \leqslant 2d_{\max}, \tag{6.5}$$

where the last inequality follows by $0 \preccurlyeq L_G, L_G^+ \preccurlyeq 2d_{\max} I_n$ and $0 \preccurlyeq B_t, C_t \preccurlyeq I_n$. For $\alpha = \frac{\varepsilon}{\sqrt{d_{\max} m}}$, it holds that $\alpha \left\| A_t^{\frac{1}{4}} P_t P_t^\top A_t^{\frac{1}{4}} \right\|_{\text{op}} \leqslant \varepsilon$. Note that this is pretty loose bound.

Then, we consider the term $\langle N_t N_t^\top, A_t \rangle$. The choice of edge e_t and Lemma 6.3.6 guarantee that

$$\langle P_t P_t^\top, A_t \rangle - \langle N_t N_t^\top, A_t \rangle = \left\langle A_t, \begin{pmatrix} L_G - mL_{e_t} & \\ & L_G^+ - mL_{e_t}^+ \end{pmatrix} \right\rangle \geqslant -\varepsilon d_{\max}.$$

As we already proved that $\langle P_t P_t^\top, A_t \rangle \leqslant 2d_{\max}$, it follows that

$$\langle N_t N_t^\top, A_t \rangle \leqslant (2+\varepsilon) d_{\max}. \tag{6.6}$$

Finally, we consider the term $\alpha \left\| A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}} \right\|_{\text{op}}$. Similar to (6.5), it hols that

$$\left\| A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}} \right\|_{\text{op}} = \max \left\{ m \left\| B_t^{\frac{1}{4}} L_{e_t} B_t^{\frac{1}{4}} \right\|_{\text{op}}, \ m \left\| C_t^{\frac{1}{4}} L_{e_t}^+ C_t^{\frac{1}{4}} \right\|_{\text{op}} \right\}.$$

We will just bound the first term, as the second term can be bounded the same way.

Let $B_t = \sum_{i=1}^n \lambda_i \cdot y_i y_i^\top$ be the eigendecomposition of B_t, and let $w = \sqrt{m} \cdot b_{e_t}$ so that $ww^\top = mL_{e_t}$ and $\|w\|_2 = \sqrt{2m}$. Then

$$\begin{aligned} m \left\| B_t^{\frac{1}{4}} L_{e_t} B_t^{\frac{1}{4}} \right\|_{\text{op}} = w^\top B_t^{\frac{1}{2}} w &= \sum_{i=1}^n \sqrt{\lambda_i} \cdot \langle w, y_i \rangle^2 \\ &\leqslant \sqrt{\sum\nolimits_{i=1}^n \langle w, y_i \rangle^2} \cdot \sqrt{\sum\nolimits_{i=1}^n \lambda_i \cdot \langle w, y_i \rangle^2} \\ &= \|w\|_2 \cdot \sqrt{w^\top B_t w} \\ &= \sqrt{2m} \cdot \sqrt{m \langle B_t, L_{e_t} \rangle}, \end{aligned}$$

where the inequality is by Cauchy-Schwartz, and the last equality follows from $w^\top B_t w = m\langle B_t, L_{e_t}\rangle$. Notice that $m\langle B_t, L_{e_t}\rangle + m\langle C_t, L_{e_t}^+\rangle = \langle N_t N_t^\top, A_t\rangle$ and both $\langle B_t, L_{e_t}\rangle, \langle C_t, L_{e_t}^+\rangle \geqslant 0$, thus it follows from (6.6) that $m\langle B_t, L_{e_t}\rangle \leqslant (2+\varepsilon)d_{\max}$, which implies

$$m\left\|B_t^{\frac{1}{4}} L_{e_t} B_t^{\frac{1}{4}}\right\|_{\mathrm{op}} \leqslant 2\sqrt{\left(1+\frac{\varepsilon}{2}\right)d_{\max}m}.$$

The same arguments gives the same upper bound on $m\|C_t^{\frac{1}{4}} L_{e_t}^+ C_t^{\frac{1}{4}}\|$. Therefore, for $\alpha = \frac{\varepsilon}{\sqrt{d_{\max}m}}$ with $\varepsilon \in (0,1)$,

$$\alpha\left\|A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}}\right\|_{\mathrm{op}} \leqslant \frac{\varepsilon}{\sqrt{d_{\max}m}} \cdot 2\sqrt{\left(1+\frac{\varepsilon}{2}\right)d_{\max}m} \leqslant 3\varepsilon. \qquad \square$$

We are ready to prove Theorem 6.3.5 with Theorem 4.2.6.

Theorem 6.3.5. *Given a graph $G = (V,E)$ with n vertices, m edges, maximum degree $d_{\max}$, and $\varepsilon \in (0, \frac{1}{10})$, there is a polynomial time deterministic algorithm that finds a* subset *F of edges with size $\widetilde{m} = |F| = O(\frac{n}{\varepsilon^2})$ such that $\widetilde{G} = (V,F)$ satisfies*

$$2\frac{m}{\widetilde{m}}D_{\widetilde{G}} - 2D_G - O(\varepsilon)d_{\max}I_n \preccurlyeq \frac{m}{\widetilde{m}}L_{\widetilde{G}} - L_G \preccurlyeq O(\varepsilon)d_{\max}I_n.$$

Proof. Let $P_t := \begin{pmatrix} L_G & \\ & L_G^+ \end{pmatrix}^{\frac{1}{2}}$ and $N_t := \begin{pmatrix} mL_{e_t} & \\ & mL_{e_t}^+ \end{pmatrix}^{\frac{1}{2}}$. By Lemma 6.3.7, when $\alpha = \frac{\varepsilon}{\sqrt{d_{\max}m}}$ with small enough ε (e.g., $\varepsilon \in (0, \frac{1}{6})$), it follows that $\alpha\|A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}}\| \leqslant 3\varepsilon \leqslant \frac{1}{2}$ for all t. Therefore, we can apply Theorem 4.2.6 and get

$$\lambda_{\min}\left(\sum_{t=0}^{\tau} F_t\right) \geqslant \sum_{t=1}^{\tau}\left(\frac{\langle P_t P_t^\top, A_t\rangle}{1+2\alpha\left\|A_t^{\frac{1}{4}} P_t P_t^\top A_t^{\frac{1}{4}}\right\|_{\mathrm{op}}} - \frac{\langle N_t N_t^\top, A_t\rangle}{1-2\alpha\left\|A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}}\right\|_{\mathrm{op}}}\right) - \frac{2\sqrt{n}}{\alpha}.$$

Since $\alpha\left\|A_t^{\frac{1}{4}} P_t P_t^\top A_t^{\frac{1}{4}}\right\|_{\mathrm{op}} \leqslant \varepsilon$ and $\alpha\left\|A_t^{\frac{1}{4}} N_t N_t^\top A_t^{\frac{1}{4}}\right\|_{\mathrm{op}} \leqslant 3\varepsilon$ by Lemma 6.3.7, it follows that

$$\begin{aligned}\lambda_{\min}\left(\sum_{t=0}^{\tau} F_t\right) &\geqslant \sum_{t=1}^{\tau}\left(\frac{\langle P_t P_t^\top, A_t\rangle}{1+2\varepsilon} - \frac{\langle N_t N_t^\top, A_t\rangle}{1-6\varepsilon}\right) - \frac{2\sqrt{n}}{\alpha} \\ &= \sum_{t=1}^{\tau}\frac{\langle P_t P_t^\top - N_t N_t^\top, A_t\rangle - 6\varepsilon\langle P_t P_t^\top, A_t\rangle - 2\varepsilon\langle N_t N_t, A_t\rangle}{(1+2\varepsilon)(1-6\varepsilon)} - \frac{2\sqrt{n}}{\alpha}.\end{aligned}$$

By Lemma 6.3.6 $\langle P_tP_t^\top - N_tN_t^\top, A_t\rangle \geqslant -\varepsilon d_{\max}$, and by Lemma 6.3.7 $\langle P_tP_t^\top, A_t\rangle \leqslant 2d_{\max}$ and $\langle N_tN_t^\top, A_t\rangle \leqslant (2+\varepsilon)d_{\max}$. Thus for $\alpha = \frac{\varepsilon}{\sqrt{d_{\max}m}}$, $\tau = \frac{n}{\varepsilon^2}$, and $\varepsilon \in (0, \frac{1}{10})$ it holds that

$$\lambda_{\min}\begin{pmatrix} \tau L_G - \sum\limits_{t=1}^{\tau} mL_{e_t} & \\ & \tau L_G^+ - \sum\limits_{t=1}^{\tau} mL_{e_t}^+ \end{pmatrix} = \lambda_{\min}\left(\sum_{t=0}^{\tau} F_t\right) \geqslant -O(\varepsilon)\tau d_{\max} - \frac{2\sqrt{d_{\max}mn}}{\varepsilon} \geqslant -O(\varepsilon)\tau d_{\max},$$

where the last inequality holds as $d_{\max}mn \leqslant d_{\max}^2 n^2 = \varepsilon^4 d_{\max}^2\tau^2$.

Let $\widetilde{m} := \tau$ and $L_{\widetilde{G}} = \sum_{t=1}^{\tau} L_{e_t}$. From the first block, we have

$$\tau L_G - \sum_{t=1}^{\tau} mL_{e_t} \succcurlyeq -O(\varepsilon)\tau d_{\max} I_n \quad \Longrightarrow \quad \frac{m}{\widetilde{m}} L_{\widetilde{G}} - L_G \preccurlyeq O(\varepsilon) d_{\max} I_n.$$

From the second block, we have

$$\tau L_G^+ - \sum_{t=1}^{\tau} mL_{e_t}^+ \succcurlyeq -O(\varepsilon)\tau d_{\max} I_n \quad \Longrightarrow \quad \frac{m}{\widetilde{m}} L_{\widetilde{G}} - L_G \succcurlyeq 2\frac{m}{\widetilde{m}} D_{\widetilde{G}} - 2D_G - O(\varepsilon) d_{\max} I_n,$$

where we used that $L_G^+ = 2D_G - L_G$ and $L_{\widetilde{G}}^+ = 2D_{\widetilde{G}} - L_{\widetilde{G}}$. □

Chapter 7

Applications of Spectral Rounding to Experimental Design

7.1 Introduction

In experimental design problems, we are given vectors $u_1, \ldots, u_n \in \mathbb{R}^d$ and a parameter $b \geqslant d$, and the goal is to choose a (multi-)subset S of b vectors so that $\sum_{i \in S} u_i u_i^\top$ optimizes some objective function. The most popular and well-studied objective functions are:

- D-design: Maximizing $\left(\det\left(\sum_{i \in S} u_i u_i^\top\right)\right)^{\frac{1}{d}}$.
- A-design: Minimizing $\operatorname{tr}\left(\left(\sum_{i \in S} u_i u_i^\top\right)^{-1}\right)$.
- E-design: Maximizing $\lambda_{\min}\left(\sum_{i \in S} u_i u_i^\top\right)$.

Two settings are studied in the literature. One is the "with repetition" setting where each vector is allowed to be chosen multiple times, and the other is the "without repetition" setting where each vector is allowed to be chosen at most once. By making b copies of each vector, we can reduce the with repetition setting to the without repetition setting easily. All the results in this chapter apply in the more general without repetition setting.

These problems of choosing a representative subset of vectors have a wide range of applications.

- Experimental design is a classical topic in statistics with extensive literature [60, 13, 121, 74], where the goal is to choose b (noisy) linear measurements from $u_1, \ldots, u_n \in \mathbb{R}^d$ so as to maximize the statistical efficiency of estimating an unknown vector in $\mathbb{R}^d$.
- In machine learning, they are used in active learning [10], feature selection [29], and data summarization [114, 34].
- In numerical linear algebra, they are used in column subset selection [15], sparse least square regression [28], and matrix approximation [50, 51].
- In signal processing, they are used in sensor placement problems [80], and optimal subsampling in graph signal processing [37, 40, 41].
- In network design, the problem of choosing a subgraph with at most b edges to minimize the total effective resistance [70, 98] is an A-design problem, and the problem of choosing a subgraph with at most b edges to maximize the algebraic connectivity [69, 86, 98] is an E-design problem.

We refer the interested readers to [140, 131, 108, 118, 6] for more discussions of these applications and further references on related work.

7.1.1 Our Results

We present both rounding algorithms and combinatorial algorithms for experimental design problems. A main contribution in this chapter is to show that these two types of algorithms can be analyzed using the same local search framework. Using this framework, we match and improve all known results and also obtain some new results.

7.1.1.1 Rounding Algorithms for Convex Programming Relaxations

There are natural convex programming relaxations for the D/A/E-design problems. The best known rounding algorithms for these three problems are all quite different, i.e. approximate positively correlated distributions for D-design [131], proportional volume sampling for A-design [118], and regret minimization for E-design [6]. Although the one-sided spectral rounding result in [6] provides a general solution for a large class of experimental design problems including D/A/E-design, this only works under the stronger assumption that $b \geq \Omega\left(\frac{d}{\varepsilon^2}\right)$ and it was unclear how to unify the best known algorithmic results.

Surprisingly, the iterative randomized swapping algorithm in Section 5.2 not only matches the best known result for E-design, but also matches and improves the previous results for D/A-design with slight modification. Moreover, the new algorithmic framework can extend previous results to handle multiple knapsack constraints. To match and improve the best known results for D/A-design, we bypass the one-sided spectral rounding problem. Instead, we perform a refined analysis for the iterative randomized swapping algorithm, in which the minimum eigenvalue of the current solution plays an unexpectedly crucial role for D/A-design as well. This provides a unified rounding algorithm to achieve the optimal results for the natural convex programming relaxations for these experimental design problems.

In D/A/E-design problems with knapsack constraints (we refer to them as *weighted experimental design problems*), we are given vectors $u_1, \ldots, u_n \in \mathbb{R}^d$, knapsack constraints $c_1, \ldots, c_m \in \mathbb{R}^n_+$ and budgets $b_1, \ldots, b_m \geqslant 0$. The goal is to find a solution $z \in \{0,1\}^n$ with $\langle c_j, z \rangle \leqslant b_j$ for $1 \leqslant j \leqslant m$ to optimize the D/A/E-design objective value. Consider the following natural convex programming relaxations for D/A/E-design.

$$\begin{aligned} \min_{x \in \mathbb{R}^d, X \in \mathbb{S}^d_{++}} \quad & f_D(X) \quad \text{or} \quad f_A(X) \quad \text{or} \quad f_E(X) \\ \text{subject to} \quad & X = \sum_{i=1}^n x(i) \cdot u_i u_i^\top, \\ & \langle c_j, x \rangle \leqslant b_j, \qquad \text{for } 1 \leqslant j \leqslant m, \\ & 0 \leqslant x(i) \leqslant 1, \qquad \text{for } 1 \leqslant i \leqslant n, \end{aligned} \tag{7.1}$$

where $f_D(X) := \det(X)^{-\frac{1}{d}}$, $f_A(X) := \operatorname{tr}(X^{-1})$, and $f_E(X) := (\lambda_{\min}(X))^{-1}$ are objective functions for D/A/E-design respectively.

The convexity of f_A follows by Fact 2.2.18. The convexity of f_D and f_E follow by the concavity of $\det(\mathsf{X})^{\frac{1}{d}}$ (Fact 2.2.16) and $\lambda_{\min}(\mathsf{X})$ (Lemma 2.2.14), and then applying the following well-known fact with $h(x) = x^{-1}$.

Fact (see, e.g., Section 3.2.4 in [31])**.** *Let $g : \mathbb{S}^d_{++} \to \mathbb{R}_{++}$ be a concave function, $h : \mathbb{R}_{++} \to \mathbb{R}$ be a convex nonincreasing function over $\mathbb{R}_{++}$. Then, the composition function $f = h \circ g$ is a convex function over $\mathbb{S}^d_{++}$.*

Remark. *The authors of [6] claimed that $f_D(\mathsf{X}) = \det(\mathsf{X})^{-\frac{1}{d}}$ is not convex. Hence, they considered another convex function $f_D(\mathsf{X}) = -\frac{1}{d}\log\det(\mathsf{X})$ instead. Although this does not make a difference for the results in [6] and this chapter, we can show that $\det(\mathsf{X})^{-\frac{1}{d}}$ is actually indeed convex as mentioned above.*

The above convex program can be solved by the ellipsoid method to inverse exponential accuracy which is sufficient for the rounding algorithm. Using the iterative randomized swapping algorithm from Section 5.2, we prove the following theorem that matches and generalizes the known results in [6].

Theorem 7.1.1. *Suppose we are given an optimal fractional solution $\mathsf{x} \in [0,1]^n$ to convex programming relaxation (7.1) of the weighted experimental design problem. For any fixed $\varepsilon \leqslant \frac{1}{5}$, if $b_j \geqslant \frac{15d\|\mathsf{c}_j\|_\infty}{\varepsilon^2}$ for all $j \in [m]$, there exists a polynomial time randomized algorithm that returns an integral vector $\mathsf{z} \in \{0,1\}^n$ such that*

$$f\left(\sum_{i=1}^{n} \mathsf{z}(i)\cdot \mathsf{u}_i\mathsf{u}_i^\top\right) \leqslant (1+O(\varepsilon))\cdot f\left(\sum_{i=1}^{n} \mathsf{x}(i)\cdot \mathsf{u}_i\mathsf{u}_i^\top\right), \quad \textit{where } f = f_D \textit{ or } f_A \textit{ or } f_E,$$

with probability at least $1 - e^{-\Omega(d)}$. Furthermore, each knapsack constraint $\langle \mathsf{c}_j, \mathsf{z}\rangle \leqslant b_j$ is satisfied with probability at least $1 - e^{-\Omega(d)}$.

Remark 7.1.2. *More generally, the above theorem holds for any convex objective function f satisfying the following conditions that were suggested in [6].*

- *Monotonicity: For any $\mathsf{A}, \mathsf{B} \in \mathbb{S}^d_{++}$, if $\mathsf{A} \preccurlyeq \mathsf{B}$, then $f(\mathsf{A}) \geqslant f(\mathsf{B})$.*
- *Reciprocal sublinearity: For any $\mathsf{A} \in \mathbb{S}^d_{++}$ and $t \in (0,1)$, it holds that $f(t\mathsf{A}) \leqslant \frac{1}{t} f(\mathsf{A})$.*

We can verify that all f_D, f_A, and f_E satisfy the above conditions.

By slightly modifying the iterative randomized swapping algorithm from Section 5.2, we achieve the following improved results for D/A-design.

Theorem 7.1.3. *Let $\mathsf{x} \in [0,1]^n$ be an optimal fractional solution to the convex programming relaxation (7.1) for D/A-design with knapsack constraints. For any $\varepsilon \leqslant \frac{1}{200}$, if each knapsack constraint budget satisfies $b_j \geqslant \frac{2d\|\mathsf{c}_j\|_\infty}{\varepsilon}$, then there is a randomized exchange algorithm which returns in polynomial time an integral solution $\sum_{i=1}^n \mathsf{z}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top$ with $\mathsf{z}(i) \in \{0,1\}$ for $1 \leqslant i \leqslant n$ such that*

$$\det\left(\sum_{i=1}^n \mathsf{z}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top\right)^{\frac{1}{d}} \geqslant \big(1 - O(\varepsilon)\big) \cdot \det\left(\sum_{i=1}^n \mathsf{x}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top\right)^{\frac{1}{d}} \text{ for D-design,}$$

$$\operatorname{tr}\left(\left(\sum_{i=1}^n \mathsf{z}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top\right)^{-1}\right) \leqslant (1+\varepsilon) \cdot \operatorname{tr}\left(\left(\sum_{i=1}^n \mathsf{x}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top\right)^{-1}\right) \text{ for A-design}$$

with probability at least $1 - O\left(\frac{k^2}{\varepsilon^2} \cdot e^{-\Omega(\sqrt{d})}\right)$ where $k = O(d^2 + m)$. Furthermore, each knapsack constraint $\langle \mathsf{c}_j, \mathsf{z}\rangle \leqslant b_j, j \in [m]$ is satisfied with probability at least $1 - e^{-\Omega(\varepsilon d)}$.

Note that D/A-design with a cardinality constraint is the special case when there is only one cost constraint ($m = 1$) and $c = \mathit{1}$. In this special case, Theorem 7.1.3 improves the previous results in [131, 118] by removing the term $O\left(\frac{1}{\varepsilon^2}\log\left(\frac{1}{\varepsilon}\right)\right)$ from their assumption $b \geqslant \Omega\left(\frac{d}{\varepsilon} + \frac{1}{\varepsilon^2}\log\left(\frac{1}{\varepsilon}\right)\right)$, and this achieves the optimal integrality gap result for D-design [131] and A-design [118]. In the general case with knapsack constraints, Theorem 7.1.3 improves the result of Theorem 7.1.1, which requires a stronger assumption that $b_j \geqslant \Omega\left(\frac{d\|\mathsf{c}_j\|_\infty}{\varepsilon^2}\right)$ to obtain the same approximation guarantee. The knapsack constraints can be used for incorporating fairness constraints in experimental design, which we will discuss in Section 7.1.1.3.

7.1.1.2 Combinatorial Algorithms

The Fedorov's exchange method [60] starts with an arbitrary initial set S_0 of b vectors, and in each step $t \geqslant 1$ it aims to exchange one of the vectors, $S_t \leftarrow S_{t-1} - \mathsf{u}_i + \mathsf{u}_j$ where $\mathsf{u}_i \in S_{t-1}$

and $u_j \notin S_{t-1}$, to improve the objective value, and stops if such an improving exchange is not possible. The simplicity of this algorithm and its good empirical performance [47, 113, 116] make the method widely used [14]. The approximation guarantee of this method is only analyzed rigorously in a recent work [108], and we extend their analysis in multiple directions. We remark that we only consider the unweighted experimental design problems with a single cardinality constraint in the discussions of combinatorial algorithms.

For D-design, it was proved in [108] that Fedorov's exchange method gives a polynomial time approximation algorithm for all inputs in the with repetition setting, and we extend their result to the without repetition setting.

Theorem 7.1.4. *The Fedorov's exchange method is a $\frac{b-d-1}{b}$-approximation polynomial time algorithm for D-design in the without repetition setting. In particular, this is a $(1-\varepsilon)$-approximation algorithm whenever $b \geqslant d+1+\frac{d}{\varepsilon}$ for any $\varepsilon > 0$.*

For A-design, it was shown in [108] that there are arbitrarily bad local optimal solutions for the Fedorov's exchange method. Interestingly, we prove that Fedorov's exchange method works well as long as there exists an almost optimal solution with good condition number. This provides a new insight about when the local search method works well, and this condition may hold in practical instances. As a corollary, this also extends the analysis of Fedorov's exchange method in [108] when all the vectors are short to the without repetition setting (see Section 7.4.2).

Theorem 7.1.5. *Let $X := \sum_{i=1}^{n} x(i) \cdot u_i u_i^\top$ with $\sum_{i=1}^{n} x(i) = b$ and $x_i \in [0,1]$ for $1 \leqslant i \leqslant n$ be a fractional solution to A-design. For any $\varepsilon \in (0,1)$, the Fedorov's exchange method returns an integral solution $Z = \sum_{i=1}^{n} z(i) \cdot u_i u_i^\top$ with $\sum_{i=1}^{n} z(i) \leqslant b$ and $z(i) \in \{0,1\}$ for $1 \leqslant i \leqslant n$ such that*

$$\operatorname{tr}\left(Z^{-1}\right) \leqslant (1+\varepsilon) \cdot \operatorname{tr}(X^{-1}) \quad \textit{whenever} \quad b \geqslant \Omega\Big(\frac{d+\sqrt{\operatorname{tr}(X)\operatorname{tr}\left(X^{-1}\right)}}{\varepsilon}\Big).$$

In particular, let $\kappa = \frac{\lambda_{\max}(X^*)}{\lambda_{\min}(X^*)}$ be the condition number of an optimal solution X^*, then the Fedorov's exchange method gives a $(1+\varepsilon)$-approximation algorithm for A-design whenever $b \geqslant \Omega\Big(\frac{\left(1+\sqrt{\kappa}\right)\cdot d}{\varepsilon}\Big)$, and the time complexity is polynomial in $n, d, \frac{1}{\varepsilon}, \kappa$.

For E-design, there are no known combinatorial local search algorithms, and there are examples showing that Fedorov's exchange method does not work even if there exists a well-conditioned optimal solution (see Section 7.4.3.2). Using the regret minimization framework in [7, 6], however, we prove that a modified local search algorithm using a "smoothed" objective function for E-design works as long as there exists an almost optimal solution with good condition number.

Theorem 7.1.6. *Let $X := \sum_{i=1}^{n} x(i) \cdot u_i u_i^\top$ with $\sum_{i=1}^{n} x(i) = b$ and $x(i) \in [0,1]$ for $1 \leqslant i \leqslant n$ be a fractional solution to E-design. For any $\varepsilon \in (0,1)$, there is a combinatorial local search algorithm which returns an integral solution $Z = \sum_{i=1}^{n} z(i) \cdot u_i u_i^\top$ with $\sum_{i=1}^{n} z(i) \leqslant b$ and $z(i) \in \{0,1\}$ for $1 \leqslant i \leqslant n$ such that*

$$\lambda_{\min}(Z) \geqslant (1 - O(\varepsilon)) \cdot \lambda_{\min}(X) \quad \textit{whenever} \quad b \geqslant \Omega\left(\frac{d}{\varepsilon^2}\sqrt{\frac{\lambda_{\text{avg}}(X)}{\lambda_{\min}(X)}}\right),$$

where $\lambda_{avg}(X) = \frac{\operatorname{tr}(X)}{d}$ is the average eigenvalue of X.

In particular, let $\kappa = \frac{\lambda_{\max}(X^*)}{\lambda_{\min}(X^*)}$ be the condition number of an optimal solution X^*, then the combinatorial local search method gives a polynomial time $(1-\varepsilon)$-approximation algorithm for E-design whenever $b \geqslant \Omega\left(\frac{d\sqrt{\kappa}}{\varepsilon^2}\right)$, and the time complexity is polynomial in $n, d, \frac{1}{\varepsilon}, \kappa$.

A combinatorial "capping" procedure was used in [108] to reduce the A-design problem to the case when every vector is "short", for which Fedorov's exchange method works. This capping procedure, however, crucially leveraged that a vector can be chosen multiple times. We do not have a preprocessing procedure to reduce A-design and E-design in the without repetition setting to the case when Theorem 7.1.5 and Theorem 7.1.6 apply. We leave it as an open problem to design a fully combinatorial algorithm for A-design and E-design in the general case.

7.1.1.3 Some Applications

We discuss some applications of our results in specific instances of experimental design problems.

Fair and Diverse Data Summarization: In the data summarization problem, we are given n data points $\mathsf{u}_1, \ldots, \mathsf{u}_n \in \mathbb{R}^d$, and the objective is to choose a subset of b data points that provides a "fair" and "diverse" summary of the data. For diversity, the D-design objective of maximizing determinant is a popular measure used in previous work [114, 34]. For fairness, the partition constraints [117, 34] for D-design are used to partition the set X of data points into p disjoint groups $X_1 \cup \cdots \cup X_p$ and to ensure that b_i data points are chosen in X_i where $\sum_{i=1}^{p} b_i = b$.

We believe that Theorem 7.1.3 for D-design with knapsack constraints provides an alternative solution for this problem. The main advantage is that the knapsack constraints are more flexible in that they do not require the groups to be disjoint. For instance, we can have knapsack constraints on arbitrary subsets $X_1, \ldots, X_p \subseteq X$ of the form $\sum_{j \in X_i} \mathsf{x}(j) \leqslant b_i$ to ensure that at most b_i data points are chosen in group X_i, so that we can handle constraints of overlapping groups such as race, age, gender (e.g., at most 50% of the chosen vectors correspond to men/women), etc. Also, the approximation guarantee in Theorem 7.1.3 is stronger than the constant factor approximation for D-design with partition constraint [117], and the convex programming relaxation used in Theorem 7.1.3 is simpler and easier to be solved than the more sophisticated one used in [117].

Minimizing Total Effective Resistance: Ghosh, Boyd and Saberi [70] studied the problem of choosing a subgraph with at most b edges to minimize the total effective resistance, and showed that this is a special case of A-design (see Section 6.2.2 for more background about the problem). The proportional volume sampling algorithm by Nikolov, Singh and Tantipongpipat [118] achieves a $(1+\varepsilon)$-approximation for this problem when $b \geqslant \Omega(\frac{n}{\varepsilon} + \frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon})$ where n is the number of vertices in the graph. In Section 6.2.2, we considered the weighted problem of choosing a subgraph with total edge cost at most b to minimize the total effective resistance, and gave a $(1+\varepsilon)$-approximation algorithm when $b \geqslant \Omega\left(\frac{n\|\mathsf{c}\|_\infty}{\varepsilon^2}\right)$ where c is the cost vector of the edges. Theorem 7.1.3 improves these two results.

Corollary 7.1.7. *For any $0 < \varepsilon < 1$, there is a polynomial time randomized $(1+\varepsilon)$-approximation algorithm for minimizing total effective resistance in an edge weighted graph whenever $b \geqslant \Omega\left(\frac{n\|\mathsf{c}\|_\infty}{\varepsilon}\right)$.*

Maximizing Algebraic Connectivity: Ghosh and Boyd [69] studied the problem of choosing a subgraph with total cost at most b that maximizes the algebraic connectivity, i.e. the second smallest eigenvalue of its Laplacian matrix. Kolla, Makarychev, Saberi and Teng [86] provided the first algorithm with non-trivial approximation guarantee in the zero-one cost setting. In Section 6.2.1, we observed that this is a special case of E-design and gave a $(1-\varepsilon)$-approximation algorithm when $b \geqslant \Omega\left(\frac{n\|c\|_\infty}{\varepsilon^2}\right)$ where c is the cost vector of the edges.

All previous results are based on convex programming. Theorem 7.1.6 provides a combinatorial algorithm for the unweighted problem, where the goal is to choose b edges to maximize the algebraic connectivity, and shows that it has a good performance as long as the optimal value is large.

Corollary 7.1.8. *For any $0 < \varepsilon < 1$, there is a polynomial time combinatorial $(1-\varepsilon)$-approximation algorithm for maximizing algebraic connectivity in an unweighted graph whenever $b \geqslant \Omega\left(\frac{n}{\varepsilon^4 \lambda_2^*}\right)$, where λ_2^* is the optimal value for the problem.*

7.1.2 Technical Overview

Rounding Algorithm for Weighted Experimental Design: In [6], Allen-Zhu, Li, Singh, and Wang observed that the experimental design problem with a general convex objective function satisfying conditions in Remark 7.1.2 and a single cardinality constraint can be reduced to the one-sided spectral rounding problem with uniform cost. More specifically, they first solve the natural convex programming relaxation for experimental design with a cardinality constraint, and obtain a solution $x \in \mathbb{R}^n$. After performing a normalization transformation to turn the input vectors u_i's into v_i's such that $\sum_{i=1}^n x(i) \cdot v_i v_i^\top = I$, they reduce the problem to one-sided spectral rounding problem with uniform cost. They use a deterministic greedy algorithm to prove Theorem 5.1.1 to solve the problem (see Section 5.1 for more details).

To solve the experimental design problem with multiple knapsack constraints, we use the same reduction. We first obtain an optimal solution x to the convex programming relaxation (7.1). Then reduce the problem to the one-sided spectral rounding problem with general cost. Finally, we use the iterative randomized swapping algorithm in Section 5.2 to solve the problem. We remark that the random sampling idea for the rounding is a key to satisfy the knapsack constraints and achieve good approximation guarantee simultaneously. See Section 7.2 for more details about the reduction.

Improved Analysis for D/A-Design: For the rounding algorithm for D/A-design with knapsack constraints, surprisingly we prove that a minor modification of the iterative randomized swapping algorithm in Section 5.2 would work with improved approximation guarantees! Essentially, we just use the same algorithm but only require that the solution to have minimum eigenvalue $\frac{3}{4}$ rather than $1-\varepsilon$. Our analysis has two phases. In the first phase, using the results in Section 5.2, we show that the randomized exchange algorithm will find a solution with minimum eigenvalue at least $\frac{3}{4}$ in polynomial time with high probability whenever $b \geqslant \Omega\left(\frac{d}{\varepsilon}\right)$ (rather than $b \geqslant \Omega\left(\frac{d}{\varepsilon^2}\right)$ in order to achieve minimum eigenvalue at least $1-\varepsilon$). In the second phase, we prove that the minimum eigenvalue will maintain to be at least $\frac{1}{4}$ with high probability when ε is not too tiny, and then the objective value for D-design and A-design will improve to $(1 \pm \varepsilon)$ times the optimal objective value in polynomial time with high probability. The condition that the minimum eigenvalue is at least $\frac{1}{4}$ is used crucially in multiple places for the analysis of the second phase. Interestingly, it is used in showing that the same sampling probability distributions in the iterative randomized swapping algorithm (which aim at improving the E-design objective) are also good for improving the objective value for D-design and A-design. Moreover, it is crucially used in the martingale concentration argument, e.g., to show that the martingale is bounded and to prove upper bounds on the variance of the changes. For the martingale concentration argument, we also use the optimality conditions for convex programs to prove that the vectors with fractional value are "short" in order to bound the quantities involved. Overall, the analysis for the rounding algorithm is quite involved, but it provides a unifying algorithm to achieve the optimal results for the natural convex relaxations for D/A/E-design. Please refer to Section 7.3 for a more detailed outline of the analysis.

Analysis of Combinatorial Algorithms: In the last part of this chapter, we use the randomized approach in Section 5.2 to analyze combinatorial algorithms. For combinatorial local search algorithms, one difference from the previous analysis in [108] is that we compare the objective of the current integral solution to that of an optimal *fractional* solution. When the objective value of the fractional solution is considerably better than that of the current integral solution, we use the fractional solution to define appropriate probability distributions similar to that in the iterative randomized sampling algorithm to sample i_t and j_t so that the expected objective value of $S_t \leftarrow S_{t-1} - u_{i_t} + u_{j_t}$ improves, and this would imply the existence of an improving pair in Fedorov's exchange method. One advantage of this approach is that this allows us having the flexibility to compare the current integral solution to a fractional solution with smaller budget which still has its objective value close to the optimal one.

Our analysis is arguably simpler than that in [108] which uses a dual fitting method while we only do a primal analysis. More importantly, our analysis shows that if the optimal fractional solution is well-conditioned (e.g., $\sum_{i=1}^{n} x(i) \cdot u_i u_i = I$), then the Fedorov's exchange method indeed performs as well as the best known rounding algorithms. This gives us a new insight that the only important step in rounding algorithms for the unweighted experimental design problems is the ability to first transform the optimal fractional solution to the identity matrix. For E-design, simply doing Fedorov's exchange method on the objective function $\lambda_{\min}\left(\sum_{i \in S_t} u_i u_i^{\top}\right)$ would not work (see Section 7.4.3.2), and instead we apply the Fedorov's exchange method to the potential function in the regret minimization framework, which is morally the same as the potential function $\operatorname{tr}\left((\sum_{i \in S_t} u_i u_i^{\top} - l I_d)^{-1}\right)$ used by Batson, Spielman and Srivastava for spectral sparsification [21].

7.1.3 Previous Work

The D/A/E experimental design problems are NP-hard [33, 142] and also APX-hard [136, 118, 33]. Despite the long history and the wide interest, strong approximation algorithms for these problems are only obtained recently.

D-design: Singh and Xie [131] designed an $(1-\varepsilon)$-approximation algorithm for D-design in the with repetition setting when $b \geqslant \frac{2d}{\varepsilon}$, and in the without repetition setting when $b = \Omega\left(\frac{d}{\varepsilon} + \frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$. Their algorithm is by rounding an optimal solution to a natural convex program relaxation using approximate positively correlated distributions.

Madan, Singh, Tantipongpipat and Xie [108] analyzed the Fedorov's exchange method and proved that it gives an $(1-\varepsilon)$-approximation algorithm for D-design as long as $b \geqslant d+\frac{d}{\varepsilon}$, which improves upon the above result. However, they only provide a polynomial time implementation of the local search algorithm to achieve this guarantee in the less general with repetition setting.

A-design: Nikolov, Singh and Tantipongpipat [118] designed an $(1+\varepsilon)$-approximation algorithm for A-design in the with repetition setting when $b \geqslant d + \frac{d}{\varepsilon}$, and in the without repetition setting when $b = \Omega\left(\frac{d}{\varepsilon} + \frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$. Their algorithm is by rounding an optimal solution to a natural convex program relaxation using proportional volume sampling. Their algorithm also works for D-design with the same guarantee.

Madan, Singh, Tantipongpipat and Xie [108] also analyzed the Fedorov's exchange method for A-design, and showed that there are arbitrarily bad local optimal solutions. On the other hand, they proved that Fedorov's exchange method works when all the input vectors are "short", and they designed a "capping procedure" to reduce the general case to the case when all vectors are short. As a result, they obtained a combinatorial $(1+\varepsilon)$-approximation algorithm, without solving convex programs, for A-design when $b \geqslant \Omega\left(\frac{d}{\varepsilon^4}\right)$ in the with repetition setting.

E-design: Allen-Zhu, Li, Singh and Wang [5, 6] designed an $(1-\varepsilon)$-approximation algorithm for E-design in the with and without repetition settings when $b \geqslant \Omega\left(\frac{d}{\varepsilon^2}\right)$. Their algorithm is by rounding an optimal solution to a natural convex program relaxation using the regret minimization framework, which was initially developed for the spectral sparsification problem [7]. They formulated and solved a "one-sided spectral rounding problem" (see Section 5.1), and showed that experimental design with any objective function satisfying some mild regularity assumptions, including D/A/E-design, can be reduced to the one-sided spectral rounding problem. Their algorithm for one-sided spectral rounding can be viewed as a local search algorithm, and this was the starting point of the current work in this chapter.

Nikolov, Singh and Tantipongpipat [118] showed that the assumption $b \geqslant \Omega\left(\frac{d}{\varepsilon^2}\right)$ is necessary to achieve $(1-\varepsilon)$-approximation for E-design using the natural convex program, and in Section 5.4 we showed that the assumption $b \geqslant \Omega\left(\frac{d}{\varepsilon^2}\right)$ is necessary for the one-sided spectral rounding problem. These suggest that the regret minimization framework may not be used to match the results for D/A-design, but we bypass the one-sided spectral rounding problem to prove Theorem 7.1.3.

Experimental design with additional constraints: Using more sophisticated convex programming relaxations, Nikolov and Singh [117] designed an approximation algorithm for D-design under partition constraints. Recently, Madan, Nikolov, Singh and Tantipongpipat [107] designed an approximation algorithm for D-design under general matroid constraints.

Organization

We first show how to use the one-sided spectral rounding to solve weighted experimental design problems in Section 7.2. Then, we present slightly modified rounding algorithms and analysis for D/A-Design in Section 7.3. Finally, we present our results about combinatorial algorithms in Section 7.4.

7.2 Rounding Algorithm for Weighted Experimental Design

In this section, we use a simple black box reduction to reduce the weighted experimental design problems to the one-sided spectral rounding problem and prove Theorem 7.1.1.

Theorem 7.1.1. *Suppose we are given an optimal fractional solution $\mathsf{x} \in [0,1]^n$ to convex programming relaxation (7.1) of the weighted experimental design problem. For any fixed $\varepsilon \leqslant \frac{1}{5}$, if $b_j \geqslant \frac{15d\|\mathsf{c}_j\|_\infty}{\varepsilon^2}$ for all $j \in [m]$, there exists a polynomial time randomized algorithm that returns an integral vector $\mathsf{z} \in \{0,1\}^n$ such that*

$$f\left(\sum_{i=1}^{n} \mathsf{z}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top\right) \leqslant (1+O(\varepsilon)) \cdot f\left(\sum_{i=1}^{n} \mathsf{x}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top\right), \quad \textit{where } f = f_D \textit{ or } f_A \textit{ or } f_E,$$

with probability at least $1 - e^{-\Omega(d)}$. *Furthermore, each knapsack constraint* $\langle c_j, z\rangle \leqslant b_j$ *is satisfied with probability at least* $1 - e^{-\Omega(d)}$.

Proof. Let $X := \sum_{i=1}^{n} x(i) \cdot u_i u_i^\top \succ 0$. We do the following transformation

$$v_i := X^{-\frac{1}{2}} u_i \qquad \text{for all } i \in [n],$$

so that $\sum_{i=1}^{n} x(i) \cdot v_i v_i^\top = I_d$.

Then, the idea is similar to the one in the proof of Corollary 5.2.3, where we scale down x and apply Theorem 5.2.2. Let $\eta = 1 - 2\varepsilon$ and set $y := \eta x$ and $w_i := \frac{1}{\sqrt{\eta}} v_i$ such that

$$\sum_{i=1}^{m} y(i) \cdot w_i w_i^\top = \sum_{i=1}^{m} x(i) \cdot v_i v_i^\top = I_d \quad \text{and} \quad \langle c_j, y\rangle = \eta \langle c_j, x\rangle \leqslant \eta b_j \quad \text{for all } j \in [m].$$

We run the iterative randomized swapping algorithm and apply Theorem 5.2.2 on the vectors $w_1, \ldots, w_m$ and y with $\delta_1 = \gamma = \varepsilon, q = \sqrt{d}$ to obtain a $z \in \{0,1\}^m$ so that

$$\begin{aligned} \sum_{i=1}^{m} z(i) \cdot w_i w_i^\top \succcurlyeq (1 - 2\varepsilon) I_d \quad &\implies \quad \sum_{i=1}^{m} z(i) \cdot v_i v_i^\top \succcurlyeq \eta(1 - 2\varepsilon) I_d \succcurlyeq (1 - 4\varepsilon) I_d \\ &\implies \quad \sum_{i=1}^{m} z(i) \cdot u_i u_i^\top \succcurlyeq (1 - 4\varepsilon) X \end{aligned}$$

The failure probability of this event is at most $e^{-\Omega(d)}$. Since the objective function f satisfies the monotonicity and the reciprocal sublinearity conditions mentioned in Remark 7.1.2, it follows that

$$f\left(\sum_{i=1}^{n} z(i) \cdot u_i u_i^\top\right) \leqslant f((1 - 4\varepsilon) X) \leqslant (1 + O(\varepsilon)) \cdot f\left(\sum_{i=1}^{n} x(i) \cdot u_i u_i^\top\right).$$

For each knapsack constraint $c_j \in \mathbb{R}_+^m$ with $b_j \geqslant \frac{15d\|c_j\|_\infty}{\varepsilon^2}$, Theorem 5.2.2 implies that

$$\langle c_j, z\rangle \leqslant (1 + \varepsilon)\langle c_j, y\rangle + \frac{15d\,\|c_j\|_\infty}{\varepsilon} \leqslant (1 + \varepsilon)(1 - 2\varepsilon) b_j + \varepsilon b_j < b_j,$$

where we used $\langle c_j, y\rangle \leqslant \eta b_j$ and the assumption that $\frac{15d\|c_j\|_\infty}{\varepsilon^2} \leqslant b_j$ for all $j \in [m]$. The failure probability is at most $e^{-\Omega(d)}$. □

7.3 Improved Analysis for D/A-Design

In this section, we propose the following randomized exchange algorithm to solve the D/A-design problems with knapsack constraints and improve the approximation guarantee in Theorem 7.1.1.

```
Randomized Exchange Algorithm

Input: $n$ vectors $u_1, \ldots, u_n \in \mathbb{R}^d$, an accuracy parameter $\varepsilon \in (0,1)$, and $m$ knapsack
       constraints $c_j \in \mathbb{R}^n_+$ with budgets $b_j \geqslant \frac{d\|c_j\|_\infty}{\varepsilon}$ for all $j \in [m]$.

    1. Solve the convex programming relaxation (7.1) for D-design or A-design and
       obtain an optimal solution $x \in [0,1]^n$ with at most $d^2 + m$ fractional entries,
       i.e. $|\{i \in [n] \mid 0 < x(i) < 1\}| \leq d^2 + m$. Let $X = \sum_{i=1}^n x(i) \cdot u_i u_i^\top$.
    2. Preprocessing: Let $v_i \leftarrow X^{-\frac{1}{2}} u_i$ for all $i \in [n]$, so that $\sum_{i=1}^n x(i) \cdot v_i v_i^\top = I_n$.
    3. Initialization: $S_0 \leftarrow \emptyset$, $\alpha \leftarrow 8\sqrt{d}$, and $k \leftarrow 16d + d^2 + m$.
    4. Add $i$ into $S_0$ independently with probability $x(i)$ for each $i \in [n]$.
    5. Let $Z_1 \leftarrow \sum_{i \in S_0} v_i v_i^\top$ and $t \leftarrow 1$.
    6. While the termination condition is not satisfied and $t = O\left(\frac{k}{\varepsilon}\right)$ do the following,
       where the termination conditions for D-design and A-design are respectively

           $\det(Z_t)^{\frac{1}{d}} \geqslant 1 - 10\varepsilon$   and   $\langle X^{-1}, Z_t^{-1} \rangle \leqslant (1+\varepsilon)\,\mathrm{tr}(X^{-1})$.

        (a) $S_t \leftarrow \mathrm{Exchange}(S_{t-1})$.
        (b) Set $Z_{t+1} \leftarrow \sum_{i \in S_t} v_i v_i^\top$ and $t \leftarrow t+1$.
    7. Return $S_{t-1}$ as the solution.
```

The exchange subroutine is described as follows.

Exchange Subroutine

1. Compute the action matrix $A_t \leftarrow (\alpha Z_t - l_t I)^{-2}$, where $Z_t = \sum_{i \in S_{t-1}} v_i v_i^\top$ and l_t is the unique scalar such that $A_t \succ 0$ and $\mathrm{tr}(A_t) = 1$.
2. Let $S'_t \leftarrow \{i \in S_{t-1} \mid 2\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle \leqslant \frac{1}{2}\}$.
3. Sample $i_t \in S'_{t-1}$ from the following probability distribution

$$\begin{aligned} \mathbb{P}[i_t = i] &= \frac{1 - x(i)}{k} \cdot \left(1 - 2\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle\right), \text{ for } i \in S'_{t-1} \text{ and} \\ \mathbb{P}[i_t = \emptyset] &= 1 - \sum_{i \in S'_{t-1}} \frac{1 - x(i)}{k} \cdot \left(1 - 2\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle\right). \end{aligned}$$

4. Sample $j_t \in [n] \backslash S_{t-1}$ from the following probability distribution

$$\begin{aligned} \mathbb{P}[j_t = j] &= \frac{x(j)}{k} \cdot \left(1 + 2\alpha\langle v_j v_j^\top, A_t^{\frac{1}{2}}\rangle\right), \text{ for } j \in [n] \backslash S_{t-1} \text{ and} \\ \mathbb{P}[j_t = \emptyset] &= 1 - \sum_{j \in [n] \backslash S_{t-1}} \frac{x(j)}{k} \cdot \left(1 + 2\alpha\langle v_j v_j^\top, A_t^{\frac{1}{2}}\rangle\right). \end{aligned}$$

5. Return $S_t \leftarrow S_{t-1} \cup \{j_t\} \backslash \{i_t\}$.

Remark 7.3.1. *The randomized exchange algorithm is almost the same as the iterative randomized swapping algorithm in Section 5.2. There are only two differences. One is that $\alpha \leftarrow 8\sqrt{d}$ instead of $\alpha \leftarrow \frac{\sqrt{d}}{\gamma}$ in Section 5.2. The other is that the termination condition $\lambda_{\min}(Z_t) \geqslant 1 - 2\gamma$, is replaced by the termination condition for D-design or the termination condition for A-design.*

The parameter α is used to control the approximation guarantee of the iterative randomized swapping algorithm. If the termination condition is $\lambda_{\min}(Z_t) \geqslant \frac{3}{4}$, then it is proved in Theorem 5.2.7 that the algorithm will terminate successfully in $O(k)$ steps with high probability.

Intuition and Proof Ideas

Based on the above remark, we see that the sampling distribution in the exchange subroutine is actually designed for the one-sided spectral rounding problem (or E-design). It is quite surprising that it also works for D/A-design.

We will use some other sampling distributions to analyze the combinatorial algorithms (Fedorov's exchange method) for D-design and A-design in Section 7.4. It appears at first sight that those distributions are more natural to use in the exchange subroutine for D/A-design.

Here, we use D-design to illustrate the difficulty of analyzing the natural distributions $\mathbb{P}[i_t = i] \propto 1 - x_i$ and $\mathbb{P}[j_t = j] \propto x(j)$ in Section 7.4.1 and to motivate the modifications made in the randomized exchange algorithm. By applying Lemma 2.1.12 repeatedly, for any $\tau \geqslant 1$,

$$\det(Z_{\tau+1}) \geqslant \det(Z_1) \cdot \prod_{t=1}^{\tau} \left(1 - v_{i_t}^\top Z_t^{-1} v_{i_t}\right) \left(1 + v_{j_t}^\top Z_t^{-1} v_{j_t}\right).$$

Using the natural distributions in Section 7.4.1, Lemma 7.4.1 and Lemma 7.4.2 shows that there exist $i_t \in S_{t-1}$ and $j_t \notin S_{t-1}$ such that setting $S_t \leftarrow S_{t-1} - i_t + j_t$ will improve the D-design objective in each iteration. However, if we randomly sample i_t and j_t from these distributions, we cannot prove that the objective value is consistently improving with good probability. For D-design, we are analyzing a product of random variables where each random variable could have a large variance, and existing martingale inequalities are not applicable to establish concentration of the product.

To bound the variance, one important observation is that when x is an optimal fractional solution, it follows from the optimality condition of the convex programming relaxation that any vector v_i with $x(i) \in (0, 1)$ satisfies $\|v_i\|_2^2 \leqslant \varepsilon$. The current algorithm is motivated by the observation that if we can also lower bound the minimum eigenvalue of Z_t, then we can upper bound $v^\top Z_t^{-1} v$ and this would allow us to establish concentration of the objective value. So our idea is to use the same algorithm in Section 5.2 to ensure that the minimum eigenvalue of Z_t is at least $\Omega(1)$ as mentioned in Remark 7.3.1. Surprisingly, we prove that sampling from the distributions for E-design can also improve the objective values for D-design and A-design, and this is particularly interesting for A-design where the minimum eigenvalue condition is needed to prove so. Having these in place, we can use Freedman's martingale inequality to prove that the objective values for D-design and A-design will be improving consistently if the minimum eigenvalue of the current solution is at least $\Omega(1)$.

Proof Outline and Organization

In the analysis of the randomized exchange algorithm, we conceptually divide the algorithm into two phases. In the first phase, we show that the minimum eigenvalue of the current solution will reach $\frac{3}{4}$ in $O(k)$ iterations with high probability. In the second phase, we prove that the objective value for D/A-design will be a $(1 \pm \varepsilon)$-approximation of the optimal in $O\left(\frac{k}{\varepsilon}\right)$ iterations with high probability. The following is an outline of the proof steps.

1. In Section 7.3.1.1, we first prove that the randomized exchange algorithm is well-defined. In particular, we show that a fractional optimal solution to the convex relaxation (7.1) with at most $O(d^2 + m)$ fractional entries can be found in polynomial time, and the probability distributions in the exchange subroutine are well-defined for $k = O(d^2 + m)$.

2. In Section 7.3.1.2, we prove that the minimum eigenvalue will reach $\frac{3}{4}$ in $O(k)$ iterations with high probability. Furthermore, the minimum eigenvalue will be at least $\frac{1}{4}$ during the next $\Theta\left(\frac{k}{\varepsilon}\right)$ iterations with good probability, for which we require the assumption that ε is not too small. The proofs are based on the regret minimization framework [7, 6] and the iterative randomized swapping algorithm in Section 5.2.

3. In Section 7.3.2 and Section 7.3.3, we prove that the objective value of D-design and A-design will improve consistently with high probability. These are the more technical parts of the proof. We use the minimum eigenvalue condition in multiple places, both in the martingale concentration arguments for D/A-design and in the expected improvement of the A-design objective.

4. In Section 7.3.1.3, we prove the main approximation results including Theorem 7.1.3 for experimental design, by combining the previous steps and using the concentration inequality for the knapsack constraints proved in Section 5.2.3. As a corollary, we slightly improve the previous results of D/A-design with a single cardinality constraint in [131, 118]. We also prove Corollary 7.1.7 as an application of the main result.

7.3.1 Analysis of the Common Algorithm

The algorithm is identical for D-design and A-design except the termination condition. In this subsection, we will present the proofs of the common parts and the main results, and then present the specific proofs for D-design and A-design in Section 7.3.2 and Section 7.3.3 respectively.

7.3.1.1 Sparse Optimal Solution and Probability Distributions in the Exchange Subroutine

In this subsection, we first show that we can find an optimal fractional solution to the convex programming relaxation (7.1) with sparse support in polynomial time. The sparsity of an optimal solution to the convex program (7.1) was proved and used in [140, 107] for experimental design problems. The following lemma is proved using similar ideas.

Lemma 7.3.2. *Given any feasible fractional solution $\hat{\mathsf{x}}$ to the convex relaxation (7.1), there exists another feasible fractional solution x with $|\{i \in [n] \mid 0 < \mathsf{x}(i) < 1\}| \leq d^2 + m$ such that*

$$\det\left(\sum_{i=1}^{n} \mathsf{x}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top\right) = \det\left(\sum_{i=1}^{n} \hat{\mathsf{x}}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top\right) \quad \textit{for D-Design, or}$$

$$\operatorname{tr}\left(\left(\sum_{i=1}^{n} \mathsf{x}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top\right)^{-1}\right) = \operatorname{tr}\left(\left(\sum_{i=1}^{n} \hat{\mathsf{x}}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top\right)^{-1}\right) \quad \textit{for A-Design.}$$

Furthermore, the solution x can be found in polynomial time.

Proof. Given the feasible fractional solution $\hat{x}$, we compute an extreme point solution x to the following polytope, which can be done in polynomial time.

$$\begin{cases} \sum_{i=1}^{n} x(i) \cdot u_i u_i^\top = \sum_{i=1}^{n} \hat{x}(i) \cdot u_i u_i^\top, \\ \langle c_j, x \rangle \leq b_j, \quad \text{for } 1 \leq j \leq m, \\ 0 \leq x(i) \leq 1, \quad \text{for } 1 \leq i \leq n. \end{cases}$$

In the extreme point solution x, the number of variables is equal to the number of linearly independent tight constraints attained by x. Clearly, the number of integral variables in x is equal to the number of linear independent tight constraints in $0 \leq x(i) \leq 1$ for $1 \leq i \leq n$ attained by x. So, the number of fractional variables in x is equal to the number of linear independent tight constraints in $\sum_{i=1}^{n} x(i) \cdot u_i u_i^\top = \sum_{i=1}^{n} \hat{x}(i) \cdot u_i u_i^\top$ and $\langle c_j, x \rangle \leq b_j$ for $1 \leq j \leq m$ attained by x. As there are only $d^2 + m$ such constraints in the above linear program, there are at most $d^2 + m$ fractional entries in x. Due to the first matrix equality constraint of the polytope, x and $\hat{x}$ have the same objective value. □

Then, we make a simple observation of the randomized exchange algorithm, that only vectors with fractional entries will be exchanged, as those vectors with $x(i) = 1$ will always be in the solution and vectors with $x(i) = 0$ will always not be in the solution.

Observation 7.3.3. *For any $t \geq 0$, it holds that $i \in S_t$ for all i with $x(i) = 1$ and $j \in [n] \backslash S_t$ for all j with $x(j) = 0$. This further implies that $\mathbb{P}(i_t = i) = 0$ for all i with $x(i) \in \{0, 1\}$ and $\mathbb{P}(j_t = j) = 0$ for all j with $x(j) \in \{0, 1\}$.*

Proof. The observation follows as all vectors with $x(i) = 1$ are selected and all vectors with $x(j) = 0$ are not selected initially. In each iteration, the probability distributions in the exchange subroutine guarantee that vectors with $x(i) = 1$ have zero probability to be removed from the solution set, and vectors with $x(j) = 0$ have zero probability to be added into the solution set. Therefore, the exchange subroutine of the algorithm would only exchange those vectors with fractional entries $x(i)$'s. □

Finally, we are ready to show that the probability distributions in the exchange subroutine are well-defined for $k = O(d^2 + m)$, which will be used to upper bound the number of iterations and the failure probability of the algorithm.

Claim 7.3.4. *The probability distributions at any t-th iteration of the randomized exchange algorithm are well-defined for $k = 16d + d^2 + m$.*

Proof. First, we verify that the probability distribution for sampling i_t is well-defined. We need to show that $\mathbb{P}(i_t = i) \geq 0$ for $i \in S'_{t-1}$ and $\sum_{i \in S'_{t-1}} \mathbb{P}(i_t = i) \leq 1$. Since $A_t \succ 0$ and $x_i \in [0, 1]$ and $2\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle \leq 1/2$ for $i \in S'_{t-1}$, it holds for $i \in S'_{t-1}$ that

$$0 \leq \mathbb{P}(i_t = i) = \frac{1}{k}(1 - x(i))(1 - 2\alpha\langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle) \leq \frac{1}{k}.$$

Thus, $\sum_{i \in S'_{t-1}} \mathbb{P}(i_t = i) \leq \frac{1}{k}|\{i \in [n] \mid 0 < x(i) < 1\}| < 1$, where the first inequality follows by Observation 7.3.3, and the second inequality follows by the the choice of $k = 16d + d^2 + m$ and Lemma 7.3.2.

Next, we verify that the probability distribution for sampling j_t is well-defined. It is clear that $\mathbb{P}(j_t = j) \geq 0$ as $A_t \succ 0$ and $x(j) \in [0, 1]$. Then, we consider

$$\begin{aligned}\sum_{j \in [n] \setminus S_{t-1}} \mathbb{P}(j_t = j) &= \frac{1}{k} \sum_{j \in [n] \setminus S_{t-1}} x(j) \cdot \left(1 + 2\alpha\langle v_j v_j^\top, A_t^{\frac{1}{2}}\rangle\right) \\ &\leq \frac{1}{k}\left(\sum_{j \in [n] \setminus S_{t-1}} x(j) + 2\alpha \operatorname{tr}\left(A_t^{\frac{1}{2}}\right)\right),\end{aligned}$$

where the inequality is by $\sum_{j=1}^{n} x(j) \cdot v_j v_j^\top = I_d$. Notice that $\sum_{j \in [n] \setminus S_{t-1}} x(j) \leq |\{i \in [n] \mid 0 < x(i) < 1\}| \leq d^2 + m$ by Observation 7.3.3 and Lemma 7.3.2. Thus,

$$\sum_{j \in [n] \setminus S_{t-1}} \mathbb{P}(j_t = j) \leq \frac{1}{k}\left(d^2 + m + 2\alpha \operatorname{tr}(A_t^{\frac{1}{2}})\right) \leq \frac{1}{k}(d^2 + m + 16d) \leq 1,$$

where the second last inequality is by $\alpha = 8\sqrt{d}$ and $\operatorname{tr}\left(A_t^{\frac{1}{2}}\right) \leq \sqrt{d}$ from Claim 2.1.10, and the last inequality is by the choice of k. □

Combining Lemma 7.3.2 and Claim 7.3.4, we have shown that the randomized exchange algorithm is well-defined.

7.3.1.2 Lower Bounding Minimum Eigenvalue

As discussed above, the minimum eigenvalue of Z_t plays a key role in our analysis of the algorithm. We conceptually divide the execution of the randomized exchange algorithm into two phases. In the first phase, we show that the minimum eigenvalue of the current solution will reach $\frac{3}{4}$ in $O(k)$ iterations with high probability.

Proposition 7.3.5. *The probability that the randomized exchange algorithm has terminated successfully within $16k$ iterations or there exists $\tau_1 \leqslant 16k$ with $\lambda_{\min}(Z_{\tau_1}) \geqslant \frac{3}{4}$ is at least $1 - \exp(-\Omega(\sqrt{d}))$.*

Proof. As noted in Remark 7.3.1, except for the termination condition, the randomized exchange algorithm is exactly the same as the algorithm in Section 5.2 with $\alpha = 8\sqrt{d}$. So, the proposition follows from Theorem 5.2.7 with $\gamma = \frac{1}{8}$ and $q = 2$. □

Recall that Proposition 5.2.8 shows that with good probability the minimum eigenvalue of Z_t remains at least $\frac{1}{4}$ for a period of time after hitting $1 - 2\gamma$.

Proposition 5.2.8. *Suppose $0 < \gamma \leqslant \frac{1}{8}$. Assume there is no termination condition in the iterative randomized swapping algorithm, and the minimum eigenvalue hit the target $\lambda_{\min}(Z_{\tau_1}) \geqslant 1 - 2\gamma$ at some time step τ_1, then the probability that $\lambda_{\min}(Z_t) \geqslant \frac{1}{4}$ for all the next τ steps $\tau_1 \leqslant t \leqslant \tau_1 + \tau$ is at least $1 - \tau^2 \cdot e^{-\Omega(\sqrt{d})}$.*

Apply Proposition 5.2.8 with $\gamma = \frac{1}{8}$ and $\tau = \frac{2k}{\varepsilon}$, we obtain the following corollary.

Corollary 7.3.6. *Suppose $\lambda_{\min}(Z_{\tau_1}) \geqslant \frac{3}{4}$ for some τ_1. In the randomized exchange algorithm, the probability that $\lambda_{\min}(Z_t) \geqslant \frac{1}{4}$ for all $\tau_1 \leqslant t \leqslant \tau_1 + \frac{2k}{\varepsilon}$ is at least $1 - \frac{4k^2}{\varepsilon^2} \cdot e^{-\Omega(\sqrt{d})}$.*

7.3.1.3 Main Approximation Results

In this subsection, we prove the main approximation results for experimental design, including Theorem 7.1.3. We will do so by first assuming the following theorem about the improvement of the objective value in the second phase, which will be proved in Section 7.3.2 for D-design and in Section 7.3.3 for A-design.

Theorem 7.3.7. *Suppose that* $\lambda_{\min}(Z_{\tau_1}) \geqslant \frac{3}{4}$ *and* $\lambda_{\min}(Z_t) \geqslant \frac{1}{4}$ *for* $t \geqslant \tau_1$*. For both D-design and A-design, if* $b_j \geqslant \frac{d\|c_j\|_\infty}{\varepsilon}$ *for all* $j \in [m]$ *for some* $\varepsilon \leqslant \frac{1}{100}$*, then the probability that the randomized exchange algorithm has not terminated by time* $\tau_1 + \frac{2k}{\varepsilon}$ *is at most* $e^{-\Omega(\sqrt{d})}$*.*

First, we prove the following bicriteria approximation result for D/A-design with knapsack constraints, by combining the previous steps and using the concentration inequality for the knapsack constraints proved in Section 5.2.

Theorem 7.3.8. *Given* $\varepsilon \leqslant \frac{1}{100}$*, if* $b_j \geqslant \frac{d\|c_j\|_\infty}{\varepsilon}$ *for all* $j \in [m]$*, then the randomized exchange algorithm returns a solution set* S *within* $16k + \frac{2k}{\varepsilon}$ *iterations such that*

$$\det\left(\sum_{i\in S} u_i u_i^\top\right)^{\frac{1}{d}} \geqslant (1-10\varepsilon)\cdot\det(X)^{\frac{1}{d}} \quad \text{or} \quad \operatorname{tr}\left(\left(\sum_{i\in S} u_i u_i^\top\right)^{-1}\right) \leqslant (1+\varepsilon)\cdot\operatorname{tr}\left(X^{-1}\right)$$

for D-design and A-design respectively with probability at least $1-O\left(\frac{k^2}{\varepsilon^2}\cdot e^{-\Omega\left(\sqrt{d}\right)}\right)$*, where* X *is an optimal fractional solution to the convex relaxation (7.1). Moreover, for each* $j \in [m]$*, the solution set* S *satisfies*

$$c_j(S) \leqslant (1+\varepsilon)b_j + 120d\left\|c_j\right\|_\infty \leqslant \big(1+O(\varepsilon)\big)b_j$$

with probability at least $1 - e^{-\Omega(\varepsilon d)}$*.*

Proof. We start with defining some bad events for the randomized exchange algorithm.

- B_1: the algorithm has not terminated successfully within $16k$ iterations and $\tau_1 > 16k$ where τ_1 is the first time such that $\lambda_{\min}(Z_{\tau_1}) \geq \frac{3}{4}$.
- B_2: there exists some $\tau_1 \leq t \leq \tau_1 + \frac{2k}{\varepsilon}$ such that $\lambda_{\min}(Z_t) < 1/4$.
- B_3: the termination condition for D/A-design is not satisfied for all $\tau_1 \leq t \leq \tau_1 + \frac{2k}{\varepsilon}$.

If none of the bad events happens, then either the algorithm has terminated successfully within $16k$ iterations or the termination condition for D/A-design will be satisfied at some time $t \leq \tau_1 + \frac{2k}{\varepsilon} \leq 16k + \frac{2k}{\varepsilon}$. So, the probability that the randomized exchange algorithm has not satisfied the termination condition within $16k + \frac{2k}{\varepsilon}$ iterations is upper bounded by

$$\begin{aligned}\mathbb{P}[B_1 \cup B_2 \cup B_3] &\leq \mathbb{P}[B_1] + \mathbb{P}[B_2 \cap \neg B_1] + \mathbb{P}[B_3 \cap \neg B_2 \cap \neg B_1] \\ &\leq O\left(e^{-\Omega(\sqrt{d})}\right) + O\left(\frac{k^2}{\varepsilon^2} \cdot e^{-\Omega(\sqrt{d})}\right) + O\left(e^{-\Omega(\sqrt{d})}\right) \\ &\leq O\left(\frac{k^2}{\varepsilon^2} \cdot e^{-\Omega(\sqrt{d})}\right),\end{aligned}$$

where $\mathbb{P}[B_1]$ is bounded in Proposition 7.3.5, $\mathbb{P}[B_2 \cap \neg B_1]$ is bounded in Corollary 7.3.6, and $\mathbb{P}[B_3 \cap \neg B_2 \cap \neg B_1]$ is bounded in Theorem 7.3.7.

For D-design, since $v_i = X^{-\frac{1}{2}} u_i$, the termination condition implies the approximation guarantee as

$$\det\left(\sum_{i \in S} v_i v_i^\top\right)^{\frac{1}{d}} > 1 - 10\varepsilon \quad \Longrightarrow \quad \det\left(\sum_{i \in S} u_i u_i^\top\right)^{\frac{1}{d}} \geq (1 - 10\varepsilon) \cdot \det(X)^{\frac{1}{d}}.$$

For A-design, note that

$$\begin{aligned}\left\langle X^{-1}, \left(\sum_{i \in S} v_i v_i^\top\right)^{-1}\right\rangle &= \left\langle X^{-1}, \left(\sum_{i \in S} X^{-\frac{1}{2}} u_i u_i^\top X^{-\frac{1}{2}}\right)^{-1}\right\rangle \\ &= \left\langle I, \left(\sum_{i \in S} u_i u_i^\top\right)^{-1}\right\rangle = \operatorname{tr}\left(\left(\sum_{i \in S} u_i u_i^\top\right)^{-1}\right),\end{aligned} \tag{7.2}$$

and so the termination condition also implies the approximation guarantee as

$$\left\langle X^{-1}, \left(\sum_{i \in S} v_i v_i^\top\right)^{-1}\right\rangle \leq (1 + \varepsilon) \operatorname{tr}(X^{-1}) \quad \Longrightarrow \quad \operatorname{tr}\left(\left(\sum_{i \in S} u_i u_i^\top\right)^{-1}\right) \leq (1 + \varepsilon) \operatorname{tr}(X^{-1}).$$

Finally, we consider the knapsack constraints. Note that the termination conditions of both D/A-design imply $\lambda_{\min}(Z_t) < 1$ before the algorithm terminates. So, we can apply Theorem 5.2.12 with $\gamma = \frac{1}{8}$ to conclude that the returned solution S satisfies

$$c_j(S) \leq (1 + \varepsilon)\langle c_j, x\rangle + 120d \left\|c_j\right\|_\infty \leq (1 + \varepsilon) b_j + 120d \left\|c_j\right\|_\infty \leq (1 + O(\varepsilon)) b_j$$

with probability at least $1 - \exp(-\Omega(\varepsilon d))$, where the last inequality follows from $b_j \geq \frac{d\|c_j\|_\infty}{\varepsilon}$. □

We are ready to prove main theorem in this section by turning the above bicriteria approximation result to a true approximation result using a simple scaling argument.

Theorem 7.1.3. *Let $x \in [0,1]^n$ be an optimal fractional solution to the convex programming relaxation (7.1) for D/A-design with knapsack constraints. For any $\varepsilon \leqslant \frac{1}{200}$, if each knapsack constraint budget satisfies $b_j \geqslant \frac{2d\|c_j\|_\infty}{\varepsilon}$, then there is a randomized exchange algorithm which returns in polynomial time an integral solution $\sum_{i=1}^n z(i) \cdot u_i u_i^\top$ with $z(i) \in \{0,1\}$ for $1 \leqslant i \leqslant n$ such that*

$$\det\left(\sum_{i=1}^n z(i) \cdot u_i u_i^\top\right)^{\frac{1}{d}} \geqslant \big(1 - O(\varepsilon)\big) \cdot \det\left(\sum_{i=1}^n x(i) \cdot u_i u_i^\top\right)^{\frac{1}{d}} \text{ for D-design,}$$

$$\operatorname{tr}\left(\left(\sum_{i=1}^n z(i) \cdot u_i u_i^\top\right)^{-1}\right) \leqslant (1+\varepsilon) \cdot \operatorname{tr}\left(\left(\sum_{i=1}^n x(i) \cdot u_i u_i^\top\right)^{-1}\right) \text{ for A-design}$$

with probability at least $1 - O\left(\frac{k^2}{\varepsilon^2} \cdot e^{-\Omega(\sqrt{d})}\right)$ where $k = O(d^2 + m)$. Furthermore, each knapsack constraint $\langle c_j, z\rangle \leqslant b_j, j \in [m]$ is satisfied with probability at least $1 - e^{-\Omega(\varepsilon d)}$.

Proof. Let $b_1, \ldots, b_m$ be the input budgets for the m knapsack constraints. We scale down the budget to $\tilde{b}_j = \frac{b_j}{1+100\varepsilon}$ for each $j \in [m]$. Since $\varepsilon \leq \frac{1}{200}$ and $b_j \geq \frac{2d\|c_j\|_\infty}{\varepsilon}$ by the assumption, the rescaled budget $\tilde{b}_j \geq \frac{d\|c_j\|_\infty}{\varepsilon}$. Therefore, the budget assumptions in Theorem 7.3.8 are satisfied by all $\tilde{b}_1, ..., \tilde{b}_m$. In the following, we prove the theorem for D-design only, as the proof for A-design follows by the same argument.

Let $\tilde{x} \in [0,1]^n$ be an optimal fractional solution of (7.1) with budget $\tilde{b}_j$ for $j \in \{1, ..., m\}$. Let $\tilde{X} := \sum_{i=1}^n \tilde{x}(i) \cdot v_i v_i^\top$ and $X = \sum_{i=1}^n x(i) \cdot v_i v_i^\top$. We run the randomized exchange algorithm with budgets $\tilde{b}_1, ..., \tilde{b}_m$. By Theorem 7.3.8, with probability at least $1 - O\Big(\frac{k^2}{\varepsilon^2} \cdot e^{-\Omega(\sqrt{d})}\Big)$, the algorithm returns a solution set S within $O(\frac{k}{\varepsilon})$ iterations such that

$$\det\left(\sum_{i\in S} u_i u_i^\top\right)^{\frac{1}{d}} \geq (1 - 10\varepsilon) \cdot \det(\tilde{X})^{\frac{1}{d}} \geq \frac{1 - 10\varepsilon}{1 + 100\varepsilon} \cdot \det(X)^{\frac{1}{d}} = \big(1 - O(\varepsilon)\big) \cdot \det(X)^{\frac{1}{d}},$$

where the second inequality holds as $\frac{1}{1+100\varepsilon} \cdot X$ is a feasible solution to (7.1) with budget $\tilde{b}_1, ..., \tilde{b}_m$. Furthermore, for each knapsack constraint $j \in [m]$, it follows from Theorem 7.3.8

that

$$\boldsymbol{c}_j(S) \leq (1+\varepsilon)\tilde{b}_j + 120d\|\boldsymbol{c}_j\|_\infty \leq \frac{1+\varepsilon}{1+100\varepsilon} \cdot b_j + 60\varepsilon b_j \leq b_j,$$

with probability at least $1-\exp(-\Omega(\varepsilon d))$, where the second inequality follows by the assumption $b_j \geq \frac{2d\|\boldsymbol{c}_j\|_\infty}{\varepsilon}$ and the last inequality follows as $\varepsilon \leq \frac{1}{200}$. □

Unweighted D/A-Design: Using the main result, we improve the previous result on D/A-design with a single cardinality constraint by replacing the assumption in [131, 118], i.e. $b \geq \Omega\left(\frac{d}{\varepsilon} + \frac{1}{\varepsilon^2}\log\left(\frac{1}{\varepsilon}\right)\right)$, with $b \geq \frac{2d}{\varepsilon}$, although there is a mild assumption on the range of ε.

Corollary 7.3.9. *For any $\frac{1}{200} \geq \varepsilon \geq e^{-\delta\sqrt{d}}$ for a small enough constant δ, if $b \geq \frac{2d}{\varepsilon}$, then there is a randomized polynomial time algorithm that returns a $\big(1+O(\varepsilon)\big)$-approximate solution for D/A-design with constant probability.*

Proof. We apply Theorem 7.1.3 on the input. The probability that the output is a $\big(1+O(\varepsilon)\big)$-approximate solution and satisfies the cardinality constraint is at least $1-e^{-\Omega(\varepsilon d)} - e^{-\Omega(\sqrt{d})}$ as $k = O(d^2)$. When $\varepsilon d = \Omega(1)$, this success probability is at least a constant for large enough d. Otherwise, this success probability can be lower bounded by

$$1 - e^{-\Omega(\varepsilon d)} - e^{-\Omega(\sqrt{d})} \geq \Omega(\varepsilon d) - e^{-\Omega(\sqrt{d})} \geq \max\left\{e^{-\Omega(\sqrt{d})}, \Omega\Big(\frac{d^2}{n}\Big)\right\} \geq \Omega\Big(\frac{d^2}{n}\Big),$$

where the first inequality is by $e^{-\Omega(\varepsilon d)} \leq 1 - \Omega(\varepsilon d)$ for $\varepsilon d = o(1)$, and the second inequality is by the assumption $\varepsilon \geq \exp(-\delta\sqrt{d})$ for a small enough δ and the fact that we can assume $\varepsilon \geq \frac{2d}{n}$ without loss of generality. Therefore, we can amplify the success probability to be a constant by applying Theorem 7.1.3 at most $O\left(\frac{n}{d^2}\right)$ times, and the total time complexity is still polynomial in n and d. □

Minimizing Total Effective Resistance: We present an application of the main result to the total effective resistance minimization problem. In this problem, we are given a graph $G = (V, E)$ with Laplacian matrix $\boldsymbol{L}_G = \sum_{e \in E} \boldsymbol{b}_e \boldsymbol{b}_e^\top$ and a cost vector $\boldsymbol{c} \in \mathbb{R}_+^m$ on the edges, and the goal is to find a subgraph H with cost at most b to minimize the sum of all pairs effective resistances $\sum_{u,v} \mathrm{Reff}_H(u,v) = n \cdot \mathrm{tr}(\boldsymbol{L}_H^\dagger)$.

Corollary 7.1.7. *For any $0 < \varepsilon < 1$, there is a polynomial time randomized $(1+\varepsilon)$-approximation algorithm for minimizing total effective resistance in an edge weighted graph whenever $b \geqslant \Omega\left(\frac{n\|c\|_\infty}{\varepsilon}\right)$.*

Proof. As observed in Section 6.2.2, total effective resistance minimization can be reduced to A-design problem with a transformation described in (6.1). Let $x^* \in [0,1]^m$ be an optimal fractional solution to the problem and let $L_{x^*} := \sum_{e \in E} x^*(e) \cdot b_e b_e^\top$. Since $b \geqslant \Omega\left(\frac{n\|c\|_\infty}{\varepsilon}\right)$, by Theorem 7.1.3, there is a randomized algorithm that returns a subgraph H with $\operatorname{tr}\left(L_H^\dagger\right) \leqslant \left(1 + O(\varepsilon)\right) \cdot \operatorname{tr}\left(L_{x^*}^\dagger\right)$ within $O\left(\frac{k}{\varepsilon}\right)$ iterations with probability at least $1 - O\left(\frac{k^2}{\varepsilon^2} \cdot e^{-\Omega(\sqrt{n})}\right)$ where $k = O(n^2)$. Moreover, the cost constraint is satisfied with probability at least $1 - e^{-\Omega(\varepsilon n)}$. Since the number of edges $m = O(n^2)$, we can assume $\varepsilon \geqslant \frac{n}{m} = \Omega(\frac{1}{n})$ without loss of generality. The running time is polynomial in the graph size and the failure probability of the algorithm is at most $O\left(\frac{k^2}{\varepsilon^2} \cdot e^{-\Omega(\sqrt{n})}\right) + e^{-\Omega(\varepsilon n)} \leqslant e^{-\Omega(\sqrt{n})} + e^{-\Omega(1)}$, a constant bounded away from 1, when n is large enough. □

7.3.2 Analysis of the D-Design Objective

We will prove Theorem 7.3.7 for D-design in this subsection. Let τ_1 be the start time of the second phase. For the ease of notation, we simply reset $\tau_1 = 1$ as the first time step in the second phase. By assumption, $\lambda_{\min}(Z_1) \geqslant \frac{3}{4}$ and $\lambda_{\min}(Z_t) \geqslant \frac{1}{4}$ for all $t \geqslant 1$, which will be crucial in the analysis.

To analyze the objective value for D-design, our plan is to transform the product of random variables in Lemma 2.1.12 into a sum of random variables in the exponent as follows,

$$\begin{aligned}\det(Z_{\tau+1}) &\geqslant \det(Z_1) \cdot \prod_{t=1}^{\tau} \left(1 - \langle v_{i_t} v_{i_t}^\top, Z_t^{-1} \rangle\right) \left(1 + \langle v_{j_t} v_{j_t}^\top, Z_t^{-1} \rangle\right) \\ &\geqslant \det(Z_1) \cdot \exp\left(\sum_{t=1}^{\tau} \left((1 - 4\varepsilon) \underbrace{\langle v_{j_t} v_{j_t}^\top, Z_t^{-1} \rangle}_{\text{gain } g_t} - (1 + 5\varepsilon) \underbrace{\langle v_{i_t} v_{i_t}^\top, Z_t^{-1} \rangle}_{\text{loss } l_t} \right) \right),\end{aligned} \quad (7.3)$$

where the inequalities $1-x \geqslant e^{(1-4\varepsilon)x}$ and $1-x \geqslant e^{-(1+5\varepsilon)x}$ only hold when $x \in [0, 4\varepsilon]$ and ε is small enough such as $\varepsilon \leqslant \frac{1}{50}$.

So, for our plan to work, we need to bound the gain term $\langle v_{j_t} v_{j_t}^\top, Z_t^{-1} \rangle$ and the loss term $\langle v_{i_t} v_{i_t}^\top, Z_t^{-1} \rangle$. To do so, we prove in Lemma 7.3.13 that in an optimal fractional solution x, every vector v_i with $0 < x(i) < 1$ satisfies the condition that $\|v_i\|_2^2 \leqslant \varepsilon$. Recall that, Observation 7.3.3 implies $0 < x(i_t), x(j_t) < 1$ for all $t \geq 1$. Therefore, Lemma 7.3.13 implies that $\|v_{i_t}\|_2^2 \leqslant \varepsilon$ and $\|v_{j_t}\|_2^2 \leqslant \varepsilon$ for all $t \geqslant 1$. Together with the assumption that $Z_t \succcurlyeq \frac{1}{4} I$ for all $t \geqslant 1$, we can ensure that $\langle v_{j_t} v_{j_t}^\top, Z_t^{-1} \rangle \leqslant 4\varepsilon$ and $\langle v_{i_t} v_{i_t}^\top, Z_t^{-1} \rangle \leqslant 4\varepsilon$ for all $t \geqslant 1$, and hence (7.3) holds.

Once this transformation is done and (7.3) is established, we can apply Freedman's martingale inequality to prove concentration of the exponent. In the following, we define the gain g_t, loss l_t and progress Γ_t in the t-th iteration as

$$g_t := \langle v_{j_t} v_{j_t}^\top, Z_t^{-1} \rangle, \qquad l_t := \langle v_{i_t} v_{i_t}^\top, Z_t^{-1} \rangle, \qquad \text{and} \qquad \Gamma_t := (1-4\varepsilon) g_t - (1+5\varepsilon) l_t.$$

In Section 7.3.2.1, we will prove that the expected progress is large if the current solution is far from optimal. Then, in Section 7.3.2.2, we will prove that the total progress is concentrated around its expectation, where the minimum eigenvalue assumption is crucial in the martingale concentration argument. Finally, we finish the proof of Theorem 7.3.7 for D-design in Section 7.3.2.3, and present the proof of the optimality condition Lemma 7.3.13 in Section 7.3.2.4.

7.3.2.1 Expected Improvement of the D-Design Objective

Here we bound the conditional expectation of progress Γ_t, and show that $\mathbb{E}_t[\Gamma_t]$ is large if the current objective value $\det(Z_t)^{\frac{1}{d}}$ is small, where we denote $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid S_{t-1}]$.

Before that, we prove a useful lemma which will be used to relate the numerator of the gain term to the current objective value $\det(Z_t)^{\frac{1}{d}}$.

Lemma 7.3.10. *For any given $d \times d$ positive definite matrices $A, B \succ 0$,*

$$\langle A, B \rangle \geqslant d \cdot \det(A)^{\frac{1}{d}} \cdot \det(B)^{\frac{1}{d}}.$$

Proof. Let $A = \sum_{i=1}^d a_i u_i u_i^\top$ and $B = \sum_{j=1}^d b_j w_j w_j^\top$ be the spectral decompositions of A and B.

$$\begin{aligned}
\frac{1}{d} \cdot \langle A, B \rangle = \sum_{1 \leqslant i,j \leqslant d} a_i b_j \cdot \frac{\langle u_i, w_j \rangle^2}{d} &\geqslant \prod_{1 \leqslant i,j \leqslant d} \left(a_i b_j\right)^{\frac{\langle u_i, w_j \rangle^2}{d}} \\
&= \left(\prod_{i=1}^d \prod_{j=1}^d a_i^{\frac{\langle u_i, w_j \rangle^2}{d}} \right) \left(\prod_{j=1}^d \prod_{i=1}^d b_j^{\frac{\langle u_i, w_j \rangle^2}{d}} \right) \\
&= \prod_{i=1}^d a_i^{\frac{1}{d}} \cdot \prod_{j=1}^d b_j^{\frac{1}{d}} \\
&= \det(A)^{\frac{1}{d}} \det(B)^{\frac{1}{d}},
\end{aligned}$$

where the inequality follows by the weighted AM-GM inequality as $\sum_{i,j=1}^d \langle u_i, w_j \rangle^2 = d$, and the second last equality follows as $\{u_i\}_{i=1}^d$ and $\{w_j\}_{j=1}^d$ are orthonormal bases. □

We are ready to analyze the conditional expectation of progress Γ_t.

Lemma 7.3.11. *Let $\gamma \geqslant 1$. Let S_{t-1} be the solution set at time t and $Z_t = \sum_{i \in S_{t-1}} v_i v_i^\top$ for $1 \leqslant t \leqslant \tau$. Suppose $\det(Z_t)^{\frac{1}{d}} \leqslant \lambda$ for $1 \leqslant t \leqslant \tau$. Then*

$$\sum_{t=1}^\tau \mathbb{E}_t[\Gamma_t] \geqslant \left(\frac{1 - 4\varepsilon}{\lambda} - (1 + 5\varepsilon) \right) \cdot \frac{d\tau}{k}.$$

Proof. Let $t \in [1, \tau]$. Using the probability distribution for sampling v_{j_t} in the randomized exchange algorithm, the expected gain of adding vector v_{j_t} is

$$\begin{aligned}
\mathbb{E}_t[g_t] &= \sum_{j \in [n] \setminus S_{t-1}} \frac{x(j)}{k} \cdot \left(1 + 2\alpha \langle v_j v_j^\top, A_t^{\frac{1}{2}} \rangle \right) \cdot \langle v_j v_j^\top, Z_t^{-1} \rangle \\
&\geqslant \sum_{j \in [n] \setminus S_{t-1}} \frac{x(j)}{k} \cdot \langle v_j v_j^\top, Z_t^{-1} \rangle \\
&= \frac{1}{k} \left(\operatorname{tr}(Z_t^{-1}) - \sum_{i \in S_{t-1}} x(i) \cdot \langle v_i v_i^\top, Z_t^{-1} \rangle \right),
\end{aligned}$$

where the last equality uses $\sum_{j=1}^n x(j) \cdot v_j v_j^\top = I$.

Using the probability distribution for sampling v_{i_t} in the randomized exchange algorithm, the expected loss of removing vector v_{i_t} is

$$\begin{aligned}\mathbb{E}_t[l_t] &= \sum_{i \in S'_{t-1}} \frac{1-\mathsf{x}(i)}{k} \cdot \left(1 - 2\alpha\langle \mathsf{v}_i\mathsf{v}_i^\top, \mathsf{A}_t^{\frac{1}{2}}\rangle\right) \cdot \langle \mathsf{v}_i\mathsf{v}_i^\top, \mathsf{Z}_t^{-1}\rangle \\ &\leqslant \frac{1}{k}\sum_{i \in S'_{t-1}} \big(1-\mathsf{x}(i)\big) \cdot \langle \mathsf{v}_i\mathsf{v}_i^\top, \mathsf{Z}_t^{-1}\rangle \\ &\leqslant \frac{1}{k}\sum_{i \in S_{t-1}} \big(1-\mathsf{x}(i)\big) \cdot \langle \mathsf{v}_i\mathsf{v}_i^\top, \mathsf{Z}_t^{-1}\rangle \\ &= \frac{1}{k}\left(d - \sum_{i \in S_{t-1}} \mathsf{x}(i) \cdot \langle \mathsf{v}_i\mathsf{v}_i^\top, \mathsf{Z}_t^{-1}\rangle\right),\end{aligned} \tag{7.4}$$

where the two inequalities hold as $1 - 2\alpha\langle \mathsf{v}_i\mathsf{v}_i^\top, \mathsf{A}_t^{\frac{1}{2}}\rangle \leqslant 1$ and $(1-x_i) \cdot \langle \mathsf{v}_i\mathsf{v}_i^\top, \mathsf{Z}_t^{-1}\rangle \geqslant 0$ for all $i \in [n]$, and the last equality holds as $\sum_{i \in S_{t-1}} \mathsf{v}_i\mathsf{v}_i^\top = \mathsf{Z}_t$.

Therefore, the expected progress is

$$\begin{aligned}&\mathbb{E}_t[\Gamma_t] = \mathbb{E}_t[(1-4\varepsilon)g_t - (1+5\varepsilon)l_t] \\ \geqslant\ & \frac{1-4\varepsilon}{k}\left(\operatorname{tr}(\mathsf{Z}_t^{-1}) - \sum_{i \in S_{t-1}} \mathsf{x}(i) \cdot \langle \mathsf{v}_i\mathsf{v}_i^\top, \mathsf{Z}_t^{-1}\rangle\right) - \frac{1+5\varepsilon}{k}\left(d - \sum_{i \in S_{t-1}} \mathsf{x}(i) \cdot \langle \mathsf{v}_i\mathsf{v}_i^\top, \mathsf{Z}_t^{-1}\rangle\right) \\ \geqslant\ & \frac{1}{k}\big((1-4\varepsilon) \cdot \operatorname{tr}(\mathsf{Z}_t^{-1}) - (1+5\varepsilon) \cdot d\big) \\ \geqslant\ & \frac{1}{k}\left((1-4\varepsilon) \cdot \frac{d}{\det(\mathsf{Z}_t)^{\frac{1}{d}}} - (1+5\varepsilon) \cdot d\right) \\ \geqslant\ & \left(\frac{1-4\varepsilon}{\lambda} - (1+5\varepsilon)\right) \cdot \frac{d}{k},\end{aligned}$$

where the second last inequality follows from Lemma 7.3.10, and the last inequality is by the assumption that $\max_t \det(\mathsf{Z}_t)^{\frac{1}{d}} \leqslant \lambda$. The lemma follows by summing over all $1 \leqslant t \leqslant \tau$. □

7.3.2.2 Martingale Concentration Argument

Here we show that the total progress is concentrated around the expectation. The proof uses the minimum eigenvalue assumption and the short vector condition from Lemma 7.3.13 to bound the variance of the random process.

Lemma 7.3.12. *Suppose $Z_t \succcurlyeq \frac{1}{4}I$ and $\|v_{i_t}\|_2^2 \leqslant \varepsilon$ and $\|v_{j_t}\|_2^2 \leqslant \varepsilon$ for $\varepsilon \leqslant \frac{1}{100}$ for all $1 \leqslant t \leqslant \tau$. Then, for any $\eta > 0$,*

$$\mathbb{P}\left[\sum_{t=1}^{\tau} \Gamma_t \leqslant \sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t] - \eta\right] \leqslant \exp\Big(-\Omega\Big(\frac{\eta^2 k}{\varepsilon\tau d^{1.5} + \varepsilon\eta k}\Big)\Big).$$

Proof. We define two sequences of random variables $\{X_t\}_t$ and $\{Y_t\}_t$, where $X_t := \mathbb{E}_t[\Gamma_t] - \Gamma_t$ and $Y_t := \sum_{l=1}^{t} X_l$. It is easy to check that $\{Y_t\}_t$ is a martingale with respect to $\{S_t\}_t$. We will use Freedman's inequality to bound $\mathbb{P}[Y_\tau \geqslant \eta]$.

To apply Freedman's inequality, we need to upper bound X_t and $\mathbb{E}_t[X_t^2]$. Note that

$$0 \leqslant g_t = \langle v_{j_t} v_{j_t}^\top, Z_t^{-1} \rangle \leqslant 4\varepsilon \quad \text{and} \quad 0 \leqslant l_t = \langle v_{i_t} v_{i_t}^\top, Z_t^{-1} \rangle \leqslant 4\varepsilon$$

by our assumptions that $Z_t \succcurlyeq \frac{1}{4}I$ and $\|v_{i_t}\|_2^2 \leqslant \varepsilon$ and $\|v_{j_t}\|_2^2 \leqslant \varepsilon$ for $1 \leqslant t \leqslant \tau$. These imply that

$$X_t = \mathbb{E}_t[\Gamma_t] - \Gamma_t \leqslant (1 - 4\varepsilon) \cdot \mathbb{E}_t[g_t] + (1 + 5\varepsilon) \cdot l_t \leqslant (2 + \varepsilon) \cdot 4\varepsilon \leqslant 10\varepsilon,$$

where the last inequality holds for $\varepsilon \leqslant \frac{1}{2}$.

To upper bound $\mathbb{E}_t[X_t^2]$, we first upper bound $\mathbb{E}_t[g_t]$ and $\mathbb{E}_t[l_t]$. Note that

$$\begin{aligned}
\mathbb{E}_t[g_t] &= \sum_{j \in [n] \setminus S_{t-1}} \frac{x(j)}{k} \cdot \Big(1 + 2\alpha \langle v_j v_j^\top, A_t^{\frac{1}{2}} \rangle\Big) \cdot \langle v_j v_j^\top, Z_t^{-1} \rangle \\
&\leqslant \frac{1 + 16\varepsilon\sqrt{d}}{k} \cdot \sum_{j \in [n] \setminus S_{t-1}} x(j) \cdot \langle v_j v_j^\top, Z_t^{-1} \rangle \\
&\leqslant \frac{1 + 16\varepsilon\sqrt{d}}{k} \cdot \mathrm{tr}(Z_t^{-1}) \\
&\leqslant \frac{4d + 64\varepsilon d^{1.5}}{k},
\end{aligned}$$

where the first inequality holds as $\alpha = 8\sqrt{d}$, $A_t \preccurlyeq I$ and $\|v_j\|_2^2 \leqslant \varepsilon$ for $j \in [n] \setminus S_{t-1}$ with $x(j) > 0$, the second inequality follows as $\sum_{i=1}^n x(i) \cdot v_i v_i^\top = I$, and the last inequality follows from the assumption that $Z_t \succcurlyeq \frac{1}{4} I$. Note also that $\mathbb{E}_t[l_t] \leqslant \frac{d}{k}$ from (7.4) in Lemma 7.3.11. So, we can upper bound $\mathbb{E}_t[X_t^2]$ by

$$\mathbb{E}_t[X_t^2] \leqslant 10\varepsilon \cdot \mathbb{E}_t[|X_t|] \leqslant 20\varepsilon \cdot \Big((1-4\varepsilon) \cdot \mathbb{E}_t[g_t] + (1+5\varepsilon) \cdot \mathbb{E}_t[l_t]\Big) \leqslant O\left(\frac{\varepsilon d^{1.5}}{k}\right),$$

where the first inequality is by the upper bound on X_t, and the last inequality is by the loose bound that $\mathbb{E}_t[g_t] \leqslant O\left(\frac{d^{1.5}}{k}\right)$.

Finally, we can apply Freedman's inequality Theorem 3.2.3 with $R = 10\varepsilon$, $\sigma_t^2 = O\left(\frac{\varepsilon d^{1.5}}{k}\right)$ for all $t \in [\tau]$, and $\sigma^2 = O\left(\frac{\varepsilon \tau d^{1.5}}{k}\right)$ to conclude that

$$\mathbb{P}[Y_\tau \geqslant \eta] \leqslant \exp\Big(-\frac{\eta^2/2}{\sigma^2 + R\eta/3}\Big) = \exp\Big(-\Omega\Big(\frac{\eta^2 k}{\varepsilon \tau d^{1.5} + \varepsilon \eta k}\Big)\Big).$$

The lemma follows by noting that $Y_\tau \geqslant \eta$ is equivalent to $\sum_{t=1}^\tau \Gamma_t \leqslant \sum_{t=1}^\tau \mathbb{E}_t[\Gamma_t] - \eta$. □

7.3.2.3 Proof of Theorem 7.3.7 for D-design

We are ready to prove Theorem 7.3.7 for D-design. Let $\tau = \frac{2k}{\varepsilon}$. Suppose the second phase of the algorithm has not terminated by time τ. Then $\lambda = \max_{1 \leqslant t \leqslant \tau+1} \det(Z_t)^{\frac{1}{d}} < 1 - 10\varepsilon$. Thus, Lemma 7.3.11 implies that

$$\sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t] \geqslant \Big(\frac{1-4\varepsilon}{\lambda} - (1+5\varepsilon)\Big) \cdot \frac{d\tau}{k} \geqslant \frac{\varepsilon d \tau}{k} = 2d.$$

On the other hand, the initial solution of the second phase satisfies $Z_1 \succcurlyeq \frac{3}{4} I$, which implies that $\det(Z_1) \geqslant \left(\frac{3}{4}\right)^d$. As the knapsack constraints satisfy $b_j \geqslant \frac{d\|c_j\|_\infty}{\varepsilon}$ for $j \in [m]$, we know from Lemma 7.3.13 that $\|v_i\|_2^2 \leqslant \varepsilon$ for each i with $0 < x(i) < 1$. Note that, in the randomized exchange algorithm, all i_t and j_t satisfy $0 < x(i_t), x(j_t) < 1$ by Observation 7.3.3. Together with the assumption that $Z_t \succcurlyeq \frac{1}{4} I$ for all $1 \leqslant t \leqslant \tau$, we have $\langle v_{j_t} v_{j_t}^\top, Z_t^{-1} \rangle \leqslant 4\varepsilon$ and $\langle v_{i_t} v_{i_t}^\top, Z_t^{-1} \rangle \leqslant 4\varepsilon$ for all $1 \leqslant t \leqslant \tau$. Hence, we can apply (7.3) to deduce that

$$1 > \det(Z_{\tau+1}) \geqslant \det(Z_1) \cdot \exp\left(\sum_{t=1}^{\tau} \Gamma_t\right) \geqslant \left(\frac{3}{4}\right)^d \exp\left(\sum_{t=1}^{\tau} \Gamma_t\right) \quad \Longrightarrow \quad \sum_{t=1}^{\tau} \Gamma_t \leqslant d \cdot \ln \frac{4}{3} \leqslant d.$$

Therefore, we can apply Lemma 7.3.12 with $\eta = d$ and $\tau = \frac{2k}{\varepsilon}$ to conclude that

$$\begin{aligned}\mathbb{P}\left[\max_{1\leqslant t\leqslant \tau+1} \det(Z_t)^{\frac{1}{d}} < 1-10\varepsilon\right] &\leqslant \mathbb{P}\left[\sum_{t=1}^{\tau}\Gamma_t < \sum_{t=1}^{\tau}\mathbb{E}_t[\Gamma_t] - d\right] \\ &\leqslant \exp\left(-\Omega\left(\frac{d^2k}{\varepsilon\left(\frac{2k}{\varepsilon}\right)d^{1.5}+\varepsilon dk}\right)\right) \\ &\leqslant \exp(-\Omega(\sqrt{d})).\end{aligned}$$

7.3.2.4 Optimality Condition of the Convex Program for D-Design

The following lemma uses the assumption about the budgets to prove that all vectors with fractional value are short.

Lemma 7.3.13. *Let $x \in [0,1]^n$ be an optimal fractional solution of the convex programming relaxation (7.1) for D-design. Let $X = \sum_{i=1}^n x(i)\cdot u_i u_i^\top$, and $v_i = X^{-\frac{1}{2}}u_i$ for $1 \leqslant i \leqslant n$. Suppose $b_j \geqslant \frac{d\|c_j\|_\infty}{\varepsilon}$ for $1 \leqslant j \leqslant m$. Then $\|v_i\|_2^2 \leqslant \varepsilon$ for each $1 \leqslant i \leqslant n$ with $0 < x(i) < 1$.*

Proof. Since both $x^{-\frac{1}{d}}$ and $\log x$ are monotone functions for $x > 0$, minimizing $f_D(X) = \det(X)^{-\frac{1}{d}}$ on $\mathbb{S}_{++}^d$ is equivalent to maximizing $\log\det(X)$ on $\mathbb{S}_{++}^d$. Thus, the following convex relaxation of D-design ($\log\det(X)$ is concave on $\mathbb{S}_+^d+$ by Fact 2.2.15) would have exactly the same optimizer and optimal solution characterization as (7.1) for D-design.

$$\begin{aligned}\max_{x\in\mathbb{R}^d, X\in\mathbb{S}_{++}^d} \quad & \log\det(X) \\ \text{subject to} \quad & X = \sum_{i=1}^n x(i)\cdot u_i u_i^\top, \\ & \langle c_j, x\rangle \leqslant b_j, \qquad \forall j\in[m], \\ & 0 \leqslant x(i) \leqslant 1, \qquad \forall i \in [n].\end{aligned}$$

As the gradient of $\nabla f_D(X) = -\frac{1}{d}\det(X)^{-\frac{1}{d}}X^{-1}$ has a more complicated form then that of $\nabla\log(X) = X^{-1}$, without loss of generality, we analyze the above convex program for D-design for the ease of notations.

We will use the Lagrangian duality (see Section 2.2.3.1) to investigate the length of the vectors v_i's. We introduce a dual variable Y for the first equality constraint, a dual variable $\mu_j \geqslant 0$ for each of the budget constraint $b_j - \langle c_j, x\rangle \geqslant 0$, a dual variable $\beta_i^- \geqslant 0$ for each non-negative constraint $x(i) \geqslant 0$, and a dual variable $\beta_i^+ \geqslant 0$ for each capacity constraint $1 - x(i) \geqslant 0$. The Lagrange function $L(x, X, Y, \mu, \beta^+, \beta^-)$ is defined as

$$L(x, X, Y, \mu, \beta^+, \beta^-) = \log\det(X) + \left\langle Y, \sum_{i=1}^{n} x(i) \cdot u_i u_i^\top - X \right\rangle \\ + \sum_{j=1}^{m} \mu_j \Big(b_j - \langle c_j, x\rangle \Big) + \sum_{i=1}^{n} \beta_i^- x(i) + \sum_{i=1}^{n} \beta_i^+ (1 - x(i)),$$

Rearrange the terms, we have

$$L(x, X, Y, \mu, \beta^+, \beta^-) = \log\det(X) - \langle Y, X\rangle + \sum_{j=1}^{m} \mu_j b_j + \sum_{i=1}^{n} \beta_i^+ \\ + \sum_{i=1}^{n} x(i) \cdot \left(\langle Y, u_i u_i^\top \rangle - \sum_{j=1}^{m} \mu_j c_j(i) + \beta_i^- - \beta_i^+ \right).$$

The Lagrangian dual function is

$$g(Y, \mu, \beta^+, \beta^-) = \max_{x, X \succ 0} L(x, X, Y, \mu, \beta^+, \beta^-).$$

It is easy to verify that $x = \delta \mathit{1}$ is a strictly feasible solution of the primal program for a small enough δ. By Theorem 2.2.26, Slater's condition implies that strong duality holds. In the primal convex program, we can relax the constraint $X \succ 0$ to $X \succcurlyeq 0$ without loss of generality, as $\log\det(X)$ blows up to $-\infty$ when X approaches the boundary of $\mathbb{S}_+^d$. Thus, the feasible solution space is closed and bounded, and the primal optimal is attained. Let $x \in [0,1]^n, X \succ 0$ be an optimal solution for the primal program, Theorem 2.2.28 says there exists a dual optimal solution $Y, \mu, \beta^+, \beta^- \geqslant 0$ together with x, X satisfy the KKT conditions. In particular, it holds that (we recall $\nabla \log\det(X) = X^{-1}$ by Fact 2.2.2)

$$\begin{aligned}
&\text{(Complementary slackness)} && \beta_i^- \cdot x(i) = 0,\ \ \beta_i^+ \cdot (1 - x(i)) = 0\ \forall i \in [n],\\
&\text{(Lagrangian optimality)} && \nabla_X L = X^{-1} - Y = 0,\\
& && \nabla_{x(i)} L = \langle Y, u_i u_i^\top \rangle - \sum_{j=1}^m \mu_j c_j(i) + \beta_i^- - \beta_i^+ = 0,\ \ \forall i \in [n].
\end{aligned}$$

By strong duality, the primal optimal and the dual optimal attain the same objective value, i.e. $\log\det(X) = g(Y, \mu, \beta^+, \beta^-) = L(x, X, Y, \mu, \beta^+, \beta^-)$. Thus, it holds that

$$\log\det(X) = L(x, X, Y, \mu, \beta^+, \beta^-) = \log\det(X) - d + \sum_{j=1}^m \mu_j b_j + \sum_{i=1}^n \beta_i^+,$$

where the last equality holds by Lagrangian optimality, $\langle Y, u_i u_i^\top \rangle = \sum_{j=1}^m \mu_j c_j(i) - \beta_i^- + \beta_i^+ = 0$ for all $i \in [n]$ and $Y = X^{-1}$. Since $\beta^+ \geqslant 0$, it further implies

$$\sum_{j=1}^m \mu_j b_j \leqslant d \quad \Longrightarrow \quad \sum_{j=1}^m \mu_j \left\| c_j \right\|_\infty \leqslant \varepsilon,$$

where the last implication follows by the assumption $b_j \geqslant \frac{d\|c_j\|_\infty}{\varepsilon}$ for each $j \in [m]$.

Finally, by the complementary slackness condition, we must have $\beta_i^+ = \beta_i^- = 0$ for each i with $0 < x(i) < 1$. Together with the Lagrangian optimality, $\langle Y, u_i u_i^\top \rangle = \sum_{j=1}^m \mu_j c_j(i) - \beta_i^- + \beta_i^+ = 0$ for all $i \in [n]$, for any $i \in [n]$ with $0 < x(i) < 1$ it holds that

$$\sum_{j=1}^m \mu_j c_j(i) = \langle Y, u_i u_i^\top \rangle = \langle X^{-1}, u_i u_i^\top \rangle = \|v_i\|_2^2 \qquad \Longrightarrow \qquad \|v_i\|_2^2 \leqslant \sum_{j=1}^m \mu_j \left\| c_j \right\|_\infty \leqslant \varepsilon. \quad \square$$

7.3.3 Analysis of the A-Design Objective

We will prove Theorem 7.3.7 for A-design in this subsection. Let τ_1 be the start time of the second phase. For ease of notation, we simply reset $\tau_1 = 1$, as the first time step in the second phase. By assumption, $\lambda_{\min}(Z_1) \geqslant \frac{3}{4}$ and $\lambda_{\min}(Z_t) \geqslant \frac{1}{4}$ for all $t \geqslant 1$, which will be crucial in the analysis.

To analyze the A-design objective $\operatorname{tr}\left(\left(\sum_{i\in S_{t-1}} u_i u_i^\top\right)^{-1}\right)$, we analyze the equivalent quantity $\langle X^{-1}, Z_t^{-1}\rangle$ after the linear transformation $v_i = X^{-\frac{1}{2}} u_i$ as shown in (7.2). By Lemma 4.2.5, if $2\langle v_{i_t} v_{i_t}^\top, Z_t^{-1}\rangle < 1$, then the change of the objective value is bounded by

$$\begin{aligned}\langle X^{-1}, Z_{t+1}^{-1}\rangle &= \langle X^{-1}, (Z_t - v_{i_t} v_{i_t}^\top + v_{j_t} v_{j_t}^\top)^{-1}\rangle \\ &\leqslant \langle X^{-1}, Z_t^{-1}\rangle + \frac{\langle X^{-1}, Z_t^{-1} v_{i_t} v_{i_t}^\top Z_t^{-1}\rangle}{1 - 2\langle v_{i_t} v_{i_t}^\top, Z_t^{-1}\rangle} - \frac{\langle X^{-1}, Z_t^{-1} v_{j_t} v_{j_t}^\top Z_t^{-1}\rangle}{1 + 2\langle v_{j_t} v_{j_t}^\top, Z_t^{-1}\rangle}.\end{aligned}$$

In Section 7.4.2, when analyzing the combinatorial algorithm for A-design, we will show in Lemma 7.4.4 and Lemma 7.4.5 that if we sample i_t and j_t from the distributions

$$\mathbb{P}[i_t = i] \propto \left(1 - x(i)\right) \cdot \left(1 - 2\langle v_i v_i^\top, Z_t^{-1}\rangle\right) \quad \text{and} \quad \mathbb{P}[j_t = j] \propto x(j) \cdot \left(1 + 2\langle v_j v_j^\top, Z_t^{-1}\rangle\right),$$

then the objective value will improve in expectation when the current objective value is far from optimal. In the randomized exchange algorithm, however, we sample i_t and j_t from the E-design distributions. An important observation is that the quantities in these two distributions can be related to each other when the minimum eigenvalue assumption holds. The following lemma will be proved in Section 7.3.3.1.

Lemma 7.3.14. *If $Z_t \succcurlyeq \frac{1}{4} I_d$, then $\langle v_i v_i^\top, Z_t^{-1}\rangle \leqslant \alpha \cdot \langle v_i v_i^\top, A_t^{\frac{1}{2}}\rangle \leqslant \alpha \lambda_{\min}(Z_t) \cdot \langle v_i v_i^\top, Z_t^{-1}\rangle$ for $1 \leqslant i \leqslant n$.*

In the exchange subroutine of the randomized exchange algorithm, only those i_t with $2\alpha \cdot \langle v_{i_t} v_{i_t}^\top, A_t^{\frac{1}{2}}\rangle \leqslant \frac{1}{2}$ are sampled. So, when the minimum eigenvalue assumption holds, Lemma 7.3.14 implies that the randomized exchange algorithm only samples i_t that satisfies $2\langle v_{i_t} v_{i_t}^\top, Z_t^{-1}\rangle \leqslant \frac{1}{2}$. Therefore, we can apply Lemma 4.2.5 repeatedly to obtain that for any $\tau \geqslant 1$,

$$\langle X^{-1}, Z_{\tau+1}^{-1}\rangle \leqslant \langle X^{-1}, Z_1^{-1}\rangle - \sum_{t=1}^{\tau} \left(\frac{\langle X^{-1}, Z_t^{-1} v_{j_t} v_{j_t}^\top Z_t^{-1}\rangle}{1 + 2\langle v_{j_t} v_{j_t}^\top, Z_t^{-1}\rangle} - \frac{\langle X^{-1}, Z_t^{-1} v_{i_t} v_{i_t}^\top Z_t^{-1}\rangle}{1 - 2\langle v_{i_t} v_{i_t}^\top, Z_t^{-1}\rangle} \right). \tag{7.5}$$

As in Section 7.3.2, we define gain g_t, loss l_t and progress Γ_t in the t-th iteration as follows

$$g_t := \frac{\langle X^{-1}, Z_t^{-1} v_{j_t} v_{j_t}^\top Z_t^{-1}\rangle}{1 + 2\langle v_{j_t} v_{j_t}^\top, Z_t^{-1}\rangle}, \quad l_t := \frac{\langle X^{-1}, Z_t^{-1} v_{i_t} v_{i_t}^\top Z_t^{-1}\rangle}{1 - 2\langle v_{i_t} v_{i_t}^\top, Z_t^{-1}\rangle}, \quad \text{and} \quad \Gamma_t := g_t - l_t.$$

In Section 7.3.3.1, we will prove Lemma 7.3.14, and use it to prove that the expected progress is large if the current objective value is far from optimal. Then, in Section 7.3.3.2, we will prove that the total progress is concentrated around its expectation, while the minimum eigenvalue assumption and the optimality condition of the convex programming relaxation are crucial in the martingale concentration argument. Finally, we complete the proof of Theorem 7.3.7 for A-design in Section 7.3.3.3, and present the proof of the optimality condition in Section 7.3.3.4.

7.3.3.1 Expected Improvement of the A-Design Objective

We first prove Lemma 7.3.14, which will be useful in bounding the expectation.

Proof of Lemma 7.3.14. Recall that $A_t = (\alpha Z_t - l_t I_d)^{-2}$ where l_t is the unique value such that $A_t \succ 0$ and $\mathrm{tr}(A_t) = 1$. Since $Z_t \succcurlyeq \lambda_{\min}(Z_t) \cdot I_d$, it follows that

$$1 = \mathrm{tr}(A_t) \leqslant (\alpha\lambda_{\min}(Z_t) - l_t)^{-2} \cdot \mathrm{tr}(I_d) \quad \Longrightarrow \quad \alpha\lambda_{\min}(Z_t) - l_t \leqslant \sqrt{d} \quad \Longrightarrow \quad l_t \geqslant 0,$$

where the last implication holds as $\alpha = 8\sqrt{d}$ and $\lambda_{\min}(Z_t) \geqslant \frac{1}{4}$. This implies that $A_t^{\frac{1}{2}} = (\alpha Z_t - l_t I_d)^{-1} \succcurlyeq \alpha^{-1} Z_t^{-1}$, proving the first inequality.

For the second inequality, consider the eigen-decomposition of $Z_t = \sum_{j=1}^{d} \lambda_j w_j w_j^\top$, where $0 < \lambda_1 \leqslant \ldots \leqslant \lambda_d$ are the eigenvalues and $\{w_j\}_j$ are the corresponding orthonormal eigenvectors. Then,

$$\frac{\langle v_i v_i^\top, A_t^{\frac{1}{2}} \rangle}{\langle v_i v_i^\top, Z_t^{-1} \rangle} = \frac{\sum_{j=1}^{d} \frac{\langle v_i, w_j \rangle^2}{\alpha\lambda_j - l_t}}{\sum_{j=1}^{d} \frac{\langle v_i, w_j \rangle^2}{\lambda_j}} \leqslant \max_{j \in [d]} \frac{\lambda_j}{\alpha\lambda_j - l_t} \leqslant \frac{\lambda_1}{\alpha\lambda_1 - l_t} \leqslant \lambda_1,$$

where the first inequality holds since $\alpha\lambda_j - l_t > 0$ as $A_t \succ 0$, the second inequality holds as $l_t \geqslant 0$ and the function $f(x) = \frac{x}{\alpha x - l_t}$ is decreasing for $x > \frac{l_t}{\alpha}$ when $l_t \geqslant 0$, and the last inequality follows as $1 = \mathrm{tr}(A_t) \geqslant (\alpha\lambda_1 - l_t)^{-2}$ which implies $\alpha\lambda_1 - l_t \geqslant 1$. □

Before analyzing the expectation, we prove another useful lemma, which will also be used in the analysis of combinatorial local search algorithm later in this chapter.

Lemma 7.3.15. *For any given $d \times d$ positive definite matrices $A, B \succ 0$,*

$$\langle A, B^2 \rangle \geqslant \frac{(\mathrm{tr}(B))^2}{\mathrm{tr}\left(A^{-1}\right)} \qquad \text{and} \tag{7.6}$$

$$\langle A, B \rangle \leqslant \sqrt{\mathrm{tr}(A) \cdot \langle A, B^2 \rangle}. \tag{7.7}$$

Proof. Let $A = \sum_{i=1}^{d} a_i u_i u_i^\top$ and $B = \sum_{j=1}^{d} b_j w_j w_j^\top$ be the eigendecomposition of A and B. Then,

$$\begin{aligned}\mathrm{tr}(B) = \sum_{j=1}^{d} b_j = \sum_{1 \leqslant i,j \leqslant d} b_j \cdot \langle u_i, w_j \rangle^2 &= \sum_{1 \leqslant i,j \leqslant d} \sqrt{a_i} b_j \langle u_i, w_j \rangle \cdot \frac{1}{\sqrt{a_i}} \langle u_i, w_j \rangle \\ &\leqslant \sqrt{\sum_{1 \leqslant i,j \leqslant d} a_i b_j^2 \langle u_i, w_j \rangle^2 \cdot \sum_{1 \leqslant i,j \leqslant d} \frac{1}{a_i} \langle u_i, w_j \rangle^2} \\ &= \sqrt{\langle A, B^2 \rangle \cdot \mathrm{tr}(A^{-1})},\end{aligned}$$

where the second equality and the last equality hold as $\{u_i\}_{i=1}^{d}$ and $\{w_j\}_{j=1}^{d}$ are orthonormal bases, and the inequality is by Cauchy-Schwarz. For the second inequality in this lemma,

$$\langle A, B \rangle = \sum_{1 \leqslant i,j \leqslant d} a_i b_j \langle u_i, w_j \rangle^2 \leqslant \sqrt{\sum_{1 \leqslant i,j \leqslant d} a_i \langle u_i, w_j \rangle^2 \cdot \sum_{1 \leqslant i,j \leqslant d} a_i b_j^2 \langle u_i, w_j \rangle^2} = \sqrt{\mathrm{tr}(A) \cdot \langle A, B^2 \rangle},$$

where the equalities hold as $\{u_i\}_i$ and $\{w_j\}_j$ are orthonormal bases and the inequality is by Cauchy-Schwarz. □

The following lemma shows that the expected progress is large if the current objective value is far from optimal. Note that, in contrast to Section 7.3.2 for D-design, the minimum eigenvalue assumption is needed in the proof.

Lemma 7.3.16. *Let S_{t-1} be the solution set at time t and $Z_t = \sum_{i \in S_{t-1}} v_i v_i^\top$ for $1 \leqslant t \leqslant \tau$. Suppose $Z_t \succcurlyeq \frac{1}{4} I$ and $\langle X^{-1}, Z_t^{-1} \rangle \geqslant \lambda \cdot \mathrm{tr}\left(X^{-1}\right)$ for $\lambda \geqslant 1$ for $1 \leqslant t \leqslant \tau$. Then*

$$\sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t] \geqslant \frac{(\lambda - 1)\tau}{k} \cdot \mathrm{tr}(X^{-1}).$$

Proof. The expected gain of adding vector v_{j_t} is

$$\begin{aligned}
\mathbb{E}_t[g_t] &= \sum_{j \in [n] \setminus S_{t-1}} \frac{x(j)}{k} \cdot \left(1 + 2\alpha \langle v_j v_j^\top, A_t^{\frac{1}{2}} \rangle\right) \cdot \frac{\langle X^{-1}, Z_t^{-1} v_j v_j^\top Z_t^{-1} \rangle}{1 + 2\langle v_j v_j^\top, Z_t^{-1} \rangle} \\
&\geqslant \sum_{j \in [n] \setminus S_{t-1}} \frac{x(j)}{k} \cdot \langle X^{-1}, Z_t^{-1} v_j v_j^\top Z_t^{-1} \rangle \\
&= \frac{1}{k}\left(\langle X^{-1}, Z_t^{-2} \rangle - \left\langle X^{-1}, Z_t^{-1} \left(\sum_{i \in S_{t-1}} x(i) \cdot v_i v_i^\top \right) Z_t^{-1} \right\rangle \right),
\end{aligned}$$

where the inequality follows from Lemma 7.3.14 and the last equality follows as $\sum_{j=1}^{n} x(j) \cdot v_j v_j^\top = I$. The expected loss of removing vector v_{i_t} is

$$\begin{aligned}
\mathbb{E}_t[l_t] &= \sum_{i \in S'_{t-1}} \frac{1 - x(i)}{k} \cdot \left(1 - 2\alpha \langle v_i v_i^\top, A_t^{\frac{1}{2}} \rangle\right) \cdot \frac{\langle X^{-1}, Z_t^{-1} v_i v_i^\top Z_t^{-1} \rangle}{1 - 2\langle v_i v_i^\top, Z_t^{-1} \rangle} \\
&\leqslant \frac{1}{k} \sum_{i \in S'_{t-1}} \left(1 - x(i)\right) \cdot \langle X^{-1}, Z_t^{-1} v_i v_i^\top Z_t^{-1} \rangle \\
&\leqslant \frac{1}{k} \sum_{i \in S_{t-1}} \left(1 - x(i)\right) \cdot \langle X^{-1}, Z_t^{-1} v_i v_i^\top Z_t^{-1} \rangle \\
&= \frac{1}{k}\left(\langle X^{-1}, Z_t^{-1} \rangle - \left\langle X^{-1}, Z_t^{-1} \left(\sum_{i \in S_{t-1}} x(i) \cdot v_i v_i^\top \right) Z_t^{-1} \right\rangle \right), \qquad (7.8)
\end{aligned}$$

where the first inequality follows from Lemma 7.3.14 and $2\alpha \langle v_i v_i^\top, A_t^{\frac{1}{2}} \rangle \leqslant \frac{1}{2}$ by the definition of S'_{t-1}, and the last equality holds as $\sum_{i \in S_{t-1}} v_i v_i^\top = Z_t$.

Therefore, the expected progress is

$$\mathbb{E}_t[\Gamma_t] = \mathbb{E}_t[g_t] - \mathbb{E}_t[l_t] \geqslant \frac{1}{k}\left(\langle X^{-1}, Z_t^{-2} \rangle - \langle X^{-1}, Z_t^{-1} \rangle \right).$$

The term $\langle X^{-1}, Z_t^{-2} \rangle$ can be lower bounded by

$$\langle X^{-1}, Z_t^{-2} \rangle \geqslant \frac{\langle X^{-1}, Z_t^{-1} \rangle^2}{\operatorname{tr}(X^{-1})} \geqslant \lambda \cdot \langle X^{-1}, Z_t^{-1} \rangle,$$

where the first inequality follows from (7.7) in Lemma 7.3.15, and the second inequality follows from our assumption of this lemma. This implies that

$$\mathbb{E}_t[\Gamma_t] \geqslant \frac{\lambda-1}{k} \cdot \langle X^{-1}, Z_t^{-1} \rangle = \frac{\lambda-1}{k} \cdot \operatorname{tr}\left(\left(\sum_{i \in S_{t-1}} u_i u_i^\top\right)^{-1}\right) \geqslant \frac{\lambda-1}{k} \cdot \operatorname{tr}(X^{-1}),$$

where the equality is from (7.2) and the last inequality is because X is an optimal solution. The lemmas follows by summing over t. □

7.3.3.2 Martingale Concentration Argument

Here we prove that the total progress is concentrated around the expectation. The proof uses the minimum eigenvalue assumption and the optimality condition in Lemma 7.3.18 in Section 7.3.3.4 to bound the variance of the random process.

Lemma 7.3.17. *Suppose $Z_t \succcurlyeq \frac{1}{4} I$ and $\langle X^{-1}, v_{i_t} v_{i_t}^\top \rangle \leqslant \frac{\varepsilon}{d} \cdot \operatorname{tr}(X^{-1})$ and $\langle X^{-1}, v_{j_t} v_{j_t}^\top \rangle \leqslant \frac{\varepsilon}{d} \cdot \operatorname{tr}(X^{-1})$ for all $1 \leqslant t \leqslant \tau$. Then, for any $\eta > 0$,*

$$\mathbb{P}\left[\sum_{t=1}^{\tau} \Gamma_t \leqslant \sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t] - \eta\right] \leqslant \exp\left(-\Omega\left(\frac{\eta^2 k d}{\varepsilon \tau \sqrt{d} \cdot \operatorname{tr}(X^{-1})^2 + \varepsilon \eta k \cdot \operatorname{tr}(X^{-1})}\right)\right).$$

Proof. We define two sequences of random variables $\{X_t\}_t$ and $\{Y_t\}_t$, where $X_t := \mathbb{E}_t[\Gamma_t] - \Gamma_t$ and $Y_t := \sum_{l=1}^{t} X_l$. It is easy to check that $\{Y_t\}_t$ is a martingale with respect to $\{S_t\}_t$. We will use Freedman's inequality to bound $\mathbb{P}[Y_\tau \geqslant \eta]$.

To apply Freedman's inequality, we need to upper bound X_t and $\mathbb{E}_t[X_t^2]$. To upper bound X_t, we first prove an upper bound on g_t and l_t. Note that

$$\begin{aligned}\langle X^{-1}, Z_t^{-1} v_{i_t} v_{i_t}^\top Z_t^{-1} \rangle = \langle Z_t^{-1} X^{-1} Z_t^{-1}, v_{i_t} v_{i_t}^\top \rangle &= \left\langle X^{\frac{1}{2}} \left(\sum_{j \in S_{t-1}} u_j u_j^\top\right)^{-2} X^{\frac{1}{2}}, v_{i_t} v_{i_t}^\top \right\rangle \\ &= \langle X^{-\frac{1}{2}} Z_t^{-2} X^{-\frac{1}{2}}, v_{i_t} v_{i_t}^\top \rangle \leqslant 16 \langle X^{-1}, v_{i_t} v_{i_t}^\top \rangle \leqslant \frac{16\varepsilon}{d} \cdot \operatorname{tr}(X^{-1}),\end{aligned}$$

where the second equality uses the fact that $Z_t = X^{-\frac{1}{2}} \left(\sum_{j \in S_{t-1}} u_j u_j^\top \right) X^{-\frac{1}{2}}$, the first inequality uses the assumption $Z_t \succcurlyeq \frac{1}{4} I$, and the last inequality follows from the assumption that $\langle X^{-1}, v_{i_t} v_{i_t}^\top \rangle \leqslant \frac{\varepsilon}{d} \cdot \mathrm{tr}(X^{-1})$. This implies that

$$g_t = \frac{\langle X^{-1}, Z_t^{-1} v_{j_t} v_{j_t}^\top Z_t^{-1} \rangle}{1 + 2\langle v_{j_t} v_{j_t}^\top, Z_t^{-1} \rangle} \leqslant \frac{16\varepsilon}{d} \cdot \mathrm{tr}(X^{-1}) \quad \text{and} \quad l_t = \frac{\langle X^{-1}, Z_t^{-1} v_{i_t} v_{i_t}^\top Z_t^{-1} \rangle}{1 - 2\langle v_{i_t} v_{i_t}^\top, Z_t^{-1} \rangle} \leqslant \frac{32\varepsilon}{d} \cdot \mathrm{tr}(X^{-1}),$$

where the second inequality holds as $2\langle v_{i_t} v_{i_t}^\top, Z_t^{-1} \rangle \leqslant 2\alpha \langle v_{i_t} v_{i_t}^\top, A_t^{\frac{1}{2}} \rangle \leqslant \frac{1}{2}$ by Lemma 7.3.14 and the definition that $i_t \in S'_{t-1}$ in the exchange subroutine. Therefore,

$$X_t = \mathbb{E}_t[\Gamma_t] - \Gamma_t \leqslant \mathbb{E}_t[g_t] + l_t \leqslant \frac{48\varepsilon}{d} \cdot \mathrm{tr}(X^{-1}).$$

Next, we upper bound $\mathbb{E}_t[X_t^2]$ by

$$\mathbb{E}_t[X_t^2] \leqslant \frac{48\varepsilon}{d} \cdot \mathrm{tr}(X^{-1}) \cdot \mathbb{E}_t[|X_t|] \leqslant \frac{96\varepsilon}{d} \cdot \mathrm{tr}(X^{-1}) \cdot \Big(\mathbb{E}_t[g_t] + \mathbb{E}_t[l_t] \Big).$$

Using (7.8), we bound the expected loss term by

$$\mathbb{E}_t[l_t] \leqslant \frac{1}{k} \cdot \langle X^{-1}, Z_t^{-1} \rangle \leqslant \frac{4}{k} \cdot \mathrm{tr}(X^{-1}),$$

where the last inequality follows by the assumption that $Z_t \succcurlyeq \frac{1}{4} I$. Then, we bound the expected gain term by

$$\begin{aligned}
\mathbb{E}_t[g_t] &= \sum_{j \in [n] \setminus S_{t-1}} \frac{x(j)}{k} \cdot \Big(1 + 2\alpha \langle v_j v_j^\top, A_t^{\frac{1}{2}} \rangle \Big) \cdot \frac{\langle X^{-1}, Z_t^{-1} v_j v_j^\top Z_t^{-1} \rangle}{1 + 2\langle v_j v_j^\top, Z_t^{-1} \rangle} \\
&\leqslant \frac{1}{k} \cdot \max_{j \in [n]} \left\{ \frac{\alpha \langle v_j v_j^\top, A_t^{\frac{1}{2}} \rangle}{\langle v_j v_j^\top, Z_t^{-1} \rangle} \right\} \cdot \sum_{j=1}^{n} x(j) \cdot \langle X^{-1}, Z_t^{-1} v_j v_j^\top Z_t^{-1} \rangle \\
&\leqslant \frac{1}{k} \cdot \alpha \lambda_{\min}(Z_t) \cdot \langle X^{-1}, Z_t^{-2} \rangle \\
&\leqslant \frac{32\sqrt{d}}{k} \cdot \mathrm{tr}\left(X^{-1} \right),
\end{aligned}$$

where the first inequality follows from the first inequality in Lemma 7.3.14, and the second inequality follows from the second inequality in Lemma 7.3.14 and $\sum_{j=1}^{n} x(j) \cdot v_j v_j^\top = I$, and the last inequality holds as $\alpha = 8\sqrt{d}$, $Z_t^{-2} \preccurlyeq \lambda_{\min}(Z_t)^{-2} I$, and $\lambda_{\min}(Z_t) \geqslant \frac{1}{4}$ by our assumption. Therefore,

$$\mathbb{E}_t[X_t^2] \leqslant O\Big(\frac{\varepsilon}{k\sqrt{d}}\Big) \cdot \operatorname{tr}\left(X^{-1}\right)^2.$$

Finally, we can apply Freedman's inequality Theorem 3.2.3 with $R = \frac{48\varepsilon}{d} \cdot \operatorname{tr}(X^{-1})$, $\sigma_t^2 = O\big(\frac{\varepsilon}{k\sqrt{d}}\big) \cdot \operatorname{tr}\left(X^{-1}\right)^2$ for all $t \in [\tau]$, and $\sigma^2 = O\big(\frac{\varepsilon\tau}{k\sqrt{d}}\big) \cdot \operatorname{tr}\left(X^{-1}\right)^2$ to conclude that

$$\mathbb{P}[Y_\tau \geqslant \eta] \leqslant \exp\left(-\frac{\eta^2/2}{\sigma^2 + R\eta/3}\right) = \exp\left(-\Omega\left(\frac{\eta^2 kd}{\varepsilon\tau\sqrt{d}\operatorname{tr}(X^{-1})^2 + \varepsilon\eta k \operatorname{tr}(X^{-1})}\right)\right).$$

The lemma follows by noting that $Y_\tau \geqslant \eta$ is equivalent to $\sum_{t=1}^{\tau} \Gamma_t \leqslant \sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t] - \eta$. □

7.3.3.3 Proof of Theorem 7.3.7 for A-Design

We are ready to prove Theorem 7.3.7 for A-design. Let $\tau = \frac{2k}{\varepsilon}$. Suppose the second phase of the algorithm has not terminated by time τ. Then $\lambda = \min_{1 \leqslant t \leqslant \tau+1} \frac{\langle X^{-1}, Z_t^{-1}\rangle}{\operatorname{tr}(X^{-1})} > (1+\varepsilon)$. Thus, Lemma 7.3.16 implies that

$$\sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t] \geqslant \frac{(\lambda-1)\tau}{k} \cdot \operatorname{tr}\left(X^{-1}\right) > 2\operatorname{tr}\left(X^{-1}\right).$$

On the other hand, the initial solution of the second phase satisfies $Z_1 \succcurlyeq \frac{3}{4} I$, which implies that $\langle X^{-1}, Z_1^{-1}\rangle \leqslant \frac{4}{3}\operatorname{tr}(X^{-1})$. By the minimum eigenvalue assumption, we know from Lemma 7.3.14 that $2\langle v_{i_t} v_{i_t}^\top, Z_t^{-1}\rangle \leqslant 2\alpha \cdot \langle v_{i_t} v_{i_t}^\top, A_t^{\frac{1}{2}}\rangle \leqslant \frac{1}{2}$, and so we can apply (7.5) to deduce that

$$\operatorname{tr}(X^{-1}) < \langle X^{-1}, Z_{\tau+1}^{-1}\rangle \leqslant \langle X^{-1}, Z_1^{-1}\rangle - \sum_{t=1}^{\tau}\Gamma_t \leqslant \frac{4}{3}\cdot\operatorname{tr}\left(X^{-1}\right) - \sum_{t=1}^{\tau}\Gamma_t$$
$$\implies \sum_{t=1}^{\tau}\Gamma_t \leqslant \frac{1}{3}\cdot\operatorname{tr}\left(X^{-1}\right).$$

As the knapsack constraints satisfy $b_j \geqslant \frac{d\|c_j\|_\infty}{\varepsilon}$ for $j \in [m]$, the optimality conditions in Lemma 7.3.18 imply that $\langle X^{-1}, v_i v_i^\top\rangle \leqslant \frac{\varepsilon}{d} \cdot \operatorname{tr}(X^{-1})$ for each i with $0 < x(i) < 1$. Note that, in the randomized exchange algorithm, all i_t and j_t satisfy $0 < x(i_t), x(j_t) < 1$ by

Observation 7.3.3. Therefore, we can apply Lemma 7.3.17 with $\eta = \frac{5}{3} \cdot \operatorname{tr}(X^{-1})$ and $\tau = \frac{2k}{\varepsilon}$ to conclude that

$$\begin{aligned}
&\mathbb{P}\left[\min_{1 \leqslant t \leqslant \tau+1} \langle X^{-1}, Z_t^{-1} \rangle > (1+\varepsilon) \operatorname{tr}(X^{-1})\right] \\
\leqslant\ & \mathbb{P}\left[\sum_{t=1}^{\tau} \Gamma_t < \sum_{t=1}^{\tau} \mathbb{E}_t[\Gamma_t] - \frac{5}{3} \cdot \operatorname{tr}(X^{-1})\right] \\
\leqslant\ & \exp\left(-\Omega\left(\frac{\operatorname{tr}(X^{-1})^2 \cdot kd}{\varepsilon\left(\frac{2k}{\varepsilon}\right)\sqrt{d} \cdot \operatorname{tr}\left(X^{-1}\right)^2 + \varepsilon k \cdot \operatorname{tr}\left(X^{-1}\right)^2}\right)\right) \\
\leqslant\ & \exp\left(-\Omega(\sqrt{d})\right).
\end{aligned}$$

7.3.3.4 Optimality Condition for the Convex Program of A-Design

This lemma follows from the optimality condition of the convex programming relaxation and the assumption about the budgets.

Lemma 7.3.18. *Let $x \in [0,1]^n$ be an optimal fractional solution of the convex programming relaxation (7.1) for A-design. Let $X = \sum_{i=1}^n x(i) \cdot u_i u_i^\top$, and $v_i = X^{-\frac{1}{2}} u_i$ for $1 \leqslant i \leqslant n$. Suppose $b_j \geqslant \frac{d\|c_j\|_\infty}{\varepsilon}$ for $1 \leqslant j \leqslant m$. Then, for each $1 \leqslant i \leqslant n$ with $0 < x(i) < 1$,*

$$\langle X^{-1}, v_i v_i^\top \rangle \leqslant \frac{\varepsilon}{d} \cdot \operatorname{tr}(X^{-1}).$$

Proof. We recall the convex relaxation (7.1) for A-design.

$$\begin{aligned}
\min_{x \in \mathbb{R}^d, X \in \mathbb{S}^d_{++}} \quad & \operatorname{tr}\left(X^{-1}\right) \\
\text{subject to} \quad & X = \sum\nolimits_{i=1}^{n} x(i) \cdot u_i u_i^\top \\
& \langle c_j, x \rangle \leqslant b_j, \qquad \forall j \in [m], \\
& 0 \leqslant x(i) \leqslant 1, \qquad \forall i \in [n].
\end{aligned}$$

The proof is very similar to the one of Lemma 7.3.13. We will use the Lagrangian duality (see Section 2.2.3.1) to investigate the length of the vectors v_i's. We introduce a dual variable Y for the first equality constraint, a dual variable $\mu_j \geqslant 0$ for each of the budget constraint $b_j - \langle c_j, x \rangle \geqslant 0$, a dual variable $\beta_i^- \geqslant 0$ for each non-negative constraint $x(i) \geqslant 0$, and a dual variable $\beta_i^+ \geqslant 0$ for each capacity constraint $1 - x(i) \geqslant 0$. The Lagrange function $L(x, X, Y, \mu, \beta^+, \beta^-)$ is defined as

$$L(x, X, Y, \mu, \beta^+, \beta^-) = \operatorname{tr}(X^{-1}) + \left\langle Y, X - \sum_{i=1}^{n} x(i) \cdot u_i u_i^\top \right\rangle + \sum_{j=1}^{m} \mu_j \Big(\langle c_j, x \rangle - b_j \Big) - \sum_{i=1}^{n} \beta_i^- x(i) + \sum_{i=1}^{n} \beta_i^+ (x(i) - 1),$$

Rearrange the terms, we have

$$L(x, X, Y, \mu, \beta^+, \beta^-) = \operatorname{tr}(X^{-1}) + \langle Y, X \rangle - \sum_{j=1}^{m} \mu_j b_j - \sum_{i=1}^{n} \beta_i^+ - \sum_{i=1}^{n} x(i) \cdot \left(\langle Y, u_i u_i^\top \rangle - \sum_{j=1}^{m} \mu_j c_j(i) + \beta_i^- - \beta_i^+ \right).$$

The Lagrangian dual function is

$$g(Y, \mu, \beta^+, \beta^-) = \max_{x, X \succ 0} L(x, X, Y, \mu, \beta^+, \beta^-).$$

It is easy to verify that $x = \delta \mathit{1}$ is a strictly feasible solution of the primal program for a small enough δ. By Theorem 2.2.26, Slater's condition implies that strong duality holds. In the primal convex program, we can relax the constraint $X \succ 0$ to $X \succcurlyeq 0$ without loss of generality, as $\operatorname{tr}(X^{-1})$ blows up when X approaches the boundary of $\mathbb{S}_+^d$. Thus, the feasible solution space is closed and bounded, and the primal optimal is attained. Let $x \in [0,1]^n, X \succ 0$ be an optimal solution for the primal program, Theorem 2.2.28 says there exists a dual optimal solution $Y, \mu, \beta^+, \beta^- \geqslant 0$ together with x, X satisfy the KKT conditions. In particular, it holds that (we recall $\nabla \operatorname{tr}(X^{-1}) = -X^{-2}$ by Fact 2.2.3)

$$\begin{aligned}
&\text{(Complementary slackness)} && \beta_i^- \cdot x(i) = 0,\ \beta_i^+ \cdot (1 - x(i)) = 0\ \forall i \in [n],\\
&\text{(Lagrangian optimality)} && \nabla_X L = -X^{-2} + Y = 0,\\
& && \nabla_{x(i)} L = \langle Y, u_i u_i^\top \rangle - \sum_{j=1}^m \mu_j c_j(i) + \beta_i^- - \beta_i^+ = 0,\ \forall i \in [n].
\end{aligned}$$

By strong duality, the primal optimal and the dual optimal attain the same objective value, i.e. $\operatorname{tr}(X^{-1}) = g(Y, \mu, \beta^+, \beta^-) = L(x, X, Y, \mu, \beta^+, \beta^-)$. Thus, it holds that

$$\operatorname{tr}(X^{-1}) = L(x, X, Y, \mu, \beta^+, \beta^-) = 2\operatorname{tr}(X^{-1}) - \sum_{j=1}^m \mu_j b_j - \sum_{i=1}^n \beta_i^+,$$

where the last equality holds by Lagrangian optimality, $\langle Y, u_i u_i^\top \rangle = \sum_{j=1}^m \mu_j c_j(i) - \beta_i^- + \beta_i^+ = 0$ for all $i \in [n]$ and $Y = X^{-2}$. Since $\beta^+ \geqslant 0$, it further implies

$$\sum_{j=1}^m \mu_j b_j \leqslant \operatorname{tr}(X^{-1}) \qquad \Longrightarrow \qquad \sum_{j=1}^m \mu_j \left\| c_j \right\|_\infty \leqslant \frac{\varepsilon}{d} \cdot \operatorname{tr}(X^{-1}),$$

where the last implication follows by the assumption $b_j \geqslant \frac{d\|c_j\|_\infty}{\varepsilon}$ for each $j \in [m]$.

Finally, by the complementary slackness condition, we must have $\beta_i^+ = \beta_i^- = 0$ for each i with $0 < x(i) < 1$. Together with the Lagrangian optimality, $Y = X^{-2}$ and $\langle Y, u_i u_i^\top \rangle = \sum_{j=1}^m \mu_j c_j(i) - \beta_i^- + \beta_i^+ = 0$ for all $i \in [n]$, for any $i \in [n]$ with $0 < x(i) < 1$ it holds that

$$\sum_{j=1}^m \mu_j c_j(i) = \langle Y, u_i u_i^\top \rangle = \langle X^{-1}, v_i v_i^\top \rangle \implies \langle X^{-1}, v_i v_i^\top \rangle \leqslant \sum_{j=1}^m \mu_j \left\| c_j \right\|_\infty \leqslant \frac{\varepsilon}{d} \cdot \operatorname{tr}(X^{-1}). \quad \square$$

7.4 Combinatorial Algorithms

In this section, we present combinatorial local search algorithms for D/A/E-design problems. In Section 7.4.1, we show that Fedorov's exchange method is a polynomial time algorithm to achieve $\frac{b-d-1}{b}$-approximation for D-design, which extends the result in [108] to the without repetition setting. In Section 7.4.2, we analyze Fedorov's exchange method for A-design, and prove that it works well as long as there is a well-conditioned optimal solution. As a corollary, this extends the result in [108] for A-design to the without repetition setting, with an arguably simpler proof. In Section 7.4.3, we show that Fedorov's exchange method does not work with the minimum eigenvalue objective, and we propose a modified local search algorithm and prove that it works well as long as there is a well-conditioned optimal solution.

A common theme in the analysis of all these algorithms is that we compare the current integral solution S to an optimal fractional solution x. As long as the objective value of x is significantly better than that of S, we use x to define two probability distributions to sample a pair of vectors $\mathsf{u}_i, \mathsf{u}_j$ so that the expected objective value of $S - i + j$ improves that of S considerably, and so we can conclude that the combinatorial algorithms will find such an improving exchange pair. One advantage of this approach is that this allows us the flexibility to compare with a fractional solution with smaller budget (which still has objective value close to the optimal one), and this makes the analysis easier and simpler.

The following notations will be used throughout this section. Given a fractional solution $\mathsf{x} \in [0,1]^n$ and an integral solution $S \subseteq [n]$, we denote

$$\mathsf{X} := \sum_{i=1}^{n} \mathsf{x}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top, \qquad \mathsf{X}_S := \sum_{i \in S} \mathsf{x}(i) \cdot \mathsf{u}_i \mathsf{u}_i^\top, \qquad \mathsf{x}(S) := \sum_{i \in S} \mathsf{x}(i), \qquad \mathsf{Z} := \sum_{i \in S} \mathsf{u}_i \mathsf{u}_i^\top.$$

7.4.1 Combinatorial Local Search Algorithm for D-Design

We analyze the following version of Fedorov's exchange method for D-design, where we always choose a pair that maximizes the improvement of the objective value and we stop as soon as the improvement is not large enough.

```
Fedorov's Exchange Method for D-Design
Input: $n$ vectors $u_1, ..., u_n \in \mathbb{R}^d$, a budget $b \geqslant d$.

1. Let $S_0 \subseteq [n]$ be an arbitrary set of full-rank vectors with $|S_0| = b$.
2. Let $t \leftarrow 1$ and $Z_1 := \sum_{i \in S_{t-1}} u_i u_i^\top$.
3. Repeat
    (a) Find $i_t \in S_{t-1}$ and $j_t \in [n] \setminus S_{t-1}$ such that
        $$(i_t, j_t) = \underset{(i,j): i \in S_{t-1}, j \in [n] \setminus S_{t-1}}{\arg\max} \det\left(Z_t - u_i u_i^\top + u_j u_j^\top\right).$$
    (b) Set $S_t \leftarrow S_{t-1} \cup \{j_t\} \setminus \{i_t\}$ and $Z_{t+1} \leftarrow Z_t - u_{i_t} u_{i_t}^\top + u_{j_t} u_{j_t}^\top$ and $t \leftarrow t+1$.
   Until $\det(Z_t) < \left(1 + \frac{d}{4b^3}\right) \det(Z_{t-1})$.
4. Return $S_{t-2}$ as the solution set.
```

To analyze the change of the objective value in each iteration, note that $\langle u_{i_t} u_{i_t}^\top, Z_t^{-1} \rangle \leqslant 1$ for any t as $i_t \in S_{t-1}$, and so it follows from Lemma 2.1.12 that

$$\det(Z_{t+1}) = \det(Z_t - u_{i_t} u_{i_t}^\top + u_{j_t} u_{j_t}^\top) \geqslant \det(Z_t) \cdot (1 - \underbrace{\langle u_{i_t} u_{i_t}^\top, Z_t^{-1} \rangle}_{\text{loss}}) \cdot (1 + \underbrace{\langle u_{j_t} u_{j_t}^\top, Z_t^{-1} \rangle}_{\text{gain}}).$$

Therefore, in order to lower bound the determinant of the solution, we lower bound the "gain" term and upper bound the "loss" term to quantify the progress in each iteration. First, we prove the existence of i_t with small loss, with respect to a fractional solution x with $\|x\|_1 = q < b$.

Lemma 7.4.1 (Loss)**.** *For any $x \in [0,1]^n$ with $\sum_{i=1}^n x(i) = q < b$ and any $S \subseteq [n]$ with $|S| = b$, there exists $i \in S$ with*

$$\langle u_i u_i^\top, Z^{-1} \rangle \leqslant \frac{d - \langle X_S, Z^{-1} \rangle}{b - x(S)}.$$

Proof. Consider the probability distribution of removing a vector u_i where each $i \in S$ is sampled with probability $\big(1 - \mathsf{x}(i)\big) / \sum_{j \in S} \big(1 - \mathsf{x}(j)\big)$, so that the "staying" probability is proportional to the value $\mathsf{x}(i)$. Note that the denominator is positive as $\mathsf{x}(S) \leqslant q < b$, and thus the probability distribution is well-defined. Then, the expected loss using this probability distribution is

$$\mathbb{E}\big[\langle \mathsf{u}_{i_t} \mathsf{u}_{i_t}^\top, \mathsf{Z}^{-1} \rangle\big] = \frac{\sum_{i \in S} \big(1 - \mathsf{x}(i)\big) \cdot \langle \mathsf{u}_i \mathsf{u}_i^\top, \mathsf{Z}^{-1} \rangle}{\sum_{j \in S} \big(1 - \mathsf{x}(j)\big)} = \frac{d - \langle \mathsf{X}_S, \mathsf{Z}^{-1} \rangle}{b - \mathsf{x}(S)},$$

where the last equality follows as $\sum_{i \in S} \mathsf{u}_i \mathsf{u}_i^\top = \mathsf{Z}$ and $|S| = b$. Therefore, there must exist one vector i with $\langle \mathsf{u}_i \mathsf{u}_i^\top, \mathsf{Z}^{-1} \rangle$ at most the expected value. □

Next, we prove the existence of j_t with large gain, again with respect to a fractional solution x with $\|\mathsf{x}\|_1 = q < b$.

Lemma 7.4.2 (Gain). *For any $\mathsf{x} \in [0,1]^n$ with $\sum_{i=1}^n \mathsf{x}(i) = q < b$ and any $S \subseteq [n]$ with $|S| = b$ and $\mathsf{x}(S) < q$, there exists $j \in [n] \backslash S$ with*

$$\langle \mathsf{u}_j \mathsf{u}_j^\top, \mathsf{Z}^{-1} \rangle \geqslant \frac{\langle \mathsf{X}, \mathsf{Z}^{-1} \rangle - \langle \mathsf{X}_S, \mathsf{Z}^{-1} \rangle}{q - \mathsf{x}(S)}.$$

Proof. Consider the probability distribution of adding a vector u_j where each $j \in [n] \setminus S$ is sampled with probability $\mathsf{x}(j) / \sum_{i \in [n] \backslash S} \mathsf{x}(j)$, so that the "adding" probability is proportional to the value $\mathsf{x}(i)$. Note that the denominator is positive by our assumption that $\mathsf{x}(S) < q$, and so the probability distribution is well-defined. Then, the expected gain using this probability distribution is

$$\mathbb{E}[\langle \mathsf{u}_j \mathsf{u}_j^\top, \mathsf{Z}^{-1} \rangle] = \frac{\sum_{j \in [n] \backslash S} \mathsf{x}(j) \cdot \langle \mathsf{u}_j \mathsf{u}_j^\top, \mathsf{Z}^{-1} \rangle}{\sum_{i \in [n] \backslash S} \mathsf{x}(i)} = \frac{\langle \mathsf{X}, \mathsf{Z}^{-1} \rangle - \langle \mathsf{X}_S, \mathsf{Z}^{-1} \rangle}{q - \mathsf{x}(S)}.$$

Therefore, there must exist one vector j with $\langle \mathsf{u}_j \mathsf{u}_j^\top, \mathsf{Z}^{-1} \rangle$ at least the expected value. □

The following is the main technical result for D-design, which lower bounds the improvement of the objective value in each iteration. In the proof, we compare our current integral solution S with size b to a fractional solution y with size $q = b - d - \frac{1}{2}$.

Proposition 7.4.3 (Progress)**.** *Let $x \in [0,1]^n$ be a feasible solution to the convex programming relaxation (7.1) for D-design with $\sum_{i=1}^n x(i) = b$ for $b \geqslant d+1$. Let Z_t be the current solution in the t-th iteration of Fedorov's exchange method. Then*

$$\det(Z_t)^{\frac{1}{d}} \leqslant \frac{b-d-1}{b} \cdot \det(X)^{\frac{1}{d}} \qquad \Longrightarrow \qquad \det(Z_{t+1}) \geqslant \Big(1 + \frac{d}{4b^3}\Big) \cdot \det(Z_t).$$

Proof. We consider the following scaled-down version y, Y of the fractional solution x, X. Define

$$q := b - d - \frac{1}{2}, \qquad y := \frac{q}{b} \cdot x, \qquad Y := \sum_{i=1}^n y(i) \cdot u_i u_i^\top = \frac{q}{b} \cdot X.$$

Note that $\det(Y)^{\frac{1}{d}} = \frac{q}{b} \cdot \det(X)^{\frac{1}{d}}$ and $\frac{1}{2} \leqslant q < b$. Let $S := S_{t-1}$ be the current solution set at time t. Note that we can assume $x(S) < b$ and hence $y(S) < q$, as otherwise $\det(Z_t) \geqslant \det(X)$ and there is nothing to prove. Hence, we can apply Lemma 7.4.1 and Lemma 7.4.2 on Y and S to ensure the existence of $i_t \in S$ and $j_t \in [n] \setminus S$ such that

$$\begin{aligned}
\det(Z_{t+1}) &\geqslant \det(Z_t) \cdot \bigg(1 - \frac{d - \langle Y_S, Z_t^{-1}\rangle}{b - y(S)}\bigg) \cdot \bigg(1 + \frac{\langle Y, Z_t^{-1}\rangle - \langle Y_S, Z_t^{-1}\rangle}{q - y(S)}\bigg) \\
&\geqslant \det(Z_t) \cdot \bigg(1 - \frac{d - \langle Y_S, Z_t^{-1}\rangle}{b - y(S)}\bigg) \cdot \bigg(1 + \frac{d \cdot \det(Y)^{\frac{1}{d}} \cdot \det(Z_t^{-1})^{\frac{1}{d}} - \langle Y_S, Z_t^{-1}\rangle}{q - y(S)}\bigg) \\
&\geqslant \det(Z_t) \cdot \bigg(1 - \frac{d - \langle Y_S, Z_t^{-1}\rangle}{b - y(S)}\bigg) \cdot \bigg(1 + \frac{d \cdot \frac{q}{b} \det(X)^{\frac{1}{d}} \cdot \frac{b}{b-d-1} \det(X)^{-\frac{1}{d}} - \langle Y_S, Z_t^{-1}\rangle}{q - y(S)}\bigg) \\
&= \det(Z_t) \cdot \bigg(1 - \frac{d - \langle Y_S, Z_t^{-1}\rangle}{b - y(S)}\bigg) \cdot \bigg(1 + \frac{(1 + \frac{1}{2q-1})d - \langle Y_S, Z_t^{-1}\rangle}{q - y(S)}\bigg),
\end{aligned}$$

where the second inequality follows from Lemma 7.3.10, the third inequality follows from $Y = \frac{q}{b} X$ and the assumption $\det(Z_t)^{\frac{1}{d}} \leqslant \frac{b-d-1}{b} \det(X)^{\frac{1}{d}}$, and the last equality is by $q = b - d - \frac{1}{2}$.

To lower bound the improvement, we write $a := d - \langle Y_S, Z_t^{-1}\rangle$ as a shorthand, and then the multiplicative factor is

$$\begin{aligned}
&\left(1-\frac{d-\langle Y_S, Z_t^{-1}\rangle}{b-y(S)}\right)\cdot\left(1+\frac{\left(1+\frac{1}{2q-1}\right)d-\langle Y_S, Z_t^{-1}\rangle}{q-y(S)}\right)\\
=\;&\left(1-\frac{a}{b-y(S)}\right)\cdot\left(1+\frac{a+\frac{d}{2q-1}}{q-y(S)}\right)\\
=\;&1+\frac{(b-q)a-a^2+\frac{(b-y(S))d}{2q-1}-\frac{ad}{2q-1}}{\big(b-y(S)\big)\cdot\big(q-y(S)\big)}\\
\geqslant\;&1+\frac{(b-q)a-a^2+\frac{(b-q)d}{2q-1}-\frac{ad}{2q-1}}{\big(b-y(S)\big)\cdot\big(q-y(S)\big)},
\end{aligned}$$

where the last inequality follows as $y(S) \leqslant q$. Let $f(x) = -x^2 + (b-q)x - \frac{dx}{2q-1} + \frac{(b-q)d}{2q-1}$ be a univariate quadratic function in x. Note that $f''(x) < 0$, and thus $\min_{x\in[x_1,x_2]} f(x)$ is attained at one of the two ends $x = x_1$ or $x = x_2$. Since $a = d - \langle Y_S, Z_t^{-1}\rangle \in [0, d]$, the numerator of the second term above is lower bounded by

$$\begin{aligned}
f(a) \geqslant \min_{x\in[0,d]} f(x) \geqslant \min\{f(0), f(d)\} &= \min\left\{\frac{(b-q)d}{2q-1}, \frac{2qd(b-q-d)}{2q-1}\right\}\\
&= \min\left\{\frac{(d+\frac{1}{2})d}{2(b-d-1)}, \frac{(b-d-\frac{1}{2})d}{2(b-d-1)}\right\} \geqslant \frac{d}{4(b-d-1)},
\end{aligned}$$

where the equality in the second line is by plugging in $q = b-d-\frac{1}{2}$, and the last inequality follows the assumption $b \geqslant d+1$. Therefore, we conclude that

$$\det(Z_{t+1}) \geqslant \det(Z_t)\cdot\left(1+\frac{d}{4(b-d-1)\big(b-y(S)\big)\big(q-y(S)\big)}\right) \geqslant \det(Z_t)\cdot\left(1+\frac{d}{4b^3}\right).\square$$

The main result in this subsection follows immediately from Proposition 7.4.3.

Theorem 7.1.4. *The Fedorov's exchange method is a $\frac{b-d-1}{b}$-approximation polynomial time algorithm for D-design in the without repetition setting. In particular, this is a $(1-\varepsilon)$-approximation algorithm whenever $b \geqslant d+1+\frac{d}{\varepsilon}$ for any $\varepsilon > 0$.*

Proof. Let $X^* = \sum_{i=1}^n x^*(i)\cdot u_i u_i^\top$ be an optimal fractional solution for D-design with budget b for $x^* \in [0,1]^n$. Let $Z_1 \succ 0$ be an arbitrary initial solution.

When the combinatorial local search algorithm terminates at the τ-th iteration, the termination condition implies that $\det(Z_{\tau+1}) < \left(1 + \frac{d}{4b^3}\right) \det(Z_\tau)$. It follows from Proposition 7.4.3 with $X = X^*$ that

$$\det(Z_\tau)^{\frac{1}{d}} \geqslant \frac{b-d-1}{b} \cdot \det(X^*)^{\frac{1}{d}},$$

and thus the returned solution of the Fedorov's exchange method is an $\frac{b-d-1}{b}$-approximate solution.

Finally, we bound the time complexity of the algorithm. If the algorithm runs for $\tau > \frac{8b^3}{d} \ln \frac{\det(X^*)}{\det(Z_1)}$ iterations, then the termination condition implies that the determinant of $Z_{\tau+1}$ is at least

$$\det(Z_{\tau+1}) \geqslant \left(1 + \frac{d}{4b^3}\right)^\tau \cdot \det(Z_1) \geqslant e^{\frac{d\tau}{8b^3}} \cdot \det(Z_1) > \det(X^*),$$

where the second inequality follows as $(1 + \frac{d}{4b^3}) \geqslant e^{\frac{d}{8b^3}}$ for $\frac{d}{4b^3} \leqslant \frac{1}{4}$. It was proved in Appendix C of [108] that $\ln \frac{\det(Z^*)}{\det(Z_1)}$ is polynomial in d, b and ℓ, where ℓ is the maximum number of bits to represent the numbers in the entries of the vectors. Specifically, they proved that $\det(Z_1) \geqslant 2^{-4(2b\ell+1)d^2}$ and $\det(X^*) \leqslant 2^{4(2n\ell+1)d^2}$, and so $\tau = O(db^3 n\ell)$ iterations of the algorithm is enough. □

7.4.2 Combinatorial Local Search Algorithm for A-Design

We analyze the following version of Fedorov's exchange method for A-design, where we always choose a pair that maximizes the improvement of the objective value and we stop as soon as the improvement is not large enough.

```
Fedorov's Exchange Method for A-Optimal Design
Input: $n$ vectors $u_1, ..., u_n \in \mathbb{R}^d$, a budget $b \geqslant d$, and an accuracy parameter $\varepsilon \in (0,1)$.

  1. Let $S_0 \subseteq [n]$ be an arbitrary set of full-rank vectors with $|S_0| = b$.
  2. Let $t \leftarrow 1$ and $Z_1 \leftarrow \sum_{i \in S_0} u_i u_i^\top$.
  3. Repeat
       (a) Find $i_t \in S_{t-1}$ and $j_t \in [n] \setminus S_{t-1}$ such that
             $(i_t, j_t) = \operatorname*{arg\,min}_{(i,j): i \in S_{t-1}, j \in [n] \setminus S_{t-1}} \operatorname{tr}\left(\left(Z_t - u_i u_i^\top + u_j u_j^\top\right)^{-1}\right).$
       (b) Set $S_t \leftarrow S_{t-1} \cup \{j_t\} \setminus \{i_t\}$ and $Z_{t+1} \leftarrow Z_t - u_{i_t} u_{i_t}^\top + u_{j_t} u_{j_t}^\top$ and $t \leftarrow t+1$.
     Until $\operatorname{tr}(Z_t^{-1}) > \left(1 - \frac{\varepsilon}{b}\right) \operatorname{tr}(Z_{t-1}^{-1})$.
  4. Return $S_{t-2}$ as the solution set.
```

To analyze the change of the objective value in each iteration, we apply Lemma 4.2.5 which states that if $2\langle u_{i_t} u_{i_t}^\top, Z_t^{-1} \rangle < 1$ then

$$\operatorname{tr}(Z_{t+1}^{-1}) - \operatorname{tr}(Z_t^{-1}) \leqslant \underbrace{\frac{\langle u_{i_t} u_{i_t}^\top, Z_t^{-2} \rangle}{1 - 2\langle u_{i_t} u_{i_t}^\top, Z_t^{-1} \rangle}}_{\text{loss}} - \underbrace{\frac{\langle u_{j_t} u_{j_t}^\top, Z_t^{-2} \rangle}{1 + 2\langle u_{j_t} u_{j_t}^\top, Z_t^{-1} \rangle}}_{\text{gain}}. \tag{7.9}$$

Therefore, to upper bound the A-design objective of the solution, we upper bound the loss term and lower bound the gain term to quantify the progress in each iteration.

In the following lemma, we first prove the existence of i_t with small loss term, with respect to a fractional solution x with $\|x\|_1 = q < b - 2d$. Note that we only restrict our choice of i_t to those vectors that satisfy $2\langle u_{i_t} u_{i_t}^\top, Z_t^{-1} \rangle < 1$ so that (7.9) applies, clearly Fedorov's exchange method could only do better by considering all possible vectors in the current solution.

Lemma 7.4.4 (Loss)**.** *For any* $x \in [0,1]^n$ *with* $\sum_{i=1}^n x(i) = q < b - 2d$ *and any* $S \subseteq [n]$ *with* $|S| = b$*, there exists* $i \in S' := \{j \in S : 2\langle u_j u_j^\top, Z^{-1}\rangle < 1\}$ *with*

$$\frac{\langle u_i u_i^\top, Z^{-2}\rangle}{1 - 2\langle u_i u_i^\top, Z^{-1}\rangle} \leqslant \frac{\operatorname{tr}(Z^{-1}) - \langle X_S, Z^{-2}\rangle}{b - x(S) - 2d}.$$

Proof. Consider the probability distribution of removing a vector u_i with probability

$$\mathbb{P}[i_t = i] = \frac{\big(1 - x(i)\big) \cdot \big(1 - 2\langle u_i u_i^\top, Z^{-1}\rangle\big)}{\sum_{j \in S'} \big(1 - x(j)\big) \cdot \big(1 - 2\langle u_j u_j^\top, Z^{-1}\rangle\big)} \quad \text{for each } i \in S'.$$

We first check that the probability distribution is well-defined. Note that the numerator is non-negative as $1 - 2\langle u_i u_i^\top, Z^{-1}\rangle > 0$ for each $i \in S'$. The denominator is

$$\begin{aligned}\sum_{j \in S'} \big(1 - x(j)\big) \cdot \big(1 - 2\langle u_j u_j^\top, Z^{-1}\rangle\big) &\geqslant \sum_{j \in S} \big(1 - x(j)\big) \cdot \big(1 - 2\langle u_j u_j^\top, Z^{-1}\rangle\big) \\ &\geqslant \sum_{j \in S} \big(1 - x(j)\big) - 2\sum_{j \in S} \langle u_j u_j^\top, Z^{-1}\rangle \\ &= b - x(S) - 2d \ > \ 0,\end{aligned}$$

where the first inequality holds as $1 - 2\langle u_j u_j^\top, Z^{-1}\rangle \leqslant 0$ for $j \in S \setminus S'$, the second inequality follows from $1 - x(j) \leqslant 1$ for each $j \in [n]$, and the equality is by $|S| = b$ and $\langle \sum_{j \in S} u_j u_j^\top, Z^{-1}\rangle = \langle Z, Z^{-1}\rangle = d$, and the strict inequality is by the assumption $b > q + 2d \geqslant x(S) + 2d$. Thus, $\mathbb{P}[i_t = i] \geqslant 0$ for each $i \in S'$, and clearly $\sum_{i \in S'} \mathbb{P}[i_t = i] = 1$.

The expected loss using this probability distribution is

$$\begin{aligned}\mathbb{E}\left[\frac{\langle u_{i_t} u_{i_t}^\top, Z^{-2}\rangle}{1 - 2\langle u_{i_t} u_{i_t}^\top, Z^{-1}\rangle}\right] &= \sum_{i \in S'} \frac{\big(1 - x(i)\big) \cdot \big(1 - 2\langle u_i u_i^\top, Z^{-1}\rangle\big)}{\sum_{j \in S'} \big(1 - x(j)\big) \cdot \big(1 - 2\langle u_j u_j^\top, Z^{-1}\rangle\big)} \cdot \frac{\langle u_i u_i^\top, Z^{-2}\rangle}{1 - 2\langle u_i u_i^\top, Z^{-1}\rangle} \\ &= \frac{\sum_{i \in S'} \big(1 - x(i)\big) \cdot \langle u_i u_i^\top, Z^{-2}\rangle}{\sum_{j \in S'} \big(1 - x(j)\big) \cdot \big(1 - 2\langle u_j u_j^\top, Z^{-1}\rangle\big)} \\ &\leqslant \frac{\operatorname{tr}(Z^{-1}) - \langle X_S, Z^{-2}\rangle}{b - x(S) - 2d},\end{aligned}$$

where the last inequality follows from the inequality above for the denominator and

$$\sum_{i\in S'} \big(1 - x(i)\big)\cdot \langle u_i u_i^\top, Z^{-2}\rangle \leqslant \sum_{i\in S} \big(1 - x(i)\big)\cdot \langle u_i u_i^\top, Z^{-2}\rangle$$
$$= \langle Z, Z^{-2}\rangle - \langle X_S, Z^{-2}\rangle = \mathrm{tr}(Z^{-1}) - \langle X_S, Z^{-2}\rangle$$

for the numerator. Therefore, there exists an $i \in S'$ with loss at most the expected value. $\square$

Next we show the existence of j_t with large gain term, again with respect to a fractional solution x.

Lemma 7.4.5 (Gain)**.** *For any $x \in [0,1]^n$ with $\sum_{i=1}^n x(i) = q < b$ and any $S \subseteq [n]$ with $|S| = b$ and $x(S) < q$, there exists $j \in [n]\setminus S$ with*

$$\frac{\langle u_j u_j^\top, Z^{-2}\rangle}{1 + 2\langle u_j u_j^\top, Z^{-1}\rangle} \geqslant \frac{\langle X, Z^{-2}\rangle - \langle X_S, Z^{-2}\rangle}{q - x(S) + 2\langle X, Z^{-1}\rangle}.$$

Proof. Consider the probability distribution of adding a vector u_j where each $j \in [n]\setminus S$ is sampled with probability

$$\mathbb{P}[j_t = j] = \frac{x(j)\cdot\big(1 + 2\langle u_j u_j^\top, Z^{-1}\rangle\big)}{\sum_{i\in[n]\setminus S}\ x(i)\cdot\big(1 + 2\langle u_i u_i^\top, Z^{-1}\rangle\big)} \quad \text{for each } j \in [n]\setminus S.$$

Note that the denominator is positive by the assumption $x(S) < q$ which implies that $x([n]\setminus S) > 0$, and so the probability distribution is well-defined.

The expected gain using this probability distribution is

$$\mathbb{E}\left[\frac{\langle u_{j_t} u_{j_t}^\top, Z^{-2}\rangle}{1 + 2\langle u_{j_t} u_{j_t}^\top, Z^{-1}\rangle}\right] = \sum_{j\in[n]\setminus S} \frac{x(j)\cdot\big(1 + 2\langle u_j u_j^\top, Z^{-1}\rangle\big)}{\sum_{i\in[n]\setminus S}\ x(i)\cdot\big(1 + 2\langle u_i u_i^\top, Z^{-1}\rangle\big)}\cdot\frac{\langle u_j u_j^\top, Z^{-2}\rangle}{1 + 2\langle u_j u_j^\top, Z^{-1}\rangle}$$
$$= \frac{\sum_{j\in[n]\setminus S}\ x(j)\cdot\langle u_j u_j^\top, Z^{-2}\rangle}{\sum_{i\in[n]\setminus S}\ x(i)\cdot\big(1 + 2\langle u_i u_i^\top, Z^{-1}\rangle\big)}$$
$$= \frac{\langle X, Z^{-2}\rangle - \langle X_S, Z^{-2}\rangle}{q - x(S) + \sum_{i\in[n]\setminus S} 2x(i)\cdot\langle u_i u_i^\top, Z^{-1}\rangle}$$
$$\geqslant \frac{\langle X, Z^{-2}\rangle - \langle X_S, Z^{-2}\rangle}{q - x(S) + 2\langle X, Z^{-1}\rangle},$$

where the third equality is by $\sum_{i=1}^{n} x(i) = q$, and the last inequality holds as $\sum_{i \in [n] \backslash S} x(i) \cdot u_i u_i^\top \preccurlyeq X$. Therefore, there exists $j \in [n] \backslash S$ with gain at least the expected value. □

We are about ready to analyze when the objective value would decrease. We will use the following simple claim to lower bound the gain term, whose proof is by checking the derivatives of $f(x)$ and $g(x)$.

Claim 7.4.6. *The functions* $f(x) = \frac{x - c_1}{c_2 + c_3 \sqrt{x}}$ *and* $g(x) = \frac{x - c_1}{c_2 + c_3 x}$ *with* $c_1, c_2, c_3 \geqslant 0$ *are monotone increasing for* $x \geqslant 0$.

The following is the main technical result for A-design, which lower bounds the improvement of the objective value in each iteration. Note that the result depends on $\mathrm{tr}(X) \cdot \mathrm{tr}(X^{-1})$.

Proposition 7.4.7 (Progress)**.** *Let* $x \in [0,1]^n$ *be a fractional solution with* $\sum_{i=1}^{n} x(i) = q$. *Let* Z_t *be the current solution in the* t*-th iteration of Fedorov's exchange method. For any* $\varepsilon > 0$, *if*

$$\mathrm{tr}(Z_t^{-1}) \geqslant (1+\varepsilon)\,\mathrm{tr}(X^{-1}) \qquad \textit{and} \qquad b \geqslant q + 2d + 2(1+\varepsilon)\sqrt{\mathrm{tr}(X) \cdot \mathrm{tr}(X^{-1})},$$

then

$$\mathrm{tr}\left(Z_{t+1}^{-1}\right) \leqslant \left(1 - \frac{\varepsilon}{b}\right) \cdot \mathrm{tr}\left(Z_t^{-1}\right).$$

Proof. Let $S := S_{t-1}$ be the current solution set at time t. Note that $x(S) < q$, as otherwise $\mathrm{tr}(Z^{-1}) \leqslant \mathrm{tr}(X^{-1})$ and the assumption does not hold. Hence, we can apply Lemma 7.4.5

to prove the existence of a $j_t \in [n] \backslash S$ such that the gain term is

$$\begin{aligned}
\frac{\langle u_{j_t} u_{j_t}^\top, Z^{-2}\rangle}{1 + 2\langle u_{j_t} u_{j_t}^\top, Z^{-1}\rangle} &\geqslant \frac{\langle X, Z^{-2}\rangle - \langle X_S, Z^{-2}\rangle}{q - x(S) + 2\langle X, Z^{-1}\rangle} \\
&\geqslant \frac{\langle X, Z^{-2}\rangle - \langle X_S, Z^{-2}\rangle}{q - x(S) + 2\sqrt{\operatorname{tr}(X) \cdot \langle X, Z^{-2}\rangle}} \\
&\geqslant \frac{\frac{\operatorname{tr}(Z^{-1})^2}{\operatorname{tr}(X^{-1})} - \langle X_S, Z^{-2}\rangle}{q - x(S) + 2\sqrt{\operatorname{tr}(X) \cdot \frac{\operatorname{tr}(Z^{-1})^2}{\operatorname{tr}(X^{-1})}}} \\
&= \frac{\frac{\operatorname{tr}(Z^{-1})}{\operatorname{tr}(X^{-1})} \cdot \operatorname{tr}(Z^{-1}) - \langle X_S, Z^{-2}\rangle}{q - x(S) + \frac{2\operatorname{tr}(Z^{-1})}{\operatorname{tr}(X^{-1})} \cdot \sqrt{\operatorname{tr}(X) \cdot \operatorname{tr}(X^{-1})}} \\
&\geqslant \frac{(1+\varepsilon)\operatorname{tr}(Z^{-1}) - \langle X_S, Z^{-2}\rangle}{q - x(S) + 2(1+\varepsilon)\sqrt{\operatorname{tr}(X) \cdot \operatorname{tr}(X^{-1})}} \\
&\geqslant \frac{(1+\varepsilon)\operatorname{tr}(Z^{-1}) - \langle X_S, Z^{-2}\rangle}{b - x(S) - 2d},
\end{aligned}$$

where the second inequality is by (7.7), the third inequality follows from $\langle X, Z^{-2}\rangle \geqslant \frac{(\operatorname{tr}(Z^{-1}))^2}{\operatorname{tr}(X^{-1})}$ by (7.6) and an application of Claim 7.4.6 with $f(x) = \frac{x - c_1}{c_2 + c_3\sqrt{x}}$ to establish monotonicity, the fourth inequality follows from the first assumption that $\operatorname{tr}(Z^{-1}) \geqslant (1+\varepsilon)\operatorname{tr}(X^{-1})$ and another application of Claim 7.4.6 with $g(x) = \frac{x - c_1}{c_2 + c_3 x}$ to establish monotonicity, and the last inequality follows from the second assumption that $b \geqslant q + 2d + 2(1+\varepsilon)\sqrt{\operatorname{tr}(X) \cdot \operatorname{tr}(X^{-1})}$.

For the loss term, note that $q < b - 2d$ by the assumption on b, and so we can apply Lemma 7.4.4 to prove the existence of an $i_t \in S' \subseteq S$ such that the loss term is

$$\frac{\langle u_{i_t} u_{i_t}^\top, Z^{-2}\rangle}{1 - 2\langle u_{i_t} u_{i_t}^\top, Z^{-1}\rangle} \leqslant \frac{\operatorname{tr}(Z^{-1}) - \langle X_S, Z^{-2}\rangle}{b - x(S) - 2d}.$$

Since $i_t \in S'$ satisfies $2\langle u_{i_t} u_{i_t}^\top, Z_t^{-1}\rangle < 1$, we can apply (7.9) to conclude that

$$\begin{aligned}
&\operatorname{tr}(Z_{t+1}^{-1}) - \operatorname{tr}(Z_t^{-1}) = \operatorname{tr}\left((Z_t - u_{i_t} u_{i_t}^\top + u_{j_t} u_{j_t}^\top)^{-1}\right) - \operatorname{tr}\left(Z_t^{-1}\right) \\
\leqslant\ & \frac{\langle u_{i_t} u_{i_t}^\top, Z^{-2}\rangle}{1 - 2\langle u_{i_t} u_{i_t}^\top, Z^{-1}\rangle} - \frac{\langle u_{j_t} u_{j_t}^\top, Z^{-2}\rangle}{1 + 2\langle u_{j_t} u_{j_t}^\top, Z^{-1}\rangle} \leqslant \frac{-\varepsilon\operatorname{tr}(Z_t^{-1})}{b - x(S) - 2d} \leqslant -\frac{\varepsilon}{b}\operatorname{tr}(Z_t^{-1}).
\end{aligned}$$

□

The main result in this subsection follows from Proposition 7.4.7 by a simple scaling argument.

Theorem 7.1.5. *Let $X := \sum_{i=1}^{n} x(i) \cdot u_i u_i^{\top}$ with $\sum_{i=1}^{n} x(i) = b$ and $x_i \in [0,1]$ for $1 \leqslant i \leqslant n$ be a fractional solution to A-design. For any $\varepsilon \in (0,1)$, the Fedorov's exchange method returns an integral solution $Z = \sum_{i=1}^{n} z(i) \cdot u_i u_i^{\top}$ with $\sum_{i=1}^{n} z(i) \leqslant b$ and $z(i) \in \{0,1\}$ for $1 \leqslant i \leqslant n$ such that*

$$\operatorname{tr}\left(Z^{-1}\right) \leqslant (1+\varepsilon) \cdot \operatorname{tr}(X^{-1}) \quad \textit{whenever} \quad b \geqslant \Omega\Big(\frac{d + \sqrt{\operatorname{tr}(X) \operatorname{tr}\left(X^{-1}\right)}}{\varepsilon}\Big).$$

In particular, let $\kappa = \frac{\lambda_{\max}(X^*)}{\lambda_{\min}(X^*)}$ be the condition number of an optimal solution X^*, then the Fedorov's exchange method gives a $(1+\varepsilon)$-approximation algorithm for A-design whenever $b \geqslant \Omega\Big(\frac{\left(1+\sqrt{\kappa}\right)\cdot d}{\varepsilon}\Big)$, and the time complexity is polynomial in $n, d, \frac{1}{\varepsilon}, \kappa$.

Proof. We consider the following scaled-down version y, Y of the fractional solution x, X. Let

$$q := b - 2d - 2(1+\varepsilon)\sqrt{\operatorname{tr}(X) \cdot \operatorname{tr}((X)^{-1})}, \qquad y := \frac{q}{b} \cdot x, \qquad Y := \sum_{i=1}^{n} y(i) \cdot u_i u_i^{\top} = \frac{q}{b} \cdot X.$$

Note that $\operatorname{tr}(Y) \cdot \operatorname{tr}\left(Y^{-1}\right) = \operatorname{tr}(X) \cdot \operatorname{tr}\left(X^{-1}\right)$ and so it holds that $b \geqslant q + 2d + 2(1+\varepsilon)\sqrt{\operatorname{tr}(Y) \cdot \operatorname{tr}(Y^{-1})}$. Thus, we can apply Proposition 7.4.7 on y to conclude that if the algorithm terminates at the τ-th iteration such that $\operatorname{tr}\left(Z_{\tau+1}^{-1}\right) > \left(1 - \frac{\varepsilon}{b}\right) \operatorname{tr}\left(Z_{\tau}^{-1}\right)$ then

$$\operatorname{tr}\left(Z_{\tau}^{-1}\right) < (1+\varepsilon) \cdot \operatorname{tr}\left(Y^{-1}\right) = \frac{(1+\varepsilon)b}{q} \cdot \operatorname{tr}\left(X^{-1}\right) \leqslant \left(1 + O(\varepsilon)\right) \cdot \operatorname{tr}\left(X^{-1}\right),$$

where the last inequality follows from the assumption $b = \Omega\Big(\frac{1}{\varepsilon}\big(d + \sqrt{\operatorname{tr}(X) \operatorname{tr}(X^{-1})}\big)\Big)$ which implies that $q \geqslant (1 - O(\varepsilon))b$. This proves the approximation guarantee of the returned solution.

Finally, we bound the time complexity of the algorithm. If the algorithm runs for $\tau > \frac{b}{\varepsilon} \ln \frac{\operatorname{tr}(Z_1^{-1})}{\operatorname{tr}(X^{-1})}$ iterations, then the termination condition implies that the objective value of $Z_{\tau+1}$ is at most

$$\operatorname{tr}(Z_{\tau+1}^{-1}) \leqslant \left(1 - \frac{\varepsilon}{b}\right)^{\tau} \cdot \operatorname{tr}\left(Z_1^{-1}\right) \leqslant e^{-\frac{\varepsilon \tau}{b}} \cdot \operatorname{tr}\left(Z_1^{-1}\right) \leqslant \operatorname{tr}\left(X^{-1}\right).$$

Note that $\ln \frac{\operatorname{tr}(Z_1^{-1})}{\operatorname{tr}(X^{-1})}$ is upper bounded by a polynomial in d, n and the input size as proved in [108] (and the corresponding bound for D-design is discussed in the proof of Theorem 7.1.4 in Section 7.4.1). □

As a corollary, we extend the analysis of Fedorov's exchange method with short input vectors in [108] to the more general without repetition setting.

Corollary 7.4.8. *Let $x \in [0,1]^n$ be a fractional solution to the convex programming relaxation (7.1) for A-design with $\sum_{i=1}^n x(i) = b$. If $\|u_i\|^2 \leqslant \frac{\varepsilon^2 b}{2\operatorname{tr}(X^{-1})}$ for each $1 \leqslant i \leqslant n$ and $b \geqslant \Omega\left(\frac{d}{\varepsilon}\right)$ for some $\varepsilon \in (0,1)$, then Fedorov's exchange method for A-design returns a solution with at most b vectors with objective value at most $\left(1 + O(\varepsilon)\right) \cdot \operatorname{tr}\left(X^{-1}\right)$ in polynomial time.*

Proof. It follows from the assumption $\|u_i\|^2 \leqslant \frac{\varepsilon^2 b}{2\operatorname{tr}(X^{-1})}$ that

$$\operatorname{tr}\left(X^{-1}\right) \cdot \operatorname{tr}(X) = \operatorname{tr}\left(X^{-1}\right) \cdot \sum_{i=1}^{n} x(i) \cdot \|u_i\|_2^2 \leqslant \operatorname{tr}(X^{-1}) \cdot \frac{\varepsilon^2 b^2}{2\operatorname{tr}(X^{-1})} = \frac{\varepsilon^2 b^2}{2}.$$

Thus, for $b \geqslant \Omega\left(\frac{d}{\varepsilon}\right)$, it holds that $b \geqslant \Omega\left(\frac{1}{\varepsilon}\left(d + \sqrt{\varepsilon^2 b^2/2}\right)\right) = \Omega\left(\frac{1}{\varepsilon}\left(d + \sqrt{\operatorname{tr}(X)\operatorname{tr}(X^{-1})}\right)\right)$, and so Theorem 7.1.5 implies that Fedorov's exchange method will find a $(1 + O(\varepsilon))$-approximate solution in polynomial time. □

7.4.3 Combinatorial Local Search Algorithm for E-Design

Unlike D-design and A-design, there are simple examples (see Section 7.4.3.2) showing that Fedorov's exchange method does not work for E-design, even if there is a well-conditioned optimal solution.

Instead, we prove that the rounding algorithm by Allen-Zhu, Li, Singh and Wang [6] for E-design can be used as a combinatorial local search algorithm as well. The only difference is that the rounding algorithm in [6] will first compute an optimal fractional solution x to the convex programming relaxation and then perform a linear transformation so that $\sum_{i=1}^{n} x(i) \cdot u_i u_i^\top = I$, before applying the following combinatorial algorithm. Our analysis will show that the combinatorial algorithm works well as long as there is an approximately optimal fractional solution with good condition number, so this tells us that the only essential use of an optimal fractional solution in the rounding algorithm is for preconditioning.

The following algorithm assumes the knowledge of the objective value λ^* of the targeted fractional solution. We will guess this value in the proof of Theorem 7.1.6.

Combinatorial Local Search Algorithm for E-Optimal Design

```
Input: $n$ vectors $u_1, ..., u_n \in \mathbb{R}^d$, a budget $b \geqslant d$, an accuracy parameter $\varepsilon \in (0,1)$, and a targeted objective value $\lambda^*$.

1. Initialization: Let $S_0 \subseteq [n]$ be an arbitrary set with $|S_0| = b$. Set $\alpha \leftarrow \frac{\sqrt{d}}{\varepsilon\lambda^*}$ and $t \leftarrow 0$.
2. Repeat
    (a) Set $t \leftarrow t+1$.
    (b) Let $Z_t := \sum_{i \in S_{t-1}} u_i u_i^\top$. Compute $A_t \leftarrow (\alpha Z_t - l_t I)^{-2}$ where $l_t \in \mathbb{R}$ is the unique scalar such that $A_t \succ 0$ and $\mathrm{tr}(A_t) = 1$.
    (c) Let $S'_{t-1} := \{i \in S_{t-1} : 2\alpha\langle u_i u_i^\top, A_t^{1/2}\rangle < 1\}$.
    (d) Find $i_t \in S'_{t-1}$ and $j_t \in [n] \setminus S_{t-1}$ such that
        $$(i_t, j_t) = \operatorname*{arg\,max}_{(i,j):\ i \in S'_{t-1},\ j \in [n]\setminus S_{t-1}} \Phi(A_t, i, j) := \frac{\langle u_j u_j^\top, A_t\rangle}{1 + 2\alpha\langle u_j u_j^\top, A_t^{\frac{1}{2}}\rangle} - \frac{\langle u_i u_i^\top, A_t\rangle}{1 - 2\alpha\langle u_i u_i^\top, A_t^{\frac{1}{2}}\rangle}.$$
    (e) Set $S_t \leftarrow S_{t-1} \cup \{j_t\} \setminus \{i_t\}$.
   Until $\Phi(A_t, i_t, j_t) < \frac{\varepsilon\lambda^*}{b}$ or $\lambda_{\min}(Z_t) \geqslant (1-2\varepsilon)\lambda^*$.
3. Return $S_{t-1}$ as the solution set.
```

The regret minimization framework developed in [7, 6] bounds the minimum eigenvalue of the current solution using the potential functions $\Phi(A_t, i, j)$ that we are optimizing in each iteration. Applying Theorem 4.2.7 with feedback matrices $F_0 = Z_1 \succcurlyeq 0$ and $F_t = u_{j_t} u_{j_t}^\top - u_{i_t} u_{i_t}^\top$ for $t \geqslant 1$, as long as $1 > 2\alpha\langle u_{i_t} u_{i_t}^\top, A_t^{\frac{1}{2}}\rangle$ for all $1 \leqslant t \leqslant \tau$, we have

$$\lambda_{\min}(Z_{\tau+1}) \geqslant \sum_{t=1}^{\tau} \left(\underbrace{\frac{\langle u_{j_t} u_{j_t}^\top, A_t\rangle}{1 + 2\alpha\langle u_{j_t} u_{j_t}^\top, A_t^{\frac{1}{2}}\rangle}}_{\text{gain}} - \underbrace{\frac{\langle u_{i_t} u_{i_t}^\top, A_t\rangle}{1 - 2\alpha\langle u_{i_t} u_{i_t}^\top, A_t^{\frac{1}{2}}\rangle}}_{\text{loss}} \right) - \frac{2\sqrt{d}}{\alpha} = \sum_{t=1}^{\tau} \Phi(A_t, i_t, j_t) - \frac{2\sqrt{d}}{\alpha}. \tag{7.10}$$

Therefore, in order to lower bound the minimum eigenvalue of the solution, we upper bound the loss term and lower bound the gain term to quantify the progress in each iteration.

First, we show the existence of a good i_t with small loss, with respect to a fractional solution x.

Lemma 7.4.9 (Loss). *Let $S := S_{t-1}$, $S' := S'_{t-1}$, $Z := Z_t$ and $A := A_t$. For any $x \in [0,1]^n$ with $\sum_{i=1}^n x(i) = q < b - 2\alpha\langle Z, A^{\frac{1}{2}}\rangle$, there exists $i \in S'$ with*

$$\frac{\langle u_i u_i^\top, A\rangle}{1 - 2\alpha\langle u_i u_i^\top, A^{\frac{1}{2}}\rangle} \leqslant \frac{\langle Z, A\rangle - \langle X_S, A\rangle}{b - x(S) - 2\alpha\langle Z, A^{\frac{1}{2}}\rangle}.$$

Proof. Consider the probability distribution of removing a vector u_i with probability

$$\mathbb{P}[i_t = i] = \frac{(1 - x(i))(1 - 2\alpha\langle u_i u_i^\top, A^{\frac{1}{2}}\rangle)}{\sum_{j \in S'}(1 - x(j))(1 - 2\alpha\langle u_j u_j^\top, A^{\frac{1}{2}}\rangle)} \quad \text{for all } i \in S'.$$

We check that the probability distribution is well-defined. Note that the numerator is non-negative as $1 - 2\alpha\langle u_i u_i^\top, A^{\frac{1}{2}}\rangle > 0$ for each $i \in S'$. The denominator is

$$\begin{aligned}\sum_{j \in S'} \big(1 - x(j)\big) \cdot \big(1 - 2\alpha\langle u_j u_j^\top, A^{\frac{1}{2}}\rangle\big) &\geqslant \sum_{j \in S} \big(1 - x(j)\big) \cdot \big(1 - 2\alpha\langle u_j u_j^\top, A^{\frac{1}{2}}\rangle\big) \\ &\geqslant \sum_{j \in S} \big(1 - x(j)\big) - 2\alpha \sum_{j \in S} \langle u_j u_j^\top, A^{\frac{1}{2}}\rangle \\ &= b - x(S) - 2\alpha\langle Z, A^{\frac{1}{2}}\rangle > 0\end{aligned}$$

where the first inequality holds as $1-2\alpha\langle u_j u_j^\top, A^{\frac{1}{2}}\rangle \leqslant 0$ for $j \in S \setminus S'$, the second inequality follows from $1 - x(j) \leqslant 1$ for each $j \in [n]$, and the equality is by $|S| = b$, and the strict inequality is by the assumption $b > q + 2\alpha\langle Z, A^{\frac{1}{2}}\rangle \geqslant x(S) + 2\alpha\langle Z, A^{\frac{1}{2}}\rangle$. Thus, $\mathbb{P}[i_t = i] \geqslant 0$ for each $i \in S'$, and clearly $\sum_{i \in S'} \mathbb{P}[i_t = i] = 1$.

The expected loss using this probability distribution is

$$\begin{aligned}
\mathbb{E}\left[\frac{\langle u_{i_t} u_{i_t}^\top, A\rangle}{1 - 2\alpha\langle u_{i_t} u_{i_t}^\top, A^{\frac{1}{2}}\rangle}\right] &= \sum_{i \in S'} \frac{\big(1 - x(i)\big) \cdot \big(1 - 2\alpha\langle u_i u_i^\top, A^{\frac{1}{2}}\rangle\big)}{\sum_{j \in S'} \big(1 - x(j)\big) \cdot \big(1 - 2\alpha\langle u_j u_j^\top, A^{\frac{1}{2}}\rangle\big)} \cdot \frac{\langle u_i u_i^\top, A\rangle}{1 - 2\alpha\langle u_i u_i^\top, A^{\frac{1}{2}}\rangle} \\
&= \frac{\sum_{i \in S'} \big(1 - x(i)\big) \cdot \langle u_i u_i^\top, A\rangle}{\sum_{j \in S'} \big(1 - x(j)\big) \cdot \big(1 - 2\alpha\langle u_j u_j^\top, A^{\frac{1}{2}}\rangle\big)} \\
&\leqslant \frac{\langle Z, A\rangle - \langle X_S, A\rangle}{b - x(S) - 2\alpha\langle Z, A^{\frac{1}{2}}\rangle},
\end{aligned}$$

where the inequality is from the above inequality for the denominator and

$$\sum_{i \in S'} \big(1 - x(i)\big) \cdot \langle u_i u_i^\top, A\rangle \leqslant \sum_{i \in S} \big(1 - x(i)\big) \cdot \langle u_i u_i^\top, A\rangle = \langle Z, A\rangle - \langle X_S, A\rangle$$

for the numerator. Therefore, there exists an $i \in S'$ with loss at most the expected value. □

Next, we show the existence of j_t with large gain term, again with respect to a fractional solution.

Lemma 7.4.10 (Gain). *Let $S := S_{t-1}$ and $A := A_t$. For any $x \in [0,1]^n$ with $\sum_{i=1}^n x(i) = q < b$ and $x(S) < q$, there exists $j \in [n] \setminus S$ with*

$$\frac{\langle u_j u_j^\top, A\rangle}{1 + 2\alpha\langle u_j u_j^\top, A^{\frac{1}{2}}\rangle} \geqslant \frac{\langle X, A\rangle - \langle X_S, A\rangle}{q - x(S) + 2\alpha\langle X, A^{\frac{1}{2}}\rangle}.$$

Proof. Consider the probability distribution of adding a vector u_j where each $j \in [n] \setminus S$ is sampled with probability

$$\mathbb{P}[j_t = j] = \frac{x(j) \cdot \big(1 + 2\alpha\langle u_j u_j^\top, A^{\frac{1}{2}}\rangle\big)}{\sum_{i \in [n] \setminus S} x(i) \cdot \big(1 + 2\alpha\langle u_i u_i^\top, A^{\frac{1}{2}}\rangle\big)} \quad \text{for each } j \in [n] \setminus S.$$

Note that the denominator is positive by the assumption $x(S) < q$ which implies that $x([n] \setminus S) > 0$.

The expected gain with respect to this probability distribution is

$$\begin{aligned}
\mathbb{E}\left[\frac{\langle u_{j_t} u_{j_t}^\top, A\rangle}{1 + 2\alpha\langle u_{j_t} u_{j_t}^\top, A^{\frac{1}{2}}\rangle}\right] &= \sum_{j \in [n]\setminus S} \frac{x(j) \cdot \left(1 + 2\alpha\langle u_j u_j^\top, A^{\frac{1}{2}}\rangle\right)}{\sum_{i\in[n]\setminus S}\ x(i)\left(1 + 2\alpha\langle u_i u_i^\top, A^{\frac{1}{2}}\rangle\right)} \cdot \frac{\langle u_j u_j^\top, A\rangle}{1 + 2\alpha\langle u_j u_j^\top, A^{\frac{1}{2}}\rangle} \\
&= \frac{\sum_{j\in[n]\setminus S}\ x(j) \cdot \langle u_j u_j^\top, A\rangle}{\sum_{i\in[n]\setminus S}\ x(i) \cdot \left(1 + 2\alpha\langle u_i u_i^\top, A^{\frac{1}{2}}\rangle\right)} \\
&= \frac{\langle X, A\rangle - \langle X_S, A\rangle}{q - x(S) + 2\alpha\sum_{i\in[n]\setminus S}\ x(i)\langle u_i u_i^\top, A^{\frac{1}{2}}\rangle} \\
&\geqslant \frac{\langle X, A\rangle - \langle X_S, A\rangle}{q - x(S) + 2\alpha\langle X, A^{\frac{1}{2}}\rangle}.
\end{aligned}$$

where the third equality is by $\sum_{i=1}^n x(i) = q$ and the last inequality holds as $\sum_{i\in[n]\setminus S} x(i) \cdot u_i u_i^\top \preccurlyeq X$. Therefore, there exist $j \in [n] \setminus S$ with gain at least the expected value. □

The following is the main technical result for E-design, which lower bounds the improvement of the potential function in each iteration. Note that the result depends on the condition number of the fractional solution.

Proposition 7.4.11 (Progress)**.** *Let $x \in [0,1]^n$ be a fractional solution with $\sum_{i=1}^n x(i) = q$. Let $Z_t = \sum_{i\in S_{t-1}} u_i u_i^\top$ be the current solution in the t-th iteration. For any $0 < \varepsilon < \frac{1}{2}$, if*

$$\alpha = \frac{\sqrt{d}}{\varepsilon \cdot \lambda_{\min}(X)}, \quad \lambda_{\min}(Z_t) \leqslant (1 - 2\varepsilon) \cdot \lambda_{\min}(X), \quad \text{and} \quad b \geqslant q + 2\Big(d + \frac{d}{\varepsilon}\Big) + \frac{2d}{\varepsilon}\sqrt{\frac{\lambda_{\text{avg}}(X)}{\lambda_{\min}(X)}}$$

where $\lambda_{\text{avg}}(X) = \frac{\operatorname{tr}(X)}{d}$ is the average eigenvalue of X, then the value of the potential function is

$$\Phi(A_t, i_t, j_t) = \frac{\langle u_{j_t} u_{j_t}^\top, A_t\rangle}{1 + 2\alpha\langle u_{j_t} u_{j_t}^\top, A_t^{\frac{1}{2}}\rangle} - \frac{\langle u_{i_t} u_{i_t}^\top, A_t\rangle}{1 - 2\alpha\langle u_{i_t} u_{i_t}^\top, A_t^{\frac{1}{2}}\rangle} \geqslant \frac{\varepsilon}{b} \cdot \lambda_{\min}(X).$$

Proof. Let $S := S_{t-1}$ be the current solution set at time t, $A = A_t$ and $Z = Z_t$. Note that $x(S) < q$, as otherwise $\lambda_{\min}(Z) \geqslant \lambda_{\min}(X)$ and the assumption does not hold. Hence, we can apply Lemma 7.4.10 to show the existence of $j_t \in [n] \backslash S$ with gain

$$\begin{aligned}\frac{\langle u_{j_t} u_{j_t}^\top, A\rangle}{1 + 2\alpha\langle u_{j_t} u_{j_t}^\top, A^{\frac{1}{2}}\rangle} &\geqslant \frac{\langle X, A\rangle - \langle X_S, A\rangle}{q - x(S) + 2\alpha\langle X, A^{\frac{1}{2}}\rangle} \geqslant \frac{\langle X, A\rangle - \langle X_S, A\rangle}{q - x(S) + 2\alpha\sqrt{\operatorname{tr}(X)\cdot\langle X, A\rangle}} \\ &\geqslant \frac{\lambda_{\min}(X) - \langle X_S, A\rangle}{q - x(S) + 2\alpha\sqrt{\operatorname{tr}(X)\cdot\lambda_{\min}(X)}} = \frac{\lambda_{\min}(X) - \langle X_S, A\rangle}{q - x(S) + \frac{2d}{\varepsilon}\sqrt{\frac{\lambda_{\text{avg}}(X)}{\lambda_{\min}(X)}}},\end{aligned}$$

where the second inequality in the first line is by (7.7) in Lemma 7.3.15, the first inequality in the second line is by Claim 7.4.6 and the fact that $\langle X, A\rangle \geqslant \langle\lambda_{\min}(X)\cdot I, A\rangle \geqslant \lambda_{\min}(X)$ as $\operatorname{tr}(A) = 1$, and the last equality is by the choice $\alpha = \frac{\sqrt{d}}{\varepsilon\lambda_{\min}(X)}$ and the definition of $\lambda_{\text{avg}}(X)$.

For the loss term, we need to check the condition that $b > q + 2\alpha\langle Z_t, A^{\frac{1}{2}}\rangle$ before applying Lemma 7.4.9. It follows from Lemma 4.2.9 and the assumptions of α, $\lambda_{\min}(Z)$, and b that

$$\alpha\langle Z, A^{\frac{1}{2}}\rangle \leqslant d + \alpha\sqrt{d}\cdot\lambda_{\min}(Z) < d + \frac{d}{\varepsilon} \quad\Longrightarrow\quad b > q + 2\alpha\cdot\langle Z, A^{\frac{1}{2}}\rangle.$$

Hence, Lemma 7.4.9 implies the existence of an $i_t \in S$ with loss

$$\begin{aligned}\frac{\langle u_{i_t} u_{i_t}^\top, A\rangle}{1 - 2\alpha\langle u_{i_t} u_{i_t}^\top, A^{\frac{1}{2}}\rangle} \leqslant \frac{\langle Z_t, A\rangle - \langle X_S, A\rangle}{b - x(S) - 2\alpha\langle Z, A^{\frac{1}{2}}\rangle} &\leqslant \frac{\lambda_{\min}(Z) + \frac{\sqrt{d}}{\alpha} - \langle X_S, A\rangle}{b - x(S) - 2\left(d + \frac{d}{\varepsilon}\right)} \\ &\leqslant \frac{(1-\varepsilon)\lambda_{\min}(X) - \langle X_S, A\rangle}{q - x(S) + \frac{2d}{\varepsilon}\sqrt{\frac{\lambda_{\text{avg}}(X)}{\lambda_{\min}(X)}}},\end{aligned}$$

where the second inequality is by Lemma 4.2.9 and the inequality above about $\alpha\langle Z, A^{\frac{1}{2}}\rangle$, and the last inequality is by our assumptions about α, $\lambda_{\min}(Z)$ and b.

Therefore, we conclude that the progress in each iteration is

$$\Phi(A, i_t, j_t) = \frac{\langle u_{j_t} u_{j_t}^\top, A\rangle}{1 + 2\alpha\langle u_{j_t} u_{j_t}^\top, A^{\frac{1}{2}}\rangle} - \frac{\langle u_{i_t} u_{i_t}^\top, A\rangle}{1 - 2\alpha\langle u_{i_t} u_{i_t}^\top, A^{\frac{1}{2}}\rangle} \geqslant \frac{\varepsilon\cdot\lambda_{\min}(X)}{q - x(S) + \frac{2d}{\varepsilon}\sqrt{\frac{\lambda_{\text{avg}}(X)}{\lambda_{\min}(X)}}} \geqslant \frac{\varepsilon}{b}\cdot\lambda_{\min}(X),$$

where the last inequality follows from the assumption about b. □

By guessing the targeted objective value, the main result in this subsection follows from Proposition 7.4.11 by a simple scaling argument.

Theorem 7.1.6. *Let $X := \sum_{i=1}^{n} x(i) \cdot u_i u_i^\top$ with $\sum_{i=1}^{n} x(i) = b$ and $x(i) \in [0,1]$ for $1 \leqslant i \leqslant n$ be a fractional solution to E-design. For any $\varepsilon \in (0,1)$, there is a combinatorial local search algorithm which returns an integral solution $Z = \sum_{i=1}^{n} z(i) \cdot u_i u_i^\top$ with $\sum_{i=1}^{n} z(i) \leqslant b$ and $z(i) \in \{0,1\}$ for $1 \leqslant i \leqslant n$ such that*

$$\lambda_{\min}(Z) \geqslant (1 - O(\varepsilon)) \cdot \lambda_{\min}(X) \quad \textit{whenever} \quad b \geqslant \Omega\left(\frac{d}{\varepsilon^2}\sqrt{\frac{\lambda_{\text{avg}}(X)}{\lambda_{\min}(X)}}\right),$$

where $\lambda_{avg}(X) = \frac{\operatorname{tr}(X)}{d}$ is the average eigenvalue of X.

In particular, let $\kappa = \frac{\lambda_{\max}(X^*)}{\lambda_{\min}(X^*)}$ be the condition number of an optimal solution X^*, then the combinatorial local search method gives a polynomial time $(1-\varepsilon)$-approximation algorithm for E-design whenever $b \geqslant \Omega\left(\frac{d\sqrt{\kappa}}{\varepsilon^2}\right)$, and the time complexity is polynomial in $n, d, \frac{1}{\varepsilon}, \kappa$.

Proof. We consider the following scaled-down version y, Y of the fractional solution x, X. Let

$$q = b - 2\left(d + \frac{d}{\varepsilon}\right) - \frac{2d}{\varepsilon}\sqrt{\frac{\lambda_{\text{avg}}(X)}{\lambda_{\min}(X)}}, \qquad y := \frac{q}{b} \cdot x, \qquad Y := \sum_{i=1}^{n} y(i) \cdot u_i u_i^\top = \frac{q}{b} \cdot X.$$

Note that $\lambda_l(Y) = \frac{q}{b} \cdot \lambda_l(X)$ for each $1 \leqslant l \leqslant d$, and this implies that $b = q + 2\left(d + \frac{d}{\varepsilon}\right) + \frac{2d}{\varepsilon}\sqrt{\frac{\lambda_{\text{avg}}(Y)}{\lambda_{\min}(Y)}}$.

Suppose the combinatorial local search algorithm is running with the accuracy parameter ε and $\lambda^* := \lambda_{\min}(Y)$ and terminates at the τ-th iteration. If $\Phi(A_\tau, i_\tau, j_\tau) < \frac{\varepsilon}{b} \cdot \lambda_{\min}(Y)$, then we can apply Proposition 7.4.11 on y to conclude that

$$\lambda_{\min}(Z_\tau) > (1 - 2\varepsilon) \cdot \lambda_{\min}(Y) = \frac{(1-2\varepsilon)q}{b} \cdot \lambda_{\min}(X) \geqslant \big(1 - O(\varepsilon)\big) \cdot \lambda_{\min}(X),$$

where the last inequality is by the assumption that $b = \Omega\left(\frac{d}{\varepsilon^2}\sqrt{\frac{\lambda_{\text{avg}}(X)}{\lambda_{\min}(X)}}\right)$. This proves the approximate guarantee of the returned solution if the algorithm is run with $\lambda^* = \lambda_{\min}(Y)$.

Our final algorithm runs the local search algorithm on different values of λ^*. Initially, we start from an upper bound on $\lambda_{\min}(X)$ by setting $\lambda^* = \lambda_{\min}\left(\sum_{i=1}^{n} u_i u_i^\top\right)$. Then it runs the local search algorithm with targeted objective value λ^*. If the returned solution Z satisfies $\lambda_{\min}(Z) \geqslant \big(1 - O(\varepsilon)\big) \cdot \lambda^*$ then it stops and returns Z as our final solution; otherwise, we set $\lambda^* \leftarrow (1-\varepsilon) \cdot \lambda^*$ and repeat until the first time that the local search algorithm finds a solution with $\lambda_{\min}(Z) \geqslant \big(1 - O(\varepsilon)\big) \cdot \lambda^*$. For correctness, it is enough to show that the algorithm will stop when $(1-\varepsilon) \cdot \lambda_{\min}(X) \leqslant \lambda^* \leqslant \lambda_{\min}(X)$. This follows by applying the argument in the previous paragraphs on $X' := \frac{\lambda^*}{\lambda_{\min}(X)} \cdot X$, so that $\lambda_{\min}(X') = \lambda^*$ and then the returned solution Z' will satisfy $\lambda_{\min}(Z') \geqslant \big(1 - O(\varepsilon)\big) \cdot \lambda_{\min}(X') \geqslant \big(1 - O(\varepsilon)\big) \cdot \lambda_{\min}(X)$.

Finally, we bound the time complexity of the algorithm. Note that $\frac{b}{n} \sum_{i=1}^{b} u_i u_i^\top$ is a feasible solution with objective value $\frac{b}{n} \cdot \lambda_{\min}\left(\sum_{i=1}^{n} u_i u_i^\top\right)$. This implies that the number of executions of the local search algorithm is at most $O\left(\frac{1}{\varepsilon} \log \frac{n}{b}\right)$. In each execution with a fixed λ^*, if the algorithm runs for $\tau \geqslant \frac{b}{\varepsilon}$ iterations, the termination condition together with (7.10) imply that

$$\lambda_{\min}(Z_{\tau+1}) \geqslant \sum_{t=1}^{\tau} \Phi(A_t, i_t, j_t) - \frac{2\sqrt{d}}{\alpha} \geqslant \tau\left(\frac{\varepsilon\lambda^*}{b}\right) - 2\varepsilon\lambda^* > (1-2\varepsilon)\lambda^*,$$

and so it would stop. Thus, the total number of iterations is at most $O\left(\frac{b}{\varepsilon^2} \log \frac{n}{b}\right)$. Each iteration can be implemented in polynomial time as shown in [6]. □

The following is a corollary in the short vector setting.

Corollary 7.4.12. *Let $x \in [0,1]^n$ be a fractional solution to the E-design problem with budget b. For any $0 < \varepsilon < 1$, if $\|u_i\|^2 \leqslant \varepsilon^2 \cdot \lambda_{\min}(X)$ for $1 \leqslant i \leqslant n$ and $b \geqslant \Omega\left(\frac{d}{\varepsilon^2}\right)$, then the combinatorial local search algorithm for E-design returns a solution with at most b vectors and objective value at least $\big(1 - O(\varepsilon)\big) \cdot \lambda_{\min}(X)$ in polynomial time.*

Proof. It follows from the assumption $\|u_i\|^2 \leqslant \varepsilon^2 \cdot \lambda_{\min}(X)$ that

$$\lambda_{\text{avg}}(X) = \frac{\operatorname{tr}(X)}{d} \leqslant \frac{b\varepsilon^2 \cdot \lambda_{\min}(X)}{d} \qquad \Longrightarrow \qquad \frac{2d}{\varepsilon^2}\sqrt{\frac{\lambda_{\text{avg}}(X)}{\lambda_{\min}(X)}} \leqslant \frac{2\sqrt{bd}}{\varepsilon}.$$

Thus, for $b \geqslant \Omega(\frac{d}{\varepsilon^2})$, it follows that $b \geqslant \Omega(\frac{\sqrt{bd}}{\varepsilon}) \geqslant \Omega\Big(\frac{d}{\varepsilon^2}\sqrt{\frac{\lambda_{\text{avg}}(X)}{\lambda_{\min}(X)}}\Big)$, and so Theorem 7.1.6 implies that the combinatorial local search algorithm will find a $\big(1 - O(\varepsilon)\big)$-approximate solution in polynomial time. □

7.4.3.1 Maximizing Algebraic Connectivity

In this problem, we are given a graph $G = (V, E)$ with Laplacian matrix $L_G = \sum_{e \in E} b_e b_e^\top$, and the goal is to find a subgraph H with at most b edges to maximize $\lambda_2(L_H)$. This problem is known as maximizing algebraic connectivity in the literature (see Section 6.2.1 for more background about the problem). It is a special case of E-design and Theorem 7.1.6 bounds the performance guarantee of a simple combinatorial local search algorithm.

Corollary 7.1.8. *For any $0 < \varepsilon < 1$, there is a polynomial time combinatorial $(1 - \varepsilon)$-approximation algorithm for maximizing algebraic connectivity in an unweighted graph whenever $b \geqslant \Omega\left(\frac{n}{\varepsilon^4 \lambda_2^*}\right)$, where λ_2^* is the optimal value for the problem.*

Proof. Note that this is an E-design problem by a similar transformation as in (6.1). Let H^* be an optimal subgraph with b edges. Note that $\lambda_{\text{avg}}(L_{H^*}) = \frac{\operatorname{tr}(L_{H^*})}{n} \leqslant \frac{2b}{n}$, and so

$$\begin{aligned} b \geqslant \Omega\left(\frac{n}{\varepsilon^4 \cdot \lambda_2(L_{H^*})}\right) \quad &\Longrightarrow \quad \sqrt{b} \geqslant \Omega\left(\frac{1}{\varepsilon^2}\sqrt{\frac{n}{\lambda_2(L_{H^*})}}\right) \\ &\Longrightarrow \quad b \geqslant \Omega\left(\frac{n}{\varepsilon^2}\sqrt{\frac{2b}{n\lambda_2(L_{H^*})}}\right) \geqslant \Omega\left(\frac{n}{\varepsilon^2}\sqrt{\frac{\lambda_{\text{avg}}(L_{H^*})}{\lambda_2(L_{H^*})}}\right). \end{aligned}$$

Therefore, by Theorem 7.1.6, the combinatorial local search algorithm for E-design returns a subgraph H with $\lambda_2(L_H) \geqslant \big(1 - O(\varepsilon)\big) \cdot \lambda_2(L_{H^*})$ in polynomial time whenever $b \geqslant \Omega\left(\frac{n}{\varepsilon^4 \lambda_2(L_{H^*})}\right)$. □

7.4.3.2 Bad Examples for Local Search Algorithms

We first present a simple example showing that Fedorov's exchange method does not work with the E-design objective function, even if there is a well-conditioned optimal solution. The reason is simply that the E-design objective function is not smooth and sometimes it is impossible to improve it by an exchange operation.

Example 7.4.13. *Suppose the input vectors $v_1, \ldots, v_n$ are in $\mathbb{R}^d$ for some $d \geqslant 3$. Suppose that we have an initial solution set $S_0 \subseteq [n]$ such that $Z_1 = \sum_{i \in S_0} v_i v_i^\top = I$. For any $i_1 \in S_0$*

and $j_1 \in [n] \backslash S_0$, note that $\lambda_{\min}(\mathsf{Z}_1 - \mathsf{v}_{i_1}\mathsf{v}_{i_1}^\top + \mathsf{v}_{j_1}\mathsf{v}_{j_1}^\top) \leqslant 1$. Therefore, Fedorov's method fails to improve the objective value even if there is a well-conditioned optimal solution say $N\mathsf{e}_1, \dots, N\mathsf{e}_d$ for a large N.

Then, we present an example where all exchanges strictly decrease the minimum eigenvalue, even though the current solution is far away from the well-conditioned optimal solution.

Example 7.4.14. *Let $N \geqslant 0$ be some large scalar. The input contains exactly $\frac{b}{2}$ copies of each $\mathsf{v}_1, \mathsf{v}_2, \mathsf{w}_1, \mathsf{w}_2 \in \mathbb{R}^2$ defined as follows:*

$$\mathsf{v}_1\mathsf{v}_1^\top = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \mathsf{v}_2\mathsf{v}_2^\top = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\mathsf{w}_1\mathsf{w}_1^\top = \frac{N}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathsf{w}_2\mathsf{w}_2^\top = \frac{N}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

The optimal solution Z^* contains $\frac{b}{2}$ copies of $\mathsf{w}_1\mathsf{w}_1^\top$ and $\mathsf{w}_2\mathsf{w}_2^\top$. Suppose the algorithm starts with the solution Z_1 containing $\frac{b}{2}$ copies of $\mathsf{v}_1\mathsf{v}_1^\top$ and $\mathsf{v}_2\mathsf{v}_2^\top$ such that

$$\mathsf{Z}^* = \begin{pmatrix} \frac{bN}{2} & \\ & \frac{bN}{2} \end{pmatrix} \text{ with } \lambda_{\min}(\mathsf{Z}^*) = \frac{bN}{2}, \qquad \mathsf{Z}_1 = \begin{pmatrix} \frac{b}{2} & \\ & \frac{b}{2} \end{pmatrix} \text{ with } \lambda_{\min}(\mathsf{Z}_1) = \frac{b}{2}.$$

Without loss of generality, we assume the exchange step removes v_1 and adds w_1. After the exchange, the solution is

$$\mathsf{Z}_2 = \begin{pmatrix} \frac{b+N}{2} - 1 & \frac{N}{2} \\ \frac{N}{2} & \frac{b+N}{2} \end{pmatrix}.$$

We can verify that the minimum eigenvalues of Z_2 is $\frac{b-1+N-\sqrt{N^2-1}}{2}$, which tends to $\frac{b-1}{2}$ when $N \to \infty$. Since all other exchanges are symmetric, we conclude that all exchanges will decrease the objective value by $\frac{1}{2}$, and thus Fedorov's exchange method fails.

Finally, we adopt an example by Madan, Singh, Tantipongpipat and Xie [108] to show that even we use a smooth objective function from the regret minimization framework, the combinatorial local search algorithm may return bad solution when there are no well-conditioned optimal solutions.

Example 7.4.15. *Let $N \geqslant 0$ be some large scalar. The input contains $M \gg b \geqslant 3$ copies of each $v_1, v_2, w_1, w_2 \in \mathbb{R}^2$ defined as follows:*

$$v_1 v_1^\top = \begin{pmatrix} N^2 & 1 \\ 1 & \frac{1}{N^2} \end{pmatrix}, \qquad v_2 v_2^\top = \begin{pmatrix} N^2 & -1 \\ -1 & \frac{1}{N^2} \end{pmatrix},$$
$$w_1 w_1^\top = \frac{1}{b} \cdot \begin{pmatrix} N^8 & N^4 \\ N^4 & 1 \end{pmatrix}, \qquad w_2 w_2^\top = \frac{1}{b} \cdot \begin{pmatrix} N^8 & -N^4 \\ -N^4 & 1 \end{pmatrix}.$$

Lemma 7.4.16. *The combinatorial local search algorithm proposed in this subsection may return a solution with an unbounded approximation ratio.*

Proof. Note that $\frac{b}{2}$ copies of $w_1 w_1^\top$ and $\frac{b}{2}$ copies of $w_2 w_2^\top$ form an optimal solution Z^* with budget b such that

$$Z^* = \begin{pmatrix} N^8 & \\ & 1 \end{pmatrix} \qquad \text{and} \qquad \lambda_{\min}(Z^*) = 1.$$

So our algorithm will choose $\alpha = \frac{\sqrt{d}}{\varepsilon \lambda_{\min}(Z^*)} = \frac{\sqrt{d}}{\varepsilon} = \frac{\sqrt{2}}{\varepsilon}$.

Consider an initial solution Z containing $\frac{b}{2}$ copies of $v_1 v_1^\top$ and $\frac{b}{2}$ copies of $v_2 v_2^\top$ such that

$$Z = \begin{pmatrix} bN^2 & \\ & \frac{b}{N^2} \end{pmatrix} \qquad \text{and} \qquad \lambda_{\min}(Z) = \frac{b}{N^2}.$$

The approximation ratio between Z and Z^* is $\frac{N^2}{b}$, which is unbounded for fixed b when $N \to \infty$.

With Z as the current solution, the action matrix A is

$$A = (\alpha Z - lI)^{-2} = \begin{pmatrix} \frac{\sqrt{2}bN^2}{\varepsilon} - l & \\ & \frac{\sqrt{2}b}{\varepsilon N^2} - l \end{pmatrix}^{-2} \approx \begin{pmatrix} \frac{\varepsilon^2}{2b^2 N^4} & \\ & 1 \end{pmatrix},$$

where the last approximate equality holds when $N \to \infty$ as $\operatorname{tr}(A) = 1$.

The loss of removing a vector v_1 (removing v_2 is similar) from the current solution is

$$\frac{\langle v_1 v_1^\top, A \rangle}{1 - 2\alpha \langle v_1 v_1^\top, A^{\frac{1}{2}} \rangle} \approx \frac{\frac{\varepsilon^2}{2b^2 N^2} + \frac{1}{N^2}}{1 - \frac{2\sqrt{2}}{\varepsilon}\left(\frac{\varepsilon}{\sqrt{2}b} + \frac{1}{N^2}\right)} \geqslant \frac{1}{N^2},$$

where we used $b \geqslant 3$ and N is large for the last inequality.

The gain of adding vector v_2 is strictly less than the loss of removing v_1

$$\frac{\langle v_2 v_2^\top, A\rangle}{1 + 2\alpha\langle v_2 v_2^\top, A^{\frac{1}{2}}\rangle} < \frac{\langle v_1 v_1^\top, A\rangle}{1 - 2\alpha\langle v_1 v_1^\top, A^{\frac{1}{2}}\rangle},$$

as $\langle v_2 v_2^\top, A\rangle = \langle v_1 v_1, A\rangle$ and $\langle v_2 v_2^\top, A^{\frac{1}{2}}\rangle = \langle v_1 v_1, A^{\frac{1}{2}}\rangle$. Also, the gain of adding a vector w_1 (adding w_2 is similar) to the current solution is

$$\frac{\langle w_1 w_1^\top, A\rangle}{1 + 2\alpha\langle w_1 w_1^\top, A^{\frac{1}{2}}\rangle} \approx \frac{\frac{\varepsilon^2 N^4}{2b^3} + \frac{1}{b}}{1 + \frac{2\sqrt{2}}{\varepsilon}\left(\frac{\varepsilon N^6}{\sqrt{2}b^2} + \frac{1}{b}\right)} \leqslant \frac{\varepsilon^2}{4bN^2} + \frac{b}{2N^6}.$$

For fixed $b \geq 3$ and $\varepsilon \leq 1$, this gain is always less than the loss when $N \to \infty$. Therefore, the combinatorial local search algorithm will stop and return the initial solution Z. □

Chapter 8

Network Design for s-t Effective Resistance

In this chapter, we study a basic problem in designing networks with a spectral requirement – the effective resistance between two vertices.

Definition (The s-t effective resistance network design problem)**.** *The input is an undirected graph $G = (V, E)$, two specified vertices $s, t \in V$, and a budget k. The goal is to find a subgraph H of G with at most k edges that minimizes* $\mathrm{Reff}_H(s,t)$*, where* $\mathrm{Reff}_H(s,t)$ *denotes the effective resistance between s and t in the subgraph H. See Section 2.4 for the definition of effective resistance and Section 8.2.1 for a mathematical formulation of the problem.*

The s-t effective resistance is an interpolation between s-t shortest path distance and s-t edge connectivity. To see this, let $f \in \mathbb{R}^{|E|}$ be a unit s-t flow in G and define the ℓ_p-energy of f as $\mathcal{E}_p(f) := (\sum_e |f(e)|^p)^{\frac{1}{p}}$, and let $\mathcal{E}_p(s,t) := \min_f \{\mathcal{E}_p(f) \mid f \text{ is a unit } s\text{-}t \text{ flow}\}$ be the minimum ℓ_p-energy of a unit s-t flow that the graph G can support. Thomson's principle (see Theorem 2.4.3) states that $\mathrm{Reff}_G(s,t) = \mathcal{E}_2^2(s,t)$, so that a graph of small s-t effective resistance can support a unit s-t flow with small ℓ_2-energy. Note that the shortest path distance between s and t is $\mathcal{E}_1(s,t)$ (as the ℓ_1-energy of a flow is just the average path length and is minimized by a shortest s-t path), and so a graph with small $\mathcal{E}_1(s,t)$ has a short path between s and t. Note also that the edge-connectivity between s and t is equal to the reciprocal of $\mathcal{E}_\infty(s,t)$ (because if there are k edge-disjoint s-t paths, we can set the flow value on each path to be $\frac{1}{k}$), and so a graph with small $\mathcal{E}_\infty(s,t)$ has many edge-disjoint s-t paths. As ℓ_2 is between ℓ_1 and ℓ_∞, the objective function $\mathrm{Reff}(s,t) = \mathcal{E}_2^2(s,t)$ takes both the s-t shortest path distance and the s-t edge-connectivity into consideration.

A simple property suggests that ℓ_2-energy may be even more desirable than ℓ_1 and ℓ_∞ as a connectivity measure. Conceptually, adding an edge e to G would make s and t more connected. For ℓ_1 and ℓ_∞, however, adding e would not yield a better energy if e does not improve the shortest path and the edge connectivity respectively. In contrast, the ℓ_2-energy would typically be improved after adding an edge, and so ℓ_2-energy provides a smoother quantitative measure that better captures our intuition how well s and t are connected in a network.

Traditionally, the effective resistance has many useful probabilistic interpretations, such as the commute time [39], the cover time [112], and the probability of an edge in a random spanning tree [85]. These interpretations suggest that the effective resistance is a useful distance function and have applications in the study of social networks. Recently, effective resistance has found surprising applications in solving problems about graph connectivity, including constructing spectral sparsifiers [132] (by using the effective resistance of an edge as the sampling probability), computing maximum flow [44, 109, 120], finding thin trees for ATSP [9], and generating random spanning trees [115, 128].

Thomson's principle also states that the electrical flow between s and t is the unique flow that minimizes the ℓ_2-energy (see Section 2.4 for a proof). So, designing a network with small s-t effective resistance has natural applications in designing electrical networks [55, 70, 78]. One natural formulation is to keep at most k wires in the input electrical network to minimize $\mathrm{Reff}(s,t)$, so that the electrical flow between s and t can still be sent with small energy while we switch off many wires in the electrical network.

Based on the above reasons, we believe that the effective resistance is a nice and natural alternative connectivity measure in network design. More generally, it is an interesting direction to develop techniques to solve network design problems with spectral requirements.

In this chapter, we explore both hardness and algorithmic results of the s-t effective resistance network design problem.

8.1 Our Contributions

8.1.1 Main Results

Unlike the classical problems of shortest path and min-cost flow (corresponding to the ℓ_1 and ℓ_∞ versions of the problem), the s-t effective resistance network design problem is NP-hard.

Theorem 8.1.1. *The s-t effective resistance network design problem is* NP-*hard.*

On the other hand, we would like to design good approximation algorithms for this problem. As we only want to connect a single pair of vertices, the budget k could be much less than the number of vertices in the graph. In this regime, spectral rounding, the main technique in this thesis, cannot provide a good approximation ratio (as discussed in Section 6.2.3). Thus, we need to go beyond spectral rounding to design a constant approximation algorithm. The following is the main algorithmic result in this chapter.

Theorem 8.1.2. *There is a convex programming based randomized algorithm that returns an 8-approximate solution in polynomial time with high probability for the s-t effective resistance network design problem.*

The algorithm crucially uses a nice characterization of the optimal solutions to the convex program (Lemma 8.2.2) to design a randomized path-rounding procedure (Section 8.2.2) for Theorem 8.1.2.

A simple example shows that the integrality gap of the convex program is at least two. When the budget k is much larger than the length of a shortest s-t path, we show how to achieve an approximation ratio close to two with a randomized “short” path rounding algorithm (Section 8.2.5).

Theorem 8.1.3. *There is a $(2+O(\varepsilon))$-approximation algorithm for the s-t effective resistance network design problem, when $k \geqslant \frac{2d_{st}}{\varepsilon^{10}}$ where d_{st} is the length of a shortest s-t path.*

8.1.2 Other Results

We consider some variants of the s-t effective resistance network design problem, including the weighted version, the dual version.

There is a natural weighted generalization of the s-t effective resistance network design problem, where we associate a cost $\mathsf{c}(e)$ and resistance $\mathsf{r}(e)$ to each edge e of the input graph.

Definition (The weighted s-t effective resistance network design problem)**.** *The input is an undirected graph $G=(V,E)$ where each edge e has a non-negative cost $\mathsf{c}(e)$ and a non-negative resistance $\mathsf{r}(e)$, two specified vertices $s,t \in V$, and a cost budget k. The goal is to find a subgraph H of G that minimizes $\mathrm{Reff}_H(s,t)$ subject to the constraint that the total edge cost of H is at most k. In the following, we may refer to this problem as the weighted problem for simplicity.*

In the weighted problem, the integrality gap of the convex program (Section 8.2.1) becomes unbounded, even when the cost on the edges are the same ($\mathsf{c}(e) = 1$ for all $e \in E$). This suggests that the weighted version may be strictly harder. Indeed, we show stronger hardness result for the weighted problem assuming the small-set expansion conjecture [123, 124].

Theorem 8.1.4. *Assuming the small-set expansion conjecture, it is NP-hard to approximate the weighted s-t effective resistance network design problem within a factor of $2-\varepsilon$ for any $\varepsilon > 0$, even when $\mathsf{c}(e) = 1$ for every edge e.*

On the other hand, when the cost on the edges are the same, the following approximation follows from the randomized path rounding algorithm in a black box manner.

Corollary 8.1.5. *There is a convex programming based $O(R)$-approximation randomized algorithm for the weighted s-t effective resistance network design problem when $\mathsf{c}(e) = 1$ for every edge e, where $R = \max_e \mathsf{r}(e) / \min_e \mathsf{r}(e)$ is the ratio between the maximum and minimum resistance.*

We also consider the "dual" problem where we set the effective resistance as a hard constraint, and the objective is to minimize the number of edges in the solution subgraph. We present similar results as the original problem in Section 8.2.6.

8.1.3 Related Work

In the survivable network design problem, we are given an undirected graph and a connectivity requirement r_{uv} for every pair of vertices u, v, and the goal is to find a minimum cost subgraph such that there are at least r_{uv} edge-disjoint paths for all u, v. This problem is extensively studied and captures many interesting special cases [71, 1, 73, 67]. The best approximation algorithm for this problem is due to Jain [79], who introduced the technique of iterative rounding to design a 2-approximation algorithm. His result has been extended in various directions, including element-connectivity [62, 43], directed graphs [66, 67], and with degree constraints [93, 56, 64, 96].

Other combinatorial connectivity requirements were also considered. A natural variation is to require $r_{u,v}$ internally vertex disjoint paths for every pair of vertices u, v. This problem is much harder to approximate [87, 92], but there are good approximation algorithms for global connectivity [58, 42] and when the maximum connectivity requirement is small [36, 46]. Another natural problem is to require a path of length $l_{u,v}$ between every pair of vertices u, v. This problem is also hard to approximate in general but there are better approximation algorithms when every edge has the same cost and the same length [53].

Spectral connectivity requirements were also studied, including algebraic connectivity [69, 86] (closely related to graph expansion), total effective resistances [70], and mixing time [30]. In particular, Ghosh, Boyd and Saberi [70] studied the related problem of minimizing the sum of effective resistances over all pairs of vertices. They gave a convex programming relaxation of the problem but did not provide any result for the discrete optimization setting. Most of the earlier works only proposed convex programming relaxations and heuristic algorithms, and approximation guarantees are only obtained recently for the more general experimental design problems. When every edge has the same cost, there is a $(1+\varepsilon)$-approximation algorithm for minimizing the total effective resistance when the budget is at least $\Omega(\frac{|V|}{\varepsilon} + \frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon})$ [118], and there is a $(1+\varepsilon)$-approximation algorithm for both maximizing the algebraic connectivity and minimizing the total effective resistance when the budget is at least $\Omega(\frac{|V|}{\varepsilon^2})$ [6]. For general edge costs, there is a randomized $(1+\varepsilon)$-approximation algorithm for both maximizing the algebraic connectivity and minimizing the total effective resistance when the budget is at least $\Omega(\frac{|V|\cdot\|c\|_\infty}{\varepsilon^2})$, where $\|c\|_\infty$ is the maximum edge cost in the input graph (see Section 6.2.1 and Section 6.2.2 in this thesis).

We remark that the one-sided spectral rounding based methods ([6] or Chapter 5) can only apply to the s-t effective resistance network design problem when $k \geqslant \Omega(\frac{|V|}{\varepsilon^2})$ for the desired precision $\varepsilon < 1$. In this chapter, the interesting regime is when k is much smaller than $|V|$, where the techniques in [6, 118, 98] cannot guarantee good approximation (including the spectral rounding technique in Chapter 5). We have developed a set of new techniques for analyzing and rounding the solutions to the convex program that will hopefully find applications for solving related problems in the regime when k is small.

8.1.4 Technical Overview

Our main technical contribution is in designing rounding techniques for a convex programming relaxation of our problem. There is a natural convex programming relaxation, by using the conductance (reciprocal of the resistance) of the edges as variables, and writing the s-t effective resistance as the objective function and noting that it is convex with respect to the variables (Section 8.2.1).

We show that optimal solutions of this convex program enjoy some nice properties[1]. Given an optimal fractional solution x^* and the unit s-t electrical flow f^* supported in

[1] We can also show that there *exists* an optimal solution such that the fractional edges form a forest, but this is not included in the paper as we have not used this property in the rounding algorithm.

x^*, we derive from the convex optimality conditions that there is a flow-conductance ratio $\alpha > 0$ such that $f^*(e) = \alpha x^*(e)$ for every fractional edge e with $0 < x^*(e) < 1$ and $f^*(e) \geqslant \alpha$ for every integral edge e with $x^*(e) = 1$. The flow-conductance ratio α is crucial in the rounding algorithm and the analysis.

The rounding techniques in recent papers on experimental design [6, 118, 98] considered each edge/vector as a unit. In [6, 98], a potential function as in spectral sparsification is used to guide a local search algorithm to swap two edges/vectors at a time to improve the current solution. In [118], a probability distribution on the edges/vectors is carefully designed for an independent randomized rounding. These techniques only work in the case when the solutions form a spanning set so that the "contribution" of each individual edge/vector is well-defined. This is basically the reason why the results in [6, 118, 98] only apply when the budget k is at least $|V|$.

Our approach is based on a randomized rounding procedure on s-t paths. Given x^*, we compute the unit s-t electrical flow f^* supported in x^*, and decompose f^* as a convex combination of s-t paths. The rounding algorithm has $\tau = 1/\alpha$ iterations (recall that α is the flow-conductance ratio of the optimal solution x^*), where we pick a random path P_i from the convex combination in each iteration, and return $H := \cup_{i=1}^{\tau} P_i$ as our solution. One difference from the previous techniques in the literature is that each unit in the rounding algorithm is a s-t path, so in particular s and t are always connected in our solution. Another difference is that our problem has some extra structure, so that we can compute the electrical flow f^* to guide our rounding procedure, where the variables $f^*(e)$ are not in the convex program. These allow us to obtain a constant factor approximation algorithm for all budget $k \geqslant d_{st}$, the shortest path distance between s and t (note that when $k < d_{st}$ there is no feasible integral solution).

In the analysis, we prove in Lemma 8.2.6 that the expected number of edges in H is at most k, and in Lemma 8.2.7 that the expected effective resistance is $\operatorname{Reff}_H(s,t) \leqslant 2 \operatorname{Reff}_{x^*}(s,t)$. To bound the expected effective resistance, we use Thomson's principle and construct a unit s-t flow f to show that $\operatorname{Reff}_H(s,t) \leqslant \mathcal{E}_H(f) \leqslant 2 \operatorname{Reff}_{x^*}(s,t)$. To construct the unit s-t flow f, we keep the flow-conductance ratio and send α units of flow on each sampled path P_i (i.e. $f(e) = \alpha$ and $x(e) = 1$). The flow-conductance ratio plays a crucial role in the proofs of both lemmas. This is because the rounding algorithm is based on the flow variables $f^*(e)$, and thus the performance guarantees are in terms of $f^*(e)$, but the ratio α allows us to relate them back to the variables $x^*(e)$ in the convex program. Combining the two lemmas give us a constant factor bicriteria approximation algorithm for the problem. This can be turned into a true approximation algorithm by scaling down the budget to $\frac{k}{2}$ and run the bicriteria approximation algorithm with some additional claims (Section 8.2.4).

The improvement on the approximation ratio when budget k is large comes from two observations. The first is that if k is much larger than the length of the shortest s-t path, then the number of independent iterations in the rounding scheme is large (Lemma 8.2.3). The second is that we can ignore some s-t paths in the flow decomposition with many fractional edges without affecting the performance much. Combining these, we can apply a Chernoff-Hoeffding bound to show that the number of edges is at most $(1+\varepsilon)k$ with high probability. Then it is not necessary to scale down the budget by a factor of 2 and we can prove a stronger bound that the effective resistance is at most $2+O(\varepsilon)$ times the optimal value.

Organization

We present the convex programming relaxation and our two rounding procedures in Section 8.2. The NP-hardness and small set expansion hardness results are provided in Section 8.3.

8.2 Convex Programming Algorithms

In this section, we analyze a convex programming relaxation for our problem. We first describe the convex program and prove a characterization of the optimal solutions in Section 8.2.1. We then present a randomized rounding algorithm using flow decomposition in Section 8.2.2, and show that it is a constant factor bicriteria approximation algorithm in Section 8.2.3. Then, we show how to convert the bicriteria approximation algorithm into a true approximation algorithm in Section 8.2.4, and how to modify the algorithm slightly to achieve a better approximation guarantee when the budget k is large in Section 8.2.5. Finally, we discuss the dual problem of minimizing the cost while satisfying the effective resistance constraint in Section 8.2.6.

8.2.1 Convex Programming Relaxation

The formulation is for the weighted problem, where each edge has a weight $w(e) := \frac{1}{r(e)}$. We introduce a variable $x(e)$ for each edge e to indicate whether e is chosen in our subgraph. Let $\mathrm{Reff}_{st}(x)$ be the s-t effective resistance of the graph with conductance $x(e)w(e)$ on edge $e \in E$. The following is a natural convex programming relaxation for the problem.

$$\begin{aligned} \underset{x \in \mathbb{R}^E}{\text{minimize}} \quad & \mathrm{Reff}_{st}(x) \\ \text{subject to} \quad & \sum_{e \in E} c(e) \cdot x(e) \leqslant k, \\ & 0 \leqslant x(e) \leqslant 1, \qquad \forall e \in E. \end{aligned} \tag{st-CP}$$

This is an exact formulation if $x(e) \in \{0,1\}$ for all $e \in E$. The objective function is convex in x by Lemma 2.4.5. The convex program can be solved in polynomial time by the ellipsoid method to inverse exponential accuracy, or by the techniques described in [6] to inverse polynomial accuracy, which are both sufficient for the rounding algorithm.

8.2.1.1 Integrality Gap Examples

We show some limitations of the convex program for general $w(e)$ and $c(e)$. The following figure shows a simple example where the integrality gap is unbounded if the cost could be arbitrary.

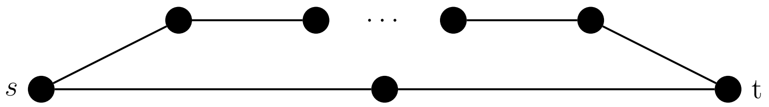

Figure 8.1: Integrality gap example with arbitrary cost and unit resistance.

In this graph, the top path has length $n-2$ where each edge has cost $\frac{1}{n-2}$. The bottom path has two edges with cost 1. The resistance of each edge is 1, and the budget is $k = 1$. The integrality gap of this example is $\Omega(n)$. To see this, the integral solution can only afford the top path, and the effective resistance is $n-2$. However, the fractional solution can set $\mathsf{x}(e) = \frac{1}{2}$ for each of the two bottom edges, and the effective resistance of this fractional solution is 4.

The following figure shows another simple example where the integrality gap is unbounded if the edge costs are the same but the resistances could be arbitrary.

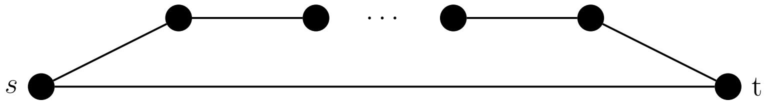

Figure 8.2: Integrality gap example with arbitrary resistance and unit cost.

In this example, the top path has length $n-1$ with each edge of resistance 1. The bottom path has only one edge with resistance R. All edges have cost 1 and the budget $k = 2$. The integral solution can only afford the bottom path, with effective resistance R. The fractional solution can set $\mathsf{x}(e) = \frac{2}{n-1}$ for each edge in the top path, with effective resistance $O(n^2)$. When $R \gg n^2$, the integrality gap could be arbitrarily large. Notice that this example also excludes the possibility of bicriteria approximation via the weighted convex program.

Even in the unit-cost unit-resistance case, the integrality gap is unbounded if k is smaller than the s-t shortest path distance. Henceforth, in view of these observations we assume the following in the rest of this section.

Assumption 8.2.1. *We assume that $\mathsf{c}(e) = \mathsf{w}(e) = \mathsf{r}(e) = 1$ for every edge $e \in E$, which is the setting of the s-t effective resistance network design problem, and the budget k is at least the shortest path distance d_{st} between s and t in the input graph.*

The integrality gap of the convex program is still at least two with Assumption 8.2.1. For a simple example, consider a graph with two vertex-disjoint s-t paths, each of length

$\frac{k}{2}+1$, and the budget is k. Then the optimal integral value is $\frac{k}{2}+1$ while the optimal fractional value is close to $\frac{k}{4}$, and so the integrality gap gets arbitrarily close to two.

We will show that the integrality gap of the convex program is at most 8 with these assumptions. Note that just to connect s and t, then k must be at least the s-t shortest path distance. It is interesting that this small additional assumption could reduce the integrality gap from unbounded to a constant.

8.2.1.2 Characterization of Optimal Solutions

In the case $\mathsf{c}(e) = \mathsf{w}(e) = \mathsf{r}(e) = 1$ for all edges $e \in E$, we will prove that the electrical flow f^* supported in the optimal solution x^* to (st-CP) satisfies a crucial property about the flow-conductance ratio $\frac{\mathsf{f}^*(e)}{\mathsf{x}^*(e)}$.

Lemma 8.2.2 (Characterization of Optimal Solution)**.** *Let $G = (V, E)$ be the input graph with $\mathsf{c}(e) = \mathsf{w}(e) = 1$ for all edges $e \in E$. Let $\mathsf{x}^* : E \to \mathbb{R}_{\geqslant 0}$ be an optimal solution to the convex program* (st-CP)*. Let $E_F \subseteq E$ be the set of fractional edges with $0 < \mathsf{x}^*(e) < 1$, and $E_I \subseteq E$ be the set of integral edges with $\mathsf{x}^*(e) = 1$. Let $\mathsf{f}^* : E \to \mathbb{R}_{\geqslant 0}$ be the unit s-t electrical flow supported in x^*. There exists $\alpha > 0$ such that*

$$\mathsf{f}^*(e) = \alpha \mathsf{x}^*(e) \quad \forall e \in E_F \quad \text{and} \quad \mathsf{f}^*(e) \geqslant \alpha \quad \forall e \in E_I.$$

Before starting the proof, we make some remarks. Notice that $\frac{\mathsf{f}(e)}{\mathsf{x}(e)}$ is the potential difference between the two endpoints of edge e. When s and t are connected in the graph, we have $\left(\frac{\mathsf{f}(e)}{\mathsf{x}(e)}\right)^2 = (\mathsf{b}_e \mathsf{L}_\mathsf{x}^\dagger \mathsf{b}_{st})^2$. Further, when $\mathrm{Reff}_{st}(\mathsf{x})$ is differentiable at x, we have the partial derivative $\partial_e \mathrm{Reff}_{st}(\mathsf{x}) = -(\mathsf{b}_e^\top \mathsf{L}_\mathsf{x}^\dagger \mathsf{b}_{st})^2$. Ideally, we want to use KKT conditions Theorem 2.2.28 to characterize these partial derivatives. However, as observed in Section 2.4, one subtle issue is that $\mathrm{Reff}_{st}(\mathsf{x})$ is not differentiable over the whole domain. For many instances, we expect that an optimal solution would have many edges with $\mathsf{x}^*(e) = 0$. Those boundary points are not differentiable as $\mathrm{Reff}_{st}(\mathsf{x})$ is undefined for nonnegative edge weights. Thus, we cannot apply Theorem 2.2.28 directly. This issue can be resolved using the more general subdifferential theory for convex functions [76, 126]. We use modified

KKT conditions (Proposition 2.2.30) as a necessary condition for optimality to bypass this issue without invoking the more advanced subdifferential theory.

Proof of Lemma 8.2.2. By Fact 2.4.2, $\operatorname{Reff}_{st}(\mathsf{x}) = +\infty$ when $\mathsf{x} \notin \mathcal{D}_{st}$, where $\mathcal{D}_{st}$ is the set of all s-t connected edge weights x defined in (2.11). Without loss of generality, we can rewrite the convex program as

$$\begin{aligned} \underset{\mathsf{x}\in\mathcal{D}_{st}}{\text{minimize}} \quad & \operatorname{Reff}_{st}(\mathsf{x}) \\ \text{subject to} \quad & \sum_{e\in E} \mathsf{x}(e) \leqslant k, \\ & 0 \leqslant \mathsf{x}(e) \leqslant 1, \qquad \forall e \in E. \end{aligned}$$

Notice that the set $\{\mathsf{x} \in \mathbb{R}^E \mid \sum_e \mathsf{x}(e) \leq k, \mathsf{x}(e) \in [0,1],\ \forall e \in E\}$ is closed and bounded, $\operatorname{Reff}_{st}(\mathsf{x})$ is continuous over $\mathcal{D}_{st}$, and $\operatorname{Reff}_{st}(\mathsf{x})$ blows up when approaching to the boundary of $\mathcal{D}_{st}$, thus there exists an optimal solution x^* that attains the optimal value of the above convex program.

Let μ be the dual variable for the budget constraint $\sum_{e\in E} \mathsf{x}(e) \leqslant k$, and $\lambda^+(e)$ and $\lambda^-(e)$ be the dual variables for the upper bound $\mathsf{x}(e) \leqslant 1$ and the nonnegative constraint $\mathsf{x}(e) \geqslant 0$ respectively.

By taking $\mathsf{x} = \delta \mathit{1}_m$ for small enough $\delta > 0$, it is obvious that the above convex program satisfies the Slater's condition. Thus, strong duality holds by Theorem 2.2.26, i.e. there exists an optimal dual solution $\lambda^{*+}, \lambda^{*-} \in \mathbb{R}_+^E$ and $\mu^* \in \mathbb{R}_+$ that attain zero duality gap together with x^*.

Given any optimal solution $\mathsf{x}^* \in \mathcal{D}_{st}$, for any $e \in E$ with $\mathsf{x}^*(e) > 0$, $\partial_e \operatorname{Reff}_{st}(\mathsf{x}^*)$ exists and $\partial_e \operatorname{Reff}_{st}(\mathsf{x}^*) = -(\mathsf{b}_{st}^\top \mathsf{L}_{\mathsf{x}^*}^\dagger \mathsf{b}_e)^2$ by Lemma 2.4.6. Since all the budget, capacity, and nonnegativity constraints are affine constraints, they are differerentiable. Applying the modified KKT necessary conditions Proposition 2.2.30 with direction $\mathsf{d} = \chi_e$ for all $e \in E$ with $\mathsf{x}^*(e) > 0$, we can show that the complement slackness condition holds

$$\lambda^{*+}(e) = \lambda^{*-}(e) = 0 \quad \forall e \in E_F \qquad \text{and} \qquad \lambda^{*-}(e) = 0 \quad \forall e \in E_I. \tag{8.1}$$

Furthermore, we also have

$$-(\mathsf{b}_{st}^\top \mathsf{L}_{\mathsf{x}^*}^\dagger \mathsf{b}_e)^2 + \lambda^{*+}(e) - \lambda^{*-}(e) + \mu^* = 0 \quad \forall e \in E_F \cup E_I. \tag{8.2}$$

Combine (8.1) and (8.2) and the fact $\lambda^{*+} \geqslant 0$, it follows that

$$(\mathsf{b}_{st}^\top \mathsf{L}_{\mathsf{x}^*}^\dagger \mathsf{b}_e)^2 = \mu^* \quad e \in E_F \qquad \text{and} \qquad (\mathsf{b}_{st}^\top \mathsf{L}_{\mathsf{x}^*}^\dagger \mathsf{b}_e)^2 \geqslant \mu^* \quad \forall e \in E_I.$$

Let φ be a potential vector of the electrical flow f^* supported in x^*. For an edge $e = uv \in E$,

$$\left(\frac{\mathsf{f}^*(e)}{\mathsf{x}^*(e)}\right)^2 = (\varphi(u) - \varphi(v))^2 = \left(\mathsf{b}_e^\top \mathsf{L}_{\mathsf{x}^*}^\dagger \mathsf{b}_{st}\right)^2,$$

where the first equality is by Ohm's law and the assumption that $\mathsf{w}(e) = 1$ for all $uv \in E$, and the second equality uses that $\mathsf{L}_{\mathsf{x}^*}\varphi = \mathsf{b}_{st}$ as explained in Section 2.4. The lemma then follows from the above paragraph and writing μ as α^2. □

The flow-conductance ratio α will be crucial in the rounding algorithm and its analysis. The following lemma shows an upper bound on α using the budget k and the shortest path distance d_{st} between s and t.

Lemma 8.2.3. *Under the conditions in Assumption 8.2.1, it holds that* $\alpha^2 \leqslant \frac{d_{st}}{k} \leqslant 1$.

Proof. Let x^* be an optimal solution to (st-CP), and f^* be the unit s-t electrical flow supported in x^*. As $k \geqslant d_{st}$, a shortest path is a feasible solution to (st-CP), and thus $\mathrm{Reff}_{\mathsf{x}^*}(s,t) \leqslant d_{st}$. On the other hand, by Thomson's principle and Lemma 8.2.2,

$$\begin{aligned}\mathrm{Reff}_{\mathsf{x}^*}(s,t) = \sum_{e \in E} \frac{\mathsf{f}^*(e)^2}{\mathsf{x}^*(e)} &= \sum_{e \in E_I} \mathsf{f}^*(e)^2 + \sum_{e \in E_F} \frac{\mathsf{f}^*(e)^2}{\mathsf{x}^*(e)} \\ &\geqslant \sum_{e \in E_I} \alpha^2 + \sum_{e \in E_F} \alpha^2 \mathsf{x}^*(e) = \alpha^2 \sum_{e \in E} \mathsf{x}^*(e) = \alpha^2 k,\end{aligned}$$

where the last equality holds since we can assume $\sum_{e \in E} \mathsf{x}^*(e) = k$ for the optimal solution x^* without loss of generality by Rayleigh's principle (or otherwise we have an integral optimal solution). The lemma follows by combining the upper bound and the lower bound. □

8.2.2 Randomized Path-Rounding Algorithm

Our rounding algorithm uses the unit electrical flow f^* supported in the optimal solution x^* to construct an integral solution. The algorithm will first decompose the flow f^* as a convex combination of flow paths, and then randomly choose the flow paths and return the union of the chosen flow paths as our solution.

The following lemma about flow decomposition is by the standard argument to remove one (fractional) flow path at a time, which holds for any unit directed acyclic s-t flow.

Lemma 8.2.4 (Flow Decomposition)**.** *Given the unit s-t electrical flow f, there is a polynomial time algorithm to find a set $\mathcal{P}$ of s-t paths with $|\mathcal{P}| \leqslant |E|$ such that the flow vector $f : \mathbb{R}^E_{\geqslant 0}$ can be written as a convex combination of the characteristic vectors of the paths in $\mathcal{P}$, i.e.*

$$f = \sum_{p \in \mathcal{P}} v_p \cdot \chi_p \quad \text{and} \quad \sum_{p \in \mathcal{P}} v_p = 1 \quad \text{and} \quad v_p > 0 \text{ for each } p \in \mathcal{P},$$

where $\chi_p \in R^E$ is the characteristic vector of the path p with one on each edge $e \in p$ and zero otherwise.

With the flow decomposition, we are ready to present the rounding algorithm.

Randomized Path Rounding Algorithm

1. Let x^* be an optimal solution to the convex program (st-CP). Let f^* be the unit s-t electrical flow supported in x^*. Let α be the flow-conductance ratio defined in Lemma 8.2.2.
2. Compute a flow decomposition $\mathcal{P}$ of f^* as defined in Lemma 8.2.4.
3. For i from 1 to $\tau := \lfloor \frac{1}{\alpha} \rfloor$ do
    - Let P_i be a random path from $\mathcal{P}$ where each path $p \in \mathcal{P}$ is sampled with probability v_p.
4. Return the subgraph H formed by the edge set $\cup_{i=1}^{\tau} P_i$.

The following lemma shows that the rounding algorithm will always return a non-empty subgraph.

Lemma 8.2.5. *Suppose the input instance satisfies the conditions in Assumption 8.2.1. Let x^* be an optimal solution to (st-CP) and $\alpha > 0$ be the flow-conductance ratio as defined in Lemma 8.2.2. Then*

$$\frac{1}{\alpha} \geqslant \tau \geqslant \frac{1}{2\alpha} > 0.$$

Proof. Since we assumed that the budget k is at least the length d_{st} of a shortest s-t path, it follows from Lemma 8.2.3 that $\alpha \leqslant 1$. This implies that

$$\frac{1}{\alpha} \geqslant \tau = \left\lfloor \frac{1}{\alpha} \right\rfloor \geqslant \max\left\{1, \frac{1}{\alpha} - 1\right\} \quad \Longrightarrow \quad 1 \geqslant \tau\alpha \geqslant \max\{\alpha, 1-\alpha\} \geqslant \frac{1}{2}. \qquad \square$$

8.2.3 Bicriteria Approximation

The analysis of the approximation guarantee goes as follows. First, we show that the expected number of edge in the returned subgraph H is at most the budget k. Then, we prove that the expected effective resistance of the returned subgraph is at most two times that of the optimal fractional solution. Both of these steps use the flow-conductance ratio α crucially. These combine to show that the randomized path rounding algorithm is a constant factor bicriteria approximation algorithm.

Let x^* be an optimal solution to (st-CP). Let E_F and E_I be the set of fractional edges and integral edges in x^*. We assume that each edge $e \in E_I$ will be included in the subgraph H returned by the rounding algorithm. We focus on bounding the number of edges in E_F that will be included in H.

Lemma 8.2.6 (Expected Budget)**.** *Let x^* be an optimal solution to (st-CP) when $\mathsf{w}(e) = 1$ for all edges $e \in E$. Let X_e be an indicator variable of whether e is included in the returned subgraph H by the rounding algorithm, Then,*

$$\mathbb{E}\left[\sum_{e \in E_F} X_e\right] \leqslant \tau\alpha \sum_{e \in E_F} \mathsf{x}^*(e) \leqslant \sum_{e \in E_F} \mathsf{x}^*(e).$$

Proof. Note that an edge e is contained in P_i with probability $\sum_{p \in \mathcal{P}: p \ni e} v_p$. By the union bound, an edge e is included in the returned subgraph H by the rounding algorithm with probability

$$\mathbb{P}[X_e = 1] \leqslant \sum_{i=1}^{\tau} \sum_{p \in \mathcal{P}: p \ni e} v_p = \tau \sum_{p \in \mathcal{P}: p \ni e} v_p = \tau f_e^*,$$

where the last equality holds by the property of the flow decomposition $\mathcal{P}$ of the electrical flow f^* in Lemma 8.2.4.

By Lemma 8.2.2, $f^*(e) = \alpha x^*(e)$ for each fractional edge $e \in E_F$, and this implies that

$$\mathbb{P}[X_e = 1] \leqslant \tau f^*(e) = \tau \alpha x^*(e) \quad \forall e \in E_F.$$

Therefore,

$$\mathbb{E}\left[\sum_{e \in E_F} X_e\right] = \sum_{e \in E_F} \mathbb{P}[X_e = 1] \leqslant \tau \alpha \sum_{e \in E_F} x^*(e) = \left\lfloor \frac{1}{\alpha} \right\rfloor \alpha \sum_{e \in E_F} x^*(e) \leqslant \sum_{e \in E_F} x^*(e). \qquad \square$$

The key step is to show that $\mathbb{E}[\mathrm{Reff}_H(s,t)] \leqslant 2\,\mathrm{Reff}_{x^*}(s,t)$. To prove this, we construct a unit s-t flow f and show that $\mathbb{E}[\mathcal{E}_H(f)] \leqslant 2\,\mathrm{Reff}_{x^*}(s,t)$, and hence by Thomson's principle $\mathbb{E}[\mathrm{Reff}_H(s,t)] \leqslant \mathbb{E}[\mathcal{E}_H(f)] \leqslant 2\,\mathrm{Reff}_{x^*}(s,t)$. To construct the flow f, the idea is to follow the ratio α in the fractional solution x^* and send α units of flow on each path P_i selected.

Lemma 8.2.7 (Expected Effective Resistance)**.** *Suppose the input instance satisfies the conditions in Assumption 8.2.1. Let x^* be an optimal solution to (st-CP) and f^* be the unit s-t electrical flow supported in x^*. The expected s-t effective resistance of the subgraph H returned by the rounding algorithm is*

$$\mathbb{E}[\mathrm{Reff}_H(s,t)] \leqslant \left(1 - \frac{1}{\tau} + \frac{1}{\tau\alpha}\right) \cdot \mathcal{E}_{x^*}(f^*) = \left(1 - \frac{1}{\tau} + \frac{1}{\tau\alpha}\right) \cdot \mathrm{Reff}_{x^*}(s,t) \leqslant 2\mathrm{Reff}_{x^*}(s,t).$$

Proof. Consider the flow vector $f : \mathbb{R}^E_{\geqslant 0}$ defined by sending α units of flow on each path P_i chosen by the rounding algorithm, i.e. the random variable $f = \sum_{i=1}^{\tau} \alpha \cdot \chi_{P_i}$ with $f(e) = \alpha \cdot |\{P_i \mid 1 \leqslant i \leqslant \tau, P_i \ni e\}|$ for each edge $e \in E$. We would like to upper bound the expected energy $\mathcal{E}_H(f)$ in order to upper bound $\mathrm{Reff}_H(s,t)$.

Each P_i is a random s-t path sampled from the flow decomposition $\mathcal{P}$ of the flow vector $f^* : \mathbb{R}^E_{\geqslant 0}$ of the unit s-t electrical flow supported in x^*, and $\chi_{P_i} \in \{0,1\}^E$ is its characteristic vector with expected value

$$\mathbb{E}\left[\chi_{P_i}\right] = \sum_{p \in \mathcal{P}} v_p \cdot \chi_p = f^*.$$

Since each edge in H is of conductance one, the expected energy of f in H is

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}_H(f)\right] = \mathbb{E}\left[\sum_{e \in E} f(e)^2\right] &= \mathbb{E}\left[\langle f, f \rangle\right] \\ &= \mathbb{E}\left[\left\langle \sum_{i=1}^{\tau} \alpha \cdot \chi_{P_i}, \sum_{j=1}^{\tau} \alpha \cdot \chi_{P_j} \right\rangle\right] = \sum_{i=1}^{\tau}\sum_{j=1}^{\tau} \alpha^2 \cdot \mathbb{E}\left[\langle \chi_{P_i}, \chi_{P_j} \rangle\right].\end{aligned}$$

As each path P_i is sampled independently, for $i \neq j$,

$$\mathbb{E}\left[\langle \chi_{P_i}, \chi_{P_j} \rangle\right] = \langle \mathbb{E}\left[\chi_{P_i}\right], \mathbb{E}\left[\chi_{P_j}\right] \rangle = \langle f^*, f^* \rangle = \sum_{e \in E} f^*(e)^2.$$

For $i = j$,

$$\mathbb{E}\left[\langle \chi_{P_i}, \chi_{P_i} \rangle\right] = \sum_{p \in \mathcal{P}} v_p \langle \chi_p, \chi_p \rangle = \sum_{p \in \mathcal{P}} v_p \sum_{e \in p} 1 = \sum_{e \in E} \sum_{p \in \mathcal{P} : p \ni e} v_p = \sum_{e \in E} f^*(e),$$

where the last equality follows from the property of the flow decomposition in Lemma 8.2.4. Combining these two terms, it follows that

$$\mathbb{E}\left[\mathcal{E}_H(f)\right] = \alpha^2 \tau \sum_{e \in E} f^*(e) + \alpha^2 \tau (\tau - 1) \sum_{e \in E} f^*(e)^2.$$

Thomson's principle states that the $\mathrm{Reff}_H(s,t)$ is upper bounded by the energy of any one unit s-t flow. Note that f is an s-t flow of $\tau\alpha$ units, and $\tau\alpha > 0$ by Lemma 8.2.5. Scaling

f to a one unit s-t flow by dividing the flow on each edge by $\tau\alpha$ gives an upper bound on

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Reff}_H(s,t)\right] \leqslant \frac{\mathbb{E}\left[\mathcal{E}_H(\mathsf{f})\right]}{\tau^2\alpha^2} &= \frac{1}{\tau}\sum_{e\in E}\mathsf{f}^*(e) + \left(1-\frac{1}{\tau}\right)\sum_{e\in E}\mathsf{f}^*(e)^2 \\
&\leqslant \frac{1}{\tau\alpha}\sum_{e\in E}\frac{\mathsf{f}^*(e)^2}{\mathsf{x}^*(e)} + \left(1-\frac{1}{\tau}\right)\sum_{e\in E}\frac{\mathsf{f}^*(e)^2}{\mathsf{x}^*(e)} \\
&= \left(1-\frac{1}{\tau}+\frac{1}{\tau\alpha}\right)\cdot\mathcal{E}_{\mathsf{x}^*}(\mathsf{f}^*) \\
&= \left(1-\frac{1}{\tau}+\frac{1}{\tau\alpha}\right)\cdot\mathrm{Reff}_{\mathsf{x}^*}(s,t),
\end{aligned}$$

where the second inequality follows from Lemma 8.2.2 that $\frac{\mathsf{f}^*(e)}{\mathsf{x}^*(e)} \geqslant \alpha$ for every edge $e \in E$ and also $\mathsf{x}^*(e) \leqslant 1$ for every edge $e \in E$, and the last equality is from Thomson's principle that $\mathrm{Reff}_{\mathsf{x}^*}(s,t) = \mathcal{E}_{\mathsf{x}^*}(\mathsf{f}^*)$. Finally, notice that $1-\frac{1}{\tau}+\frac{1}{\tau\alpha} \leqslant 2$ as $\frac{1}{\alpha}-1 \leqslant \lfloor\frac{1}{\alpha}\rfloor = \tau$. □

Combining Lemma 8.2.6 and Lemma 8.2.7, it follows from a simple application of Markov's inequality that there is an outcome of the randomized path-rounding algorithm which uses at most $2k$ edges with s-t effective resistance at most $4\,\mathrm{Reff}_{\mathsf{x}^*}(s,t)$. In the following, we apply Markov's inequality more carefully to show that the success probability is at least $\Omega(\alpha)$. In the next subsection, we will argue that α can be assumed to be $\Omega(\frac{1}{m})$ and so the path-rounding algorithm is a randomized polynomial time algorithm.

Theorem 8.2.8 (Bicriteria Approximation)**.** *Suppose the input instance satisfies the conditions in Assumption 8.2.1. Let x^* be an optimal solution to (st-CP). Given x^*, the randomized path rounding algorithm will return a subgraph H with at most $2k$ edges and $\mathrm{Reff}_H(s,t) \leqslant 4\,\mathrm{Reff}_{\mathsf{x}^*}(s,t)$ with probability at least $\Omega(\alpha)$.*

Proof. First, we bound the probability that the subgraph H has more than $2k$ edges. Let X_e be an indicator variable of whether the edge e is included in the returned subgraph H. Recall that E_F and E_I denote the set of fractional edges and integral edges in x^* respectively. We assume pessimistically that all edges in E_I will be included in the subgraph

H returned by the rounding algorithm. Then, by Markov's inequality and Lemma 8.2.6,

$$\mathbb{P}\left[\sum_{e\in E} X_e > 2k\right] \leqslant \mathbb{P}\left[\sum_{e\in E_F} X_e > 2k - |E_I|\right] \leqslant \frac{\mathbb{E}\left[\sum_{e\in E_F} X_e\right]}{2k-|E_I|} \leqslant \frac{\tau\alpha \sum_{e\in E_F} \mathsf{x}^*(e)}{2k-|E_I|} \leqslant \frac{\tau\alpha}{2},$$

where the last inequality is by $\sum_{e\in E_F} \mathsf{x}^*(e) \leqslant k - |E_I|$.

Next, we bound the probability that $\operatorname{Reff}_H(s,t) > 4\operatorname{Reff}_{\mathsf{x}^*}(s,t)$. By Markov's inequality and Lemma 8.2.7,

$$\mathbb{P}\left[\operatorname{Reff}_H(s,t) > 4\operatorname{Reff}_{\mathsf{x}^*}(s,t)\right] \leqslant \frac{1}{4}\left(1 - \frac{1}{\tau} + \frac{1}{\tau\alpha}\right) = \frac{\tau\alpha+1}{4\tau\alpha} - \frac{1}{4\tau} \leqslant \frac{\tau\alpha+1}{4\tau\alpha} - \Omega(\alpha),$$

where the last inequality is because $\tau = \lfloor \frac{1}{\alpha} \rfloor \leqslant \frac{1}{\alpha}$.

To prove the lemma, it remains to show that

$$\frac{\tau\alpha}{2} + \frac{\tau\alpha+1}{4\tau\alpha} \leqslant 1 \quad \iff \quad 2(\tau\alpha)^2 - 3(\tau\alpha) + 1 = (2\tau\alpha - 1)(\tau\alpha - 1) \leqslant 0,$$

which follows from Lemma 8.2.5. □

8.2.4 Constant Factor Approximation

We showed that the randomized path rounding algorithm is a bicriteria approximation algorithm. To achieve a true approximation algorithm, a natural idea is to scale down the budget from k to $\frac{k}{2}$ and apply the randomized path rounding algorithm. The following lemma takes care of the case of $\frac{k}{2} < d_{st}$, when the shortest path assumption does not hold after scaling, by showing that simply returning a shortest s-t path is already a good enough approximation.

Lemma 8.2.9. *When the budget k is at least the length d_{st} of a shortest s-t path, any s-t shortest path is a $(\frac{k}{d_{st}})$-approximate solution for the s-t effective resistance network design problem.*

Proof. When $k \geqslant d_{st}$, a s-t shortest path is a feasible solution to the problem with s-t effective resistance at most d_{st}. To prove the lemma, we will show that $\operatorname{Reff}_{\mathsf{x}}(s,t) \geqslant \frac{d_{st}^2}{k}$

for any feasible solution x to (st-CP), and so an s-t shortest path is already a $(\frac{k}{d_{st}})$-approximation.

Let G_{x} be the graph G with fractional weight $\mathsf{x}(e)$ on each edge $e \in E$. To show a lower bound on $\text{Reff}_{\mathsf{x}}(s,t)$, we identify the vertices in G_{x} to a form a path graph P_{x} as follows: For each $i \geqslant 0$, let U_i be the set of vertices in G with shortest path distance i to s, where the shortest path distance is defined where each edge in G is of length one. First, for each $0 \leqslant i \leqslant d_{st} - 1$, we identify the vertices in U_i to a single vertex u_i. Then, we identify all the vertices in $\cup_{i \geqslant d_{st}} U_i$ to a single vertex $u_{d_{st}}$. The path graph P_{x} has vertex set $\{u_0, \ldots, u_{d_{st}}\}$ and edge set $\{ab \in E \mid a \in U_i \text{ and } b \in U_{i+1} \text{ for } 0 \leqslant i \leqslant d_{st} - 1\}$. For each edge e in P_{x}, its weight $\mathsf{x}(e)$ in P_{x} is the same as that in G_{x}. As an electrical network, identifying two vertices uv is equivalent to adding an edge of resistance zero between u and v. So, it follows from Rayleigh's monotonicity principle (Theorem 2.4.4) that $\text{Reff}_{G_{\mathsf{x}}}(s,t) \geqslant \text{Reff}_{P_{\mathsf{x}}}(u_0, u_{d_{st}})$ as $s \in U_0$ and $t \in U_{d_{st}}$.

As P_{x} is a series-parallel graph, we can compute $\text{Reff}_{P_{\mathsf{x}}}(s,t)$ directly. For each $1 \leqslant i \leqslant d_{st}$, let E_i be the set of parallel edges connecting u_{i-1} and u_i in P_{x}, and $c_i = \sum_{e \in E_i} \mathsf{x}(e)$ be the effective conductance between u_{i-1} and u_i in P_{x}. Then, by Fact 2.4.1,

$$\text{Reff}_{P_{\mathsf{x}}}(u_{i-1}, u_i) = \frac{1}{c_i} \quad \text{and} \quad \text{Reff}_{P_{\mathsf{x}}}(u_0, u_{d_{st}}) = \sum_{i=1}^{d_{st}} \text{Reff}_{P_{\mathsf{x}}}(u_{i-1}, u_i) = \sum_{i=1}^{d_{st}} \frac{1}{c_i}.$$

Note that $\sum_{i=1}^{d_{st}} c_i = \sum_{i=1}^{d_{st}} \sum_{e \in E_i} \mathsf{x}(e) \leqslant \sum_{e \in E} \mathsf{x}(e) \leqslant k$ for any feasible solution x. Using Cauchy-Schwarz inequality,

$$d_{st} = \sum_{i=1}^{d_{st}} \sqrt{c_i} \cdot \frac{1}{\sqrt{c_i}} \leqslant \sqrt{\sum_{i=1}^{d_{st}} c_i} \cdot \sqrt{\sum_{i=1}^{d_{st}} \frac{1}{c_i}} \leqslant \sqrt{k} \cdot \sqrt{\text{Reff}_{P_{\mathsf{x}}}(u_0, u_{d_{st}})}.$$

Therefore, we conclude that $\text{Reff}_{G_{\mathsf{x}}}(s,t) \geqslant \text{Reff}_{P_{\mathsf{x}}}(u_0, u_{d_{st}}) \geqslant \frac{d_{st}^2}{k}$. □

We are ready to prove our main approximation result.

Theorem 8.2.10. *Suppose the input instance satisfies the conditions in Assumption 8.2.1. There is a polynomial time 8-approximation algorithm for the s-t effective resistance network design problem.*

Proof. If the budget $k \leqslant 2d_{st}$, then Lemma 8.2.9 shows that simply returning an s-t shortest path would give a 2-approximation. Henceforth, we assume $k \geqslant 2d_{st}$.

Let $\mathsf{opt}(k)$ be the objective value of an optimal solution x^* to the convex program (st-CP) with budget k, so $\mathrm{Reff}_{\mathsf{x}^*}(s,t) = \mathsf{opt}(k)$. As $\frac{1}{2}\mathsf{x}^*$ is a feasible solution to (st-CP) with budget $\frac{1}{k}$, by Thomson's principle,

$$\mathsf{opt}\left(\frac{k}{2}\right) \leqslant \mathrm{Reff}_{\frac{1}{2}\mathsf{x}^*}(s,t) = \mathsf{b}_{st}^\top \left(\sum_{e\in E} \frac{\mathsf{x}_e^*}{2} b_e b_e^\top \right)^\dagger \mathsf{b}_{st}$$
$$= 2\mathsf{b}_{st}^\top \Big(\sum_{e\in E} \mathsf{x}^*(e) \cdot \mathsf{b}_e \mathsf{b}_e^\top \Big)^\dagger \mathsf{b}_{st} = 2\,\mathrm{Reff}_{\mathsf{x}^*}(s,t) = 2\mathsf{opt}(k).$$

Given the original budget $k \geqslant 2d_{st}$, our algorithm is to find an optimal solution z^* to (st-CP) with budget $\frac{k}{2} \geqslant d_{st}$, and use the path-rounding algorithm with input z^* to return a subgraph H. By Theorem 8.2.8, with probability $\Omega(\alpha)$, the subgraph H satisfies

$$|E(H)| \leqslant 2\sum_{e\in E} \mathsf{z}^*(e) \leqslant 2\left(\frac{k}{2}\right) = k \quad \text{and} \quad \mathrm{Reff}_H(s,t) \leqslant 4\mathsf{opt}\left(\frac{k}{2}\right) \leqslant 8\mathsf{opt}(k),$$

and so H is an 8-approximate solution to the s-t effective resistance network design problem.

Finally, we consider the time complexity of the algorithm. The number of iterations in the path rounding algorithm is $O\big(\frac{1}{\alpha}\big)$, and we need to run the path rounding algorithm $O\big(\frac{1}{\alpha}\big)$ times to boost the success probability to a constant. This is a randomized polynomial time algorithm when $\alpha = \Omega\big(\frac{1}{m}\big)$.

In the following, we show that when $\alpha \leqslant \frac{1}{4m}$, it is easy to obtain a 2-approximate solution without running the path-rounding algorithm. Let x^* be an optimal solution to (st-CP) with budget k, and f^* be the unit s-t electrical flow supported in x^*. Let $\mathcal{P}$ be the flow decomposition of f^* as in Lemma 8.2.4. We call a path $p \in \mathcal{P}$ an integral path if every edge $e \in p$ has $\mathsf{x}^*(e) = 1$; otherwise we call p a fractional path. When $\alpha \leqslant \frac{1}{4m}$, we simply return the union of all integral paths as our solution H. Clearly, H has at most k edges as it only contains integral edges. Next, we bound $\mathrm{Reff}_H(s,t)$ by the energy of the flow supported in the integral paths. By Lemma 8.2.2, an edge e with $\mathsf{x}^*(e) < 1$ has

$f^*(e) = \alpha x^*(e) < \alpha \leqslant \frac{1}{4m}$. This implies that each fractional path p has $v_p \leqslant \frac{1}{4m}$. Since $\mathcal{P}$ has at most m paths (Lemma 8.2.4), the total flow in the fractional paths is at most $\frac{1}{4}$, and thus the total flow in the integral paths is at least $\frac{3}{4}$. By scaling the flow supported in the integral paths to a one unit s-t flow, we see that

$$\mathrm{Reff}_H(s,t) \leqslant \frac{\mathcal{E}_{x^*}(f^*)}{(\frac{3}{4})^2} \leqslant 2\mathcal{E}_{x^*}(f^*) = 2\,\mathrm{Reff}_{x^*}(s,t).$$

To summarize, in all cases including $k < 2d_{st}$ and $\alpha \leqslant \frac{1}{4m}$, there is a polynomial time algorithm to return an 8-approximate solution to the s-t effective resistance network design problem. □

We make two remarks about improvements of Theorem 8.2.10.

Remark 8.2.11 (Approximation Ratio)**.** *The analysis of the 8-approximation algorithm is not tight. By a more careful analysis of the expected energy in Lemma 8.2.7 and the short path idea used in the next subsection, we can show that the approximation guarantee of the same algorithm in Theorem 8.2.10 is less than 5. However, the analysis is quite involved and not very insightful, so we have decided to omit those details and only keep the current analysis.*

Remark 8.2.12 (Deterministic Algorithm)**.** *Using the standard pessimistic estimator technique, we can derandomize the path-rounding algorithm to obtain an 8-approximation deterministic algorithm. The analysis is standard and we omit the details that would take a few pages.*

8.2.5 The Large Budget Case

In this subsection, we show how to modify the algorithm in Theorem 8.2.10 to achieve a better approximation ratio when the budget is much larger than the s-t shortest path distance.

The observation is that when $k \gg d_{st}$, then α is small by Lemma 8.2.3, and so there are many iterations in the path-rounding algorithm. Since each iteration is independent, we can use Chernoff-Hoeffding's bound to prove a stronger bound on the probability that the number of edges in the returned solution is significantly more than k (which outperforms the bound proved in Lemma 8.2.6 using Markov's inequality). We can then show that the expected s-t effective resistance is close to two times the optimal value by arguments similar to the proof of Lemma 8.2.7.

Modified Rounding Algorithm

For our analysis, we slightly modify the path-rounding algorithm to ignore "long" paths in the flow decomposition, so that we have a worst case bound to apply Chernoff-Hoeffding's bound. Unlike the flow decomposition in Lemma 8.2.4, the short path flow decomposition definition is specific to the electrical flow of an optimal solution to (st-CP). In the following definition, c is a parameter which will be set as $\frac{1}{\varepsilon} > 1$ to achieve a $(2+O(\varepsilon))$-approximation.

Definition 8.2.13 (Short Path Decomposition of Electrical Flow of Optimal Solution)**.** *Let x^* be an optimal solution to the convex program (st-CP). Let f^* be the unit s-t electrical flow supported in x^*. Let α be the flow-conductance ratio defined in Lemma 8.2.2.*

Let $\mathcal{P}^*$ be a flow decomposition of f^* as defined in Lemma 8.2.4. Let $x_F^* := \sum_{e \in E_F} \mathsf{x}^*(e)$ be the total fractional value on the fractional edges E_F in the optimal solution x^*.

We call a path $p \in \mathcal{P}^*$ a long path if p has at least $c\alpha x_F^*$ edges in E_F, i.e. $|p \cap E_F| \geqslant c\alpha x_F^*$. Otherwise we call a path $p \in \mathcal{P}^*$ a short path.

Let $\mathcal{P} := \{p \in \mathcal{P}^* \mid p \text{ is a short path}\}$ be the collection of short paths in $\mathcal{P}^*$. Let $\mathsf{f}_{\mathcal{P}} := \sum_{p \in \mathcal{P}} v_p \chi_p$ be the s-t flow defined by the short paths, and $v_{\mathcal{P}} := \sum_{p \in \mathcal{P}} v_p$ be the total flow value of $\mathsf{f}_{\mathcal{P}}$.

The modified algorithm is very similar to the randomized path-rounding algorithm in Section 8.2.3. The only difference is that we only sample the paths in the short path flow decomposition in Definition 8.2.13, and we adjust the sampling probability of a path p to $\frac{v_p}{v_{\mathcal{P}}}$ so that the sum is one.

Randomized Short Path Rounding Algorithm

1. Let x^* be an optimal solution to the convex program (st-CP). Let f^* be the unit s-t electrical flow supported in x^*. Let α be the flow-conductance ratio defined in Lemma 8.2.2.
2. Compute a short path flow decomposition $\mathcal{P}$ of f^* as described in Definition 8.2.13.
3. For i from 1 to $\tau = \lfloor \frac{1}{\alpha} \rfloor$ do
   - Let P_i be a random path from $\mathcal{P}$ where each path $p \in \mathcal{P}$ is sampled with probability $\frac{v_p}{v_{\mathcal{P}}}$.
4. Return the subgraph H formed by the edge set $\cup_{i=1}^{\tau} P_i$.

The following simple lemma shows that the total flow on the long paths is negligible when c is large, which will be useful in the analysis.

Lemma 8.2.14. *For the short path flow decomposition in Definition 8.2.13, $v_{\mathcal{P}} \geqslant 1 - \frac{1}{c}$.*

Proof. Using $\alpha \mathsf{x}(e)^* = \mathsf{f}^*(e)$ for $e \in E_F$ from Lemma 8.2.2 and the properties of the flow decomposition $\mathcal{P}^*$ of f^* in Lemma 8.2.4,

$$\alpha x_F^* = \sum_{e \in E_F} \mathsf{f}^*(e) = \sum_{p \in \mathcal{P}^*} v_p \cdot |p \cap E_F| \geqslant \sum_{p \in \mathcal{P}^* - \mathcal{P}} v_p \cdot |p \cap E_F| \geqslant c\alpha x_F^* \sum_{p \in \mathcal{P}^* - \mathcal{P}} v_p = c\alpha x_F^* (1 - v_{\mathcal{P}}),$$

where the last inequality is by the definition of long paths and the last equality is because f^* is a unit s-t flow. □

Analysis of Approximation Guarantee

First, we consider the expected s-t effective resistance of the returned subgraph H. For intuition, we can think of the modified rounding algorithm as applying the rounding algorithm in the scaled flow $f_{\mathcal{P}}/v_{\mathcal{P}}$, and so it should follow from Lemma 8.2.7 that

$$\mathbb{E}\left[\mathrm{Reff}_H(s,t)\right] \leqslant 2\mathcal{E}_{x^*}\left(\frac{f_{\mathcal{P}}}{v_{\mathcal{P}}}\right) = \frac{2}{v_{\mathcal{P}}^2}\mathcal{E}_{x^*}(f_{\mathcal{P}}) \leqslant \frac{2}{v_{\mathcal{P}}^2}\mathcal{E}_{x^*}(f^*) = \frac{2}{v_{\mathcal{P}}^2}\,\mathrm{Reff}_{x^*}(s,t),$$

which will be at most $(2+O(\varepsilon))\,\mathrm{Reff}_{x^*}(s,t)$ when $c = 1/\varepsilon$ from Lemma 8.2.14.

We cannot directly apply Lemma 8.2.7 as stated, as the flow $f_{\mathcal{P}}$ does not satisfy the flow-conductance ratio α in Lemma 8.2.2, but essentially the same proof will work to get the same conclusion (but not exactly the same intermediate step).

Lemma 8.2.15. *Suppose the input instance satisfies the conditions in Assumption 8.2.1. Let x^* be an optimal solution to* (st-CP) *and f^* be the unit s-t electrical flow supported in x^*. The expected s-t effective resistance of the subgraph H returned by the randomized short path rounding algorithm is*

$$\mathbb{E}\left[\mathrm{Reff}_H(s,t)\right] \leqslant \frac{2}{v_{\mathcal{P}}^2}\mathcal{E}_{x^*}(f^*) = \frac{2}{v_{\mathcal{P}}^2}\,\mathrm{Reff}_{x^*}(s,t),$$

where $\mathcal{P}$ is the short path flow decomposition of f^* as described in Definition 8.2.13.

The main difference of the analysis is to apply the Hoeffding's inequality (instead of Markov's inequality) to bound the probability that the returned subgraph has significantly more than k edges.

Lemma 8.2.16. *Suppose the input instance satisfies the conditions in Assumption 8.2.1. Let x^* be an optimal solution to* (st-CP) *and f^* be the unit s-t electrical flow supported in x^*. Let H be the subgraph returned by the randomized short path rounding algorithm given x^* as input, and $|E(H)|$ be the number of edges in H. Then, for any $\delta > 0$,*

$$\mathbb{P}\left[|E(H)| \geqslant (1+\delta)k\right] \leqslant \exp\left(-\frac{2\delta^2}{c^2\alpha}\right),$$

where c is the parameter in the short path flow decomposition in Definition 8.2.13 and α is the flow-conductance ratio of f^* and x^* as defined in Lemma 8.2.2.

Proof. As in Lemma 8.2.6, we assume pessimistically that all integral edges E_I will be included in H, and so we focus on the fractional edges E_F. Let $X_{i,e}$ be the indicator variable of whether the edge e is sampled in the i-th iteration of the short path rounding algorithm, and $X_{i,F} := \sum_{e \in E_F} X_{i,e}$ be the total number of fractional edges sampled in the i-th iteration. Let X_F be the total number of fractional edges in H. Note that $X_F \leqslant \sum_{i=1}^{\tau} X_{i,F}$, since if some fractional edge was sampled in different iterations, we only count it once in X_F. By linearity of expectation, $\mathbb{E}\left[X_F\right] \leqslant \sum_{i=1}^{\tau} \mathbb{E}\left[X_{i,F}\right]$.

Let $\mathcal{P}^*$ be the flow path decomposition of f^* in Lemma 8.2.4, and $\mathcal{P}$ be the short path flow decomposition of f^* as described in Definition 8.2.13. For an edge e, recall that $\mathsf{f}_{\mathcal{P}}(e) = \sum_{p \in \mathcal{P}: p \ni e} v_p$ is the total flow value on e from the short paths in $\mathcal{P}$. As we scaled the probability of each path by $1/v_{\mathcal{P}}$ in the rounding algorithm, the probability that edge e is sampled in the i-th iteration is $\mathsf{f}_{\mathcal{P}}(e)/v_{\mathcal{P}}$. Let $\overline{\mathsf{f}}(e) := \mathsf{f}^*(e) - \mathsf{f}_{\mathcal{P}}(e)$ be the total flow value on e from the long paths in $\mathcal{P}^* - \mathcal{P}$. The expected value of $X_{i,F}$ is

$$\mathbb{E}\left[X_{i,F}\right] = \sum_{e \in E_F} \mathbb{E}\left[X_{i,e}\right] = \sum_{e \in E_F} \frac{\mathsf{f}_{\mathcal{P}}(e)}{v_{\mathcal{P}}} = \sum_{e \in E_F} \frac{\mathsf{f}^*(e) - \overline{\mathsf{f}}(e)}{v_{\mathcal{P}}} = \sum_{e \in E_F} \frac{\alpha \mathsf{x}^*(e) - \overline{\mathsf{f}}(e)}{v_{\mathcal{P}}}$$

By the definition of the long paths,

$$\sum_{e \in E_F} \overline{\mathsf{f}}(e) = \sum_{e \in E_F} \sum_{p \in \mathcal{P}^* - \mathcal{P}} v_p = \sum_{p \in \mathcal{P}^* - \mathcal{P}} v_p \cdot |p \cap E_F| \geqslant c\alpha x_F^* \sum_{p \in \mathcal{P}^* - \mathcal{P}} v_p = c\alpha x_F^* (1 - v_{\mathcal{P}}),$$

where we recall that $x_F^* = \sum_{e \in E_F} x_e^*$. Therefore,

$$\mathbb{E}\left[X_{i,F}\right] = \sum_{e \in E_F} \frac{\alpha \mathsf{x}^*(e) - \overline{\mathsf{f}}(e)}{v_{\mathcal{P}}} \leqslant \alpha x_F^* \cdot \frac{1 - c + c v_{\mathcal{P}}}{v_{\mathcal{P}}} = \alpha x_F^* \cdot (c - \frac{c-1}{v_{\mathcal{P}}}) \leqslant \alpha x_F^*,$$

where the last inequality uses that $v_{\mathcal{P}} \leqslant 1$ and $c > 1$. It follows that $\mathbb{E}\left[X_F\right] \leqslant \tau \alpha x_F^* \leqslant x_F^*$.

As each iteration is independent, the random variables $X_{i,F}$ for $1 \leqslant i \leqslant \tau$ are independent. Since we only use short paths, the maximum value of each $X_{i,F}$ is at most $c\alpha x_F^*$. So we can apply Hoeffding's inequality Theorem 3.1.3 to show that

$$\mathbb{P}\left[X_F \geqslant (1+\delta) x_F^*\right] \leqslant \exp\left(-\frac{2\delta^2 (x_F^*)^2}{\tau c^2 \alpha^2 (x_F^*)^2}\right) \leqslant \exp\left(-\frac{2\delta^2}{c^2 \alpha}\right).$$

Let X_I be the total number of integral edges in H. As $X_I \leqslant |E_I|$, we conclude that

$$\begin{aligned}\mathbb{P}\left[|E(H)| \geqslant (1+\delta)k\right] &= \mathbb{P}\left[X_I + X_F \geqslant (1+\delta)(|E_I| + x_F^*)\right] \\ &\leqslant \mathbb{P}\left[X_F \geqslant (1+\delta)x_F^*\right] \leqslant \exp\left(-\frac{2\delta^2}{c^2\alpha}\right). \qquad \square\end{aligned}$$

As in Section 8.2.3, we can combine Lemma 8.2.16 and Lemma 8.2.15 to show that the randomized short path rounding algorithm is a bicriteria approximation algorithm.

Theorem 8.2.17. *Suppose the input instance satisfies the conditions in Assumption 8.2.1. Suppose further that $k \geqslant d_{st}/\varepsilon^{10}$, where $\varepsilon > 0$ is an error parameter satisfying $\varepsilon \leqslant \eta$ for a small constant η. Let x^* be an optimal solution to (st-CP). Given x^*, the randomized short path rounding algorithm with $c = \frac{1}{\varepsilon}$ will return a subgraph H with at most $(1+\varepsilon)k$ edges and $\mathrm{Reff}_H(s,t) \leqslant (2+10\varepsilon)\cdot \mathrm{Reff}_{\mathsf{x}^*}(s,t)$ with probability at least ε.*

Proof. The additional assumption $k \geqslant d_{st}/\varepsilon^{10}$ implies that $\alpha \leqslant \varepsilon^5$ by Lemma 8.2.3.

Setting $c = \frac{1}{\varepsilon}$ and $\delta = \varepsilon$, it follows from Lemma 8.2.16 that

$$\mathbb{P}\left[|E(H)| \geqslant (1+\varepsilon)k\right] \leqslant \exp\left(-\frac{2\delta^2}{c^2\alpha}\right) \leqslant \exp\left(-\frac{2}{\varepsilon}\right) < \varepsilon,$$

where the last inequality holds for $\varepsilon > 0$.

Since $c = \frac{1}{\varepsilon}$, Lemma 8.2.14 implies that $v_{\mathcal{P}} \geqslant 1-\varepsilon$ for the short path flow decomposition in Definition 8.2.13. Using Markov's inequality and Lemma 8.2.15, for sufficiently small ε we have

$$\begin{aligned}\mathbb{P}\left[\mathrm{Reff}_H(s,t) \geqslant (2+10\varepsilon)\cdot\mathrm{Reff}_{\mathsf{x}^*}(s,t)\right] &\leqslant \frac{\mathbb{E}\left[\mathrm{Reff}_H(s,t)\right]}{(2+10\varepsilon)\cdot\mathrm{Reff}_{\mathsf{x}^*}(s,t)} \\ &\leqslant \frac{2}{v_{\mathcal{P}}^2(2+10\varepsilon)} \leqslant \frac{2}{(1-\varepsilon)^2(2+10\varepsilon)} < 1-2\varepsilon.\end{aligned}$$

Therefore, with probability at least ε, the subgraph H returned by the randomized short path rounding algorithm satisfies both properties. $\square$

Using the same arguments as in Section 8.2.4, we can turn the above bicriteria approximation algorithm into a true approximation algorithm.

Theorem 8.2.18. *Suppose the input instance satisfies the conditions in Assumption 8.2.1. Suppose further that $k \geqslant 2d_{st}/\varepsilon^{10}$, where $\varepsilon > 0$ is an error parameter satisfying $\varepsilon \leqslant \eta$ for a small constant η. There is a polynomial time $(2 + O(\varepsilon))$-approximation algorithm for the s-t effective resistance network design problem.*

Proof. As in the proof of Theorem 8.2.10, we apply the bicriteria approximation algorithm in Theorem 8.2.17 with input x^*, an optimal solution to (st-CP) with the scaled-down budget $\frac{k}{1+\varepsilon}$, to return a subgraph H. As the new budget $\frac{k}{1+\varepsilon}$ is still greater than d_{st}/ε^{10}, by Theorem 8.2.17, with probability at least ε the subgraph H satisfies $|E(H)| \leqslant \frac{(1+\varepsilon)k}{1+\varepsilon} = k$ and

$$\mathrm{Reff}_H(s,t) \leqslant \big(2 + O(\varepsilon)\big) \cdot \mathsf{opt}\left(\frac{k}{1+\varepsilon}\right) \leqslant \big(2 + O(\varepsilon)\big)(1+\varepsilon) \cdot \mathsf{opt}(k) \leqslant \big(2 + O(\varepsilon)\big) \cdot \mathsf{opt}(k),$$

where we used the notations and arguments in Theorem 8.2.10.

For the time complexity, note that $\alpha \leqslant \varepsilon^5$ by Lemma 8.2.3 and the large budget assumption, and so we can assume that $\varepsilon^5 \geqslant \alpha \geqslant \frac{1}{4m}$, as otherwise there is a simple 2-approximation algorithm in the case $\alpha \leqslant \frac{1}{4m}$ described in Theorem 8.2.10. Therefore, the success probability can be boosted to a constant in polynomial number of executions of the bicriteria algorithm in Theorem 8.2.17. □

8.2.6 Cost Minimization with s-t Effective Resistance Constraint

In this subsection, we consider a "dual" problem of the s-t effective resistance minimization problem. In the dual problem, we are given a graph $G = (V, E)$ and a target effective resistance R, and the objective is to find a subgraph H of minimum number of edges such that $\mathrm{Reff}_H(s,t) \leqslant R$. The same NP-hardness proof in Section 8.3.1 can be used to show that the dual problem is NP-complete.

Using the same techniques for the s-t effective resistance minimization problem, we can obtain a constant factor bicriteria approximation algorithm for this problem. As the proofs are very similar, we will just state the results and highlight the differences. The main difference is that the convex program has unbounded integrality gap, and as a consequence we cannot turn the bicriteria approximation algorithm into a true approximation algorithm as in the s-t effective resistance network design problem. Using the same technique as in Theorem 8.2.10, however, we can return an 8-approximation to the optimal number of edges without violating the effective resistance constraint, if we are allowed to buy up to four copies of the same edge (see Theorem 8.2.19).

Convex Programming Relaxation

We consider the following natural convex programming relaxation for the dual problem.

$$\begin{aligned} \underset{x \in \mathbb{R}^E}{\text{minimize}} \quad & \sum_{e \in E} x(e) \\ \text{subject to} \quad & \mathrm{Reff}_{st}(x) \leqslant R, \\ & 0 \leqslant x(e) \leqslant 1 \qquad \forall e \in E. \end{aligned} \tag{DCP}$$

Integrality Gap Examples

Unlike the s-t effective resistance network design problem, the convex program (DCP) has unbounded integrality gap. Consider the following example in Figure 8.3, where the top path has length $n-1$, and the bottom path has only one edge. The target effective resistance is $R = \frac{(n-1)^2}{(n-1)^2+\varepsilon}$ for some constant $\varepsilon > 0$. Since $R < 1$, to satisfy the effective resistance constraint, any integral solution must contain both paths and thus has cost n. However, the fractional solution can set $x(e) = \frac{\varepsilon}{n-1}$ for each edge in the top path and set $x(e) = 1$ for the bottom edge. It can be checked that this fractional solution satisfies the constraint, and the total cost is $1+\varepsilon$. Therefore, the integrality gap of this example is $\Omega(n)$.

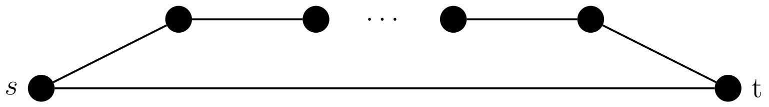

Figure 8.3: $\Omega(n)$ integrality gap example.

Optimal Solutions

Although the convex program (DCP) has a large integrality gap, the same rounding technique can be used to obtain a constant factor bicriteria approximation algorithm.

Exactly the same characterization of the optimality conditions as in the s-t effective resistance network design problem holds, such that any optimal solution satisfies the flow-conductance ratio $\alpha > 0$ as described in Lemma 8.2.2.

Analogous to Lemma 8.2.3, we can prove an upper bound on α that

$$\alpha^2 \leqslant \frac{R}{d_{st}}.$$

Analogous to Lemma 8.2.9, we can prove a lower bound on any optimal solution x that

$$\mathsf{opt} := \sum_{e \in E} \mathsf{x}(e) \geqslant \frac{d_{st}^2}{R}.$$

We can assume that $R < d_{st}$, as otherwise a shortest s-t path is an optimal solution, and so we can assume that $0 < \alpha < 1$.

Rounding Algorithm

The rounding algorithm is exactly the same as in Section 8.2.3. The same proofs as in Lemma 8.2.6 and Lemma 8.2.7 will imply that, with probability $\Omega(\alpha)$, the subgraph H returned by the randomized path rounding algorithm satisfies

$$|E(H)| \leqslant 2 \sum_{e \in E} \mathsf{x}^*(e) \quad \text{and} \quad \mathrm{Reff}_H(s,t) \leqslant 4\, \mathrm{Reff}_{\mathsf{x}^*}(s,t),$$

where x^* is an optimal solution to (DCP) and so $|E(H)| \leqslant 2\mathsf{opt}$. The same lower bound on $\alpha = \Omega(\frac{1}{m})$ as described in Theorem 8.2.10 applies, and so this is a randomized polynomial time algorithm.

An Alternative Bicriteria Approximation Algorithm

In the s-t effective resistance network design problem, we turn a bicriteria approximation algorithm into a true approximation algorithm, by scaling down the budget k by a factor of two and running the bicriteria approximation algorithm. For the proof, we argue $\mathsf{opt}(\frac{k}{2}) \leqslant 2\mathsf{opt}(k)$ by scaling down an optimal solution x^* with budget k to a solution $\frac{1}{2}\mathsf{x}^*$ with budget $\frac{k}{2}$.

In the dual problem, we can also try a similar approach, by scaling down the target effective resistance R by a factor of 4 and run the bicriteria approximation algorithm. However, we cannot argue that $\mathsf{opt}(\frac{R}{4}) \leqslant 4\mathsf{opt}(R)$, as an optimal solution x^* with effective resistance R may not be able to scale up to $4\mathsf{x}^*$ with effective resistance $R/4$ because of the capacity constraints $0 \leqslant \mathsf{x}(e) \leqslant 1$ for $e \in E$. This approach would work if we are allowed to violate the capacity constraint by a factor of 4.

Theorem 8.2.19. *Given an weighted input graph $G = (V, E)$, there is a polynomial time algorithm for the dual problem which returns a multi-subgraph H with $|E(H)| \leqslant 8\mathsf{opt}$ and $\mathrm{Reff}_H(s,t) \leqslant R$ where there are at most 4 parallel copies of each edge.*

8.3 Hardness

In this section, we first prove that the s-t effective resistance network design problem is NP-hard in Section 8.3.1. Then, we prove that the weighted problem is APX-hard assuming the small-set expansion conjecture in Section 8.3.2.

8.3.1 NP-Hardness

We will prove Theorem 8.1.1 in this subsection. The following is the decision version of the problem.

Problem 8.3.1 (s-t effective resistance network design)**.**

Input: *An undirected graph $G = (V, E)$, two vertices $s, t \in V$, and two parameters k and R.*

Question: *Does there exist a subgraph H of G with at most k edges and* $\mathrm{Reff}_H(s,t) \leqslant R$*?*

We will show that this problem is NP-complete by a reduction from the 3-Dimensional Matching (3DM) problem.

Problem 8.3.2 (3-Dimensional Matching)**.**

Input: *Three disjoint sets of elements $X = \{x_1, \ldots, x_q\}, Y = \{y_1, \ldots, y_q\}, Z = \{z_1, \ldots, z_q\}$; a set of triples $\mathcal{T} \subseteq X \times Y \times Z$ where each triple contains exactly one element in X, Y, Z.*

Question: *Does there exist a subset of q pairwise disjoint triples in $\mathcal{T}$?*

Reduction: Given an instance of 3DM with $\{(X, Y, Z), \mathcal{T}\}$, let $\tau = |\mathcal{T}|$ and denote the triples by $\mathcal{T} = \{T_1, \ldots, T_\tau\}$. We construct a graph $G = (V, E)$ as follows:

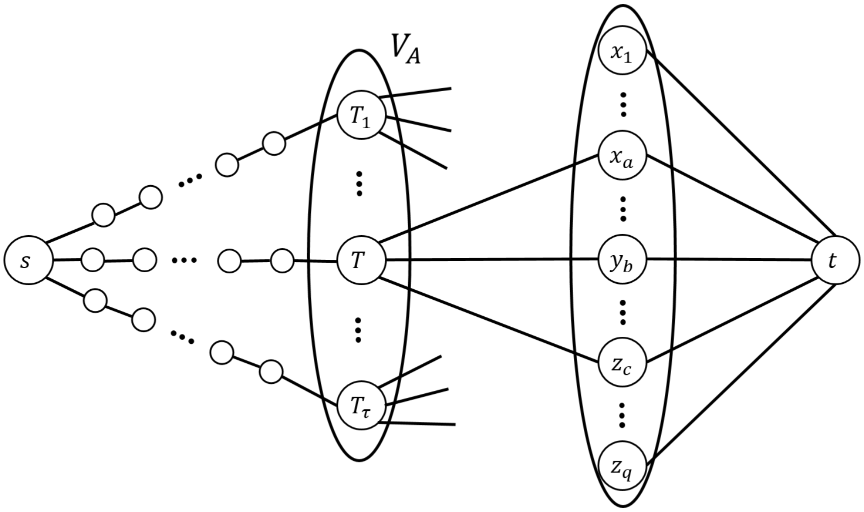

Figure 8.4: An illustration of the construction of the graph G from a 3DM instance.

Vertex Set: The vertex set V is the disjoint union of five sets $\{s\}, \{t\}, V_A$, V_B, and D. Each vertex in V_A corresponds to a triple in $\mathcal{T}$, that is $V_A = \{T_1, \ldots, T_\tau\}$. Each vertex in V_B corresponds to an element in $X \cup Y \cup Z$, that is $V_B = \{x_1, \ldots, x_q, y_1, \ldots, y_q, z_1, \ldots, z_q\}$. Let $l = 3\tau + 3q$. The set D consists of $\tau \cdot l$ "dummy" vertices $\{d_{i,j} \mid 1 \leqslant i \leqslant \tau, 1 \leqslant j \leqslant l\}$. So, there are totally $\tau + 3q + 2 + \tau(3\tau + 3q)$ vertices in G, which is polynomial in the input size of the 3DM instance.

Edge Set: The edge set E is the disjoint union of three edge sets F_1, F_2 and P. There are 3τ edges in F_1, where we have three edges (T, x_a), (T, y_b) and (T, z_c) for each triple $T = (x_a, y_b, z_c) \in \mathcal{T}$. There are $3q$ edges in F_2, where there is an edge from each vertex in V_B to t. There are $\tau(l+1)$ edges in P, where there is a path $P_i := (s, d_{i,1}, d_{i,2}, \ldots, d_{i,l}, T_i)$ for each triple $T_i \in \mathcal{T}$, $1 \leqslant i \leqslant \tau$. So, there are totally $3\tau + 3q + \tau(3\tau + 3q + 1)$ edges in E, which is polynomial in the input size of the 3DM instance.

The following claim completes the proof of Theorem 8.1.1.

Lemma 8.3.3. *Let $k = q(l+1) + 3\tau + 3q$ and $R = (3(l+1)+2)/3q$. The 3DM instance has q disjoint triples if and only if the graph G has a subgraph H with at most k edges and* $\mathrm{Reff}_H(s,t) \leqslant R$.

Proof. One direction is easy. If there are q disjoint triples in the 3DM instance, say $\{T_1, \ldots, T_q\}$, then H will consist of the q paths $P_1, \ldots, P_q$, the $3q$ edges in F_1 incident on $T_1, \ldots, T_q$, and all the $3q$ edges in F_2. There are $(l+1)q + 3q + 3q \leqslant k$ edges in H, and $\mathrm{Reff}_H(s,t) = \frac{l+1}{q} + \frac{1}{3q} + \frac{1}{3q} = \frac{3(l+1)+2}{3q} = R$, as in the graph in Figure 8.5.

The other direction is more interesting. If there do not exist q disjoint triples in the 3DM instance, then we need to argue that $\mathrm{Reff}_H(s,t) > R$ for any H with at most k edges. First, note that $k < (q+1)(l+1)$, and so the budget is not enough for us to buy more than q paths. As it is useless to buy only a proper subset of a path, we can thus assume that H consists of q paths and all the edges in F_1, F_2. H has a total of exactly $q(l+1)+3\tau+3q = k$ edges. For any such H, we will argue that $\mathrm{Reff}_H(s,t) > R$. Without loss of generality, assume that H consists of $P_1, \ldots, P_q$ and all edges in F_1 and F_2. As $T_1, \ldots, T_q$ are not disjoint, there are some vertices in V_B that are not neighbors of $T_1 \cup \ldots \cup T_q$. Call those vertices U.

We consider the following modifications of H to obtain H', and use $\text{Reff}_{H'}(s,t)$ to lower bound $\text{Reff}_H(s,t)$. For every pair of vertices in V_B, we add an edge of zero resistance. For each edge incident on $T_{q+1}, \ldots, T_\tau$, we decrease its resistance to zero. By the monotonicity principle, the modifications will not increase the s-t effective resistance, as we either add edges with zero resistance or decrease the resistance of existing edges. The modifications are equivalent to contracting the vertices with zero resistance edges in between, and so H' is equivalent to the graph in Figure 8.5. Therefore, we have $\text{Reff}_H(s,t) \geqslant \text{Reff}_{H'}(s,t) \geqslant R$.

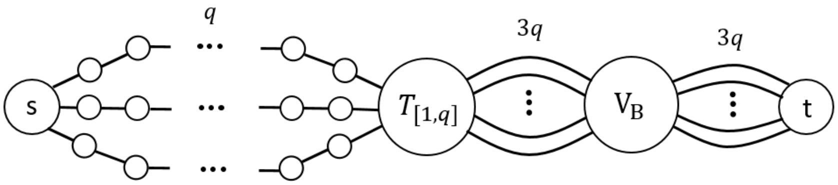

Figure 8.5: The subgraph H when the 3DM instance has q disjoint triples.

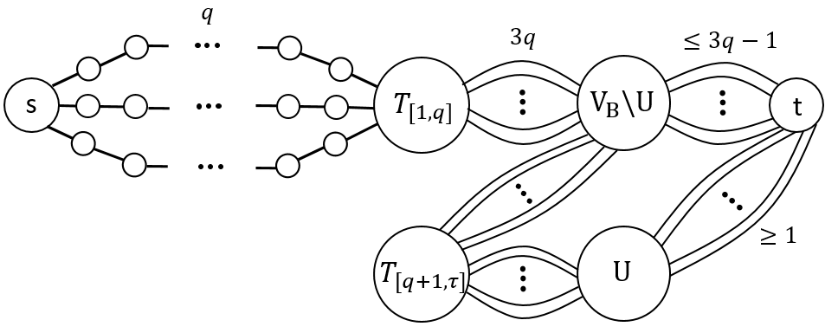

Figure 8.6: The subgraph H when U is non-empty.

We will prove that one of the inequalities in $\text{Reff}_H(s,t) \geqslant \text{Reff}_{H'}(s,t) \geqslant R$ must be strict when $U \neq \emptyset$ (Figure 8.6). To argue the strict inequality, we look at the unit s-t electrical flow f in H and consider two cases.

- If there exists some vertex $u \in U$ with no incoming electrical flow, then we can delete such a vertex without changing $\mathrm{Reff}_H(s,t)$. But then in the modified graph H', the number of parallel edges to t is now strictly smaller than $3q$, and therefore $\mathrm{Reff}_{H'}(s,t) > R$.

- If there exists some vertex $u \in U$ with some incoming electrical flow, then $f(T_j u) > 0$ for some $j \geqslant q+1$. Since we have decreased the resistance of such an edge $T_j u$ to 0, the energy of f in H' is strictly smaller than the energy of f in H. By Thomson's principle, we have $\mathrm{Reff}_{H'}(s,t) \leqslant \mathcal{E}_{H'}(f) < \mathcal{E}_H(f) = \mathrm{Reff}_H(s,t)$.

Since the 3DM instance has no q disjoint triples, it follows that $U \neq \emptyset$ and thus one of the above two cases must apply. In either case, we have $\mathrm{Reff}_H(s,t) > R$ and this completes the proof of the other direction. □

8.3.2 Improved Hardness Assuming Small-Set Expansion Conjecture

In this subsection, we will prove Theorem 8.1.4 that it is NP-hard to approximate the weighted s-t effective resistance network design problem within a factor smaller than 2. First, we will state the small-set expansion conjecture and its variant on bipartite graphs, and present an overview of the proof in Section 8.3.2.1. Next, we will reduce the bipartite small-set expansion problem to the weighted s-t effective resistance network design problem in Section 8.3.2.2, and then reduce the small-set expansion problem to the bipartite small-set expansion problem in Section 8.3.2.3 to complete the proof.

8.3.2.1 The Small-Set Expansion Conjecture and Proof Overview

The gap small-set expansion problem is formulated by Raghavendra and Steurer [123]. We use the version stated in [124].

Definition 8.3.4 (Gap Small-Set Expansion Problem [123, 124])**.** *Given an undirected graph $G = (V, E)$, two parameters $0 < \beta < \alpha < 1$ and $\delta > 0$, the (α, β)-gap δ-small-set expansion problem, denoted by $\mathrm{SSE}_\delta(\alpha, \beta)$, is to distinguish between the following two cases.*

- Yes*: There exists a subset $S \subseteq V$ with $\mathrm{vol}(S) = \delta \,\mathrm{vol}(V)$ and $\phi(S) \leqslant \beta$.*
- No*: Every subset $S \subseteq V$ with $\mathrm{vol}(S) = \delta \,\mathrm{vol}(V)$ has $\phi(S) \geqslant \alpha$.*

It is conjectured in [123] that the gap small-set expansion problem becomes harder when δ becomes smaller.

Conjecture 8.3.5 (Small-Set Expansion Conjecture [123, 124])**.** *For any $\varepsilon \in (0, \frac{1}{2})$, there exists sufficiently small $\delta > 0$ such that $\mathrm{SSE}_\delta(1 - \varepsilon, \varepsilon)$ is NP-hard even for regular graphs.*

It is known that the small-set expansion conjecture implies the Unique Game conjecture [123] and is equivalent to some variant of the Unique Game Conjecture [124].

We will show the SSE-hardness of the weighted s-t effective resistance network design problem in two steps, and use the small-set expansion problem on regular *bipartite* graphs as an intermediate problem.

Proposition 8.3.6. *For any $\varepsilon > 0$, there is a polynomial time reduction from $\mathrm{SSE}_\delta(1-\varepsilon, \varepsilon)$ on d-regular graphs to $\mathrm{SSE}_\delta(1 - 16\varepsilon, \varepsilon)$ on d-regular bipartite graphs.*

Proposition 8.3.7. *Given an instance of $\mathrm{SSE}_\delta(\alpha, \beta)$ on a d-regular bipartite graph B, there is a polynomial time algorithm to construct an instance of the weighted s-t effective resistance network design problem with graph G and cost budget k satisfying the following properties.*

- *If B is a* Yes*-instance, then there is a subgraph H of G with cost at most k and*

$$\mathrm{Reff}_H(s, t) \leqslant \frac{2}{(1 - \beta)dk}.$$

- *if B is a* No*-instance, then every subgraph H of G with cost at most k has*

$$\mathrm{Reff}_H(s,t) \geqslant \frac{2}{(1-\frac{\alpha}{2})dk}.$$

Theorem 8.1.4 will follow immediately from the two propositions.

Theorem 8.3.8. *For any $\varepsilon' > 0$, it is NP-hard to approximate the weighted s-t effective resistance network design problem to within a factor of $2-\varepsilon'$, assuming that $\mathrm{SSE}_\delta(1-\varepsilon,\varepsilon)$ is NP-hard on regular graphs for sufficiently small $\varepsilon > 0$.*

Proof. First, given a d-regular instance of $\mathrm{SSE}_\delta(1-\varepsilon,\varepsilon)$, we apply Proposition 8.3.6 to obtain a d-regular bipartite instance of $\mathrm{SSE}_\delta(1-16\varepsilon,\varepsilon)$. Then, we apply Proposition 8.3.7 with $\alpha = 1-16\varepsilon$ and $\beta = \varepsilon$ and see that the ratio between the s-t effective resistance of the No-case and the Yes-case is at least

$$\frac{(1-\beta)dk}{(1-\frac{\alpha}{2})dk} = \frac{1-\varepsilon}{\frac{1}{2}+8\varepsilon} = \frac{2(1-\varepsilon)}{1+16\varepsilon} > 2-\varepsilon',$$

for sufficiently small ε. □

We prove Proposition 8.3.7 in Section 8.3.2.2 and Proposition 8.3.6 in Section 8.3.2.3.

8.3.2.2 From Bipartite Small-Set Expansion to weighted s-t Effective Resistance Network Design

We prove Proposition 8.3.7 in this subsection. In the Yes-case of bipartite SSE, we use the small dense subgraph (from the small low conductance set) to construct a small subgraph with small s-t effective resistance. In the No-case of bipartite SSE, we argue that every small subgraph has considerably larger s-t effective resistance.

Construction: Given an $\mathrm{SSE}_\delta(\alpha,\beta)$ instance with a d-regular bipartite graph $B = (V_X, V_Y; E_B)$, we construct an instance of the weighted s-t effective resistance network design problem with graph $G = (V,E)$ as follows. See Figure 8.7 for an illustration.

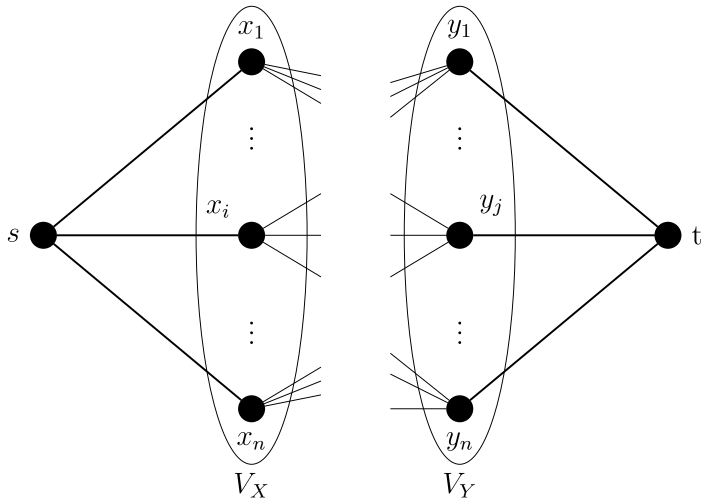

Figure 8.7: Reduction from bipartite small set expansion to weighted s-t effective resistance network design.

Vertex Set: The vertex set V of G is simply the disjoint union of $\{s\}, V_X, V_Y, \{t\}$.

Edge Set: The edge set E of G is the disjoint union of three edge sets E_s, E_B, E_t. The edge set E_s has $|V_X|$ edges, where there is an edge from s to each vertex $v \in V_X$. The edge set E_t has $|V_Y|$ edges, where there is an edge from each vertex $v \in V_Y$ to t.

Costs and Resistances: Every edge e in E_B has $c_e = 0$ and $r_e = 1$. Every edge $e \in E_s \cup E_t$ has $c_e = 1$ and $r_e = 0$.

Budget: The cost budget k is $\delta|V_X \cup V_Y|$.

YES-**case:** Suppose B is a YES-instance of $\text{SSE}_\delta(\alpha, \beta)$. Since B is regular, there exist subsets $X \subseteq V_X$ and $Y \subseteq V_Y$ such that $|X \cup Y| = \delta|V_X \cup V_Y| = k$ and $\phi_B(X \cup Y) \leqslant \beta$. We construct the subgraph H of G as follows.

Subgraph H: The subgraph H includes all the edges from s to X, all the edges from X to Y, and all the edges from Y to t. Since edges from X to Y are of cost zero, the total cost in H is equal to $|X| + |Y| = k$.

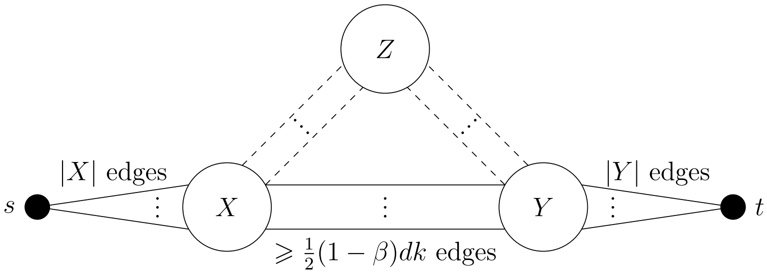

Figure 8.8: In the YES-case, the solid edges are included in H and the dashed edges are deleted.

The following claim will complete the proof of the first item of Proposition 8.3.7.

Lemma 8.3.9. $\mathrm{Reff}_H(s,t) \leqslant \frac{2}{(1-\beta)dk}$.

Proof. Since B is a d-regular bipartite graph, we have

$$d(|X|+|Y|) = \mathrm{vol}_B(X \cup Y) = |\delta_B(X \cup Y)| + 2|E_B(X,Y)|,$$

where $E_B(X,Y)$ denotes the set of edges with one endpoint in X and one endpoint in Y. Since $\phi_B(X \cup Y) \leqslant \beta$, we have $|\delta_B(X \cup Y)| \leqslant \beta \cdot \mathrm{vol}_B(X \cup Y) = d\beta(|X|+|Y|)$. Hence, the number of edges between X and Y is

$$|E_B(X,Y)| = \frac{d(|X|+|Y|) - |\delta_B(X \cup Y)|}{2} \geqslant \frac{1}{2}(1-\beta)d(|X|+|Y|) = \frac{1}{2}(1-\beta)dk.$$

In terms of s-t effective resistance, H is equivalent to the graph in Figure 8.8, where $Z = (V_X \backslash X) \cup (V_Y \backslash Y)$ is the set of vertices not in X and Y. Since the edges from s to X and from Y to t have zero resistance and edges between X and Y have resistance one, we have $\mathrm{Reff}_H(s,t) \leqslant \frac{2}{(1-\beta)dk}$. □

NO**-case:** We will prove the second item of Proposition 8.3.7 by arguing that every subgraph of B with total cost at most k has considerably larger s-t effective resistance. Since all the edges between V_X and V_Y have zero cost and adding edges never increases s-t effective resistance (by Rayleigh's monotonicity principle), we can assume without loss of generality that any solution H to the weighted s-t effective resistance network design problem takes all edges between V_X and V_Y and also takes exactly k edges from $E_s \cup E_t$. Consider an arbitrary subgraph H with the above properties. Let $X \subseteq V_X$ be the set of neighbors of s and $Y \subseteq V_Y$ be the set of neighbors of t, with $|X| + |Y| = k$. Let $\phi := \phi_B(X \cup Y)$. Note that $\phi \geqslant \alpha$ as we are in the No-case where $\phi_B(X \cup Y) \geqslant \alpha$ for every $|X \cup Y| = k$. Using the same calculation as above, we have

$$|E_B(X,Y)| = \frac{1}{2}(1 - \phi_B(X \cup Y))dk = \frac{1}{2}(1-\phi)dk.$$

The subgraph H is shown in Figure 8.9, where $Z = (V_X \backslash X) \cup (V_Y \backslash Y)$ is the set of vertices not in X and Y, and the edges within Z are not shown. To lower bound $\mathrm{Reff}_H(s,t)$, we modify H to obtain H' and argue that $\mathrm{Reff}_H(s,t) \geqslant \mathrm{Reff}_{H'}(s,t)$ and then show a lower bound on $\mathrm{Reff}_{H'}(s,t)$.

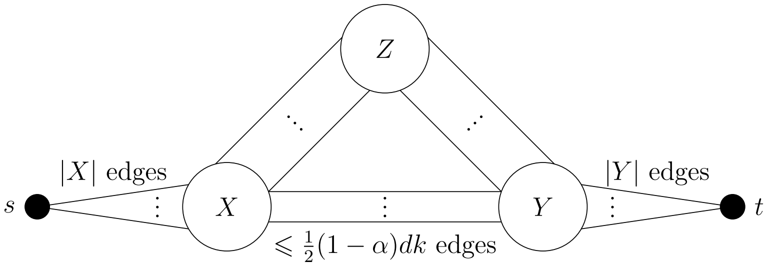

Figure 8.9: The subgraph H' is obtained by identifying the subsets X, Y, Z into single vertices.

To obtain H' from H, we simply identify the three subsets of vertices X, Y, Z to three vertices, which is equivalent to adding a clique of zero resistance edges to each of these three subsets. By Rayleigh's monotonicity principle, this could only decrease the s-t effective resistance and so we have $\mathrm{Reff}_H(s,t) \geqslant \mathrm{Reff}_{H'}(s,t)$.

In terms of s-t effective resistance, the subgraph H' is equivalent to the graph with two paths between X and Y (with parallel edges): one path P_1 of length one with $|E_B(X,Y)|$ parallel edges between X and Y, another path P_2 of length two with $|E_B(X,Z)|$ parallel edges between X and Z and $|E_B(Z,Y)|$ parallel edges between Z and Y. To lower bound $\mathrm{Reff}_{H'}(s,t)$, we lower bound the resistance of P_1 and P_2, denoted by $r(P_1)$ and $r(P_2)$. Note that

$$r(P_1) = \frac{1}{E_B(X,Y)} = \frac{2}{(1-\phi)dk}.$$

For $r(P_2)$, let $x = |\delta_B(X,Z)|$ and $y = |\delta_B(Y,Z)|$, then

$$r(P_2) = \frac{1}{x} + \frac{1}{y} = \frac{1}{x+y} \cdot \frac{(x+y)^2}{xy} = \frac{1}{x+y} \cdot \left(\frac{x}{y} + \frac{y}{x} + 2\right) \geqslant \frac{4}{x+y} = \frac{4}{\phi dk},$$

where the inequality holds since $a + 1/a \geqslant 2$ for any $a > 0$, and the last equality holds because $x + y = |\delta_B(X \cup Y)| = \phi dk$. Finally, by Fact 2.4.1,

$$\begin{aligned} \mathrm{Reff}_H(s,t) \geqslant \mathrm{Reff}_{H'}(s,t) = & \frac{1}{1/r(P_1) + 1/r(P_2)} \\ & \geqslant \frac{1}{\frac{1}{2}(1-\phi)dk + \frac{1}{4}\phi dk} = \frac{2}{(1-\phi/2)dk} \geqslant \frac{2}{(1-\alpha/2)dk}, \end{aligned}$$

where the last inequality is because we are in the No-case. This completes the proof of the second item of Proposition 8.3.7.

Remark 8.3.10. *In this subsection, we show the hardness of the weighted s-t effective resistance network design problem, when the edge cost and the edge resistance could be arbitrary. Using a similar argument as in the proof of Theorem 8.1.1, the reduction can be modified to the unit-cost case if we replace the edges from s to V_X and V_Y to t by sufficiently long paths (so that the cost of connecting s to a vertex in V_X is much larger than the cost of connecting a vertex in V_X to a vertex in V_Y). Therefore, the same $(2-\varepsilon)$-SSE-hardness also holds in the case when every edge has the same cost.*

8.3.2.3 From Small Set Expansion to Bipartite Small Set Expansion

We prove Proposition 8.3.6 in this subsection.

Construction: Given an instance $\mathrm{SSE}_\delta(1-\varepsilon,\varepsilon)$ on a d-regular graph $G = (V,E)$, we construct a d-regular bipartite graph $B = (V_X, V_Y; E_B)$ as follows. For each vertex v in V, we create a vertex $v_X \in V_X$ and a vertex $v_Y \in V_Y$, so that $|V_X| = |V_Y| = |V|$. For each edge $uv \in E$, we add two edges $u_X v_Y$ and $u_Y v_X$ to E_B. It is clear from the construction that B is d-regular.

Correctness: To prove Proposition 8.3.6, we will establish the following two claims.

1. Yes-**case:** If there is a set $S \subseteq V$ with $|S| = \delta|V|$ and $\phi_G(S) \leqslant \varepsilon$ in G, then there exist $X \subseteq V_X$ and $Y \subseteq V_Y$ with $|X| + |Y| = \delta(|V_X| + |V_Y|)$ and $\phi_B(X \cup Y) \leqslant \varepsilon$ in B.
2. No-**case:** If every set $S \subseteq V$ with $|S| = \delta|V|$ has $\phi_G(S) \geqslant 1 - \varepsilon$ in G, then every sets $X \subseteq V_X$ and $Y \subseteq V_Y$ with $|X| + |Y| = \delta(|V_X| + |V_Y|)$ has $\phi_B(X \cup Y) \geqslant 1 - 16\varepsilon$ in B.

Yes-**case:** Let $S \subseteq V$ be a subset with $|S| = \delta|V|$ and $\phi_G(S) \leqslant \varepsilon$ in G. Let $S_X := \{v_X \mid v \in S\}$ and $S_Y := \{v_Y \mid v \in S\}$, with $|S| = |S_X| = |S_Y|$. By construction, an edge $uv \in \delta_G(S)$ if and only if both $u_X v_Y$ and $v_X u_Y$ are in $\delta_B(S_X \cup S_Y)$, and thus $|\delta_B(S_X \cup S_Y)| = 2|\delta_G(S)|$. Since $|S_X \cup S_Y| = |S_X| + |S_Y| = 2|S|$ and B is d-regular, we have

$$\phi_B(S_X \cup S_Y) = \frac{|\delta_B(S_X \cup S_Y)|}{\mathrm{vol}_B(S_X \cup S_Y)} = \frac{|\delta_B(S_X \cup S_Y)|}{d(|S_X| + |S_Y|)} = \frac{2|\delta_G(S)|}{2d|S|} = \phi_G(S) \leqslant \varepsilon.$$

No-**case:** Consider arbitrary subsets $X \subseteq V_X$ and $Y \subseteq V_Y$ with $|X| + |Y| = \delta(|V_X| + |V_Y|) = 2\delta|V|$. To lower bound $\phi_B(X \cup Y)$, we will upper bound $|E_B(X, Y)|$. We partition X into groups $X_1, \ldots, X_a$ where every group except the last group is of size $\delta|V|/2$ and the last group is of size at most $\delta|V|/2$. We partition Y into groups $Y_1, \ldots, Y_b$ in a similar way. The following claim uses the small-set expansion property in G to show that there is no small dense subset in B.

Lemma 8.3.11. *Suppose G is a* No*-instance of* $\mathrm{SSE}_\delta(1 - \varepsilon, \varepsilon)$*. Then, for any* $1 \leqslant i \leqslant a$ *and* $1 \leqslant j \leqslant b$*,*

$$|E_B(X_i, Y_j)| \leqslant \varepsilon\delta d|V|.$$

Proof. We first argue that there is no small dense subset in G, and then we will use it to bound $|E_B(X_i, Y_j)|$. Suppose $S \subseteq V$ with $|S| = \delta|V|$. As G is a No-instance, we know that $\phi_G(S) \geqslant 1 - \varepsilon$ and thus $|\delta_G(S)| \geqslant (1 - \varepsilon)\,\mathrm{vol}_G(S) = (1 - \varepsilon)d|S|$. Since $d|S| = \mathrm{vol}_G(S) = |\delta_G(S)| + 2|E_G(S, S)|$, it follows that $|E_G(S, S)| \leqslant \varepsilon d|S|/2 = \varepsilon\delta d|V|/2$. Note that this also implies trivially that $|E_G(Z, Z)| \leqslant \varepsilon\delta d|V|/2$ for any Z with $|Z| \leqslant \delta|V|$.

Given X_i and Y_j, let $Z := \{v \in G \mid v_X \in X_i \text{ or } v_Y \in Y_j\}$. In words, Z is the set of vertices in G which have at least one copy in $X_i \cup Y_j$ in B. Since each X_i and Y_j is of size at most $\delta|V|/2$, it follows that $|Z| \leqslant \delta|V|$. Also, note that $|E_B(X_i, Y_j)| \leqslant 2|E_G(Z, Z)|$, as each edge in $E_B(X_i, Y_j)$ corresponds to one edge in $E_G(Z, Z)$ while each edge in $E_G(Z, Z)$ is corresponded to at most two edges in $E_B(X_i, Y_j)$. Therefore, we can apply the bound in the previous paragraph to conclude that $|E(X_i, Y_j)| \leqslant 2|E_G(Z, Z)| \leqslant \varepsilon\delta d|V|$. □

We now use the lemma to bound $|E_B(X, Y)|$. Since $|X| + |Y| = 2\delta|V|$, it follows that $a \leqslant 4$ and $b \leqslant 4$, and therefore

$$|E_B(X, Y)| \leqslant \sum_{i=1}^{a} \sum_{j=1}^{b} |E_B(X_i, Y_j)| \leqslant ab\varepsilon\delta d|V| \leqslant 16\varepsilon\delta d|V|.$$

As B is bipartite,

$$|\delta_B(X \cup Y)| = \mathrm{vol}_B(X \cup Y) - 2|E_B(X, Y)| \geqslant 2\delta d|V| - 32\varepsilon\delta d|V| = 2(1 - 16\varepsilon)\delta d|V|.$$

Therefore, we have

$$\phi_B(X \cup Y) = \frac{|\delta_B(X \cup Y)|}{\mathrm{vol}_B(X \cup Y)} \geqslant \frac{2(1 - 16\varepsilon)\delta d|V|}{2\delta d|V|} = 1 - 16\varepsilon.$$

This completes the proof of Proposition 8.3.6. We remark that a more careful argument gives $|E_B(X, Y)| \leqslant 6\varepsilon\delta d|V|$ and thus $\phi_B(X \cup Y) \geqslant 1 - 6\varepsilon$, but this constant does not matter for the proof of Theorem 8.3.8.

Chapter 9

Conclusion and Future Work

In this thesis, we studied the spectral rounding problem, which is an extension of the spectral sparsification problem introduced by Spielman and Teng [133]. We demonstrated that the spectral rounding problem underlies a large family of network design and experimental design problems. Our results showed that this spectral approach significantly extends the scope of the traditional survivable network design, and also finds applications in various other graph problems (e.g., spectral network design problems, additive spectral sparsification) in the literature. We also showed that the techniques developed for spectral rounding provide a unified framework for the experimental design problems, which matches and improves the state-of-the-art. Going beyond spectral rounding, we studied the s-t effective resistance network design problem. We provided a constant approximation algorithm that works in the regime where spectral rounding does not apply.

We believe that the linear algebraic perspective and spectral approach will bring new techniques and stronger results to network design and potentially other combinatorial optimization problems. The spectral approaches that we discussed in this thesis opens up many interesting new directions to investigate.

- Our current result of the two-sided spectral rounding is based on Theorem 2.6.10 from [91], whose argument relies on the nonconstructive method of interlacing polynomials. Can we find an efficient algorithm for the two-sided spectral rounding result

Theorem 5.1.4? It is a major open problem to achieve constructive results for the method of interlacing polynomials.

- For network design problems in Section 6.1, there are several interesting questions that one can investigate. Can we recover Jain's iterative rounding result with a spectral approach for survivable network design on undirected graphs? Can we extend the spectral approach to settings such as directed graphs, vertex-connectivity, etc.? Note that Jain's iterative rounding has good performance under these settings.

- We can also ask another question regarding to the generalized network design problem. As shown by the tight example in Section 5.4, the additive error term in the approximation guarantee of Theorem 6.1.9 is optimal even if there is only an algebraic connectivity constraint. Nevertheless, can we improve the additive error term if there are only effective resistance constraints involved? In particular, can we improve the dependence of ε?

- For the experimental design problems in Chapter 7, it is natural to consider minimizing $\left(\operatorname{tr}\left(\frac{1}{d}(\sum_{i\in S} \boldsymbol{u}_i \boldsymbol{u}_i^\top)^{-p}\right)\right)^{\frac{1}{p}}$ as the objective. Note that this objective function is an interpolation between D-design ($p \to 0$), A-design ($p = 1$) and E-design ($p \to \infty$) objectives. Can we extend the current techniques to solve experimental design with this general objective function?

- Similarly, for the s-t effective resistance network design problem in Chapter 8, can we find a good approximation algorithm for minimizing the generalized ℓ_p-energy for all $p \geq 1$?

References

[1] Ajit Agrawal, Philip Klein, and R. Ravi. When trees collide: an approximation algorithm for the generalized Steiner problem on networks. *SIAM J. Comput.*, 24(3):440–456, 1995.

[2] Rudolf Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, 2002.

[3] Arthur Albert. Conditions for positive and nonnegative definiteness in terms of pseudoinverses. *SIAM J. Appl. Math.*, 17:434–440, 1969.

[4] Vedat Levi Alev, Nima Anari, Lap Chi Lau, and Shayan Oveis Gharan. Graph clustering using effective resistance. In Anna R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94, pages 41:1–41:16, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[5] Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal design of experiments via regret minimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *ICML'17*, pages 126–135. JMLR.org, 2017.

[6] Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal discrete optimization for experimental design: A regret minimization approach. *arXiv preprint arXiv:1711.05174*, 2017.

[7] Zeyuan Allen-Zhu, Zhenyu Liao, and Lorenzo Orecchia. Spectral sparsification and regret minimization beyond matrix multiplicative updates. In *Proceedings of the 47th*

Annual ACM Symposium on Theory of Computing, STOC '15, pages 237–245, New York, NY, USA, 2015. ACM.

[8] Nima Anari and Shayan Oveis Gharan. The Kadison-Singer problem for strongly Rayleigh measures and applications to asymmetric TSP. *arXiv preprint arXiv:1412.1143*, 2014.

[9] Nima Anari and Shayan Oveis Gharan. Effective-resistance-reducing flows, spectrally thin trees, and asymmetric TSP. In *Proceedings of the 56th Annual Symposium on Foundations of Computer Science*, FOCS '15, pages 20–39, Washington, DC, USA, 2015. IEEE Computer Society.

[10] Dana Angluin. Queries and concept learning. *Mach. Learn.*, 2(4):319–342, 1988.

[11] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory Comput.*, 8:121–164, 2012.

[12] Arash Asadpour, Michel X. Goemans, Aleksander Mądry, Shayan Oveis Gharan, and Amin Saberi. An $O(\log n/\log\log n)$-approximation algorithm for the asymmetric traveling salesman problem. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, page 379–389, USA, 2010. Society for Industrial and Applied Mathematics.

[13] A. C. Atkinson and A. N. Donev. *Optimum experimental designs.* Clarendon Press, 1992.

[14] A. C. Atkinson, A. N. Donev, and R. D. Tobias. *Optimum experimental designs, with SAS*, volume 34 of *Oxford Statistical Science Series.* Oxford University Press, Oxford, 2007.

[15] Haim Avron and Christos Boutsidis. Faster subset selection for matrices and applications. *SIAM J. Matrix Anal. Appl.*, 34(4):1464–1499, 2013.

[16] Nikhil Bansal. On a generalization of iterated and randomized rounding. In *Proceedings of the 51st Annual ACM Symposium on Theory of Computing*, STOC '19, pages 1125–1135, New York, NY, USA, 2019. ACM.

[17] Nikhil Bansal, Daniel Dadush, Shashwat Garg, and Shachar Lovett. The Gram–Schmidt walk: A cure for the banaszczyk blues. *Theory of Computing*, 15(21):1–27, 2019.

[18] Nikhil Bansal and Shashwat Garg. Algorithmic discrepancy beyond partial coloring. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC '17, page 914–926, New York, NY, USA, 2017. Association for Computing Machinery.

[19] Nikhil Bansal, Rohit Khandekar, Jochen Könemann, Viswanath Nagarajan, and Britta Peis. On generalizations of network design problems with degree bounds. *Math. Program.*, 141(1-2, Ser. A):479–506, 2013.

[20] Nikhil Bansal, Ola Svensson, and Luca Trevisan. New notions and constructions of sparsification for graphs and hypergraphs. In *Proceedings of the 60th Annual Symposium on Foundations of Computer Science*, FOCS '19, pages 910–928, Washington, DC, USA, 2019. IEEE Computer Society.

[21] Joshua Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. *SIAM J. Comput.*, 41(6):1704–1721, 2012.

[22] Heinz H. Bauschke and Jonathan M. Borwein. Legendre functions and the method of random Bregman projections. *J. Convex Anal.*, 4(1):27–67, 1997.

[23] András A. Benczúr and David R. Karger. Approximating s-t minimum cuts in $\tilde{O}(n^2)$ time. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '96, page 47–55, New York, NY, USA, 1996. Association for Computing Machinery.

[24] Rajendra Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[25] Vittorio Bilò, Vineet Goyal, R. Ravi, and Mohit Singh. On the crossing spanning tree problem. In Klaus Jansen, Sanjeev Khanna, José D. P. Rolim, and Dana Ron, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 51–60. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[26] Julius Borcea and Petter Brändén. Applications of stable polynomials to mixed determinants: Johnson's conjectures, unimodality, and symmetrized Fischer products. *Duke Math. J.*, 143(2):205–223, 2008.

[27] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities.* Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[28] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Sparse features for PCA-like linear regression. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 2285–2293, Red Hook, NY, USA, 2011. Curran Associates Inc.

[29] Christos Boutsidis and Malik Magdon-Ismail. Deterministic feature selection for k-means clustering. *IEEE Trans. Inform. Theory*, 59(9):6099–6110, 2013.

[30] Stephen Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing Markov chain on a graph. *SIAM Rev.*, 46(4):667–689, 2004.

[31] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, 2004.

[32] Jarosł aw Byrka, Fabrizio Grandoni, Thomas Rothvoss, and Laura Sanità. Steiner tree approximation via iterative randomized rounding. *J. ACM*, 60(1):Art. 6, 33, 2013.

[33] Ali Çivril and Malik Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoret. Comput. Sci.*, 410(47-49):4801–4811, 2009.

[34] Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth Vishnoi. Fair and diverse DPP-based data summarization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 716–725, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[35] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge, 2006.

[36] Tanmoy Chakraborty, Julia Chuzhoy, and Sanjeev Khanna. Network design for vertex connectivity. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, STOC '08, pages 167–176, New York, NY, USA, 2008. ACM.

[37] Luiz F. O. Chamon and Alejandro Ribeiro. Greedy sampling of graph signals. *IEEE Trans. Signal Process.*, 66(1):34–47, 2018.

[38] Pak Hay Chan, Lap Chi Lau, Aaron Schild, Sam Chiu-wai Wong, and Hong Zhou. Network design for s-t effective resistance. *arXiv preprint arXiv:1904.03219*, 2019.

[39] Ashok K. Chandra, Prabhakar Raghavan, Walter L. Ruzzo, Roman Smolensky, and Prasoon Tiwari. The electrical resistance of a graph captures its commute and cover times. *Comput. Complexity*, 6(4):312–340, 1996/97.

[40] Siheng Chen, Aliaksei Sandryhaila, José M. F. Moura, and Jelena Kovačević. Signal recovery on graphs: variation minimization. *IEEE Trans. Signal Process.*, 63(17):4609–4624, 2015.

[41] Siheng Chen, Rohan Varma, Aarti Singh, and Jelena Kovačević. Signal recovery on graphs: fundamental limits of sampling strategies. *IEEE Trans. Signal Inform. Process. Netw.*, 2(4):539–554, 2016.

[42] Joseph Cheriyan and László A. Végh. Approximating minimum-cost k-node connected subgraphs via independence-free graphs. *SIAM J. Comput.*, 43(4):1342–1362, 2014.

[43] Joseph Cheriyan, Santosh Vempala, and Adrian Vetta. Network design via iterative rounding of setpair relaxations. *Combinatorica*, 26(3):255–275, 2006.

[44] Paul Christiano, Jonathan A. Kelner, Aleksander Mądry, Daniel A. Spielman, and Shang-Hua Teng. Electrical flows, Laplacian systems, and faster approximation of

maximum flow in undirected graphs. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, STOC '11, pages 273–282, New York, NY, USA, 2011. ACM.

[45] Maria Chudnovsky and Paul Seymour. The roots of the independence polynomial of a clawfree graph. *J. Combin. Theory Ser. B*, 97(3):350–357, 2007.

[46] Julia Chuzhoy and Sanjeev Khanna. An $O(k^3 \log n)$-approximation algorithm for vertex-connectivity survivable network design. *Theory Comput.*, 8:401–413, 2012.

[47] R. Dennis Cook and Christopher J. Nachtrheim. A comparison of algorithms for constructing exact D-optimal designs. *Technometrics*, 22(3):315–324, 1980.

[48] Marcel K. de Carli Silva, Nicholas J. A. Harvey, and Cristiane M. Sato. Sparse sums of positive semidefinite matrices. *ACM Trans. Algorithms*, 12(1):Art. 9, 17, 2016.

[49] Jean-Pierre Dedieu. Obreschkoff's theorem revisited: what convex sets are contained in the set of hyperbolic polynomials? *J. Pure Appl. Algebra*, 81(3):269–278, 1992.

[50] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. *Theory Comput.*, 2:225–247, 2006.

[51] Amit Deshpande and Santosh Vempala. Adaptive sampling and fast low-rank matrix approximation. In *Proceedings of the 9th International Conference on Approximation Algorithms for Combinatorial Optimization Problems, and 10th International Conference on Randomization and Computation*, APPROX'06/RANDOM'06, page 292–303, Berlin, Heidelberg, 2006. Springer-Verlag.

[52] Jian Ding, James R. Lee, and Yuval Peres. Cover times, blanket times, and majorizing measures. *Ann. of Math. (2)*, 175(3):1409–1471, 2012.

[53] Yevgeniy Dodis and Sanjeev Khanna. Design networks with bounded pairwise distance. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, STOC '99, pages 750–759, New York, NY, USA, 1999. ACM.

[54] David Durfee, Rasmus Kyng, John Peebles, Anup B. Rao, and Sushant Sachdeva. Sampling random spanning trees faster than matrix multiplication. In *Proceedings of the 49th Annual ACM Symposium on Theory of Computing*, STOC '17, pages 730–742, New York, NY, USA, 2017. ACM.

[55] Jeremy Elson, Richard M. Karp, Christos H. Papadimitriou, and Scott Shenker. Global synchronization in sensornets. In Martín Farach-Colton, editor, *LATIN 2004: Theoretical Informatics*, pages 609–624, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[56] Alina Ene and Ali Vakilian. Improved approximation algorithms for degree-bounded network design problems with node connectivity requirements. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 754–763, New York, NY, USA, 2014. ACM.

[57] David Eppstein, Zvi Galil, Giuseppe F. Italiano, and Amnon Nissenzweig. Sparsification—a technique for speeding up dynamic graph algorithms. *J. ACM*, 44(5):669–696, 1997.

[58] Jittat Fakcharoenphol and Bundit Laekhanukit. An $O(\log^2 k)$-approximation algorithm for the k-vertex connected spanning subgraph problem. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, STOC '08, pages 153–158, New York, NY, USA, 2008. ACM.

[59] Shaun M. Fallat, Steve Kirkland, and Sukanta Pati. On graphs with algebraic connectivity equal to minimum edge density. *Linear Algebra Appl.*, 373:31–50, 2003. Special issue on the Combinatorial Matrix Theory Conference (Pohang, 2002).

[60] V. V. Fedorov. *Theory of optimal experiments*. Academic Press, New York-London, 1972. Translated from the Russian and edited by W. J. Studden and E. M. Klimko, Probability and Mathematical Statistics, No. 12.

[61] H. J. Fell. On the zeros of convex combinations of polynomials. *Pacific J. Math.*, 89(1):43–50, 1980.

[62] Lisa Fleischer, Kamal Jain, and David P Williamson. An iterative rounding 2-approximation algorithm for the element connectivity problem. In *Proceedings of the 42nd Annual Symposium on Foundations of Computer Science*, FOCS '01, pages 339–347, Las Vegas, NV, USA, 2001. IEEE, IEEE Computer Society.

[63] David A. Freedman. On tail probabilities for martingales. *Ann. Probability*, 3:100–118, 1975.

[64] Takuro Fukunaga, Zeev Nutov, and R. Ravi. Iterative rounding approximation algorithms for degree-bounded node-connectivity network design. *SIAM J. Comput.*, 44(5):1202–1229, 2015.

[65] Martin Fürer and Balaji Raghavachari. Approximating the minimum-degree Steiner tree to within one of optimal. *J. Algorithms*, 17(3):409–423, 1994. Third Annual ACM-SIAM Symposium on Discrete Algorithms (Orlando, FL, 1992).

[66] Harold N. Gabow. On the L_∞-norm of extreme points for crossing supermodular directed network LPs. *Math. Program.*, 110(1, Ser. B):111–144, 2007.

[67] Harold N. Gabow, Michel X. Goemans, Éva Tardos, and David P. Williamson. Approximating the smallest k-edge connected spanning subgraph by LP-rounding. *Networks*, 53(4):345–357, 2009.

[68] Naveen Garg, Goran Konjevod, and R. Ravi. A polylogarithmic approximation algorithm for the group Steiner tree problem. *J. Algorithms*, 37(1):66–84, 2000. Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 1998).

[69] Arpita Ghosh and Stephen Boyd. Growing well-connected graphs. In *Proceedings of the 45th IEEE Conference on Decision and Control*, CDC '06, pages 6605–6611. IEEE, IEEE, 2006.

[70] Arpita Ghosh, Stephen Boyd, and Amin Saberi. Minimizing effective resistance of a graph. *SIAM Rev.*, 50(1):37–66, 2008.

[71] M. X. Goemans, A. V. Goldberg, S. Plotkin, D. B. Shmoys, É. Tardos, and D. P. Williamson. Improved approximation algorithms for network design problems. In

Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '94, pages 223–232, Philadelphia, PA, USA, 1994. Society for Industrial and Applied Mathematics.

[72] Michel X. Goemans. Minimum bounded degree spanning trees. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, page 273–282, USA, 2006. IEEE Computer Society.

[73] Michel X. Goemans and David P. Williamson. A general approximation technique for constrained forest problems. *SIAM J. Comput.*, 24(2):296–317, 1995.

[74] Peter Goos and Bradley Jones. *Optimal design of experiments: a case study approach.* John Wiley & Sons, 2011.

[75] Fabrizio Grandoni, Bundit Laekhanukit, and Shi Li. $O(\log^2 k/\log\log k)$-approximation algorithm for directed steiner tree: A tight quasi-polynomial-time algorithm. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC '19, page 253–264, New York, NY, USA, 2019. Association for Computing Machinery.

[76] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis.* Grundlehren Text Editions. Springer-Verlag, Berlin, 2001. Abridged version of ıt Convex analysis and minimization algorithms. I [Springer, Berlin, 1993; MR1261420 (95m:90001)] and ıt II [ibid.; MR1295240 (95m:90002)].

[77] Roger A. Horn and Charles R. Johnson. *Matrix analysis.* Cambridge University Press, Cambridge, second edition, 2013.

[78] Rabih A Jabr, Ravindra Singh, and Bikash C Pal. Minimum loss network reconfiguration using mixed-integer convex programming. *IEEE Transactions on Power systems*, 27(2):1106–1115, 2012.

[79] Kamal Jain. A factor 2 approximation algorithm for the generalized Steiner network problem. *Combinatorica*, 21(1):39–60, 2001.

[80] Siddharth Joshi and Stephen Boyd. Sensor selection via convex optimization. *IEEE Trans. Signal Process.*, 57(2):451–462, 2009.

[81] Richard V. Kadison and I. M. Singer. Extensions of pure states. *Amer. J. Math.*, 81:383–400, 1959.

[82] David R. Karger. Using randomized sparsification to approximate minimum cuts. In *Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '94, page 424–432, USA, 1994. Society for Industrial and Applied Mathematics.

[83] David R. Karger. Random sampling in cut, flow, and network design problems. *Math. Oper. Res.*, 24(2):383–413, 1999.

[84] William Thomson Baron Kelvin and Peter Guthrie Tait. *Treatise on natural philosophy*, volume 1. Clarendon Press, 1867.

[85] Gustav Kirchhoff. Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird. *Annalen der Physik*, 148(12):497–508, 1847.

[86] Alexandra Kolla, Yury Makarychev, Amin Saberi, and Shang-Hua Teng. Subgraph sparsification and nearly optimal ultrasparsifiers. In *Proceedings of the 42nd ACM Symposium on Theory of Computing*, STOC '10, pages 57–66, New York, NY, USA, 2010. ACM.

[87] Guy Kortsarz, Robert Krauthgamer, and James R. Lee. Hardness of approximation for vertex-connectivity network design problems. *SIAM J. Comput.*, 33(3):704–720, 2004.

[88] Ioannis Koutis, Alex Levin, and Richard Peng. Faster spectral sparsification and numerical algorithms for SDD matrices. *ACM Trans. Algorithms*, 12(2):Art. 17, 16, 2016.

[89] Ioannis Koutis, Gary L. Miller, and Richard Peng. A nearly-$m \log n$ time solver for SDD linear systems. In *Proceedings of the 2011 IEEE 52nd Annual Symposium*

on Foundations of Computer Science, FOCS '11, page 590–598, USA, 2011. IEEE Computer Society.

[90] Rasmus Kyng, Yin Tat Lee, Richard Peng, Sushant Sachdeva, and Daniel A. Spielman. Sparsified cholesky and multigrid solvers for connection Laplacians. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 842–850, New York, NY, USA, 2016. Association for Computing Machinery.

[91] Rasmus Kyng, Kyle Luh, and Zhao Song. Four deviations suffice for rank 1 matrices. *Adv. Math.*, 375:107366, 17, 2020.

[92] Bundit Laekhanukit. Parameters of two-prover-one-round game and the hardness of connectivity problems. In *Proceedings of the Twenty-fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1626–1643, Philadelphia, PA, USA, 2014. Society for Industrial and Applied Mathematics.

[93] Lap Chi Lau, Joseph Naor, Mohammad R. Salavatipour, and Mohit Singh. Survivable network design with degree or order constraints. *SIAM J. Comput.*, 39(3):1062–1087, 2009.

[94] Lap Chi Lau, R. Ravi, and Mohit Singh. *Iterative methods in combinatorial optimization.* Cambridge Texts in Applied Mathematics. Cambridge University Press, New York, 2011.

[95] Lap Chi Lau and Mohit Singh. Additive approximation for bounded degree survivable network design. *SIAM J. Comput.*, 42(6):2217–2242, 2013.

[96] Lap Chi Lau and Hong Zhou. A unified algorithm for degree bounded survivable network design. *Math. Program.*, 154(1-2, Ser. B):515–532, 2015.

[97] Lap Chi Lau and Hong Zhou. A local search framework for experimental design. *arXiv preprint arXiv:2010.15805*, 2020.

[98] Lap Chi Lau and Hong Zhou. A spectral approach to network design. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC '20, page 826–839, New York, NY, USA, 2020. Association for Computing Machinery.

[99] James R. Lee. University of Washington Computer Science, CSE 599S, Lecture Notes: Entropy optimality, 2016. URL: https://homes.cs.washington.edu/~jrl/teaching/cse599swi16/. Last visited on 2020/09/19.

[100] Yin Tat Lee and He Sun. An SDP-based algorithm for linear-sized spectral sparsification. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC '17, page 678–687, New York, NY, USA, 2017. Association for Computing Machinery.

[101] Yin Tat Lee and He Sun. Constructing linear-sized spectral sparsification in almost-linear time. *SIAM J. Comput.*, 47(6):2315–2336, 2018.

[102] A. S. Lewis. Convex analysis on the Hermitian matrices. *SIAM J. Optim.*, 6(1):164–177, 1996.

[103] Huan Li and Zhongzhi Zhang. Kirchhoff index as a measure of edge centrality in weighted networks: Nearly linear time algorithms. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, page 2377–2396, USA, 2018. Society for Industrial and Applied Mathematics.

[104] Elliott H. Lieb. Convex trace functions and the Wigner-Yanase-Dyson conjecture. *Advances in Math.*, 11:267–288, 1973.

[105] André Linhares and Chaitanya Swamy. Approximating min-cost chain-constrained spanning trees: a reduction from weighted to unweighted problems. *Math. Program.*, 172(1-2, Ser. B):17–34, 2018.

[106] Anand Louis and Nisheeth K. Vishnoi. Improved algorithm for degree bounded survivable network design problem. In Haim Kaplan, editor, *Algorithm Theory - SWAT 2010*, pages 408–419, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[107] Vivek Madan, Aleksandar Nikolov, Mohit Singh, and Uthaipon Tantipongpipat. Maximizing determinants under matroid constraints. *arXiv preprint arXiv:2004.07886*, 2020.

[108] Vivek Madan, Mohit Singh, Uthaipon Tantipongpipat, and Weijun Xie. Combinatorial algorithms for optimal design. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2210–2258, Phoenix, USA, 25–28 Jun 2019. PMLR.

[109] Aleksander Mądry. Navigating central path with electrical flows: From flows to matchings, and back. In *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, FOCS '13, page 253–262, USA, 2013. IEEE Computer Society.

[110] Adam W. Marcus, Daniel A. Spielman, and Nikhil Srivastava. Interlacing families I: Bipartite Ramanujan graphs of all degrees. *Ann. of Math. (2)*, 182(1):307–325, 2015.

[111] Adam W. Marcus, Daniel A. Spielman, and Nikhil Srivastava. Interlacing families II: Mixed characteristic polynomials and the Kadison-Singer problem. *Ann. of Math. (2)*, 182(1):327–350, 2015.

[112] Peter Matthews. Covering problems for Brownian motion on spheres. *Ann. Probab.*, 16(1):189–199, 1988.

[113] Alan J. Miller and Nam-Ky Nguyen. A Fedorov exchange algorithm for d-optimal design. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(4):669–677, 1994.

[114] Baharan Mirzasoleiman. *Big Data Summarization Using Submodular Functions*. PhD thesis, ETH Zurich, Zurich, 2017.

[115] Aleksander Mądry, Damian Straszak, and Jakub Tarnawski. Fast generation of random spanning trees and the effective resistance metric. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, page 2019–2036, USA, 2015. Society for Industrial and Applied Mathematics.

[116] Nam-Ky Nguyen and Alan J. Miller. A review of some exchange algorithms for constructing discrete D-optimal designs. *Comput. Statist. Data Anal.*, 14(4):489–498, 1992.

[117] Aleksandar Nikolov and Mohit Singh. Maximizing determinants under partition constraints. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 192–201, New York, NY, USA, 2016. Association for Computing Machinery.

[118] Aleksandar Nikolov, Mohit Singh, and Uthaipon Tao Tantipongpipat. Proportional volume sampling and approximation algorithms for A-optimal design. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '19, pages 1369–1386, Philadelphia, PA, USA, 2019. Society for Industrial and Applied Mathematics.

[119] Neil Olver and Rico Zenklusen. Chain-constrained spanning trees. *Math. Program.*, 167(2, Ser. A):293–314, 2018.

[120] Richard Peng. Approximate undirected maximum flows in $O(m \operatorname{polylog} n)$ time. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '16, page 1862–1867, USA, 2016. Society for Industrial and Applied Mathematics.

[121] Friedrich Pukelsheim. *Optimal design of experiments*, volume 50 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006. Reprint of the 1993 original.

[122] Balaji Raghavachari. Algorithms for finding low degree structures. *Approximation algorithms for NP-hard problems*, pages 266–295, 1997.

[123] Prasad Raghavendra and David Steurer. Graph expansion and the unique games conjecture. In *Proceedings of the 42nd ACM Symposium on Theory of Computing*, STOC '10, pages 755–764, New York, NY, USA, 2010. ACM.

[124] Prasad Raghavendra, David Steurer, and Madhur Tulsiani. Reductions between expansion problems. In *Proceedings of the 2012 IEEE Conference on Computational Complexity*, CCC '12, pages 64–73, Washington, DC, USA, 2012. IEEE Computer Society.

[125] R. Ravi, M. V. Marathe, S. S. Ravi, D. J. Rosenkrantz, and H. B. Hunt, III. Approximation algorithms for degree-constrained minimum-cost network-design problems. *Algorithmica*, 31(1):58–78, 2001.

[126] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[127] Walter Rudin. *Principles of mathematical analysis*. McGraw-Hill Book Co., New York-Auckland-Düsseldorf, third edition, 1976. International Series in Pure and Applied Mathematics.

[128] Aaron Schild. An almost-linear time algorithm for uniform random spanning tree generation. In *Proceedings of the 50th Annual ACM Symposium on Theory of Computing*, STOC '18, pages 214–227, New York, NY, USA, 2018. ACM.

[129] Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Statistics*, 21:124–127, 1950.

[130] Mohit Singh and Lap Chi Lau. Approximating minimum bounded degree spanning trees to within one of optimal. *J. ACM*, 62(1):Art. 1, 19, 2015.

[131] Mohit Singh and Weijun Xie. Approximate positive correlated distributions and approximation algorithms for d-optimal design. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, page 2240–2255, USA, 2018. Society for Industrial and Applied Mathematics.

[132] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM J. Comput.*, 40(6):1913–1926, 2011.

[133] Daniel A. Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *SIAM J. Comput.*, 40(4):981–1025, 2011.

[134] Daniel A. Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM J. Matrix Anal. Appl.*, 35(3):835–885, 2014.

[135] A. Srinivasan. Distributions on level-sets with applications to approximation algorithms. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, FOCS '01, page 588, USA, 2001. IEEE Computer Society.

[136] Marco Di Summa, Friedrich Eisenbrand, Yuri Faenza, and Carsten Moldenhauer. On largest volume simplices and sub-determinants. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, page 315–323, USA, 2015. Society for Industrial and Applied Mathematics.

[137] Joel A. Tropp. Freedman's inequality for matrix martingales. *Electron. Commun. Probab.*, 16:262–270, 2011.

[138] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[139] David G. Wagner. Multivariate stable polynomials: theory and applications. *Bull. Amer. Math. Soc. (N.S.)*, 48(1):53–84, 2011.

[140] Yining Wang, Adams Wei Yu, and Aarti Singh. On computationally tractable selection of experiments in measurement-constrained regression models. *J. Mach. Learn. Res.*, 18:Paper No. 143, 41, 2017.

[141] Nik Weaver. The Kadison-Singer problem in discrepancy theory. *Discrete Math.*, 278(1-3):227–239, 2004.

[142] William J. Welch. Algorithmic complexity: three NP-hard problems in computational statistics. *J. Statist. Comput. Simulation*, 15(1):17–25, 1982.

[143] Max A. Woodbury. *Inverting modified matrices.* Statistical Research Group, Memo. Rep. no. 42. Princeton University, Princeton, N. J., 1950.

Appendix

Appendix A

Omitted Proofs in the Main Text

A.1 Calculation of Gradients

Fact 2.2.2. *Let $f : \mathbb{S}^d_{++} \to \mathbb{R}$ be defined as $f(X) = \log\det(X)$. Then, f is differentiable at any $X \succ 0$ with $\nabla f(X) = X^{-1}$.*

Proof. For any $X \succ 0$, it suffices to show

$$\lim_{H\to 0} \frac{\log\det(X+H) - \log\det(X) - \langle X^{-1}, H\rangle}{\|H\|_{\mathrm{op}}} = 0.$$

We consider the term $\log\det(X+H)$. It follows that

$$\begin{aligned}
\log\det(X+H) &= \log\det(X)\det(I_d + X^{-\frac{1}{2}} H X^{-\frac{1}{2}}) \\
&= \log\det(X) + \log\prod_{i=1}^{d}(1+\lambda_i) \\
&= \log\det(X) + \sum_{i=1}^{d}\log(1+\lambda_i),
\end{aligned}$$

where $\lambda_1, \ldots, \lambda_d$ are eigenvalues of the matrix $X^{-\frac{1}{2}} H X^{-\frac{1}{2}}$. Thus,

$$\begin{aligned}\log\det(X+H) - \log\det(X) - \langle X^{-1}, H\rangle &= \sum_{i=1}^{d} \log(1+\lambda_i) - \operatorname{tr}(X^{-\frac{1}{2}} H X^{-\frac{1}{2}}) \\ &= \sum_{i=1}^{d} (\log(1+\lambda_i) - \lambda_i) = \sum_{i=1}^{d} \Big(-\frac{\lambda_i^2}{2} + o(\lambda_i^2)\Big),\end{aligned}$$

where the last equality follows by the Taylor series of $\log(1+x)$. Let $\alpha = \|H\|_{\mathrm{op}}$. Since $X \succ 0$, it follows that $|\lambda_i| \leqslant \alpha \cdot \lambda_{\min}(X)^{-1}$ for all $i \in [d]$. Therefore, when $\alpha \to 0$

$$\frac{|\log\det(X+H) = \log\det(X) \cdot \det(I_d + X^{-\frac{1}{2}} H X^{-\frac{1}{2}})|}{\|H\|_{\mathrm{op}}} \leqslant \sum_{i=1}^{d} \frac{O(\alpha)}{\lambda_{\min}(X)^2} \to 0. \qquad \square$$

Fact 2.2.3. *Let $f : \mathbb{S}^d_{++} \to \mathbb{R}$ be defined as $f(X) = \operatorname{tr}(X^{-1})$. Then, f is differentiable at any $X \succ 0$ with $\nabla f(X) = -X^{-2}$.*

Proof. The proof is similar to the previous one, where we verify the solution by the definition. Let $H \in \mathbb{S}^d$ with rank r. Let $H = U\Lambda U^\top$ be the eigendecomposition with $\Lambda \in \mathbb{R}^{r\times r}$ being the diagonal matrix contains all the nonzero eigenvalues. We consider the matrix $(X+H)^{-1}$. For small enough H, $X+H$ is invertible as $X \succ 0$, thus

$$(X+H)^{-1} = (X + U\Lambda U^\top)^{-1} = X^{-1} - X^{-1}U(\Lambda^{-1} + U^\top X^{-1} U)^{-1} U^\top X^{-1},$$

where the last equality follows by Woodburry matrix identity Lemma 2.1.15. Thus,

$$\begin{aligned}\operatorname{tr}((X+H)^{-1}) - \operatorname{tr}(X^{-1}) + \langle X^{-2}, H\rangle &= -\operatorname{tr}(X^{-1}U(\Lambda^{-1} + U^\top X^{-1} U)^{-1} U^\top X^{-1}) + \langle X^{-2}, H\rangle \\ &= -\langle U^\top X^{-2} U, (\Lambda^{-1} + U^\top X^{-1} U)^{-1}\rangle + \langle U^\top X^{-2} U, \Lambda\rangle \\ &= \langle U^\top X^{-2} U, \Lambda - (\Lambda^{-1} + U^\top X^{-1} U)^{-1}\rangle \\ &= \langle U^\top X^{-2} U, \Lambda U^\top (X + U\Lambda U^\top)^{-1} U \Lambda\rangle,\end{aligned}$$

where we apply Woodburry matrix identity Lemma 2.1.15 again in the last equality.

$$(\Lambda^{-1} + U^\top X^{-1} U)^{-1} = \Lambda - \Lambda U^\top (X + U\Lambda U^\top)^{-1} U \Lambda.$$

Let $\alpha = \|H\|_{\mathrm{op}} = \|\Lambda\|_{\mathrm{op}}$. For small enough α, it holds that $X + U\Lambda U^\top \succcurlyeq (\lambda_{\min}(X) - \alpha) I_d \succ 0$. Thus,

$$\left\|\Lambda U^\top (X + U\Lambda U^\top)^{-1} U\Lambda\right\|_{\mathrm{op}} \leqslant \frac{\alpha^2}{\lambda_{\min}(X) - \alpha},$$

which further implies that, when $\alpha \to 0$,

$$\frac{|\operatorname{tr}((X+H)^{-1}) - \operatorname{tr}(X^{-1}) + \langle X^{-2}, H\rangle|}{\|H\|_{\mathrm{op}}} \leqslant \frac{\alpha \operatorname{tr}(U^\top X^{-2} U)}{\lambda_{\min}(X) - \alpha} \to 0. \qquad \square$$

Fact 2.2.4. *Let $f : \mathbb{S}^d_+ \to \mathbb{R}$ be defined as $f(X) = \operatorname{tr}(X^{\frac{1}{2}})$. Then, f is differentiable at any $X \succ 0$ with $\nabla f(X) = \frac{1}{2} X^{-\frac{1}{2}}$.*

Proof. The proof for this fact needs a bit more work then the previous two, since we do not have a nice formula to expand $(X + H)^{\frac{1}{2}}$. Instead of verifying by definition, we are going to calculate the partial derivatives directly, and then use the continuity of the the partial derivatives together with Theorem 2.2.1 to prove the fact.

For $i, j \in [d]$, let $E_{i,j} \in \mathbb{R}^{d\times d}$ denote a matrix with (i, j)-th entry being one and all other entries being 0.

Given a matrix $X \succ 0$, let $\lambda_1 \geqslant \ldots \geqslant \lambda_d > 0$ be the eigenvalues of X, and let $X = U\Lambda U^\top$ be the eigendecomposition of X, where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$, and $U \in \mathbb{R}^{d\times d}$ is an orthonormal matrix. Let $D_{ij} := U(E_{i,j} + E_{j,i})U^\top$ for all $1 \leqslant i < j \leqslant d$, and $D_{ii} := UE_{i,i}U^\top$ for $i \in [d]$. We notice that the $\frac{d(d+1)}{2}$ matrices $\{D_{ij}\}_{1\leqslant i\leqslant j\leqslant d}$ form a basis for the $\frac{d(d+1)}{2}$-dimensional vector space $\mathbb{S}^d$.

We first calculate the directional derivative $f'(X; D_{ii})$ for $i \in [d]$. By definition

$$\begin{aligned} f'(X; D_{ii}) &= \lim_{\delta\to 0} \frac{\operatorname{tr}((X + \delta D_{ii})^{\frac{1}{2}}) - \operatorname{tr}(X^{\frac{1}{2}})}{\delta} = \lim_{\delta\to 0} \frac{\operatorname{tr}((\Lambda + \delta E_{i,i})^{\frac{1}{2}}) - \operatorname{tr}(\Lambda^{\frac{1}{2}})}{\delta} \\ &= \lim_{\delta\to 0} \frac{\sqrt{\lambda_i + \delta} - \sqrt{\lambda_i}}{\delta} = \lim_{\delta\to 0} \frac{1}{2\sqrt{\lambda_i + \delta}} = \frac{1}{2\sqrt{\lambda_i}}, \end{aligned}$$

where the second last inequality follows by L'Hôpital's rule.

Then, we consider the directional derivative $f'(X; D_{ij})$ for $i \neq j \in [d]$. Without loss of generality, we assume $i = 1$ and $j = 2$. We consider the difference of function values of moving in D_{12} direction by δ. Since Λ is a diagonal matrix, and D_{12} only have two nonzero entries, we have

$$\operatorname{tr}((X+\delta D_{12})^{\frac{1}{2}}) - \operatorname{tr}(X^{\frac{1}{2}}) = \operatorname{tr}((\Lambda + \delta(E_{1,2}+E_{2,1}))^{\frac{1}{2}}) - \operatorname{tr}(\Lambda^{\frac{1}{2}}) = \operatorname{tr}(M) - \sqrt{\lambda_1} - \sqrt{\lambda_2},$$

where $M = \left(\begin{smallmatrix} \lambda_1 & \delta \\ \delta & \lambda_2 \end{smallmatrix}\right)^{\frac{1}{2}}$. Notice that

$$\big(\operatorname{tr}(M)\big)^2 = (\lambda_1(M)+\lambda_2(M))^2 = \lambda_1(M)^2+\lambda_2(M)^2+2\lambda_1(M)\lambda_2(M) = \operatorname{tr}(M^2)+2\sqrt{\det(M^2)}.$$

Since $M^2 = \left(\begin{smallmatrix} \lambda_1 & \delta \\ \delta & \lambda_2 \end{smallmatrix}\right)$, we have $\operatorname{tr}(M^2) = \lambda_1+\lambda_2$ and $\det(M^2) = \lambda_1\lambda_2 - \delta^2$. It follows that

$$\operatorname{tr}((X+\delta D_{12})^{\frac{1}{2}}) - \operatorname{tr}(X^{\frac{1}{2}}) = \sqrt{\lambda_1+\lambda_2+2\sqrt{\lambda_1\lambda_2-\delta^2}} - \sqrt{\lambda_1} - \sqrt{\lambda_2}.$$

By definition, the directional derivative is equal to

$$\begin{aligned} f'(X; D_{12}) &= \lim_{\delta\to 0} \frac{\operatorname{tr}((X+\delta D_{12})^{\frac{1}{2}}) - \operatorname{tr}(X^{\frac{1}{2}})}{\delta} \\ &= \lim_{\delta\to 0} \frac{\sqrt{\lambda_1+\lambda_2+2\sqrt{\lambda_1\lambda_2-\delta^2}} - \sqrt{\lambda_1} - \sqrt{\lambda_2}}{\delta} \\ &= \lim_{\delta\to 0} \frac{\mathrm{d}}{\mathrm{d}\delta}\Big(\sqrt{\lambda_1+\lambda_2+2\sqrt{\lambda_1\lambda_2-\delta^2}} - \sqrt{\lambda_1} - \sqrt{\lambda_2}\Big), \end{aligned}$$

where the last equality is by L'Hôpital's rule. When $\delta \to 0$, we have

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\delta}\Big(\sqrt{\lambda_1+\lambda_2+2\sqrt{\lambda_1\lambda_2-\delta^2}} - \sqrt{\lambda_1} - \sqrt{\lambda_2}\Big) \\ &\qquad\qquad = -(\lambda_1+\lambda_2+2\sqrt{\lambda_1\lambda_2-\delta^2})^{-\frac{1}{2}}(\lambda_1\lambda_2-\delta^2)^{-\frac{1}{2}}\delta \to 0. \end{aligned}$$

Thus, we have proved $f'(X; D_{ij}) = 0$ for all $i \neq j \in [d]$, and $f'(X; D_{ii}) = \frac{1}{2\lambda_i}$ for all $i \in [d]$. Since the eigenvalues of a matrix is a continuous function with respect to the entries of the matrix (see, e.g. [77]), all the directional derivatives with respect to the basis $\{D_{ij}\}_{i,j}$ exist and are continuous. By Theorem 2.2.1, the function f is continuously differentiable at X. The gradient with respect to the basis $\{D_{ij}\}_{i,j}$ can be written as $\frac{1}{2}\Lambda^{-\frac{1}{2}}$. Transforming back to the standard basis, the gradient can be represented by $\nabla f(X) = \frac{1}{2}U\Lambda^{-\frac{1}{2}}U^\top = \frac{1}{2}X^{-\frac{1}{2}}$. □

With a similar analysis as in the proof of Fact 2.2.4, we can also prove Fact 2.2.5 for the negative entropy function.

A.2 Convexity and Concavity

Fact 2.2.15. *The function* $f(X) = \log\det(X)$ *is concave on* $\mathbb{S}^d_{++}$.

Proof. According to Lemma 2.2.8, it suffices to show $g(t) = \log\det(X+t(Y-X))$ is concave on $[0,1]$ for any $X, Y \in \mathbb{S}^d_{++}$. Since $X \succ 0$, we rewrite $g(t)$

$$g(t) = \log\det(X + t(Y-X)) = \log\det(X^{\frac{1}{2}}(I + t(X^{-\frac{1}{2}}YX^{-\frac{1}{2}} - I)X^{\frac{1}{2}})$$
$$= \log\det(X) + \log\det(I + t(X^{-\frac{1}{2}}YX^{-\frac{1}{2}} - I)) = \log\det(X) + \sum_{i=1}^{d} \log(1 + t\lambda_i),$$

where $\lambda_1, \ldots, \lambda_d \in \mathbb{R}$ are eigenvalues of $X^{-\frac{1}{2}}YX^{-\frac{1}{2}} - I$. Since both $X, Y \succ 0$, we have $\lambda_1, \ldots, \lambda_d > -1$, and $1 + t\lambda_i > 0$ for $t \in [0,1]$ and all $i = 1, \ldots, d$. Thus, we have

$$g'(t) = \sum_{i=1}^{d} \frac{\lambda_i}{1+t\lambda_i} \qquad \text{and} \qquad g''(t) = -\sum_{i=1}^{d} \frac{\lambda_i^2}{(1+t\lambda_i)^2}.$$

Hence, $g''(t)$ is non-positive on $t \in [0,1]$. Thus, g is concave on $[0,1]$ by Lemma 2.2.11. □

Fact 2.2.16. *The function* $f(X) = \det(X)^{\frac{1}{d}}$ *is concave on* $\mathbb{S}^d_{+}$.

Proof. Given $X, Y \succcurlyeq 0$, assume one of them is singular, say $\det(X) = 0$. Then, for any $\lambda \in [0,1]$, Jensen's inequality holds

$$\lambda\det(X)^{\frac{1}{d}} + (1-\lambda)\det(Y)^{\frac{1}{d}} = (1-\lambda)\det(Y)^{\frac{1}{d}} = \det((1-\lambda)Y)^{\frac{1}{d}} \leqslant \det(\lambda X + (1-\lambda)Y)^{\frac{1}{d}},$$

where the last inequality follows by $X, Y \succcurlyeq 0$.

It remains to consider the case where $X, Y \succ 0$. By Lemma 2.2.8, it suffices to show $g(t) = \det(X + t(Y-X))^{\frac{1}{d}}$ is concave on $[0,1]$. With similar calculations as in the proof of Fact 2.2.15, we have

$$g(t) = \det(X)^{\frac{1}{d}} \cdot \det(I + t(X^{-\frac{1}{2}}YX^{-\frac{1}{2}} - I))^{\frac{1}{d}} = \det(X)^{\frac{1}{d}} \cdot \prod_{i=1}^{d}(1+t\lambda_i)^{\frac{1}{d}},$$

where $\lambda_1, \ldots, \lambda_d > -1$ are eigenvalues of $X^{-\frac{1}{2}} Y X^{-\frac{1}{2}} - I$. Note that $1 + t\lambda_i > 0$ for $t \in [0, 1]$ for all $i = 1, \ldots, d$. Then, we consider the first and second derivative of g.

$$g'(t) = \det(X)^{\frac{1}{d}} \prod_{i=1}^{d} (1 + t\lambda_i)^{\frac{1}{d}} \cdot \sum_{i=1}^{d} \frac{\lambda_i}{d(1 + t\lambda_i)},$$

$$g''(t) = \det(X)^{\frac{1}{d}} \prod_{i=1}^{d} (1 + t\lambda_i)^{\frac{1}{d}} \cdot \left(\left(\sum_{i=1}^{d} \frac{\lambda_i}{d(1 + t\lambda_i)} \right)^2 - \sum_{i=1}^{d} \frac{\lambda_i^2}{d(1 + t\lambda_i)^2} \right).$$

By Cauchy-Schwartz, we have $\left(\sum_{i=1}^{d} \frac{\lambda_i}{d(1+t\lambda_i)} \right)^2 \leqslant d \cdot \sum_{i=1}^{d} \frac{\lambda_i^2}{d^2(1+t\lambda_i)^2} = \sum_{i=1}^{d} \frac{\lambda_i^2}{d(1+t\lambda_i)^2}$. Together with the fact that $1 + t\lambda_i > 0$ for all $t \in [0, 1]$, we have $g''(t) \leqslant 0$ on $[0, 1]$.

Have considered all cases, $f(X)$ is concave on $\mathbb{S}_+^d$. □

A.3 Legendre Function

Lemma A.3.1. *The $\ell_{\frac{1}{2}}$-regularizer $w(X) = -2\operatorname{tr}(X^{\frac{1}{2}})$ is a Legendre function.*

Proof. $w(X) = -2\operatorname{tr}(X^{\frac{1}{2}})$ is a continuous function with a closed domain $\mathcal{D} = \mathbb{S}_+^d$. The differentiability on $\mathbb{S}_{++}^d$ follows from Fact 2.2.4. The strict convexity on $\mathbb{S}_{++}^d$ follows from Fact 2.2.19. It remains to verify the boundary barrier condition.

Let $X \in \mathbb{S}_+^d$ be a singular matrix with rank $< d$, let $Y \in \mathbb{S}_{++}^d$ be an arbitrary positive definite matrix. For any $t \in (0, 1]$, $X + t(Y - X) \succ 0$. Thus, $\nabla w(X + t(Y - X)) = -(X + t(Y - X))^{-\frac{1}{2}}$ by Fact 2.2.4, and

$$\langle \nabla w(X + t(Y - X)), Y - X \rangle = -\langle (X + t(Y - X))^{-\frac{1}{2}}, Y - X \rangle.$$

Let $\lambda_1 \geqslant \ldots \geqslant \lambda_d > 0$ be the eigenvalues of $X + t(Y - X)$, and $\sum_{i=1}^{d} \lambda_i v_i v_i^\top$ be the corresponding eigendecomposition. We first upper bound the smallest eigenvalue λ_d. Take any unit vector v from the null space of X. It holds that

$$\lambda_d \leqslant v^\top (X + t(Y - X)) v = t \cdot v^\top Y v \leqslant t \cdot \lambda_{\max}(Y), \tag{A.1}$$

where the first and the last inequality hold by Theorem 2.1.3.

Let $S := \{i \in [d] \mid \lambda_i < \frac{1}{2}\lambda_{\min}(Y)\}$ be set of small eigenvalues and $L := \{i \in [d] \mid \lambda_i \geqslant \frac{1}{2}\lambda_{\min}(Y)\}$ be the set of large eigenvalues. Note that for small enough t, $\lambda_d \leqslant t \cdot \lambda_{\max}(Y) < \frac{1}{2}\lambda_{\min}(Y)$. Thus, S is not empty. Now, we further rewrite the directional derivative,

$$\begin{aligned} & \langle \nabla w(X + t(Y - X)), Y - X\rangle \\ = \; & -\left\langle \left(\sum\nolimits_{i=1}^{d} \lambda_i v_i v_i^\top\right)^{-\frac{1}{2}}, Y - X \right\rangle = \sum_{i=1}^{d} \frac{-1}{\sqrt{\lambda_i}} \cdot \left(v_i^\top (Y - X) v_i\right) \\ = \; & \sum_{i \in S} \frac{-1}{\sqrt{\lambda_i}} \cdot \left(v_i^\top (Y - X) v_i\right) + \sum_{i \in L} \frac{-1}{\sqrt{\lambda_i}} \cdot \left(v_i^\top (Y - X) v_i\right). \end{aligned} \tag{A.2}$$

Consider the first summation in (A.2). For the small eigenvalues $i \in S$, it holds that

$$\frac{1}{2}\lambda_{\min}(Y) > \lambda_i = v_i^\top (X + t(Y - X)) v_i = (1 - t) v_i^\top X v_i + t v_i^\top Y v_i > (1 - t) v_i^\top X v_i,$$

where the last inequality follows as $Y \succ 0$ and $t > 0$. However, $v_i^\top Y v_i \geqslant \lambda_{\min}(Y)$ for all $i \in [d]$. Thus, $v_i^\top (Y - X) v_i \geqslant \frac{1}{3}\lambda_{\min}(Y) > 0$ for $t \leqslant \frac{1}{4}$ for all $i \in S$. Therefore, when $t \downarrow 0$, the first summation in (A.2)

$$\sum_{i \in S} \frac{-1}{\sqrt{\lambda_i}} \cdot v_i^\top (Y - X) v_i \leqslant \frac{-1}{\sqrt{\lambda_d}} \cdot v_d^\top (Y - X) v_d \leqslant -\frac{\lambda_{\min}(Y)/3}{\sqrt{t \cdot \lambda_{\max}(Y)}} \to -\infty,$$

where the first inequality follows as we have proved $v_i^\top (Y - X) v_i > 0$ for all $i \in S$, the last inequality follows by (A.1).

For the second summation in (A.2), we can upper bound it by

$$\sum_{i \in L} \frac{-1}{\sqrt{\lambda_i}} \cdot v_i^\top (Y - X) v_i \leqslant \sum_{i \in L} \frac{v_i^\top X v_i}{\sqrt{\lambda_i}} \leqslant \frac{\sqrt{2} d \lambda_{\max}(X)}{\sqrt{\lambda_{\min}(Y)}},$$

where the first inequality follows as $Y \succ 0$, and the second inequality follos as $|L| \leqslant d$, $X \succcurlyeq 0$, and $\lambda_i \geqslant \frac{1}{2}\lambda_{\min}(Y) > 0$ for $i \in L$. Notice that the above upper bound does not depend on t.

Combining the first and second summation in (A.2), $\lim_{t \downarrow 0} \langle \nabla w(X + t(Y - X)), Y - X\rangle = -\infty$ as desired. □