Graphs, Principal Minors, and Eigenvalue Problems

by

John C. Urschel

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Mathematics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2021

Author . . . . . .
Department of Mathematics
August 6, 2021

Certified by . . . . . .
Michel X. Goemans
RSA Professor of Mathematics and Department Head
Thesis Supervisor

Accepted by . . . . . .
Jonathan Kelner
Chairman, Department Committee on Graduate Theses

Graphs, Principal Minors, and Eigenvalue Problems

by

John C. Urschel

Submitted to the Department of Mathematics
on August 6, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Mathematics

Abstract

This thesis considers four independent topics within linear algebra: determinantal point processes, extremal problems in spectral graph theory, force-directed layouts, and eigenvalue algorithms. For determinantal point processes (DPPs), we consider the classes of symmetric and signed DPPs, respectively, and in both cases connect the problem of learning the parameters of a DPP to a related matrix recovery problem. Next, we consider two conjectures in spectral graph theory regarding the spread of a graph, and resolve both. For force-directed layouts of graphs, we connect the layout of the boundary of a Tutte spring embedding to trace theorems from the theory of elliptic PDEs, and we provide a rigorous theoretical analysis of the popular Kamada-Kawai objective, proving hardness of approximation and structural results regarding optimal layouts, and providing a polynomial time randomized approximation scheme for low diameter graphs. Finally, we consider the Lanczos method for computing extremal eigenvalues of a symmetric matrix and produce new error estimates for this algorithm.

Thesis Supervisor: Michel X. Goemans
Title: RSA Professor of Mathematics and Department Head

Acknowledgments

I would like to thank my advisor, Michel Goemans. From the moment I met him, I was impressed and inspired by the way he thinks about mathematics. Despite his busy schedule as head of the math department, he always made time when I needed him. He has been the most supportive advisor a student could possibly ask for. He's the type of mathematician I aspire to be. I would also like to thank the other members of my thesis committee, Alan Edelman and Philippe Rigollet. They have been generous with their time and their mathematical knowledge. In my later years as a PhD candidate, Alan has served as a secondary advisor of sorts, and I'm thankful for our weekly Saturday meetings.

I've also been lucky to work with a number of great collaborators during my time at MIT. This thesis would not have been possible without my mathematical collaborations and conversations with Jane Breen, Victor-Emmanuel Brunel, Erik Demaine, Adam Hesterberg, Fred Koehler, Jayson Lynch, Ankur Moitra, Alex Riasanovksy, Michael Tait, and Ludmil Zikatanov. And, though the associated work does not appear in this thesis, I am also grateful for collaborations with Xiaozhe Hu, Dhruv Rohatgi, and Jake Wellens during my PhD. I've been the beneficiary of a good deal of advice, mentorship, and support from the larger mathematics community. This group is too long to list, but I'm especially grateful to Rediet Abebe, Michael Brenner, David Mordecai, Jelani Nelson, Peter Sarnak, Steven Strogatz, and Alex Townsend, among many others, for always being willing to give me some much needed advice during my time at MIT.

I would like to thank my office mates, Fred Koehler and Jake Wellens (and Megan Fu), for making my time in office 2-342 so enjoyable. With Jake's wall mural and very large and dusty rug, our office felt lived in and was a comfortable place to do math. I've had countless interesting conversations with fellow grad students, including Michael Cohen, Younhun Kim, Meehtab Sawhney, and Jonathan Tidor, to name a few. My interactions with MIT undergrads constantly reminded me just how fulfilling and rewarding mentorship can be. I'm thankful to all the support staff members at MIT math, and have particularly fond memories of grabbing drinks with André at the Muddy, constantly bothering Michele with

whatever random MIT problem I could not solve, and making frequent visits to Rosalee for more espresso pods. More generally, I have to thank the entire math department for fostering such a warm and welcoming environment to do math. MIT math really does feel like home.

I'd be remiss if I did not thank my friends and family for their non-academic contributions. I have to thank Ty Howle for supporting me every step of the way, and trying to claim responsibility for nearly all of my mathematical work. This is a hard claim to refute, as it was often only once I took a step back and looked at things from a different perspective that breakthroughs were made. I'd like to thank Robert Hess for constantly encouraging and believing in me, despite knowing absolutely no math. I'd like to thank my father and mother for instilling a love of math in me from a young age. My fondest childhood memories of math were of going through puzzle books at the kitchen table with my mother, and leafing through the math section in the local Chapters in St. Catherines while my father worked in the cafe. I'd like to thank Louisa for putting up with me, moving with me (eight times during this PhD), and proofreading nearly all of my academic papers. I would thank Joanna, but, frankly, she was of little to no help.

This research was supported in part by ONR Research Contract N00014-17-1-2177. During my time at MIT, I was previously supported by a Dean of Science Fellowship and am currently supported by a Mathworks Fellowship, both of which have allowed me to place a greater emphasis on research, and without which this thesis, in its current form, would not have been possible.

For John and Venita

Contents

Chapter 1

Introduction

In this thesis, we consider a number of different mathematical topics in linear algebra, with a focus on determinantal point processes and eigenvalue problems. Below we provide a brief non-technical summary of each topic in this thesis, as well as a short summary of other interesting projects that I had the pleasure of working on during my PhD that do not appear in this thesis.

Determinantal Point Processes and Principal Minor Assignment Problems

A determinantal point process is a probability distribution over the subsets of a finite ground set where the distribution is characterized by the principal minors of some fixed matrix. Given a finite set, say, $[N] = \{1, \ldots, N\}$ where N is a positive integer, a DPP is a random subset $Y \subseteq [N]$ such that $P(J \subseteq Y) = \det(K_J)$, for all fixed $J \subseteq [N]$, where $K \in \mathbb{R}^{N \times N}$ is a given matrix called the kernel of the DPP and $K_J = (K_{i,j})_{i,j \in J}$ is the principal submatrix of K associated with the set J. In Chapter 2, we focus on DPPs with kernels that have two different types of structure. First, we restrict ourselves to DPPs with symmetric kernels, and produce an algorithm that learns the kernel of a DPP from its samples. Then, we consider the slightly broader class of DPPs with kernels that have corresponding off-diagonal entries equal in magnitude, i.e., K satisfies $|K_{i,j}| = |K_{j,i}|$ for all $i, j \in [N]$, and produce an algorithm to recover such a matrix K from (possibly perturbed versions of) its principal minors. Sections 2.2 and 2.3 are written so that they may be read

independently of each other. This chapter is joint work with Victor-Emmanuel Brunel, Michel Goemans, Ankur Moitra, and Philippe Rigollet, and based on [119, 15].

The Spread and Bipartite Spread Conjecture

Given a graph $G = ([n], E)$, $[n] = \{1, ..., n\}$, the adjacency matrix $A \in \mathbb{R}^{n \times n}$ of G is defined as the matrix with $A_{i,j} = 1$ if $\{i, j\} \in E$ and $A_{i,j} = 0$ otherwise. A is a real symmetric matrix, and so has real eigenvalues $\lambda_1 \geq ... \geq \lambda_n$. One interesting quantity is the difference between extreme eigenvalues $\lambda_1 - \lambda_n$, commonly referred to as the spread of the graph G. In [48], the authors proposed two conjectures. First, the authors conjectured that if a graph G of size $|E| \leq n^2/4$ maximizes the spread over all graphs of order n and size $|E|$, then the graph is bipartite. The authors also conjectured that if a graph G maximizes the spread over all graphs of order n, then the graph is the join (i.e., with all edges in between) $K_{\lfloor 2n/3 \rfloor} \vee en\bar{K}_{\lceil n/3 \rceil}$ of a clique of order $\lfloor 2n/3 \rfloor$ and an independent set of order $\lceil n/3 \rceil$. We refer to these conjectures as the bipartite spread and spread conjectures, respectively. In Chapter 3, we consider both of these conjectures. We provide an infinite class of counterexamples to the bipartite spread conjecture, and prove an asymptotic version of the bipartite spread conjecture that is as tight as possible. In addition, for the spread conjecture, we make a number of observations regarding any spread-optimal graph, and use these observations to sketch a proof of the spread conjecture for all n sufficiently large. The full proof of the spread conjecture for all n sufficiently large is rather long and technical, and can be found in [12]. This chapter is joint work with Jane Breen, Alex Riasanovsky, and Michael Tait, and based on [12].

Force-Directed Layouts

A force-directed layout, broadly speaking, is a technique for drawing a graph in a low-dimensional Euclidean space (usually dimension ≤ 3) by applying "forces" between the set of vertices and/or edges. In a force-directed layout, vertices connected by an edge (or at a small graph distance from each other) tend to be close to each other in the resulting layout. Two well-known examples of a force-directed layout are Tutte's spring embedding

and metric multidimensional scaling. In his 1963 work titled "How to Draw a Graph," Tutte found an elegant technique to produce planar embeddings of planar graphs that also minimize the sum of squared edge lengths in some sense. In particular, for a three-connected planar graph, he showed that if the outer face of the graph is fixed as the complement of some convex region in the plane, and every other point is located at the mass center of its neighbors, then the resulting embedding is planar. This result is now known as Tutte's spring embedding theorem, and is considered by many to be the first example of a force-directed layout [117]. One of the major questions that this result does not treat is how to best embed the outer face. In Chapter 4, we investigate this question, consider connections to a Schur complement, and provide some theoretical results for this Schur complement using a discrete energy-only trace theorem. In addition, we also consider the later proposed Kamada-Kawai objective, which can provide a force-directed layout of an arbitrary graph. In particular, we prove a number of structural results regarding layouts that minimize this objective, provide algorithmic lower bounds for the optimization problem, and propose a polynomial time approximation scheme for drawing low diameter graphs. Sections 4.2 and 4.3 are written so that they may be read independently of each other. This chapter is joint work with Erik Demaine, Adam Hesterberg, Fred Koehler, Jayson Lynch, and Ludmil Zikatanov, and is based on [124, 31].

Error Estimates for the Lanczos Method

The Lanczos method is one of the most powerful and fundamental techniques for solving an extremal symmetric eigenvalue problem. Convergence-based error estimates depend heavily on the eigenvalue gap, but in practice, this gap is often relatively small, resulting in significant overestimates of error. One way to avoid this issue is through the use of uniform error estimates, namely, bounds that depend only on the dimension of the matrix and the number of iterations. In Chapter 5, we prove upper uniform error estimates for the Lanczos method, and provide a number of lower bounds through the use of orthogonal polynomials. In addition, we prove more specific results for matrices that possess some level of eigenvalue regularity or whose eigenvalues converge to some limiting empirical spectral

distribution. Through numerical experiments, we show that the theoretical estimates of this chapter do apply to practical computations for reasonably sized matrices. This chapter has no collaborators, and is based on [122].

Topics not in this Thesis

During my PhD, I have had the chance to work on a number of interesting topics that do not appear in this thesis. Below I briefly describe some of these projects. In [52] (joint with X. Hu and L. Zikatanov), we consider graph disaggregation, a technique to break high degree nodes of a graph into multiple smaller degree nodes, and prove a number of results regarding the spectral approximation of a graph by a disaggregated version of it. In [17] (joint with V.E. Brunel, A. Moitra, and P. Rigollet), we analyze the curvature of the expected log-likelihood around its maximum and characterize of when the maximum likelihood estimator converges at a parametric rate. In [120], we study centroidal Voronoi tessellations from a variational perspective and show that, for any density function, there does not exist a unique two generator centroidal Voronoi tessellation for dimensions greater than one. In [121], we prove a generalization of Courant's theorem for discrete graphs, namely, that for the k^{th} eigenvalue of a generalized Laplacian of a discrete graph, there exists a set of corresponding eigenvectors such that each eigenvector can be decomposed into at most k nodal domains, and that this set is of co-dimension zero with respect to the entire eigenspace. In [123] (joint with J. Wellens), we show that, given a graph with local crossing number either at most k or at least $2k$, it is NP-complete to decide whether the local crossing number is at most k or at least $2k$. In [97] (joint with D. Rohatgi and J. Wellens), we treat two conjectures – one regarding the maximum possible value of $cp(G) + cp(\bar{G})$ (where $cp(G)$ is the minimum number of cliques needed to cover the edges of G exactly once), due to de Caen, Erdos, Pullman and Wormald, and the other regarding $bp_k(K_n)$ (where $bp_k(G)$ is the minimum number of bicliques needed to cover each edge of G exactly k times), due to de Caen, Gregory and Pritikin. We disprove the first, obtaining improved lower and upper bounds on $max_G\, cp(G) + cp(\bar{G})$, and we prove an asymptotic version of the second, showing that $bp_k(K_n) = (1 + o(1))n$. If any of the topics (very briefly)

described here interest you, I encourage you to take a look at the associated paper. This thesis was constructed to center around a topic and to be of a manageable length, and I am no less proud of some of the results described here that do not appear in this thesis.

Chapter 2

Determinantal Point Processes and Principal Minor Assignment Problems

2.1 Introduction

A determinantal point process (DPP) is a probability distribution over the subsets of a finite ground set where the distribution is characterized by the principal minors of some fixed matrix. DPPs emerge naturally in many probabilistic setups such as random matrices and integrable systems [10]. They also have attracted interest in machine learning in the past years for their ability to model random choices while being tractable and mathematically elegant [71]. Given a finite set, say, $[N] = \{1, \ldots, N\}$ where N is a positive integer, a DPP is a random subset $Y \subseteq [N]$ such that $P(J \subseteq Y) = \det(K_J)$, for all fixed $J \subseteq [N]$, where $K \in \mathbb{R}^{N \times N}$ is a given matrix called the kernel of the DPP and $K_J = (K_{i,j})_{i,j \in J}$ is the principal submatrix of K associated with the set J. Assumptions on K that yield the existence of a DPP can be found in [14] and it is easily seen that uniqueness of the DPP is then automatically guaranteed. For instance, if $I - K$ is invertible (I being the identity matrix), then K is the kernel of a DPP if and only if $K(I - K)^{-1}$ is a P_0-matrix, i.e., all its principal minors are nonnegative. We refer to [54] for further definitions and properties of P_0 matrices. Note that in that case, the DPP is also an L-ensemble, with probability mass function given by $P(Y = J) = \det(L_J)/\det(I + L)$, for all $J \subseteq [N]$,

where $L = K(I - K)^{-1}$.

DPPs with a symmetric kernel have attracted a lot of interest in machine learning because they satisfy a property called negative association, which models repulsive interactions between items [8]. Following the seminal work of Kulesza and Taskar [72], discrete symmetric DPPs have found numerous applications in machine learning, including in document and timeline summarization [76, 132], image search [70, 1] and segmentation [73], audio signal processing [131], bioinformatics [6] and neuroscience [106]. What makes such models appealing is that they exhibit repulsive behavior and lend themselves naturally to tasks where returning a diverse set of objects is important. For instance, when applied to recommender systems, DPPs enforce diversity in the items within one basket [45].

From a statistical point of view, DPPs raise two essential questions. First, what are the families of kernels that give rise to one and the same DPP? Second, given observations of independent copies of a DPP, how to recover its kernel (which, as foreseen by the first question, is not necessarily unique)? These two questions are directly related to the *principal minor assignment problem* (PMA). Given a class of matrices, (1) describe the set of all matrices in this class that have a prescribed list of principal minors (this set may be empty), (2) find one such matrix. While the first task is theoretical in nature, the second one is algorithmic and should be solved using as few queries of the prescribed principal minors as possible. In this work, we focus on the question of recovering a kernel from observations, which is closely related to the problem of recovering a matrix with given principal minors. In this chapter, we focus on symmetric DPPs (which, when clear from context, we simply refer to as DPPs) and signed DPPs, a slightly larger class of kernels that only require corresponding off-diagonal entries be symmetric in magnitude.

2.1.1 Symmetric Matrices

In the symmetric case, there are fast algorithms for sampling (or approximately sampling) from a DPP [32, 94, 74, 75]. Marginalizing the distribution on a subset $I \subseteq [N]$ and conditioning on the event that $J \subseteq Y$ both result in new DPPs and closed form expressions for their kernels are known [11]. All of this work pertains to how to use a DPP once we

have learned its parameters. However, there has been much less work on the problem of learning the parameters of a symmetric DPP. A variety of heuristics have been proposed, including Expectation-Maximization [46], MCMC [1], and fixed point algorithms [80]. All of these attempt to solve a non-convex optimization problem, and no guarantees on their statistical performance are known. Recently, Brunel *et al.* [16] studied the rate of estimation achieved by the maximum likelihood estimator, but the question of efficient computation remains open. Apart from positive results on sampling, marginalization and conditioning, most provable results about DPPs are actually negative. It is conjectured that the maximum likelihood estimator is NP-hard to compute [69]. Actually, approximating the mode of size k of a DPP to within a c^k factor is known to be NP-hard for some $c > 1$ [22, 108]. The best known algorithms currently obtain a $e^k + o(k)$ approximation factor [88, 89].

In Section 2.2, we bypass the difficulties associated with maximum likelihood estimation by using the method of moments to achieve optimal sample complexity. In the setting of DPPs, the method of moments has a close connection to a principal minor assignment problem, which asks, given some set of principal minors of a matrix, to recover the original matrix, up to some equivalence class. We introduce a parameter ℓ, called the cycle sparsity of the graph induced by the kernel K, which governs the number of moments that need to be considered and, thus, the sample complexity. The cycle sparsity of a graph is the smallest integer ℓ so that the cycles of length at most ℓ yield a basis for the cycle space of the graph. We use a refined version of Horton's algorithm [51, 2] to implement the method of moments in polynomial time. Even though there are in general exponentially many cycles in a graph to consider, Horton's algorithm constructs a minimum weight cycle basis and, in doing so, also reveals the parameter ℓ together with a collection of at most ℓ induced cycles spanning the cycle space. We use such cycles in order to construct our method of moments estimator. For any fixed $\ell \geq 2$ and kernel K satisfying either $|K_{i,j}| \geq \alpha$ or $K_{i,j} = 0$ for all $i, j \in [N]$, our algorithm has sample complexity

$$n = O\left(\left(\frac{C}{\alpha}\right)^{2\ell} + \frac{\log N}{\alpha^2 \varepsilon^2}\right)$$

for some constant $C > 1$, runs in time polynomial in n and N, and learns the parameters up to an additive ε with high probability. The $(C/\alpha)^{2\ell}$ term corresponds to the number of samples needed to recover the signs of the entries in K. We complement this result with a minimax lower bound (Theorem 2) to show that this sample complexity is in fact near optimal. In particular, we show that there is an infinite family of graphs with cycle sparsity ℓ (namely length ℓ cycles) on which any algorithm requires at least $(C'\alpha)^{-2\ell}$ samples to recover the signs of the entries of K for some constant $C' > 1$. We also provide experimental results that confirm many quantitative aspects of our theoretical predictions. Together, our upper bounds, lower bounds, and experiments present a nuanced understanding of which symmetric DPPs can be learned provably and efficiently.

2.1.2 Magnitude-Symmetric Matrices

Recently, DPPs with non-symmetric kernels have gained interest in the machine learning community [14, 44, 3, 93]. However, for these classes, questions of recovery are significantly harder in general. In [14], Brunel proposes DPPs with kernels that are symmetric in magnitudes, i.e., $|K_{i,j}| = |K_{j,i}|$, for all $i, j = 1, \ldots, N$. This class is referred to as signed DPPs. A signed DPP is a generalization of DPPs which allows for both attractive and repulsive behavior. Such matrices are relevant in machine learning applications, because they increase the modeling power of DPPs. An essential assumption made in [14] is that K is dense, which simplifies the combinatorial analysis. We consider magnitude-symmetric matrices, but without any density assumption. The problem becomes significantly harder because it requires a fine analysis of the combinatorial properties of a graphical representation of the sparsity of the matrix and the products of corresponding off-diagonal entries. In Section 2.3, we treat both theoretical and algorithmic questions around recovering a magnitude-symmetric matrix from its principal minors. First, in Subsection 2.3.1, we show that, for a given magnitude-symmetric matrix K, the principal minors of length at most ℓ, for some graph invariant ℓ depending only on principal minors of order one and two, uniquely determine principal minors of all orders. Next, in Subsection 2.3.2, we describe an efficient algorithm that, given the principal minors of a magnitude-symmetric

matrix, computes a matrix with those principal minors. This algorithm queries only $O(n^2)$ principal minors, all of a bounded order that depends solely on the sparsity of the matrix. Finally, in Subsection 2.3.3, we consider the question of recovery when principal minors are only known approximately, and construct an algorithm that, for magnitude-symmetric matrices that are sufficiently generic in a sense, recovers the matrix almost exactly. This algorithm immediately implies an procedure for learning a signed DPP from its samples, as basic probabilistic techniques (e.g., a union bound, etc.) can be used to show that with high probability all estimators are sufficiently close to the principal minor they are approximating.

2.2 Learning Symmetric Determinantal Point Processes

Let $Y_1, \ldots, Y_n$ be n independent copies of $Y \sim \mathsf{DPP}(K)$, for some unknown kernel K such that $0 \preceq K \preceq I_N$. It is well known that K is identified by $\mathsf{DPP}(K)$ only up to flips of the signs of its rows and columns: If K' is another symmetric matrix with $0 \preceq K' \preceq I_N$, then $\mathsf{DPP}(K')$=$\mathsf{DPP}(K)$ if and only if $K' = DKD$ for some $D \in \mathcal{D}_N$, where $\mathcal{D}_N$ denotes the class of all $N \times N$ diagonal matrices with only 1 and -1 on their diagonals [69, Theorem 4.1]. We call such a transform a $\mathcal{D}_N$-similarity of K.

In view of this equivalence class, we define the following pseudo-distance between kernels K and K':

$$\rho(K, K') = \inf_{D \in \mathcal{D}_N} |DKD - K'|_\infty \,,$$

where for any matrix K, $|K|_\infty = \max_{i,j \in [N]} |K_{i,j}|$ denotes the entrywise sup-norm.

For any $S \subset [N]$, we write $\Delta_S = \det(K_S)$, where K_S denotes the $|S| \times |S|$ sub-matrix of K obtained by keeping rows and columns with indices in S. Note that for $1 \leq i \neq j \leq N$, we have the following relations:

$$K_{i,i} = \mathbb{P}[i \in Y], \qquad \Delta_{\{i,j\}} = \mathbb{P}[\{i,j\} \subseteq Y],$$

and $|K_{i,j}| = \sqrt{K_{i,i}K_{j,j} - \Delta_{\{i,j\}}}$. Therefore, the principal minors of size one and two of K determine K up to the sign of its off-diagonal entries. What remains is to compute the signs of the off-diagonal entries. In fact, for any K, there exists an ℓ depending only on the graph G_K induced by K, such that K can be recovered up to a $\mathcal{D}_N$-similarity with only the knowledge of its principal minors of size at most ℓ. We will show that this ℓ is exactly the cycle sparsity.

2.2.1 An Associated Principal Minor Assignment Problem

In this subsection, we consider the following related principal minor assignment problem:

> Given a symmetric matrix $K \in \mathbb{R}^{N \times N}$, what is the minimal ℓ such that K can be recovered (up to $\mathcal{D}_N$-similarity), using principal minors Δ_S, of size at most ℓ?

This problem has a clear relation to learning DPPs, as, in our setting, we can approximate the principal minors of K by empirical averages. However the accuracy of our estimator deteriorates with the size of the principal minor, and we must therefore estimate the smallest possible principal minors in order to achieve optimal sample complexity. Answering the above question tells us how small of principal minors we can consider. The relationship between the principal minors of K and recovery of $\mathsf{DPP}(K)$ has also been considered elsewhere. There has been work regarding the symmetric principal minor assignment problem, namely the problem of computing a matrix given an oracle that gives any principal minor in constant time [96]. Here, we prove that the smallest ℓ such that all the principal minors of K are uniquely determined by those of size at most ℓ is exactly the cycle sparsity of the graph induced by K.

We begin by recalling some standard graph theoretic notions. Let $G = ([N], E)$, $|E| = m$. A *cycle* C of G is any connected subgraph in which each vertex has even degree. Each cycle C is associated with an incidence vector $x \in GF(2)^m$ such that $x_e = 1$ if e is an edge in C and $x_e = 0$ otherwise. The *cycle space* $\mathcal{C}$ of G is the subspace of $GF(2)^m$ spanned by the incidence vectors of the cycles in G. The dimension ν_G of the cycle space is called the *cyclomatic number*, and it is well known that $\nu_G := m - N + \kappa(G)$, where $\kappa(G)$ denotes the number of connected components of G. Recall that a simple cycle is a graph where every vertex has either degree two or zero and the set of vertices with degree two form a connected set. A *cycle basis* is a basis of $\mathcal{C} \subset GF(2)^m$ such that every element is a simple cycle. It is well known that every cycle space has a cycle basis of induced cycles.

Definition 1. *The cycle sparsity of a graph G is the minimal ℓ for which G admits a cycle basis of induced cycles of length at most ℓ, with the convention that $\ell = 2$ whenever the cycle space is empty. A corresponding cycle basis is called a shortest maximal cycle basis.*

A shortest maximal cycle basis of the cycle space was also studied for other reasons by [21]. We defer a discussion of computing such a basis to a later subsection. For any subset $S \subseteq [N]$, denote by $G_K(S) = (S, E(S))$ the subgraph of G_K induced by S. A matching of $G_K(S)$ is a subset $M \subseteq E(S)$ such that any two distinct edges in M are not adjacent in $G(S)$. The set of vertices incident to some edge in M is denoted by $V(M)$. We denote by

$\mathcal{M}(S)$ the collection of all matchings of $G_K(S)$. Then, if $G_K(S)$ is an induced cycle, we can write the principal minor $\Delta_S = \det(K_S)$ as follows:

$$\Delta_S = \sum_{M \in \mathcal{M}(S)} (-1)^{|M|} \prod_{\{i,j\} \in M} K_{i,j}^2 \prod_{i \notin V(M)} K_{i,i} + 2 \times (-1)^{|S|+1} \prod_{\{i,j\} \in E(S)} K_{i,j}. \tag{2.1}$$

Proposition 1. *Let $K \in \mathbb{R}^{N \times N}$ be a symmetric matrix, G_K be the graph induced by K, and $\ell \geq 3$ be some integer. The kernel K is completely determined up to $\mathcal{D}_N$-similarity by its principal minors of size at most ℓ if and only if the cycle sparsity of G_K is at most ℓ.*

Proof. Note first that all the principal minors of K completely determine K up to a $\mathcal{D}_N$-similarity [96, Theorem 3.14]. Moreover, recall that principal minors of degree at most 2 determine the diagonal entries of K as well as the magnitude of its off-diagonal entries. In particular, given these principal minors, one only needs to recover the signs of the off-diagonal entries of K. Let the sign of cycle C in K be the product of the signs of the entries of K corresponding to the edges of C.

Suppose G_K has cycle sparsity ℓ and let $(C_1, \ldots, C_\nu)$ be a cycle basis of G_K where each $C_i, i \in [\nu]$ is an induced cycle of length at most ℓ. By (2.1), the sign of any $C_i, i \in [\nu]$ is completely determined by the principal minor Δ_S, where S is the set of vertices of C_i and is such that $|S| \leq \ell$. Moreover, for $i \in [\nu]$, let $x_i \in GF(2)^m$ denote the incidence vector of C_i. By definition, the incidence vector x of any cycle C is given by $\sum_{i \in \mathcal{I}} x_i$ for some subset $\mathcal{I} \subset [\nu]$. The sign of C is then given by the product of the signs of $C_i, i \in \mathcal{I}$ and thus by corresponding principal minors. In particular, the signs of all cycles are determined by the principal minors Δ_S with $|S| \leq \ell$. In turn, by Theorem 3.12 in [96], the signs of all cycles completely determine K, up to a $\mathcal{D}_N$-similarity.

Next, suppose the cycle sparsity of G_K is at least $\ell + 1$, and let $\mathcal{C}_\ell$ be the subspace of $GF(2)^m$ spanned by the induced cycles of length at most ℓ in G_K. Let $x_1, \ldots, x_\nu$ be a basis of $\mathcal{C}_\ell$ made of the incidence column vectors of induced cycles of length at most ℓ in G_K and form the matrix $A \in GF(2)^{m \times \nu}$ by concatenating the x_i's. Since $\mathcal{C}_\ell$ does not span the cycle space of G_K, $\nu < \nu_{G_K} \leq m$. Hence, the rank of A is less than m, so the null space of $A^\top$ is non trivial. Let $\bar{x}$ be the incidence column vector of an induced cycle $\bar{C}$ that is not in $\mathcal{C}_\ell$, and

let $h \in GL(2)^m$ with $A^\top h = 0$, $h \neq 0$ and $\bar{x}^\top h = 1$. These three conditions are compatible because $\bar{C} \notin \mathcal{C}_\ell$. We are now in a position to define an alternate kernel K' as follows: Let $K'_{i,i} = K_{i,i}$ and $|K'_{i,j}| = |K_{i,j}|$ for all $i, j \in [N]$. We define the signs of the off-diagonal entries of K' as follows: For all edges $e = \{i, j\}, i \neq j$, $\operatorname{sgn}(K'_e) = \operatorname{sgn}(K_e)$ if $h_e = 0$ and $\operatorname{sgn}(K'_e) = -\operatorname{sgn}(K_e)$ otherwise. We now check that K and K' have the same principal minors of size at most ℓ but differ on a principal minor of size larger than ℓ. To that end, let x be the incidence vector of a cycle C in $\mathcal{C}_\ell$ so that $x = Aw$ for some $w \in GL(2)^\nu$. Thus the sign of C in K is given by

$$\prod_{e\,:\,x_e=1} K_e = (-1)^{x^\top h} \prod_{e\,:\,x_e=1} K'_e = (-1)^{w^\top A^\top h} \prod_{e\,:\,x_e=1} K'_e = \prod_{e\,:\,x_e=1} K'_e$$

because $A^\top h = 0$. Therefore, the sign of any $C \in \mathcal{C}_\ell$ is the same in K and K'. Now, let $S \subseteq [N]$ with $|S| \leq \ell$, and let $G = G_{K_S} = G_{K'_S}$ be the graph corresponding to K_S (or, equivalently, to K'_S). For any induced cycle C in G, C is also an induced cycle in G_K and its length is at most ℓ. Hence, $C \in \mathcal{C}_\ell$ and the sign of C is the same in K and K'. By [96, Theorem 3.12], $\det(K_S) = \det(K'_S)$. Next observe that the sign of $\bar{C}$ in K is given by

$$\prod_{e\,:\,\bar{x}_e=1} K_e = (-1)^{\bar{x}^\top h} \prod_{e\,:\,\bar{x}_e=1} K'_e = -\prod_{e\,:\,x_e=1} K'_e.$$

Note also that since $\bar{C}$ is an induced cycle of $G_K = G_{K'}$, the above quantity is nonzero. Let $\bar{S}$ be the set of vertices in $\bar{C}$. By (2.1) and the above display, we have $\det(K_{\bar{S}}) \neq \det(K'_{\bar{S}})$. Together with [96, Theorem 3.14], it yields $K \neq DK'D$ for all $D \in \mathcal{D}_N$.

□

2.2.2 Definition of the Estimator

Our procedure is based on the previous result and can be summarized as follows. We first estimate the diagonal entries (i.e., the principal minors of size one) of K by the method of moments. By the same method, we estimate the principal minors of size two of K, and we deduce estimates of the magnitude of the off-diagonal entries. To use these estimates to

deduce an estimate $\hat{G}$ of G_K, we make the following assumption on the kernel K.

Assumption 1. *Fix $\alpha \in (0,1)$. For all $1 \leq i < j \leq N$, either $K_{i,j} = 0$, or $|K_{i,j}| \geq \alpha$.*

Finally, we find a shortest maximal cycle basis of $\hat{G}$, and we set the signs of our non-zero off-diagonal entry estimates by using estimators of the principal minors induced by the elements of the basis, again obtained by the method of moments.

For $S \subseteq [N]$, set $\hat{\Delta}_S = \dfrac{1}{n}\sum_{p=1}^{n} \mathbb{1}_{S \subseteq Y_p}$, and define

$$\hat{K}_{i,i} = \hat{\Delta}_{\{i\}} \quad \text{and} \quad \hat{B}_{i,j} = \hat{K}_{i,i}\hat{K}_{j,j} - \hat{\Delta}_{\{i,j\}},$$

where $\hat{K}_{i,i}$ and $\hat{B}_{i,j}$ are our estimators of $K_{i,i}$ and $K_{i,j}^2$, respectively. Define $\hat{G} = ([N], \hat{E})$, where, for $i \neq j$, $\{i,j\} \in \hat{E}$ if and only if $\hat{B}_{i,j} \geq \frac{1}{2}\alpha^2$. The graph $\hat{G}$ is our estimator of G_K. Let $\{\hat{C}_1, ..., \hat{C}_{\nu_{\hat{G}}}\}$ be a shortest maximal cycle basis of the cycle space of $\hat{G}$. Let $\hat{S}_i \subseteq [N]$ be the subset of vertices of $\hat{C}_i$, for $1 \leq i \leq \nu_{\hat{G}}$. We define

$$\hat{H}_i = \hat{\Delta}_{\hat{S}_i} - \sum_{M \in \mathcal{M}(\hat{S}_i)} (-1)^{|M|} \prod_{\{i,j\} \in M} \hat{B}_{i,j} \prod_{i \notin V(M)} \hat{K}_{i,i},$$

for $1 \leq i \leq \nu_{\hat{G}}$. In light of (2.1), for large enough n, this quantity should be close to

$$H_i = 2 \times (-1)^{|\hat{S}_i|+1} \prod_{\{i,j\} \in E(\hat{S}_i)} K_{i,j}\,.$$

We note that this definition is only symbolic in nature, and computing $\hat{H}_i$ in this fashion is extremely inefficient. Instead, to compute it in practice, we will use the determinant of an auxiliary matrix, computed via a matrix factorization. Namely, let us define the matrix $\widetilde{K} \in \mathbb{R}^{N \times N}$ such that $\widetilde{K}_{i,i} = \hat{K}_{i,i}$ for $1 \leq i \leq N$, and $\widetilde{K}_{i,j} = \hat{B}_{i,j}^{1/2}$. We have

$$\det \widetilde{K}_{\hat{S}_i} = \sum_{M \in \mathcal{M}} (-1)^{|M|} \prod_{\{i,j\} \in M} \hat{B}_{i,j} \prod_{i \notin V(M)} \hat{K}_{i,i} + 2 \times (-1)^{|\hat{S}_i|+1} \prod_{\{i,j\} \in \hat{E}(\hat{S}_i)} \hat{B}_{i,j}^{1/2},$$

so that we may equivalently write

$$\hat{H}_i = \hat{\Delta}_{\hat{S}_i} - \det(\widetilde{K}_{\hat{S}_i}) + 2 \times (-1)^{|\hat{S}_i|+1} \prod_{\{i,j\} \in \hat{E}(\hat{S}_i)} \hat{B}_{i,j}^{1/2}.$$

Finally, let $\hat{m} = |\hat{E}|$. Set the matrix $A \in GF(2)^{\nu_{\hat{G}} \times \hat{m}}$ with i-th row representing $\hat{C}_i$ in $GF(2)^m$, $1 \leq i \leq \nu_{\hat{G}}$, $b = (b_1, \ldots, b_{\nu_{\hat{G}}}) \in GF(2)^{\nu_{\hat{G}}}$ with $b_i = \frac{1}{2}[\mathrm{sgn}(\hat{H}_i) + 1]$, $1 \leq i \leq \nu_{\hat{G}}$, and let $x \in GF(2)^m$ be a solution to the linear system $Ax = b$ if a solution exists, $x = \mathbb{1}_m$ otherwise. We define $\hat{K}_{i,j} = 0$ if $\{i,j\} \notin \hat{E}$ and $\hat{K}_{i,j} = \hat{K}_{j,i} = (2x_{\{i,j\}} - 1)\hat{B}_{i,j}^{1/2}$ for all $\{i,j\} \in \hat{E}$.

Next, we prove the following lemma which relates the quality of estimation of K in terms of ρ to the quality of estimation of the principal minors Δ_S.

Lemma 1. *Let K satisfy Assumption 1, and let ℓ be the cycle sparsity of G_K. Let $\varepsilon > 0$. If $|\hat{\Delta}_S - \Delta_S| \leq \varepsilon$ for all $S \subseteq [N]$ with $|S| \leq 2$ and if $|\hat{\Delta}_S - \Delta_S| \leq (\alpha/4)^{|S|}$ for all $S \subseteq [N]$ with $3 \leq |S| \leq \ell$, then*

$$\rho(\hat{K}, K) < 4\varepsilon/\alpha\,.$$

Proof. We can bound $|\hat{B}_{i,j} - K_{i,j}^2|$, namely,

$$\begin{aligned}\hat{B}_{i,j} &\leq (K_{i,i} + \alpha^2/16)(K_{j,j} + \alpha^2/16) - (\Delta_{\{i,j\}} - \alpha^2/16) \\ &\leq K_{i,j}^2 + \alpha^2/4\end{aligned}$$

and

$$\begin{aligned}\hat{B}_{i,j} &\geq (K_{i,i} - \alpha^2/16)(K_{j,j} - \alpha^2/16) - (\Delta_{\{i,j\}} + \alpha^2/16) \\ &\geq K_{i,j}^2 - 3\alpha^2/16,\end{aligned}$$

giving $|\hat{B}_{i,j} - K_{i,j}^2| < \alpha^2/4$. Thus, we can correctly determine whether $K_{i,j} = 0$ or $|K_{i,j}| \geq \alpha$, yielding $\hat{G} = G_K$. In particular, the cycle basis $\hat{C}_1, \ldots, \hat{C}_{\nu_{\hat{G}}}$ of $\hat{G}$ is a cycle basis of G_K.

Let $1 \leq i \leq \nu_{\hat{G}}$. Denote by $t = (\alpha/4)^{|S_i|}$. We have

$$\begin{aligned}
\left|\hat{H}_i - H_i\right| &\leq |\hat{\Delta}_{\hat{S}_i} - \Delta_{\hat{S}_i}| + |\mathcal{M}(\hat{S}_i)| \max_{x \in \pm 1} \left[(1+4tx)^{|\hat{S}_i|} - 1\right] \\
&\leq (\alpha/4)^{|\hat{S}_i|} + |\mathcal{M}(\hat{S}_i)| \left[(1+4t)^{|\hat{S}_i|} - 1\right] \\
&\leq (\alpha/4)^{|\hat{S}_i|} + T\left(|\hat{S}_i|, \left\lfloor \frac{|\hat{S}_i|}{2} \right\rfloor\right) 4t \; T(|\hat{S}_i|, |\hat{S}_i|) \\
&\leq (\alpha/4)^{|\hat{S}_i|} + 4t \; (2^{\frac{|\hat{S}_i|}{2}} - 1)(2^{|\hat{S}_i|} - 1) \\
&\leq (\alpha/4)^{|\hat{S}_i|} + t 2^{2|\hat{S}_i|} \\
&< 2\alpha^{|\hat{S}_i|} \leq |H_i|,
\end{aligned}$$

where, for positive integers $p < q$, we denote by $T(q,p) = \sum_{i=1}^{p} \binom{q}{i}$. Therefore, we can determine the sign of the product $\prod_{\{i,j\} \in E(\hat{S}_i)} K_{i,j}$ for every element in the cycle basis and recover the signs of the non-zero off-diagonal entries of $K_{i,j}$. Hence,

$$\rho(\hat{K}, K) = \max_{1 \leq i,j \leq N} \left| |\hat{K}_{i,j}| - |K_{i,j}| \right|.$$

For $i = j$, $\left| |\hat{K}_{i,j}| - |K_{i,j}| \right| = |\hat{K}_{i,i} - K_{i,i}| \leq \varepsilon$. For $i \neq j$ with $\{i,j\} \in \hat{E} = E$, one can easily show that $\left| \hat{B}_{i,j} - K_{i,j}^2 \right| \leq 4\varepsilon$, yielding

$$|\hat{B}_{i,j}^{1/2} - |K_{i,j}|| \leq \frac{4\varepsilon}{\left|\hat{B}_{i,j}^{1/2} + |K_{i,j}|\right|} \leq \frac{4\varepsilon}{\alpha},$$

which completes the proof. □

We are now in a position to establish a sufficient sample size to estimate K within distance ε.

Theorem 1. *Let K satisfy Assumption 1, and let ℓ be the cycle sparsity of G_K. Let $\varepsilon > 0$. For any $A > 0$, there exists $C > 0$ such that*

$$n \geq C\Big(\frac{1}{\alpha^2 \varepsilon^2} + \ell\big(\frac{4}{\alpha}\big)^{2\ell}\Big) \log N \,,$$

yields $\rho(\hat{K}, K) \leq \varepsilon$ *with probability at least* $1 - N^{-A}$.

Proof. Using the previous lemma, and applying a union bound,

$$\begin{aligned}\mathbb{P}\left[\rho(\hat{K}, K) > \varepsilon\right] &\leq \sum_{|S|\leq 2} \mathbb{P}\left[|\hat{\Delta}_S - \Delta_S| > \alpha\varepsilon/4\right] + \sum_{2\leq|S|\leq\ell} \mathbb{P}\left[|\hat{\Delta}_S - \Delta_S| > (\alpha/4)^{|S|}\right] \\ &\leq 2N^2 e^{-n\alpha^2\varepsilon^2/8} + 2N^{\ell+1} e^{-2n(\alpha/4)^{2\ell}}, \end{aligned} \tag{2.2}$$

where we used Hoeffding's inequality. □

2.2.3 Information Theoretic Lower Bounds

We prove an information-theoretic lower bound that holds already if G_K is an ℓ-cycle. Let $D(K\|K')$ and $\mathbb{H}(K, K')$ denote respectively the Kullback-Leibler divergence and the Hellinger distance between $\mathsf{DPP}(K)$ and $\mathsf{DPP}(K')$.

Lemma 2. *For* $\eta \in \{-, +\}$*, let* K^η *be the* $\ell \times \ell$ *matrix with elements given by*

$$K_{i,j} = \begin{cases} 1/2 & \text{if } j = i \\ \alpha & \text{if } j = i \pm 1 \\ \eta\alpha & \text{if } (i,j) \in \{(1,\ell),(\ell,1)\} \\ 0 & \text{otherwise} \end{cases}.$$

Then, for any $\alpha \leq 1/8$*, it holds*

$$D(K\|K') \leq 4(6\alpha)^\ell, \qquad \text{and} \qquad \mathbb{H}(K, K') \leq (8\alpha^2)^\ell.$$

Proof. It is straightforward to see that

$$\det(K^+_J) - \det(K^-_J) = \begin{cases} 2\alpha^\ell & \text{if } J = [\ell] \\ 0 & \text{else} \end{cases}.$$

If Y is sampled from $\mathsf{DPP}(K^\eta)$, we denote by $p_\eta(S) = \mathbb{P}[Y = S]$, for $S \subseteq [\ell]$. It follows

from the inclusion-exclusion principle that for all $S \subseteq [\ell]$,

$$\begin{aligned} p_+(S) - p_-(S) &= \sum_{J \subseteq [\ell] \setminus S} (-1)^{|J|} (\det K^+_{S \cup J} - \det K^-_{S \cup J}) \\ &= (-1)^{\ell - |S|} (\det K^+ - \det K^-) = \pm 2\alpha^\ell \,, \end{aligned} \tag{2.3}$$

where $|J|$ denotes the cardinality of J. The inclusion-exclusion principle also yields that $p_\eta(S) = |\det(K^\eta - I_{\bar{S}})|$ for all $S \subseteq [l]$, where $I_{\bar{S}}$ stands for the $\ell \times \ell$ diagonal matrix with ones on its entries (i,i) for $i \notin S$, zeros elsewhere.

We denote by $D(K^+ \| K^-)$ the Kullback Leibler divergence between $\mathsf{DPP}(K^+)$ and $\mathsf{DPP}(K^-)$:

$$\begin{aligned} D(K^+ \| K^-) &= \sum_{S \subseteq [\ell]} p_+(S) \log \left(\frac{p_+(S)}{p_-(S)} \right) \\ &\leq \sum_{S \subseteq [\ell]} \frac{p_+(S)}{p_-(S)} (p_+(S) - p_-(S)) \\ &\leq 2\alpha^\ell \sum_{S \subseteq [\ell]} \frac{|\det(K^+ - I_{\bar{S}})|}{|\det(K^- - I_{\bar{S}})|}, \end{aligned} \tag{2.4}$$

by (2.3). Using the fact that $0 < \alpha \leq 1/8$ and the Gershgorin circle theorem, we conclude that the absolute value of all eigenvalues of $K^\eta - I_{\bar{S}}$ are between $1/4$ and $3/4$. Thus we obtain from (2.4) the bound $D(K^+ \| K^-) \leq 4(6\alpha)^\ell$.

Using the same arguments as above, the Hellinger distance $\mathbb{H}(K^+, K^-)$ between $\mathsf{DPP}(K^+)$ and $\mathsf{DPP}(K^-)$ satisfies

$$\begin{aligned} \mathbb{H}(K^+, K^-) &= \sum_{J \subseteq [\ell]} \left(\frac{p_+(J) - p_-(J)}{\sqrt{p_+(J)} + \sqrt{p_-(J)}} \right)^2 \\ &\leq \sum_{J \subseteq [\ell]} \frac{4\alpha^{2\ell}}{2 \cdot 4^{-\ell}} = (8\alpha^2)^\ell \end{aligned}$$

which completes the proof. □

The sample complexity lower bound now follows from standard arguments.

Theorem 2. *Let* $0 < \varepsilon \leq \alpha \leq 1/8$ *and* $3 \leq \ell \leq N$. *There exists a constant* $C > 0$ *such that if*

$$n \leq C\Big(\frac{8^\ell}{\alpha^{2\ell}} + \frac{\log(N/\ell)}{(6\alpha)^\ell} + \frac{\log N}{\varepsilon^2}\Big),$$

then the following holds: for any estimator $\hat{K}$ *based on* n *samples, there exists a kernel* K *that satisfies Assumption 1 and such that the cycle sparsity of* G_K *is* ℓ *and for which* $\rho(\hat{K}, K) \geq \varepsilon$ *with probability at least* $1/3$.

Proof. Recall the notation of Lemma 2. First consider the $N \times N$ block diagonal matrix K (resp. K') where its first block is K^+ (resp. K^-) and its second block is $I_{N-\ell}$. By a standard argument, the Hellinger distance $\mathbb{H}_n(K, K')$ between the product measures $\mathsf{DPP}(K)^{\otimes n}$ and $\mathsf{DPP}(K')^{\otimes n}$ satisfies

$$1 - \frac{\mathbb{H}_n^2(K, K')}{2} = \big(1 - \frac{\mathbb{H}^2(K, K')}{2}\big)^n \geq \big(1 - \frac{\alpha^{2\ell}}{2 \times 8^\ell}\big)^n,$$

which yields the first term in the desired lower bound.

Next, by padding with zeros, we can assume that $L = N/\ell$ is an integer. Let $K^{(0)}$ be a block diagonal matrix where each block is K^+ (using the notation of Lemma 2). For $j = 1, \ldots, L$, define the $N \times N$ block diagonal matrix $K^{(j)}$ as the matrix obtained from $K^{(0)}$ by replacing its jth block with K^- (again using the notation of Lemma 2).

Since $\mathsf{DPP}(K^{(j)})$ is the product measure of L lower dimensional DPPs that are each independent of each other, using Lemma 2 we have $D(K^{(j)} \| K^{(0)}) \leq 4(6\alpha)^\ell$. Hence, by Fano's lemma (see, e.g., Corollary 2.6 in [115]), the sample complexity to learn the kernel of a DPP within a distance $\varepsilon \leq \alpha$ is

$$\Omega\left(\frac{\log(N/\ell)}{(6\alpha)^\ell}\right)$$

which yields the second term.

The third term follows from considering $K_0 = (1/2)I_N$ and letting K_j be obtained from K_0 by adding ε to the jth entry along the diagonal. It is easy to see that $D(K_j \| K_0) \leq 8\varepsilon^2$. Hence, a second application of Fano's lemma yields that the sample complexity to learn the

kernel of a DPP within a distance ε is $\Omega(\frac{\log N}{\varepsilon^2})$. $\square$

The third term in the lower bound is the standard parametric term and is unavoidable in order to estimate the magnitude of the coefficients of K. The other terms are more interesting. They reveal that the cycle sparsity of G_K, namely, ℓ, plays a key role in the task of recovering the sign pattern of K. Moreover the theorem shows that the sample complexity of our method of moments estimator is near optimal.

2.2.4 Algorithmic Aspects and Experiments

We first give an algorithm to compute the estimator $\hat{K}$ defined in Subsection 2.2.2. A well-known algorithm of Horton [51] computes a cycle basis of minimum total length in time $O(m^3N)$. Subsequently, the running time was improved to $O(m^2N/\log N)$ time [2]. Also, it is known that a cycle basis of minimum total length is a shortest maximal cycle basis [21]. Together, these results imply the following.

Lemma 3. *Let $G = ([N], E)$, $|E| = m$. There is an algorithm to compute a shortest maximal cycle basis in $O(m^2N/\log N)$ time.*

In addition, we recall the following standard result regarding the complexity of Gaussian elimination [47].

Lemma 4. *Let $A \in GF(2)^{\nu\times m}$, $b \in GF(2)^{\nu}$. Then Gaussian elimination will find a vector $x \in GF(2)^m$ such that $Ax = b$ or conclude that none exists in $O(\nu^2 m)$ time.*

We give our procedure for computing the estimator $\hat{K}$ in Algorithm 1. In the following theorem, we bound the running time of Algorithm 1 and establish an upper bound on the sample complexity needed to solve the recovery problem as well as the sample complexity needed to compute an estimate $\hat{K}$ that is close to K.

Theorem 3. *Let $K \in \mathbb{R}^{N\times N}$ be a symmetric matrix satisfying $0 \preceq K \preceq I$, and satisfying Assumption 1. Let G_K be the graph induced by K and ℓ be the cycle sparsity of G_K. Let*

Algorithm 1 Compute Estimator $\hat{K}$

Input: samples $Y_1, ..., Y_n$, parameter $\alpha > 0$.

Compute $\hat{\Delta}_S$ for all $|S| \leq 2$.
Set $\hat{K}_{i,i} = \hat{\Delta}_{\{i\}}$ for $1 \leq i \leq N$.
Compute $\hat{B}_{i,j}$ for $1 \leq i < j \leq N$.
Form $\widetilde{K} \in \mathbb{R}^{N \times N}$ and $\hat{G} = ([N], \hat{E})$.
Compute a shortest maximal cycle basis $\{\hat{v}_1, ..., \hat{v}_{\nu_{\hat{G}}}\}$.
Compute $\hat{\Delta}_{\hat{S}_i}$ for $1 \leq i \leq \nu_{\hat{G}}$.
Compute $\hat{C}_{\hat{S}_i}$ using $\det \widetilde{K}_{\hat{S}_i}$ for $1 \leq i \leq \nu_{\hat{G}}$.
Construct $A \in GF(2)^{\nu_{\hat{G}} \times m}$, $b \in GF(2)^{\nu_{\hat{G}}}$.
Solve $Ax = b$ using Gaussian elimination.
Set $\hat{K}_{i,j} = \hat{K}_{j,i} = (2x_{\{i,j\}} - 1)\hat{B}_{i,j}^{1/2}$, for all $\{i, j\} \in \hat{E}$.

$Y_1, ..., Y_n$ be samples from DPP(K) and $\delta \in (0, 1)$. If

$$n > \frac{\log(N^{\ell+1}/\delta)}{(\alpha/4)^{2\ell}},$$

then with probability at least $1 - \delta$, Algorithm 1 computes an estimator $\hat{K}$ which recovers the signs of K up to a $\mathcal{D}_N$-similarity and satisfies

$$\rho(K, \hat{K}) < \frac{1}{\alpha} \left(\frac{8 \log(4N^{\ell+1}/\delta)}{n} \right)^{1/2} \tag{2.5}$$

in $O(m^3 + nN^2)$ time.

Proof. (2.5) follows directly from (2.2) in the proof of Theorem 1. That same proof also shows that with probability at least $1-\delta$, the support of G_K and the signs of K are recovered up to a $\mathcal{D}_N$-similarity. What remains is to upper bound the worst case run time of Algorithm 1. We will perform this analysis line by line. Initializing $\hat{K}$ requires $O(N^2)$ operations. Computing Δ_S for all subsets $|S| \leq 2$ requires $O(nN^2)$ operations. Setting $\hat{K}_{i,i}$ requires $O(N)$ operations. Computing $\hat{B}_{i,j}$ for $1 \leq i < j \leq N$ requires $O(N^2)$ operations. Forming $\widetilde{K}$ requires $O(N^2)$ operations. Forming G_K requires $O(N^2)$ operations. By Lemma 3, computing a shortest maximal cycle basis requires $O(mN)$ operations. Constructing the subsets S_i, $1 \leq i \leq \nu_{\hat{G}}$, requires $O(mN)$ operations. Computing $\hat{\Delta}_{S_i}$ for $1 \leq i \leq \nu_{\hat{G}}$

requires $O(nm)$ operations. Computing $\hat{C}_{S_i}$ using $\det(\widetilde{K}[S_i])$ for $1 \leq i \leq \nu_{\hat{G}}$ requires $O(m\ell^3)$ operations, where a factorization of each $\widetilde{K}[S_i]$ is used to compute each determinant in $O(\ell^3)$ operations. Constructing A and b requires $O(m\ell)$ operations. By Lemma 4, finding a solution x using Gaussian elimination requires $O(m^3)$ operations. Setting $\hat{K}_{i,j}$ for all edges $\{i,j\} \in E$ requires $O(m)$ operations. Put this all together, Algorithm 1 runs in $O(m^3 + nN^2)$ time. □

Chordal Graphs

Here we show that it is possible to obtain faster algorithms by exploiting the structure of G_K. Specifically, in the case where G_K chordal, we give an $O(m)$ time algorithm to determine the signs of the off-diagonal entries of the estimator $\hat{K}$, resulting in an improved overall runtime of $O(m + nN^2)$. Recall that a graph $G = ([N], E)$ is said to be chordal if every induced cycle in G is of length three. Moreover, a graph $G = ([N], E)$ has a perfect elimination ordering (PEO) if there exists an ordering of the vertex set $\{v_1, ..., v_N\}$ such that, for all i, the graph induced by $\{v_i\} \cup \{v_j | \{i,j\} \in E, j > i\}$ is a clique. It is well known that a graph is chordal if and only if it has a PEO. A PEO of a chordal graph with m edges can be computed in $O(m)$ operations using lexicographic breadth-first search [98].

Lemma 5. *Let $G = ([N], E)$, be a chordal graph and $\{v_1, ..., v_n\}$ be a PEO. Given i, let $i^* := \min\{j | j > i, \{v_i, v_j\} \in E\}$. Then the graph $G' = ([N], E')$, where $E' = \{\{v_i, v_{i^*}\}\}_{i=1}^{N-\kappa(G)}$, is a spanning forest of G.*

Proof. We first show that there are no cycles in G'. Suppose to the contrary, that there is an induced cycle C of length k on the vertices $\{v_{j_1}, ..., v_{j_k}\}$. Let v be the vertex of smallest index. Then v is connected to two other vertices in the cycle of larger index. This is a contradiction to the construction.

What remains is to show that $|E'| = N - \kappa(G)$. It suffices to prove the case $\kappa(G) = 1$. Suppose to the contrary, that there exists a vertex v_i, $i < N$, with no neighbors of larger index. Let P be the shortest path in G from v_i to v_N. By connectivity, such a path exists. Let v_k be the vertex of smallest index in the path. However, it has two neighbors in the path of larger index, which must be adjacent to each other. Therefore, there is a shorter path. □

Algorithm 2 Compute Signs of Edges in Chordal Graph

Input: $G_K = ([N], E)$ chordal, $\hat{\Delta}_S$ for $|S| \leq 3$.

Compute a perfect elimination ordering $\{v_1, ..., v_N\}$.
Compute the spanning forest $G' = ([N], E')$.
Set all edges in E' to have positive sign.
Compute $\hat{C}_{\{i,j,i^*\}}$ for all $\{i, j\} \in E \setminus E'$, $j < i$.
Order edges $E \setminus E' = \{e_1, ..., e_\nu\}$ such that $i > j$ if $\max e_i < \max e_j$.
Visit edges in sorted order and for $e = \{i, j\}$, $j > i$, set

$$\operatorname{sgn}(\{i, j\}) = \operatorname{sgn}(\hat{C}_{\{i,j,i^*\}}) \operatorname{sgn}(\{i, i^*\}) \operatorname{sgn}(\{j, i^*\}).$$

Now, given the chordal graph G_K induced by K and the estimates of principal minors of size at most three, we provide an algorithm to determine the signs of the edges of G_K, or, equivalently, the off-diagonal entries of K.

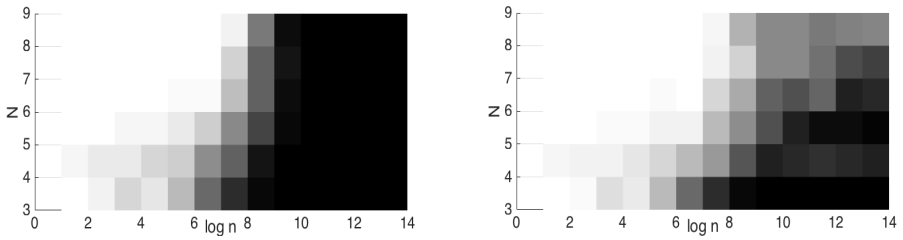
Theorem 4. *If G_K is chordal, Algorithm 2 correctly determines the signs of the edges of G_K in $O(m)$ time.*

Proof. We will simultaneously perform a count of the operations and a proof of the correctness of the algorithm. Computing a PEO requires $O(m)$ operations. Computing the spanning forest requires $O(m)$ operations. The edges of the spanning tree can be given arbitrary sign, because it is a cycle-free graph. This assigns a sign to two edges of each 3-cycle. Computing each $\hat{C}_{\{i,j,i^*\}}$ requires a constant number of operations because $\ell = 3$, requiring a total of $O(m - N)$ operations. Ordering the edges requires $O(m)$ operations. Setting the signs of each remaining edge requires $O(m)$ operations. □

Therefore, when G_K is chordal, the overall complexity required by our algorithm to compute $\hat{K}$ is reduced to $O(m + nN^2)$.

Experiments

Here we present experiments to supplement the theoretical results of the section. We test our algorithm on two types of random matrices. First, we consider the matrix $K \in \mathbb{R}^{N \times N}$



(a) graph recovery, cycle

(b) graph and sign recovery, cycle

(c) graph recovery, clique

(d) graph and sign recovery, clique

Figure 2-1: Plots of the proportion of successive graph recovery, and graph and sign recovery, for random matrices with cycle and clique graph structure, respectively. The darker the box, the higher the proportion of trials that were recovered successfully.

corresponding to the cycle on N vertices,

$$K = \frac{1}{2}I + \frac{1}{4}A,$$

where A is symmetric and has non-zero entries only on the edges of the cycle, either $+1$ or -1, each with probability $1/2$. By the Gershgorin circle theorem, $0 \preceq K \preceq I$. Next, we consider the matrix $K \in \mathbb{R}^{N\times N}$ corresponding to the clique on N vertices,

$$K = \frac{1}{2}I + \frac{1}{4\sqrt{N}}A,$$

where A is symmetric and has all entries either $+1$ or -1, each with probability $1/2$. It is well known that $-2\sqrt{N} \preceq A \preceq 2\sqrt{N}$ with high probability, implying $0 \preceq K \preceq I$.

For both cases and for a range of values of matrix dimension N and samples n, we run our algorithm on 50 randomly generated instances. We record the proportion of trials where we recover the graph induced by K, and the proportion of the trials where we recover

both the graph and correctly determine the signs of the entries. In Figure 2-1, the shade of each box represents the proportion of trials where recovery was successful for a given pair N, n. A completely white box corresponds to zero success rate, black to a perfect success rate. The plots corresponding to the cycle and the clique are telling. We note that for the clique, recovering the sparsity pattern and recovering the signs of the off-diagonal entries come hand-in-hand. However, for the cycle, there is a noticeable gap between the number of samples required to recover the sparsity pattern and the number of samples required to recover the signs of the off-diagonal entries. This empirically confirms the central role that cycle sparsity plays in parameter estimation, and further corroborates our theoretical results.

2.3 Recovering a Magnitude-Symmetric Matrix from its Principal Minors

In this section, we consider matrices $K \in \mathbb{R}^{N\times N}$ satisfying $|K_{i,j}| = |K_{j,i}|$ for $i, j \in [N]$, which we refer to as magnitude-symmetric matrices, and investigate the algorithmic question of recovering such a matrix from its principal minors. First, we require a number of key definitions and notation. For completeness (and to allow independent reading), we are including terms and notations that we have defined in Section 2.2. If $K \in \mathbb{R}^{N\times N}$ and $S \subseteq [N]$, $K_S := (K_{i,j})_{i,j\in S}$ and $\Delta_S(K) := \det K_S$ is the principal minor of K associated with the set S ($\Delta_\emptyset(K) = 1$ by convention). When it is clear from the context, we simply write Δ_S instead of $\Delta_S(K)$, and for sets of order at most four, we replace the set itself by its elements, i.e., write $\Delta_{\{1,2,3\}}$ as $\Delta_{1,2,3}$.

In addition, we recall a number of relevant graph-theoretic definitions. In this work, all graphs G are simple, undirected graphs. An *articulation point* or *cut vertex* of a graph G is a vertex whose removal disconnects the graph. A graph G is said to be *two-connected* if it has at least two vertices and has no articulation points. A maximal two-connected subgraph H of G is called a *block* of G. Given a graph G, a *cycle* C is a subgraph of G in which every vertex of C has even degree (within C). A *simple cycle* C is a connected subgraph of G in which every vertex of C has degree two. For simplicity, we may sometimes describe C by a traversal of its vertices along its edges, i.e., $C = i_1\ i_2\ ...\ i_k\ i_1$; notice we have the choice of the start vertex and the orientation of the cycle in this description. We denote the subgraph induced by a vertex subset $S \subset V$ by $G[S]$. A simple cycle C of a graph G on vertex set $V(C)$ is not necessarily induced, and while the cycle itself has all vertices of degree two, this cycle may contain some number of chords in the original graph G, and we denote this set $E(G[S])\backslash E(C)$ of chords by $\gamma(C)$.

Given any subgraph H of G, we can associate with H an incidence vector $\chi_H \in GF(2)^m$, $m = |E|$, where $\chi_H(e) = 1$ if and only if $e \in H$, and $\chi_H(e) = 0$ otherwise. Given two subgraphs $H_1, H_2 \subset G$ of G, we define their *sum* $H_1 + H_2$ as the graph containing all edges in exactly one of $E(H_1)$ and $E(H_2)$ (i.e., their symmetric difference) and no isolated

vertices. This corresponds to the graph resulting from the sum of their incidence vectors. The cycle space of G is given by

$$\mathcal{C}(G) = \text{span}\{\chi_C \,|\, C \text{ is a cycle of } G\} \subset GF(2)^m,$$

and has dimension $\nu = m - N + \kappa(G)$, where $\kappa(G)$ denotes the number of connected components of G. The quantity ν is commonly referred to as the *cyclomatic number*. The *cycle sparsity* ℓ of the graph G is the smallest number for which the set of incidence vectors χ_C of cycles of edge length at most ℓ spans $\mathcal{C}(G)$.

In this section, we require the use and analysis of graphs $G = ([N], E)$ endowed with a linear Boolean function ϵ that maps subgraphs of G to $\{-1, +1\}$, i.e., $\epsilon(e) \in \{-1, +1\}$ for all $e \in E(G)$ and

$$\epsilon(H) = \prod_{e \in E(H)} \epsilon(e)$$

for all subgraphs H. The graph G combined with a linear Boolean function ϵ is denoted by $G = ([N], E, \epsilon)$ and referred to as a *charged* graph. If $\epsilon(H) = +1$ (resp. $\epsilon(H) = -1$), then we say the subgraph H is positive (resp. negative). For the sake of space, we often denote $\epsilon(\{i, j\})$, $\{i, j\} \in E(G)$, by $\epsilon_{i,j}$. Given a magnitude-symmetric matrix $K \in \mathbb{R}^{N \times N}$, $|K_{i,j}| = |K_{j,i}|$ for $i, j \in [N]$, we define the charged sparsity graph G_K as $G_K = ([N], E, \epsilon)$, $E := \{i, j \in [N] \,|\, i \neq j, |K_{i,j}| \neq 0\}$, $\epsilon(\{i, j\}) := \text{sgn}(K_{i,j} K_{j,i})$, and when clear from context, we simply write G instead of G_K.

We define the span of the incidence vectors of positive cycles as

$$\mathcal{C}^+(G) = \text{span}\{\chi_C \,|\, C \text{ is a cycle of } G,\ \epsilon(C) = +1\}.$$

As we will see, when H is a block, the space $\mathcal{C}^+(H)$ is spanned by positive simple cycles (Proposition 3). We define the *simple cycle sparsity* ℓ_+ of $\mathcal{C}^+(G)$ to be the smallest number for which the set of incidence vectors χ_C of positive simple cycles of edge length at most ℓ_+ spans $\mathcal{C}^+(G)$ (or, equivalently, the set $\mathcal{C}^+(H)$ for every block H of G). Unlike $\mathcal{C}(G)$, for $\mathcal{C}^+(G)$ this quantity ℓ_+ depends on whether we consider a basis of cycles or of only simple

cycles. The study of bases of $\mathcal{C}^+(H)$, for blocks H, consisting of incidence vectors of simple cycles constitutes the major graph-theoretic subject of this work, and has connections to the recovery of a magnitude-symmetric matrix from its principal minors.

2.3.1 Principal Minors and Magnitude-Symmetric Matrices

In this subsection, we are primarily concerned with recovering a magnitude-symmetric matrix K from its principal minors. Using only principal minors of order one and two, the quantities $K_{i,i}$, $K_{i,j}K_{j,i}$, and the charged graph G_K can be computed immediately, as $K_{i,i} = \Delta_i$ and $K_{i,j}K_{j,i} = \Delta_i\Delta_j - \Delta_{i,j}$ for all $i \neq j$. The main focus of this subsection is to obtain further information on K using principal minors of order greater than two, and to quantify the extent to which K can be identified from its principal minors. To avoid the unintended cancellation of terms in a principal minor, in what follows we assume that K is *generic* in the sense that

$$\begin{aligned}&\text{If } K_{i,j}K_{j,k}K_{k,\ell}K_{\ell,i} \neq 0 \text{ for } i,j,k,\ell \in [N] \text{ distinct, then } |K_{i,j}K_{k,\ell}| \neq |K_{j,k}K_{\ell,i}|, \quad (2.6)\\ &\text{and } \phi_1\, K_{i,j}K_{j,k}K_{k,\ell}K_{\ell,i} + \phi_2 K_{i,j}K_{j,\ell}K_{\ell,k}K_{k,i} + \phi_3 K_{i,k}K_{k,j}K_{j,\ell}K_{\ell,i} \neq 0\\ &\text{for all } \phi_1,\phi_2,\phi_3 \in \{-1,1\}.\end{aligned}$$

The first part of the condition in (2.6) implies that the three terms (corresponding to the three distinct cycles on 4 vertices) in the second part are all distinct in magnitude; the second part strengthens this requirement. Condition (2.6) can be thought of as a no-cancellation requirement for four-cycles of principal minors of order four. As we will see, this condition is quite important for the recovery of a magnitude-symmetric matrix from its principal minors. Though, the results of this section, slightly modified, hold under slightly weaker conditions than (2.6), albeit at the cost of simplicity. This condition rules out dense matrices with a high degree of symmetry, as well as the large majority of rank one matrices, but this condition is satisfied for almost all matrices.

We denote the set of $N \times N$ magnitude-symmetric matrices satisfying Property (2.6) by $\mathcal{K}_N$, and, when the dimension is clear from context, we often simply write $\mathcal{K}$. We note

that if $K \in \mathcal{K}$, then any matrix K' satisfying $|K'_{i,j}| = |K_{i,j}|$, $i, j \in [N]$, is also in $\mathcal{K}$. In this subsection, we answer the following question:

$$\begin{aligned} &\text{Given } K \in \mathcal{K}, \text{ what is the minimal } \ell_+ \text{ such that any } K' \in \mathcal{K} \text{ with} \\ &\Delta_S(K) = \Delta_S(K') \text{ for all } |S| \leq \ell_+ \text{ also satisfies } \Delta_S(K) = \Delta_S(K') \\ &\text{for all } S \subset [N]? \end{aligned} \tag{2.7}$$

Question (2.7) asks for the smallest ℓ such that the principal minors of order at most ℓ uniquely determines principal minors of all orders. In Theorem 5, we show that the answer is the simple cycle sparsity of $\mathcal{C}^+(G)$. In Subsections 2.3.2 and 2.3.3, we build on the analysis of this subsection, and produce a polynomial-time algorithm for recovering a matrix K with prescribed principal minors (possibly given up to some error term). The polynomial-time algorithm (of Subsection 2.3.3) for recovery given perturbed principal minors has key connections to learning signed DPPs.

In addition, it is also reasonable to ask:

$$\begin{aligned} &\text{Given } K \in \mathcal{K}, \text{ what is the set of } K' \in \mathcal{K} \text{ that satisfy } \Delta_S(K) = \Delta_S(K') \\ &\text{for all } S \subset [N]? \end{aligned} \tag{2.8}$$

Question (2.8) treats the extent to which we can hope to recover a matrix from its principal minors. For instance, the transpose operation K^T and the similarity transformation DKD, where D is a diagonal matrix with entries $\{-1, +1\}$ on the diagonal, both clearly preserve principal minors. In fact, these two operations suitably combined completely define this set. This result follows fairly quickly from a more general result of Loewy [79, Theorem 1] regarding matrices with all principal minors equal. In particular, Loewy shows that if two $n \times n$, $n \geq 4$, matrices A and B have all principal minors equal, and A is irreducible and $\text{rank}(A_{S,T}) \geq 2$ or $\text{rank}(A_{T,S}) \geq 2$ (where $A_{S,T}$ is the submatrix of A containing the rows in S and the columns in T) for all partitions $S \cup T = [n]$, $S \cap T = \emptyset$, $|S|, |T| \geq 2$, then either B or B^T is diagonally similar to A. In Proposition 4, we include an alternate proof

answering Question (2.8), as Loewy's result and proof, though more general and quite nice, is more involved and not as illuminating for the specific case that we consider here.

To answer Question (2.7), we study properties of certain bases of the space $\mathcal{C}^+(G)$. We recall that, given a matrix $K \in \mathcal{K}$, we can define the charged sparsity graph $G = ([N], E, \epsilon)$ of K, where G is the simple graph on vertex set $[N]$ with an edge $\{i, j\} \in E$, $i \neq j$, if $K_{i,j} \neq 0$ and endowed with a function ϵ mapping subgraphs of G to $\{-1, +1\}$. Viewing the possible values ± 1 for ϵ as the two elements of $GF(2)$, we note that $\epsilon(H)$ is additive over $GF(2)$, i.e. $\epsilon(H_1 + H_2) = \epsilon(H_1)\epsilon(H_2)$. We can make a number of observations regarding the connectivity of G. If G is disconnected, with connected components given by vertex sets $V_1, ..., V_k$, then any principal minor Δ_S satisfies

$$\Delta_S = \prod_{j=1}^{k} \Delta_{S \cap V_j}. \tag{2.9}$$

In addition, if G has a cut vertex i, whose removal results in connected components with vertex sets $V_1, ..., V_k$, then the principal minor Δ_S satisfies

$$\Delta_S = \sum_{j_1=1}^{k} \Delta_{[\{i\} \cup V_{j_1}] \cap S} \prod_{j_2 \neq j_1} \Delta_{V_{j_2} \cap S} - (k-1)\Delta_{\{i\} \cap S} \prod_{j=1}^{k} \Delta_{V_j \cap S}. \tag{2.10}$$

This implies that the principal minors of K are uniquely determined by principal minors corresponding to subsets of blocks of G. Given this property, we focus on matrices K without an articulation point, i.e. G is two-connected. Given results for matrices without an articulation point and Equations (2.9) and (2.10), we can then answer Question (2.7) in the more general case.

Next, we make an important observation regarding the contribution of certain terms in the Laplace expansion of a principal minor. Recall that the Laplace expansion of the determinant is given by

$$\det(K) = \sum_{\sigma \in \mathcal{S}_N} \text{sgn}(\sigma) \prod_{i=1}^{N} K_{i,\sigma(i)},$$

where $\text{sgn}(\sigma)$ is multiplicative over the (disjoint) cycles forming the permutation, and is preserved when taking the inverse of the permutation, or of any cycle therein. Consider now an arbitrary, possibly non-induced, simple cycle C (in the graph-theoretic sense) of the sparsity graph G, without loss of generality given by $C = 1\ 2\ ...\ k\ 1$, that satisfies $\epsilon(C) = -1$. Consider the sum of all terms in the Laplace expansion that contains either the cycle $(1\ 2\ ...\ k)$ or its inverse $(k\ k-1\ ...\ 1)$ in the associated permutation. Because $\epsilon(C) = -1$,

$$K_{k,1}\prod_{i=1}^{k-1} K_{i,i+1} + K_{1,k}\prod_{i=2}^{k} K_{i,i-1} = 0,$$

and the sum over all permutations containing the cyclic permutations $(k\ k-1\ ...\ 1)$ or $(1\ 2\ ...\ k)$ is zero. Therefore, terms associated with permutations containing negative cycles ($\epsilon(C) = -1$) do not contribute to principal minors.

To illustrate the additional information contained in higher order principal minors Δ_S, $|S| > 2$, we first consider principal minors of order three and four. Consider the principal minor $\Delta_{1,2,3}$, given by

$$\begin{aligned}
\Delta_{1,2,3} &= K_{1,1}K_{2,2}K_{3,3} - K_{1,1}K_{2,3}K_{3,2} - K_{2,2}K_{1,3}K_{3,1} - K_{3,3}K_{1,2}K_{2,1} \\
&\quad + K_{1,2}K_{2,3}K_{3,1} + K_{3,2}K_{2,1}K_{1,3} \\
&= \Delta_1\Delta_2\Delta_3 - \Delta_1\big[\Delta_2\Delta_3 - \Delta_{2,3}\big] - \Delta_2\big[\Delta_1\Delta_3 - \Delta_{1,3}\big] - \Delta_3\big[\Delta_1\Delta_2 - \Delta_{1,2}\big] \\
&\quad + \big[1 + \epsilon_{1,2}\epsilon_{2,3}\epsilon_{3,1}\big]K_{1,2}K_{2,3}K_{3,1}.
\end{aligned}$$

If the corresponding graph $G[\{1,2,3\}]$ is not a cycle, or if the cycle is negative ($\epsilon_{1,2}\epsilon_{2,3}\epsilon_{3,1} = -1$), then $\Delta_{1,2,3}$ can be written in terms of principal minors of order at most two, and contains no additional information about K. If $G[\{1,2,3\}]$ is a positive cycle, then we can write $K_{1,2}K_{2,3}K_{3,1}$ as a function of principal minors of order at most three,

$$K_{1,2}K_{2,3}K_{3,1} = \Delta_1\Delta_2\Delta_3 - \frac{1}{2}\big[\Delta_1\Delta_{2,3} + \Delta_2\Delta_{1,3} + \Delta_3\Delta_{1,2}\big] + \frac{1}{2}\Delta_{1,2,3},$$

which allows us to compute $\text{sgn}(K_{1,2}K_{2,3}K_{3,1})$. This same procedure holds for any positively charged induced simple cycle in G. However, when a simple cycle is not induced,

further analysis is required. To illustrate some of the potential issues for a non-induced positive simple cycle, we consider principal minors of order four.

Consider the principal minor $\Delta_{1,2,3,4}$. By our previous analysis, all terms in the Laplace expansion of $\Delta_{1,2,3,4}$ corresponding to permutations with cycles of length at most three can be written in terms of principal minors of size at most three. What remains is to consider the sum of the terms in the Laplace expansion corresponding to permutations containing a cycle of length four (which there are three pairs of, each pair corresponding to the two orientations of a cycle of length four in the graph sense), which we denote by Z, given by

$$\begin{aligned} Z = &-K_{1,2}K_{2,3}K_{3,4}K_{4,1} - K_{1,2}K_{2,4}K_{4,3}K_{3,1} - K_{1,3}K_{3,2}K_{2,4}K_{4,1} \\ &- K_{4,3}K_{3,2}K_{2,1}K_{1,4} - K_{4,2}K_{2,1}K_{1,3}K_{3,4} - K_{4,2}K_{2,3}K_{3,1}K_{1,4}. \end{aligned}$$

If $G[\{1,2,3,4\}]$ does not contain a positive simple cycle of length four, then $Z = 0$ and $\Delta_{1,2,3,4}$ can be written in terms of principal minors of order at most three. If $G[\{1,2,3,4\}]$ has exactly one positive cycle of length four, without loss of generality given by $C := 1\ 2\ 3\ 4\ 1$, then

$$Z = -\big[1 + \epsilon(C)\big] K_{1,2}K_{2,3}K_{3,4}K_{4,1} = -2K_{1,2}K_{2,3}K_{3,4}K_{4,1},$$

and we can write $K_{1,2}K_{2,3}K_{3,4}K_{4,1}$ as a function of principal minors of order at most four, which allows us to compute $\text{sgn}(K_{1,2}K_{2,3}K_{3,4}K_{4,1})$. Finally, we treat the case in which there is more than one positive simple cycle of length four. This implies that all simple cycles of length four are positive and Z is given by

$$Z = -2\big[K_{1,2}K_{2,3}K_{3,4}K_{4,1} + K_{1,2}K_{2,4}K_{4,3}K_{3,1} + K_{1,3}K_{3,2}K_{2,4}K_{4,1}\big]. \tag{2.11}$$

By Condition (2.6), the magnitude of each of these three terms is distinct, $Z \neq 0$, and we can compute the sign of each of these terms, as there is a unique choice of $\phi_1, \phi_2, \phi_3 \in \{-1, +1\}$

satisfying

$$\phi_1|K_{1,2}K_{2,3}K_{3,4}K_{4,1}| + \phi_2|K_{1,2}K_{2,4}K_{4,3}K_{3,1}| + \phi_3|K_{1,3}K_{3,2}K_{2,4}K_{4,1}| = -\frac{Z}{2}.$$

In order to better understand the behavior of principal minors of order greater than four, we investigate the set of positive simple cycles (i.e., simple cycles C satisfying $\epsilon(C) = +1$). In the following two propositions, we compute the dimension of $\mathcal{C}^+(G)$, and note that when G is two-connected, this space is spanned by incidence vectors corresponding to positive simple cycles.

Proposition 2. *Let $G = ([N], E, \epsilon)$ be a charged graph. Then $\dim(\mathcal{C}^+(G)) = \nu - 1$ if G contains at least one negative cycle, and equals ν otherwise.*

Proof. Consider a basis $\{x_1, ..., x_\nu\}$ for $\mathcal{C}(G)$. If all associated cycles are positive, then $\dim(\mathcal{C}^+(G)) = \nu$ and we are done. If G has a negative cycle, then, by the linearity of ϵ, at least one element of $\{x_1, ..., x_\nu\}$ must be the incidence vector of a negative cycle. Without loss of generality, suppose that $x_1, ..., x_i$ are incidence vectors of negative cycles. Then $x_1 + x_2, ..., x_1 + x_i, x_{i+1}, ..., x_\nu$ is a linearly independent set of incidence vectors of positive cycles. Therefore, $\dim(\mathcal{C}^+(G)) \geq \nu - 1$. However, x_1 is the incidence vector of a negative cycle, and so cannot be in the span of $x_1 + x_2, ..., x_1 + x_i, x_{i+1}, ..., x_\nu$. □

Proposition 3. *Let $G = ([N], E, \epsilon)$ be a two-connected charged graph. Then*

$$\mathcal{C}^+(G) = \textit{span}\{\chi_C \mid C \textit{ is a simple cycle of } G, \epsilon(C) = +1\}.$$

Proof. If G does not contain a negative cycle, then the result follows immediately, as then $\mathcal{C}^+(G) = \mathcal{C}(G)$. Suppose then that G has at least one negative cycle. Decomposing this negative cycle into the union of edge-disjoint simple cycles, we note that G must also have at least one negative simple cycle.

Since G is a two-connected graph, it admits a proper (also called open) ear decomposition with $\nu - 1$ proper ears (Whitney [128], see also [63]) starting from any simple cycle. We choose our initial simple cycle to be a negative simple cycle denoted by G_0. Whitney's proper

ear decomposition says that we can obtain a sequence of graphs G_0, $G_1, \cdots, G_{\nu-1} = G$ where G_i is obtained from G_{i-1} by adding a path P_i between two distinct vertices u_i and v_i of $V(G_{i-1})$ with its internal vertices not belonging to $V(G_{i-1})$ (a proper or open ear). By construction, P_i is also a path between u_i and v_i in G_i.

We will prove a stronger statement by induction on i, namely that for suitably constructed positive simple cycles C_j, $j = 1, ..., \nu - 1$, we have that, for every $i = 0, 1, ..., \nu - 1$:

(i) $\mathcal{C}^+(G_i) = \text{span}\{\chi_{C_j} : 1 \leq j \leq i\}$,

(ii) for every pair of distinct vertices $u, v \in V(G_i)$, there exists both a positive path in G_i between u and v and a negative path in G_i between u and v.

For $i = 0$, (i) is clear and (ii) follows from the fact that the two paths between u and v in the cycle G_0 must have opposite charge since their sum is G_0 and $\epsilon(G_0) = -1$. For $i > 0$, we assume that we have constructed $C_1, \cdots, C_{i-1}$ satisfying (i) and (ii) for smaller values of i. To construct C_i, we take the path P_i between u_i and v_i and complete it with a path of charge $\epsilon(P_i)$ between u_i and v_i in G_{i-1}. The existence of this latter path follows from (ii), and these two paths together form a simple cycle C_i with $\epsilon(C_i) = +1$. It is clear that this C_i is linearly independent from all the previously constructed cycles C_j since C_i contains P_i but P_i was not even part of G_{i-1}. This implies (i) for i.

To show (ii) for G_i, we need to consider three cases for $u \neq v$. If $u, v \in V(G_{i-1})$, we can use the corresponding positive and negative paths in G_{i-1} (whose existence follows from (ii) for $i - 1$). If $u, v \in V(P_i)$, one of the paths can be the subpath P_{uv} of P_i between u and v and the other path can be $P_i \setminus P_{uv}$ together with a path in G_{i-1} between u_i and v_i of charge equal to $-\epsilon(P_i)$ (so that together they form a negative cycle). Otherwise, we must have $u \in V(P_i) \setminus \{u_i, v_i\}$ and $v \in V(G_{i-1}) \setminus \{u_i, v_i\}$ (or vice versa), and we can select the paths to be the path in P_i from u to u_i together with two oppositely charged paths in G_{i-1} between u_i and v. In all these cases, we have shown property (ii) holds for G_i. □

For the remainder of the analysis in this subsection, we assume that G contains a negative cycle, otherwise $\mathcal{C}^+(G) = \mathcal{C}(G)$ and we inherit all of the desirable properties of $\mathcal{C}(G)$. Next, we study the properties of simple cycle bases (bases consisting of incidence

vectors of simple cycles) of $\mathcal{C}^+(G)$ that are minimal in some sense. We say that a simple cycle basis $\{\chi_{C_1}, ..., \chi_{C_{\nu-1}}\}$ for $\mathcal{C}^+(G)$ is lexicographically minimal if it lexicographically minimizes $(|C_1|, |C_2|, ..., |C_{\nu-1}|)$, i.e., minimizes $|C_1|$, minimizes $|C_2|$ conditional on $|C_1|$ being minimal, etc. Any lexicographical minimal basis also minimizes $\sum |C_i|$ and $\max |C_i|$ (by optimality of the greedy algorithm for (linear) matroids). For brevity, we will simply refer to such a basis as "minimal." One complicating issue is that, while a minimal cycle basis for $\mathcal{C}(G)$ always consists of induced simple cycles, this is no longer the case for $\mathcal{C}^+(G)$. This for example can happen for an appropriately constructed graph with only two negative, short simple cycles, far away from each other; a minimal cycle basis for $\mathcal{C}^+(G)$ then contains a cycle consisting of the union of these 2 negative simple cycles.

However, we can make a number of statements regarding chords of simple cycles corresponding to elements of a minimal simple cycle basis of $\mathcal{C}^+(G)$. In the following lemma, we show that a minimal simple cycle basis satisfies a number of desirable properties. However, before we do so, we introduce some useful notation. Given a simple cycle C, it is convenient to fix an orientation say $C = i_1\ i_2\ ...\ i_k\ i_1$. Now for any chord $\{i_{k_1}, i_{k_2}\} \in \gamma(C)$, $k_1 < k_2$, we denote the two cycles created by this chord by

$$C(k_1, k_2) = i_{k_1}\ i_{k_1+1}\ ...\ i_{k_2}\ i_{k_1} \quad \text{and} \quad C(k_2, k_1) = i_{k_2}\ i_{k_2+1}\ ...\ i_k\ i_1\ ...\ i_{k_1}\ i_{k_2}.$$

We have the following result.

Lemma 6. *Let $G = ([N], E, \epsilon)$ be a two-connected charged graph, and $C := i_1\ i_2\ ...\ i_k\ i_1$ be a cycle corresponding to an incidence vector of a minimal simple cycle basis $\{x_1, ..., x_{\nu-1}\}$ of $\mathcal{C}^+(G)$. Then*

(i) $\epsilon\big(C(k_1, k_2)\big) = \epsilon\big(C(k_2, k_1)\big) = -1$ *for all* $\{i_{k_1}, i_{k_2}\} \in \gamma(C)$,

(ii) if $\{i_{k_1}, i_{k_2}\}, \{i_{k_3}, i_{k_4}\} \in \gamma(C)$ *satisfy* $k_1 < k_3 < k_2 < k_4$ *(crossing chords), then either* $k_3 - k_1 = k_4 - k_2 = 1$ *or* $k_1 = 1$, $k_2 - k_3 = 1$, $k_4 = k$, *i.e. these two chords form a four-cycle with two edges of the cycle,*

(iii) there does not exist $\{i_{k_1}, i_{k_2}\}, \{i_{k_3}, i_{k_4}\}, \{i_{k_5}, i_{k_6}\} \in \gamma(C)$ *satisfying either* $k_1 < k_2 \leq$

$k_3 < k_4 \leq k_5 < k_6$ *or* $k_6 \leq k_1 < k_2 \leq k_3 < k_4 \leq k_5$.

Proof. Without loss of generality, suppose that $C = 1\ 2\ ...\ k$. Given a chord $\{k_1, k_2\}$, $C(k_1, k_2) + C(k_2, k_1) = C$, and so $\epsilon(C) = +1$ implies that $\epsilon\big(C(k_1, k_2)\big) = \epsilon\big(C(k_2, k_1)\big)$. If both these cycles were positive, then this would contradict the minimality of the basis, as $|C(k_1, k_2)|, |C(k_2, k_1)| < k$. This completes the proof of Property (i).

Given two chords $\{k_1, k_2\}$ and $\{k_3, k_4\}$ satisfying $k_1 < k_3 < k_2 < k_4$, we consider the cycles

$$C_1 := C(k_1, k_2) + C(k_3, k_4),$$
$$C_2 := C(k_1, k_2) + C(k_4, k_3).$$

By Property (i), $\epsilon(C_1) = \epsilon(C_2) = +1$. In addition, $C_1 + C_2 = C$, and $|C_1|, |C_2| \leq |C|$. By the minimality of the basis, either $|C_1| = |C|$ or $|C_2| = |C|$, which implies that either $|C_1| = 4$ or $|C_2| = 4$ (or both). This completes the proof of Property (ii).

Given three non-crossing chords $\{k_1, k_2\}$, $\{k_3, k_4\}$, and $\{k_5, k_6\}$, with $k_1 < k_2 \leq k_3 < k_4 \leq k_5 < k_6$ (the other case is identical, up to rotation), we consider the cycles

$$C_1 := C(k_3, k_4) + C(k_5, k_6) + C,$$
$$C_2 := C(k_1, k_2) + C(k_5, k_6) + C,$$
$$C_3 := C(k_1, k_2) + C(k_3, k_4) + C.$$

By Property (i), $\epsilon(C_1) = \epsilon(C_2) = \epsilon(C_3) = +1$. In addition, $C_1 + C_2 + C_3 = C$ and $|C_1|, |C_2|, |C_3| < |C|$, a contradiction to the minimality of the basis. This completes the proof of Property (iii). □

The above lemma tells us that in a minimal simple cycle basis for $\mathcal{C}^+(G)$, in each cycle two crossing chords always form a positive cycle of length four (and thus any chord has at most one other crossing chord), and there doesn't exist three non-crossing chords without one chord in between the other two. Using a simple cycle basis of $\mathcal{C}^+(G)$ that satisfies the three properties of the above lemma, we aim to show that the principal minors of order at

most the length of the longest cycle in the basis uniquely determine the principal minors of all orders. To do so, we prove a series of three lemmas that show that from the principal minor corresponding to a cycle $C = i_1\ ...\ i_k\ i_1$ in our basis and $O(1)$ principal minors of subsets of $V(C)$, we can compute the quantity

$$s(C) := \prod_{\substack{\{i,j\}\in E(C),\\ i<j}} \operatorname{sgn}\big(K_{i,j}\big).$$

Once we have computed $s(C_i)$ for every simple cycle in our basis, we can compute $s(C)$ for any simple positive cycle by writing C as a sum of some subset of the simple cycles $C_1, ..., C_{\nu-1}$ corresponding to incidence vectors in our basis, and taking the product of $s(C_i)$ for all indices i in the aforementioned sum. To this end, we prove the following three lemmas.

Lemma 7. *Let C be a positive simple cycle of $G = ([N], E, \epsilon)$ with $V(C) = \{i_1, i_2, \cdots, i_k\}$ whose chords $\gamma(C)$ satisfy Properties (i)-(iii) of Lemma 6. Then there exist two edges of C, say, $\{i_1, i_k\}$ and $\{i_\ell, i_{\ell+1}\}$, where $C := i_1\ i_2\ ...\ i_k\ i_1$ such that*

(i) all chords $\{i_p, i_q\} \in \gamma(C)$, $p < q$, satisfy $p \le \ell < q$,

(ii) any positive simple cycle in $G[V(C)]$ containing $\{i_1, i_k\}$ spans $V(C)$ and contains only edges in $E(C)$ and pairs of crossed chords.

Proof. We first note that Property (i) of Lemma 6 implies that the charge of a cycle $C' \subset G[V(C)]$ corresponds to the parity of $E(C') \cap \gamma(C)$, i.e., if C' contains an even number of chords then $\epsilon(C') = +1$, otherwise $\epsilon(C') = -1$. This follows from constructing a cycle basis for $G[V(C)]$ consisting of incidence vectors corresponding to C and $C(u, v)$ for every chord $\{u, v\} \in \gamma(C)$.

Let the cycle $C(i, j)$ be a shortest cycle in $G[V(C)]$ containing exactly one chord of C. If this is a non-crossed chord, then we take an arbitrary edge in $E(C(i, j))\backslash\{i, j\}$ to be $\{i_1, i_k\}$. If this chord crosses another, denoted by $\{i', j'\}$, then either $C(i', j')$ or $C(j', i')$ is also a shortest cycle, and we take an arbitrary edge in the intersection of these two shortest

cycles. Without loss of generality, let $\{1, k\}$ denote this edge and let the cycle be $1\ 2\ ...\ k\ 1$, where for simplicity we have assumed that $V(C) = [k]$. Because of Property (iii) of Lemma 6, there exists ℓ such that all chords $\{i, j\} \in \gamma(C)$ satisfy $i \leq \ell$ and $j > \ell$. This completes the proof of Property (i).

Next, consider an arbitrary positive simple cycle C' in $G[V(C)]$ containing $\{1, k\}$. We aim to show that this cycle contains only edges in $E(G)$ and pairs of crossed chords, from which the equality $V(C') = V(C)$ immediately follows. Since C' is positive, it contains an even number of chords $\gamma(C)$, and either all the chords $\gamma(C) \cap E(C')$ are pairs of crossed chords or there exist two chords $\{p_1, q_1\}, \{p_2, q_2\} \in \gamma(C) \cap E(C')$, w.l.o.g. $p_1 \leq p_2 \leq \ell < q_2 < q_1$ (we can simply reverse the orientation if $q_1 = q_2$), that do not cross any other chord in $\gamma(C) \cap E(C')$. In this case, there would be a $1 - q_2$ path in C' neither containing p_1 nor q_1 (as $\{p_1, q_1\}$ would be along the other path between 1 and q_2 in C' and $q_1 \neq q_2$). However, this is a contradiction as no such path exists in C', since $\{p_1, q_1\}$ does not cross any chord in $\gamma(C) \cap E(C')$. Therefore, the chords $\gamma(C) \cap E(C')$ are all pairs of crossed chords. Furthermore, since all pairs of crossed chords must define cycles of length 4, we have $V(C') = V(C)$. This completes the proof of Property (ii). □

Although this is not needed in what follows, observe that, in the above Lemma 7, $\{i_\ell, i_{\ell+1}\}$ also plays the same role as $\{i_1, i_k\}$ in the sense that any positive simple cycle containing $\{i_\ell, i_{\ell+1}\}$ also spans $V(C)$ and contains only pairs of crossed chords and edges in $E(C)$.

Lemma 8. *Let $K \in \mathcal{K}$ have charged sparsity graph $G = ([N], E, \epsilon)$, and $C = i_1\ ...\ i_k\ i_1$ be a positive simple cycle of G whose chords $\gamma(C)$ satisfy Properties (i)-(iii) of Lemma 6, and with vertices ordered as in Lemma 7. Then the principal minor corresponding to*

$S = V(C)$ *is given by*

$$\begin{aligned}\Delta_S = \Delta_{i_1}\Delta_{S\setminus i_1} + \Delta_{i_k}\Delta_{S\setminus i_k} + \big[\Delta_{i_1,i_k} - 2\Delta_{i_1}\Delta_{i_k}\big]\Delta_{S\setminus i_1,i_k} \\ - 2K_{i_1,i_{k-1}}K_{i_{k-1},i_k}K_{i_k,i_2}K_{i_2,i_1}\Delta_{S\setminus i_1,i_2,i_{k-1},i_k} \\ - \big[\Delta_{i_1,i_2} - \Delta_{i_1}\Delta_{i_2}\big]\big[\Delta_{i_{k-1},i_k} - \Delta_{i_{k-1}}\Delta_{i_k}\big]\Delta_{S\setminus i_1,i_2,i_{k-1},i_k} \\ + \big[\Delta_{i_1,i_{k-1}} - \Delta_{i_1}\Delta_{i_{k-1}}\big]\big[\Delta_{i_2,i_k} - \Delta_{i_2}\Delta_{i_k}\big]\Delta_{S\setminus i_1,i_2,i_{k-1},i_k} \\ + \big[\Delta_{i_1,i_2} - \Delta_{i_1}\Delta_{i_2}\big]\big[\Delta_{S\setminus i_1,i_2} - \Delta_{i_k}\Delta_{S\setminus i_1,i_2,i_k}\big] \\ + \big[\Delta_{i_{k-1},i_k} - \Delta_{i_{k-1}}\Delta_{i_k}\big]\big[\Delta_{S\setminus i_{k-1},i_k} - \Delta_{i_2}\Delta_{S\setminus i_1,i_{k-1},i_k}\big] \\ + Z.\end{aligned}$$

where Z is the sum of terms in the Laplace expansion of Δ_S corresponding to a permutation where $\sigma(i_1) = i_k$ or $\sigma(i_k) = i_1$, but not both.

Proof. Without loss of generality, suppose that $C = 1\ 2\ ...\ k\ 1$. Each term in Δ_S corresponds to a partition of $S = V(C)$ into a disjoint collection of vertices, pairs corresponding to edges of $G[S]$, and simple cycles C_i of $E(S)$ which can be assumed to be positive. We can decompose the Laplace expansion of Δ_S into seven sums of terms, based on how the associated permutation for each term treats the elements 1 and k. Let X_{j_1,j_2}, $j_1, j_2 \in \{1, k, *\}$, equal the sum of terms corresponding to permutations where $\sigma(1) = j_1$ and $\sigma(k) = j_2$, where $*$ denotes any element in $\{2, ..., k-1\}$. The case $\sigma(1) = \sigma(k)$ obviously cannot occur, and so

$$\Delta_S = X_{1,k} + X_{1,*} + X_{*,k} + X_{k,1} + X_{k,*} + X_{*,1} + X_{*,*}.$$

By definition, $Z = X_{k,*} + X_{*,1}$. What remains is to compute the remaining five terms. We have

$$\begin{aligned}X_{k,1} &= -K_{1,k}K_{k,1}\Delta_{S\setminus 1,k} \\ &= \big[\Delta_{1,k} - \Delta_1\Delta_k\big]\Delta_{S\setminus 1,k},\end{aligned}$$

$$X_{1,k} = \Delta_1 \Delta_k \Delta_{S\setminus 1,k},$$
$$X_{1,k} + X_{1,*} = \Delta_1 \Delta_{S\setminus 1},$$
$$X_{1,k} + X_{*,k} = \Delta_k \Delta_{S\setminus k},$$

and so

$$\Delta_S = \Delta_1 \Delta_{S\setminus 1} + \Delta_k \Delta_{S\setminus k} + \big[\Delta_{1,k} - 2\Delta_1 \Delta_k\big] \Delta_{S\setminus 1,k} + X_{*,*} + Z.$$

The sum $X_{*,*}$ corresponds to permutations where $\sigma(1) \notin \{1, k\}$ and $\sigma(k) \notin \{1, k\}$. We first show that the only permutations that satisfy these two properties take the form $\sigma^2(1) = 1$ or $\sigma^2(k) = k$ (or both), or contain both 1 and k in the same cycle. Suppose to the contrary, that there exists some permutation where 1 and k are not in the same cycle, and both are in cycles of length at least three. Then in $G[S]$ both vertices 1 and k contain a chord emanating from them. By Property (ii) of Lemma 6, each has exactly one chord and these chords are $\{1, k-1\}$ and $\{2, k\}$. The cycle containing 1 must also contain 2 and $k-1$, and the cycle containing k must also contain 2 and $k-1$, a contradiction.

In addition, we note that, also by the above analysis, the only cycles (of a permutation) that can contain both 1 and k without having $\sigma(1) = k$ or $\sigma(k) = 1$ are given by $(1\ 2\ k\ k-1)$ and $(1\ k-1\ k\ 2)$. Therefore, we can decompose $X_{*,*}$ further into the sum of five terms. Let

- $Y_1 =$ the sum of terms corresponding to permutations containing either $(1\ 2\ k\ k-1)$ or $(1\ k-1\ k\ 2)$,
- $Y_2 =$ the sum of terms corresponding to permutations containing $(1\ 2)$ and $(k-1\ k)$,
- $Y_3 =$ the sum of terms corresponding to permutations containing $(1\ k-1)$ and $(2\ k)$,
- $Y_4 =$ the sum of terms corresponding to permutations containing $(1\ 2)$ and where $\sigma(k) \neq k-1, k$,
- $Y_5 =$ the sum of terms corresponding to permutations containing $(k-1\ k)$ and where $\sigma(1) \neq 1, 2$.

Y_1 and Y_3 are only non-zero if $\{1, k-1\}$ and $\{2, k\}$ are both chords of C. Y_4 is only non-zero if there is a chord incident to k, and Y_5 is only non-zero if there is a chord incident to 1. We have $X_{*,*} = Y_1 + Y_2 + Y_3 + Y_4 + Y_5$, and

$$\begin{aligned}
Y_1 &= -2K_{1,k-1}K_{k-1,k}K_{k,2}K_{2,1}\Delta_{S\setminus 1,2,k-1,k},\\
Y_2 &= K_{1,2}K_{2,1}K_{k-1,k}K_{k,k-1}\Delta_{S\setminus 1,2,k-1,k}\\
&= \big[\Delta_{1,2} - \Delta_1\Delta_2\big]\big[\Delta_{k-1,k} - \Delta_{k-1}\Delta_k\big]\Delta_{S\setminus 1,2,k-1,k},\\
Y_3 &= K_{1,k-1}K_{k-1,1}K_{2,k}K_{k,2}\Delta_{S\setminus 1,2,k-1,k}\\
&= \big[\Delta_{1,k-1} - \Delta_1\Delta_{k-1}\big]\big[\Delta_{2,k} - \Delta_2\Delta_k\big]\Delta_{S\setminus 1,2,k-1,k},\\
Y_2 + Y_4 &= -K_{1,2}K_{2,1}\big[\Delta_{S\setminus 1,2} - \Delta_k\Delta_{S\setminus 1,2,k}\big]\\
&= \big[\Delta_{1,2} - \Delta_1\Delta_2\big]\big[\Delta_{S\setminus 1,2} - \Delta_k\Delta_{S\setminus 1,2,k}\big],\\
Y_2 + Y_5 &= -K_{k-1,k}K_{k,k-1}\big[\Delta_{S\setminus k-1,k} - \Delta_1\Delta_{S\setminus 1,k-1,k}\big]\\
&= \big[\Delta_{k-1,k} - \Delta_{k-1}\Delta_k\big]\big[\Delta_{S\setminus k-1,k} - \Delta_2\Delta_{S\setminus 1,k-1,k}\big].
\end{aligned}$$

Combining all of these terms gives us

$$\begin{aligned}
X_{*,*} = &-2K_{1,k-1}K_{k-1,k}K_{k,2}K_{2,1}\Delta_{S\setminus 1,2,k-1,k}\\
&- \big[\Delta_{1,2} - \Delta_1\Delta_2\big]\big[\Delta_{k-1,k} - \Delta_{k-1}\Delta_k\big]\Delta_{S\setminus 1,2,k-1,k}\\
&+ \big[\Delta_{1,k-1} - \Delta_1\Delta_{k-1}\big]\big[\Delta_{2,k} - \Delta_2\Delta_k\big]\Delta_{S\setminus 1,2,k-1,k}\\
&+ \big[\Delta_{1,2} - \Delta_1\Delta_2\big]\big[\Delta_{S\setminus 1,2} - \Delta_k\Delta_{S\setminus 1,2,k}\big]\\
&+ \big[\Delta_{k-1,k} - \Delta_{k-1}\Delta_k\big]\big[\Delta_{S\setminus k-1,k} - \Delta_2\Delta_{S\setminus 1,k-1,k}\big].
\end{aligned}$$

Combining our formula for $X_{*,*}$ with our formula for Δ_S leads to the desired result

$$\begin{aligned}\Delta_S = \Delta_1\Delta_{S\setminus 1} + \Delta_k\Delta_{S\setminus k} + \left[\Delta_{1,k} - 2\Delta_1\Delta_k\right]\Delta_{S\setminus 1,k} \\ - 2K_{1,k-1}K_{k-1,k}K_{k,2}K_{2,1}\Delta_{S\setminus 1,2,k-1,k} \\ - \left[\Delta_{1,2} - \Delta_1\Delta_2\right]\left[\Delta_{k-1,k} - \Delta_{k-1}\Delta_k\right]\Delta_{S\setminus 1,2,k-1,k} \\ + \left[\Delta_{1,k-1} - \Delta_1\Delta_{k-1}\right]\left[\Delta_{2,k} - \Delta_2\Delta_k\right]\Delta_{S\setminus 1,2,k-1,k} \\ + \left[\Delta_{1,2} - \Delta_1\Delta_2\right]\left[\Delta_{S\setminus 1,2} - \Delta_k\Delta_{S\setminus 1,2,k}\right] \\ + \left[\Delta_{k-1,k} - \Delta_{k-1}\Delta_k\right]\left[\Delta_{S\setminus k-1,k} - \Delta_2\Delta_{S\setminus 1,k-1,k}\right] + Z.\end{aligned}$$

□

Lemma 9. *Let $K \in \mathcal{K}$ have two-connected charged sparsity graph $G = ([N], E, \epsilon)$, and $C = i_1 \ldots i_k\ i_1$ be a positive simple cycle of G whose chords $\gamma(C)$ satisfy Properties (i)-(iii) of Lemma 6, with vertices ordered as in Lemma 7. Let Z equal the sum of terms in the Laplace expansion of Δ_S, $S = V(C)$, corresponding to a permutation where $\sigma(i_1) = i_k$ or $\sigma(i_k) = i_1$, but not both. Let $U \subset S^2$ be the set of pairs (a, b), $a < b$, for which $\{i_a, i_{b-1}\}, \{i_{a+1}, i_b\} \in \gamma(C)$, i.e., cycle edges $\{i_a, i_{a+1}\}$ and $\{i_{b-1}, i_b\}$) have a pair of crossed chords between them. Then Z is equal to*

$$2\,(-1)^{k+1}\,K_{i_k,i_1}\prod_{j=1}^{k-1}K_{i_j,i_{j+1}}\prod_{(a,b)\in U}\left[1 - \frac{\epsilon_{i_a,i_{a+1}}K_{i_{b-1},i_a}K_{i_{a+1},i_b}}{K_{i_a,i_{a+1}}K_{i_{b-1},i_b}}\right].$$

Proof. Without loss of generality, suppose that G is labelled so that $C = 1\ 2\ \ldots\ k\ 1$. Edge $\{1, k\}$ satisfies Property (ii) of Lemma 7, and so every positive simple cycle of $G[S]$ containing $\{1, k\}$ spans S and contains only edges in $E(C)$ and pairs of crossing chords. Therefore, we may assume without loss of generality that $G[S]$ contains p pairs of crossing chords, and no other chords. If $p = 0$, then C is an induced cycle and the result follows immediately.

There are 2^p Hamiltonian $1 - k$ paths in $G[S]$ not containing $\{1, k\}$, corresponding to the p binary choices of whether to use each pair of crossing chords or not. Let us

denote these paths from 1 to k by $P_\theta^{1,k}$, where $\theta \in \mathbb{Z}_2^p$, and θ_i equals one if the i^{th} crossing is used in the path (ordered based on increasing distance to $\{1,k\}$), and zero otherwise. In particular, $P_0^{1,k}$ corresponds to the path $= 1\ 2\ ...\ k$. We only consider paths with the orientation $1 \to k$, as all cycles under consideration are positive, and the opposite orientation has the same sum. Denoting the product of the terms of K corresponding to the path $P_\theta^{1,k} = 1 = \ell_1\ \ell_2\ ...\ \ell_k = k$, by

$$K\big(P_\theta^{1,k}\big) = \prod_{j=1}^{k-1} K_{\ell_j,\ell_{j+1}},$$

we have

$$Z = 2\,(-1)^{k+1}\,K_{k,1} K\left(P_0^{1,k}\right) \sum_{\theta\in\mathbb{Z}_2^p} \frac{K\big(P_\theta^{1,k}\big)}{K\big(P_0^{1,k}\big)},$$

where

$$K\big(P_0^{1,k}\big) = \prod_{j=1}^{k-1} K_{j,j+1}.$$

Suppose that the first possible crossing occurs at cycle edges $\{a, a+1\}$ and $\{b-1, b\}$, $a < b$, i.e., $\{a, b-1\}$ and $\{a+1, b\}$ are crossing chords. Then

$$Z = 2\,(-1)^{k+1}\,K_{k,1} \prod_{j=1}^{k-1} K_{j,j+1} \sum_{\theta\in\mathbb{Z}_2^p} \frac{K\big(P_\theta^{a,b}\big)}{K\big(P_0^{a,b}\big)}.$$

We have

$$\sum_{\theta\in\mathbb{Z}_2^p} K\big(P_\theta^{a,b}\big) = \sum_{\theta_1=0} K\big(P_\theta^{a,b}\big) + \sum_{\theta_1=1} K\big(P_\theta^{a,b}\big),$$

and

$$\sum_{\theta_1=0} K\big(P_\theta^{a,b}\big) = K_{a,a+1}K_{b-1,b} \sum_{\theta'\in\mathbb{Z}_2^{p-1}} K\big(P_{\theta'}^{a+1,b-1}\big),$$

$$\sum_{\theta_1=1} K\big(P_\theta^{a,b}\big) = K_{a,b-1}K_{a+1,b} \sum_{\theta' \in \mathbb{Z}_2^{p-1}} K\big(P_{\theta'}^{b-1,a+1}\big)$$
$$= K_{a,b-1}K_{a+1,b} \sum_{\theta' \in \mathbb{Z}_2^{p-1}} K\big(P_{\theta'}^{a+1,b-1}\big)\epsilon\big(P_{\theta'}^{a+1,b-1}\big).$$

By Property (i) of Lemma 6, $\epsilon\big(C(a,b-1)\big) = -1$, and so

$$\epsilon\big(P_{\theta'}^{a+1,b-1}\big) = -\epsilon_{b-1,a}\epsilon_{a,a+1} \quad \text{and} \quad K_{a,b-1}\epsilon\big(P_{\theta'}^{a+1,b-1}\big) = -\epsilon_{a,a+1}K_{b-1,a}.$$

This implies that

$$\sum_{\theta \in \mathbb{Z}_2^p} K\big(P_\theta^{a,b}\big) = \big(K_{a,a+1}K_{b-1,b} - \epsilon_{a,a+1}K_{b-1,a}K_{a+1,b}\big) \sum_{\theta' \in \mathbb{Z}_2^{p-1}} K\left(P_{\theta'}^{a+1,b-1}\right),$$

and that

$$\sum_{\theta \in \mathbb{Z}_2^p} \frac{K\big(P_\theta^{a,b}\big)}{K\big(P_0^{a,b}\big)} = \left(1 - \frac{\epsilon_{a,a+1}K_{b-1,a}K_{a+1,b}}{K_{a,a+1}K_{b-1,b}}\right) \sum_{\theta' \in \mathbb{Z}_2^{p-1}} \frac{K\big(P_{\theta'}^{a+1,b-1}\big)}{K\big(P_0^{a+1,b-1}\big)}.$$

Repeating the above procedure for the remaining $p-1$ crossings completes the proof. $\square$

Equipped with Lemmas 6, 7, 8, and 9, we can now make a key observation. Suppose that we have a simple cycle basis $\{x_1, ..., x_{\nu-1}\}$ for $\mathcal{C}^+(G)$ whose corresponding cycles all satisfy Properties (i)-(iii) of Lemma 6. Of course, a minimal simple cycle basis satisfies this, but in Subsections 2.3.2 and 2.3.3 we will consider alternate bases that also satisfy these conditions and may be easier to compute in practice. For cycles C of length at most four, we have already detailed how to compute $s(C)$, and this computation requires only principal minors corresponding to subsets of $V(C)$. When C is of length greater than four, by Lemmas 8 and 9, we can also compute $s(C)$, using only principal minors corresponding to subsets of $V(C)$, as the quantity $\text{sgn}(K_{i_1,i_{k-1}}K_{i_{k-1},i_k}K_{i_k,i_2}K_{i_2,i_1})$ in Lemma 8 corresponds to a positive four-cycle, and in Lemma 9 the quantity

$$\text{sgn}\left(1 - \frac{\epsilon_{i_a,i_{a+1}}K_{i_{b-1},i_a}K_{i_{a+1},i_b}}{K_{i_a,i_{a+1}}K_{i_{b-1},i_b}}\right)$$

equals $+1$ if $|K_{i_a,i_{a+1}}K_{i_{b-1},i_b}| > |K_{i_{b-1},i_a}K_{i_{a+1},i_b}|$ and equals

$$-\epsilon_{i_a,i_{a+1}}\epsilon_{i_{b-1},i_b}\,\mathrm{sgn}\big(K_{i_a,i_{a+1}}K_{i_{a+1},i_b}K_{i_b,i_{b-1}}K_{i_{b-1},i_a}\big)$$

if $|K_{i_a,i_{a+1}}K_{i_{b-1},i_b}| < |K_{i_{b-1},i_a}K_{i_{a+1},i_b}|$. By Condition (2.6), we have $|K_{i_a,i_{a+1}}K_{i_{b-1},i_b}| \neq |K_{i_{b-1},i_a}K_{i_{a+1},i_b}|$. Therefore, given a simple cycle basis $\{x_1, ..., x_{\nu-1}\}$ whose corresponding cycles satisfy Properties (i)-(iii) of Lemma 6, we can compute $s(C)$ for every such cycle in the basis using only principal minors of size at most the length of the longest cycle in the basis. Given this fact, we are now prepared to answer Question (2.7) and provide an alternate proof for Question (2.8) through the following proposition and theorem.

Proposition 4. *Let $K \in \mathcal{K}$, with charged sparsity graph $G = ([N], E, \epsilon)$. The set of $K' \in \mathcal{K}$ that satisfy $\Delta_S(K) = \Delta_S(K')$ for all $S \subset [N]$ is exactly the set generated by K and the operations*

$\mathcal{D}_N$-similarity: $K \to DKD$, *where $D \in \mathbb{R}^{N\times N}$ is an arbitrary involutory diagonal matrix, i.e., D has entries ± 1 on the diagonal, and*

block transpose: $K \to K'$, *where $K'_{i,j} = K_{j,i}$ for all $i, j \in V(H)$ for some block H, and $K'_{i,j} = K_{i,j}$ otherwise.*

Proof. We first verify that the $\mathcal{D}_N$-similarity and block transpose operations both preserve principal minors. The determinant is multiplicative, and so principal minors are immediately preserved under $\mathcal{D}_N$-similarity, as the principal submatrix of DKD corresponding to S is equal to the product of the principal submatrices corresponding to S of the three matrices D, K, D, and so

$$\Delta_S(DKD) = \Delta_S(D)\Delta_S(K)\Delta_S(D) = \Delta_S(K).$$

For the block transpose operation, we note that the transpose leaves the determinant unchanged. By equation (2.10), the principal minors of a matrix are uniquely determined by the principal minors of the matrices corresponding to the blocks of G. As the transpose of any block also leaves the principal minors corresponding to subsets of other blocks unchanged, principal minors are preserved under $\mathcal{D}_N$-similarity.

What remains is to show that if K' satisfies $\Delta_S(K) = \Delta_S(K')$ for all $S \subset [N]$, then K' is generated by K and the above two operations. Without loss of generality, we may suppose that the shared charged sparsity graph $G = ([N], E, \epsilon)$ of K and K' is two-connected, as the general case follows from this one. We will transform K' to K by first making them agree for entries corresponding to a spanning tree and a negative edge of G, and then observing that by Lemmas 8 and 9, this property implies that all entries agree.

Let C' be an arbitrary negative simple cycle of G, $e \in E(C')$ be a negative edge in C', and T be a spanning tree of G containing the edges $E(C') \setminus e$. By applying $\mathcal{D}_N$-similarity and block transpose to K', we can produce a matrix that agrees with K for all entries corresponding to edges in $E(T) \cup e$. We perform this procedure in three steps, first by making K' agree with K for edges in $E(T)$, then for edge e, and then finally fixing any edges in $E(T)$ for which the matrices no longer agree.

First, we make K' and K agree for edges in $E(T)$. Suppose that $K'_{p,q} = -K_{p,q}$ for some $\{p, q\} \in E(T)$. Let $U \subset [N]$ be the set of vertices connected to p in the forest $T \setminus \{p, q\}$ (the removal of edge $\{p, q\}$ from T), and $\hat{D}$ be the diagonal matrix with $\hat{D}_{i,i} = +1$ if $i \in U$ and $\hat{D}_{i,i} = -1$ if $i \notin U$. The matrix $\hat{D}K'\hat{D}$ satisfies

$$\left[\hat{D}K'\hat{D}\right]_{p,q} = \hat{D}_{p,p}K'_{p,q}\hat{D}_{q,q} = -K'_{p,q} = K_{p,q},$$

and $\left[\hat{D}K'\hat{D}\right]_{i,j} = K'_{i,j}$ for any i, j either both in U or neither in U (and therefore for all edges $\{i, j\} \in E(T) \setminus \{p, q\}$). Repeating this procedure for every edge $\{p, q\} \in E(T)$ for which $K'_{p,q} = -K_{p,q}$ results in a matrix that agrees with K for all entries corresponding to edges in $E(T)$.

Next, we make our matrix and K agree for the edge e. If our matrix already agrees for edge e, then we are done with this part of the proof, and we denote the resulting matrix by $\hat{K}$. If our resulting matrix does not agree, then, by taking the transpose of this matrix, we have a new matrix that now agrees with K for all edges $E(T) \cup e$, except for negative edges in $E(T)$. By repeating the $\mathcal{D}_N$-similarity operation again on negative edges of $E(T)$, we again obtain a matrix that agrees with K on the edges of $E(T)$, but now also agrees on the

edge e, as there is an even number of negative edges in the path between the vertices of e in the tree T. We now have a matrix that agrees with K for all entries corresponding to edges in $E(T) \cup e$, and we denote this matrix by $\hat{K}$.

Finally, we aim to show that agreement on the edges $E(T) \cup e$ already implies that $\hat{K} = K$. Let $\{i, j\}$ be an arbitrary edge not in $E(T) \cup e$, and $\hat{C}$ be the simple cycle containing edge $\{i, j\}$ and the unique $i - j$ path in the tree T. By Lemmas 8 and 9, the value of $s_K(C)$ can be computed for every cycle corresponding to an incidence vector in a minimal simple cycle basis of $\mathcal{C}^+(G)$ using only principal minors. Then $s_K(\hat{C})$ can be computed using principal minors and $s_K(C')$, as the incidence vector for C' combined with a minimal positive simple cycle basis forms a basis for $\mathcal{C}(G)$. By assumption, $\Delta_S(K) = \Delta_S(\hat{K})$ for all $S \subset [N]$, and, by construction, $s_K(C') = s_{\hat{K}}(C')$. Therefore, $s_K(\hat{C}) = s_{\hat{K}}(\hat{C})$ for all $\{i, j\}$ not in $E(T) \cup e$, and so $\hat{K} = K$. □

Theorem 5. *Let $K \in \mathcal{K}$, with charged sparsity graph $G = ([N], E, \epsilon)$. Let $\ell_+ \geq 3$ be the simple cycle sparsity of $\mathcal{C}^+(G)$. Then any matrix $K' \in \mathcal{K}$ with $\Delta_S(K) = \Delta_S(K')$ for all $|S| \leq \ell_+$ also satisfies $\Delta_S(K) = \Delta_S(K')$ for all $S \subset [N]$, and there exists a matrix $\hat{K} \in \mathcal{K}$ with $\Delta_S(\hat{K}) = \Delta_S(K)$ for all $|S| < \ell_+$ but $\Delta_S(\hat{K}) \neq \Delta_S(K)$ for some S.*

Proof. The first part of theorem follows almost immediately from Lemmas 8 and 9. It suffices to consider a matrix $K \in \mathcal{K}$ with a two-connected charged sparsity graph $G = ([N], E, \epsilon)$, as the more general case follows from this one. By Lemmas 8, and 9, the quantity $s(C)$ is computable for all simple cycles C in a minimal cycle basis for $\mathcal{C}^+(G)$ (and therefore for all positive simple cycles), using only principal minors of size at most the length of the longest cycle in the basis, which in this case is the simple cycle sparsity ℓ_+ of $\mathcal{C}^+(G)$. The values $s(C)$ for positive simple cycles combined with the magnitude of the entries of K and the charge function ϵ uniquely determines all principal minors, as each term in the Laplace expansion of some $\Delta_S(K)$ corresponds to a partitioning of S into a disjoint collection of vertices, pairs corresponding to edges, and oriented versions of positive simple cycles of $E(S)$. This completes the first portion of the proof.

Next, we explicitly construct a matrix $\hat{K}$ that agrees with K in the first $\ell_+ - 1$ principal minors ($\ell_+ \geq 3$), but disagrees on a principal minor of order ℓ_+. To do so, we consider a

minimal simple cycle basis $\left\{\chi_{C_1}, ..., \chi_{C_{\nu-1}}\right\}$ of $\mathcal{C}^+(G)$, ordered so that $|C_1| \geq |C_2| \geq ... \geq |C_{\nu-1}|$. By definition, $\ell_+ = |C_1|$. Let $\hat{K}$ be a matrix satisfying

$$\hat{K}_{i,i} = K_{i,i} \text{ for } i \in [N], \qquad \hat{K}_{i,j}\hat{K}_{j,i} = K_{i,j}K_{j,i} \text{ for } i, j \in [N],$$

and

$$s_{\hat{K}}(C_i) = \begin{cases} s_K(C_i) & \text{if } i > 1 \\ -s_K(C_i) & \text{if } i = 1 \end{cases}.$$

The matrix $\hat{K}$ and K agree on all principal minors of order at most $\ell_+ - 1$, for if there was a principal minor where they disagreed, this would imply that there is a positive simple cycle of length less than ℓ_+ whose incidence vector is not in the span of $\{\chi_{C_2}, ..., \chi_{C_{\nu-1}}\}$, a contradiction. To complete the proof, it suffices to show that $\Delta_{V(C_1)}(\hat{K}) \neq \Delta_{V(C_1)}(K)$.

To do so, we consider three different cases, depending on the length of C_1. If C_1 is a three-cycle, then C_1 is an induced cycle and the result follows immediately. If C_1 is a four-cycle, $G[V(C_1)]$ may possibly have multiple positive four-cycles. However, in this case the quantity Z from Equation (2.11) is distinct for $\hat{K}$ and K, as $K \in \mathcal{K}$. Finally, we consider the general case when $|C_1| > 4$. By Lemma 8, all terms of $\Delta_{V(C)}(K)$ (and $\Delta_{V(C)}(\hat{K})$) except for Z depend only on principal minors of order at most $\ell_+ - 1$. We denote the quantity Z from Lemma 8 for K and $\hat{K}$ by Z and $\hat{Z}$ respectively. By Lemma 9, the magnitude of Z and $\hat{Z}$ depend only on principal minors of order at most four, and so $|Z| = |\hat{Z}|$. In addition, because $K \in \mathcal{K}$, $Z \neq 0$. The quantities Z and $\hat{Z}$ are equal in sign if and only if $s_{\hat{K}}(C_1) = s_K(C_1)$, and therefore $\Delta_{V(C_1)}(\hat{K}) \neq \Delta_{V(C_1)}(K)$. □

2.3.2 Efficiently Recovering a Matrix from its Principal Minors

In Subsection 2.3.1, we characterized both the set of magnitude-symmetric matrices in $\mathcal{K}$ that share a given set of principal minors, and noted that principal minors of order $\leq \ell_+$ uniquely determine principal minors of all orders, where ℓ_+ is the simple cycle sparsity of $\mathcal{C}^+(G)$ and is computable using principal minors of order one and two. In this subsection, we make use of a number of theoretical results of Subsection 2.3.1 to formally describe a

polynomial time algorithm to produce a magnitude-symmetric matrix K with prescribed principal minors. We provide a high-level description and discussion of this algorithm below, and save the description of a key subroutine for computing a positive simple cycle basis for after the overall description and proof.

An Efficient Algorithm

Our formal algorithm proceeds by completing a number of high-level tasks, which we describe below. This procedure is very similar in nature to the algorithm implicitly described and used in the proofs of Proposition 4 and Theorem 5. The main difference between the algorithmic procedure alluded to in Subsection 2.3.1 and this algorithm is the computation of a positive simple cycle basis. Unlike $\mathcal{C}(G)$, there is no known polynomial time algorithm to compute a minimal simple cycle basis for $\mathcal{C}^+(G)$, and a decision version of this problem may indeed be NP-hard. Our algorithm, which we denote by $\textsc{RecoverK}\big(\{\Delta_S\}_{S\subset[N]}\big)$, proceeds in five main steps:

Step 1: Compute $|K_{i,j}|$, $i,j \in [N]$, and the charged sparsity graph $G = ([N], E, \epsilon)$.

We recall that $K_{i,i} = \Delta_i$ and $|K_{i,j}| = |K_{j,i}| = \sqrt{|\Delta_i\Delta_j - \Delta_{i,j}|}$. The edges $\{i,j\} \in E(G)$ correspond to non-zero off-diagonal entries $|K_{i,j}| \neq 0$, and the function ϵ is given by $\epsilon_{i,j} = \text{sgn}(\Delta_i\Delta_j - \Delta_{i,j})$.

Step 2: For every block H of G, compute a simple cycle basis $\{x_1, ..., x_k\}$ of $\mathcal{C}^+(H)$.

In the proof of Proposition 3, we defined an efficient algorithm to compute a simple cycle basis of $\mathcal{C}^+(H)$ for any two-connected graph H. This algorithm makes use of the property that every two-connected graph has an open ear decomposition. Unfortunately, this algorithm has no provable guarantees on the length of the longest cycle in the basis. For this reason, we introduce an alternate efficient algorithm below that computes a simple cycle basis of $\mathcal{C}^+(H)$ consisting of cycles all of length at most $3\phi_H$, where ϕ_H is

the maximum length of a shortest cycle between any two edges in H, i.e.,

$$\phi_H := \max_{e,e' \in E(H)} \min_{\substack{\text{simple cycle } C \\ s.t.\ e,e' \in E(C)}} |C|$$

with $\phi_H := 2$ if H is acyclic. The existence of a simple cycle through any two edges of a 2-connected graph can be deduced from the existence of two vertex-disjoint paths between newly added midpoints of these two edges. The resulting simple cycle basis also maximizes the number of three- and four-cycles it contains. This is a key property which allows us to limit the number of principal minors that we query.

Step 3: For every block H of G, convert $\{x_1, ..., x_k\}$ into a simple cycle basis satisfying Properties (i)-(iii) of Lemma 6.

If there was an efficient algorithm for computing a minimal simple cycle basis for $\mathcal{C}^+(H)$, by Lemma 6, we would be done (and there would be no need for the previous step). However, there is currently no known algorithm for this, and a decision version of this problem may be NP-hard. By using the simple cycle basis from Step 2 and iteratively removing chords, we can create a basis that satisfies the same three key properties (in Lemma 6) that a minimal simple cycle basis does. In addition, the lengths of the cycles in this basis are no larger than those of Step 2, i.e., no procedure in Step 3 ever increases the length of a cycle.

The procedure for this is quite intuitive. Given a cycle C in the basis, we efficiently check that $\gamma(C)$ satisfies Properties (i)-(iii) of Lemma 6. If all properties hold, we are done, and check another cycle in the basis, until all cycles satisfy the desired properties. In each case, if a given property does not hold, then, by the proof of Lemma 6, we can efficiently compute an alternate cycle C', $|C'| < |C|$, that can replace C in the simple cycle basis for $\mathcal{C}^+(H)$, decreasing the sum of cycle lengths in the basis by at least one. Because of this, the described procedure is a polynomial time algorithm.

Step 4: For every block H of G, compute $s(C)$ for every cycle C in the basis.

This calculation relies heavily on the results of Subsection 2.3.1. The simple cycle basis for $\mathcal{C}^+(H)$ satisfies Properties (i)-(iii) of Lemma 6, and so we may apply Lemmas 8 and 9. We compute the quantities $s(C)$ iteratively based on cycle length, beginning with the shortest cycle in the basis and finishing with the longest. By the analysis at the beginning of Subsection 2.3.1, we recall that when C is a three- or four-cycle, $s(C)$ can be computed efficiently using $O(1)$ principal minors, all corresponding to subsets of $V(C)$. When $|C| > 4$, by Lemma 8, the quantity Z (defined in Lemma 8) can be computed efficiently using $O(1)$ principal minors, all corresponding to subsets of $V(C)$. By Lemma 9, the quantity $s(C)$ can be computed efficiently using Z, $s(C')$ for positive four-cycles $C' \subset G[V(C)]$, and $O(1)$ principal minors all corresponding to subsets of $V(C)$. Because our basis maximizes the number of three- and four-cycles it contains, any such four-cycle C' is either in the basis (and so $s(C')$ has already been computed) or is a linear combination of three- and four-cycles in the basis, in which case $s(C')$ can be computed using Gaussian elimination, without any additional querying of principal minors.

Step 5: Output a matrix K satisfying $\Delta_S(K) = \Delta_S$ for all $S \subset [n]$.

The procedure for producing this matrix is quite similar to the proof of Proposition 4. It suffices to fix the signs of the upper triangular entries of K, as the lower triangular entries can be computed using ϵ. For each block H, we find a negative simple cycle C (if one exists), fix a negative edge $e \in E(C)$, and extend $E(C)\backslash e$ to a spanning tree T of H, i.e., $\big[E(C)\backslash e\big] \subset E(T)$. We give the entries $K_{i,j}$, $i < j$, $\{i,j\} \in E(T) \cup e$ an arbitrary sign, and note our choice of $s_K(C)$. We extend the simple cycle basis for $\mathcal{C}^+(H)$ to a basis for $\mathcal{C}(H)$ by adding C. On the other hand, if no negative cycle exists, then we simply fix an arbitrary spanning tree T, and give the entries $K_{i,j}$, $i < j$, $\{i,j\} \in E(T)$ an arbitrary sign. In both cases, we have a basis for $\mathcal{C}(H)$ consisting of cycles C_i for

which we have computed $s(C_i)$. For each edge $\{i,j\} \in E(H)$ corresponding to an entry $K_{i,j}$ for which we have not fixed the sign of, we consider the cycle C' consisting of the edge $\{i,j\}$ and the unique $i-j$ path in T. Using Gaussian elimination, we can write C' as a sum of a subset of the cycles in our basis. As noted in Subsection 2.3.1, the quantity $s(C')$ is simply the product of the quantities $s(C_i)$ for cycles C_i in this sum.

A few comments are in order. Conditional on the analysis of Step 2, the above algorithm runs in time polynomial in N. Of course, the set $\{\Delta_S\}_{S\subset[N]}$ is not polynomial in N, but rather than take the entire set as input, we assume the existence of some querying operation, in which the value of any principal minor can be queried in polynomial time. Combining the analysis of each step, we can give the following guarantee for the RECOVERK$\big(\{\Delta_S\}_{S\subset[N]}\big)$ algorithm.

Theorem 6. *Let $\{\Delta_S\}_{S\subset[N]}$ be the principal minors of some matrix in $\mathcal{K}$. The algorithm* RECOVERK$\big(\{\Delta_S\}_{S\subset[N]}\big)$ *computes a matrix $K \in \mathcal{K}$ satisfying $\Delta_S(K) = \Delta_S$ for all $S \subset [N]$. This algorithm runs in time polynomial in N, and queries at most $O(N^2)$ principal minors, all of order at most $3\,\phi_G$, where G is the sparsity graph of $\{\Delta_S\}_{|S|\leq 2}$ and ϕ_G is the maximum of ϕ_H over all blocks H of G. In addition, there exists a matrix $K' \in \mathcal{K}$, $|K'_{i,j}| = |K_{i,j}|$, $i,j \in [N]$, such that any algorithm that computes a matrix with principal minors $\{\Delta_S(K')\}_{S\subset[N]}$ must query a principal minor of order at least ϕ_G.*

Proof. Conditional on the existence of the algorithm described in Step 2, i.e., an efficient algorithm to compute a simple cycle basis for $\mathcal{C}^+(H)$ consisting of cycles of length at most $3\phi_H$ that maximizes the number of three- and four-cycles it contains, by the above analysis we already have shown that the RECOVERK$\big(\{\Delta_S\}_{S\subset[N]}\big)$ algorithm runs in time polynomial in n and queries at most $O(N^2)$ principal minors.

To construct K', we first consider the uncharged sparsity graph $G = ([N], E)$ of K, and let $e, e' \in E(G)$ be a pair of edges for which the quantity ϕ_G is achieved (if $\phi_G = 2$, we are done). We aim to define an alternate charge function for G, show that the simple cycle sparsity of $\mathcal{C}^+(G)$ is at least ϕ_G, find a matrix K' that has this charged sparsity graph, and

then make use of Theorem 5. Consider the charge function ϵ satisfying $\epsilon(e) = \epsilon(e') = -1$, and equal to $+1$ otherwise. Any simple cycle basis for the block containing e, e' must have a simple cycle containing both edges e, e'. By definition, the length of this cycle is at least ϕ_G, and so the simple cycle sparsity of $\mathcal{C}^+(G)$ is at least ϕ_G. Next, let K' be an arbitrary matrix with $|K'_{i,j}| = |K_{i,j}|$, $i, j \in [N]$, and charged sparsity graph $G = ([N], E, \epsilon)$, with ϵ as defined above. $K \in \mathcal{K}$, and so $K' \in \mathcal{K}$, and therefore, by Theorem 5, any algorithm that computes a matrix with principal minors $\{\Delta_S(K')\}_{S \subset [N]}$ must query a principal minor of order at least $\ell_+ \geq \phi_G$. $\square$

Computing a Simple Cycle Basis with Provable Guarantees

Next, we describe an efficient algorithm to compute a simple cycle basis for $\mathcal{C}^+(H)$ consisting of cycles of length at most $3\phi_H$. Suppose we have a two-connected graph $H = ([N], E, \epsilon)$. We first compute a minimal cycle basis $\{\chi_{C_1}, ..., \chi_{C_\nu}\}$ of $\mathcal{C}(H)$, where each C_i is an induced simple cycle (as argued previously, any lexicographically minimal basis for $\mathcal{C}(H)$ consists only of induced simple cycles). Without loss of generality, suppose that C_1 is the shortest negative cycle in the basis (if no negative cycle exists, we are done). The set of incidence vectors $\chi_{C_1} + \chi_{C_i}$, $\epsilon(C_i) = -1$, combined with vectors χ_{C_j}, $\epsilon(C_j) = +1$, forms a basis for $\mathcal{C}^+(H)$, which we denote by $\mathcal{B}$. We will build a simple cycle basis for $\mathcal{C}^+(H)$ by iteratively choosing incidence vectors of the form $\chi_{C_1} + \chi_{C_i}$, $\epsilon(C_i) = -1$, replacing each with an incidence vector $\chi_{\widetilde{C}_i}$ corresponding to a positive simple cycle $\widetilde{C}_i$, and iteratively updating $\mathcal{B}$.

Let $C := C_1 + C_i$ be a positive cycle in our basis $\mathcal{B}$. If C is a simple cycle, we are done, otherwise C is the union of edge-disjoint simple cycles $F_1, ..., F_p$ for some $p > 1$. If one of these simple cycles is positive and satisfies

$$\chi_{F_j} \notin \text{span}\big\{\mathcal{B} \backslash \{\chi_{C_1} + \chi_{C_i}\}\big\},$$

we are done. Otherwise there is a negative simple cycle, without loss of generality given by F_1. We can construct a set of $p - 1$ positive cycles by adding F_1 to each negative simple

cycle, and at least one of these cycles is not in span$\{\mathcal{B}\backslash\{\chi_{C_1} + \chi_{C_i}\}\}$. We now have a positive cycle not in this span, given by the sum of two *edge-disjoint* negative simple cycles F_1 and F_j. These two cycles satisfy, by construction, $|F_1| + |F_j| \leq |C| \leq 2\ell$, where ℓ is the cycle sparsity of $\mathcal{C}(H)$. If $|V(F_1) \cap V(F_j)| > 1$, then $F_1 \cup F_j$ is two-connected, and we may compute a simple cycle basis for $\mathcal{C}^+(F_1 \cup F_j)$ using the ear decomposition approach of Proposition 3. At least one positive simple cycle in this basis satisfies our desired condition, and the length of this positive simple cycle is at most $|E(F_1 \cup F_j)| \leq 2\ell$. If $F_1 \cup F_j$ is not two-connected, then we compute a shortest cycle $\widetilde{C}$ in $E(H)$ containing both an edge in $E(F_1)$ and $E(F_j)$ (computed using Suurballe's algorithm, see [109] for details). The graph

$$H' = \big(V(F_1) \cup V(F_j) \cup V(\widetilde{C}), E(F_1) \cup E(F_j) \cup E(\widetilde{C})\big)$$

is two-connected, and we can repeat the same procedure as above for H', where, in this case, the resulting cycle is of length at most $|E(H')| \leq 2\ell + \phi_H$. The additional guarantee for maximizing the number of three- and four-cycles can be obtained easily by simply listing all positive cycles of length three and four, combining them with the computed basis, and performing a greedy algorithm. What remains is to note that $\ell \leq \phi_H$. This follows quickly from the observation that for any vertex u in any simple cycle C of a minimal cycle basis for $\mathcal{C}(H)$, there exists an edge $\{v, w\} \in E(C)$ such that C is the disjoint union of a shortest $u - v$ path, shortest $u - w$ path, and $\{v, w\}$ [51, Theorem 3].

2.3.3 An Algorithm for Principal Minors with Noise

So far, we have exclusively considered the situation in which principal minors are known (or can be queried) exactly. In applications related to this work, this is often not the case. The key application of a large part of this work is signed determinantal point processes, and the algorithmic question of learning the kernel of a signed DPP from some set of samples. Here, for the sake of readability, we focus on the non-probabilistic setting in which, given some matrix K with spectral radius $\rho(K) \leq 1$, each principal minor Δ_S can be queried/estimated

up to some absolute error term $0 < \delta < 1$, i.e., we are given a set $\{\hat{\Delta}_S\}_{S\subset[N]}$, satisfying

$$\left|\hat{\Delta}_S - \Delta_S\right| < \delta \qquad \text{for all } S \subset [N],$$

where $\{\Delta_S\}_{S\subset[N]}$ are the principal minors of some magnitude-symmetric matrix K, and asked to compute a matrix K' minimizing

$$\rho(K, K') := \min_{\substack{\hat{K} \\ \Delta_S(\hat{K})=\Delta_S(K), \\ S\subset[N]}} \text{s.t.} \quad |\hat{K} - K'|_\infty,$$

where $|K - K'|_\infty = \max_{i,j} |K_{i,j} - K'_{i,j}|$.

In this setting, we require both a separation condition for the entries of K and a stronger version of Condition (2.6). Let $\mathcal{K}_N(\alpha, \beta)$ be the set of matrices $K \in \mathbb{R}^{N\times N}$, $\rho(K) \le 1$, satisfying: $|K_{i,j}| = |K_{j,i}|$ for $i, j \in [K]$, $|K_{i,j}| > \alpha$ or $|K_{i,j}| = 0$ for all $i, j \in [N]$, and

$$\begin{aligned} &\text{If } K_{i,j}K_{j,k}K_{k,\ell}K_{\ell,i} \neq 0 \text{ for } i, j, k, \ell \in [N] \text{ distinct,} \\ &\text{then } \left||K_{i,j}K_{k,\ell}| - |K_{j,k}K_{\ell,i}|\right| > \beta, \text{ and the sum} \\ &\left|\phi_1\, K_{i,j}K_{j,k}K_{k,\ell}K_{\ell,i} + \phi_2 K_{i,j}K_{j,\ell}K_{\ell,k}K_{k,i}\right. \\ &\left. \quad + \phi_3 K_{i,k}K_{k,j}K_{j,\ell}K_{\ell,i}\right| > \beta \quad \forall\, \phi \in \{-1, 1\}^3. \end{aligned} \tag{2.12}$$

The algorithm RECOVERK$\big(\{\Delta_S\}_{S\subset[N]}\big)$, slightly modified, performs well for this more general problem. Here, we briefly describe a variant of the above algorithm that performs nearly optimally on any $K \in \mathcal{K}(\alpha, \beta)$, $\alpha, \beta > 0$, for δ sufficiently small, and quantify the relationship between α and β. The algorithm to compute a K' sufficiently close to K, which we denote by RECOVERNOISYK$\big(\{\hat{\Delta}_S\}_{S\subset[N]}; \delta\big)$, proceeds in three main steps:

Step 1: Define $|K'_{i,j}|$, $i, j \in [N]$, and the charged sparsity graph $G = ([N], E, \epsilon)$.

We define

$$K'_{i,i} = \begin{cases} \hat{\Delta}_i & \text{if} \quad |\hat{\Delta}_i| > \delta \\ 0 & \text{otherwise} \end{cases},$$

and

$$K'_{i,j}K'_{j,i} = \begin{cases} \hat{\Delta}_i\hat{\Delta}_j - \hat{\Delta}_{i,j} & \text{if} \quad |\hat{\Delta}_i\hat{\Delta}_j - \hat{\Delta}_{i,j}| > 3\delta + \delta^2 \\ 0 & \text{otherwise} \end{cases}.$$

The edges $\{i, j\} \in E(G)$ correspond to non-zero off-diagonal entries $|K'_{i,j}| \neq 0$, and the function ϵ is given by $\epsilon_{i,j} = \text{sgn}(K'_{i,j}K'_{j,i})$.

Step 2: For every block H of G, compute a simple cycle basis $\{x_1, ..., x_k\}$ of $\mathcal{C}^+(H)$ satisfying Properties (i)-(iii) of Lemma 6, and define $s_{K'}(C)$ for every cycle C in the basis.

The computation of the simple cycle basis depends only on the graph G, and so is identical to Steps 1 and 2 of the RECOVERK algorithm. The simple cycle basis for $\mathcal{C}^+(H)$ satisfies Properties (i)-(iii) of Lemma 6, and so we can apply Lemmas 8 and 9. We define the quantities $s_{K'}(C)$ iteratively based on cycle length, beginning with the shortest cycle. We begin by detailing the cases of $|C| = 3$, then $|C| = 4$, and then finally $|C| > 4$.

Case I: $|C| = 3$.

Suppose that the three-cycle $C = i\, j\, k\, i$, $i < j < k$, is in our basis. If G was the charged sparsity graph of K, then the equation

$$K_{i,j}K_{j,k}K_{k,i} = \Delta_i\Delta_j\Delta_k - \frac{1}{2}\big[\Delta_i\Delta_{j,k} + \Delta_j\Delta_{i,k} + \Delta_k\Delta_{i,j}\big] + \frac{1}{2}\Delta_{i,j,k}$$

would hold. For this reason, we define $s_{K'}(C)$ as

$$s_{K'}(C) = \epsilon_{i,k}\ \text{sgn}\big(2\hat{\Delta}_i\hat{\Delta}_j\hat{\Delta}_k - \big[\hat{\Delta}_i\hat{\Delta}_{j,k} + \hat{\Delta}_j\hat{\Delta}_{i,k} + \hat{\Delta}_k\hat{\Delta}_{i,j}\big] + \hat{\Delta}_{i,j,k}\big).$$

Case II: $|C| = 4$.

Suppose that the four-cycle $C = i\,j\,k\,\ell\,i$, with (without loss of generality) $i < j < k < \ell$, is in our basis. We note that

$$\begin{aligned}\Delta_{i,j,k,\ell} = \Delta_{i,j}\Delta_{k,\ell} + \Delta_{i,k}\Delta_{j,\ell} + \Delta_{i,\ell}\Delta_{j,k} + \Delta_i\Delta_{j,k,\ell} + \Delta_j\Delta_{i,k,\ell} + \Delta_k\Delta_{i,j,\ell} \\ + \Delta_\ell\Delta_{i,j,k} - 2\Delta_i\Delta_j\Delta_{k,\ell} - 2\Delta_i\Delta_k\Delta_{j,\ell} - 2\Delta_i\Delta_\ell\Delta_{j,k} - 2\Delta_j\Delta_k\Delta_{i,\ell} \\ - 2\Delta_j\Delta_\ell\Delta_{i,k} - 2\Delta_k\Delta_\ell\Delta_{i,j} + 6\Delta_i\Delta_j\Delta_k\Delta_\ell + Z,\end{aligned}$$

where Z is the sum of the terms in the Laplace expansion of $\Delta_{i,j,k,\ell}$ corresponding to a four-cycle of $G[\{i,j,k,\ell\}]$, and define the quantity $\hat{Z}$ analogously:

$$\begin{aligned}\hat{Z} = \hat{\Delta}_{i,j,k,\ell} - \hat{\Delta}_{i,j}\hat{\Delta}_{k,\ell} - \hat{\Delta}_{i,k}\hat{\Delta}_{j,\ell} - \hat{\Delta}_{i,\ell}\hat{\Delta}_{j,k} - \hat{\Delta}_i\hat{\Delta}_{j,k,\ell} - \hat{\Delta}_j\hat{\Delta}_{i,k,\ell} \\ - \hat{\Delta}_k\hat{\Delta}_{i,j,\ell} - \hat{\Delta}_\ell\hat{\Delta}_{i,j,k} + 2\hat{\Delta}_i\hat{\Delta}_j\hat{\Delta}_{k,\ell} + 2\hat{\Delta}_i\hat{\Delta}_k\hat{\Delta}_{j,\ell} + 2\hat{\Delta}_i\hat{\Delta}_\ell\hat{\Delta}_{j,k} \\ + 2\hat{\Delta}_j\hat{\Delta}_k\hat{\Delta}_{i,\ell} + 2\hat{\Delta}_j\hat{\Delta}_\ell\hat{\Delta}_{i,k} + 2\hat{\Delta}_k\hat{\Delta}_\ell\hat{\Delta}_{i,j} - 6\hat{\Delta}_i\hat{\Delta}_j\hat{\Delta}_k\hat{\Delta}_\ell.\end{aligned}$$

We consider two cases, depending on the number of positive four-cycles in $G[\{i,j,k,\ell\}]$. First, suppose that C is the only positive four-cycle in $G[\{i,j,k,\ell\}]$. If G was the charged sparsity graph of K, then Z would equal $-2K_{i,j}K_{j,k}K_{k,\ell}K_{\ell,i}$. For this reason, we define $s_{K'}(C)$ as

$$s_{K'}(C) = \epsilon_{i,\ell}\ \mathrm{sgn}\big(-\hat{Z}\big)$$

in this case. Next, we consider the second case, in which there is more than one positive four-cycle in $G[\{i,j,k,\ell\}]$, which actually implies that all three four-cycles in $G[\{i,j,k,\ell\}]$ are positive. If G was the charged sparsity graph of K then Z would be

given by

$$Z = -2\big(K_{i,j}K_{j,k}K_{k,\ell}K_{\ell,i} + K_{i,j}K_{j,\ell}K_{\ell,k}K_{k,i} + K_{i,k}K_{k,j}K_{j,\ell}K_{\ell,i}\big).$$

For this reason, we compute an assignment of values $\phi_1, \phi_2, \phi_3 \in \{-1, +1\}$ that minimizes (not necessarily uniquely) the quantity

$$\left|\hat{Z}/2 + \phi_1\big|\hat{K}_{i,j}\hat{K}_{j,k}\hat{K}_{k,\ell}\hat{K}_{\ell,i}\big| + \phi_2\big|\hat{K}_{i,j}\hat{K}_{j,\ell}\hat{K}_{\ell,k}\hat{K}_{k,i}\big| + \phi_3\big|\hat{K}_{i,k}\hat{K}_{k,j}\hat{K}_{j,\ell}\hat{K}_{\ell,i}\big|\right|,$$

and define $s_{K'}(C) = \epsilon_{i,\ell}\ \phi_1$ in this case.

Case III: $|C| > 4$.

Suppose that the cycle $C = i_1 \ldots i_k\, i_1$, $k > 4$, is in our basis, with vertices ordered to match the ordering of Lemma 8, and define $S = \{i_1, ..., i_k\}$. Based on the equation in Lemma 8, we define $\hat{Z}$ to equal

$$\begin{aligned}
\hat{Z} = \hat{\Delta}_S &- \hat{\Delta}_{i_1}\hat{\Delta}_{S\setminus i_1} - \hat{\Delta}_{i_k}\hat{\Delta}_{S\setminus i_k} - \big[\hat{\Delta}_{i_1,i_k} - 2\hat{\Delta}_{i_1}\hat{\Delta}_{i_k}\big]\hat{\Delta}_{S\setminus i_1,i_k} \\
&+ 2K'_{i_1,i_{k-1}}K'_{i_{k-1},i_k}K'_{i_k,i_2}K'_{i_2,i_1}\hat{\Delta}_{S\setminus i_1,i_2,i_{k-1},i_k} \\
&+ \big[\hat{\Delta}_{i_1,i_2} - \hat{\Delta}_{i_1}\hat{\Delta}_{i_2}\big]\big[\hat{\Delta}_{i_{k-1},i_k} - \hat{\Delta}_{i_{k-1}}\hat{\Delta}_{i_k}\big]\hat{\Delta}_{S\setminus i_1,i_2,i_{k-1},i_k} \\
&- \big[\hat{\Delta}_{i_1,i_{k-1}} - \hat{\Delta}_{i_1}\hat{\Delta}_{i_{k-1}}\big]\big[\hat{\Delta}_{i_2,i_k} - \hat{\Delta}_{i_2}\hat{\Delta}_{i_k}\big]\hat{\Delta}_{S\setminus i_1,i_2,i_{k-1},i_k} \\
&- \big[\hat{\Delta}_{i_1,i_2} - \hat{\Delta}_{i_1}\hat{\Delta}_{i_2}\big]\big[\hat{\Delta}_{S\setminus i_1,i_2} - \hat{\Delta}_{i_k}\hat{\Delta}_{S\setminus i_1,i_2,i_k}\big] \\
&- \big[\hat{\Delta}_{i_{k-1},i_k} - \hat{\Delta}_{i_{k-1}}\hat{\Delta}_{i_k}\big]\big[\hat{\Delta}_{S\setminus i_{k-1},i_k} - \hat{\Delta}_{i_2}\hat{\Delta}_{S\setminus i_1,i_{k-1},i_k}\big],
\end{aligned}$$

and note that the quantity

$$\text{sgn}(K'_{i_1,i_{k-1}}K'_{i_{k-1},i_k}K'_{i_k,i_2}K'_{i_2,i_1})$$

is, by construction, computable using the signs of three- and four-cycles in our basis.

Based on the equation in Lemma 9, we set

$$\operatorname{sgn}\left(2\,(-1)^{k+1}\,K'_{i_k,i_1}\prod_{j=1}^{k-1}K'_{i_j,i_{j+1}}\prod_{(a,b)\in U}\left[1-\frac{\epsilon_{i_a,i_{a+1}}K'_{i_{b-1},i_a}K'_{i_{a+1},i_b}}{K'_{i_a,i_{a+1}}K'_{i_{b-1},i_b}}\right]\right)=\operatorname{sgn}(\hat{Z}),$$

with the set U defined as in Lemma 9. From this equation, we can compute $s_{K'}(C)$, as the quantity

$$\operatorname{sgn}\left[1-\frac{\epsilon_{i_a,i_{a+1}}K'_{i_{b-1},i_a}K'_{i_{a+1},i_b}}{K'_{i_a,i_{a+1}}K'_{i_{b-1},i_b}}\right]$$

is computable using the signs of three- and four-cycles in our basis (see Subsection 2.3.1 for details). In the case of $|C|>4$, we note that $s_{K'}(C)$ was defined using $O(1)$ principal minors, all corresponding to subsets of $V(C)$, and previously computed information regarding three- and four-cycles (which requires no additional querying).

Step 3: Define K'.

We have already defined $K'_{i,i}$, $i\in[N]$, $|K'_{i,j}|$, $i,j\in[N]$, and $s_{K'}(C)$ for every cycle in a simple cycle basis. The procedure for producing this matrix is identical to Step 5 of the RECOVERK algorithm.

The RECOVERNOISYK$\big(\{\hat{\Delta}_S\}_{S\subset[N]},\,\delta\big)$ algorithm relies quite heavily on δ being small enough so that the sparsity graph of K can be recovered. In fact, we can quantify the size of δ required so that the complete signed structure of K can be recovered, and only uncertainty in the magnitude of entries remains. In the following theorem, we show that, in this regime, the RECOVERNOISYK algorithm is nearly optimal.

Theorem 7. *Let $K\in\mathcal{K}_N(\alpha,\beta)$ and $\{\hat{\Delta}_S\}_{S\subset[N]}$ satisfy $|\hat{\Delta}_S-\Delta_S(K)|<\delta<1$ for all $S\subset[N]$, for some*

$$\delta<\min\{\alpha^{3\phi_G},\beta^{3\phi_G/2},\alpha^2\}/400,$$

where G is the sparsity graph of $\{\Delta_S\}_{|S|\leq 2}$ and ϕ_G is the maximum of ϕ_H over all blocks H

of G. The RECOVERNOISYK$\big(\{\hat{\Delta}_S\}_{S\subset[N]},\, \delta\big)$ *algorithm outputs a matrix K' satisfying*

$$\rho(K, K') < 2\delta/\alpha.$$

This algorithm runs in time polynomial in N, and queries at most $O(N^2)$ principal minors, all of order at most $3\,\phi_G$.

Proof. This proof consists primarily of showing that, for δ sufficiently small, the charged sparsity graph $G = ([N], E, \epsilon)$ and the cycle signs $s(C)$ of K' and K agree, and quantifying the size of δ required to achieve this. If this holds, then the quantity $\rho(K, K')$ is simply given by $\max_{i,j} \big||K_{i,j}| - |K'_{i,j}|\big|$. We note that, because $\rho(K) \leq 1$ and $\delta < 1$,

$$\big|\Delta_{S_1}...\Delta_{S_k} - \hat{\Delta}_{S_1}...\hat{\Delta}_{S_k}\big| < (1+\delta)^k - 1 < (2^k - 1)\delta \tag{2.13}$$

for any $S_1, ..., S_k \subset [N]$, $k \in \mathbb{N}$. This is the key error estimate we will use throughout the proof.

We first consider error estimates for the magnitudes of the entries of K and K'. We have $|K_{i,i} - \hat{\Delta}_i| < \delta$ and either $|K_{i,i}| > \alpha$ or $K_{i,i} = 0$. If $K_{i,i} = 0$, then $|\hat{\Delta}_i| < \delta$, and $K_{i,i} = K'_{i,i}$. If $|K_{i,i}| > \alpha$, then $|\hat{\Delta}_i| > \delta$, as $\alpha > 2\delta$, and $K_{i,i} = \hat{\Delta}_i$. Combining these two cases, we have

$$|K_{i,i} - K'_{i,i}| < \delta \qquad \text{for all } i \in [N].$$

Next, we consider off-diagonal entries $K'_{i,j}$. By Equation (2.13),

$$\big|(\hat{\Delta}_i\hat{\Delta}_j - \hat{\Delta}_{i,j}) - K_{i,j}K_{j,i}\big| \leq [(1+\delta)^2 - 1] + \delta < 4\delta.$$

Therefore, if $K_{i,j} = 0$, then $K'_{i,j} = 0$. If $|K_{i,j}| > \alpha$, then $|K'_{i,j}| = \sqrt{|\hat{\Delta}_i\hat{\Delta}_j - \hat{\Delta}_{i,j}|}$, as $\alpha^2 > 8\delta$. Combining these two cases, we have

$$|K_{i,j}K_{j,i} - K'_{i,j}K'_{j,i}| < 4\delta \qquad \text{for all } i, j \in [N],$$

and therefore

$$\big|\,|K_{i,j}| - |K'_{i,j}|\,\big| < 2\delta/\alpha \qquad \text{for all } i,j \in [N],\ i \neq j. \tag{2.14}$$

Our above analysis implies that K and K' share the same charged sparsity graph $G = ([N], E, \epsilon)$. What remains is to show that the cycle signs $s(C)$ of K and K' agree for the positive simple cycle basis of our algorithm. Similar to the structure in Step 2 of the description of the RECOVERNOISYK algorithm, we will treat the cases of $|C| = 3$, $|C| = 4$, and $|C| > 4$ separately, and rely heavily on the associated notation defined above.

Case I: $|C| = 3$.

We have $|K_{i,j}K_{j,k}K_{k,i}| > \alpha^3$, and, by Equation (2.13), the difference between $K_{i,j}K_{j,k}K_{k,i}$ and its estimated value

$$\hat{\Delta}_i\hat{\Delta}_j\hat{\Delta}_k - \frac{1}{2}\big[\hat{\Delta}_i\hat{\Delta}_{j,k} + \hat{\Delta}_j\hat{\Delta}_{i,k} + \hat{\Delta}_k\hat{\Delta}_{i,j}\big] + \frac{1}{2}\hat{\Delta}_{i,j,k}$$

is at most

$$[(1+\delta)^3 - 1] + \frac{3}{2}[(1+\delta)^2 - 1] + \frac{1}{2}[(1+\delta) - 1] < 12\delta < \alpha^3.$$

Therefore, $s_K(C)$ equals $s_{K'}(C)$.

Case II: $|C| = 4$.

By repeated application of Equation (2.13),

$$|Z - \hat{Z}| < 6[(1+\delta)^4 - 1] + 12[(1+\delta)^3 - 1] + 7[(1+\delta)^2 - 1] + [(1+\delta) - 1] < 196\,\delta.$$

If C is the only positive four-cycle of $G[\{i,j,k,\ell\}]$, then $s_K(C)$ equals $s_{K'}(C)$, as $2\beta >$

196δ. If $G[\{i,j,k,\ell\}]$ has three positive four-cycles, then

$$Z = -2\big(K_{i,j}K_{j,k}K_{k,\ell}K_{\ell,i} + K_{i,j}K_{j,\ell}K_{\ell,k}K_{k,i} + K_{i,k}K_{k,j}K_{j,\ell}K_{\ell,i}\big),$$

and, by Condition (2.12), any two signings of the three terms of the above equation for Z are at distance at least 4β from each other. As $2\beta > 196\delta$, $s_K(C)$ equals $s_{K'}(C)$ in this case as well.

Case III: $|C| > 4$.

Using equations (2.13) and (2.14), and noting that $\rho(K) < 1$ implies $|K_{i,j}| < \sqrt{2}$, we can bound the difference between Z and $\hat{Z}$ by

$$\begin{aligned}|Z - \hat{Z}| &< [(1+\delta) - 1] + 5[(1+\delta)^2 - 1] + 8[(1+\delta)^3 - 1] + 6[(1+\delta)^4 - 1] \\ &\quad + 2[(1+\delta)^5 - 1] + 2[(\sqrt{2} + 2\delta/\alpha)^4 - 4][(1+\delta) - 1] \\ &\leq (224 + 120\delta/\alpha)\delta \leq 344\delta.\end{aligned}$$

Based on the equation of Lemma 9 for Z, the quantity Z is at least

$$\min\{\alpha^{|C|}, \beta^{|C|/2}\} > 344\delta,$$

and so $s_K(C)$ equals $s_{K'}(C)$. This completes the analysis of the final case.

This implies that $\rho(K, K')$ is given by $\max_{i,j} \big||K_{i,j}| - |K'_{i,j}|\big| < 2\delta/\alpha$. □

Chapter 3

The Spread and Bipartite Spread Conjecture

3.1 Introduction

The spread $s(M)$ of an arbitrary $n \times n$ complex matrix M is the diameter of its spectrum; that is,

$$s(M) := \max_{i,j} |\lambda_i - \lambda_j|,$$

where the maximum is taken over all pairs of eigenvalues of M. This quantity has been well-studied in general, see [33, 53, 83, 129] for details and additional references. Most notably, Johnson, Kumar, and Wolkowicz produced the lower bound

$$s(M) \geq \left| \sum_{i \neq j} m_{i,j} \right| / (n-1)$$

for normal matrices $M = (m_{i,j})$ [53, Theorem 2.1], and Mirsky produced the upper bound

$$s(M) \leq \sqrt{2 \sum_{i,j} |m_{i,j}|^2 - (2/n) \left| \sum_i m_{i,i} \right|^2}$$

for any $n \times n$ matrix M, which is tight for normal matrices with $n-2$ of its eigenvalues all equal and equal to the arithmetic mean of the other two [83, Theorem 2].

The spread of a matrix has also received interest in certain particular cases. Consider a simple undirected graph $G = (V, E)$ of order n. The adjacency matrix A of a graph G is the $n \times n$ matrix whose rows and columns are indexed by the vertices of G, with entries satisfying

$$A_{u,v} = \begin{cases} 1 & \text{if } \{u, v\} \in E(G) \\ 0 & \text{otherwise} \end{cases}.$$

This matrix is real and symmetric, and so its eigenvalues are real, and can be ordered $\lambda_1(G) \geq \lambda_2(G) \geq \cdots \geq \lambda_n(G)$. When considering the spread of the adjacency matrix A of some graph G, the spread is simply the distance between $\lambda_1(G)$ and $\lambda_n(G)$, denoted by

$$s(G) := \lambda_1(G) - \lambda_n(G).$$

In this instance, $s(G)$ is referred to as the *spread of the graph*. The spectrum of a bipartite graph is symmetric with respect to the origin, and, in particular, $\lambda_n(G) = -\lambda_1(G)$ and so $s(G) = 2\,\lambda_1(G)$.

In [48], the authors investigated a number of properties regarding the spread of a graph, determining upper and lower bounds on $s(G)$. Furthermore, they made two key conjectures. Let us denote the maximum spread over all n vertex graphs by $s(n)$, the maximum spread over all n vertex graphs of size m by $s(n, m)$, and the maximum spread over all n vertex bipartite graphs of size m by $s_b(n, m)$. Let K_k be the clique of order k and $G(n, k) := K_k \vee \overline{K_{n-k}}$ be the join of the clique K_k and the independent set $\overline{K_{n-k}}$ (i.e., the disjoint union of K_k and $\overline{K_{n-k}}$, with all possible edges between the two graphs added). The conjectures addressed in this article are as follows.

Conjecture 1 ([48], Conjecture 1.3)**.** *For any positive integer n, the graph of order n with maximum spread is $G(n, \lfloor 2n/3 \rfloor)$; that is, $s(n)$ is attained only by $G(n, \lfloor 2n/3 \rfloor)$.*

Conjecture 2 ([48], Conjecture 1.4)**.** *If G is a graph with n vertices and m edges attaining the maximum spread $s(n, m)$, and if $m \leq \lfloor n^2/4 \rfloor$, then G must be bipartite. That is, $s_b(n, m) = s(n, m)$ for all $m \leq \lfloor n^2/4 \rfloor$.*

Conjecture 1 is referred to as the Spread Conjecture, and Conjecture 2 is referred to as the Bipartite Spread Conjecture. In this chapter, we investigate both conjectures. In Section 3.2, we prove an asymptotic version of the Bipartite Spread Conjecture, and provide an infinite family of counterexamples to illustrate that our asymptotic version is as tight as possible, up to lower order terms. In Section 3.3, we provide a high-level sketch of a proof that the Spread Conjecture holds for all n sufficiently large (the full proof can be found in [12]). The exact associated results are given by Theorems 8 and 9.

Theorem 8. *There exists a constant N so that the following holds: Suppose G is a graph on $n \geq N$ vertices with maximum spread; then G is the join of a clique on $\lfloor 2n/3 \rfloor$ vertices and an independent set on $\lceil n/3 \rceil$ vertices.*

Theorem 9.

$$s(n,m) - s_b(n,m) \leq \frac{1 + 16\, m^{-3/4}}{m^{3/4}}\, s(n,m)$$

for all $n, m \in \mathbb{N}$ satisfying $m \leq \lfloor n^2/4 \rfloor$. In addition, for any $\varepsilon > 0$, there exists some n_ε such that

$$s(n,m) - s_b(n,m) \geq \frac{1 - \varepsilon}{m^{3/4}}\, s(n,m)$$

for all $n \geq n_\varepsilon$ and some $m \leq \lfloor n^2/4 \rfloor$ depending on n.

The proof of Theorem 8 is quite involved, and for this reason we provide only a sketch, illustrating the key high-level details involved in the proof, and referring some of the exact details to the main paper [12]. The general technique consists of showing that a spread-extremal graph has certain desirable properties, considering and solving an analogous problem for graph limits, and then using this result to say something about the Spread Conjecture for sufficiently large n. In comparison, the proof of Theorem 9 is surprisingly short, making use of the theory of equitable decompositions and a well-chosen class of counterexamples.

3.2 The Bipartite Spread Conjecture

In [48], the authors investigated the structure of graphs which maximize the spread over all graphs with a fixed number of vertices n and edges m, denoted by $s(n, m)$. In particular, they proved the upper bound

$$s(G) \leq \lambda_1 + \sqrt{2m - \lambda_1^2} \leq 2\sqrt{m}, \tag{3.1}$$

and noted that equality holds throughout if and only if G is the union of isolated vertices and $K_{p,q}$, for some $p + q \leq n$ satisfying $m = pq$ [48, Thm. 1.5]. This led the authors to conjecture that if G has n vertices, $m \leq \lfloor n^2/4 \rfloor$ edges, and spread $s(n, m)$, then G is bipartite [48, Conj. 1.4]. In this section, we prove a weak asymptotic form of this conjecture and provide an infinite family of counterexamples to the exact conjecture which verifies that the error in the aforementioned asymptotic result is of the correct order of magnitude. Recall that $s_b(n, m)$, $m \leq \lfloor n^2/4 \rfloor$, is the maximum spread over all bipartite graphs with n vertices and m edges. To explicitly compute the spread of certain graphs, we make use of the theory of equitable partitions. In particular, we note that if ϕ is an automorphism of G, then the quotient matrix of $A(G)$ with respect to ϕ, denoted by A_ϕ, satisfies $\Lambda(A_\phi) \subset \Lambda(A)$, and therefore $s(G)$ is at least the spread of A_ϕ (for details, see [13, Section 2.3]). Additionally, we require two propositions, one regarding the largest spectral radius of subgraphs of $K_{p,q}$ of a given size, and another regarding the largest gap between sizes which correspond to a complete bipartite graph of order at most n.

Let $K_{p,q}^m$, $0 \leq pq - m < \min\{p, q\}$, be the subgraph of $K_{p,q}$ resulting from removing $pq - m$ edges all incident to some vertex in the larger side of the bipartition (if $p = q$, the vertex can be from either set). In [78], the authors proved the following result.

Proposition 5. *If $0 \leq pq - m < \min\{p, q\}$, then $K_{p,q}^m$ maximizes λ_1 over all subgraphs of $K_{p,q}$ of size m.*

We also require estimates regarding the longest sequence of consecutive sizes $m < \lfloor n^2/4 \rfloor$ for which there does not exist a complete bipartite graph on at most n vertices and

exactly e edges. As pointed out by [4], the result follows quickly by induction. However, for completeness, we include a brief proof.

Proposition 6. *The length of the longest sequence of consecutive sizes $m < \lfloor n^2/4 \rfloor$ for which there does not exist a complete bipartite graph on at most n vertices and exactly m edges is zero for $n \leq 4$ and at most $\sqrt{2n-1} - 1$ for $n \geq 5$.*

Proof. We proceed by induction. By inspection, for every $n \leq 4$, $m \leq \lfloor n^2/4 \rfloor$, there exists a complete bipartite graph of size m and order at most n, and so the length of the longest sequence is trivially zero for $n \leq 4$. When $n = m = 5$, there is no complete bipartite graph of order at most five with exactly five edges. This is the only such instance for $n = 5$, and so the length of the longest sequence for $n = 5$ is one.

Now, suppose that the statement holds for graphs of order at most $n-1$, for some $n > 5$. We aim to show the statement for graphs of order at most n. By our inductive hypothesis, it suffices to consider only sizes $m \geq \lfloor (n-1)^2/4 \rfloor$ and complete bipartite graphs on n vertices. We have

$$\left(\frac{n}{2} + k\right)\left(\frac{n}{2} - k\right) \geq \frac{(n-1)^2}{4} \qquad \text{for} \quad |k| \leq \frac{\sqrt{2n-1}}{2}.$$

When $1 \leq k \leq \sqrt{2n-1}/2$, the difference between the sizes of $K_{n/2+k-1,n/2-k+1}$ and $K_{n/2+k,n/2-k}$ is at most

$$\left|E\left(K_{\frac{n}{2}+k-1,\frac{n}{2}-k+1}\right)\right| - \left|E\left(K_{n/2+k,n/2-k}\right)\right| = 2k - 1 \leq \sqrt{2n-1} - 1.$$

Let k^* be the largest value of k satisfying $k \leq \sqrt{2n-1}/2$ and $n/2 + k \in \mathbb{N}$. Then

$$\begin{aligned}\left|E\left(K_{\frac{n}{2}+k^*,\frac{n}{2}-k^*}\right)\right| &< \left(\frac{n}{2} + \frac{\sqrt{2n-1}}{2} - 1\right)\left(\frac{n}{2} - \frac{\sqrt{2n-1}}{2} + 1\right) \\ &= \sqrt{2n-1} + \frac{(n-1)^2}{4} - 1,\end{aligned}$$

and the difference between the sizes of $K_{n/2+k^*,n/2-k^*}$ and $K_{\lceil\frac{n-1}{2}\rceil,\lfloor\frac{n-1}{2}\rfloor}$ is at most

$$\left|E\left(K_{\frac{n}{2}+k^*,\frac{n}{2}-k^*}\right)\right| - \left|E\left(K_{\lceil\frac{n-1}{2}\rceil,\lfloor\frac{n-1}{2}\rfloor}\right)\right| < \sqrt{2n-1} + \frac{(n-1)^2}{4} - \left\lfloor\frac{(n-1)^2}{4}\right\rfloor - 1$$
$$< \sqrt{2n-1}.$$

Combining these two estimates completes our inductive step, and the proof. □

We are now prepared to prove an asymptotic version of [48, Conjecture 1.4], and provide an infinite class of counterexamples that illustrates that the asymptotic version under consideration is the tightest version of this conjecture possible.

Theorem 10.

$$s(n,m) - s_b(n,m) \leq \frac{1+16\,m^{-3/4}}{m^{3/4}}\,s(n,m)$$

for all $n, m \in \mathbb{N}$ *satisfying* $m \leq \lfloor n^2/4\rfloor$. *In addition, for any* $\epsilon > 0$, *there exists some* n_ϵ *such that*

$$s(n,m) - s_b(n,m) \geq \frac{1-\epsilon}{m^{3/4}}\,s(n,m)$$

for all $n \geq n_\epsilon$ *and some* $m \leq \lfloor n^2/4\rfloor$ *depending on* n.

Proof. The main idea of the proof is as follows. To obtain an upper bound on $s(n,m) - s_b(n,m)$, we upper bound $s(n,m)$ by $2\sqrt{m}$ (using Inequality 3.1) and lower bound $s_b(n,m)$ by the spread of some specific bipartite graph. To obtain a lower bound on $s(n,m)-s_b(n,m)$ for a specific n and m, we explicitly compute $s_b(n,m)$ using Proposition 5, and lower bound $s(n,m)$ by the spread of some specific non-bipartite graph.

First, we analyze the spread of $K^m_{p,q}$, $0 < pq - m < q \leq p$, a quantity that will be used in the proof of both the upper and lower bound. Let us denote the vertices in the bipartition of $K^m_{p,q}$ by $u_1, ..., u_p$ and $v_1, ..., v_q$, and suppose without loss of generality that u_1 is not adjacent to $v_1, ..., v_{pq-m}$. Then

$$\phi = (u_1)(u_2, ..., u_p)(v_1, ..., v_{pq-m})(v_{pq-m+1}, ..., v_q)$$

is an automorphism of $K_{p,q}^m$. The corresponding quotient matrix is given by

$$A_\phi = \begin{pmatrix} 0 & 0 & 0 & m-(p-1)q \\ 0 & 0 & pq-m & m-(p-1)q \\ 0 & p-1 & 0 & 0 \\ 1 & p-1 & 0 & 0 \end{pmatrix},$$

has characteristic polynomial

$$Q(p,q,m) = \det[A_\phi - \lambda I] = \lambda^4 - m\lambda^2 + (p-1)(m-(p-1)q)(pq-m),$$

and, therefore,

$$s\left(K_{p,q}^m\right) \geq 2\left(\frac{m+\sqrt{m^2-4(p-1)(m-(p-1)q)(pq-m)}}{2}\right)^{1/2}. \tag{3.2}$$

For $pq = \Omega(n^2)$ and n sufficiently large, this upper bound is actually an equality, as $A(K_{p,q}^m)$ is a perturbation of the adjacency matrix of a complete bipartite graph with $\Omega(n)$ bipartite sets by an $O(\sqrt{n})$ norm matrix. For the upper bound, we only require the inequality, but for the lower bound, we assume n is large enough so that this is indeed an equality.

Next, we prove the upper bound. For some fixed n and $m \leq \lfloor n^2/4 \rfloor$, let $m = pq - r$, where $p, q, r \in \mathbb{N}$, $p+q \leq n$, and r is as small as possible. If $r = 0$, then by [48, Thm. 1.5] (described above), $s(n,m) = s_b(n,m)$ and we are done. Otherwise, we note that $0 < r < \min\{p,q\}$, and so Inequality 3.2 is applicable (in fact, by Proposition 6, $r = O(\sqrt{n})$). Using the upper bound $s(n,m) \leq 2\sqrt{m}$ and Inequality 3.2, we have

$$\frac{s(n,pq-r) - s\left(K_{p,q}^m\right)}{s(n,pq-r)} \leq 1 - \left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - \frac{4(p-1)(q-r)r}{(pq-r)^2}}\right)^{1/2}. \tag{3.3}$$

To upper bound r, we use Proposition 6 with $n' = \lceil 2\sqrt{m} \rceil \leq n$ and m. This implies that

$$r \leq \sqrt{2\lceil 2\sqrt{m}\rceil - 1} - 1 < \sqrt{2(2\sqrt{m}+1) - 1} - 1 = \sqrt{4\sqrt{m}+1} - 1 \leq 2m^{1/4}.$$

Recall that $\sqrt{1-x} \geq 1 - x/2 - x^2/2$ for all $x \in [0,1]$, and so

$$\begin{aligned} 1 - \left(\tfrac{1}{2} + \tfrac{1}{2}\sqrt{1-x}\right)^{1/2} &\leq 1 - \left(\tfrac{1}{2} + \tfrac{1}{2}(1 - \tfrac{1}{2}x - \tfrac{1}{2}x^2)\right)^{1/2} = 1 - \left(1 - \tfrac{1}{4}(x + x^2)\right) \\ &\leq 1 - \left(1 - \tfrac{1}{8}(x + x^2) - \tfrac{1}{32}(x + x^2)^2\right) \\ &\leq \tfrac{1}{8}x + \tfrac{1}{4}x^2 \end{aligned}$$

for $x \in [0,1]$. To simplify Inequality 3.3, we observe that

$$\frac{4(p-1)(q-r)r}{(pq-r)^2} \leq \frac{4r}{m} \leq \frac{8}{m^{3/4}}.$$

Therefore,

$$\frac{s(n, pq-r) - s\left(K^m_{p,q}\right)}{s(n, pq-r)} \leq \frac{1}{m^{3/4}} + \frac{16}{m^{3/2}}.$$

This completes the proof of the upper bound.

Finally, we proceed with the proof of the lower bound. Let us fix some $0 < \epsilon < 1$, and consider some arbitrarily large n. Let $m = (n/2+k)(n/2-k)+1$, where k is the smallest number satisfying $n/2 + k \in \mathbb{N}$ and $\hat{\epsilon} := 1 - 2k^2/n < \epsilon/2$ (here we require $n = \Omega(1/\epsilon^2)$). Denote the vertices in the bipartition of $K_{n/2+k,n/2-k}$ by $u_1, ..., u_{n/2+k}$ and $v_1, ..., v_{n/2-k}$, and consider the graph $K^+_{n/2+k,n/2-k} := K_{n/2+k,n/2-k} \cup \{(v_1, v_2)\}$ resulting from adding one edge to $K_{n/2+k,n/2-k}$ between two vertices in the smaller side of the bipartition. Then

$$\phi = (u_1, ..., u_{n/2+k})(v_1, v_2)(v_3, ..., v_{n/2-k})$$

is an automorphism of $K^+_{n/2+k,n/2-k}$, and

$$A_\phi = \begin{pmatrix} 0 & 2 & n/2-k-2 \\ n/2+k & 1 & 0 \\ n/2+k & 0 & 0 \end{pmatrix}$$

has characteristic polynomial

$$\begin{aligned}\det[A_\phi - \lambda I] &= -\lambda^3 + \lambda^2 + \left(n^2/4 - k^2\right)\lambda - (n/2+k)(n/2-k-2) \\ &= -\lambda^3 + \lambda^2 + \left(\frac{n^2}{4} - \frac{(1-\hat{\epsilon})n}{2}\right)\lambda - \left(\frac{n^2}{4} - \frac{(3-\hat{\epsilon})n}{2} - \sqrt{2(1-\hat{\epsilon})n}\right).\end{aligned}$$

By matching higher order terms, we obtain

$$\lambda_{max}(A_\phi) = \frac{n}{2} - \frac{1-\hat{\epsilon}}{2} + \frac{(8-(1-\hat{\epsilon})^2)}{4n} + o(1/n),$$

$$\lambda_{min}(A_\phi) = -\frac{n}{2} + \frac{1-\hat{\epsilon}}{2} + \frac{(8+(1-\hat{\epsilon})^2)}{4n} + o(1/n),$$

and

$$s(K^+_{n/2+k,n/2-k}) \geq n - (1-\hat{\epsilon}) - \frac{(1-\hat{\epsilon})^2}{2n} + o(1/n).$$

Next, we aim to compute $s_b(n,m)$, $m = (n/2+k)(n/2-k)+1$. By Proposition 5, $s_b(n,m)$ is equal to the maximum of $s(K^m_{n/2+\ell,n/2-\ell})$ over all $\ell \in [0, k-1]$, $k-\ell \in \mathbb{N}$. As previously noted, for n sufficiently large, the quantity $s(K^m_{n/2+\ell,n/2-\ell})$ is given exactly by Equation (3.2), and so the optimal choice of ℓ minimizes

$$\begin{aligned}f(\ell) &:= (n/2+\ell-1)(k^2-\ell^2-1)(n/2-\ell-(k^2-\ell^2-1)) \\ &= (n/2+\ell)\big((1-\hat{\epsilon})n/2 - \ell^2\big)\big(\hat{\epsilon}n/2 + \ell^2 - \ell\big) + O(n^2).\end{aligned}$$

We have

$$f(k-1) = (n/2+k-2)(2k-2)(n/2-3k+3),$$

and if $\ell \leq \frac{4}{5}k$, then $f(\ell) = \Omega(n^3)$. Therefore the minimizing ℓ is in $[\frac{4}{5}k, k]$. The derivative of $f(\ell)$ is given by

$$\begin{aligned}f'(\ell) = &(k^2-\ell^2-1)(n/2-\ell-k^2+\ell^2+1) \\ &- 2\ell(n/2+\ell-1)(n/2-\ell-k^2+\ell^2+1) \\ &+ (2\ell-1)(n/2+\ell-1)(k^2-\ell^2-1).\end{aligned}$$

For $\ell \in [\frac{4}{5}k, k]$,

$$\begin{aligned} f'(\ell) &\leq \frac{n(k^2-\ell^2)}{2} - \ell n(n/2-\ell-k^2+\ell^2) + 2\ell(n/2+\ell)(k^2-\ell^2) \\ &\leq \frac{9k^2n}{50} - \tfrac{4}{5}kn(n/2-k-\tfrac{9}{25}k^2) + \tfrac{18}{25}(n/2+k)k^3 \\ &= \frac{81k^3n}{125} - \frac{2kn^2}{5} + O(n^2) \\ &= kn^2\left(\frac{81(1-\hat{\epsilon})}{250} - \frac{2}{5}\right) + O(n^2) < 0 \end{aligned}$$

for sufficiently large n. This implies that the optimal choice is $\ell = k-1$, and $s_b(n,m) = s(K^m_{n/2+k-1,n/2-k+1})$. The characteristic polynomial $Q(n/2+k-1, n/2-k+1, n^2/4-k^2+1)$ equals

$$\lambda^4 - \left(n^2/4 - k^2 + 1\right)\lambda^2 + 2(n/2+k-2)(n/2-3k+3)(k-1).$$

By matching higher order terms, the extreme root of Q is given by

$$\lambda = \frac{n}{2} - \frac{1-\hat{\epsilon}}{2} - \sqrt{\frac{2(1-\hat{\epsilon})}{n}} + \frac{27-14\hat{\epsilon}-\hat{\epsilon}^2}{4n} + o(1/n),$$

and so

$$s_b(n,m) = n - (1-\hat{\epsilon}) - 2\sqrt{\frac{2(1-\hat{\epsilon})}{n}} + \frac{27-14\hat{\epsilon}-\hat{\epsilon}^2}{2n} + o(1/n),$$

and

$$\begin{aligned} \frac{s(n,m)-s_b(n,m)}{s(n,m)} &\geq \frac{2^{3/2}(1-\hat{\epsilon})^{1/2}}{n^{3/2}} - \frac{14-8\hat{\epsilon}}{n^2} + o(1/n^2) \\ &= \frac{(1-\hat{\epsilon})^{1/2}}{m^{3/4}} + \frac{(1-\hat{\epsilon})^{1/2}}{(n/2)^{3/2}}\left[1 - \frac{(n/2)^{3/2}}{m^{3/4}}\right] - \frac{14-8\hat{\epsilon}}{n^2} + o(1/n^2) \\ &\geq \frac{1-\epsilon/2}{m^{3/4}} + o(1/m^{3/4}). \end{aligned}$$

This completes the proof. □

3.3 A Sketch for the Spread Conjecture

Here, we provide a concise, high-level description of an asymptotic proof of the Spread Conjecture, proving only the key graph-theoretic structural results for spread-extremal graphs. The full proof itself is quite involved (and long), making use of interval arithmetic and a number of fairly complicated symbolic calculations, but conceptually, is quite intuitive. For the full proof, we refer the reader to [12]. The proof consists of four main steps:

Step 1: Graph-Theoretic Results

In Subsection 3.3.1, we observe a number of important structural properties of any graph that maximizes the spread for a given order n. In particular, in Theorem 11, we show that

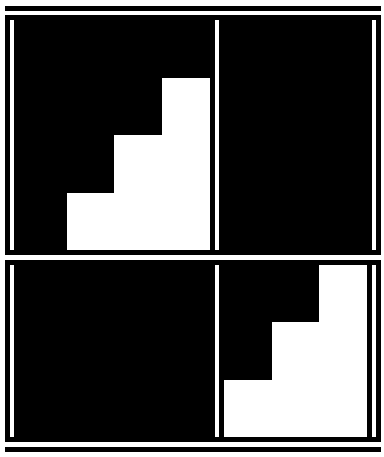
- any graph that maximizes spread is be the join of two threshold graphs, each with order linear in n,
- the unit eigenvectors $\mathbf{x}$ and $\mathbf{z}$ corresponding to $\lambda_1(G)$ and $\lambda_n(G)$ have infinity norms of order $n^{-1/2}$,
- the quantities $\lambda_1 \mathbf{x}_u^2 - \lambda_n \mathbf{z}_u^2$, $u \in V$, are all nearly equal, up to a term of order n^{-1}.

This last structural property serves as the key backbone of our proof. In addition, we note that, by a tensor argument (replacing vertices by independent sets $\overline{K_t}$ and edges by bicliques $K_{t,t}$), an asymptotic upper bound for $s(n)$ implies a bound for all n.

Step 2: Graph Limits, Averaging, and a Finite-Dimensional Eigenvalue Problem

For this step, we provide only a brief overview, and refer the reader to [12] for additional details. In this step, we can make use of graph limits to understand how spread-extremal graphs behave as n tends to infinity. By proving a continuous version of Theorem 11 for graphons, and using an averaging argument inspired by [112], we can show that

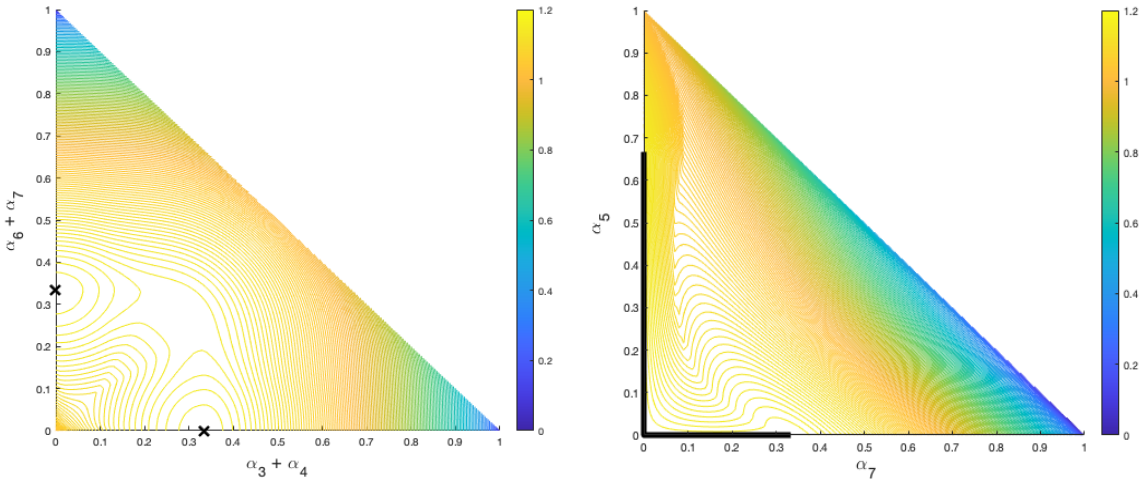
the spread-extremal graphon takes the form of a step-graphon with a fixed structure of symmetric seven by seven blocks, illustrated below.



The lengths $\alpha = (\alpha_1, ..., \alpha_7)$, $\alpha^T \mathbf{1} = 1$, of each row/column in the spread-optimal step-graphon is unknown. For any choice of lengths α, we can associate a seven by seven matrix whose spread is identical to that of the associated step-graphon pictured above. Let B be the seven by seven matrix with $B_{i,j}$ equal to the value of the above step-graphon on block i, j, and $D = \text{diag}(\alpha_1, ..., \alpha_7)$ be a diagonal matrix with α on the diagonal. Then the matrix $D^{1/2}BD^{1/2}$, given by

$$\text{diag}(\alpha_1, ..., \alpha_7)^{1/2} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \text{diag}(\alpha_1, ..., \alpha_7)^{1/2}$$

has spread equal to the spread of the associated step-graphon. This process has reduced the graph limit version of our problem to a finite-dimensional eigenvalue optimization problem, which we can endow with additional structural constraints resulting from the graphon version of Theorem 11.



(a) $\alpha_i \neq 0$ for all i

(b) $\alpha_2 = \alpha_3 = \alpha_4 = 0$

Figure 3-1: Contour plots of the spread for some choices of α. Each point (x, y) of Plot (a) illustrates the maximum spread over all choices of α satisfying $\alpha_3 + \alpha_4 = x$ and $\alpha_6 + \alpha_7 = y$ (and therefore, $\alpha_1 + \alpha_2 + \alpha_5 = 1 - x - y$) on a grid of step size $1/100$. Each point (x, y) of Plot (b) illustrates the maximum spread over all choices of α satisfying $\alpha_2 = \alpha_3 = \alpha_4 = 0$, $\alpha_5 = y$, and $\alpha_7 = x$ on a grid of step size $1/100$. The maximum spread of Plot (a) is achieved at the black x, and implies that, without loss of generality, $\alpha_3 + \alpha_4 = 0$, and therefore $\alpha_2 = 0$ (indices α_1 and α_2 can be combined when $\alpha_3 + \alpha_4 = 0$). Plot (b) treats this case when $\alpha_2 = \alpha_3 = \alpha_4 = 0$, and the maximum spread is achieved on the black line. This implies that either $\alpha_5 = 0$ or $\alpha_7 = 0$. In both cases, this reduces to the block two by two case $\alpha_1, \alpha_7 \neq 0$ (or, if $\alpha_7 = 0$, then $\alpha_1, \alpha_6 \neq 0$).

Step 3: Computer-Assisted Proof of a Finite-Dimensional Eigenvalue Problem

Using a computer-assisted proof, we can show that the optimizing choice of α_i, $i = 1, ..., 7$, is, without loss of generality, given by $\alpha_1 = 2/3$, $\alpha_6 = 1/3$, and all other $\alpha_i = 0$. This is exactly the limit of the conjectured spread optimal graph as n tends to infinity. The proof of this fact is extremely technical. Though not a proof, in Figure 1 we provide intuitive visual justification that this result is true. In this figure, we provide contour plots resulting from numerical computations of the spread of the above matrix for various values of α. The numerical results suggest that the $2/3 - 1/3$ two by two block step-graphon is indeed optimal. See Figure 1 and the associated caption for details. The

actual proof of this fact consists of three parts. First, we can reduce the possible choices of non-zero α_i from 2^7 to 17 different cases by removing different representations of the same matrix. Second, using eigenvalue equations, the graph limit version of $\lambda_1 \mathbf{x}_u^2 - \lambda_n \mathbf{z}_u^2$ all nearly equal, and interval arithmetic, we can prove that, of the 17 cases, only the cases $\alpha_1, \alpha_7 \neq 0$ or $\alpha_4, \alpha_5, \alpha_7 \neq 0$ need to be considered. Finally, using basic results from the theory of cubic polynomials and computer-assisted symbolic calculations, we can restrict ourselves to the case $\alpha_1, \alpha_7 \neq 0$. This case corresponds to a quadratic polynomial and can easily be shown to be maximized by $\alpha_1 = 2/3$, $\alpha_7 = 1/3$. We refer the reader to [118] for the code (based on [95]) and output of the interval arithmetic algorithm (the key portion of this step), and [12] for additional details regarding this step.

Step 4: From Graph Limits to an Asymptotic Proof of the Spread Conjecture

We can convert our result for the spread-optimal graph limit to a statement for graphs. This process consists of showing that any spread optimal graph takes the form $(K_{n_1} \cup \overline{K_{n_2}}) \vee \overline{K_{n_3}}$ for $n_1 = (2/3 + o(1))n$, $n_2 = o(n)$, and $n_3 = (1/3 + o(1))n$, i.e. any spread optimal graph is equal up to a set of $o(n)$ vertices to the conjectured optimal graph $K_{\lfloor 2n/3 \rfloor} \vee \overline{K_{\lceil n/3 \rceil}}$, and then showing that, for n sufficiently large, the spread of $(K_{n_1} \cup \overline{K_{n_2}}) \vee \overline{K_{n_3}}$, $n_1 + n_2 + n_3 = n$, is maximized when $n_2 = 0$. This step is fairly straightforward, and we refer the reader to [12] for details.

3.3.1 Properties of Spread-Extremal Graphs

In this subsection, we consider properties of graphs that maximize the quantity $\lambda_1(G) - c\,\lambda_n(G)$ over all graphs of order n, for some fixed $c > 0$. When $c = 1$, this is exactly the class of spread-extremal graphs. Let $\mathbf{x}$ be the unit eigenvector of λ_1 and $\mathbf{z}$ be the unit eigenvector of λ_n. As noted in [48] (for the case of $c = 1$), we have

$$\lambda_1 - c\,\lambda_n = \sum_{u \sim v} \mathbf{x}_u \mathbf{x}_v - c\,\mathbf{z}_u \mathbf{z}_v,$$

and if a graph G maximizes $\lambda_1 - c\,\lambda_n$ over all n vertex graphs, then $u \sim v$, $u \neq v$, if $\mathbf{x}_u\mathbf{x}_v - c\,\mathbf{z}_u\mathbf{z}_v > 0$ and $u \not\sim v$ if $\mathbf{x}_u\mathbf{x}_v - c\,\mathbf{z}_u\mathbf{z}_v < 0$. We first produce some weak lower bounds for the maximum of $\lambda_1 - c\,\lambda_n$ over all n vertex graphs.

Proposition 7.

$$\max_G \lambda_1(G) - c\,\lambda_n(G) \geq n\big[1 + \min\{1, c^2\}/50\big]$$

for all $c > 0$ *and* $n \geq 100\max\{1, c^{-2}\}$.

Proof. We consider two different graphs, depending on the choice of c. When $c \geq 5/4$, $n \geq 100$, we consider the bicliques $K_{\lfloor n/2\rfloor,\lceil n/2\rceil}$ and note that

$$\lambda_1(K_{\lfloor n/2\rfloor,\lceil n/2\rceil}) - c\,\lambda_n(K_{\lfloor n/2\rfloor,\lceil n/2\rceil}) \geq \frac{(c+1)(n-1)}{2} > n[1 + 1/50].$$

When $c \leq 5/4$, we consider the graph $G(n,k) = K_k \vee \overline{K_{n-k}}$, $k > 0$, with characteristic polynomial $\lambda^{n-k-1}(\lambda+1)^{k-1}\big(\lambda^2 - (k-1)\lambda - k(n-k)\big)$, and note that

$$\lambda_1(G(n,k)) - c\,\lambda_n(G(n,k)) = \frac{(1-c)(k-1)}{2} + \frac{1+c}{2}\sqrt{(k-1)^2 + 4(n-k)k}.$$

Setting $\hat{k} = \lceil (1-c/3)n\rceil + 1$ and using the inequalities $(1-c/3)n+1 \leq \hat{k} < (1-c/3)n+2$, we have

$$\frac{(1-c)(\hat{k}-1)}{2} \geq \frac{(1-c)(1-\frac{c}{3})n}{2} = n\left[\frac{1}{2} - \frac{2c}{3} + \frac{c^2}{6}\right],$$

and

$$\begin{aligned}(\hat{k}-1)^2 + 4(n-\hat{k})\hat{k} &\geq \left(1-\frac{c}{3}\right)^2 n^2 + 4\left(\frac{cn}{3} - 2\right)\left(n - \frac{cn}{3} + 1\right)\\ &= n^2\left[1 + \frac{2c}{3} - \frac{c^2}{3} + \frac{4c-8}{n} - \frac{8}{n^2}\right] \geq 1 + \frac{2c}{3} - \frac{c^2}{2}\end{aligned}$$

for $n \geq 100/c^2$. Therefore,

$$\lambda_1(G(n,\hat{k})) - c\,\lambda_n(G(n,\hat{k})) \geq n\left[\frac{1}{2} - \frac{2c}{3} + \frac{c^2}{6} + \frac{1+c}{2}\sqrt{1 + \frac{2c}{3} - \frac{c^2}{2}}\right].$$

Using the inequality $\sqrt{1+x} \geq 1 + x/2 - x^2/8$ for all $x \in [0,1]$,

$$\sqrt{1 + \frac{2c}{3} - \frac{c^2}{2}} \geq 1 + \left(\frac{c}{3} - \frac{c^2}{4}\right) - \frac{1}{8}\left(\frac{2c}{3} - \frac{c^2}{2}\right)^2,$$

and, after a lengthy calculation, we obtain

$$\begin{aligned}\lambda_1(G(n,\hat{k})) - c\,\lambda_n(G(n,\hat{k})) &\geq n\left[1 + \frac{13c^2}{72} - \frac{c^3}{9} + \frac{5c^4}{192} - \frac{c^5}{64}\right] \\ &\geq n\left[1 + c^2/50\right].\end{aligned}$$

Combining these two estimates completes the proof. □

This implies that for n sufficiently large and c fixed, any extremal graph must satisfy $|\lambda_n| = \Omega(n)$. In the following theorem, we make a number of additional observations regarding the structure of these extremal graphs. Similar results for the specific case of $c = 1$ can be found in the full paper [12].

Theorem 11. *Let $G = (V, E)$, $n \geq 3$, maximize the quantity $\lambda_1(G) - c\,\lambda_n(G)$, for some $c > 0$, over all n vertex graphs, and $\mathbf{x}$ and $\mathbf{z}$ be unit eigenvectors corresponding to λ_1 and λ_n, respectively. Then*

1. *$G = G[P] \vee G[N]$, where $P = \{u \in V \mid \mathbf{z}_u \geq 0\}$ and $N = V \backslash P$,*
2. *$\mathbf{x}_u\mathbf{x}_v - c\,\mathbf{z}_u\mathbf{z}_v \neq 0$ for all $u \neq v$,*
3. *$G[P]$ and $G[N]$ are threshold graphs, i.e., there exists an ordering $u_1, ..., u_{|P|}$ of the vertices of $G[P]$ (and $G[N]$) such that if $u_i \sim u_j$, $i < j$, then $u_j \sim u_k$, $k \neq i$, implies that $u_i \sim u_k$,*
4. *$|P|, |N| > \min\{1, c^4\}\, n/2500$ for $n \geq 100 \max\{1, c^{-2}\}$,*
5. *$\|\mathbf{x}\|_\infty \leq 6/n^{1/2}$ and $\|\mathbf{z}\|_\infty \leq 10^3 \max\{1, c^{-3}\}/n^{1/2}$ for $n \geq 100 \max\{1, c^{-2}\}$,*
6. *$\left|n(\lambda_1\mathbf{x}_u^2 - c\,\lambda_n\mathbf{z}_u^2) - (\lambda_1 - c\,\lambda_n)\right| \leq 5 \cdot 10^{12} \max\{c, c^{-12}\}$ for all $u \in V$ for $n \geq 4 \cdot 10^6 \max\{1, c^{-6}\}$.*

Proof. Property 1 follows immediately, as $\mathbf{x}_u\mathbf{x}_v - c\,\mathbf{z}_u\mathbf{z}_v > 0$ for all $u \in P$, $v \in N$. To prove Property 2, we proceed by contradiction. Suppose that this is not the case, that $\mathbf{x}_{u_1}\mathbf{x}_{u_2} - c\,\mathbf{z}_{u_1}\mathbf{z}_{u_2} = 0$ for some $u_1 \neq u_2$. Let G' be the graph resulting from either adding edge $\{u_1, u_2\}$ to G (if $\{u_1, u_2\} \notin E(G)$) or removing edge $\{u_1, u_2\}$ from G (if $\{u_1, u_2\} \in E(G)$). Then,

$$\lambda_1(G) - c\,\lambda_n(G) = \sum_{\{u,v\}\in E(G')} \mathbf{x}_u\mathbf{x}_v - c\,\mathbf{z}_u\mathbf{z}_v \leq \lambda_1(G') - c\,\lambda_n(G'),$$

and so, by the optimality of G, $\mathbf{x}$ and $\mathbf{z}$ are also eigenvectors of the adjacency matrix of G'. This implies that

$$\left|(\lambda_1(G) - \lambda_1(G'))\mathbf{x}_{u_1}\right| = \left|\mathbf{x}_{u_2}\right| \quad \text{and} \quad (\lambda_1(G) - \lambda_1(G'))\mathbf{x}_v = 0$$

for any $v \neq u_1, u_2$. If $n \geq 3$, then there exists some v such that $\mathbf{x}_v = 0$, and so, by the Perron-Frobenius theorem, G is disconnected. By Property 1, either $P = \emptyset$ or $N = \emptyset$, and so $\lambda_n \geq 0$, a contradiction, as the trace of A is zero. This completes the proof of Property 2. To show Property 3, we simply order the vertices of P so that the quantity $\mathbf{z}_{u_i}/\mathbf{x}_{u_i}$ is non-decreasing in i. If $u_i \sim u_j$, $i < j$, and $u_j \sim u_k$, without loss of generality with $i < k$, then

$$\mathbf{x}_{u_i}\mathbf{x}_{u_k} - c\,\mathbf{z}_{u_i}\mathbf{z}_{u_k} \geq \mathbf{x}_{u_i}\mathbf{x}_{u_k} - c\,\frac{\mathbf{x}_{u_i}}{\mathbf{x}_{u_j}}\mathbf{z}_{u_j}\mathbf{z}_{u_k} = \frac{\mathbf{x}_{u_i}}{\mathbf{x}_{u_j}}(\mathbf{x}_{u_j}\mathbf{x}_{u_k} - c\,\mathbf{z}_{u_j}\mathbf{z}_{u_k}) > 0.$$

To show Properties 4-6, we assume $n \geq 100\max\{1, c^{-2}\}$, and so, by Proposition 7, $\lambda_n(G) < -\min\{1, c^2\}\,n/50$. We begin with Property 4. Suppose, to the contrary, that $|P| \leq \min\{1, c^4\}\,n/2500$, and let v maximize $|\mathbf{z}_v|$. If $v \in N$, then

$$\lambda_n = \sum_{u\sim v} \frac{\mathbf{z}_u}{\mathbf{z}_v} \geq \sum_{u\in P} \frac{\mathbf{z}_u}{\mathbf{z}_v} \geq -|P| \geq -\min\{1, c^4\}\,n/2500,$$

a contradiction to $\lambda_n(G) < -\min\{1, c^2\}\, n/50$. If $v \in P$, then

$$\lambda_n^2 = \sum_{u \sim v} \frac{\lambda_n \mathbf{z}_u}{\mathbf{z}_v} = \sum_{u \sim v} \sum_{w \sim u} \frac{\mathbf{z}_w}{\mathbf{z}_v} \leq \sum_{u \sim v} \sum_{\substack{w \in P, \\ w \sim u}} \frac{\mathbf{z}_w}{\mathbf{z}_v} \leq n\,|P| \leq \min\{1, c^4\}\, n^2/2500,$$

again, a contradiction. This completes the proof of Property 4.

We now prove Property 5. Suppose that $\hat{u}$ maximizes $|\mathbf{x}_u|$ and $\hat{v}$ maximizes $|\mathbf{z}_v|$, and, without loss of generality, $\hat{v} \in N$. Let

$$A = \left\{ w \,\middle|\, \mathbf{x}_w > \frac{\mathbf{x}_{\hat{u}}}{3} \right\} \quad \text{and} \quad B = \left\{ w \,\middle|\, \mathbf{z}_w > \min\{1, c^2\} \frac{|\mathbf{z}_{\hat{v}}|}{100} \right\}.$$

We have

$$\lambda_1 \mathbf{x}_{\hat{u}} = \sum_{v \sim \hat{u}} \mathbf{x}_v \leq \mathbf{x}_{\hat{u}} \big(|A| + (n - |A|)/3 \big)$$

and $\lambda_1 \geq n/2$, and so $|A| \geq n/4$. In addition,

$$1 \;\geq\; \sum_{v \in A} \mathbf{x}_v^2 \;\geq\; |A| \frac{\|\mathbf{x}\|_\infty^2}{9} \;\geq\; \frac{n\,\|\mathbf{x}\|_\infty^2}{36},$$

so $\|\mathbf{x}\|_\infty \leq 6/n^{1/2}$. Similarly,

$$\lambda_n \mathbf{z}_{\hat{v}} = \sum_{u \sim \hat{v}} \mathbf{z}_u \leq |\mathbf{z}_{\hat{v}}| \left(|B| + \frac{\min\{1, c^2\}}{100} (n - |B|) \right)$$

and $|\lambda_n(G)| \geq \min\{1, c^2\}\, n/50$, and so $|B| \geq \min\{1, c^2\}\, n/100$. In addition,

$$1 \geq \sum_{u \in B} \mathbf{z}_u^2 \geq |B| \min\{1, c^4\} \frac{\|\mathbf{z}\|_\infty^2}{10^4} \geq \min\{1, c^6\}\, n\, \frac{\|\mathbf{z}\|_\infty^2}{10^6},$$

so $\|\mathbf{z}\|_\infty \leq 10^3 \max\{1, c^{-3}\}/n^{1/2}$. This completes the proof of Property 5.

Finally, we prove Property 6. Let $\tilde{G} = (V(\tilde{G}), E(\tilde{G}))$ be the graph resulting from deleting u from G and cloning v, namely,

$$V(\tilde{G}) = \{v'\} \cup \big[V(G) \setminus \{u\} \big] \quad \text{and} \quad E(\tilde{G}) = E(G \setminus \{u\}) \cup \{\{v', w\} \,|\, \{v, w\} \in E(G)\}.$$

Let $\tilde{A}$ be the adjacency matrix of $\tilde{G}$, and $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{z}}$ equal

$$\tilde{\mathbf{x}}_w = \begin{cases} \mathbf{x}_w & w \neq v' \\ \mathbf{x}_v & w = v', \end{cases} \qquad \text{and} \qquad \tilde{\mathbf{z}}_w = \begin{cases} \mathbf{z}_w & w \neq v' \\ \mathbf{z}_v & w = v'. \end{cases}.$$

Then

$$\tilde{\mathbf{x}}^T\tilde{\mathbf{x}} = 1 - \mathbf{x}_u^2 + \mathbf{x}_v^2,$$
$$\tilde{\mathbf{z}}^T\tilde{\mathbf{z}} = 1 - \mathbf{z}_u^2 + \mathbf{z}_v^2,$$

and

$$\tilde{\mathbf{x}}^T\tilde{A}\tilde{\mathbf{x}} = \lambda_1 - 2\lambda_1\mathbf{x}_u^2 + 2\lambda_1\mathbf{x}_v^2 - 2A_{uv}\mathbf{x}_u\mathbf{x}_v,$$
$$\tilde{\mathbf{z}}^T\tilde{A}\tilde{\mathbf{z}} = \lambda_n - 2\lambda_n\mathbf{z}_u^2 + 2\lambda_n\mathbf{z}_v^2 - 2A_{uv}\mathbf{z}_u\mathbf{z}_v.$$

By the optimality of G,

$$\lambda_1(G) - c\,\lambda_n(G) - \left(\frac{\tilde{\mathbf{x}}^T\tilde{A}\tilde{\mathbf{x}}}{\tilde{\mathbf{x}}^T\tilde{\mathbf{x}}} - c\,\frac{\tilde{\mathbf{z}}^T\tilde{A}\tilde{\mathbf{z}}}{\tilde{\mathbf{z}}^T\tilde{\mathbf{z}}}\right) \geq 0.$$

Rearranging terms, we have

$$\frac{\lambda_1\mathbf{x}_u^2 - \lambda_1\mathbf{x}_v^2 + 2A_{u,v}\mathbf{x}_u\mathbf{x}_v}{1 - \mathbf{x}_u^2 + \mathbf{x}_v^2} - c\,\frac{\lambda_n\mathbf{z}_u^2 - \lambda_n\mathbf{z}_v^2 + 2A_{u,v}\mathbf{z}_u\mathbf{z}_v}{1 - \mathbf{z}_u^2 + \mathbf{z}_v^2} \geq 0,$$

and so

$$(\lambda_1\mathbf{x}_u^2 - c\,\lambda_n\mathbf{z}_u^2) - (\lambda_1\mathbf{x}_v^2 - c\,\lambda_n\mathbf{z}_v^2) \geq -\left[\frac{\lambda_1(\mathbf{x}_u^2 - \mathbf{x}_v^2)^2 + 2A_{u,v}\mathbf{x}_u\mathbf{x}_v}{1 - \mathbf{x}_u^2 + \mathbf{x}_v^2} - c\,\frac{\lambda_n(\mathbf{z}_u^2 - \mathbf{z}_v^2)^2 + 2A_{u,v}\mathbf{z}_u\mathbf{z}_v}{1 - \mathbf{z}_u^2 + \mathbf{z}_v^2}\right].$$

The infinity norms of $\mathbf{x}$ and $\mathbf{z}$ are at most $6/n^{1/2}$ and $10^3\max\{1, c^{-3}\}/n^{1/2}$, respectively,

and so we can upper bound

$$\left|\frac{\lambda_1(\mathbf{x}_u^2-\mathbf{x}_v^2)^2+2A_{u,v}\mathbf{x}_u\mathbf{x}_v}{1-\mathbf{x}_u^2+\mathbf{x}_v^2}\right| \leq \left|\frac{4n\|\mathbf{x}\|_\infty^4+2\|\mathbf{x}\|_\infty^2}{1-2\|\mathbf{x}\|_\infty^2}\right| \leq \frac{4\cdot 6^4+2\cdot 6^2}{n/2},$$

and

$$\left|\frac{\lambda_n(\mathbf{z}_u^2-\mathbf{z}_v^2)^2+2A_{u,v}\mathbf{z}_u\mathbf{z}_v}{1-\mathbf{z}_u^2+\mathbf{z}_v^2}\right| \leq \left|\frac{2n\|\mathbf{z}\|_\infty^4+2\|\mathbf{z}\|_\infty^2}{1-2\|\mathbf{z}\|_\infty^2}\right| \leq \frac{2(10^{12}+10^6)\max\{1,c^{-12}\}}{n/2}$$

for $n \geq 4\cdot 10^6 \max\{1,c^{-6}\}$. Our choice of u and v was arbitrary, and so

$$\left|(\lambda_1\mathbf{x}_u^2-c\,\lambda_n\mathbf{z}_u^2)-(\lambda_1\mathbf{x}_v^2-c\,\lambda_n\mathbf{z}_v^2)\right| \leq 5\cdot 10^{12}\max\{c,c^{-12}\}/n$$

for all $u,v\in V$. Noting that $\sum_u(\lambda_1\mathbf{x}_u^2-c\,\lambda_n\mathbf{z}_u^2)=\lambda_1-c\,\lambda_n$ completes the proof. □

The generality of this result implies that similar techniques to those used to prove the spread conjecture could be used to understand the behavior of graphs that maximize $\lambda_1-c\,\lambda_n$ and how the graphs vary with the choice of $c\in(0,\infty)$.

Chapter 4

Force-Directed Layouts

4.1 Introduction

Graph drawing is an area at the intersection of mathematics, computer science, and more qualitative fields. Despite the extensive literature in the field, in many ways the concept of what constitutes the optimal drawing of a graph is heuristic at best, and subjective at worst. For a general review of the major areas of research in graph drawing, we refer the reader to [7, 57]. In this chapter, we briefly introduce the broad class of force-directed layouts of graphs and consider two prominent force-based techniques for drawing a graph in low-dimensional Euclidean space: Tutte's spring embedding and metric multidimensional scaling.

The concept of a force-directed layout is somewhat ambiguous, but, loosely defined, it is a technique for drawing a graph in a low-dimensional Euclidean space (usually dimension ≤ 3) by applying "forces" between the set of vertices and/or edges. In a force-directed layout, vertices connected by an edge (or at a small graph distance from each other) tend to be close to each other in the resulting layout. Below we briefly introduce a number of prominent force-directed layouts, including Tutte's spring embedding, Eades' algorithm, the Kamada-Kawai objective and algorithm, and the more recent UMAP algorithm.

In his 1963[1] work titled "How to Draw a Graph," Tutte found an elegant technique

[1]The paper was received in May 1962, but was published in 1963. For this reason, in [124], the year is

to produce planar embeddings of planar graphs that also minimize "energy" (the sum of squared edge lengths) in some sense [116]. In particular, for a three-connected planar graph, he showed that if the outer face of the graph is fixed as the complement of some convex region in the plane, and every other point is located at the mass center of its neighbors, then the resulting embedding is planar. This result is now known as Tutte's spring embedding theorem, and is considered by many to be the first example of a force-directed layout. One of the major questions that this result does not treat is how to best embed the outer face. In Section 4.2, we investigate this question, consider connections to a Schur complement, and provide some theoretical results for this Schur complement using a discrete energy-only trace theorem.

An algorithm for general graphs was later proposed by Eades in 1984, in which vertices are placed in a random initial position in the plane, a logarithmic spring force is applied to each edge and a repulsive inverse square-root law force is applied to each pair of non-adjacent vertices [38]. The algorithm proceeds by iteratively moving each vertex towards its local equilibrium. Five years later, Kamada and Kawai proposed an objective function corresponding to the situation in which each pair of vertices are connected by a spring with equilibrium length given by their graph distance and spring constant given by the inverse square of their graph distance [55]. Their algorithm for locally minimizing this objective consists of choosing vertices iteratively based on the magnitude of the associated derivative of the objective, and locally minimizing with respect to that vertex. The associated objective function is a specific example of metric multidimensional scaling, and both the objective function and the proposed algorithm are quite popular in practice (see, for instance, popular packages such as GRAPHVIZ [39] and the SMACOF package in R). In Section 4.3, we provide a theoretical analysis of the Kamada-Kawai objective function. In particular, we prove a number of structural results regarding optimal layouts, provide algorithmic lower bounds for the optimization problem, and propose a polynomial time randomized approximation scheme for drawing low diameter graphs.

Most recently, a new force-based layout algorithm called UMAP was proposed as an

---

referred to as 1962.

alternative to the popular t-SNE algorithm [82]. The UMAP algorithm takes a data set as input, constructs a weighted graph, and performs a fairly complex force-directed layout in which attractive and repulsive forces governed by a number of hyper-parameters are applied. For the exact details of this approach, we refer the reader to [82, Section 3.2]. Though the UMAP algorithm is not specifically analyzed in this chapter, it is likely that some the techniques proposed in this chapter can be applied (given a much more complicated analysis) to more complicated force-directed layout objectives, such as that of the UMAP algorithm. In addition, the popularity of the UMAP algorithm illustrates the continued importance and relevance of force-based techniques and their analysis.

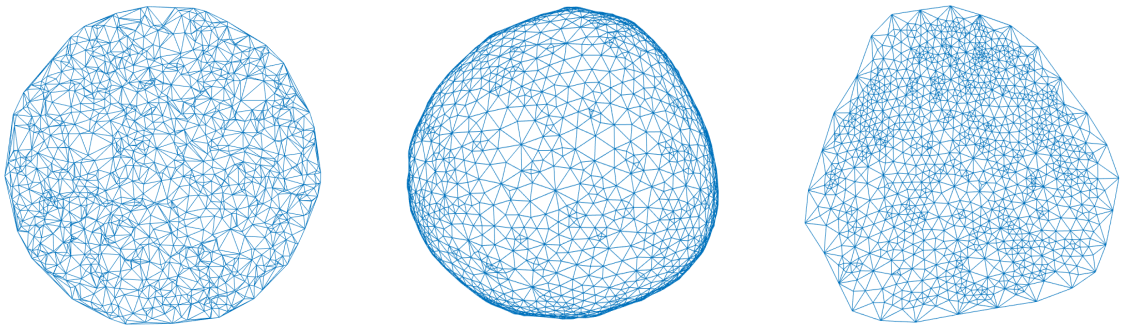
In the remainder of this chapter, we investigate both Tutte's spring embedding and the Kamada-Kawai objective function. In Section 4.2, we formally describe Tutte's spring embedding and consider theoretical questions regarding the best choice of boundary. In particular, we consider connections to a Schur complement and, using a discrete version of a trace theorem from the theory of elliptic PDEs, provide a spectral equivalence result for this matrix. In Section 4.3, we formally define the Kamada-Kawai objective, and prove a number of theoretical results for this optimization problem. We lower bound the objective value and upper bound the diameter of any optimal layout, prove that a gap version of the optimization problem is NP-hard even for bounded diameter graph metrics, and provide a polynomial time randomized approximation scheme for low diameter graphs.

4.2 Tutte's Spring Embedding and a Trace Theorem

The question of how to "best" draw a graph in the plane is not an easy one to answer, largely due to the ambiguity in the objective. When energy (i.e. Hall's energy, the sum of squared distances between adjacent vertices) minimization is desired, the optimal embedding in the plane is given by the two-dimensional diffusion map induced by the eigenvectors of the two smallest non-zero eigenvalues of the graph Laplacian [60, 61, 62]. This general class of graph drawing techniques is referred to as spectral layouts. When drawing a planar graph, often a planar embedding (a drawing in which edges do not intersect) is desirable. However, spectral layouts of planar graphs are not guaranteed to be planar. When looking at a triangulation of a given domain, it is commonplace for the near-boundary points of the spectral layout to "grow" out of the boundary, or lack any resemblance to a planar embedding. For instance, see the spectral layout of a random triangulation of a disk in Figure 4-1.

Tutte's spring embedding is an elegant technique to produce planar embeddings of planar graphs that also minimize energy in some sense [116]. In particular, for a three-connected planar graph, if the outer face of the graph is fixed as the complement of some convex region in the plane, and every other point is located at the mass center of its neighbors, then the resulting embedding is planar. This embedding minimizes Hall's energy, conditional on the embedding of the boundary face. This result is now known as Tutte's spring embedding theorem. While this result is well known (see [59], for example), it is not so obvious how to embed the outer face. This, of course, should vary from case to case, depending on the dynamics of the interior.

In this section, we investigate how to embed the boundary face so that the resulting drawing is planar and Hall's energy is minimized (subject to some normalization). In what follows, we produce an algorithm with theoretical guarantees for a large class of three-connected planar graphs. Our analysis begins by observing that the Schur complement of the graph Laplacian with respect to the interior vertices is, in some sense, the correct matrix to consider when choosing an optimal embedding of boundary vertices. See Figure 4-2

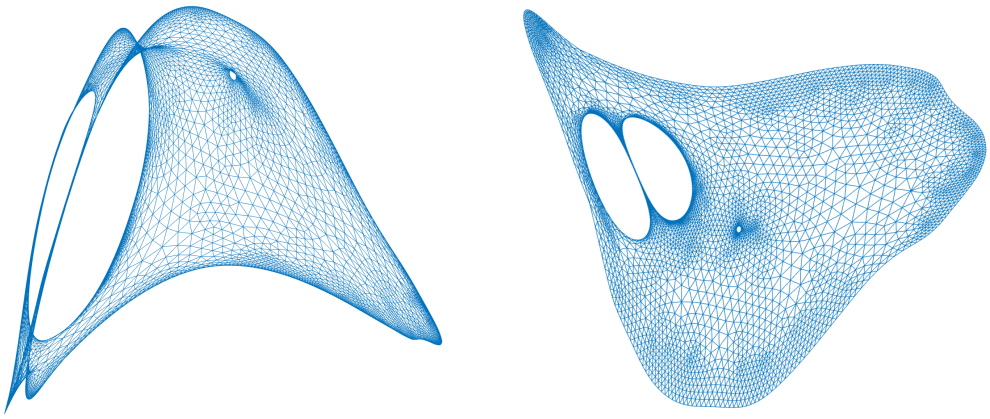


(a) Circle (b) Spectral Layout (c) Schur Complement Layout

Figure 4-1: A Delaunay triangulation of 1250 points randomly generated on the disk (A), its non-planar spectral layout (B), and a planar layout using a spring embedding of the Schur complement of the graph Laplacian with respect to the interior vertices (C).

for a visual example of a spring embedding using the two minimal non-trivial eigenvectors of the Schur complement. In order to theoretically understand the behavior of the Schur complement, we prove a discrete trace theorem. Trace theorems are a class of results in the theory of partial differential equations relating norms on the domain to norms on the boundary, which are used to provide a priori estimates on the Dirichlet integral of functions with given data on the boundary. We construct a discrete version of a trace theorem in the plane for energy-only semi-norms. Using a discrete trace theorem, we show that this Schur complement is spectrally equivalent to the boundary Laplacian to the one-half power. This spectral equivalence proves the existence of low energy (w.r.t. the Schur complement) convex and planar embeddings of the boundary, but is also of independent interest and applicability in the study of spectral properties of planar graphs. These theoretical guarantees give rise to a simple graph drawing algorithm with provable guarantees.

The remainder of this section is as follows. In Subsection 4.2.1, we formally introduce Tutte's spring embedding theorem, describe the optimization problem under consideration, illustrate the connection to a Schur complement, and, conditional on a spectral equivalence, describe a simple algorithm with provable guarantees. In Subsection 4.2.2, we prove the aforementioned spectral equivalence for a large class of three-connected planar graphs. In particular, we consider trace theorems for Lipschitz domains from the theory of elliptic



(a) Laplacian Embedding (b) Schur Complement Embedding

Figure 4-2: A visual example of embeddings of the 2D finite element discretization graph 3elt, taken from the SuiteSparse Matrix Collection [27]. Figure (A) is the non-planar spectral layout of this 2D mesh, and Figure (B) is a planar spring embedding of the mesh using the minimal non-trivial eigenvectors of the Schur complement to embed the boundary.

partial differential equations, prove discrete energy-only variants of these results for a large class of graphs in the plane, and show that the Schur complement with respect to the interior is spectrally equivalent to the boundary Laplacian to the one-half power.

Definitions and Notation

Let $G = (V, E)$, $V = \{1, ..., n\}$, $E \subset \{e \subset V \,|\, |e| = 2\}$, be a simple, connected, undirected graph. A graph G is k-connected if it remains connected upon the removal of any $k - 1$ vertices, and is planar if it can be drawn in the plane such that no edges intersect (save for adjacent edges at their mutual endpoint). A face of a planar embedding of a graph is a region of the plane bounded by edges (including the outer infinite region, referred to as the outer face). Let $\mathcal{G}_n$ be the set of all ordered pairs (G, Γ), where G is a simple, undirected, planar, three-connected graph of order $n > 4$, and $\Gamma \subset V$ is the set of vertices of some face of G. Three-connectedness is an important property for planar graphs, which, by Steinitz's theorem, guarantees that the graph is the skeleton of a convex polyhedron [107]. This characterization implies that, for three-connected graphs, the edges corresponding to each face in a planar embedding are uniquely determined by the graph. In particular, the set of

faces is simply the set of induced cycles, so we may refer to faces of the graph without specifying an embedding. One important corollary of this result is that, for $n \geq 3$, the vertices of any face form an induced simple cycle.

Let $N_G(i)$ be the neighborhood of vertex i, $N_G(S)$ be the union of the neighborhoods of the vertices in S, and $d_G(i,j)$ be the distance between vertices i and j in the graph G. When the associated graph is obvious, we may remove the subscript. Let $d(i)$ be the degree of vertex i. Let $G[S]$ be the graph induced by the vertices S, and $d_S(i,j)$ be the distance between vertices i and j in $G[S]$. If H is a subgraph of G, we write $H \subset G$. The Cartesian product $G_1 \square G_2$ between $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is the graph with vertices $(v_1, v_2) \in V_1 \times V_2$ and edge $\{(u_1, u_2), (v_1, v_2)\} \in E$ if $(u_1, v_1) \in E_1$ and $u_2 = v_2$, or $u_1 = v_1$ and $(u_2, v_2) \in E_2$. The graph Laplacian $L_G \in \mathbb{R}^{n \times n}$ of G is the symmetric positive semi-definite matrix defined by

$$\langle L_G x, x \rangle = \sum_{\{i,j\} \in E} (x_i - x_j)^2,$$

and, in general, a matrix is the graph Laplacian of some weighted graph if it is symmetric diagonally dominant, has non-positive off-diagonal entries, and the vector $\mathbf{1} := (1, ..., 1)^T$ lies in its nullspace. Given a matrix A, we denote the i^{th} row by $A_{i,\cdot}$, the j^{th} column by $A_{\cdot,j}$, and the entry in the i^{th} row and j^{th} column by $A_{i,j}$.

4.2.1 Spring Embeddings and a Schur Complement

Here and in what follows, we refer to Γ as the "boundary" of the graph G, $V \backslash \Gamma$ as the "interior," and generally assume $n_\Gamma := |\Gamma|$ to be relatively large (typically $n_\Gamma = \Theta(n^{1/2})$). The concept of a "boundary" face is somewhat arbitrary, but, depending on the application from which the graph originated (i.e., a discretization of some domain), one face is often already designated as the boundary face. If a face has not been designated, choosing the largest induced cycle is a reasonable choice. By producing a planar drawing of G in the plane and traversing the embedding, one can easily find all the induced cycles of G in linear time and space [20].

Without loss of generality, suppose that $\Gamma = \{n - n_\Gamma + 1, ..., n\}$. A matrix $X \in \mathbb{R}^{n\times 2}$ is said to be a planar embedding of G if the drawing of G using straight lines and with vertex i located at coordinates $X_{i,\cdot}$ for all $i \in V$ is a planar drawing. A matrix $X_\Gamma \in \mathbb{R}^{n_\Gamma \times 2}$ is said to be a convex embedding of Γ if the embedding is planar and every point is an extreme point of the convex hull $\text{conv}(\{[X_\Gamma]_{i,\cdot}\}_{i=1}^{n_\Gamma})$. Tutte's spring embedding theorem states that if X_Γ is a convex embedding of Γ, then the system of equations

$$X_{i,\cdot} = \begin{cases} \frac{1}{d(i)} \sum_{j\in N(i)} X_{j,\cdot} & i = 1, ..., n - n_\Gamma \\ [X_\Gamma]_{i-(n-n_\Gamma),\cdot} & i = n - n_\Gamma + 1, ..., n \end{cases}$$

has a unique solution X, and this solution is a planar embedding of G [116].

We can write both the Laplacian and embedding of G in block-notation, differentiating between interior and boundary vertices as follows:

$$L_G = \begin{pmatrix} L_o + D_o & -A_{o,\Gamma} \\ -A_{o,\Gamma}^T & L_\Gamma + D_\Gamma \end{pmatrix} \in \mathbb{R}^{n\times n}, \quad X = \begin{pmatrix} X_o \\ X_\Gamma \end{pmatrix} \in \mathbb{R}^{n\times 2},$$

where $L_o, D_o \in \mathbb{R}^{(n-n_\Gamma)\times(n-n_\Gamma)}$, $L_\Gamma, D_\Gamma \in \mathbb{R}^{n_\Gamma\times n_\Gamma}$, $A_{o,\Gamma} \in \mathbb{R}^{(n-n_\Gamma)\times n_\Gamma}$, $X_o \in \mathbb{R}^{(n-n_\Gamma)\times 2}$, $X_\Gamma \in \mathbb{R}^{n_\Gamma\times 2}$, and L_o and L_Γ are the Laplacians of $G[V\backslash\Gamma]$ and $G[\Gamma]$, respectively. Using block notation, the system of equations for the Tutte spring embedding of some convex embedding X_Γ is given by

$$X_o = (D_o + D[L_o])^{-1}[(D[L_o] - L_o)X_o + A_{o,\Gamma}X_\Gamma],$$

where $D[A]$ is the diagonal matrix with diagonal entries given by the diagonal of A. The unique solution to this system is

$$X_o = (L_o + D_o)^{-1}A_{o,\Gamma}X_\Gamma.$$

We note that this choice of X_o not only guarantees a planar embedding of G, but also

minimizes Hall's energy, namely,

$$\arg\min_{X_o} h(X) = (L_o + D_o)^{-1} A_{o,\Gamma} X_\Gamma,$$

where $h(X) := \text{Tr}(X^T L X)$ (see [62] for more on Hall's energy).

While Tutte's theorem is a very powerful result, guaranteeing that, given a convex embedding of any face, the energy minimizing embedding of the remaining vertices results in a planar embedding, it gives no direction as to how this outer face should be embedded.

We consider embeddings of the boundary that minimize Hall's energy, given some normalization. We consider embeddings that satisfy $X_\Gamma^T X_\Gamma = I$ and $X_\Gamma^T \mathbf{1} = 0$, though other normalizations, such as $X^T X = I$ and $X^T \mathbf{1} = 0$, would be equally appropriate. The analysis that follows in this section can be readily applied to this alternate normalization, but it does require the additional step of verifying a norm equivalence between V and Γ for the harmonic extension of low energy vectors, which can be produced relatively easily for the class of graphs considered in Subsection 4.2.2. In addition, alternate normalization techniques, such as the introduction of negative weights on the boundary cycle, can also be considered. Let $\mathcal{X}$ be the set of all planar embeddings X_Γ that satisfy $X_\Gamma^T X_\Gamma = I$, $X_\Gamma^T \mathbf{1} = 0$, and for which the spring embedding X with $X_o = (L_o + D_o)^{-1} A_{o,\Gamma} X_\Gamma$ is planar. We consider the optimization problem

$$\min \; h(X) \quad s.t. \quad X_\Gamma \in \text{cl}(\mathcal{X}), \tag{4.1}$$

where $\text{cl}(\cdot)$ is the closure of a set. The set $\mathcal{X}$ is not closed, and the minimizer of (4.1) may be non-planar, but must be arbitrarily close to a planar embedding. The normalizations $X_\Gamma^T \mathbf{1} = 0$ and $X_\Gamma^T X_\Gamma = I$ ensure that the solution does not degenerate into a single point or line. In what follows we are primarily concerned with approximately solving this optimization problem and connecting this problem to the Schur complement of L_G with respect to $V \backslash \Gamma$. It is unclear whether there exists an efficient algorithm to solve (4.1) exactly or if the associated decision problem is NP-hard. If (4.1) is NP-hard, it seems rather difficult to verify that this is indeed the case.

A Schur Complement

Given some choice of X_Γ, by Tutte's theorem the minimum value of $h(X)$ is attained when $X_o = (L_o + D_o)^{-1} A_{o,\Gamma} X_\Gamma$, and given by

$$\begin{aligned}
&\text{Tr}\left[\begin{pmatrix} [(L_o + D_o)^{-1} A_{o,\Gamma} X_\Gamma]^T & X_\Gamma^T \end{pmatrix} \begin{pmatrix} L_o + D_o & -A_{o,\Gamma} \\ -A_{o,\Gamma}^T & L_\Gamma + D_\Gamma \end{pmatrix} \begin{pmatrix} (L_o + D_o)^{-1} A_{o,\Gamma} X_\Gamma \\ X_\Gamma \end{pmatrix} \right] \\
&\quad = \text{Tr}\big(X_\Gamma^T \big[L_\Gamma + D_\Gamma - A_{o,\Gamma}^T (L_o + D_o)^{-1} A_{o,\Gamma} \big] X_\Gamma \big) \\
&\quad = \text{Tr}\left(X_\Gamma^T S_\Gamma X_\Gamma \right),
\end{aligned}$$

where S_Γ is the Schur complement of L_G with respect to $V \backslash \Gamma$,

$$S_\Gamma = L_\Gamma + D_\Gamma - A_{o,\Gamma}^T (L_o + D_o)^{-1} A_{o,\Gamma}.$$

For this reason, we can instead consider the optimization problem

$$\min \; h_\Gamma(X_\Gamma) \quad s.t. \quad X_\Gamma \in \text{cl}(\mathcal{X}), \tag{4.2}$$

where

$$h_\Gamma(X_\Gamma) := \text{Tr}\big(X_\Gamma^T S_\Gamma X_\Gamma \big).$$

Therefore, if the minimal two non-trivial eigenvectors of S_Γ produce a planar spring embedding, then this is the exact solution of (4.2). However, a priori, there is no reason to think that this drawing would be planar. It turns out that, for a large class of graphs with some macroscopic structure, this is often the case (see, for example, Figures 4-1 and 4-2), and, in the rare instance that it is not, through a spectral equivalence result, a low energy planar and convex embedding of the boundary always exists. This fact follows from the spectral equivalence of S_Γ and $L_\Gamma^{1/2}$, which is shown in Subsection 4.2.2.

First, we present a number of basic properties of the Schur complement of a graph Laplacian. For more information on the Schur complement, we refer the reader to [18, 40, 134].

Proposition 8. *Let $G = (V, E)$, $n = |V|$, be a graph and $L_G \in \mathbb{R}^{n\times n}$ the associated graph Laplacian. Let L_G and vectors $v \in \mathbb{R}^n$ be written in block form*

$$L(G) = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix}, \quad v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix},$$

where $L_{22} \in \mathbb{R}^{m\times m}$, $v_2 \in \mathbb{R}^m$, and $L_{12} \neq 0$. Then

(1) $S = L_{22} - L_{21}L_{11}^{-1}L_{12}$ is a graph Laplacian,

(2) $\sum_{i=1}^{m}(e_i^T L_{22}\mathbf{1}_m)e_i e_i^T - L_{21}L_{11}^{-1}L_{12}$ is a graph Laplacian,

(3) $\langle Sw, w\rangle = \inf\{\langle Lv, v\rangle | v_2 = w\}$.

Proof. Let $P = \begin{pmatrix} -L_{11}^{-1}L_{12} \\ I \end{pmatrix} \in \mathbb{R}^{n\times m}$. Then

$$P^T LP = \begin{pmatrix} -L_{21}L_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} -L_{11}^{-1}L_{12} \\ I \end{pmatrix} = L_{22} - L_{21}L_{11}^{-1}L_{12} = S.$$

Because $L_{11}\mathbf{1}_{n-m} + L_{12}\mathbf{1}_m = 0$, we have $\mathbf{1}_{n-m} = -L_{11}^{-1}L_{12}\mathbf{1}_m$. Therefore $P\mathbf{1}_m = \mathbf{1}_n$, and, as a result,

$$S\mathbf{1}_m = P^T LP\mathbf{1}_m = P^T L\mathbf{1}_n = P^T 0 = 0.$$

In addition,

$$\begin{aligned} \left[\sum_{i=1}^{m}(e_i^T L_{22}\mathbf{1}_m)e_i e_i^T - L_{21}L_{11}^{-1}L_{12}\right]\mathbf{1}_m &= \left[\sum_{i=1}^{m}(e_i^T L_{22}\mathbf{1}_m)e_i e_i^T - L_{22}\right]\mathbf{1}_m + S\mathbf{1}_m \\ &= \sum_{i=1}^{m}(e_i^T L_{22}\mathbf{1}_m)e_i - L_{22}\mathbf{1}_m \\ &= \left[\sum_{i=1}^{m} e_i e_i^T - I_m\right]L_{22}\mathbf{1}_m \;=\; 0. \end{aligned}$$

L_{11} is an M-matrix, so L_{11}^{-1} is a non-negative matrix. $L_{21}L_{11}^{-1}L_{12}$ is the product of three non-negative matrices, and so must also be non-negative. Therefore, the off-diagonal entries of

S and $\sum_{i=1}^{m}(e_i^T L_{22}\mathbf{1})e_i e_i^T - L_{21}L_{11}^{-1}L_{12}$ are non-positive, and so both are graph Laplacians.

Consider

$$\langle Lv, v\rangle = \langle L_{11}v_1, v_1\rangle + 2\langle L_{12}v_2, v_1\rangle + \langle L_{22}v_2, v_2\rangle,$$

with v_2 fixed. Because L_{11} is symmetric positive definite, the minimum occurs when

$$\frac{\partial}{\partial v_1}\langle Lv, v\rangle = 2L_{11}v_1 + 2L_{12}v_2 = 0.$$

Setting $v_1 = -L_{11}^{-1}L_{12}v_2$, the desired result follows. □

The above result illustrates that the Schur complement Laplacian S_Γ is the sum of two Laplacians L_Γ and $D_\Gamma - A_{o,\Gamma}^T(L_o + D_o)^{-1}A_{o,\Gamma}$, where the first is the Laplacian of $G[\Gamma]$, and the second is a Laplacian representing the dynamics of the interior.

An Algorithm

Given the above analysis for S_Γ, and assuming a spectral equivalence result of the form

$$\frac{1}{c_1}\langle L_\Gamma^{1/2}x, x\rangle \leq \langle S_\Gamma x, x\rangle \leq c_2\langle L_\Gamma^{1/2}x, x\rangle, \tag{4.3}$$

for all $x \in \mathbb{R}^{n_\Gamma}$ and some constants c_1 and c_2 that are not too large, we can construct a simple algorithm for producing a spring embedding. Constants c_1 and c_2 can be computed explicitly using the results of Subsection 4.2.2, and are reasonably sized if the graph resembles a discretization of some planar domain (i.e., a graph that satisfies the conditions of Theorem 15). Given these two facts, a very natural algorithm consists of the following steps. First we compute the two eigenvectors corresponding to the two minimal non-trivial eigenvalues of the Schur complement S_Γ. If these two eigenvectors produce a planar embedding, we are done, and have solved the optimization problem (4.2) exactly. Depending on the structure of the graph, this may often be the case. For instance, this occurs for the graphs in Figures 4-1 and 4-2, and the numerical results of [124] confirm that this also occurs often for discretizations of circular and rectangular domains. If this is not the case, then we consider

Algorithm 3 A Schur Complement-Based Spring Embedding Algorithm

function SchurSPRING(G, Γ)
  Compute the two minimal non-trivial eigenpairs $X_{alg} \in \mathbb{R}^{n_\Gamma \times 2}$ of S_Γ
  if spring embedding of X_{alg} is non-planar **then**
    $X_{new} \leftarrow \left\{ \frac{2}{n_\Gamma} \left(\cos \frac{2\pi j}{n_\Gamma}, \sin \frac{2\pi j}{n_\Gamma} \right) \right\}_{j=1}^{n_\Gamma}$; gap $\leftarrow 1$
    while spring embedding of X_{new} is planar, gap > 0 **do**
      $\hat{X} \leftarrow$ smooth(X_{new}); $X_{alg} \leftarrow X_{new}$
      $\hat{X} \leftarrow \hat{X} - \mathbf{1}\mathbf{1}^T \hat{X}/n_\Gamma$
      Solve $[\hat{X}^T \hat{X}]Q = Q\Lambda$, Q orthogonal, Λ diagonal; set $X_{new} \leftarrow \hat{X} Q \Lambda^{-1/2}$
      gap $\leftarrow h_\Gamma(X_{alg}) - h_\Gamma(X_{new})$
    end while
  end if
  Return X_{alg}
end function

the boundary embedding

$$[X_C]_{j,\cdot} = \frac{2}{n_\Gamma} \left(\cos \frac{2\pi j}{n_\Gamma}, \sin \frac{2\pi j}{n_\Gamma} \right), \qquad j = 1, ..., n_\Gamma,$$

namely, the embedding of the two minimal non-trivial eigenvectors of $L_\Gamma^{1/2}$. This choice of boundary is convex and planar, and so the associated spring embedding is planar. By spectral equivalence,

$$h_\Gamma(X_C) \leq 4c_2 \sin \frac{\pi}{n_\Gamma} \leq c_1 c_2 \min_{X_\Gamma \in \mathrm{cl}(\mathcal{X})} h_\Gamma(X_\Gamma),$$

and therefore, this algorithm already produces a $c_1 c_2$ approximation guarantee for (4.2). The choice of X_C can often be improved, and a natural technique consists of smoothing X_C using S_Γ and renormalizing until either the resulting drawing is no longer planar or the objective value no longer decreases. We describe this procedure in Algorithm 3.

We now discuss some of the finer details of the SchurSPRING(G, Γ) algorithm (Algorithm 3). Determining whether a drawing is planar can be done in near-linear time using the sweep line algorithm [102]. Also, in practice, it is advisable to replace conditions of the form gap > 0 in Algorithm 3 by gap $>$ tol for some small value of tol, in order to ensure that the algorithm terminates after some finite number of steps. For a reasonable choice of

smoother, a constant tolerance, and $n_\Gamma = \Theta(n^{1/2})$, the SchurSPRING$(G, \Gamma)$ algorithm has complexity near-linear in n. The main cost of this procedure is due to the computations that involve S_Γ.

The SchurSPRING(G, Γ) algorithm requires the repeated application of S_Γ or S_Γ^{-1} in order to compute the minimal eigenvectors of S_Γ and also to perform smoothing. The Schur complement S_Γ is a dense matrix and requires the inversion of a $(n - n_\Gamma) \times (n - n_\Gamma)$ matrix, but can be represented as the composition of functions of sparse matrices. In practice, S_Γ should never be formed explicitly. Rather, the operation of applying S_Γ to a vector x should occur in two steps. First, the sparse Laplacian system $(L_o + D_o)y = A_{o,\Gamma}x$ should be solved for y, and then the product Sx is given by $S_\Gamma x = (L_\Gamma + D_\Gamma)x - A_{o,\Gamma}^T y$. Each application of S_Γ is therefore an $\tilde{O}(n)$ procedure (using a nearly-linear time Laplacian solver). The application of the inverse S_Γ^{-1} defined on the subspace $\{x \mid \langle x, \mathbf{1}\rangle = 0\}$ also requires the solution of a Laplacian system. As noted in [130], the action of S_Γ^{-1} on a vector $x \in \{x \mid \langle x, \mathbf{1}\rangle = 0\}$ is given by

$$S_\Gamma^{-1} x = \begin{pmatrix} 0 & I \end{pmatrix} \begin{pmatrix} L_o + D_o & -A_{o,\Gamma} \\ -A_{o,\Gamma}^T & L_\Gamma + D_\Gamma \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ x \end{pmatrix},$$

as verified by the computation

$$\begin{aligned} S_\Gamma \left[S_\Gamma^{-1} x \right] &= S_\Gamma \begin{pmatrix} 0 & I \end{pmatrix} \left[\begin{pmatrix} I & 0 \\ -A_{o,\Gamma}^T (L_o + D_o)^{-1} & I \end{pmatrix} \begin{pmatrix} L_o + D_o & -A_{o,\Gamma} \\ 0 & S_\Gamma \end{pmatrix} \right]^{-1} \begin{pmatrix} 0 \\ x \end{pmatrix} \\ &= S_\Gamma \begin{pmatrix} 0 & I \end{pmatrix} \begin{pmatrix} L_o + D_o & -A_{o,\Gamma} \\ 0 & S_\Gamma \end{pmatrix}^{-1} \begin{pmatrix} I & 0 \\ A_{o,\Gamma}^T (L_o + D_o)^{-1} & I \end{pmatrix} \begin{pmatrix} 0 \\ x \end{pmatrix} \\ &= S_\Gamma \begin{pmatrix} 0 & I \end{pmatrix} \begin{pmatrix} L_o + D_o & -A_{o,\Gamma} \\ 0 & S_\Gamma \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ x \end{pmatrix} = x. \end{aligned}$$

Given that the application of S_Γ^{-1} has the same complexity as an application S_Γ, the inverse power method is naturally preferred over the shifted power method for both smoothing and the computation of low energy eigenvectors.

4.2.2 A Discrete Trace Theorem and Spectral Equivalence

The main result of this subsection takes classical trace theorems from the theory of partial differential equations and extends them to a class of planar graphs. However, for our purposes, we require a stronger form of trace theorem, one between energy semi-norms (i.e., no ℓ^2 term), which we refer to as "energy-only" trace theorems. These energy-only trace theorems imply their classical variants with ℓ^2 terms almost immediately. We then use these new results to prove the spectral equivalence of S_Γ and $L_\Gamma^{1/2}$ for the class of graphs under consideration. This class of graphs is rigorously defined below, but includes planar three-connected graphs that have some regular structure (such as graphs of finite element discretizations). For the sake of space and readibility, a number of the results of this subsection are stated for arbitrary constants, or constants larger than necessary. More complicated proofs with explicit constants (or, in some cases, improved constants) can be found in [124]. We begin by formally describing a classical trace theorem.

Let $\Omega \subset \mathbb{R}^d$ be a domain with boundary $\Gamma = \delta\Omega$ that, locally, is a graph of a Lipschitz function. $H^1(\Omega)$ is the Sobolev space of square integrable functions with square integrable weak gradient, with norm

$$\|u\|_{1,\Omega}^2 = \|\nabla u\|_{L^2(\Omega)}^2 + \|u\|_{L^2(\Omega)}^2, \quad \text{where} \quad \|u\|_{L_2(\Omega)}^2 = \int_\Omega u^2 \, dx.$$

Let

$$\|\varphi\|_{1/2,\Gamma}^2 = \|\varphi\|_{L_2(\Gamma)}^2 + \iint_{\Gamma\times\Gamma} \frac{(\varphi(x) - \varphi(y))^2}{|x-y|^d} \, dx \, dy$$

for functions defined on Γ, and denote by $H^{1/2}(\Gamma)$ the Sobolev space of functions defined on the boundary Γ for which $\|\cdot\|_{1/2,\Gamma}$ is finite. The trace theorem for functions in $H^1(\Omega)$ is one of the most important and frequently used trace theorems in the theory of partial differential equations. More general results for traces on boundaries of Lipschitz domains, which involve L^p norms and fractional derivatives, are due to E. Gagliardo [42] (see also [24]). Gagliardo's theorem, when applied to the case of $H^1(\Omega)$ and $H^{1/2}(\Gamma)$, states that if $\Omega \subset \mathbb{R}^d$

is a Lipschitz domain, then the norm equivalence

$$\|\varphi\|_{1/2,\Gamma} \eqsim \inf\{\|u\|_{1,\Omega} \,\big|\, u|_\Gamma = \varphi\}$$

holds (the right hand side is indeed a norm on $H^{1/2}(\Gamma)$). These results are key tools in proving a priori estimates on the Dirichlet integral of functions with given data on the boundary of a domain Ω. Roughly speaking, a trace theorem gives a bound on the energy of a harmonic function via norm of the trace of the function on $\Gamma = \partial\Omega$. In addition to the classical references given above, further details on trace theorems and their role in the analysis of PDEs (including the case of Lipschitz domains) can be found in [77, 87]. There are several analogues of this theorem for finite element spaces (finite dimensional subspaces of $H^1(\Omega)$). For instance, in [86] it is shown that the finite element discretization of the Laplace-Beltrami operator on the boundary to the one-half power provides a norm which is equivalent to the $H^{1/2}(\Gamma)$-norm. Here we prove energy-only analogues of the classical trace theorem for graphs $(G, \Gamma) \in \mathcal{G}_n$, using energy semi-norms

$$|u|_G^2 = \langle L_G u, u\rangle \qquad \text{and} \qquad |\varphi|_\Gamma^2 = \sum_{\substack{p,q\in\Gamma,\\ p<q}} \frac{\left(\varphi(p)-\varphi(q)\right)^2}{d_G^2(p,q)}.$$

The energy semi-norm $|\cdot|_G$ is a discrete analogue of $\|\nabla u\|_{L^2(\Omega)}$, and the boundary semi-norm $|\cdot|_\Gamma$ is a discrete analogue of the quantity $\iint_{\Gamma\times\Gamma} \frac{(\varphi(x)-\varphi(y))^2}{|x-y|^2}\, dx\, dy$. In addition, by connectivity, $|\cdot|_G$ and $|\cdot|_\Gamma$ are norms on the quotient space orthogonal to $\mathbf{1}$. We aim to prove that for any $\varphi \in \mathbb{R}^{n_\Gamma}$,

$$\frac{1}{c_1}\,|\varphi|_\Gamma \le \min_{u|_\Gamma=\varphi} |u|_G \le c_2\,|\varphi|_\Gamma$$

for some constants c_1, c_2 that do not depend on n_Γ, n. We begin by proving these results for a simple class of graphs, and then extend our analysis to more general graphs. Some of the proofs of the below results are rather technical, and are omitted for the sake of readibility and space.

Trace Theorems for a Simple Class of Graphs

Let $G_{k,\ell} = C_k \,\square\, P_\ell$ be the Cartesian product of the k vertex cycle C_k and the ℓ vertex path P_ℓ, where $4\ell < k < 2c\ell$ for some constant $c \in \mathbb{N}$. The lower bound $4\ell < k$ is arbitrary in some sense, but is natural, given that the ratio of boundary length to in-radius of a convex region is at least 2π. Vertex (i, j) in $G_{k,\ell}$ corresponds to the product of $i \in C_k$ and $j \in P_\ell$, $i = 1, ..., k$, $j = 1, ..., \ell$. The boundary of $G_{k,\ell}$ is defined to be $\Gamma = \{(i, 1)\}_{i=1}^{k}$. Let $u \in \mathbb{R}^{k\times\ell}$ and $\varphi \in \mathbb{R}^k$ be functions on $G_{k,\ell}$ and Γ, respectively, with $u[(i, j)]$ denoted by $u_{i,j}$ and $\varphi[(i, 1)]$ denoted by φ_i. For the remainder of the section, we consider the natural periodic extension of the vertices (i, j) and the functions $u_{i,j}$ and φ_i to the indices $i \in \mathbb{Z}$. In particular, if $i \notin \{1, ..., k\}$, then $(i, j) := (i^*, j)$, $\varphi_i := \varphi_{i^*}$, and $u_{i,j} := u_{i^*,j}$, where $i^* \in \{1, ..., k\}$ and $i^* = i \mod k$. Let $G^*_{k,\ell}$ be the graph resulting from adding to $G_{k,\ell}$ all edges of the form $\{(i, j), (i - 1, j + 1)\}$ and $\{(i, j), (i + 1, j + 1)\}$, $i = 1, ..., k$, $j = 1, ..., \ell - 1$. We provide a visual example of $G_{k,\ell}$ and $G^*_{k,\ell}$ in Figure 4-3. First, we prove a trace theorem for $G_{k,\ell}$. The proof of the trace theorem consists of two lemmas. Lemma 10 shows that the discrete trace operator is bounded, and Lemma 11 shows that it has a continuous right inverse. Taken together, these lemmas imply our desired result for $G_{k,\ell}$. We then extend this result to all graphs H satisfying $G_{k,\ell} \subset H \subset G^*_{k,\ell}$.

Lemma 10. *Let $G = G_{k,\ell}$, $4\ell < k < 2c\ell$, $c \in \mathbb{N}$, with boundary $\Gamma = \{(i, 1)\}_{i=1}^{k}$. For any $u \in \mathbb{R}^{k\times\ell}$, the vector $\varphi = u|_\Gamma$ satisfies $|\varphi|_\Gamma \le 4\sqrt{c}\,|u|_G$.*

Proof. We can decompose $\varphi_{p+h} - \varphi_h$ into a sum of differences, given by

$$\varphi_{p+h} - \varphi_p = \sum_{i=1}^{s-1} u_{p+h,i} - u_{p+h,i+1} + \sum_{i=1}^{h} u_{p+i,s} - u_{p+i-1,s} + \sum_{i=1}^{s-1} u_{p,s-i+1} - u_{p,s-i},$$

where $s = \left\lceil \frac{h}{c} \right\rceil$ is a function of h. By Cauchy-Schwarz,

$$\sum_{p=1}^{k} \sum_{h=1}^{\lfloor k/2 \rfloor} \left(\frac{\varphi_{p+h} - \varphi_p}{h} \right)^2 \leq 3 \sum_{p=1}^{k} \sum_{h=1}^{\lfloor k/2 \rfloor} \left(\frac{1}{h} \sum_{i=1}^{s-1} u_{p+h,i} - u_{p+h,i+1} \right)^2 + 3 \sum_{p=1}^{k} \sum_{h=1}^{\lfloor k/2 \rfloor} \left(\frac{1}{h} \sum_{i=1}^{h} u_{p+i,s} - u_{p+i-1,s} \right)^2 + 3 \sum_{p=1}^{k} \sum_{h=1}^{\lfloor k/2 \rfloor} \left(\frac{1}{h} \sum_{i=1}^{s-1} u_{p,s-i+1} - u_{p,s-i} \right)^2 .$$

We bound the first and the second term separately. The third term is identical to the first. Using Hardy's inequality [50, Theorem 326], we can bound the first term by

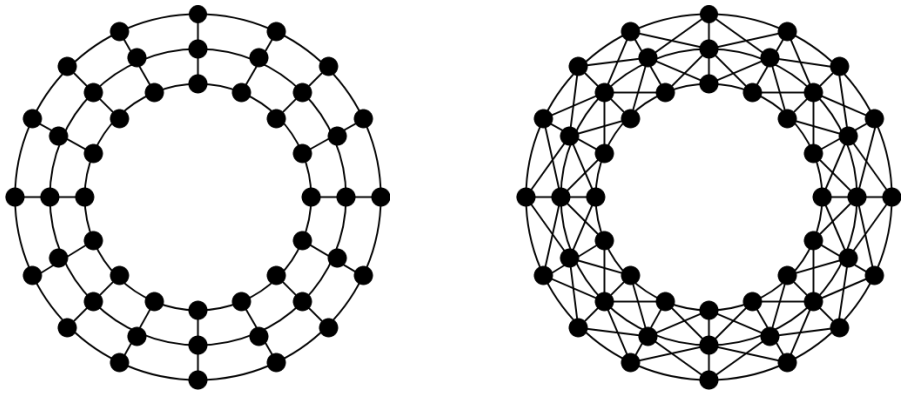
$$\sum_{p=1}^{k} \sum_{h=1}^{\lfloor k/2 \rfloor} \left(\frac{1}{h} \sum_{i=1}^{s-1} u_{p,i} - u_{p,i+1} \right)^2 = \sum_{p=1}^{k} \sum_{s=1}^{\ell} \left(\frac{1}{s} \sum_{i=1}^{s-1} u_{p,i} - u_{p,i+1} \right)^2 \sum_{\substack{h:\lceil h/c \rceil = s \\ 1 \leq h \leq \lfloor k/2 \rfloor}} \frac{s^2}{h^2} \leq 4 \sum_{p=1}^{k} \sum_{s=1}^{\ell-1} \left(u_{p,s} - u_{p,s+1} \right)^2 \sum_{\substack{h:\lceil h/c \rceil = s \\ 1 \leq h \leq \lfloor k/2 \rfloor}} \frac{s^2}{h^2}.$$

We have

$$\sum_{\substack{h:\lceil h/c \rceil = s \\ 1 \leq h \leq \lfloor k/2 \rfloor}} \frac{s^2}{h^2} \leq s^2 \sum_{i=c(s-1)+1}^{cs} \frac{1}{i^2} \leq \frac{s^2(c-1)}{(c(s-1)+1)^2} \leq \frac{4(c-1)}{(c+1)^2} \leq \frac{1}{2}$$

for $s \geq 2$ ($c \geq 3$, by definition), and for $s = 1$,

$$\sum_{\substack{h:\lceil h/c \rceil = 1 \\ 1 \leq h \leq \lfloor k/2 \rfloor}} \frac{1}{h^2} \leq \sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}.$$



(a) $G_{16,3} = C_{16} \square P_3$ (b) $G^*_{16,3}$

Figure 4-3: A visual example of $G_{k,\ell}$ and $G^*_{k,\ell}$ for $k = 16$, $\ell = 3$. The boundary Γ is given by the outer (or, by symmetry, inner) cycle.

Therefore, we can bound the first term by

$$\sum_{p=1}^{k} \sum_{h=1}^{\lfloor k/2 \rfloor} \left(\frac{1}{h} \sum_{i=1}^{s-1} u_{p,i} - u_{p,i+1} \right)^2 \leq \frac{2\pi^2}{3} \sum_{p=1}^{k} \sum_{s=1}^{\ell-1} \left(u_{p,s} - u_{p,s+1} \right)^2.$$

For the second term, we have

$$\begin{aligned} \sum_{p=1}^{k} \sum_{h=1}^{\lfloor k/2 \rfloor} \left(\frac{1}{h} \sum_{i=1}^{h} u_{p+i,s} - u_{p+i-1,s} \right)^2 &\leq \sum_{p=1}^{k} \sum_{h=1}^{\lfloor k/2 \rfloor} \frac{1}{h} \sum_{i=1}^{h} \left(u_{p+i,s} - u_{p+i-1,s} \right)^2 \\ &\leq c \sum_{p=1}^{k} \sum_{s=1}^{\ell} \left(u_{p+1,s} - u_{p,s} \right)^2. \end{aligned}$$

Combining these bounds produces the result $|\varphi|_\Gamma^2 \leq \max\{4\pi^2, 3c\} \, |u|_G^2$. Noting that $c \geq 3$ gives the desired result. $\square$

In order to show that the discrete trace operator has a continuous right inverse, we need to produce a provably low-energy extension of an arbitrary function on Γ. Let

$$a = \frac{1}{k} \sum_{p=1}^{k} \varphi_p \qquad \text{and} \qquad a_{i,j} = \frac{1}{2j-1} \sum_{h=1-j}^{j-1} \varphi_{i+h}.$$

We consider the extension

$$u_{i,j} = \frac{j-1}{\ell-1}\, a + \left(1 - \frac{j-1}{\ell-1}\right) a_{i,j}. \tag{4.4}$$

The proof of following inverse result for the discrete trace operator is similar in technique to that of Lemma 10, but significantly more involved.

Lemma 11. *Let $G = G_{k,\ell}$, $4\ell < k < 2c\ell$, $c \in \mathbb{N}$, with boundary $\Gamma = \{(i,1)\}_{i=1}^{k}$. For any $\varphi \in \mathbb{R}^k$, the vector u defined by (4.4) satisfies $|u|_G \leq 5\sqrt{c}\,|\varphi|_\Gamma$.*

Proof. We can decompose $|u|_G^2$ into two parts, namely,

$$|u|_G^2 = \sum_{i=1}^{k}\sum_{j=1}^{\ell}(u_{i+1,j} - u_{i,j})^2 + \sum_{i=1}^{k}\sum_{j=1}^{\ell-1}(u_{i,j+1} - u_{i,j})^2.$$

We bound each sum separately, beginning with the first. We have

$$u_{i+1,j} - u_{i,j} = \left(1 - \frac{j-1}{\ell-1}\right)(a_{i+1,j} - a_{i,j}) = \left(1 - \frac{j-1}{\ell-1}\right)\frac{\varphi_{i+j} - \varphi_{i+1-j}}{2j-1}.$$

Squaring both sides and noting that $4\ell < k$, we have

$$\sum_{i=1}^{k}\sum_{j=1}^{\ell}(u_{i+1,j} - u_{i,j})^2 \leq \sum_{i=1}^{k}\sum_{j=1}^{\ell}\left[\frac{\varphi_{i+j} - \varphi_{i+1-j}}{2j-1}\right]^2 \leq \sum_{p=1}^{k}\sum_{h=1}^{2\ell-1}\left[\frac{\varphi_{p+h} - \varphi_p}{h}\right]^2 \leq |\varphi|_\Gamma^2.$$

We now consider the second sum. Each term can be decomposed as

$$u_{i,j+1} - u_{i,j} = \frac{a - a_{i,j}}{\ell-1} + \left(1 - \frac{j}{\ell-1}\right)[a_{i,j+1} - a_{i,j}],$$

which leads to the upper bound

$$\sum_{i=1}^{k}\sum_{j=1}^{\ell-1}(u_{i,j+1} - u_{i,j})^2 \leq 2\sum_{i=1}^{k}\sum_{j=1}^{\ell-1}\left[\frac{a - a_{i,j}}{\ell-1}\right]^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{\ell-1}(a_{i,j+1} - a_{i,j})^2.$$

We estimate the two terms in the previous equation separately, beginning with the first. The

difference $a - a_{i,j}$ can be written as

$$a - a_{i,j} = \frac{1}{k}\sum_{p=1}^{k}\varphi_p - \frac{1}{2j-1}\sum_{h=1-j}^{j-1}\varphi_{i+h} = \frac{1}{k(2j-1)}\sum_{p=1}^{k}\sum_{h=1-j}^{j-1}\varphi_p - \varphi_{i+h}.$$

Squaring both sides,

$$(a - a_{i,j})^2 = \frac{1}{k^2(2j-1)^2}\left(\sum_{p=1}^{k}\sum_{h=1-j}^{j-1}\varphi_p - \varphi_{i+h}\right)^2 \leq \frac{1}{k(2j-1)}\sum_{p=1}^{k}\sum_{h=1-j}^{j-1}(\varphi_p - \varphi_{i+h})^2.$$

Summing over all i and j gives

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{\ell-1}\left[\frac{a - a_{i,j}}{\ell - 1}\right]^2 &\leq \frac{1}{(\ell-1)^2}\sum_{i=1}^{k}\sum_{j=1}^{\ell-1}\frac{1}{k(2j-1)}\sum_{p=1}^{k}\sum_{h=1-j}^{j-1}(\varphi_p - \varphi_{i+h})^2 \\ &= \frac{k}{4(\ell-1)^2}\sum_{j=1}^{\ell-1}\frac{1}{2j-1}\sum_{h=1-j}^{j-1}\sum_{i,p=1}^{k}\frac{(\varphi_p - \varphi_{i+h})^2}{k^2/4} \\ &\leq \frac{k}{4(\ell-1)}|\varphi|_\Gamma^2 \;\leq\; c|\varphi|_\Gamma^2.\end{aligned}$$

This completes the analysis of the first term. For the second term, we have

$$a_{i,j+1} - a_{i,j} = \frac{1}{2j+1}\left[\varphi_{i+j} + \varphi_{i-j} - \frac{2}{2j-1}\sum_{h=1-j}^{j-1}\varphi_{i+h}\right].$$

Next, we note that

$$\begin{aligned}\left|\varphi_{i+j} - \frac{\varphi_i}{2j-1} - \frac{2}{2j-1}\sum_{h=1}^{j-1}\varphi_{i+h}\right| &= \left|\frac{\varphi_{i+j} - \varphi_i}{2j-1} + 2\sum_{h=1}^{j-1}\frac{\varphi_{i+j} - \varphi_{i+h}}{2j-1}\right| \\ &\leq 2\sum_{h=0}^{j-1}\frac{|\varphi_{i+j} - \varphi_{i+h}|}{2j-1},\end{aligned}$$

and, similarly,

$$\left|\varphi_{i-j} - \frac{\varphi_i}{2j-1} - \frac{2}{2j-1}\sum_{h=1}^{j-1}\varphi_{i-h}\right| = \left|\frac{\varphi_{i-j}-\varphi_i}{2j-1} + 2\sum_{h=1}^{j-1}\frac{\varphi_{i-j}-\varphi_{i-h}}{2j-1}\right| \leq 2\sum_{h=0}^{j-1}\frac{|\varphi_{i-j}-\varphi_{i-h}|}{2j-1}.$$

Hence,

$$\sum_{j=1}^{l-1}(a_{i,j+1}-a_{i,j})^2 \leq \sum_{j=1}^{l-1}\frac{8}{(2j+1)^2}\left[\left(\sum_{h=0}^{j-1}\frac{|\varphi_{i+j}-\varphi_{i+h}|}{2j-1}\right)^2 + \left(\sum_{h=0}^{j-1}\frac{|\varphi_{i-j}-\varphi_{i-h}|}{2j-1}\right)^2\right].$$

Once we sum over all i, the sum of the first and second term are identical, and therefore

$$\sum_{i=1}^{k}\sum_{j=1}^{l-1}(a_{i,j+1}-a_{i,j})^2 \leq 16\sum_{i=1}^{k}\sum_{j=1}^{l-1}\left(\sum_{h=0}^{j-1}\frac{|\varphi_{i+j}-\varphi_{i+h}|}{(2j-1)(2j+1)}\right)^2.$$

We have

$$\sum_{h=0}^{j-1}\frac{|\varphi_{i+j}-\varphi_{i+h}|}{(2j-1)(2j+1)} \leq \frac{1}{3j}\sum_{p=i}^{i+j-1}\frac{|\varphi_{i+j}-\varphi_p|}{j} \leq \frac{1}{3j}\sum_{p=i}^{i+j-1}\frac{|\varphi_{i+j}-\varphi_p|}{i+j-p},$$

which implies that

$$\begin{aligned}16\sum_{i=1}^{k}\sum_{j=1}^{l-1}\left(\sum_{h=0}^{j-1}\frac{|\varphi_{i+j}-\varphi_{i+h}|}{(2j-1)(2j+1)}\right)^2 &\leq \frac{16}{9}\sum_{i=1}^{k}\sum_{j=1}^{l-1}\left(\frac{1}{j}\sum_{p=i}^{i+j-1}\frac{|\varphi_{i+j}-\varphi_p|}{i+j-p}\right)^2 \\ &\leq \frac{16}{9}\sum_{q=1}^{k+\ell-1}\sum_{m=1}^{q-1}\left(\frac{1}{q-m}\sum_{p=m}^{q-1}\frac{|\varphi_q-\varphi_p|}{q-p}\right)^2,\end{aligned}$$

where $q = i+j$ and $m = i$. Letting $r = q-m$, $s = q-p$, and using Hardy's inequality [50,

Theorem 326], we obtain

$$\begin{aligned}\frac{16}{9}\sum_{q=1}^{k+\ell-1}\sum_{m=1}^{q-1}\left(\frac{1}{q-m}\sum_{p=m}^{q-1}\frac{|\varphi_q-\varphi_p|}{q-p}\right)^2 &= \frac{16}{9}\sum_{q=1}^{k+\ell-1}\sum_{r=1}^{q-1}\left(\frac{1}{r}\sum_{s=1}^{r}\frac{|\varphi_q-\varphi_{q-s}|}{s}\right)^2\\ &\le \frac{64}{9}\sum_{q=1}^{k+\ell-1}\sum_{r=1}^{q-1}\left[\frac{\varphi_q-\varphi_{q-r}}{r}\right]^2\\ &\le \frac{64}{9}\sum_{q=1}^{k+\ell-1}\sum_{r=1}^{q-1}\left[\frac{\varphi_q-\varphi_{q-r}}{d_G\big((q,1),(q-r,1)\big)}\right]^2\\ &\le \frac{256}{9}|\varphi|_\Gamma^2,\end{aligned}$$

where we note that the sum over the indices q and r consists of some amount of over-counting, with some terms $(\varphi(q)-\varphi(q-r))^2$ appearing up to four times. Improved estimates can be obtained by noting that the choice of indexing for Γ is arbitrary (see [124] for details), but, for the sake of readability, we focus on simplicity over tight constants.

Combining our estimates and noting that $c \ge 3$, we obtain the desired result

$$|u|_G \le \sqrt{1+2\left(c+\frac{256}{9}\right)}\,|\varphi|_\Gamma = \sqrt{2c+\frac{521}{9}}\,|\varphi|_\Gamma \le 5\sqrt{c}\,|\varphi|_\Gamma.$$

□

Combining Lemmas 10 and 11, we obtain our desired trace theorem.

Theorem 12. *Let $G = G_{k,\ell}$, $4\ell < k < 2c\ell$, $c \in \mathbb{N}$, with boundary $\Gamma = \{(i,1)\}_{i=1}^k$. For any $\varphi \in \mathbb{R}^k$,*

$$\frac{1}{4\sqrt{c}}\,|\varphi|_\Gamma \le \min_{u|_\Gamma=\varphi}|u|_G \le 5\sqrt{c}\,|\varphi|_\Gamma.$$

With a little more work, we can prove a similar result for a slightly more general class of graphs. Using Theorem 12, we can almost immediately prove a trace theorem for any graph H satisfying $G_{k,\ell} \subset H \subset G_{k,\ell}^*$. In fact, Lemma 10 carries over immediately. In order to prove a new version of Lemma 11, it suffices to bound the energy of u on the edges in

$G^*_{k,\ell}$ not contained in $G_{k,\ell}$. By Cauchy-Schwarz,

$$\begin{aligned} |u|^2_{G^*} &= |u|^2_G + \sum_{i=1}^{k}\sum_{j=1}^{\ell-1}\Big[(u_{i,j+1}-u_{i-1,j})^2 + (u_{i,j+1}-u_{i+1,j})^2\Big] \\ &\le 3\sum_{i=1}^{k}\sum_{j=1}^{\ell}(u_{i+1,j}-u_{i,j})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{\ell-1}(u_{i,j+1}-u_{i,j})^2 \le 3\,|u|^2_G, \end{aligned}$$

and therefore Corollary 1 follows immediately from Lemmas 10 and 11.

Corollary 1. *Let H satisfy $G_{k,\ell} \subset H \subset G^*_{k,\ell}$, $4\ell < k < 2c\ell$, $c \in \mathbb{N}$, with boundary $\Gamma = \{(i,1)\}_{i=1}^k$. For any $\varphi \in \mathbb{R}^k$,*

$$\frac{1}{4\sqrt{c}}\,|\varphi|_\Gamma \le \min_{u|_\Gamma=\varphi} |u|_H \le 5\sqrt{3c}\,|\varphi|_\Gamma.$$

Trace Theorems for General Graphs

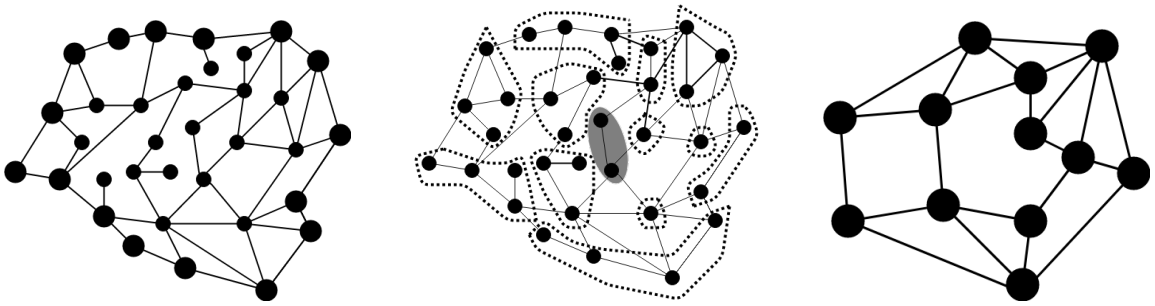
In order to extend Corollary 1 to more general graphs, we introduce a graph operation which is similar in concept to an aggregation (a partition of V into connected subsets) in which the size of aggregates is bounded. In particular, we give the following definition.

Definition 2. *The graph H, $G_{k,\ell} \subset H \subset G^*_{k,\ell}$, is said to be an M-aggregation of $(G,\Gamma) \in \mathcal{G}_n$ if there exists a partition $\mathcal{A} = a_* \cup \{a_{i,j}\}_{i=1,\dots,k}^{j=1,\dots,\ell}$ of $V(G)$ satisfying*

1. *$G[a_{i,j}]$ is connected and $|a_{i,j}| \le M$ for all $i = 1,\dots,k$, $j = 1,\dots,\ell$,*
2. *$\Gamma \subset \bigcup_{i=1}^k a_{i,1}$, and $\Gamma \cap a_{i,1} \neq \emptyset$ for all $i = 1,\dots,k$,*
3. *$N_G(a_*) \subset a_* \cup \bigcup_{i=1}^k a_{i,\ell}$,*
4. *the aggregation graph of $\mathcal{A}\backslash a_*$, given by*

$$(\mathcal{A}\backslash a_*, \{(a_{i_1,j_1}, a_{i_2,j_2}) \mid N_G(a_{i_1,j_1}) \cap a_{i_2,j_2} \neq 0\}),$$

is isomorphic to H.



(a) graph (G, Γ) (b) partition $\mathcal{A}$ (c) $G_{6,2} \subset H \subset G^*_{6,2}$

Figure 4-4: An example of an M-aggregation. Figure (A) provides a visual representation of a graph G, with boundary vertices Γ enlarged. Figure (B) shows a partition $\mathcal{A}$ of G, in which each aggregate (enclosed by dotted lines) has order at most four. The set a_* is denoted by a shaded region. Figure (C) shows the aggregation graph H of $\mathcal{A}\backslash a_*$. The graph H satisfies $G_{6,2} \subset H \subset G^*_{6,2}$, and is therefore a 4-aggregation of (G, Γ).

We provide a visual example in Figure 4-4, and later show that this operation applies to a fairly large class of graphs. However, the M-aggregation procedure is not the only operation for which we can control the behavior of the energy and boundary semi-norms. For instance, the behavior of our semi-norms under the deletion of some number of edges can be bounded easily if each edge can be replaced by a path of constant length, and no remaining edge is contained in more than a constant number of such paths. In addition, the behavior of these semi-norms under the disaggregation of large degree vertices is also relatively well-behaved (see [52] for details). Here, we focus on the M-aggregation procedure. We give the following result regarding graphs (G, Γ) for which some H, $G_{k,\ell} \subset H \subset G^*_{k,\ell}$, is an M-aggregation of (G, Γ), but note that a large number of minor refinements are possible, such as the two briefly mentioned above.

Theorem 13. *If H, $G_{k,\ell} \subset H \subset G^*_{k,\ell}$, $4\ell < k < 2c\ell$, $c \in \mathbb{N}$, is an M-aggregation of $(G, \Gamma) \in \mathcal{G}_n$, then for any $\varphi \in \mathbb{R}^{n_\Gamma}$,*

$$\frac{1}{C_1} |\varphi|_\Gamma \leq \min_{u|_\Gamma = \varphi} |u|_G \leq C_2 |\varphi|_\Gamma$$

for some fixed constants $C_1, C_2 > 0$ that depend only on c and M.

Proof. We first prove that there is an extension u of φ which satisfies $|u|_G \leq C_2|\varphi|_\Gamma$ for some $C_2 > 0$ that depends only on c and M. To do so, we define auxiliary functions $\widehat{u}$ and $\widehat{\varphi}$ on $(G^*_{2k,\ell}, \Gamma_{2k,\ell})$. Let

$$\widehat{\varphi}(p) = \begin{cases} \max_{q \in \Gamma \cap a_{(p+1)/2,1}} \varphi(q) & \text{if } p \text{ is odd,} \\ \min_{q \in \Gamma \cap a_{p/2,1}} \varphi(q) & \text{if } p \text{ is even,} \end{cases}$$

and $\widehat{u}$ be extension (4.4) of $\widehat{\varphi}$. The idea is to upper bound the semi-norm for u by $\widehat{u}$, for $\widehat{u}$ by $\widehat{\varphi}$ (using Corollary 1), and for $\widehat{\varphi}$ by φ. On each aggregate $a_{i,j}$, let u take values between $\widehat{u}(2i-1,j)$ and $\widehat{u}(2i,j)$, and let u equal a on a_*. We can decompose $|u|^2_G$ into

$$\begin{aligned} |u|^2_G \;=\; & \sum_{i=1}^{k}\sum_{j=1}^{\ell} \sum_{\substack{p,q \in a_{i,j}, \\ p \sim q}} (u(p)-u(q))^2 + \sum_{i=1}^{k}\sum_{j=1}^{\ell} \sum_{\substack{p \in a_{i,j}, \\ q \in a_{i+1,j}, \\ p \sim q}} (u(p)-u(q))^2 \\ & + \sum_{i=1}^{k}\sum_{j=1}^{\ell-1} \sum_{\substack{p \in a_{i,j}, \\ q \in a_{i-1,j+1}, \\ p \sim q}} (u(p)-u(q))^2 + \sum_{i=1}^{k}\sum_{j=1}^{\ell-1} \sum_{\substack{p \in a_{i,j}, \\ q \in a_{i+1,j+1}, \\ p \sim q}} (u(p)-u(q))^2 \\ & + \sum_{i=1}^{k}\sum_{j=1}^{\ell-1} \sum_{\substack{p \in a_{i,j}, \\ q \in a_{i,j+1}, \\ p \sim q}} (u(p)-u(q))^2, \end{aligned}$$

and bound each term of $|u|^2_G$ separately, beginning with the first. The maximum energy semi-norm of an m vertex graph that takes values in the range $[a,b]$ is bounded above by $(m/2)^2(b-a)^2$. Therefore,

$$\sum_{\substack{p,q \in a_{i,j}, \\ p \sim q}} (u(p)-u(q))^2 \leq \frac{M^2}{4} \left(\widehat{u}(2i-1,j) - \widehat{u}(2i,j)\right)^2.$$

For the second term,

$$\begin{aligned}\sum_{\substack{p\in a_{i,j},\\ q\in a_{i+1,j},\\ p\sim q}} (u(p)-u(q))^2 &\leq M^2 \max_{\substack{i_1\in\{2i-1,2i\},\\ i_2\in\{2i+1,2i+2\}}} (\widehat{u}(i_1,j)-\widehat{u}(i_2,j))^2 \\ &\leq 3M^2\big[(\widehat{u}(2i-1,j)-\widehat{u}(2i,j))^2+(\widehat{u}(2i,j)-\widehat{u}(2i+1,j))^2 \\ &\quad +(\widehat{u}(2i+1,j)-\widehat{u}(2i+2,j))^2\big].\end{aligned}$$

The exact same type of bound holds for the third and fourth terms. For the fifth term,

$$\sum_{\substack{p\in a_{i,j},\\ q\in a_{i,j+1},\\ p\sim q}} (u(p)-u(q))^2 \leq M^2 \max_{\substack{i_1\in\{2i-1,2i\},\\ i_2\in\{2i-1,2i\}}} (\widehat{u}(i_1,j)-\widehat{u}(i_2,j+1))^2,$$

and, unlike terms two, three, and four, this maximum appears in $|\widehat{u}|^2_{G^*_{2k,\ell}}$. Combining these three estimates, we obtain the upper bound $|u|^2_G \leq (6\times 3+\frac{1}{4})M^2\,|\widehat{u}|^2_{G^*_{2k,\ell}} = \frac{73}{4}M^2\,|\widehat{u}|^2_{G^*_{2k,\ell}}$.

Next, we lower bound $|\varphi|_\Gamma$ by a constant times $|\widehat{\varphi}|_{\Gamma_{2k,\ell}}$. By definition, in $\Gamma\cap a_{i,1}$ there is a vertex that takes value $\widehat{\varphi}(2i-1)$ and a vertex that takes value $\widehat{\varphi}(2i)$. This implies that every term in $|\widehat{\varphi}|_{\Gamma_{2k,\ell}}$ is a term in $|\varphi|_\Gamma$, with a possibly different denominator. Distances between vertices on Γ can be decreased by at most a factor of $2M$ on $\Gamma_{2k,\ell}$. In addition, it may be the case that an aggregate contains only one vertex of Γ, which results in $\widehat{\varphi}(2i-1)=\widehat{\varphi}(2i)$. Therefore, a given term in $|\varphi|^2_\Gamma$ could appear four times in $|\widehat{\varphi}|^2_{\Gamma_{2k,\ell}}$. Combining these two facts, we immediately obtain the bound $|\widehat{\varphi}|^2_{\Gamma_{2k,\ell}}\leq 16M^2|\varphi|^2_\Gamma$. Combining these two estimates with Corollary 1 completes the first half of the proof.

All that remains is to show that for any u, $|\varphi|_\Gamma\leq C_1|u|_G$ for some $C_1>0$ that depends only on c and M. To do so, we define auxiliary functions $\widetilde{u}$ and $\widetilde{\varphi}$ on $(G_{2k,2\ell},\Gamma_{2k,2\ell})$. Let

$$\widetilde{u}(i,j)=\begin{cases}\max_{p\in a_{\lceil i/2\rceil,\lceil j/2\rceil}} u(p) & \text{if } i=j \mod 2,\\ \min_{p\in a_{\lceil i/2\rceil,\lceil j/2\rceil}} u(p) & \text{if } i\neq j \mod 2.\end{cases}$$

Here, the idea is to lower bound the semi-norm for u by $\widetilde{u}$, for $\widetilde{u}$ by $\widetilde{\varphi}$ (using Corollary 1),

and for $\widetilde{\varphi}$ by φ. We can decompose $|\widetilde{u}|^2_{G_{2k,2\ell}}$ into

$$\begin{aligned}|\widetilde{u}|^2_{G_{2k,2\ell}} &= 4\sum_{i=1}^{k}\sum_{j=1}^{\ell}(\widetilde{u}(2i-1,2j-1)-\widetilde{u}(2i,2j-1))^2 \\ &+\sum_{i=1}^{k}\sum_{j=1}^{\ell}(\widetilde{u}(2i,2j-1)-\widetilde{u}(2i+1,2j-1))^2+(\widetilde{u}(2i,2j)-\widetilde{u}(2i+1,2j))^2 \\ &+\sum_{i=1}^{k}\sum_{j=1}^{\ell-1}(\widetilde{u}(2i-1,2j)-\widetilde{u}(2i-1,2j+1))^2+(\widetilde{u}(2i,2j)-\widetilde{u}(2i,2j+1))^2.\end{aligned}$$

The minimum squared energy semi-norm of an m vertex graph that takes value a at some vertex and value b at some vertex is bounded below by $(b-a)^2/m$. Therefore,

$$(\widetilde{u}(2i-1,2j-1)-\widetilde{u}(2i,2j-1))^2 \leq M\sum_{\substack{p,q\in a_{i,j},\\ p\sim q}}(u(p)-u(q))^2,$$

and

$$\max_{\substack{i_1\in\{2i-1,2i\},\\ i_2\in\{2i+1,2i+2\}}}(\widetilde{u}(i_1,2j)-\widetilde{u}(i_2,2j))^2 \leq 2M\sum_{\substack{p,q\in a_{i,j}\cup a_{i+1,j},\\ p\sim q}}(u(p)-u(q))^2.$$

Repeating the above estimate for each term results in the desired bound.

Next, we upper bound $|\varphi|_\Gamma$ by a constant multiple of $|\widetilde{\varphi}|_{\Gamma_{2k,2\ell}}$. We can write $|\varphi|^2_\Gamma$ as

$$|\varphi|^2_\Gamma=\sum_{i=1}^{k}\sum_{p,q\in\Gamma\cap a_{i,1}}\frac{(\varphi(p)-\varphi(q))^2}{d^2_G(p,q)}+\sum_{i_1=1}^{k-1}\sum_{i_2=i_1+1}^{k}\sum_{\substack{p\in\Gamma\cap a_{i_1,1},\\ q\in\Gamma\cap a_{i_2,1}}}\frac{(\varphi(p)-\varphi(q))^2}{d^2_G(p,q)},$$

and bound each term separately. The first term is bounded by

$$\sum_{p,q\in\Gamma\cap a_{i,1}}\frac{(\varphi(p)-\varphi(q))^2}{d^2_G(p,q)}\leq\frac{M^2}{4}(\widetilde{\varphi}(2i-1)-\widetilde{\varphi}(2i))^2.$$

For the second term, we first note that $d_G(p,q)\geq\frac{1}{3}d_{\Gamma_{2k,2\ell}}((m_1,1),(m_2,1))$ for $p\in\Gamma\cap a_{i_1,1}$, $q\in\Gamma\cap a_{i_2,1}$, $m_1\in\{2i_1-1,2i_1\}$, $m_2\in\{2i_2-1,2i_2\}$, which allows us to bound the

second term by

$$\sum_{\substack{p\in\Gamma\cap a_{i_1,1},\\ q\in\Gamma\cap a_{i_2,1}}} \frac{(\varphi(p)-\varphi(q))^2}{d_G^2(p,q)} \leq 9M^2 \max_{\substack{m_1\in\{2i_1-1,2i_1\},\\ m_2\in\{2i_2-1,2i_2\}}} \frac{(\widetilde{\varphi}(m_1)-\widetilde{\varphi}(m_2))^2}{d^2_{\Gamma_{2k,2\ell}}(m_1,m_2)}.$$

Combining the lower bound for u in terms of $\widetilde{u}$, the upper bound for ϕ in terms of $\widetilde{\phi}$, and Corollary 1 completes the second half of the proof. □

The proof of Theorem 13 also immediately implies a similar result. Let $\widetilde{L} \in \mathbb{R}^{n_\Gamma \times n_\Gamma}$ be the Laplacian of the complete graph on Γ with weights $w(i,j) = d_\Gamma^{-2}(i,j)$. The same proof implies the following.

Corollary 2. *If H, $G_{k,\ell} \subset H \subset G^*_{k,\ell}$, $4\ell < k < 2c\ell$, $c \in \mathbb{N}$, is an M-aggregation of $(G,\Gamma) \in \mathcal{G}_n$, then for any $\varphi \in \mathbb{R}^{n_\Gamma}$,*

$$\frac{1}{C_1}\langle \widetilde{L}\varphi, \varphi\rangle^{1/2} \leq \min_{u|_\Gamma=\varphi} |u|_G \leq C_2 \langle \widetilde{L}\varphi,\varphi\rangle^{1/2},$$

for some fixed constants $C_1, C_2 > 0$ that depend only on c and M.

Spectral Equivalence of S_Γ and $L_\Gamma^{1/2}$

By Corollary 2, and the property $\langle \varphi, S_\Gamma\varphi\rangle = \min_{u|_\Gamma=\varphi}|u|_G^2$ (see Proposition 8), in order to prove spectral equivalence between S_Γ and $L_\Gamma^{1/2}$, it suffices to show that $L_\Gamma^{1/2}$ and $\widetilde{L}$ are spectrally equivalent. This can be done relatively easily, and leads to a proof of the main result of the section.

Theorem 14. *If H, $G_{k,\ell} \subset H \subset G^*_{k,\ell}$, $4\ell < k < 2c\ell$, $c \in \mathbb{N}$, is an M-aggregation of $(G,\Gamma) \in \mathcal{G}_n$, then for any $\varphi \in \mathbb{R}^{n_\Gamma}$,*

$$\frac{1}{C_1}\langle L_\Gamma^{1/2}\varphi,\varphi\rangle \leq \langle S_\Gamma\varphi,\varphi\rangle \leq C_2\langle L_\Gamma^{1/2}\varphi,\varphi\rangle,$$

for some fixed constants $C_1, C_2 > 0$ that depend only on c and M.

Proof. Let $\phi(i,j) = \min\{i-j \mod n_\Gamma,\ j-i \mod n_\Gamma\}$. $G[\Gamma]$ is a cycle, so $\widetilde{L}(i,j) = -\phi(i,j)^{-2}$ for $i \neq j$. The spectral decomposition of L_Γ is well known, namely,

$$L_\Gamma = \sum_{k=1}^{\lfloor \frac{n_\Gamma}{2} \rfloor} \lambda_k(L_\Gamma) \left[\frac{x_k x_k^T}{\|x_k\|^2} + \frac{y_k y_k^T}{\|y_k\|^2} \right],$$

where $\lambda_k(L_\Gamma) = 2 - 2\cos\frac{2\pi k}{n_\Gamma}$ and $x_k(j) = \sin\frac{2\pi kj}{n_\Gamma}$, $y_k(j) = \cos\frac{2\pi kj}{n_\Gamma}$, $j = 1, ..., n_\Gamma$. If n_Γ is odd, then $\lambda_{(n_\Gamma-1)/2}$ has multiplicity two, but if n_Γ is even, then $\lambda_{n_\Gamma/2}$ has only multiplicity one, as $x_{n_\Gamma/2} = 0$. If $k \neq n_\Gamma/2$, we have

$$\begin{aligned} \|x_k\|^2 &= \sum_{j=1}^{n_\Gamma} \sin^2\left(\frac{2\pi kj}{n_\Gamma}\right) = \frac{n_\Gamma}{2} - \frac{1}{2}\sum_{j=1}^{n_\Gamma} \cos\left(\frac{4\pi kj}{n_\Gamma}\right) \\ &= \frac{n_\Gamma}{2} - \frac{1}{4}\left[\frac{\sin(2\pi k(2+\frac{1}{n_\Gamma}))}{\sin\frac{2\pi k}{n_\Gamma}} - 1\right] = \frac{n_\Gamma}{2}, \end{aligned}$$

and so $\|y_k\|^2 = \frac{n_\Gamma}{2}$ as well. If $k = n_\Gamma/2$, then $\|y_k\|^2 = n_\Gamma$. If n_Γ is odd,

$$\begin{aligned} L_\Gamma^{1/2}(i,j) &= \frac{2\sqrt{2}}{n_\Gamma} \sum_{k=1}^{\frac{n_\Gamma-1}{2}} \left[1 - \cos\frac{2k\pi}{n_\Gamma}\right]^{1/2} \left[\sin\frac{2\pi ki}{n_\Gamma}\sin\frac{2\pi kj}{n_\Gamma} - \cos\frac{2\pi ki}{n_\Gamma}\cos\frac{2\pi kj}{n_\Gamma}\right] \\ &= \frac{4}{n_\Gamma} \sum_{k=1}^{\frac{n_\Gamma-1}{2}} \sin\left(\frac{\pi}{2}\frac{2k}{n_\Gamma}\right)\cos\left(\phi(i,j)\pi\frac{2k}{n_\Gamma}\right) \\ &= \frac{2}{n_\Gamma} \sum_{k=0}^{n_\Gamma} \sin\left(\frac{\pi}{2}\frac{2k}{n_\Gamma}\right)\cos\left(\phi(i,j)\pi\frac{2k}{n_\Gamma}\right), \end{aligned}$$

and if n_Γ is even,

$$\begin{aligned} L_\Gamma^{1/2}(i,j) &= \frac{2}{n_\Gamma}(-1)^{i+j} + \frac{4}{n_\Gamma} \sum_{k=1}^{\frac{n_\Gamma}{2}-1} \sin\left(\frac{\pi}{2}\frac{2k}{n_\Gamma}\right)\cos\left(\phi(i,j)\pi\frac{2k}{n_\Gamma}\right) \\ &= \frac{2}{n_\Gamma} \sum_{k=0}^{n_\Gamma} \sin\left(\frac{\pi}{2}\frac{2k}{n_\Gamma}\right)\cos\left(\phi(i,j)\pi\frac{2k}{n_\Gamma}\right). \end{aligned}$$

$L_\Gamma^{1/2}(i,j)$ is simply the trapezoid rule applied to the integral of $\sin(\frac{\pi}{2}x)\cos(\phi(i,j)\pi x)$ on

the interval $[0, 2]$. Therefore,

$$\left| L_\Gamma^{1/2}(i,j) + \frac{2}{\pi(4\phi(i,j)^2 - 1)} \right| = \left| L_\Gamma^{1/2}(i,j) - \int_0^2 \sin\left(\frac{\pi}{2}x\right) \cos\left(\phi(i,j)\pi x\right) dx \right| \leq \frac{2}{3n_\Gamma^2},$$

where we have used the fact that if $f \in C^2([a,b])$, then

$$\left| \int_a^b f(x)dx - \frac{f(a)+f(b)}{2}(b-a) \right| \leq \frac{(b-a)^3}{12} \max_{\xi\in[a,b]} |f''(\xi)|.$$

Noting that $n_\Gamma \geq 3$, it quickly follows that

$$\left(\frac{1}{2\pi} - \frac{\sqrt{2}}{12}\right)\langle \tilde{L}\varphi, \varphi\rangle \leq \langle L_\Gamma^{1/2}\varphi, \varphi\rangle \leq \left(\frac{2}{3\pi} + \frac{\sqrt{2}}{27}\right)\langle \tilde{L}\varphi, \varphi\rangle.$$

Combining this result with Corollary 2, and noting that $\langle \varphi, S_\Gamma \varphi\rangle = |\widehat{u}|_G^2$, where $\widehat{u}$ is the harmonic extension of φ, we obtain the desired result. □

An Illustrative Example

While the concept of a graph (G, Γ) having some H, $G_{k,\ell} \subset H \subset G_{k,\ell}^*$, as an M-aggregation seems somewhat abstract, this simple formulation in itself is quite powerful. As an example, we illustrate that this implies a trace theorem (and, therefore, spectral equivalence) for all three-connected planar graphs with bounded face degree (number of edges in the associated induced cycle) and for which there exists a planar spring embedding with a convex hull that is not too thin (a bounded distance to Hausdorff distance ratio for the boundary with respect to some point in the convex hull) and satisfies bounded edge length and small angle conditions. Let $\mathcal{G}_n^{f\leq c}$ be the elements of $(G, \Gamma) \in \mathcal{G}_n$ for which every face other than the outer face Γ has at most c edges. The exact proof the following theorem is rather long, and can be found in the Appendix of [124]. The intuition is that by taking the planar spring embedding, rescaling it so that the vertices of the boundary face lie on the unit circle, and overlaying a natural embedding of $G_{k,\ell}^*$ on the unit disk, an intuitive M-aggregation of the graph G can be formed.

Theorem 15. *If there exists a planar spring embedding X of $(G, \Gamma) \in \mathcal{G}_n^{f\leq c_1}$ for which*

(1) $K = \mathit{conv}\left(\{[X_\Gamma]_{i,.}\}_{i=1}^{n_\Gamma}\right)$ *satisfies*

$$\sup_{u \in K} \inf_{v \in \partial K} \sup_{w \in \partial K} \frac{\|u - v\|}{\|u - w\|} \geq c_2 > 0,$$

(2) X *satisfies*

$$\max_{\substack{\{i_1,i_2\} \in E \\ \{j_1,j_2\} \in E}} \frac{\|X_{i_1,.} - X_{i_2,.}\|}{\|X_{j_1,.} - X_{j_2,.}\|} \leq c_3 \quad \textit{and} \quad \min_{\substack{i \in V \\ j_1, j_2 \in N(i)}} \angle X_{j_1,.}\, X_{i,.}\, X_{j_2,.} \geq c_4 > 0,$$

then there exists an H, $G_{k,\ell} \subset H \subset G^*_{k,\ell}$, $\ell \leq k < 2c\ell$, $c \in \mathbb{N}$, *such that* H *is an* M*-aggregation of* (G, Γ), *where* c *and* M *are constants that depend on* c_1, c_2, c_3, *and* c_4.

4.3 The Kamada-Kawai Objective and Optimal Layouts

Given the distances between data points in a high dimensional space, how can we meaningfully visualize their relationships? This is a fundamental task in exploratory data analysis, for which a variety of different approaches have been proposed. Many of these techniques seek to visualize high-dimensional data by embedding it into lower dimensional, e.g. two or three-dimensional, space. Metric multidimensional scaling (MDS or mMDS) [64, 66] is a classical approach that attempts to find a low-dimensional embedding that accurately represents the distances between points. Originally motivated by applications in psychometrics, MDS has now been recognized as a fundamental tool for data analysis across a broad range of disciplines. See [66, 9] for more details, including a discussion of applications to data from scientific, economic, political, and other domains. Compared to other classical visualization tools like PCA[2], metric multidimensional scaling has the advantage of not being restricted to linear projections of the data, and of being applicable to data from an arbitrary metric space, rather than just Euclidean space. Because of this versatility, MDS has also become one of the most popular algorithms in the field of graph drawing, where the goal is to visualize relationships between nodes (e.g. people in a social network). In this context, MDS was independently proposed by Kamada and Kawai [55] as a force-directed graph drawing method.

In this section, we consider the algorithmic problem of computing an optimal drawing under the MDS/Kamada-Kawai objective, and structural properties of optimal drawings. The Kamada-Kawai objective is to minimize the following energy/stress functional $E : \mathbb{R}^{rn} \to \mathbb{R}$

$$E(\vec{x}_1, \ldots, \vec{x}_n) = \sum_{i<j} \left(\frac{\|\vec{x}_i - \vec{x}_j\|}{d(i,j)} - 1 \right)^2, \tag{4.5}$$

which corresponds to the physical situation where $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^r$ are particles and for each $i \neq j$, particles $\vec{x}_i$ and $\vec{x}_j$ are connected by an idealized spring with equilibrium length $d(i,j)$ following Hooke's law with spring constant $k_{ij} = \frac{1}{d(i,j)^2}$. In applications to visualization,

[2] In the literature, PCA is sometimes referred to as classical multidimensional scaling, in contrast to metric multidimensional scaling, which we study in this work.

the choice of dimension is often small, i.e. $r \leq 3$. We also note that in (4.5) the terms in the sum are sometimes re-weighted with vertex or edge weights, which we discuss in more detail later.

In practice, the MDS/Kamada-Kawai objective (4.5) is optimized via a heuristic procedure like gradient descent [65, 135] or stress majorization [29, 43]. Because the objective is non-convex, these algorithms may not reach the global minimum, but instead may terminate at approximate critical points of the objective function. Heuristics such as restarting an algorithm from different initializations and using modified step size schedules have been proposed to improve the quality of results. In practice, these heuristic methods do seem to work well for the Kamada-Kawai objective and are implemented in popular packages like GRAPHVIZ [39] and the SMACOF package in R.

We revisit this problem from an approximation algorithms perspective. First, we resolve the computational complexity of minimizing (4.5) by proving that finding the global minimum is NP-hard, even for graph metrics (where the metric is the shortest path distance on a graph). Consider the gap decision version of stress minimization over graph metrics, which we formally define below:

GAP STRESS MINIMIZATION

Input: Graph $G = ([n], E)$, $r \in \mathbb{N}$, $L \geq 0$.

Output: TRUE if there exists $\vec{x} = (\vec{x}_1, \ldots, \vec{x}_n) \in \mathbb{R}^{nr}$ such that $E(\vec{x}) \leq L$; FALSE if for every $\vec{x}$, $E(\vec{x}) \geq L + 1$.

Theorem 16. Gap Stress Minimization *in dimension $r = 1$ is NP-hard. Furthermore, the problem is hard even restricted to input graphs with diameter bounded by an absolute constant.*

As a gap problem, the output is allowed to be arbitrary if neither case holds; the hardness of the gap formulation shows that there cannot exist a Fully-Polynomial Randomized Approximation Scheme (FPRAS) for this problem if P $\neq$ NP, i.e. the runtime cannot be polynomial in the desired approximation guarantee. Our reduction shows this problem is

hard even when the input graph has diameter bounded above by an absolute constant. This is a natural setting to consider, since many real world graphs (for example, social networks [35]) and random graph models [127] indeed have low diameter due to the "small-world phenomena." Other key aspects of this hardness proof are that we show the problem is hard even when the input d is a graph metric, and we show it is hard even in its canonical unweighted formulation (4.5).

Given that computing the minimizer is NP-hard, a natural question is whether there exist polynomial time approximation algorithms for minimizing (4.5). We show that if the input graph has bounded diameter $D = O(1)$, then there indeed exists a Polynomial-Time Approximation Scheme (PTAS) to minimize (4.5), i.e. for fixed $\epsilon > 0$ and fixed D there exists an algorithm to approximate the global minimum of a n vertex diameter D graph up to multiplicative error $(1+\epsilon)$ in time $f(\epsilon, D) \cdot poly(n)$. More generally, we show the following result, where KKSCHEME is a simple greedy algorithm described in Subsection 4.3.3 below.

Theorem 17. *For any input metric over $[n]$ with $\min_{i,j\in[n]} d(i,j) = 1$ and any $R > \epsilon > 0$, Algorithm* KKSCHEME *with $\epsilon_1 = O(\epsilon/R)$ and $\epsilon_2 = O(\epsilon/R^2)$ runs in time $n^2(R/\epsilon)^{O(rR^4/\epsilon^2)}$ and outputs $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^r$ with $\|\vec{x}_i\| \leq R$ such that*

$$\mathbb{E}\left[E(\vec{x}_1, \ldots, \vec{x}_n)\right] \leq E(\vec{x}_1^*, \ldots, \vec{x}_n^*) + \epsilon n^2$$

for any $\vec{x}_1^*, \ldots, \vec{x}_n^*$ with $\|\vec{x}_i^*\| \leq R$ for all i, where $\mathbb{E}$ is the expectation over the randomness of the algorithm.

The fact that this result is a PTAS for bounded diameter graphs follows from combining it with the two structural results regarding optimal Kamada-Kawai embeddings, which are of independent interest. The first (Lemma 13) shows that the optimal objective value for low diameter graphs must be of order $\Omega(n^2)$, and the second (Lemma 15) shows that the optimal KK embedding is "contractive" in the sense that the diameter of the output is never much larger than the diameter of the input. Both lemmas are proven in Subsection 4.3.1.

In the multidimensional scaling literature, there has been some study of the local convergence of algorithms like stress majorization (for example [28]) which shows that

stress majorization will converge quickly if in a sufficiently small neighborhood of a local minimum. This work seems to propose the first provable guarantees for global optimization. The closest previous hardness result is the work of [19], where they showed that a similar problem is hard. In their problem, the terms in (4.5) are weighted by $d(i,j)$, absolute value loss replaces the squared loss, and the input is an arbitrary pseudometric where nodes in the input are allowed to be at distance zero from each other. The second assumption makes the diameter (ratio of max to min distance in the input) infinite, which is a major obstruction to modifying their approach to show Theorem 16. See Remark 1 for further discussion. In the approximation algorithms literature, there has been a great deal of interest in optimizing the worst-case distortion of metric embeddings into various spaces, see e.g. [5] for approximation algorithms for embeddings into one dimension, and [34, 85] for more general surveys of low distortion metric embeddings. Though conceptually related, the techniques used in this literature are not generally targeted for minimizing a measure of average pairwise distortion like (4.5).

In what follows, we generally assume the input is given as an unweighted graph to simplify notation. However, for the upper bound (Theorem 17) we do handle the general case of arbitrary metrics with distances in $[1, D]$ (the lower bound of 1 is without loss of generality after re-scaling). In the lower bound (Theorem 16), we prove the (stronger) result that the problem is hard when restricted to graph metrics, instead of just for arbitrary metrics.

Throughout this section, we use techniques involving estimating different components of the objective function $E(\vec{x}_1, \ldots, \vec{x}_n)$ given by (4.5). For convenience, we use the notation

$$E_{i,j}(\vec{x}) := \left(\frac{\|\vec{x}_i - \vec{x}_j\|}{d(i,j)} - 1 \right)^2, \quad E_S(\vec{x}) := \sum_{\substack{i,j \in S \\ i<j}} E_{i,j}(\vec{x}), \quad E_{S,T}(\vec{x}) := \sum_{i \in S} \sum_{j \in T} E_{i,j}(\vec{x}).$$

The remainder of this section is as follows. In Subsection 4.3.1, we prove two structural results regarding the objective value and diameter of an optimal layout. In Subsection 4.3.2, we provide a proof of Theorem 16, the main algorithmic lower bound of the section. Finally, in Subsection 4.3.3, we formally describe an approximation algorithm for low diameter graphs and prove Theorem 17.

4.3.1 Structural Results for Optimal Embeddings

In this subsection, we present two results regarding optimal layouts of a given graph. In particular, we provide a lower bound for the energy of a graph layout and an upper bound for the diameter of an optimal layout. First, we recall the following standard ϵ-net estimate.

Lemma 12 (Corollary 4.2.13 of [126]). *Let $B_R = \{x : \|x\| \leq R\} \subset \mathbb{R}^r$ be the origin-centered radius R ball in r dimensions. For any $\epsilon \in (0, R)$ there exists a subset $S_\epsilon \subset B_R$ with $|S_\epsilon| \leq (3R/\epsilon)^r$ such that $\max_{\|x\| \leq R} \min_{y \in S_\epsilon} \|x - y\| \leq \epsilon$, i.e. S_ϵ is an ϵ-net of B_R.*

We use this result to prove a lower bound for the objective value of any layout of a diameter D graph in $\mathbb{R}^r$.

Lemma 13. *Let $G = ([n], E)$ have diameter*

$$D \leq \frac{(n/2)^{1/r}}{10}.$$

Then any layout $\vec{x} \in \mathbb{R}^{rn}$ has energy

$$E(\vec{x}) \geq \frac{n^2}{81(10D)^r}.$$

Proof. Let $G = ([n], E)$ have diameter $D \leq (n/2)^{1/r}/10$, and suppose that there exists a layout $\vec{x} \subset \mathbb{R}^r$ of G in dimension r with energy $E(\vec{x}) = cn^2$ for some $c \leq 1/810$. If no such layout exists, then we are done. We aim to lower bound the possible values of c. For each vertex $i \in [n]$, we consider the quantity $E_{i,V\setminus i}(\vec{x})$. The sum

$$\sum_{i \in [n]} E_{i,V\setminus i}(\vec{x}) = 2cn^2,$$

and so there exists some $i' \in [n]$ such that $E_{i',V\setminus i'}(\vec{x}) \leq 2cn$. By Markov's inequality,

$$\left|\{j \in [n] \,|\, E_{i',j}(\vec{x}) > 10c\}\right| < n/5,$$

and so at least $4n/5$ vertices (including i') in $[n]$ satisfy

$$\left(\frac{\|\vec{x}_{i'} - \vec{x}_j\|}{d(i', j)} - 1\right)^2 \leq 10c,$$

and also

$$\|\vec{x}_{i'} - \vec{x}_j\| \leq d(i', j)(1 + \sqrt{10c}) \leq \frac{10}{9}D.$$

The remainder of the proof consists of taking the d-dimensional ball with center $\vec{x}_{i'}$ and radius $10D/9$ (which contains $\geq 4n/5$ vertices), partitioning it into smaller sub-regions, and then lower bounding the energy resulting from the interactions between vertices within each sub-region.

By applying Lemma 12 with $R := 10D/9$ and $\epsilon := 1/3$, we may partition the r dimensional ball with center $\vec{x}_{i'}$ and radius $10D/9$ into $(10D)^r$ disjoint regions, each of diameter at most $2/3$. For each of these regions, we denote by $S_j \subset [n]$, $j \in [(10D)^r]$, the subset of vertices whose corresponding point lies in the corresponding region. As each region is of diameter at most $2/3$ and the graph distance between any two distinct vertices is at least one, either

$$E_{S_j}(\vec{x}) \geq \binom{|S_j|}{2}(2/3 - 1)^2 = \frac{|S_j|(|S_j| - 1)}{18}$$

or $|S_j| = 0$. Empty intervals provide no benefit and can be safely ignored. A lower bound on the total energy can be produced by the following optimization problem

$$\min \sum_{k=1}^{\ell} m_k(m_k - 1) \qquad \text{s.t.} \quad \sum_{k=1}^{\ell} m_k = m,\ m_k \geq 1,\ k \in [\ell],$$

where the m_k have been relaxed to real values. This optimization problem has a non-empty feasible region for $m \geq \ell$, and the solution is given by $m(m/\ell - 1)$ (achieved when $m_k = m/\ell$ for all k). In our situation, $m := 4n/5$ and $\ell := (10D)^r$, and, by assumption,

$m \geq \ell$. This leads to the lower bound

$$cn^2 = E(\vec{x}) \geq \sum_{j=1}^{\ell} E_{S_j}(\vec{x}) \geq \frac{4n}{90}\left[\frac{4n}{5(10D)^r} - 1\right],$$

which implies that

$$c \geq \frac{16}{450(10D)^r}\left(1 - \frac{5(10D)^r}{4n}\right) \geq \frac{1}{75(10D)^r}$$

for $D \leq (n/2)^{1/r}/10$. This completes the proof. □

The above estimate has the correct dependence for $r = 1$. For instance, consider the lexicographical product of a path P_D and a clique $K_{n/D}$ (i.e., a graph with D cliques in a line, and complete bipartite graphs between neighboring cliques). This graph has diameter D, and the layout in which the "vertices" (each corresponding to a copy of $K_{n/D}$) of P_D lie exactly at the integer values $[D]$ has objective value $\frac{n}{2}(n/D - 1)$. This estimate is almost certainly not tight for dimensions $r > 1$, as there is no higher dimensional analogue of the path (i.e., a graph with $O(D^r)$ vertices and diameter D that embeds isometrically in $\mathbb{R}^r$).

Next, we provide a proof of Lemma 15, which upper bounds the diameter of any optimal layout of a diameter D graph. However, before we proceed with the proof, we first prove an estimate for the concentration of points $\vec{x}_i$ at some distance from the marginal median in any optimal layout. The marginal median $\vec{m} \in \mathbb{R}^r$ of a set of points $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^r$ is the vector whose i^{th} component $\vec{m}(i)$ is the univariate median of $\vec{x}_1(i), \ldots, \vec{x}_n(i)$ (with the convention that, if n is even, the univariate median is given by the mean of the $(n/2)^{th}$ and $(n/2+1)^{th}$ ordered values). The below result, by itself, is not strong enough to prove our desired diameter estimate, but serves as a key ingredient in the proof of Lemma 15.

Lemma 14. *Let $G = ([n], E)$ have diameter D. Then, for any optimal layout $\vec{x} \in \mathbb{R}^{rn}$, i.e., $\vec{x}$ such that $E(\vec{x}) \leq E(\vec{y})$ for all $\vec{y} \in \mathbb{R}^{rn}$,*

$$\left|\{i \in [n] \,|\, \|\vec{m} - \vec{x}_i\|_\infty \geq (C+k)D\}\right| \leq 2rnC^{-\sqrt{2}^k}$$

for all $C > 0$ and $k \in \mathbb{N}$, where $\vec{m}$ is the marginal median of $\vec{x}_1, \ldots, \vec{x}_n$.

Proof. Let $G = ([n], E)$ have diameter D, and $\vec{x}$ be an optimal layout. Without loss of generality, we order the vertices so that $\vec{x}_1(1) \leq \cdots \leq \vec{x}_n(1)$, and shift $\vec{x}$ so that $\vec{x}_{\lfloor n/2 \rfloor}(1) = 0$. Next, we fix some $C > 0$, and define the subsets $S_0 := \big[\lfloor n/2 \rfloor\big]$ and

$$S_k := \big\{ i \in [n] \,|\, \vec{x}_i(1) \in \big[(C+k)D, (C+k+1)D\big) \big\},$$
$$T_k := \big\{ i \in [n] \,|\, \vec{x}_i(1) \geq (C+k)D \big\},$$

$k \in \mathbb{N}$. Our primary goal is to estimate the quantity $|T_k|$, for an arbitrary k, from which the desired result will follow quickly.

The objective value $E(\vec{x})$ is at most $\binom{n}{2}$; otherwise we could replace $\vec{x}$ by a layout with all $\vec{x}_i$ equal. To obtain a first estimate on $|T_k|$, we consider a crude lower bound on $E_{S_0,T_k}(\vec{x})$. We have

$$E_{S_0,T_k}(\vec{x}) \geq |S_0||T_k| \left(\frac{(C+k)D}{D} - 1 \right)^2 = (C+k-1)^2 \lfloor n/2 \rfloor |T_k|,$$

and therefore

$$|T_k| \leq \binom{n}{2} \frac{1}{(C+k-1)^2 \lfloor n/2 \rfloor} \leq \frac{n}{(C+k-1)^2}.$$

From here, we aim to prove the following claim

$$|T_k| \leq \frac{n}{\big(C+k-(2\ell-1)\big)^{2^\ell}}, \quad k, \ell \in \mathbb{N},\ k \geq 2\ell - 1, \tag{4.6}$$

by induction on ℓ. The above estimate serves as the base case for $\ell = 1$. Assuming the above statement holds for some fixed ℓ, we aim to prove that this holds for $\ell + 1$.

To do so, we consider the effect of collapsing the set of points in S_k, $k \geq 2\ell$, into a

hyperplane with a fixed first value. In particular, consider the alternate layout $\vec{x}'$ given by

$$\vec{x}'_i(j) = \begin{cases} (C+k)D & j = 1,\, i \in S_k \\ \vec{x}_i(j) - D & j = 1,\, i \in T_{k+1} \\ \vec{x}_i(j) & \text{otherwise} \end{cases}.$$

For this new layout, we have

$$E_{S_{k-1} \cup S_k \cup S_{k+1}}(\vec{x}') \leq E_{S_{k-1} \cup S_k \cup S_{k+1}}(\vec{x}) + \frac{|S_k|^2}{2} + \big(|S_{k-1}| + |S_{k+1}|\big)|S_k|$$

and

$$\begin{aligned} E_{S_0, T_{k+1}}(\vec{x}') &\leq E_{S_0, T_{k+1}}(\vec{x}) - |S_0||T_{k+1}| \left(\big((C+k+1)-1\big)^2 - \big((C+k)-1\big)^2 \right) \\ &= E_{S_0, T_{k+1}}(\vec{x}) - (2C + 2k - 1)\lfloor n/2 \rfloor |T_{k+1}|. \end{aligned}$$

Combining these estimates, we note that this layout has objective value bounded above by

$$\begin{aligned} E(\vec{x}') &\leq E(\vec{x}) + \big(|S_{k-1}| + |S_k|/2 + |S_{k+1}|\big)|S_k| - (2C + 2k - 1)\lfloor n/2 \rfloor |T_{k+1}| \\ &\leq E(\vec{x}) + |T_{k-1}||T_k| - (2C + 2k - 1)\lfloor n/2 \rfloor |T_{k+1}|, \end{aligned}$$

and so

$$\begin{aligned} |T_{k+1}| &\leq \frac{|T_{k-1}||T_k|}{(2C + 2k - 1)\lfloor n/2 \rfloor} \\ &\leq \frac{n^2}{\big(C + (k-1) - (2\ell - 1)\big)^{2^\ell} \big(C + k - (2\ell - 1)\big)^{2^\ell} (2C + 2k - 1)\lfloor n/2 \rfloor} \\ &\leq \frac{n}{2\lfloor n/2 \rfloor} \frac{\big(C + (k-1) - (2\ell - 1)\big)^{2^\ell}}{\big(C + k - (2\ell - 1)\big)^{2^\ell}} \frac{n}{\big(C + (k-1) - (2\ell - 1)\big)^{2^{\ell+1}}} \\ &\leq \frac{n}{\big(C + (k+1) - (2(\ell+1) - 1)\big)^{2^{\ell+1}}} \end{aligned}$$

for all $k+1 \geq 2(\ell+1)-1$ and $C+k \leq n+1$. If $C+k > n+1$, then

$$|T_{k+1}| \leq \frac{|T_{k-1}||T_k|}{(2C+2k-1)\lfloor n/2 \rfloor} \leq \frac{|T_{k-1}||T_k|}{(2n+1)\lfloor n/2 \rfloor} < 1,$$

and so $|T_{k+1}| = 0$. This completes the proof of claim (4.6), and implies that $|T_k| \leq n\, C^{-\sqrt{2}^k}$. Repeating this argument, with indices reversed, and also for the remaining dimensions $j = 2, \ldots, r$, leads to the desired result

$$\big|\{i \in [n] \,|\, \|\vec{m} - \vec{x}_i\|_\infty \geq (C+k)D\}\big| \leq 2rnC^{-\sqrt{2}^k}$$

for all $k \in \mathbb{N}$ and $C > 0$. □

Using the estimates in the proof of Lemma 14 (in particular, the bound $|T_k| \leq nC^{-\sqrt{2}^k}$), we are now prepared to prove the following diameter bound.

Lemma 15. *Let $G = ([n], E)$ have diameter D. Then, for any optimal layout $\vec{x} \in \mathbb{R}^{rn}$, i.e., $\vec{x}$ such that $E(\vec{x}) \leq E(\vec{y})$ for all $\vec{y} \in \mathbb{R}^{rn}$,*

$$\|\vec{x}_i - \vec{x}_j\|_2 \leq 8D + 4D \log_2 \log_2 2D$$

for all $i, j \in [n]$.

Proof. Let $G = ([n], E)$, have diameter D, and $\vec{x}$ be an optimal layout. Without loss of generality, suppose that the largest distance between any two points of $\vec{x}$ is realized between two points lying in span$\{e_1\}$ (i.e., on the axis of the first dimension). In addition, we order the vertices so that $\vec{x}_1(1) \leq \cdots \leq \vec{x}_n(1)$ and translate $\vec{x}$ so that $\vec{x}_{\lfloor n/2 \rfloor}(1) = 0$.

Let $x_1(1) = -\alpha e_1$, and suppose that $\alpha \geq 5D$. If this is not the case, then we are done. By differentiating E with respect to $\vec{x}_1(1)$, we obtain

$$\frac{\partial E}{\partial \vec{x}_1(1)} = -\sum_{j=2}^{n} \left(\frac{\|\vec{x}_j - \vec{x}_1\|}{d(1,j)} - 1 \right) \frac{\vec{x}_j(1) + \alpha}{d(1,j)\|\vec{x}_j - \vec{x}_1\|} = 0.$$

Let

$$T_1 := \{j \in [n] \,:\, \|\vec{x}_j - \vec{x}_1\| \geq d(1,j)\},$$

$$T_2 := \{j \in [n] \,:\, \|\vec{x}_j - \vec{x}_1\| < d(1,j)\},$$

and compare the sum of the terms in the above equation corresponding to T_1 and T_2, respectively. We begin with the former. Note that for all $j \geq \lfloor n/2 \rfloor$ we must have $j \in T_1$, because $\|x_j - x_1\| \geq \alpha \geq 5D > d(1,j)$. Therefore we have

$$\begin{aligned}
\sum_{j \in T_1} \left(\frac{\|\vec{x}_j - \vec{x}_1\|}{d(1,j)} - 1 \right) \frac{\vec{x}_j(1) + \alpha}{d(1,j)\|\vec{x}_j - \vec{x}_1\|} &\geq \sum_{j=\lfloor n/2 \rfloor}^{n} \left(\frac{\|\vec{x}_j - \vec{x}_1\|}{d(1,j)} - 1 \right) \frac{\vec{x}_j(1) + \alpha}{d(1,j)\|\vec{x}_j - \vec{x}_1\|} \\
&\geq \lceil n/2 \rceil \left(\frac{\|\vec{x}_j - \vec{x}_1\|}{D} - 1 \right) \frac{\alpha}{D\|\vec{x}_j - \vec{x}_1\|} \\
&\geq \frac{\lceil n/2 \rceil (\alpha - D)}{D^2}.
\end{aligned}$$

where in the last inequality we used $\|x_j - x_1\| \geq \alpha$ and $\|x_j - x_1\|/D - 1 \geq \alpha/D - 1$. Next, we estimate the sum of terms corresponding to T_2. Let

$$T_3 := \{i \in [n] \,:\, \vec{x}_i(1) + \alpha \leq D\}$$

and note that $T_2 \subset T_3$. We have

$$\begin{aligned}
\sum_{j \in T_2} \left(1 - \frac{\|\vec{x}_j - \vec{x}_1\|}{d(1,j)} \right) \frac{\vec{x}_j(1) + \alpha}{d(1,j)\|\vec{x}_j - \vec{x}_1\|} &= \sum_{j \in T_3} \left| 1 - \frac{\|\vec{x}_j - \vec{x}_1\|}{d(1,j)} \right|_+ \frac{\vec{x}_j(1) + \alpha}{d(1,j)\|\vec{x}_j - \vec{x}_1\|} \\
&\leq \sum_{j \in T_3} \left| 1 - \frac{\|\vec{x}_j - \vec{x}_1\|}{d(1,j)} \right|_+ \frac{1}{d(1,j)} \\
&\leq \sum_{j \in T_3} \frac{1}{d(1,j)} \;\leq\; |T_3|,
\end{aligned}$$

where $|\cdot|_+ := \max\{\cdot, 0\}$. Combining these estimates, we have

$$|T_3| - \frac{\lceil n/2 \rceil (\alpha - D)}{D^2} \geq 0,$$

or equivalently,

$$\alpha \leq D + \frac{|T_3|D^2}{\lceil n/2 \rceil}.$$

By the estimate in the proof of Lemma 14, $|T_3| \leq nC^{-\sqrt{2}^k}$ for any $k \in \mathbb{N}$ and $C > 0$ satisfying $(C+k)D \leq \alpha - D$. Taking $C = 2$, and $k = \lfloor \alpha/D - 3 \rfloor > 2D \log_2 \log_2(2D)$, we have

$$\alpha \leq D + 2D^2 2^{-\sqrt{2}^k} \leq D + \frac{2D^2}{2D} = 2D,$$

a contradiction. Therefore, α is at most $4D + 2D \log_2 \log_2 2D$. Repeating this argument with indices reversed completes the proof.

□

While the above estimate is sufficient for our purposes, we conjecture that it is not tight, and that the diameter of an optimal layout of a diameter D graph is always at most $2D$.

4.3.2 Algorithmic Lower Bounds

In this subsection, we provide a proof of Theorem 16, the algorithmic lower bound of this section. Our proof is based on a reduction from a version of Max All-Equal 3SAT. The Max All-Equal 3SAT decision problem asks whether, given variables $t_1, \ldots, t_\ell$, clauses $C_1, \ldots, C_m \subset \{t_1, \ldots, t_\ell, \bar{t}_1, \ldots, \bar{t}_\ell\}$ each consisting of at most three literals (variables or their negation), and some value L, there exists an assignment of variables such that at least L clauses have all literals equal. The Max All-Equal 3SAT decision problem is known to be APX-hard, as it does not satisfy the conditions of the Max CSP classification theorem for a polynomial time optimizable Max CSP [58] (setting all variables true or all variables false does not satisfy all clauses, and all clauses cannot be written in disjunctive normal form as two terms, one with all unnegated variables and one with all negated variables).

However, we require a much more restrictive version of this problem. In particular, we require a version in which all clauses have exactly three literals, no literal appears in a clause more than once, the number of copies of a clause is equal to the number of copies of its complement (defined as the negation of all its elements), and each literal appears in exactly

k clauses. We refer to this restricted version as Balanced Max All-Equal EU3SAT-Ek. This is indeed still APX-hard, even for $k = 6$. The proof adds little to the understanding of the hardness of the Kamada-Kawai objective, and for this reason is omitted. The proof can be found in [31].

Lemma 16. *Balanced Max All-Equal EU3SAT-E6 is APX-hard.*

Reduction. Suppose we have an instance of Balanced Max All-Equal EU3SAT-Ek with variables $t_1, \ldots, t_\ell$, and clauses $C_1, \ldots, C_{2m}$. Let $\mathcal{L} = \{t_1, \ldots, t_\ell, \bar{t}_1, \ldots, \bar{t}_\ell\}$ be the set of literals and $\mathcal{C} = \{C_1, \ldots, C_{2m}\}$ be the multiset of clauses. Consider the graph $G = (V, E)$, with

$$V = \{v^i\}_{i\in[N_v]} \cup \{t^i\}_{\substack{t\in\mathcal{L}\\ i\in[N_t]}} \cup \{C^i\}_{\substack{C\in\mathcal{C}\\ i\in[N_c]}},$$
$$E = V^{(2)} \setminus \left[\{(t^i, \bar{t}^j)\}_{\substack{t\in\mathcal{L}\\ i,j\in[N_t]}} \cup \{(t^i, C^j)\}_{\substack{t\in C, C\in\mathcal{C}\\ i\in[N_t], j\in[N_c]}}\right],$$

where $V^{(2)} := \{U \subset V \,|\, |U| = 2\}$, $N_v = N_c^{20}$, $N_t = N_c^2$, $N_c = 2000\ell^{20}m^{20}$. The graph G consists of a central clique of size N_v, a clique of size N_t for each literal, and a clique of size N_c for each clause; the central clique is connected to all other cliques (via bicliques); the clause cliques are connected to each other; each literal clique is connected to all other literals, save for the clique corresponding to its negation; and each literal clique is connected to the cliques of all clauses it does not appear in. Note that the distance between any two nodes in this graph is at most two.

Our goal is to show that, for some fixed $L = L(\ell, m, n, p) \in \mathbb{N}$, if our instance of Balanced Max All-Equal EU3SAT-Ek can satisfy $2p$ clauses, then there exists a layout $\vec{x}$ with $E(\vec{x}) \leq L$, and if it can satisfy at most $2p - 2$ clauses, then all layouts $\vec{x}'$ have objective value at least $E(\vec{x}') \geq L + 1$. To this end, we propose a layout and calculate its objective function up to lower-order terms. Later we will show that, for any layout of an instance that satisfies at most $2p - 2$ clauses, the objective value is strictly higher than our proposed construction for the previous case. The anticipated difference in objective value is of order $\Theta(N_t N_c)$, and so we will attempt to correctly place each literal vertex up to

accuracy $o\big(N_c/(\ell n)\big)$ and each clause vertex up to accuracy $o\big(N_t/(mn)\big)$.

The general idea of this reduction is that the clique $\{v^i\}$ serves as an "anchor" of sorts that forces all other vertices to be almost exactly at the correct distance from its center. Without loss of generality, assume this anchor clique is centered at 0. Roughly speaking, this forces each literal clique to roughly be at either -1 or $+1$, and the distance between negations forces negations to be on opposite sides, i.e., $\vec{x}_{t^i} \approx -\vec{x}_{\bar{t}^i}$. Clause cliques are also roughly at either -1 or $+1$, and the distance to literals forces clauses to be opposite the side with the majority of its literals, i.e., clause $C = \{t_1, t_2, t_3\}$ lies at $\vec{x}_{C^i} \approx -\chi\{\vec{x}_{t_1^i} + \vec{x}_{t_2^i} + \vec{x}_{t_3^i} \geq 0\}$, where χ is the indicator variable. The optimal embedding of G, i.e. the location of variable cliques at either $+1$ or -1, corresponds to an optimal assignment for the Max All-Equal 3SAT instance.

Remark 1 (Comparison to [19])**.** *As mentioned in the Introduction, the reduction here is significantly more involved than the hardness proof for a related problem in [19]. At a high level, the key difference is that in [19] they were able to use a large number of distance-zero vertices to create a simple structure around the origin. This is no longer possible in our setting (in particular, with bounded diameter graph metrics), which results in graph layouts with much less structure. For this reason, we require a graph that exhibits as much structure as possible. To this end, a reduction from Max All-Equal 3SAT using both literals and clauses in the graph is a much more suitable technique than a reduction from Not All-Equal 3SAT using only literals. In fact, it is not at all obvious that the same approach in [19], applied to unweighted graphs, would lead to a computationally hard instance.*

Structure of optimal layout. Let $\vec{x}$ be a globally optimal layout of G, and let us label the vertices of G based on their value in $\vec{x}$, i.e., such that $\vec{x}_1 \leq \cdots \leq \vec{x}_n$. In addition, we define

$$\hat{n} := n - N_v = 2\ell N_t + 2mN_c.$$

Without loss of generality, we assume that $\sum_i \vec{x}_i = 0$. We consider the first-order conditions for an arbitrary vertex i. Let $S := \{(i,j) \,|\, i < j,\, d(i,j) = 2\}$, $S_i := \{j \in [n] \,|\, d(i,j) = 2\}$,

and

$$S_i^{<} := \{j < i \,|\, d(i,j) = 2\}, \qquad S_i^{>} := \{j > i \,|\, d(i,j) = 2\},$$
$$S_i^{<,+} := \{j < i \,|\, d(i,j) = 2, \vec{x}_j \geq 0\}, \qquad S_i^{>,+} := \{j > i \,|\, d(i,j) = 2, \vec{x}_j \geq 0\},$$
$$S_i^{<,-} := \{j < i \,|\, d(i,j) = 2, \vec{x}_j < 0\}, \qquad S_i^{>,-} := \{j > i \,|\, d(i,j) = 2, \vec{x}_j < 0\}.$$

We have

$$\begin{aligned}
\frac{\partial E}{\partial \vec{x}_i} &= 2\sum_{j<i}(\vec{x}_i - \vec{x}_j - 1) - 2\sum_{j>i}(\vec{x}_j - \vec{x}_i - 1) - 2\sum_{j\in S_i^{<}}(\vec{x}_i - \vec{x}_j - 1) \\
&\quad + 2\sum_{j\in S_i^{>}}(\vec{x}_j - \vec{x}_i - 1) + \frac{1}{2}\sum_{j\in S_i^{<}}(\vec{x}_i - \vec{x}_j - 2) - \frac{1}{2}\sum_{j\in S_i^{>}}(\vec{x}_j - \vec{x}_i - 2) \\
&= \big(2n - \tfrac{3}{2}|S_i|\big)\vec{x}_i + 2(n+1-2i) + \frac{1}{2}\left[\sum_{j\in S_i^{<}}(3\vec{x}_j + 2) + \sum_{j\in S_i^{>}}(3\vec{x}_j - 2)\right] \\
&= \big(2n - \tfrac{3}{2}|S_i|\big)\vec{x}_i + 2(n+1-2i) + \tfrac{1}{2}\big(|S_i^{>,+}| - 5|S_i^{>,-}| + 5|S_i^{<,+}| - |S_i^{<,-}|\big) + \tfrac{3}{2}C_i \\
&= 0,
\end{aligned}$$

where

$$C_i := \sum_{j\in S_i^{<,-}\cup S_i^{>,-}}(\vec{x}_j + 1) + \sum_{j\in S_i^{<,+}\cup S_i^{>,+}}(\vec{x}_j - 1).$$

The i^{th} vertex has location

$$\vec{x}_i = \frac{2i-(n+1)}{n - \frac{3}{4}|S_i|} - \frac{|S_i^{>,+}| - 5|S_i^{>,-}| + 5|S_i^{<,+}| - |S_i^{<,-}|}{4n - 3|S_i|} - \frac{3C_i}{4n - 3|S_i|}. \tag{4.7}$$

By Lemma 15,

$$|C_i| \leq |S_i| \max_{j\in[n]} \max\{|\vec{x}_j - 1|, |\vec{x}_j + 1|\} \leq 75N_t,$$

which implies that $|\vec{x}_i| \leq 1 + 1/N_t^5$ for all $i \in [n]$.

Good layout for positive instances. Next, we will formally describe a layout that we will then show to be nearly optimal (up to lower-order terms). Before giving the formal description, we describe the layout's structure at a high level first. Given some assignment of variables that corresponds to $2p$ clauses being satisfied, for each literal t, we place the clique $\{t^i\}_{i\in[N_t]}$ at roughly $+1$ if the corresponding literal t is true; otherwise we place the clique at roughly -1. For each clause C, we place the corresponding clique $\{C^i\}_{i\in[N_C]}$ at roughly $+1$ if the majority of its literals are near -1; otherwise we place it at roughly -1. The anchor clique $\{v^i\}_{i\in[N_v]}$ lies in the middle of the interval $[-1,+1]$, separating the literal and clause cliques on both sides.

The order of the vertices, from most negative to most positive, consists of

$$T_1,\ T_2,\ T_3^0,\ \ldots,\ T_3^k,\ T_4,\ T_5^k,\ \ldots,\ T_5^0,\ T_6,\ T_7,$$

where

$$\begin{aligned}
T_1 &= \{\text{ clause cliques near } -1 \text{ with all corresponding literal cliques near } +1\ \},\\
T_2 &= \{\text{ clause cliques near } -1 \text{ with two corresponding literal cliques near } +1\},\\
T_3^\phi &= \{\text{ literal cliques near } -1 \text{ with } \phi \text{ corresponding clauses near } -1\ \}, \quad \phi = 0,\ldots,k,\\
T_4 &= \{\text{ the anchor clique }\},\\
T_5^\phi &= \{\text{ literal cliques near } +1 \text{ with } \phi \text{ corresponding clauses near } +1\ \}, \quad \phi = 0,\ldots,k,\\
T_6 &= \{\text{ clause cliques near } +1 \text{ with two corresponding literal cliques near } -1\},\\
T_7 &= \{\text{ clause cliques near } +1 \text{ with all corresponding literal cliques near } -1\ \},
\end{aligned}$$

$T_3 := \bigcup_{\phi=0}^k T_3^\phi$, $T_6 := \bigcup_{\phi=0}^k T_6^\phi$, $T_c := T_1 \cup T_2 \cup T_6 \cup T_7$, and $T_t := T_3 \cup T_5$. Let us define $\vec{y}_i := \frac{2i-(n+1)}{n}$, i.e., the optimal layout of a clique K_n in one dimension. We can write our proposed optimal layout as a perturbation of $\vec{y}_i$.

Using the above Equation 4.7 for $\vec{x}_i$, we obtain $\vec{x}_i = \vec{y}_i$ for $i \in T_4$, and, by ignoring both

the contribution of C_i and $o(1/n)$ terms, we obtain

$$\begin{aligned}
&\vec{x}_i = \vec{y}_i - \frac{3N_t}{n}, && i \in T_1, && \vec{x}_i = \vec{y}_i + \frac{3N_t}{n}, && i \in T_7,\\
&\vec{x}_i = \vec{y}_i - \frac{3N_t}{2n}, && i \in T_2, && \vec{x}_i = \vec{y}_i + \frac{3N_t}{2n}, && i \in T_6,\\
&\vec{x}_i = \vec{y}_i - \frac{N_t + (k-\phi/2)N_c}{n}, && i \in T_3, && \vec{x}_i = \vec{y}_i + \frac{N_t + (k-\phi/2)N_c}{n}, && i \in T_5.
\end{aligned}$$

Next, we upper bound the objective value of $\vec{x}$. We proceed in steps, estimating different components up to $o(N_c N_t)$. We recall the useful formulas

$$\sum_{i=1}^{r-1}\sum_{j=i+1}^{r}\left(\frac{2(j-i)}{n}-1\right)^2 = \frac{(r-1)r\big(3n^2-4n(r+1)+2r(r+1)\big)}{6n^2}, \tag{4.8}$$

and

$$\begin{aligned}
\sum_{i=1}^{r}\sum_{j=1}^{s}\left(\frac{2(j-i)+q}{n}-1\right)^2 &= \frac{rs}{3n^2}\big(3n^2+3q^2+4r^2+4s^2-6nq+6nr-6ns\\
&\qquad -6qr+6qs-6rs-2\big)\\
&\le \frac{rs^3}{3n^2} + \frac{13rs}{3n^2}\,\max\{(n-s)^2,(r-q)^2,r^2\}.
\end{aligned} \tag{4.9}$$

For $i \in T_c$, $|1-|\vec{x}_i|| \le 4N_t/n$, and so

$$E_{T_c}(\vec{x}) \le |T_c|^2 \max_{i,j\in T_c}(|\vec{x}_i - \vec{x}_j| - 1)^2 \le 8m^2N_c^2.$$

In addition, for $i \in T_t$, $|1-|\vec{x}_i|| \le 3\ell N_t/n$, and so

$$\begin{aligned}
E_{T_t}(\vec{x}) &\le |\{i,j\in T_t \,|\, d(i,j)=1\}|\left(1+\frac{6\ell N_t}{n}\right)^2 + |\{i,j\in T_t \,|\, d(i,j)=2\}|\,\frac{1}{4}\left(\frac{6\ell N_t}{n}\right)^2\\
&\le (2\ell-1)\ell N_t^2 + 40\frac{\ell^3 N_t^3}{n}
\end{aligned}$$

and

$$\begin{aligned}
E_{T_c,T_t}(\vec{x}) &\le \left(1+\frac{6\ell N_t}{N}\right)^2 |\{i\in T_c, j\in T_t \,|\, d(i,j)=1\}| \\
&\quad + 2\times\frac{1}{4}\left(2+\frac{6\ell N_t}{n}\right)^2 |\{i\in T_2, j\in T_3 \,|\, d(i,j)=2\}| \\
&\quad + 2\times\frac{1}{4}\left(\frac{6\ell N_t}{n}\right)^2 |\{i\in T_1\cup T_2, j\in T_5 \,|\, d(i,j)=2\}| \\
&\le \left(1+\frac{18\ell N_t}{n}\right)(2\ell-3)N_t\, 2mN_c \\
&\quad + \frac{1}{2}\left(4+\frac{30\ell N_t}{n}\right)(m-p)N_cN_t \\
&\quad + \frac{1}{2}\frac{6\ell N_t}{n}(2m+p)N_cN_t \\
&\le (4\ell-6)mN_tN_c + 2(m-p)N_tN_c + 100\frac{m\ell^2N_t^2N_c}{n}.
\end{aligned}$$

The quantity $E_{T_4}(\vec{x})$ is given by (4.8) with $r=N_v$, and so

$$\begin{aligned}
E_{T_4}(\vec{x}) &= \frac{N_v(N_v-1)(3n^2-4n(N_v+1)+2N_v(N_v+1))}{6n^2} \\
&\le \frac{(n-1)(n-2)-2\hat{n}n+3\hat{n}^2+3\hat{n}}{6}.
\end{aligned}$$

The quantity $E_{T_1,T_4}(\vec{x})$ is given by (4.9) with

$$q = 3N_t+\hat{n}/2, \qquad r=pN_c, \qquad \text{and} \qquad s=N_v,$$

and so

$$E_{T_1,T_4}(\vec{x}) \le \frac{pN_cN_v^3}{3n^2}+\frac{13pN_cN_v\hat{n}^2}{3n^2} \le \frac{pN_c}{3}(n-3\hat{n})+\frac{16pN_c\hat{n}^2}{3n}.$$

Similarly, the quantity $E_{T_2,T_4}(\vec{x})$ is given by (4.9) with

$$q = \tfrac{3}{2}N_t+\hat{n}/2-pN_c, \qquad r=(m-p)N_c, \qquad \text{and} \qquad s=N_v,$$

and so

$$\begin{aligned} E_{T_1,T_4}(\vec{x}) &\leq \frac{(m-p)N_cN_v^3}{3n^2} + \frac{13(m-p)N_cN_v\hat{n}^2}{3n^2} \\ &\leq \frac{(m-p)N_c}{3}(n-3\hat{n}) + \frac{16(m-p)N_c\hat{n}^2}{3n}. \end{aligned}$$

Next, we estimate $E_{T_3^\phi,T_4}(\vec{x})$. Using formula (4.9) with

$$q = N_t + (k-\phi/2)N_c + \sum_{j=\phi}^{k} \ell_j, \qquad r = \ell_\phi, \qquad \text{and} \qquad s = N_v,$$

where $\ell_\phi := |T_3^\phi|$, we have

$$E_{T_3^\phi,T_4} \leq \frac{\ell_\phi N_v^3}{3n^2} + \frac{13\ell_\phi N_v\hat{n}^2}{3n^2} \leq \frac{\ell_\phi}{3}(n-3\hat{n}) + \frac{16\ell_\phi\hat{n}^2}{3n}.$$

Combining the individual estimates for each T_3^ϕ gives us

$$E_{T_3,T_4} \leq \frac{\ell N_t}{3}(n-3\hat{n}) + \frac{16\ell N_t\hat{n}^2}{3n}.$$

Finally, we can combine all of our above estimates to obtain an upper bound on $E(\vec{x})$. We have

$$\begin{aligned} E(\vec{x}) &\leq (2\ell-1)\ell N_t^2 + (4\ell-6)mN_tN_c + 2(m-p)N_tN_c + \frac{(n-1)(n-2)}{6} - \tfrac{1}{3}n\hat{n} \\ &\qquad + \tfrac{1}{2}\hat{n}^2 + \tfrac{2}{3}mN_c(n-3\hat{n}) + \tfrac{2}{3}\ell N_t(n-3\hat{n}) + 200m^2N_c^2 \\ &= \frac{(n-1)(n-2)}{6} - \tfrac{1}{2}\hat{n}^2 + (2\ell-1)\ell N_t^2 \\ &\qquad + \big[(4\ell-6)m + 2(m-p)\big]N_tN_c + 200m^2N_c^2 \\ &\leq \frac{(n-1)(n-2)}{6} - \ell N_t^2 - 2(2m+p)N_tN_c + 200m^2N_c^2. \end{aligned}$$

We define the ceiling of this final upper bound to be the quantity L. The remainder of the proof consists of showing that if our given instance satisfies at most $2p-2$ clauses, then any layout has objective value at least $L+1$.

Suppose, to the contrary, that there exists some layout $\vec{x}'$ (shifted so that $\sum_i \vec{x}'_i = 0$), with $E(\vec{x}') < L + 1$.

From the analysis above, $|\vec{x}'_i| \leq 1 + 1/N_t^5$ for all i. Intuitively, an optimal layout should have a large fraction of the vertices at distance two on opposite sides. To make this intuition precise, we first note that

Lemma 17. *Let $\vec{x} \in \mathbb{R}^n$ be a layout of the clique K_n. Then $E(\vec{x}) \geq (n-1)(n-2)/6$.*

Proof. The first order conditions (4.7) imply that the optimal layout (up to translation and vertex reordering) is given by $\vec{x}'_i = \big(2i-(n+1)\big)/n$. By (4.8), $E(\vec{x}') = (n-1)(n-2)/6$. □

Using Lemma 17, we can lower bound $E(\vec{x}')$ by

$$\begin{aligned} E(\vec{x}') &= \sum_{i<j} \big(\vec{x}'_j - \vec{x}'_i - 1\big)^2 - \sum_{(i,j)\in S} \big(\vec{x}'_j - \vec{x}'_i - 1\big)^2 + \frac{1}{4} \sum_{(i,j)\in S} \big(\vec{x}'_j - \vec{x}'_i - 2\big)^2 \\ &\geq \frac{(n-1)(n-2)}{6} - \sum_{(i,j)\in S} \big[\tfrac{3}{4}(\vec{x}'_j - \vec{x}'_i)^2 - (\vec{x}'_j - \vec{x}'_i)\big]. \end{aligned}$$

Therefore, by assumption,

$$\begin{aligned} \sum_{(i,j)\in S} \big[\tfrac{3}{4}(\vec{x}'_j - \vec{x}'_i)^2 - (\vec{x}'_j - \vec{x}'_i)\big] &\geq \frac{(n-1)(n-2)}{6} - (L+1) \\ &\geq \ell N_t^2 + 2(2m+p)N_t N_c - 200m^2 N_c^2 - 2. \end{aligned}$$

We note that the function $\frac{3}{4}x^2 - x$ equals one at $x = 2$ and is negative for $x \in \big[0, \frac{4}{3}\big)$. Because $|\vec{x}'_i| \leq 1 + 1/N_t^5$ for all i,

$$\max_{(i,j)\in S} \big[\tfrac{3}{4}(\vec{x}'_j - \vec{x}'_i)^2 - (\vec{x}'_j - \vec{x}'_i)\big] \leq \frac{3}{4}\left(2 + \frac{2}{N_t^5}\right)^2 - \left(2 + \frac{2}{N_t^5}\right) \leq 1 + \frac{7}{N_t^5}.$$

Let

$$T' := \{(i,j) \in S \,|\, \vec{x}'_i \leq -\tfrac{1}{6} \text{ and } \vec{x}'_j \geq \tfrac{1}{6}\}.$$

By assumption, $|T'|$ is at most $\ell N_t^2 + 2(2m+p-1)N_t N_c$, otherwise the corresponding instance could satisfy at least $2p$ clauses, a contradiction.

However, the quantity $\left[\frac{3}{4}(\vec{x}'_j - \vec{x}'_i)^2 - (\vec{x}'_j - \vec{x}'_i)\right]$ is negative for all $(i,j) \in S \backslash T'$. Therefore,

$$\left(1 + \frac{7}{N_t^5}\right)|T'| \geq \ell N_t^2 + 2(2m+p)N_t N_c - 200m^2 N_c^2 - 2,$$

which implies that

$$\begin{aligned}
|T'| &\geq \left(1 - \frac{7}{N_t^5}\right)\left(\ell N_t^2 + 2(2m+p)N_t N_c - 200m^2 N_c^2 - 2\right) \\
&\geq \ell N_t^2 + 2(2m+p)N_t N_c - 200m^2 N_c^2 - 2000 \\
&> \ell N_t^2 + 2(2m+p-1)N_t N_c + N_t N_c,
\end{aligned}$$

a contradiction. This completes the proof, with a gap of one.

4.3.3 An Approximation Algorithm

In this subsection, we formally describe an approximation algorithm using tools from the Dense CSP literature, and prove theoretical guarantees for the algorithm.

Preliminaries: Greedy Algorithms for Max-CSP

A long line of work studies the feasibility of solving the Max-CSP problem under various related pseudorandomness and density assumptions. In our case, an algorithm with mild dependence on the alphabet size is extremely important. A very simple greedy approach, proposed and analyzed by Mathieu and Schudy [81, 101] (see also [133]), satisfies this requirement.

Theorem 18 ([81, 101])**.** *Suppose that Σ is a finite alphabet, $n \geq 1$ is a positive integer, and for every $i,j \in \binom{n}{2}$ we have a function $f_{ij} : \Sigma \times \Sigma \to [-M, M]$. Then for any $\epsilon > 0$, Algorithm* GREEDYCSP *with $t_0 = O(1/\epsilon^2)$ runs in time $n^2|\Sigma|^{O(1/\epsilon^2)}$ and returns*

Algorithm 4 Greedy Algorithm for Dense CSPs [81, 101]

function GreedyCSP($\Sigma, n, t_0, \{f_{ij}\}$)
  Shuffle the order of variables $x_1, \ldots, x_n$ by a random permutation.
  for all assignments $x_1, \ldots, x_{t_0} \in \Sigma^{t_0}$ **do**
    for $(t_0 + 1) \leq i \leq n$ **do**
      Choose $x_i \in \Sigma$ to maximize

$$\sum_{j<i} f_{ji}(x_j, x_i)$$

    end for
    Record x and objective value $\sum_{i \neq j} f_{ij}(x_i, x_j)$.
  end for
  Return the assignment x found with maximum objective value.
end function

$x_1, \ldots, x_n \in \Sigma$ *such that*

$$\mathbb{E} \sum_{i \neq j} f_{ij}(x_i, x_j) \geq \sum_{i \neq j} f_{ij}(x_i^*, x_j^*) - \epsilon M n^2$$

for any $x_1^*, \ldots, x_n^* \in \Sigma$*, where* $\mathbb{E}$ *denotes the expectation over the randomness of the algorithm.*

In comparison, we note that computing the maximizer using brute force would run in time $|\Sigma|^n$, i.e. exponentially slower in terms of n. This guarantee is stated in expectation but, if desired, can be converted to a high probability guarantee by using Markov's inequality and repeating the algorithm multiple times. We use GREEDYCSP to solve a minimization problem instead of maximization, which corresponds to negating all of the functions f_{ij}. To do so, we require estimates on the error due to discretization of our domain.

Lemma 18. *For* $c, R > 0$ *and* $y \in \mathbb{R}^r$ *with* $\|y\| \leq R$*, the function* $x \mapsto (\|x - y\|/c - 1)^2$ *is* $\frac{2}{c} \max(1, 2R/c)$*-Lipschitz on* $B_R = \{x : \|x\| \leq R\}$.

Proof. We first prove that the function $x \mapsto (x/c - 1)^2$ is $\frac{2}{c} \max(1, R/c)$-Lipschitz on the interval $[0, R]$. Because the derivative of the function is $\frac{2}{c}(x/c - 1)$ and $\left|\frac{2}{c}(x/c - 1)\right| \leq \frac{2}{c} \max(1, R/c)$ on $[0, R]$, this result follows from the mean value theorem.

Algorithm 5 Approximation Algorithm KKSCHEME

function KKSCHEME($\epsilon_1, \epsilon_2, R$):
Build an ϵ_1-net S_{ϵ_1} of $B_R = \{x : \|x\| \leq R\} \subset \mathbb{R}^r$ as in Lemma 12.
Apply the GREEDYCSP algorithm of Theorem 18 with $\epsilon = \epsilon_2$ to approximately minimize $E(\vec{x}_1, \ldots, \vec{x}_n)$ over $\vec{x}_1, \ldots, \vec{x}_n \in S_{\epsilon_1}^n$.
Return $\vec{x}_1, \ldots, \vec{x}_n$.
end function

Because the function $\|x - y\|$ is 1-Lipschitz and $\|x - y\| \leq \|x\| + \|y\| \leq 2R$ by the triangle inequality, the result follows from the fact that $x \mapsto (x/c - 1)^2$ is $\frac{2}{c}\max(1, R/c)$-Lipschitz and a composition of Lipschitz functions is Lipschitz. □

Combining this result with Lemma 12 we can bound the loss in objective value due to restricting to a well-chosen discretization.

Lemma 19. *Let $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^r$ be arbitrary vectors such that $\|\vec{x}_i\| \leq R$ for all i and $\epsilon > 0$ be arbitrary. Define S_ϵ to be an ϵ-net of B_R as in Lemma 12, so $|S_\epsilon| \leq (3R/\epsilon)^r$. For any input metric over $[n]$ with $\min_{i,j\in[n]} d(i, j) = 1$, there exists $\vec{x}'_1, \ldots, \vec{x}'_n \in S_\epsilon$ such that*

$$E(\vec{x}'_1, \ldots, \vec{x}'_n) \leq E(\vec{x}_1, \ldots, \vec{x}_n) + 4\epsilon R n^2$$

where E is (4.5) *defined with respect to an arbitrary graph with n vertices.*

Proof. By Lemma 18 the energy $E(\vec{x}_1, \ldots, \vec{x}_n)$ is the sum of $\binom{n}{2} \leq n^2/2$ many terms, which, for each i and j, are individually $4R$-Lipschitz in $\vec{x}_i$ and $\vec{x}_j$. Therefore, defining $\vec{x}'_i$ to be the closest point in S_ϵ for all i gives the desired result. □

This result motivates a straightforward algorithm for producing nearly-optimal layouts of a graph. Given some graph, by Lemma 15, a sufficiently large radius can be chosen so that there exists an optimal layout in a ball of that radius. By constructing an ϵ-net of that ball, and applying the GREEDYCSP algorithm to the resulting discretization, we obtain a nearly-optimal layout with theoretical guarantees. This technique is formally described in the algorithm KKSCHEME. We are now prepared to prove Theorem 17.

By combining Lemma 19 with Theorem 18 (used as a minimization instead of maximization algorithm), the output $\vec{x}_1, \ldots, \vec{x}_n$ of KKSCHEME satisfies

$$E(\vec{x}_1, \ldots, \vec{x}_n) \leq E(\vec{x}_1^*, \ldots, \vec{x}_n^*) + 4\epsilon_1 R n^2 + \epsilon_2 R^2 n^2$$

and runs in time $n^2 2^{O(1/\epsilon_2^2) r \log(3R/\epsilon_1)}$. Taking $\epsilon_2 = O(\epsilon/R^2)$ and $\epsilon_1 = O(\epsilon/R)$ completes the proof of Theorem 17.

The runtime can be improved to $n^2 + (R/\epsilon)^{O(dR^4/\epsilon^2)}$ using a slightly more complex greedy CSP algorithm [81]. Also, by the usual argument, a high probability guarantee can be derived by repeating the algorithm $O(\log(2/\delta))$ times, where $\delta > 0$ is the desired failure probability.

Chapter 5

Error Estimates for the Lanczos Method

5.1 Introduction

The computation of extremal eigenvalues of matrices is one of the most fundamental problems in numerical linear algebra. When a matrix is large and sparse, methods such as the Jacobi eigenvalue algorithm and QR algorithm become computationally infeasible, and, therefore, techniques that take advantage of the sparsity of the matrix are required. Krylov subspace methods are a powerful class of techniques for approximating extremal eigenvalues, most notably the Arnoldi iteration for non-symmetric matrices and the Lanczos method for symmetric matrices.

The Lanczos method is a technique that, given a symmetric matrix $A \in \mathbb{R}^{n\times n}$ and an initial vector $b \in \mathbb{R}^n$, iteratively computes a tridiagonal matrix $T_m \in \mathbb{R}^{m\times m}$ that satisfies $T_m = Q_m^T A Q_m$, where $Q_m \in \mathbb{R}^{n\times m}$ is an orthonormal basis for the Krylov subspace

$$\mathcal{K}_m(A, b) = \text{span}\{b, Ab, ..., A^{m-1}b\}.$$

The eigenvalues of T_m, denoted by $\lambda_1^{(m)}(A, b) \geq ... \geq \lambda_m^{(m)}(A, b)$, are the Rayleigh-Ritz approximations to the eigenvalues $\lambda_1(A) \geq ... \geq \lambda_n(A)$ of A on $\mathcal{K}_m(A, b)$, and, therefore,

are given by

$$\lambda_i^{(m)}(A,b) \quad = \quad \min_{\substack{U \subset \mathcal{K}_m(A,b) \\ \dim(U)=m+1-i}} \max_{\substack{x \in U \\ x \neq 0}} \frac{x^T A x}{x^T x}, \qquad i = 1, ..., m, \tag{5.1}$$

or, equivalently,

$$\lambda_i^{(m)}(A,b) \quad = \quad \max_{\substack{U \subset \mathcal{K}_m(A,b) \\ \dim(U)=i}} \min_{\substack{x \in U \\ x \neq 0}} \frac{x^T A x}{x^T x}, \qquad i = 1, ..., m. \tag{5.2}$$

This description of the Lanczos method is sufficient for a theoretical analysis of error (i.e., without round-off error), but, for completeness, we provide a short description of the Lanczos method (when $\mathcal{K}_m(A,b)$ is full-rank) in Algorithm 6 [113]. For a more detailed discussion of the nuances of practical implementation and techniques to minimize the effects of round-off error, we refer the reader to [47, Section 10.3]. If A has νn non-zero entries, then Algorithm 6 outputs T_m after approximately $(2\nu + 8)mn$ floating point operations. A number of different techniques, such as the divide-and-conquer algorithm with the fast multipole method [23], can easily compute the eigenvalues of a tridiagonal matrix. The complexity of this computation is negligible compared to the Lanczos algorithm, as, in practice, m is typically orders of magnitude less than n.

Equation (5.1) for the Ritz values $\lambda_i^{(m)}$ illustrates the significant improvement that the Lanczos method provides over the power method for extremal eigenvalues. Whereas the power method uses only the iterate $A^m b$ as an approximation of an eigenvector associated with the largest magnitude eigenvalue, the Lanczos method uses the span of all of the iterates of the power method (given by $\mathcal{K}_{m+1}(A,b)$). However, the analysis of the Lanczos method is significantly more complicated than that of the power method. Error estimates for extremal eigenvalue approximation using the Lanczos method have been well studied, most notably by Kaniel [56], Paige [91], Saad [99], and Kuczynski and Wozniakowski [67] (other notable work includes [30, 36, 68, 84, 92, 104, 105, 125]). The work of Kaniel, Paige, and Saad focused on the convergence of the Lanczos method as m increases, and, therefore, their results have strong dependence on the spectrum of the matrix A and the choice of initial

vector b. For example, a standard result of this type is the estimate

$$\frac{\lambda_1(A) - \lambda_1^{(m)}(A,b)}{\lambda_1(A) - \lambda_n(A)} \leq \left(\frac{\tan \angle(b, \varphi_1)}{T_{m-1}(1+2\gamma)} \right)^2, \qquad \gamma = \frac{\lambda_1(A) - \lambda_2(A)}{\lambda_2(A) - \lambda_n(A)}, \tag{5.3}$$

where φ_1 is the eigenvector corresponding to λ_1, and T_{m-1} is the $(m-1)^{th}$ degree Chebyshev polynomial of the first kind [100, Theorem 6.4]. Kuczynski and Wozniakowski took a quite different approach, and estimated the maximum expected relative error $(\lambda_1 - \lambda_1^{(m)})/\lambda_1$ over all $n \times n$ symmetric positive definite matrices, resulting in error estimates that depend only on the dimension n and the number of iterations m. They produced the estimate

$$\sup_{A \in \mathcal{S}_{++}^n} \mathbb{E}_{b \sim \mathcal{U}(S^{n-1})} \left[\frac{\lambda_1(A) - \lambda_1^{(m)}(A,b)}{\lambda_1(A)} \right] \leq .103 \, \frac{\ln^2(n(m-1)^4)}{(m-1)^2}, \tag{5.4}$$

for all $n \geq 8$ and $m \geq 4$, where $\mathcal{S}_{++}^n$ is the set of all $n \times n$ symmetric positive definite matrices, and the expectation is with respect to the uniform probability measure on the hypersphere $S^{n-1} = \{x \in \mathbb{R}^n \,|\, \|x\| = 1\}$ [67, Theorem 3.2]. One can quickly verify that equation (5.4) also holds when the $\lambda_1(A)$ term in the denominator is replaced by $\lambda_1(A) - \lambda_n(A)$, and the supremum over $\mathcal{S}_{++}^n$ is replaced by the maximum over the set of all $n \times n$ symmetric matrices, denoted by $\mathcal{S}^n$.

Both of these approaches have benefits and drawbacks. If an extremal eigenvalue is known to have a reasonably large eigenvalue gap (based on application or construction), then a distribution dependent estimate provides a very good approximation of error, even for small m. However, if the eigenvalue gap is not especially large, then distribution dependent estimates can significantly overestimate error, and estimates that depend only on n and m are preferable. This is illustrated by the following elementary, yet enlightening, example.

Example 1. *Let $A \in S_{++}^n$ be the tridiagonal matrix resulting from the discretization of the Laplacian operator on an interval with Dirichlet boundary conditions, namely, $A_{i,i} = 2$ for $i = 1, ..., n$ and $A_{i,i+1} = A_{i+1,i} = -1$ for $i = 1, ..., n-1$. The eigenvalues of A are given by $\lambda_i(A) = 2 + 2\cos(i\pi/(n+1))$, $i = 1, ..., n$. Consider the approximation of $\lambda_1(A)$ by m iterations of the Lanczos method. For a random choice of b, the expected*

Algorithm 6 Lanczos Method

Input: symmetric matrix $A \in \mathbb{R}^{n\times n}$, vector $b \in \mathbb{R}^n$, number of iterations m.

Output: symmetric tridiagonal matrix $T_m \in \mathbb{R}^{m\times m}$, $T_m(i,i) = \alpha_i$,

$T_m(i,i+1) = \beta_i$, satisfying $T_m = Q_m^T A Q_m$, where $Q_m = [q_1 \ ... \ q_m]$.

Set $\beta_0 = 0$, $q_0 = 0$, $q_1 = b/\|b\|$

For $i = 1, ..., m$

$v = Aq_i$

$\alpha_i = q_i^T v$

$v = v - \alpha_i q_i - \beta_{i-1} q_{i-1}$

$\beta_i = \|v\|$

$q_{i+1} = v/\beta_i$

---

value of $\tan^2 \angle(b, \varphi_1)$ *is* $(1+o(1))n$. *If* $\tan^2 \angle(b, \varphi_1) = Cn$ *for some constant* C, *then (5.3) produces the estimate*

$$\frac{\lambda_1(A) - \lambda_1^{(m)}(A,b)}{\lambda_1(A) - \lambda_n(A)} \leq Cn\, T_{m-1}\left(1 + 2\tan\left(\frac{\pi}{2(n+1)}\right)\tan\left(\frac{3\pi}{2(n+1)}\right)\right)^{-2}$$
$$\simeq n(1 + O(n^{-1}))^{-m} \simeq n.$$

In this instance, the estimate is a trivial one for all choices of m, *which varies greatly from the error estimate (5.4) of order* $\ln^2 n/m^2$. *The exact same estimate holds for the smallest eigenvalue* $\lambda_n(A)$ *when the corresponding bounds are applied, since* $4I - A$ *is similar to* A.

Now, consider the approximation of the largest eigenvalue of $B = A^{-1}$ *by* m *iterations of the Lanczos method. The matrix* B *possesses a large gap between the largest and second-largest eigenvalue, which results in a value of* γ *for (5.3) that remains bounded below by a constant independent of* n, *namely*

$$\gamma = \left(2\cos\left(\pi/(n+1)\right) + 1\right) / \left(2\cos\left(\pi/(n+1)\right) - 1\right).$$

Therefore, in this instance, the estimate (5.3) illustrates a constant convergence rate, produces non-trivial bounds (for typical b) for $m = \Theta(\ln n)$, and is preferable to the error estimate (5.4) of order $\ln^2 n/m^2$.

More generally, if a matrix A has eigenvalue gap $\gamma \lesssim n^{-\alpha}$ and the initial vector b satisfies $\tan^2 \angle(b, \varphi_1) \gtrsim n$, then the error estimate (5.3) is a trivial one for $m \lesssim n^{\alpha/2}$. This implies that the estimate (5.3) is most useful when the eigenvalue gap is constant or tends to zero very slowly as n increases. When the gap is not especially large (say, $n^{-\alpha}$, α constant), then uniform error estimates are preferable for small values of m. In this work, we focus on uniform bounds, namely, error estimates that hold uniformly over some large set of matrices (typically $\mathcal{S}^n_{++}$ or $\mathcal{S}^n$). We begin by recalling some of the key existing uniform error estimates for the Lanczos method.

5.1.1 Related Work

Uniform error estimates for the Lanczos method have been produced almost exclusively for symmetric positive definite matrices, as error estimates for extremal eigenvalues of symmetric matrices can be produced from estimates for S^n_{++} relatively easily. The majority of results apply only to either $\lambda_1(A)$, $\lambda_n(A)$, or some function of the two (i.e., condition number). All estimates are probabilistic and take the initial vector b to be uniformly distributed on the hypersphere. Here we provide a brief description of some key uniform error estimates previously produced for the Lanczos method.

In [67], Kuczynski and Wozniakowski produced a complete analysis of the power method and provided a number of upper bounds for the Lanczos method. Most notably, they produced error estimate (5.4) and provided the following upper bound for the probability that the relative error $(\lambda_1 - \lambda_1^{(m)})/\lambda_1$ is greater than some value ϵ:

$$\sup_{A \in \mathcal{S}^n_{++}} \mathbb{P}_{b \sim \mathcal{U}(S^{n-1})} \left[\frac{\lambda_1(A) - \lambda_1^{(m)}(A, b)}{\lambda_1(A)} > \epsilon \right] \leq 1.648\sqrt{n} e^{-\sqrt{\epsilon}(2m-1)}. \tag{5.5}$$

However, the authors were unable to produce any lower bounds for (5.4) or (5.5), and stated that a more sophisticated analysis would most likely be required. In the same paper, they

performed numerical experiments for the Lanczos method that produced errors of the order m^{-2}, which led the authors to suggest that the error estimate (5.4) may be an overestimate, and that the $\ln^2 n$ term may be unnecessary.

In [68], Kuczynski and Wozniakowski noted that the above estimates immediately translate to relative error estimates for minimal eigenvalues when the error $\lambda_m^{(m)} - \lambda_n$ is considered relative to λ_1 or $\lambda_1 - \lambda_n$ (both normalizations can be shown to produce the same bound). However, we can quickly verify that there exist sequences in $\mathcal{S}_{++}^n$ for which the quantity $\mathbb{E}\big[(\lambda_m^{(m)} - \lambda_n)/\lambda_n\big]$ is unbounded. These results for minimal eigenvalues, combined with (5.4), led to error bounds for estimating the condition number of a matrix. Unfortunately, error estimates for the condition number face the same issue as the quantity $(\lambda_m^{(m)} - \lambda_n)/\lambda_n$, and therefore, only estimates that depend on the value of the condition number can be produced.

The proof technique used to produce (5.4) works specifically for the quantity $\mathbb{E}\big[(\lambda_1 - \lambda_1^{(m)})/\lambda_1\big]$ (i.e., the 1-norm), and does not carry over to more general p-norms of the form $\mathbb{E}\big[((\lambda_1 - \lambda_1^{(m)})/\lambda_1)^p\big]^{1/p}$, $p \in (1, \infty)$. Later, in [30, Theorem 5.2, $r = 1$], Del Corso and Manzini produced an upper bound of the order $(\sqrt{n}/m)^{1/p}$ for arbitrary p-norms, given by

$$\sup_{A \in \mathcal{S}_{++}^n} \mathbb{E}_{b \sim \mathcal{U}(S^{n-1})} \left[\left(\frac{\lambda_1(A) - \lambda_1^{(m)}(A, b)}{\lambda_1(A)} \right)^p \right]^{\frac{1}{p}} \lesssim \frac{1}{m^{1/p}} \left(\frac{\Gamma\left(p - \frac{1}{2}\right) \Gamma\left(\frac{n}{2}\right)}{\Gamma(p) \Gamma\left(\frac{n-1}{2}\right)} \right)^{\frac{1}{p}}. \tag{5.6}$$

This bound is clearly worse than (5.4), and better bounds can be produced for arbitrary p simply by making use of (5.5). Again, the authors were unable to produce any lower bounds.

More recently, the machine learning and optimization community has become increasingly interested in the problem of approximating the top singular vectors of a symmetric matrix by randomized techniques (for a review, see [49]). This has led to a wealth of new results in recent years regarding classical techniques, including the Lanczos method. In [84], Musco and Musco considered a block Krylov subspace algorithm similar to the block Lanczos method and showed that with high probability their algorithm achieves an error of ϵ in $m = O(\epsilon^{-1/2} \ln n)$ iterations, matching the bound (5.5) for a block size of one. Very recently, a number of lower bounds for a wide class of randomized algorithms were shown

in [104, 105], all of which can be applied to the Lanczos method as a corollary. First, the authors showed that the dependence on n in the above upper bounds was in fact necessary. In particular, it follows from [105, Theorem A.1] that $O(\log n)$ iterations are necessary to obtain a non-trivial error. In addition, as a corollary of their main theorem [105, Theorem 1], there exist $c_1, c_2 > 0$ such that for all $\epsilon \in (0, 1)$ there exists an $n_o = \text{poly}(\epsilon^{-1})$ such that for all $n \geq n_o$ there exists a random matrix $A \in \mathcal{S}^n$ such that

$$\mathbb{P}\left[\frac{\rho(A) - \rho(T)}{\rho(A)} \geq \frac{\epsilon}{12}\right] \geq 1 - e^{-n^{c_2}} \qquad \text{for all } m \leq c_1 \epsilon^{-1/2} \ln n, \tag{5.7}$$

where $\rho(\cdot)$ is the spectral radius of a matrix, and the randomness is over both the initial vector and matrix.

5.1.2 Contributions and Remainder of Chapter

In what follows, we prove improved upper bounds for the maximum expected error of the Lanczos method in the p-norm, $p \geq 1$, and combine this with nearly-matching asymptotic lower bounds with explicit constants. These estimates can be found in Theorem 19. The upper bounds result from using a slightly different proof technique than that of (5.4), which is more robust to arbitrary p-norms. Comparing the lower bounds of Theorem 19 to the estimate (5.7), we make a number of observations. Whereas (5.7) results from a statistical analysis of random matrices, our estimates follow from taking an infinite sequence of non-random matrices with explicit eigenvalues and using the theory of orthogonal polynomials. Our estimate for $m = O(\ln n)$ (in Theorem 19) is slightly worse than (5.7) by a factor of $\ln^2 \ln n$, but makes up for this in the form of an explicit constant. The estimate for $m = o(n^{1/2} \ln^{-1/2} n)$ (also in Theorem 19) does not have n dependence, but illustrates a useful lower bound, as it has an explicit constant and the $\ln n$ term becomes negligible as m increases. The results (5.7) do not fully apply to this regime, as the dependence of the required lower bound on n is at least cubic in the inverse of the eigenvalue gap (see [105, Theorem 6.1] for details).

To complement these bounds, we also provide an error analysis for matrices that have a

certain structure. In Theorem 20, we produce improved dimension-free upper bounds for matrices that have some level of eigenvalue "regularity" near λ_1. In addition, in Theorem 21, we prove a powerful result that can be used to determine, for any fixed m, the asymptotic relative error for any sequence of matrices $X_n \in \mathcal{S}^n$, $n = 1, 2, ...$, that exhibits suitable convergence of its empirical spectral distribution. Later, we perform numerical experiments that illustrate the practical usefulness of this result. Theorem 21, combined with estimates for Jacobi polynomials (see Proposition 13), illustrates that the inverse quadratic dependence on the number of iterations m in the estimates produced throughout this chapter does not simply illustrate the worst case, but is actually indicative of the typical case in some sense.

In Corollary 3 and Theorem 22, we produce results similar to Theorem 19 for arbitrary eigenvalues λ_i. The lower bounds follow relatively quickly from the estimates for λ_1, but the upper bounds require some mild assumptions on the eigenvalue gaps of the matrix. These results mark the first uniform-type bounds for estimating arbitrary eigenvalues by the Lanczos method. In addition, in Corollary 4, we translate our error estimates for the extremal eigenvalues $\lambda_1(A)$ and $\lambda_n(A)$ into error bounds for the condition number of a symmetric positive definite matrix, but, as previously mentioned, the relative error of the condition number of a matrix requires estimates that depend on the condition number itself. Finally, we present numerical experiments that support the accuracy and practical usefulness of the theoretical estimates detailed above.

The remainder of the chapter is as follows. In Section 5.2, we prove basic results regarding relative error and make a number of fairly straightforward observations. In Section 5.3, we prove asymptotic lower bounds and improved upper bounds for the relative error in an arbitrary p-norm. In Section 5.4, we produce a dimension-free error estimate for a large class of matrices and prove a theorem that can be used to determine the asymptotic relative error for any fixed m and sequence of matrices $X_n \in \mathcal{S}^n$, $n = 1, 2, ...$, with suitable convergence of its empirical spectral distribution. In Section 5.5, under some mild additional assumptions, we prove a version of Theorem 19 for arbitrary eigenvalues λ_i, and extend our results for λ_1 and λ_n to the condition number of a symmetric positive definite matrix. Finally, in Section 5.6, we perform a number of experiments and discuss how the numerical

results compare to the theoretical estimates in this work.

5.2 Preliminary Results

Because the Lanczos method applies only to symmetric matrices, all matrices in this chapter are assumed to belong to $\mathcal{S}^n$. The Lanczos method and the quantity $(\lambda_1(A) - \lambda_1^{(m)}(A,b))/(\lambda_1(A) - \lambda_n(A))$ are both unaffected by shifting and scaling, and so any maximum over $A \in \mathcal{S}^n$ can be replaced by a maximum over all $A \in \mathcal{S}^n$ with $\lambda_1(A)$ and $\lambda_n(A)$ fixed. Often for producing upper bounds, it is convenient to choose $\lambda_1(A) = 1$ and $\lambda_n(A) = 0$. For the sake of brevity, we will often write $\lambda_i(A)$ and $\lambda_j^{(m)}(A,b)$ as λ_i and $\lambda_j^{(m)}$ when the associated matrix A and vector b are clear from context.

We begin by rewriting expression (5.2) for $\lambda_1^{(m)}$ in terms of polynomials. The Krylov subspace $\mathcal{K}_m(A,b)$ can be alternatively defined as

$$\mathcal{K}_m(A,b) = \{P(A)b \,|\, P \in \mathcal{P}_{m-1}\},$$

where $\mathcal{P}_{m-1}$ is the set of all real-valued polynomials of degree at most $m-1$. Suppose $A \in \mathcal{S}^n$ has eigendecomposition $A = Q\Lambda Q^T$, where $Q \in \mathbb{R}^{n\times n}$ is an orthogonal matrix and $\Lambda \in \mathbb{R}^{n\times n}$ is the diagonal matrix satisfying $\Lambda(i,i) = \lambda_i(A)$, $i = 1, ..., n$. Then we have the relation

$$\lambda_1^{(m)}(A,b) = \max_{\substack{x\in\mathcal{K}_m(A,b)\\ x\neq 0}} \frac{x^TAx}{x^Tx} = \max_{\substack{P\in\mathcal{P}_{m-1}\\ P\neq 0}} \frac{b^TP^2(A)Ab}{b^TP^2(A)b} = \max_{\substack{P\in\mathcal{P}_{m-1}\\ P\neq 0}} \frac{\tilde{b}^TP^2(\Lambda)\Lambda\tilde{b}}{\tilde{b}^TP^2(\Lambda)\tilde{b}},$$

where $\tilde{b} = Q^Tb$. The relative error is given by

$$\frac{\lambda_1(A) - \lambda_1^{(m)}(A,b)}{\lambda_1(A) - \lambda_n(A)} = \min_{\substack{P\in\mathcal{P}_{m-1}\\ P\neq 0}} \frac{\sum_{i=2}^n \tilde{b}_i^2P^2(\lambda_i)(\lambda_1 - \lambda_i)}{(\lambda_1 - \lambda_n)\sum_{i=1}^n \tilde{b}_i^2P^2(\lambda_i)},$$

and the expected p^{th} moment of the relative error is given by

$$\mathbb{E}_{b\sim\mathcal{U}(S^{n-1})}\left[\left(\frac{\lambda_1-\lambda_1^{(m)}}{\lambda_1-\lambda_n}\right)^p\right]=\int_{S^{n-1}}\min_{\substack{P\in\mathcal{P}_{m-1}\\P\neq 0}}\left[\frac{\sum_{i=2}^n\tilde{b}_i^2P^2(\lambda_i)(\lambda_1-\lambda_i)}{(\lambda_1-\lambda_n)\sum_{i=1}^n\tilde{b}_i^2P^2(\lambda_i)}\right]^p d\sigma(\tilde{b}),$$

where σ is the uniform probability measure on S^{n-1}. Because the relative error does not depend on the norm of $\tilde{b}$ or the sign of any entry, we can replace the integral over S^{n-1} by an integral of $y=(y_1,...,y_n)$ over $[0,\infty)^n$ with respect to the joint chi-square probability density function

$$f_Y(y)=\frac{1}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^n y_i\right\}\prod_{i=1}^n y_i^{-\frac{1}{2}} \tag{5.8}$$

of n independent chi-square random variables $Y_1,...,Y_n\sim\chi_1^2$ with one degree of freedom each. In particular, we have

$$\mathbb{E}_{b\sim\mathcal{U}(S^{n-1})}\left[\left(\frac{\lambda_1-\lambda_1^{(m)}}{\lambda_1-\lambda_n}\right)^p\right]=\int_{[0,\infty)^n}\min_{\substack{P\in\mathcal{P}_{m-1}\\P\neq 0}}\left[\frac{\sum_{i=2}^n y_iP^2(\lambda_i)(\lambda_1-\lambda_i)}{(\lambda_1-\lambda_n)\sum_{i=1}^n y_iP^2(\lambda_i)}\right]^p f_Y(y)\,dy. \tag{5.9}$$

Similarly, probabilistic estimates with respect to $b\sim\mathcal{U}(S^{n-1})$ can be replaced by estimates with respect to $Y_1,...,Y_n\sim\chi_1^2$, as

$$\mathbb{P}_{b\sim\mathcal{U}(S^{n-1})}\left[\frac{\lambda_1-\lambda_1^{(m)}}{\lambda_1-\lambda_n}\geq\epsilon\right]=\mathbb{P}_{Y_i\sim\chi_1^2}\left[\min_{\substack{P\in\mathcal{P}_{m-1}\\P\neq 0}}\frac{\sum_{i=2}^n Y_iP^2(\lambda_i)(\lambda_1-\lambda_i)}{(\lambda_1-\lambda_n)\sum_{i=1}^n Y_iP^2(\lambda_i)}\geq\epsilon\right]. \tag{5.10}$$

For the remainder of the chapter, we almost exclusively use equation (5.9) for expected relative error and equation (5.10) for probabilistic bounds for relative error. If P minimizes the expression in equation (5.9) or (5.10) for a given y or Y, then any polynomial of the form αP, $\alpha\in\mathbb{R}\backslash\{0\}$, is also a minimizer. Therefore, without any loss of generality, we can alternatively minimize over the set $\mathcal{P}_{m-1}(1)=\{Q\in\mathcal{P}_{m-1}\,|\,Q(1)=1\}$. For the sake of brevity, we will omit the subscripts under $\mathbb{E}$ and $\mathbb{P}$ when the underlying distribution is clear from context. In this work, we make use of asymptotic notation to express the limiting behavior of a function with respect to n. A function $f(n)$ is $O(g(n))$ if there exists $M,n_0>0$ such that $|f(n)|\leq Mg(n)$ for all $n\geq n_0$, $o(g(n))$ if for every $\epsilon>0$ there exists

a n_ϵ such that $|f(n)| \leq \epsilon g(n)$ for all $n \geq n_0$, $\omega(g(n))$ if $|g(n)| = o(|f(n)|)$, and $\Theta(g(n))$ if $f(n) = O(g(n))$ and $g(n) = O(f(n))$.

5.3 Asymptotic Lower Bounds and Improved Upper Bounds

In this section, we obtain improved upper bounds for $\mathbb{E}\left[((\lambda_1 - \lambda_1^{(m)})/\lambda_1)^p\right]^{1/p}$, $p \geq 1$, and produce nearly-matching lower bounds. In particular, we prove the following theorem for the behavior of m as a function of n as n tends to infinity.

Theorem 19.

$$\max_{A \in \mathcal{S}^n} \mathop{\mathbb{P}}_{b \sim \mathcal{U}(S^{n-1})} \left[\frac{\lambda_1(A) - \lambda_1^{(m)}(A,b)}{\lambda_1(A) - \lambda_n(A)} \geq 1 - o(1) \right] \geq 1 - o(1/n)$$

for $m = o(\ln n)$*,*

$$\max_{A \in \mathcal{S}^n} \mathop{\mathbb{P}}_{b \sim \mathcal{U}(S^{n-1})} \left[\frac{\lambda_1(A) - \lambda_1^{(m)}(A,b)}{\lambda_1(A) - \lambda_n(A)} \geq \frac{.015 \ln^2 n}{m^2 \ln^2 \ln n} \right] \geq 1 - o(1/n)$$

for $m = \Theta(\ln n)$*,*

$$\max_{A \in \mathcal{S}^n} \mathop{\mathbb{P}}_{b \sim \mathcal{U}(S^{n-1})} \left[\frac{\lambda_1(A) - \lambda_1^{(m)}(A,b)}{\lambda_1(A) - \lambda_n(A)} \geq \frac{1.08}{m^2} \right] \geq 1 - o(1/n)$$

for $m = o\left(n^{1/2} \ln^{-1/2} n\right)$ *and* $\omega(1)$*, and*

$$\max_{A \in \mathcal{S}^n} \mathbb{E}_{b \sim \mathcal{U}(S^{n-1})} \left[\left(\frac{\lambda_1(A) - \lambda_1^{(m)}(A,b)}{\lambda_1(A) - \lambda_n(A)} \right)^p \right]^{1/p} \leq .068 \, \frac{\ln^2 \left(n(m-1)^{8p}\right)}{(m-1)^2}$$

for $n \geq 100$*,* $m \geq 10$*,* $p \geq 1$*.*

Proof. The first result is a corollary of [105, Theorem A.1], and the remaining results follow from Lemmas 20, 21, and 22. □

By Hölder's inequality, the lower bounds in Theorem 19 also hold for arbitrary p-norms,

$p \geq 1$. All results are equally applicable to λ_n, as the Krylov subspace is unaffected by shifting and scaling, namely $\mathcal{K}_m(A, b) = \mathcal{K}_m(\alpha A + \beta I, b)$ for all $\alpha \neq 0$.

We begin by producing asymptotic lower bounds. The general technique is as follows. We choose an infinite sequence of matrices $\{A_n\}_{n=1}^{\infty}$, $A_n \in \mathcal{S}^n$, treat m as a function of n, and show that, as n tends to infinity, for "most" choices of an initial vector b, the relative error of this sequence of matrices is well approximated by an integral polynomial minimization problem for which bounds can be obtained. First, we recall a number of useful propositions regarding Gauss-Legendre quadrature, orthogonal polynomials, and Chernoff bounds for chi-square random variables.

Proposition 9. *([41],[114]) Let $P \in \mathcal{P}_{2k-1}$, $\{x_i\}_{i=1}^k$ be the zeros of the k^{th} degree Legendre polynomial $P_k(x)$, $P_k(1) = 1$, in descending order ($x_1 > ... > x_k$), and $w_i = 2(1 - x_i^2)^{-1}[P_k'(x_i)]^{-2}$, $i = 1, ..., k$ be the corresponding weights. Then*

1. $\displaystyle\int_{-1}^{1} P(x)\,dx = \sum_{i=1}^{k} w_i P(x_i),$

2. $x_i = \left(1 - \frac{1}{8k^2}\right)\cos\left(\frac{(4i-1)\pi}{4k+2}\right) + O(k^{-3}), \quad i = 1, ..., k,$

3. $\displaystyle\frac{\pi}{k+\frac{1}{2}}\sqrt{1-x_1^2}\left(1 - \frac{1}{8(k+\frac{1}{2})^2(1-x_1^2)}\right) \leq w_1 < ... < w_{\lfloor\frac{k+1}{2}\rfloor} \leq \frac{\pi}{k+\frac{1}{2}}.$

Proposition 10. *([110, Section 7.72], [111]) Let $\omega(x)\,dx$ be a measure on $[-1, 1]$ with infinitely many points of increase, with orthogonal polynomials $\{p_k(x)\}_{k\geq 0}$, $p_k \in \mathcal{P}_k$. Then*

$$\max_{\substack{P\in\mathcal{P}_k \\ P\neq 0}} \frac{\int_{-1}^{1} xP^2(x)\,\omega(x)\,dx}{\int_{-1}^{1} P^2(x)\,\omega(x)\,dx} = \max\left\{x \in [-1, 1] \,\middle|\, p_{k+1}(x) = 0\right\}.$$

Proposition 11. *Let $Z \sim \chi_k^2$. Then $\mathbb{P}[Z \leq x] \leq \left[\frac{x}{k}\exp\left\{1 - \frac{x}{k}\right\}\right]^{\frac{k}{2}}$ for $x \leq k$ and $\mathbb{P}[Z \geq x] \leq \left[\frac{x}{k}\exp\left\{1 - \frac{x}{k}\right\}\right]^{\frac{k}{2}}$ for $x \geq k$.*

Proof. The result follows from taking [26, Lemma 2.2] and letting $d \to \infty$. □

We are now prepared to prove a lower bound of order m^{-2}.

Lemma 20. *If* $m = o\big(n^{1/2}\ln^{-1/2} n\big)$ *and* $\omega(1)$*, then*

$$\max_{A\in\mathcal{S}^n}\mathop{\mathbb{P}}_{b\sim\mathcal{U}(S^{n-1})}\left[\frac{\lambda_1(A)-\lambda_1^{(m)}(A,b)}{\lambda_1(A)-\lambda_n(A)}\geq\frac{1.08}{m^2}\right]\geq 1-o(1/n).$$

Proof. The structure of the proof is as follows. We choose a matrix with eigenvalues based on the zeros of a Legendre polynomial, and show that a large subset of $[0,\infty)^n$ (with respect to $f_Y(y)$) satisfies conditions that allow us to lower bound the relative error using an integral minimization problem. The choice of Legendre polynomials is based on connections to Gaussian quadrature, but, as we will later see (in Theorem 21 and experiments), the $1/m^2$ dependence is indicative of a large class of matrices with connections to orthogonal polynomials. Let $x_1 > ... > x_{2m}$ be the zeros of the $(2m)^{th}$ degree Legendre polynomial. Let $A\in\mathcal{S}^n$ have eigenvalue x_1 with multiplicity one, and remaining eigenvalues given by x_j, $j=2,...,2m$, each with multiplicity at least $\lfloor(n-1)/(2m-1)\rfloor$. By equation (5.10),

$$\begin{aligned}\mathbb{P}\left[\frac{\lambda_1-\lambda_1^{(m)}}{\lambda_1-\lambda_n}\geq\epsilon\right] &= \mathbb{P}\left[\min_{\substack{P\in\mathcal{P}_{m-1}\\P\neq 0}}\frac{\sum_{i=2}^n Y_iP^2(\lambda_i)(\lambda_1-\lambda_i)}{(\lambda_1-\lambda_n)\sum_{i=1}^n Y_iP^2(\lambda_i)}\geq\epsilon\right]\\ &= \mathbb{P}\left[\min_{\substack{P\in\mathcal{P}_{m-1}\\P\neq 0}}\frac{\sum_{j=2}^{2m}\hat{Y}_jP^2(x_j)\,(x_1-x_j)}{(x_1-x_{2m})\sum_{j=1}^{2m}\hat{Y}_jP^2\,(x_j)}\geq\epsilon\right],\end{aligned}$$

where $Y_1,...,Y_n\sim\chi_1^2$, and $\hat{Y}_j$ is the sum of the Y_i that satisfy $\lambda_i=x_j$. $\hat{Y}_1$ has one degree of freedom, and $\hat{Y}_j$, $j=2,...,2m$, each have at least $\lfloor(n-1)/(2m-1)\rfloor$ degrees of freedom. Let $w_1,...,w_{2m}$ be the weights of Gaussian quadrature associated with $x_1,...,x_{2m}$. By Proposition 9,

$$x_1=1-\frac{4+9\pi^2}{128m^2}+O(m^{-3}),\qquad x_2=1-\frac{4+49\pi^2}{128m^2}+O(m^{-3}),$$

and, therefore, $1-x_1^2=\frac{4+9\pi^2}{64m^2}+O(m^{-3})$. Again, by Proposition 9, we can lower bound

the smallest ratio between weights by

$$\begin{aligned}\frac{w_1}{w_m} &\geq \sqrt{1-x_1^2}\left(1-\frac{1}{8(2m+1/2)^2(1-x_1^2)}\right)\\ &= \left(\frac{\sqrt{4+9\pi^2}}{8m}+O(m^{-2})\right)\left(1-\frac{1}{(4+9\pi^2)/2+O(m^{-1})}\right)\\ &= \frac{2+9\pi^2}{8\sqrt{4+9\pi^2}\,m}+O(m^{-2}).\end{aligned}$$

Therefore, by Proposition 11,

$$\begin{aligned}\mathbb{P}\left[\min_{j\geq 2}\hat{Y}_j \geq \frac{w_m}{w_1}\hat{Y}_1\right] &\geq \mathbb{P}\left[\hat{Y}_j \geq \frac{1}{3}\left\lfloor\frac{n-1}{2m-1}\right\rfloor,\, j\geq 2\right]\times\mathbb{P}\left[\hat{Y}_1 \leq \frac{w_1}{3w_m}\left\lfloor\frac{n-1}{2m-1}\right\rfloor\right]\\ &\geq \left[1-\left(\frac{e^{2/3}}{3}\right)^{\frac{1}{2}\left\lfloor\frac{n-1}{2m-1}\right\rfloor}\right]^{2m-1}\left[1-\left(\frac{w_1}{3w_m}\left\lfloor\frac{n-1}{2m-1}\right\rfloor e^{1-\frac{w_1}{3w_m}\left\lfloor\frac{n-1}{2m-1}\right\rfloor}\right)^{\frac{1}{2}}\right]\\ &= 1-o(1/n).\end{aligned}$$

We now restrict our attention to values of $Y=(Y_1,...,Y_n)\in[0,\infty)^n$ that satisfy the inequality $w_1\min_{j\geq 2}\hat{Y}_j \geq w_m\hat{Y}_1$. If, for some fixed choice of Y,

$$\min_{\substack{P\in\mathcal{P}_{m-1}\\P\neq 0}}\frac{\sum_{j=2}^{2m}\hat{Y}_jP^2(x_j)(x_1-x_j)}{\sum_{j=1}^{2m}\hat{Y}_jP^2(x_j)} \leq x_1-x_2, \tag{5.11}$$

then, by Proposition 9 and Proposition 10 for $\omega(x)=1$,

$$\begin{aligned}\min_{\substack{P\in\mathcal{P}_{m-1}\\P\neq 0}}\frac{\sum_{j=2}^{2m}\hat{Y}_jP^2(x_j)(x_1-x_j)}{\sum_{j=1}^{2m}\hat{Y}_jP^2(x_j)} &\geq \min_{\substack{P\in\mathcal{P}_{m-1}\\P\neq 0}}\frac{\sum_{j=2}^{2m}w_jP^2(x_j)(x_1-x_j)}{\sum_{j=1}^{2m}w_jP^2(x_j)}\\ &= \min_{\substack{P\in\mathcal{P}_{m-1}\\P\neq 0}}\frac{\sum_{j=1}^{2m}w_jP^2(x_j)(1-x_j)}{\sum_{j=1}^{2m}w_jP^2(x_j)}-(1-x_1)\\ &= \min_{\substack{P\in\mathcal{P}_{m-1}\\P\neq 0}}\frac{\int_{-1}^{1}P^2(y)(1-y)\,dy}{\int_{-1}^{1}P^2(y)\,dy}-\frac{4+9\pi^2}{128m^2}+O(m^{-3})\\ &= \frac{4+9\pi^2}{32m^2}-\frac{4+9\pi^2}{128m^2}+O(m^{-3})=\frac{12+27\pi^2}{128m^2}+O(m^{-3}).\end{aligned}$$

Alternatively, if equation (5.11) does not hold, then we can lower bound the left hand side of (5.11) by $x_1 - x_2 = \frac{5\pi^2}{16m^2} + O(m^{-3})$. Noting that $x_1 - x_{2m} \leq 2$ completes the proof, as $\frac{12+27\pi^2}{256}$ and $\frac{5\pi^2}{32}$ are both greater than 1.08. □

The previous lemma illustrates that the inverse quadratic dependence of the relative error on the number of iterations m persists up to $m = o(n^{1/2}/\ln^{1/2} n)$. As we will see in Sections 4 and 6, this m^{-2} error is indicative of the behavior of the Lanczos method for a wide range of typical matrices encountered in practice. Next, we aim to produce a lower bound of the form $\ln^2 n/(m^2 \ln^2 \ln n)$. To do so, we will make use of more general Jacobi polynomials instead of Legendre polynomials. In addition, in order to obtain a result that contains some dependence on n, the choice of Jacobi polynomial must vary with n (i.e., α and β are a function of n). However, due to the distribution of eigenvalues required to produce this bound, Gaussian quadrature is no longer exact, and we must make use of estimates for basic composite quadrature. We recall the following propositions regarding composite quadrature and Jacobi polynomials.

Proposition 12. *If $f \in \mathcal{C}^1[a,b]$, then for $a = x_0 < ... < x_n = b$,*

$$\left| \int_a^b f(x)\,dx - \sum_{i=1}^{n} (x_i - x_{i-1}) f(x_i^*) \right| \leq \frac{b-a}{2} \max_{x\in[a,b]} |f'(x)| \max_{i=1,...,n} (x_i - x_{i-1}),$$

where $x_i^* \in [x_{i-1}, x_i]$, $i = 1, ..., n$.

Proposition 13. *([103, Chapter 3.2]) Let $\{P_k^{(\alpha,\beta)}(x)\}_{k=0}^{\infty}$, $\alpha, \beta > -1$, be the orthogonal system of Jacobi polynomials over $[-1,1]$ with respect to weight function $\omega^{\alpha,\beta}(x) = (1-x)^\alpha (1+x)^\beta$, namely,*

$$P_k^{(\alpha,\beta)}(x) = \frac{\Gamma(k+\alpha+1)}{k!\,\Gamma(k+\alpha+\beta+1)} \sum_{i=0}^{k} \binom{k}{i} \frac{\Gamma(k+i+\alpha+\beta+1)}{\Gamma(i+\alpha+1)} \left(\frac{x-1}{2}\right)^i.$$

Then

(i) $$\int_{-1}^{1} \left[P_k^{(\alpha,\beta)}(x)\right]^2 \omega^{\alpha,\beta}(x)\,dx = \frac{2^{\alpha+\beta+1}\Gamma(k+\alpha+1)\Gamma(k+\beta+1)}{(2k+\alpha+\beta+1)k!\,\Gamma(k+\alpha+\beta+1)},$$

(ii) $\max_{x\in[-1,1]} \left|P_k^{(\alpha,\beta)}(x)\right| = \max\left\{\frac{\Gamma(k+\alpha+1)}{k!\,\Gamma(\alpha+1)}, \frac{\Gamma(k+\beta+1)}{k!\,\Gamma(\beta+1)}\right\}$ *for* $\max\{\alpha,\beta\} \geq -\frac{1}{2}$,

(iii) $\frac{d}{dx}P_k^{(\alpha,\beta)}(x) = \frac{k+\alpha+\beta+1}{2}P_{k-1}^{(\alpha+1,\beta+1)}(x)$,

(iv) $\max\{x \in [-1,1] \,|\, P_k^{(\alpha,\beta)}(x) = 0\} \leq \sqrt{1-\left(\frac{\alpha+3/2}{k+\alpha+1/2}\right)^2}$ *for* $\alpha \geq \beta > -\frac{11}{12}$ *and* $k > 1$.

Proof. Equations (i)-(iii) are standard results, and can be found in [103, Chapter 3.2] or [110]. What remains is to prove (iv). By [90, Lemma 3.5], the largest zero x_1 of the k^{th} degree Gegenbauer polynomial (i.e., $P_k^{(\lambda-1/2,\lambda-1/2)}(x)$), with $\lambda > -5/12$, satisfies

$$x_1^2 \leq \frac{(k-1)(k+2\lambda+1)}{(k+\lambda)^2+3\lambda+\frac{5}{4}+3(\lambda+\frac{1}{2})^2/(k-1)} \leq \frac{(k-1)(k+2\lambda+1)}{(k+\lambda)^2} = 1-\left(\frac{\lambda+1}{k+\lambda}\right)^2.$$

By [37, Theorem 2.1], the largest zero of $P_k^{(\alpha,\beta+t)}(x)$ is strictly greater than the largest zero of $P_k^{(\alpha,\beta)}(x)$ for any $t > 0$. As $\alpha \geq \beta$, combining these two facts provides our desired result. □

For the sake of brevity, the inner product and norm on $[-1,1]$ with respect to $\omega^{\alpha,\beta}(x) = (1-x)^\alpha(1+x)^\beta$ will be denoted by $\langle\cdot,\cdot\rangle_{\alpha,\beta}$ and $\|\cdot\|_{\alpha,\beta}$. We are now prepared to prove the following lower bound.

Lemma 21. *If* $m = \Theta(\ln n)$*, then*

$$\max_{A\in\mathcal{S}^n} \mathbb{P}_{b\sim\mathcal{U}(S^{n-1})}\left[\frac{\lambda_1(A)-\lambda_1^{(m)}(A,b)}{\lambda_1(A)-\lambda_n(A)} \geq \frac{.015\ln^2 n}{m^2\ln^2\ln n}\right] \geq 1-o(1/n).$$

Proof. The structure of the proof is similar in concept to that of Lemma 20. We choose a matrix with eigenvalues based on a function corresponding to an integral minimization problem, and show that a large subset of $[0,\infty)^n$ (with respect to $f_Y(y)$) satisfies conditions that allow us to lower bound the relative error using the aforementioned integral minimization problem. The main difficulty is that, to obtain improved bounds, we must use Proposition 10

with a weight function that requires quadrature for functions that are no longer polynomials, and, therefore, cannot be represented exactly using Gaussian quadrature. In addition, the required quadrature is for functions whose derivative has a singularity at $x = 1$, and so Proposition 12 is not immediately applicable. Instead, we perform a two-part error analysis. In particular, if a function is $C^1[0, a]$, $a < 1$, and monotonic on $[a, 1]$, then using Proposition 12 on $[0, a]$ and a monotonicity argument on $[a, 1]$ results in an error bound for quadrature.

Let $\ell = \left\lfloor \frac{.2495 \ln n}{\ln \ln n} \right\rfloor$, $k = \left\lfloor m^{4.004\ell} \right\rfloor$, and $m = \Theta(\ln n)$. We assume that n is sufficiently large, so that $\ell > 0$. ℓ is quite small for practical values of n, but the growth of ℓ with respect to n is required for a result with dependence on n. Consider a matrix $A \in \mathcal{S}^n$ with eigenvalues given by $f(x_j)$, $j = 1, ..., k$, each with multiplicity either $\lfloor n/k \rfloor$ or $\lceil n/k \rceil$, where $f(x) = 1 - 2(1 - x)^{\frac{1}{\ell}}$, and $x_j = j/k$, $j = 1, ..., k$. Then

$$\begin{aligned}
\mathbb{P}\left[\frac{\lambda_1 - \lambda_1^{(m)}}{\lambda_1 - \lambda_n} \geq \epsilon\right] &= \mathbb{P}\left[\min_{\substack{P \in \mathcal{P}_{m-1} \\ P \neq 0}} \frac{\sum_{i=2}^n Y_i P^2(\lambda_i)(\lambda_1 - \lambda_i)}{(\lambda_1 - \lambda_n)\sum_{i=1}^n Y_i P^2(\lambda_i)} \geq \epsilon\right] \\
&\geq \mathbb{P}\left[\min_{\substack{P \in \mathcal{P}_{m-1} \\ P \neq 0}} \frac{\sum_{j=1}^k \hat{Y}_j P^2(f(x_j))(1 - f(x_j))}{2\sum_{j=1}^k \hat{Y}_j P^2(f(x_j))} \geq \epsilon\right],
\end{aligned}$$

where $Y_1, ..., Y_n \sim \chi_1^2$, and $\hat{Y}_j$ is the sum of the Y_i's that satisfy $\lambda_i = f(x_j)$. Each $\hat{Y}_j$, $j = 1, ..., k$, has either $\lfloor n/k \rfloor$ or $\lceil n/k \rceil$ degrees of freedom. Because $m = \Theta(\ln n)$, we have $k = o(n^{.999})$ and, by Proposition 11,

$$\begin{aligned}
\mathbb{P}\left[.999\lfloor \tfrac{n}{k} \rfloor \leq \hat{Y}_j \leq 1.001\lceil \tfrac{n}{k} \rceil,\ j = 1, ..., k\right] &\geq \left(1 - \left(.999e^{.001}\right)^{\lfloor \frac{n}{k} \rfloor} - \left(1.001e^{-.001}\right)^{\lceil \frac{n}{k} \rceil}\right)^k \\
&= 1 - o(1/n).
\end{aligned}$$

Therefore, with probability $1 - o(1/n)$,

$$\min_{\substack{P \in \mathcal{P}_{m-1} \\ P \neq 0}} \frac{\sum_{j=1}^k \hat{Y}_j P^2(f(x_j))(1 - f(x_j))}{2\sum_{j=1}^k \hat{Y}_j P^2(f(x_j))} \geq \frac{.998}{2} \min_{\substack{P \in \mathcal{P}_{m-1} \\ P \neq 0}} \frac{\sum_{j=1}^k P^2(f(x_j))(1 - f(x_j))}{\sum_{j=1}^k P^2(f(x_j))}.$$

Let $P(y) = \sum_{r=0}^{m-1} c_r P_r^{(\ell,0)}(y)$, $c_r \in \mathbb{R}$, $r = 0, ..., m-1$, and define

$$g_{r,s}(x) = P_r^{(\ell,0)}(f(x)) P_s^{(\ell,0)}(f(x)) \quad \text{and} \quad \hat{g}_{r,s}(x) = g_{r,s}(x)(1 - f(x)),$$

$r, s = 0, ..., m-1$. Then we have

$$\frac{\sum_{j=1}^{k} P^2(f(x_j))(1 - f(x_j))}{\sum_{j=1}^{k} P^2(f(x_j))} = \frac{\sum_{r,s=0}^{m-1} c_r c_s \sum_{j=1}^{k} \hat{g}_{r,s}(x_j)}{\sum_{r,s=0}^{m-1} c_r c_s \sum_{j=1}^{k} g_{r,s}(x_j)}.$$

We now replace each quadrature $\sum_{j=1}^{k} \hat{g}_{r,s}(x_j)$ or $\sum_{j=1}^{k} g_{r,s}(x_j)$ in the previous expression by its corresponding integral, plus a small error term. The functions $g_{r,s}$ and $\hat{g}_{r,s}$ are not elements of $\mathcal{C}^1[0,1]$, as $f'(x)$ has a singularity at $x = 1$, and, therefore, we cannot use Proposition 12 directly. Instead we break the error analysis of the quadrature into two components. We have $g_{r,s}, \hat{g}_{r,s} \in \mathcal{C}^1[0,a]$ for any $0 < a < 1$, and, if a is chosen to be large enough, both $g_{r,s}$ and $\hat{g}_{r,s}$ will be monotonic on the interval $[a,1]$. In that case, we can apply Proposition 12 to bound the error over the interval $[0,a]$ and use basic properties of Riemann sums of monotonic functions to bound the error over the interval $[a,1]$.

The function $f(x)$ is increasing on $[0,1]$, and, by equations (iii) and (iv) of Proposition 13, the function $P_r^{(\ell,0)}(y)$, $r = 2, ..., m-1$, is increasing on the interval

$$\left[\sqrt{1 - \left(\frac{\ell + 5/2}{m + \ell - 1/2} \right)^2}, \ 1 \right]. \tag{5.12}$$

The functions $P_0^{(\ell,0)}(y) = 1$ and $P_1^{(\ell,0)}(y) = \ell + 1 + (\ell+2)(y-1)/2$ are clearly non-decreasing over interval (5.12). By the inequality $\sqrt{1-y} \leq 1 - y/2$, $y \in [0,1]$, the function $P_r^{(\ell,0)}(f(x))$, $r = 0, ..., m-1$, is non-decreasing on the interval

$$\left[1 - \left(\frac{\ell + 5/2}{2m + 2\ell - 1} \right)^{2\ell}, \ 1 \right]. \tag{5.13}$$

Therefore, the functions $g_{r,s}(x)$ are non-decreasing and $\hat{g}_{r,s}(x)$ are non-increasing on the interval (5.13) for all $r, s = 0, ..., m-1$. The term $\left(\frac{\ell+5/2}{2m+2\ell-1} \right)^{2\ell} = \omega(1/k)$, and so, for

sufficiently large n, there exists an index $j^* \in \{1, ..., k-2\}$ such that

$$1 - \left(\frac{\ell + 5/2}{2m + 2\ell - 1}\right)^{2\ell} \leq x_{j^*} < 1 - \left(\frac{\ell + 5/2}{2m + 2\ell - 1}\right)^{2\ell} + \frac{1}{k} < 1.$$

We now upper bound the derivatives of $g_{r,s}$ and $\hat{g}_{r,s}$ on the interval $[0, x_{j^*}]$. By equation (iii) of Proposition 13, the first derivatives of $g_{r,s}$ and $\hat{g}_{r,s}$ are given by

$$\begin{aligned} g'_{r,s}(x) &= f'(x)\left(\left[\frac{d}{dy}P_r^{(\ell,0)}(y)\right]_{y=f(x)} P_s^{(\ell,0)}(f(x)) + P_r^{(\ell,0)}(f(x))\left[\frac{d}{dy}P_s^{(\ell,0)}(y)\right]_{y=f(x)}\right) \\ &= \frac{(1-x)^{-\frac{\ell-1}{\ell}}}{\ell}\Bigg((r+\ell+1)P_{r-1}^{(\ell+1,1)}(f(x))P_s^{(\ell,0)}(f(x)) \\ &\qquad +(s+\ell+1)P_r^{(\ell,0)}(f(x))P_{s-1}^{(\ell+1,1)}(f(x))\Bigg), \end{aligned}$$

and

$$\hat{g}'_{r,s}(x) = g'_{r,s}(x)(1-x)^{\frac{1}{\ell}} - 2g_{r,s}(x)\frac{(1-x)^{-\frac{\ell-1}{\ell}}}{\ell}.$$

By equation (ii) of Proposition 13, and the inequality $\binom{x}{y} \leq \left(\frac{ex}{y}\right)^y$, $x, y \in \mathbb{N}$, $y \leq x$,

$$\begin{aligned} \max_{x\in[0,x_{j^*}]} |g'_{r,s}(x)| &\leq \frac{(1-x_{j^*})^{-\frac{\ell-1}{\ell}}}{\ell}\Bigg((r+\ell+1)\binom{r+\ell}{\ell+1}\binom{s+\ell}{\ell} \\ &\qquad +(s+\ell+1)\binom{r+\ell}{\ell}\binom{s+\ell}{\ell+1}\Bigg) \\ &\leq 2\,\frac{m+\ell}{\ell}\left(\left(\frac{\ell+5/2}{2m+2\ell-1}\right)^{2\ell} - \frac{1}{k}\right)^{-\frac{\ell-1}{\ell}}\left(\frac{e(m+\ell-1)}{\ell}\right)^{2\ell+1}, \end{aligned}$$

and

$$\begin{aligned} \max_{x\in[0,x_{j^*}]} \left|\hat{g}'_{r,s}(x)\right| &\leq \max_{x\in[0,x_{j^*}]} |g'_{r,s}(x)| + 2\frac{(1-x_{j^*})^{-\frac{\ell-1}{\ell}}}{\ell}\binom{r+\ell}{\ell}\binom{s+\ell}{\ell} \\ &\leq 4\,\frac{m+\ell}{\ell}\left(\left(\frac{\ell+5/2}{2m+2\ell-1}\right)^{2\ell} - \frac{1}{k}\right)^{-\frac{\ell-1}{\ell}}\left(\frac{e(m+\ell-1)}{\ell}\right)^{2\ell+1}. \end{aligned}$$

Therefore, both $\max_{x\in[0,x_{j^*}]} |g'_{r,s}(x)|$ and $\max_{x\in[0,x_{j^*}]} |\hat{g}'_{r,s}(x)|$ are $o(m^{4.002\ell})$. Then, by

Proposition 12, equation (ii) of Proposition 13 and monotonicity on the interval $[x_{j^*}, 1]$, we have

$$\begin{aligned}
\left|\frac{1}{k}\sum_{j=1}^{k} g_{r,s}(x_j) - \int_0^1 g_{r,s}(x)\,dx\right| &\leq \left|\frac{1}{k}\sum_{j=1}^{j^*} g_{r,s}(x_j) - \int_0^{x_{j^*}} g_{r,s}(x)\,dx\right| \\
&\quad + \left|\frac{1}{k}\sum_{j=j^*+1}^{k} g_{r,s}(x_j) - \int_a^1 g_{r,s}(x)\,dx\right| \\
&\leq \frac{1+o(1)}{2km^{4.002\ell}} + \frac{g_{r,s}(1)}{k} \\
&\leq \frac{1+o(1)}{2m^{.002\ell}} + \frac{1}{k}\left(\frac{e(m+\ell+1)}{\ell+1}\right)^{2(\ell+1)} \\
&\leq \frac{1+o(1)}{2m^{.002\ell}},
\end{aligned}$$

and, similarly,

$$\begin{aligned}
\left|\frac{1}{k}\sum_{j=1}^{k} \hat{g}_{r,s}(x_j) - \int_0^1 \hat{g}_{r,s}(x)\,dx\right| &\leq \left|\frac{1}{k}\sum_{j=1}^{j^*} \hat{g}_{r,s}(x_j) - \int_0^{x_{j^*}} \hat{g}_{r,s}(x)\,dx\right| \\
&\quad + \left|\frac{1}{k}\sum_{j=j^*+1}^{k} \hat{g}_{r,s}(x_j) - \int_{x_{j^*}}^1 \hat{g}_{r,s}(x)\,dx\right| \\
&\leq \frac{1+o(1)}{2m^{.002\ell}} + \frac{\hat{g}_{r,s}(x_{j^*})}{k} \\
&\leq \frac{1+o(1)}{2m^{.002\ell}} + \frac{g_{r,s}(1)}{k} \\
&\leq \frac{1+o(1)}{2m^{.002\ell}}.
\end{aligned}$$

Let us denote this upper bound by $M = (1+o(1))/(2m^{.002\ell})$. By using the substitution $x = 1 - (\frac{1-y}{2})^\ell$, we have

$$\int_0^1 \hat{g}_{r,s}(x)\,dx = \frac{\ell}{2^\ell}\langle P_r^{(\ell,0)}, P_s^{(\ell,0)}\rangle_{\ell,0} \quad \text{and} \quad \int_0^1 g_{r,s}(x)\,dx = \frac{\ell}{2^\ell}\langle P_r^{(\ell,0)}, P_s^{(\ell,0)}\rangle_{\ell-1,0}.$$

Then

$$\max_{\substack{P\in\mathcal{P}_{m-1}\\P\neq 0}} \frac{\sum_{j=1}^k P^2(f(x_j))}{\sum_{j=1}^k P^2(f(x_j))(1-f(x_j))} \leq \max_{\substack{c\in\mathbb{R}^m\setminus 0\\|\epsilon_{r,s}|\leq M\\|\hat{\epsilon}_{r,s}|\leq M}} \frac{\sum_{r,s=0}^{m-1} c_r c_s \left[\frac{\ell}{2^\ell}\langle P_r^{(\ell,0)}, P_s^{(\ell,0)}\rangle_{\ell-1,0} + \epsilon_{r,s}\right]}{\sum_{r,s=0}^{m-1} c_r c_s \left[\frac{\ell}{2^\ell}\langle P_r^{(\ell,0)}, P_s^{(\ell,0)}\rangle_{\ell,0} + \hat{\epsilon}_{r,s}\right]}.$$

Letting $\tilde{c}_r = c_r\|P_r^{(\ell,0)}\|_{\ell,0}$, $r = 0, ..., m-1$, and noting that, by equation (i) of Proposition 13, $\|P_r^{(\ell,0)}\|_{\ell,0} = \left(\frac{2^{\ell+1}}{2r+\ell+1}\right)^{1/2}$, we obtain the bound

$$\max_{\substack{P\in\mathcal{P}_{m-1}\\P\neq 0}} \frac{\sum_{j=1}^k P^2(f(x_j))}{\sum_{j=1}^k P^2(f(x_j))(1-f(x_j))} \leq \max_{\substack{\tilde{c}\in\mathbb{R}^m\setminus 0\\|\epsilon_{r,s}|\leq \epsilon\\|\hat{\epsilon}_{r,s}|\leq \epsilon}} \frac{\sum_{r,s=0}^{m-1} \tilde{c}_r\tilde{c}_s \left[\frac{\langle P_r^{(\ell,0)}, P_s^{(\ell,0)}\rangle_{\ell-1,0}}{\|P_r^{(\ell,0)}\|_{\ell,0}\|P_s^{(\ell,0)}\|_{\ell,0}} + \epsilon_{r,s}\right]}{\sum_{r=0}^{m-1}\tilde{c}_r^2 + \sum_{r,s=0}^{m-1}\tilde{c}_r\tilde{c}_s\hat{\epsilon}_{r,s}},$$

where

$$\epsilon = \frac{M(2m+\ell-1)}{\ell} = (1+o(1))\frac{2m+\ell-1}{2m^{.002\ell}\ell} = o(1/m).$$

Let $B \in \mathbb{R}^{m\times m}$ be given by $B(r,s) = \langle P_r^{(\ell,0)}, P_s^{(\ell,0)}\rangle_{\ell-1,0}\|P_r^{(\ell,0)}\|_{\ell,0}^{-1}\|P_s^{(\ell,0)}\|_{\ell,0}^{-1}$, $r, s = 0, ..., m-1$. By Proposition 10 applied to $\omega^{\ell-1,0}(x)$ and equation (iv) of Proposition 13, we have

$$\max_{\tilde{c}\in\mathbb{R}^m\setminus 0}\frac{\tilde{c}^T B\tilde{c}}{\tilde{c}^T\tilde{c}} \leq \frac{1}{1-\sqrt{1-\left(\frac{\ell+1/2}{m+\ell-1/2}\right)^2}} \leq 2\left(\frac{m+\ell-1/2}{\ell+1/2}\right)^2.$$

This implies that

$$\begin{aligned}
\max_{\substack{P\in\mathcal{P}_{m-1}\\P\neq 0}} \frac{\sum_{j=1}^k P^2(f(x_j))}{\sum_{j=1}^k P^2(f(x_j))(1-f(x_j))} &\leq \max_{\substack{\tilde{c}\in\mathbb{R}^m\\\tilde{c}\neq 0}} \frac{\tilde{c}^T\left(B+\epsilon\mathbb{1}\mathbb{1}^T\right)\tilde{c}}{\tilde{c}^T\left(I-\epsilon\mathbb{1}\mathbb{1}^T\right)\tilde{c}}\\
&\leq \frac{1}{1-\epsilon m}\max_{\substack{\tilde{c}\in\mathbb{R}^m\\\tilde{c}\neq 0}}\left[\frac{\tilde{c}^T B\tilde{c}}{\tilde{c}^T\tilde{c}}\right] + \frac{\epsilon m}{1-\epsilon m}\\
&\leq \frac{2}{1-\epsilon m}\left(\frac{m+\ell-1/2}{\ell+1/2}\right)^2 + \frac{\epsilon m}{1-\epsilon m},
\end{aligned}$$

and that, with probability $1 - o(1/n)$,

$$\min_{\substack{P \in \mathcal{P}_{m-1} \\ P \neq 0}} \frac{\sum_{j=1}^{k} \hat{Y}_j P^2\left(f(x_j)\right)\left(1 - f(x_j)\right)}{\sum_{j=1}^{k} \hat{Y}_j P^2\left(f(x_j)\right)} \geq (1 - o(1)) \frac{.998}{4} \left(\frac{\ell + 1/2}{m + \ell - 1/2}\right)^2 \geq \frac{.015 \ln^2 n}{m^2 \ln^2 \ln n}.$$

This completes the proof. □

In the previous proof, a number of very small constants were used, exclusively for the purpose of obtaining an estimate with constant as close to $1/64$ as possible. The constants used in the definition of ℓ and k can be replaced by more practical numbers (that begin to exhibit asymptotic convergence for reasonably sized n), at the cost of a constant worse than $.015$. This completes the analysis of asymptotic lower bounds.

We now move to upper bounds for relative error in the p-norm. Our estimate for relative error in the one-norm is of the same order as the estimate (5.4), but with an improved constant. Our technique for obtaining these estimates differs from the technique of [67] in one key way. Rather than integrating first by b_1 and using properties of the arctan function, we replace the ball B^n by n chi-square random variables on $[0, \infty)^n$, and iteratively apply Cauchy-Schwarz to our relative error until we obtain an exponent c for which the inverse chi-square distribution with one degree of freedom has a convergent c^{th} moment.

Lemma 22. *Let $n \geq 100$, $m \geq 10$, and $p \geq 1$. Then*

$$\max_{A \in \mathcal{S}^n} \mathbb{E}_{b \sim \mathcal{U}(S^{n-1})} \left[\left(\frac{\lambda_1(A) - \lambda_1^{(m)}(A, b)}{\lambda_1(A) - \lambda_n(A)} \right)^p \right]^{\frac{1}{p}} \leq .068 \, \frac{\ln^2 \left(n(m-1)^{8p}\right)}{(m-1)^2}.$$

Proof. Without loss of generality, suppose $\lambda_1(A) = 1$ and $\lambda_n(A) = 0$. By repeated application of Cauchy Schwarz,

$$\frac{\sum_{i=1}^{n} y_i Q^2(\lambda_i)(1 - \lambda_i)}{\sum_{i=1}^{n} y_i Q^2(\lambda_i)} \leq \left[\frac{\sum_{i=1}^{n} y_i Q^2(\lambda_i)(1 - \lambda_i)^2}{\sum_{i=1}^{n} y_i Q^2(\lambda_i)} \right]^{\frac{1}{2}} \leq \left[\frac{\sum_{i=1}^{n} y_i Q^2(\lambda_i)(1 - \lambda_i)^{2^q}}{\sum_{i=1}^{n} y_i Q^2(\lambda_i)} \right]^{\frac{1}{2^q}}$$

for $q \in \mathbb{N}$. Choosing q to satisfy $2p < 2^q \leq 4p$, and using equation (5.9) with polynomial

normalization $Q(1) = 1$, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\lambda_1 - \lambda_1^{(m)}}{\lambda_1 - \lambda_n}\right)^p\right]^{\frac{1}{p}} &\leq \min_{Q\in\mathcal{P}_{m-1}(1)} \left[\int_{[0,\infty)^n} \left(\frac{\sum_{i=2}^n y_i Q^2(\lambda_i)(1-\lambda_i)^{2^q}}{\sum_{i=1}^n y_i Q^2(\lambda_i)}\right)^{\frac{p}{2^q}} f_Y(y)\, dy\right]^{\frac{1}{p}} \\
&\leq \min_{Q\in\mathcal{P}_{m-1}(1)} \left[\int_{[0,\infty)^n} \left(\frac{\sum_{i:\lambda_i<\beta} y_i Q^2(\lambda_i)(1-\lambda_i)^{2^q}}{\sum_{i=1}^n y_i Q^2(\lambda_i)}\right)^{\frac{p}{2^q}} f_Y(y)\, dy\right]^{\frac{1}{p}} \\
&\quad + \left[\int_{[0,\infty)^n} \left(\frac{\sum_{i:\lambda_i\geq\beta} y_i Q^2(\lambda_i)(1-\lambda_i)^{2^q}}{\sum_{i=1}^n y_i Q^2(\lambda_i)}\right)^{\frac{p}{2^q}} f_Y(y)\, dy\right]^{\frac{1}{p}}
\end{aligned}$$

for any $\beta \in (0,1)$. The integrand of the first term satisfies

$$\begin{aligned}
\left(\frac{\sum_{i:\lambda_i<\beta} y_i Q^2(\lambda_i)(1-\lambda_i)^{2^q}}{\sum_{i=1}^n y_i Q^2(\lambda_i)}\right)^{\frac{p}{2^q}} &\leq \left(\frac{\sum_{i:\lambda_i<\beta} y_i Q^2(\lambda_i)(1-\lambda_i)^{2^q}}{\sum_{i=1}^n y_i Q^2(\lambda_i)}\right)^{\frac{1}{4}} \\
&\leq \max_{x\in[0,\beta]} |Q(x)|^{1/2}(1-x)^{2^{q-2}} \left(\frac{\sum_{i:\lambda_i<\beta} y_i}{y_1}\right)^{\frac{1}{4}},
\end{aligned}$$

and the second term is always bounded above by $\max_{\lambda\in[\beta,1]}(1-\lambda) = 1-\beta$. We replace the minimizing polynomial in $\mathcal{P}_{m-1}(1)$ by $\widehat{Q}(x) = \frac{T_{m-1}\left(\frac{2}{\beta}x-1\right)}{T_{m-1}\left(\frac{2}{\beta}-1\right)}$, where $T_{m-1}(\cdot)$ is the Chebyshev polynomial of the first kind. The Chebyshev polynomials $T_{m-1}(\cdot)$ are bounded by one in magnitude on the interval $[-1,1]$, and this bound is tight at the endpoints. Therefore, our maximum is achieved at $x = 0$, and

$$\max_{x\in[0,\beta]} |\widehat{Q}(x)|^{1/2}(1-x)^{2^{q-2}} = \left|T_{m-1}\left(\frac{2}{\beta}-1\right)\right|^{-1/2}.$$

By the definition $T_{m-1}(x) = 1/2\left(\left(x-\sqrt{x^2-1}\right)^{m-1} + \left(x+\sqrt{x^2-1}\right)^{m-1}\right)$, $|x| \geq 1$,

and the standard inequality $e^{2x} \leq \frac{1+x}{1-x}$, $x \in [0,1]$,

$$\begin{aligned} T_{m-1}\left(\frac{2}{\beta}-1\right) &\geq \frac{1}{2}\left(\frac{2}{\beta}-1+\sqrt{\left(\frac{2}{\beta}-1\right)^2-1}\right)^{m-1} = \frac{1}{2}\left(\frac{1+\sqrt{1-\beta}}{1-\sqrt{1-\beta}}\right)^{m-1} \\ &\geq \frac{1}{2}\exp\left\{2\sqrt{1-\beta}\,(m-1)\right\}. \end{aligned}$$

In addition,

$$\int_{[0,\infty)^n}\left(\frac{\sum_{i=2}^n y_i}{y_1}\right)^{1/4} f_Y(y)\,dy = \frac{\Gamma\left(n/2-1/4\right)\Gamma(1/4)}{\Gamma(n/2-1/2)\Gamma(1/2)} \leq \frac{\Gamma(1/4)}{2^{1/4}\Gamma(1/2)}n^{1/4},$$

which gives us

$$\mathbb{E}\left[\left(\frac{\lambda_1-\lambda_1^{(m)}}{\lambda_1-\lambda_n}\right)^p\right]^{\frac{1}{p}} \leq \left[\frac{2^{1/4}\Gamma(1/4)}{\Gamma(1/2)}n^{1/4}\right]^{1/p} e^{-\gamma(m-1)/p}+\gamma^2,$$

where $\gamma = \sqrt{1-\beta}$. Setting $\gamma = \frac{p}{m-1}\ln\left(n^{1/4p}(m-1)^2\right)$ (assuming $\gamma < 1$, otherwise our bound is already greater than one, and trivially holds), we obtain

$$\begin{aligned} \mathbb{E}\left[\left(\frac{\lambda_1-\lambda_1^{(m)}}{\lambda_1-\lambda_n}\right)^p\right]^{1/p} &\leq \frac{\left(\frac{2^{1/4}\Gamma(1/4)}{\Gamma(1/2)}\right)^{1/p}+p^2\ln^2(n^{1/4p}(m-1)^2)}{(m-1)^2} \\ &= \frac{\left(\frac{2^{1/4}\Gamma(1/4)}{\Gamma(1/2)}\right)^{1/p}+\frac{1}{16}\ln^2\left(n(m-1)^{8p}\right)}{(m-1)^2} \\ &\leq .068\,\frac{\ln^2\left(n(m-1)^{8p}\right)}{(m-1)^2}, \end{aligned}$$

for $m \geq 10$, $n \geq 100$. The constant of $.068$ is produced by upper bounding $\frac{2^{1/4}\Gamma(1/4)}{\Gamma(1/2)}$ by a constant[1] times $\ln^2\left[n(m-1)^{8p}\right]$. This completes the proof.

□

A similar proof, paired with probabilistic bounds on the quantity $\sum_{i=2}^n Y_i/Y_1$, where

[1]The best constant in this case is given by the ratio of $\frac{2^{1/4}\Gamma(1/4)}{\Gamma(1/2)}$ to the minimum value of $\ln^2\left[n(m-1)^{8p}\right]$ over $m \geq 10$, $n \geq 100$.

$Y_1, ..., Y_n \sim \chi_1^2$, gives a probabilistic upper estimate. Combining the lower bounds in Lemmas 20 and 21 with the upper bound of Lemma 22 completes the proof of Theorem 19.

5.4 Distribution Dependent Bounds

In this section, we consider improved estimates for matrices with certain specific properties. First, we show that if a matrix A has a reasonable number of eigenvalues near $\lambda_1(A)$, then we can produce an error estimate that depends only on the number of iterations m. The intuition is that if there are not too few eigenvalues near the maximum eigenvalue, then the initial vector has a large inner product with a vector with Rayleigh quotient close to the maximum eigenvalue. In particular, we suppose that the eigenvalues of our matrix A are such that, once scaled, there are at least $n/(m-1)^\alpha$ eigenvalues in the range $[\beta, 1]$, for a specific value of β satisfying $1 - \beta = O\left(m^{-2} \ln^2 m\right)$. For a large number of matrices for which the Lanczos method is used, this assumption holds true. Under this assumption, we prove the following error estimate.

Theorem 20. *Let $A \in \mathcal{S}^n$, $m \geq 10$, $p \geq 1$, $\alpha > 0$, and $n \geq m(m-1)^\alpha$. If*

$$\#\left\{\lambda_i(A) \,\middle|\, \frac{\lambda_1(A) - \lambda_i(A)}{\lambda_1(A) - \lambda_n(A)} \leq \left(\frac{(2p + \alpha/4)\ln(m-1)}{m-1}\right)^2\right\} \geq \frac{n}{(m-1)^\alpha},$$

then

$$\mathbb{E}_{b \sim \mathcal{U}(S^{n-1})}\left[\left(\frac{\lambda_1(A) - \lambda_1^{(m)}(A,b)}{\lambda_1(A) - \lambda_n(A)}\right)^p\right]^{\frac{1}{p}} \leq .077\,\frac{(2p+\alpha/4)^2 \ln^2(m-1)}{(m-1)^2}.$$

In addition, if

$$\#\left\{\lambda_i(A) \,\middle|\, \frac{\lambda_1(A) - \lambda_i(A)}{\lambda_1(A) - \lambda_n(A)} \leq \left(\frac{(\alpha+2)\ln(m-1)}{4(m-1)}\right)^2\right\} \geq \frac{n}{(m-1)^\alpha},$$

then

$$\mathop{\mathbb{P}}_{b\sim\mathcal{U}(S^{n-1})}\left[\frac{\lambda_1(A)-\lambda_1^{(m)}(A,b)}{\lambda_1(A)-\lambda_n(A)}\le .126\,\frac{(\alpha+2)^2\ln^2(m-1)}{(m-1)^2}\right]\ge 1-O(e^{-m}).$$

Proof. We begin by bounding expected relative error. The main idea is to proceed as in the proof of Lemma 22, but take advantage of the number of eigenvalues near λ_1. For simplicity, let $\lambda_1 = 1$ and $\lambda_n = 0$. We consider eigenvalues in the ranges $[0, 2\beta-1)$ and $[2\beta-1, 1]$ separately, $1/2 < \beta < 1$, and then make use of the lower bound for the number of eigenvalues in $[\beta, 1]$. From the proof of Lemma 22, we have

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\lambda_1-\lambda_1^{(m)}}{\lambda_1-\lambda_n}\right)^p\right]^{\frac{1}{p}} \;\le\; &\min_{Q\in\mathcal{P}_{m-1}(1)}\left[\int_{[0,\infty)^n}\left(\frac{\sum_{i:\lambda_i<2\beta-1}y_iQ^2(\lambda_i)(1-\lambda_i)^{2^q}}{\sum_{i=1}^n y_iQ^2(\lambda_i)}\right)^{\frac{p}{2^q}} f_Y(y)\,dy\right]^{\frac{1}{p}}\\ &+\left[\int_{[0,\infty)^n}\left(\frac{\sum_{i:\lambda_i\ge 2\beta-1}y_iQ^2(\lambda_i)(1-\lambda_i)^{2q}}{\sum_{i=1}^n y_iQ^2(\lambda_i)}\right)^{\frac{p}{2^q}} f_Y(y)\,dy\right]^{\frac{1}{p}},\end{aligned}$$

where $q\in\mathbb{N}$, $2p<2^q\le 4p$, $f_Y(y)$ is given by (5.8), and $\beta\in(1/2,1)$. The second term is at most $2(1-\beta)$, and the integrand in the first term is bounded above by

$$\begin{aligned}\left(\frac{\sum_{i:\lambda_i<2\beta-1}y_iQ^2(\lambda_i)(1-\lambda_i)^{2^q}}{\sum_{i=1}^n y_iQ^2(\lambda_i)}\right)^{p/2^q} &\le \left(\frac{\sum_{i:\lambda_i<2\beta-1}y_iQ^2(\lambda_i)(1-\lambda_i)^{2^q}}{\sum_{i=1}^n y_iQ^2(\lambda_i)}\right)^{1/4}\\ &\le \left(\frac{\sum_{i:\lambda_i<2\beta-1}y_iQ^2(\lambda_i)(1-\lambda_i)^{2^q}}{\sum_{i:\lambda_i\ge\beta}y_iQ^2(\lambda_i)}\right)^{1/4}\\ &\le \frac{\max_{x\in[0,2\beta-1]}|Q(x)|^{1/2}\,(1-x)^{2^{q-2}}}{\min_{x\in[\beta,1]}|Q(x)|^{1/2}}\left(\frac{\sum_{i:\lambda_i<2\beta-1}y_i}{\sum_{i:\lambda_i\ge\beta}y_i}\right)^{1/4}.\end{aligned}$$

Let $\sqrt{1-\beta}=\frac{(2p+\alpha/4)\ln(m-1)}{m-1}<1/4$ (if $\beta\le 1/2$, then our estimate is a trivial one, and we are already done). By the condition of the theorem, there are at least $n/(m-1)^\alpha$ eigenvalues

in the interval $[\beta, 1]$. Therefore,

$$\begin{aligned}\int_{[0,\infty)^n} \left(\frac{\sum_{i:\lambda_i<2\beta-1} y_i}{\sum_{i:\lambda_i\geq\beta} y_i}\right)^{1/4} f_Y(y)\,dy &\leq \mathbb{E}_{\hat{Y}\sim\chi^2_n}\left[\hat{Y}^{1/4}\right]\mathbb{E}_{\tilde{Y}\sim\chi^2_{\lceil n/(m-1)^\alpha\rceil}}\left[\tilde{Y}^{-1/4}\right] \\ &= \frac{\Gamma(n/2+1/4)\Gamma(\lceil n/(m-1)^\alpha\rceil/2-1/4)}{\Gamma(n/2)\Gamma(\lceil n/(m-1)^\alpha\rceil/2)} \\ &\leq \frac{\Gamma(m(m-1)^\alpha/2+1/4)\Gamma(m/2-1/4)}{\Gamma(m(m-1)^\alpha/2)\Gamma(m/2)} \\ &\leq 1.04\,(m-1)^{\alpha/4}\end{aligned}$$

for $n \geq m(m-1)^\alpha$ and $m \geq 10$. Replacing the minimizing polynomial by $\widehat{Q}(x) = \frac{T_{m-1}\left(\frac{2x}{2\beta-1}-1\right)}{T_{m-1}\left(\frac{2}{2\beta-1}-1\right)}$, we obtain

$$\frac{\max_{x\in[0,2\beta-1]}\left|\widehat{Q}(x)\right|^{\frac{1}{2}}(1-x)^{2^{q-2}}}{\min_{x\in[\beta,1]}\left|\widehat{Q}(x)\right|^{\frac{1}{2}}} = \frac{1}{T^{1/2}_{m-1}\left(\frac{2\beta}{2\beta-1}-1\right)} \leq \frac{1}{T^{1/2}_{m-1}\left(\frac{2}{\beta}-1\right)} \leq \sqrt{2}e^{-\sqrt{1-\beta}(m-1)}.$$

Combining our estimates results in the bound

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\lambda_1-\lambda_1^{(m)}}{\lambda_1-\lambda_n}\right)^p\right]^{\frac{1}{p}} &\leq \left(1.04\sqrt{2}(m-1)^{\alpha/4}\right)^{1/p} e^{-\sqrt{1-\beta}(m-1)/p} + 2(1-\beta) \\ &= \frac{(1.04\sqrt{2})^{1/p}}{(m-1)^2} + \frac{(2p+\alpha/4)^2\ln^2(m-1)}{(m-1)^2} \\ &\leq .077\,\frac{(2p+\alpha/4)^2\ln^2(m-1)}{(m-1)^2}\end{aligned}$$

for $m \geq 10$, $p \geq 1$, and $\alpha > 0$. This completes the bound for expected relative error. We

now produce a probabilistic bound for relative error. Let $\sqrt{1-\beta} = \frac{(\alpha+2)\ln(m-1)}{4(m-1)}$. We have

$$\begin{aligned}
\frac{\lambda_1 - \lambda_1^{(m)}}{\lambda_1 - \lambda_n} &= \min_{Q \in \mathcal{P}_{m-1}(1)} \frac{\sum_{i=2}^n Y_i Q^2(\lambda_i)(1-\lambda_i)}{\sum_{i=1}^n Y_i Q^2(\lambda_i)} \\
&\leq \min_{Q \in \mathcal{P}_{m-1}(1)} \frac{\max_{x\in[0,2\beta-1]} Q^2(x)(1-x)}{\min_{x\in[\beta,1]} Q^2(x)} \frac{\sum_{i:\lambda_i<2\beta-1} Y_i}{\sum_{i:\lambda_i\geq\beta} Y_i} + 2(1-\beta) \\
&\leq T_{m-1}^{-2}\left(\frac{2}{\beta} - 1\right) \frac{\sum_{i:\lambda_i<2\beta-1} Y_i}{\sum_{i:\lambda_i\geq\beta} Y_i} + 2(1-\beta) \\
&\leq 4\exp\{-4\sqrt{1-\beta}(m-1)\} \frac{\sum_{i:\lambda_i<2\beta-1} Y_i}{\sum_{i:\lambda_i\geq\beta} Y_i} + 2(1-\beta).
\end{aligned}$$

By Proposition 11,

$$\begin{aligned}
\mathbb{P}\left[\frac{\sum_{i:\lambda_i<2\beta-1} Y_i}{\sum_{i:\lambda_i\geq\beta} Y_i} \geq 4(m-1)^\alpha\right] &\leq \mathbb{P}\left[\sum_{i:\lambda_i<2\beta-1} Y_i \geq 2n\right] + \mathbb{P}\left[\sum_{i:\lambda_i\geq\beta} Y_i \leq \frac{n}{2(m-1)^\alpha}\right] \\
&\leq (2/e)^{n/2} + \left(\sqrt{e}/2\right)^{n/2(m-1)^\alpha} = O(e^{-m}).
\end{aligned}$$

Then, with probability $1 - O(e^{-m})$,

$$\frac{\lambda_1 - \lambda_1^{(m)}}{\lambda_1 - \lambda_n} \leq 16(m-1)^\alpha e^{-4\sqrt{1-\beta}(m-1)} + 2(1-\beta) = \frac{16}{(m-1)^2} + \frac{(\alpha+2)^2\ln^2(m-1)}{8(m-1)^2}.$$

The $16/(m-1)^2$ term is dominated by the log term as m increases, and, therefore, with probability $1 - O(e^{-m})$,

$$\frac{\lambda_1 - \lambda_1^{(m)}}{\lambda_1 - \lambda_n} \leq .126\,\frac{(\alpha+2)^2\ln^2(m-1)}{(m-1)^2}.$$

This completes the proof. □

The above theorem shows that, for matrices whose distribution of eigenvalues is independent of n, we can obtain dimension-free estimates. For example, the above theorem holds for the matrix from Example 1, for $\alpha = 2$.

When a matrix has eigenvalues known to converge to a limiting distribution as the dimension increases, or a random matrix X_n exhibits suitable convergence of its empirical

spectral distribution $L_{X_n} := 1/n \sum_{i=1}^{n} \delta_{\lambda_i(X_n)}$, improved estimates can be obtained by simply estimating the corresponding integral polynomial minimization problem. However, to do so, we first require a law of large numbers for weighted sums of independent identically distributed (i.i.d.) random variables. We recall the following result.

Proposition 14. *([25]) Let $a_1, a_2, ... \in [a, b]$ and $X_1, X_2, ...$ be i.i.d. random variables, with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^2] < \infty$. Then $\frac{1}{n} \sum_{i=1}^{n} a_i X_i \to 0$ almost surely.*

We present the following theorem regarding the error of random matrices that exhibit suitable convergence.

Theorem 21. *Let $X_n \in \mathcal{S}^n$, $\Lambda(X_n) \in [a, b]$, $n = 1, 2, ...$ be a sequence of random matrices, such that L_{X_n} converges in probability to $\sigma(x)\, dx$ in $L_2([a, b])$, where $\sigma(x) \in C([a, b])$, $a, b \in$ supp(σ). Then, for all $m \in \mathbb{N}$ and $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \frac{\lambda_1(X_n) - \lambda_1^{(m)}(X_n, b)}{\lambda_1(X_n) - \lambda_n(X_n)} - \frac{b - \xi(m)}{b - a} \right| > \epsilon \right) = 0,$$

where $\xi(m)$ is the largest zero of the m^{th} degree orthogonal polynomial of $\sigma(x)$ in the interval $[a, b]$.

Proof. The main idea of the proof is to use Proposition 14 to control the behavior of Y, and the convergence of L_{X_n} to $\sigma(x)\, dx$ to show convergence to our integral minimization problem. We first write our polynomial $P \in \mathcal{P}_{m-1}$ as $P(x) = \sum_{j=0}^{m-1} c_j x^j$ and our unnormalized error as

$$\begin{aligned} \lambda_1(X_n) - \lambda_1^{(m)}(X_n, b) &= \min_{\substack{P \in \mathcal{P}_{m-1} \\ P \neq 0}} \frac{\sum_{i=2}^{n} Y_i P^2(\lambda_i)(\lambda_1 - \lambda_i)}{\sum_{i=1}^{n} Y_i P^2(\lambda_i)} \\ &= \min_{\substack{c \in \mathbb{R}^m \\ c \neq 0}} \frac{\sum_{j_1, j_2 = 0}^{m-1} c_{j_1} c_{j_2} \sum_{i=2}^{n} Y_i \lambda_i^{j_1 + j_2}(\lambda_1 - \lambda_i)}{\sum_{j_1, j_2 = 0}^{m-1} c_{j_1} c_{j_2} \sum_{i=1}^{n} Y_i \lambda_i^{j_1 + j_2}}, \end{aligned}$$

where $Y_1, ..., Y_n$ are i.i.d. chi-square random variables with one degree of freedom each. The functions x^j, $j = 0, ..., 2m - 2$, are bounded on $[a, b]$, and so, by Proposition 14, for

any $\epsilon_1, \epsilon_2 > 0$,

$$\left| \frac{1}{n} \sum_{i=2}^{n} Y_i \lambda_i^j (\lambda_1 - \lambda_i) - \frac{1}{n} \sum_{i=2}^{n} \lambda_i^j (\lambda_1 - \lambda_i) \right| < \epsilon_1, \qquad j = 0, ..., 2m-2,$$

and

$$\left| \frac{1}{n} \sum_{i=1}^{n} Y_i \lambda_i^j - \frac{1}{n} \sum_{i=1}^{n} \lambda_i^j \right| < \epsilon_2, \qquad j = 0, ..., 2m-2,$$

with probability $1-o(1)$. L_{X_n} converges in probability to $\sigma(x)\,dx$, and so, for any $\epsilon_3, \epsilon_4 > 0$,

$$\left| \frac{1}{n} \sum_{i=2}^{n} \lambda_i^j (\lambda_1 - \lambda_i) - \int_a^b x^j (b-x) \sigma(x)\,dx \right| < \epsilon_3, \qquad j = 0, ..., 2m-2,$$

and

$$\left| \frac{1}{n} \sum_{i=1}^{n} \lambda_i^j - \int_a^b x^j \sigma(x)\,dx \right| < \epsilon_4, \qquad j = 0, ..., 2m-2,$$

with probability $1 - o(1)$. This implies that

$$\lambda_1(X_n) - \lambda_1^{(m)}(X_n, b) = \min_{\substack{c \in \mathbb{R}^m \\ c \neq 0}} \frac{\sum_{j_1, j_2 = 0}^{m-1} c_{j_1} c_{j_2} \left[\int_a^b x^{j_1 + j_2} (b-x) \sigma(x)\,dx + \hat{E}(j_1, j_2) \right]}{\sum_{j_1, j_2 = 0}^{m-1} c_{j_1} c_{j_2} \left[\int_a^b x^{j_1 + j_2} \sigma(x)\,dx + E(j_1, j_2) \right]},$$

where $|\hat{E}(j_1, j_2)| < \epsilon_1 + \epsilon_3$ and $|E(j_1, j_2)| < \epsilon_2 + \epsilon_4$, $j_1, j_2 = 0, ..., m-1$, with probability $1 - o(1)$.

By the linearity of integration, the minimization problem

$$\min_{\substack{P \in \mathcal{P}_{m-1} \\ P \neq 0}} \frac{\int_a^b P^2(x) \left(\frac{b-x}{b-a} \right) \sigma(x)\,dx}{\int_a^b P^2(x) \sigma(x)\,dx},$$

when rewritten in terms of the polynomial coefficients c_i, $i = 0, ..., m-1$, corresponds to a generalized Rayleigh quotient $\frac{c^T A c}{c^T B c}$, where $A, B \in \mathcal{S}_{++}^m$ and $\lambda_{max}(A)$, $\lambda_{max}(B)$, and $\lambda_{min}(B)$ are all constants independent of n, and $c = (c_0, ..., c_{m-1})^T$. By choosing

$\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ sufficiently small,

$$\begin{aligned}\left|\frac{c^T(A+\hat{E})c}{c^T(B+E)c} - \frac{c^TAc}{c^TBc}\right| &= \left|\frac{(c^T\hat{E}c)c^TBc - (c^TEc)c^TAc}{(c^TBc + c^TEc)c^TBc}\right| \\ &\leq \frac{(\epsilon_1+\epsilon_3)m\lambda_{max}(B) + (\epsilon_2+\epsilon_4)m\lambda_{max}(A)}{(\lambda_{min}(B) - (\epsilon_2+\epsilon_4)m)\lambda_{min}(B)} \leq \epsilon\end{aligned}$$

for all $c \in \mathbb{R}^m$ with probability $1 - o(1)$. Applying Proposition 10 to the above integral minimization problem completes the proof. □

The above theorem is a powerful tool for explicitly computing the error in the Lanczos method for certain types of matrices, as the computation of the extremal eigenvalue of an $m \times m$ matrix (corresponding to the largest zero) is a nominal computation compared to one application of an n dimensional matrix. In Section 6, we will see that this convergence occurs quickly in practice. In addition, the above result provides strong evidence that the inverse quadratic dependence on m in the error estimates throughout this chapter is not so much a worst case estimate, but actually indicative of error rates in practice. For instance, if the eigenvalues of a matrix are sampled from a distribution bounded above and below by some multiple of a Jacobi weight function, then, by equation (iv) Proposition 13 and Theorem 21, it immediately follows that the error is of order m^{-2}. Of course, we note that Theorem 21 is equally applicable for estimating λ_n.

5.5 Estimates for Arbitrary Eigenvalues and Condition Number

Up to this point, we have concerned ourselves almost exclusively with the extremal eigenvalues λ_1 and λ_n of a matrix. In this section, we extend the techniques of this chapter to arbitrary eigenvalues, and also obtain bounds for the condition number of a positive definite matrix. The results of this section provide the first uniform error estimates for arbitrary eigenvalues and the first lower bounds for the relative error with respect to the condition number. Lower bounds for arbitrary eigenvalues follow relatively quickly from our previous

work. However, our proof technique for upper bounds requires some mild assumptions regarding the eigenvalue gaps of the matrix. We begin with asymptotic lower bounds for an arbitrary eigenvalue, and present the following corollary of Theorem 19.

Corollary 3.

$$\max_{A\in\mathcal{S}^n} \mathop{\mathbb{P}}_{b\sim\mathcal{U}(S^{n-1})} \left[\frac{\lambda_i(A) - \lambda_i^{(m)}(A,b)}{\lambda_1(A) - \lambda_n(A)} \geq 1 - o(1) \right] \geq 1 - o(1/n)$$

for $m = o(\ln n)$,

$$\max_{A\in\mathcal{S}^n} \mathop{\mathbb{P}}_{b\sim\mathcal{U}(S^{n-1})} \left[\frac{\lambda_i(A) - \lambda_i^{(m)}(A,b)}{\lambda_1(A) - \lambda_n(A)} \geq \frac{.015\ln^2 n}{m^2 \ln^2 \ln n} \right] \geq 1 - o(1/n)$$

for $m = \Theta(\ln n)$, *and*

$$\max_{A\in\mathcal{S}^n} \mathop{\mathbb{P}}_{b\sim\mathcal{U}(S^{n-1})} \left[\frac{\lambda_i(A) - \lambda_i^{(m)}(A,b)}{\lambda_1(A) - \lambda_n(A)} \geq \frac{1.08}{m^2} \right] \geq 1 - o(1/n)$$

for $m = o\left(n^{1/2}\ln^{-1/2} n\right)$ *and* $\omega(1)$.

Proof. Let $\varphi_1, ..., \varphi_n$ be orthonormal eigenvectors corresponding to $\lambda_1(A) \geq ... \geq \lambda_n(A)$, and $\hat{b} = b - \sum_{j=1}^{i-1}(\varphi_j^T b)b$. By the inequalities

$$\lambda_i^{(m)}(A,b) \leq \max_{\substack{x\in\mathcal{K}_m(A,b)\setminus 0 \\ x^T\varphi_j = 0, \\ j=1,...,i-1}} \frac{x^TAx}{x^Tx} \leq \max_{x\in\mathcal{K}_m\left(A,\hat{b}\right)\setminus 0} \frac{x^TAx}{x^Tx} = \lambda_1^{(m)}\left(A,\hat{b}\right),$$

the relative error

$$\frac{\lambda_i(A) - \lambda_i^{(m)}(A,b)}{\lambda_1(A) - \lambda_n(A)} \geq \frac{\lambda_i(A) - \lambda_1^{(m)}(A,\hat{b})}{\lambda_1(A) - \lambda_n(A)}.$$

The right-hand side corresponds to an extremal eigenvalue problem of dimension $n - i + 1$. Setting the largest eigenvalue of A to have multiplicity i, and defining the eigenvalues $\lambda_i, ..., \lambda_n$ based on the eigenvalues (corresponding to dimension $n-i+1$) used in the proofs of Lemmas 20 and 21 completes the proof. □

Next, we provide an upper bound for the relative error in approximating λ_i under the assumption of non-zero gaps between eigenvalues $\lambda_1, ..., \lambda_i$.

Theorem 22. *Let $n \geq 100$, $m \geq 9 + i$, $p \geq 1$, and $A \in \mathcal{S}^n$. Then*

$$\mathbb{E}_{b\sim\mathcal{U}(S^{n-1})}\left[\left(\frac{\lambda_i(A) - \lambda_i^{(m)}(A,b)}{\lambda_i(A) - \lambda_n(A)}\right)^p\right]^{1/p} \leq .068\,\frac{\ln^2\left(\delta^{-2(i-1)}n(m-i)^{8p}\right)}{(m-i)^2}$$

and

$$\mathbb{P}_{b\sim\mathcal{U}(S^{n-1})}\left[\frac{\lambda_i(A) - \lambda_i^{(m)}(A,b)}{\lambda_i(A) - \lambda_n(A)} \leq .571\,\frac{\ln^2\left(\delta^{-2(i-1)/3}n(m-i)^{2/3}\right)}{(m-i)^2}\right] \geq 1 - o(1/n),$$

where

$$\delta = \frac{1}{2}\min_{k=2,...,i}\frac{\lambda_{k-1}(A) - \lambda_k(A)}{\lambda_1(A) - \lambda_n(A)}.$$

Proof. As in previous cases, it suffices to prove the theorem for matrices A with $\lambda_i(A) = 1$ and $\lambda_n(A) = 0$ (if $\lambda_i = \lambda_n$, we are done). Similar to the polynomial representation of $\lambda_1^{(m)}(A,b)$, the Ritz value $\lambda_i^{(m)}(A,b)$ corresponds to finding the polynomial in $\mathcal{P}_{m-1}$ with zeros $\lambda_k^{(m)}$, $k = 1, ..., i-1$, that maximizes the corresponding Rayleigh quotient. For the sake of brevity, let $\phi_i(x) = \prod_{k=1}^{i-1}(\lambda_k^{(m)} - x)^2$. Then $\lambda_i^{(m)}(A,b)$ can be written as

$$\lambda_i^{(m)}(A,b) = \max_{P\in\mathcal{P}_{m-i}}\frac{\sum_{j=1}^n b_j^2P^2(\lambda_j)\phi_i(\lambda_j)\lambda_j}{\sum_{j=1}^n b_j^2P^2(\lambda_j)\phi_i(\lambda_j)},$$

and, therefore, the error is bounded above by

$$\lambda_i(A) - \lambda_i^{(m)}(A,b) \leq \max_{P\in\mathcal{P}_{m-i}}\frac{\sum_{j=i+1}^n b_j^2P^2(\lambda_j)\phi_i(\lambda_j)(1-\lambda_j)}{\sum_{j=1}^n b_j^2P^2(\lambda_j)\phi_i(\lambda_j)}.$$

The main idea of the proof is very similar to that of Lemma 22, paired with a pigeonhole principle. The intervals $[\lambda_j(A) - \delta\lambda_1, \lambda_j(A) + \delta\lambda_1]$, $j = 1, ..., i$, are disjoint, and so there exists some index j^* such that the corresponding interval does not contain any of the Ritz values $\lambda_k^{(m)}(A,b)$, $k = 1, ..., i-1$. We begin by bounding expected relative error. As in the proof of Lemma 22, by integrating over chi-square random variables and using

Cauchy-Schwarz, we have

$$\begin{aligned}\mathbb{E}\left[\left(\lambda_i-\lambda_i^{(m)}\right)^p\right]^{\frac{1}{p}} &\le \min_{P\in\mathcal{P}_{m-i}}\left[\int_{[0,\infty)^n}\left(\frac{\sum_{j=i+1}^n y_j P^2(\lambda_j)\phi_i(\lambda_j)(1-\lambda_j)^{2^q}}{\sum_{j=1}^n y_j P^2(\lambda_j)\phi_i(\lambda_j)}\right)^{\frac{p}{2^q}} f_Y(y)\,dy\right]^{\frac{1}{p}} \\ &\le \min_{P\in\mathcal{P}_{m-i}}\left[\int_{[0,\infty)^n}\left(\frac{\sum_{j:\lambda_j<\beta} y_j P^2(\lambda_j)\phi_i(\lambda_j)(1-\lambda_j)^{2^q}}{\sum_{j=1}^n y_j P^2(\lambda_j)\phi_i(\lambda_j)}\right)^{\frac{p}{2^q}} f_Y(y)\,dy\right]^{\frac{1}{p}} \\ &+ \left[\int_{[0,\infty)^n}\left(\frac{\sum_{j:\lambda_j\in[\beta,1]} y_j P^2(\lambda_j)\phi_i(\lambda_j)(1-\lambda_j)^{2^q}}{\sum_{j=1}^n y_j P^2(\lambda_j)\phi_i(\lambda_j)}\right)^{\frac{p}{2^q}} f_Y(y)\,dy\right]^{\frac{1}{p}}\end{aligned}$$

for $q\in\mathbb{N}$, $2p<2^q\le 4p$, and any $\beta\in(0,1)$. The second term on the right-hand side is bounded above by $1-\beta$ (the integrand is at most $\max_{\lambda\in[\beta,1]}(1-\lambda)=1-\beta$), and the integrand in the first term is bounded above by

$$\begin{aligned}\left(\frac{\sum_{j:\lambda_j<\beta} y_j P^2(\lambda_j)\phi_i(\lambda_j)(1-\lambda_j)^{2^q}}{\sum_{j=1}^n y_j P^2(\lambda_j)\phi_i(\lambda_j)}\right)^{\frac{p}{2^q}} &\le \left(\frac{\sum_{j:\lambda_j<\beta} y_j P^2(\lambda_j)\phi_i(\lambda_j)(1-\lambda_j)^{2^q}}{\sum_{j=1}^n y_j P^2(\lambda_j)\phi_i(\lambda_j)}\right)^{\frac{1}{4}} \\ &\le \max_{x\in[0,\beta]}\frac{|P(x)|^{\frac{1}{2}}\phi_i^{\frac{1}{4}}(x)(1-x)^{2^{q-2}}}{|P(\lambda_{j^*})|^{\frac{1}{2}}\phi_i^{\frac{1}{4}}(\lambda_{j^*})}\left(\frac{\sum_{i:\lambda_i<\beta} y_i}{y_{j^*}}\right)^{\frac{1}{4}}.\end{aligned}$$

By replacing the minimizing polynomial in $\mathcal{P}_{m-i}$ by $T_{m-i}\left(\frac{2}{\beta}x-1\right)$, the maximum is achieved at $x=0$, and, by the monotonicity of T_{m-i} on $[1,\infty)$,

$$T_{m-i}\left(\frac{2}{\beta}\lambda_{j^*}-1\right)\ge T_{m-i}\left(\frac{2}{\beta}-1\right)\ge\frac{1}{2}\exp\left\{2\sqrt{1-\beta}\,(m-i)\right\}.$$

In addition,

$$\frac{\phi_i^{1/4}(0)}{\phi_i^{1/4}(\lambda_{j^*})}=\prod_{k=1}^{i-1}\left|\frac{\lambda_k^{(m)}}{\lambda_k^{(m)}-\lambda_{j^*}}\right|^{1/2}\le\delta^{-(i-1)/2}.$$

We can now bound the p-norm by

$$\mathbb{E}\left[\left(\lambda_i - \lambda_i^{(m)}\right)^p\right]^{\frac{1}{p}} \leq \left[\frac{2^{1/4}\Gamma(1/4)}{\Gamma(1/2)}\frac{n^{1/4}}{\delta^{(i-1)/2}}\right]^{1/p} e^{-\gamma(m-i)/p} + \gamma^2,$$

where $\gamma = \sqrt{1-\beta}$. Setting $\gamma = \frac{p}{m-i}\ln\left(\delta^{-(i-1)/2p}n^{1/4p}(m-i)^2\right)$ (assuming $\gamma < 1$, otherwise our bound is already greater than one, and trivially holds), we obtain

$$\begin{aligned}\mathbb{E}\left[\left(\lambda_i - \lambda_i^{(m)}\right)^p\right]^{\frac{1}{p}} &\leq \frac{\left(\frac{2^{1/4}\Gamma(1/4)}{\Gamma(1/2)}\right)^{1/p} + \frac{1}{16}\ln^2\left(\delta^{-2(i-1)}n(m-i)^{8p}\right)}{(m-i)^2} \\ &\leq .068\,\frac{\ln^2\left(\delta^{-2(i-1)}n(m-i)^{8p}\right)}{(m-i)^2},\end{aligned}$$

for $m \geq 9+i$, $n \geq 100$. This completes the proof of the expected error estimate. We now focus on the probabilistic estimate. We have

$$\begin{aligned}\lambda_i(A) - \lambda_i^{(m)}(A,b) &\leq \min_{P\in\mathcal{P}_{m-i}} \frac{\sum_{j=i+1}^n Y_j P^2(\lambda_j)\phi_i(\lambda_j)(1-\lambda_j)}{\sum_{j=1}^n Y_j P^2(\lambda_j)\phi_i(\lambda_j)} \\ &\leq \min_{P\in\mathcal{P}_{m-i}}\max_{x\in[0,\beta]} \frac{P^2(x)\phi_i(x)(1-x)}{P^2(\lambda_{j^*})\phi_i(\lambda_{j^*})}\frac{\sum_{j:\lambda_j<\beta} Y_j}{Y_{j^*}} + (1-\beta) \\ &\leq \delta^{-2(i-1)}\,T_{m-i}^{-2}\left(\frac{2}{\beta}-1\right)\frac{\sum_{j:\lambda_j<\beta} Y_j}{Y_{j^*}} + (1-\beta) \\ &\leq 4\,\delta^{-2(i-1)}\exp\{-4\sqrt{1-\beta}(m-i)\}\frac{\sum_{j:\lambda_j<\beta} Y_j}{Y_{j^*}} + (1-\beta).\end{aligned}$$

By Proposition 11,

$$\begin{aligned}\mathbb{P}\left[\frac{\sum_{j:\lambda_j<\beta} Y_j}{Y_{j^*}} \geq n^{3.02}\right] &\leq \mathbb{P}\left[Y_{j^*} \leq n^{-2.01}\right] + \mathbb{P}\left[\sum_{j\neq j^*} Y_j \geq n^{1.01}\right] \\ &\leq \left(e/n^{2.01}\right)^{1/2} + \left(n^{.01}e^{1-n^{.01}}\right)^{(n-1)/2} = o(1/n).\end{aligned}$$

Let $\sqrt{1-\beta} = \frac{\ln\left(\delta^{-2(i-1)}n^{3.02}(m-i)^2\right)}{4(m-i)}$. Then, with probability $1-o(1/n)$,

$$\lambda_i(A) - \lambda_i^{(m)}(A) \leq \frac{4}{(m-i)^2} + \frac{\ln^2\left(\delta^{-2(i-1)}n^{3.02}(m-i)^2\right)}{16(m-i)^2}.$$

The $4/(m-i)^2$ term is dominated by the log term as n increases, and, therefore, with probability $1 - o(1/n)$,

$$\lambda_i(A) - \lambda_i^{(m)}(A) \leq .571 \, \frac{\ln^2\left(\delta^{-2(i-1)/3} n (m-i)^{2/3}\right)}{(m-i)^2}.$$

This completes the proof. □

For typical matrices with no repeated eigenvalues, δ is usually a very low degree polynomial in n, and, for i constant, the estimates for λ_i are not much worse than that of λ_1. In addition, it seems likely that, given any matrix A, a small random perturbation of A before the application of the Lanczos method will satisfy the condition of the theorem with high probability, and change the eigenvalues by a negligible amount. Of course, the bounds from Theorem 22 for maximal eigenvalues λ_i also apply to minimal eigenvalues λ_{n-i}.

Next, we make use of Theorem 19 to produce error estimates for the condition number of a symmetric positive definite matrix. Let $\kappa(A) = \lambda_1(A)/\lambda_n(A)$ denote the condition number of a matrix $A \in \mathcal{S}_{++}^n$ and $\kappa^{(m)}(A,b) = \lambda_1^{(m)}(A,b)/\lambda_m^{(m)}(A,b)$ denote the condition number of the tridiagonal matrix $T_m \in \mathcal{S}_{++}^m$ resulting from m iterations of the Lanczos method applied to a matrix $A \in \mathcal{S}_{++}^n$ with initial vector b. Their difference can be written as

$$\begin{aligned}
\kappa(A) - \kappa^{(m)}(A,b) &= \frac{\lambda_1}{\lambda_n} - \frac{\lambda_1^{(m)}}{\lambda_m^{(m)}} = \frac{\lambda_1 \lambda_m^{(m)} - \lambda_n \lambda_1^{(m)}}{\lambda_n \lambda_m^{(m)}} \\
&= \frac{\lambda_m^{(m)}\left(\lambda_1 - \lambda_1^{(m)}\right) + \lambda_1^{(m)}\left(\lambda_m^{(m)} - \lambda_n\right)}{\lambda_n \lambda_m^{(m)}} \\
&= (\kappa(A) - 1)\left[\frac{\lambda_1 - \lambda_1^{(m)}}{\lambda_1 - \lambda_n} + \kappa^{(m)}(A,b)\frac{\lambda_m^{(m)} - \lambda_n}{\lambda_1 - \lambda_n}\right],
\end{aligned}$$

which leads to the bounds

$$\frac{\kappa(A) - \kappa^{(m)}(A,b)}{\kappa(A)} \leq \frac{\lambda_1 - \lambda_1^{(m)}}{\lambda_1 - \lambda_n} + \kappa(A)\frac{\lambda_m^{(m)} - \lambda_n}{\lambda_1 - \lambda_n}$$

and

$$\frac{\kappa(A) - \kappa^{(m)}(A,b)}{\kappa^{(m)}(A,b)} \geq (\kappa(A) - 1)\frac{\lambda_m^{(m)} - \lambda_n}{\lambda_1 - \lambda_n}.$$

Using Theorem 19 and Minkowski's inequality, we have the following corollary.

Corollary 4.

$$\sup_{\substack{A \in \mathcal{S}_{++}^n \\ \kappa(A) = \bar{\kappa}}} \mathbb{P}_{b \sim \mathcal{U}(S^{n-1})} \left[\frac{\kappa(A) - \kappa^{(m)}(A,b)}{\kappa^{(m)}(A,b)} \geq (1 - o(1))(\bar{\kappa} - 1) \right] \geq 1 - o(1/n)$$

for $m = o(\ln n)$*,*

$$\sup_{\substack{A \in \mathcal{S}_{++}^n \\ \kappa(A) = \bar{\kappa}}} \mathbb{P}_{b \sim \mathcal{U}(S^{n-1})} \left[\frac{\kappa(A) - \kappa^{(m)}(A,b)}{\kappa^{(m)}(A,b)} \geq .015\, \frac{(\bar{\kappa} - 1)\ln^2 n}{m^2 \ln^2 \ln n} \right] \geq 1 - o(1/n)$$

for $m = \Theta(\ln n)$*,*

$$\sup_{\substack{A \in \mathcal{S}_{++}^n \\ \kappa(A) = \bar{\kappa}}} \mathbb{P}_{b \sim \mathcal{U}(S^{n-1})} \left[\frac{\kappa(A) - \kappa^{(m)}(A,b)}{\kappa^{(m)}(A,b)} \geq 1.08\, \frac{\bar{\kappa} - 1}{m^2} \right] \geq 1 - o(1/n)$$

for $m = o\left(n^{1/2} \ln^{-1/2} n\right)$ *and* $\omega(1)$*, and*

$$\sup_{\substack{A \in \mathcal{S}_{++}^n \\ \kappa(A) = \bar{\kappa}}} \mathbb{E}_{b \sim \mathcal{U}(S^{n-1})} \left[\left(\frac{\kappa(A) - \kappa^{(m)}(A,b)}{\kappa(A)} \right)^p \right]^{1/p} \leq .068\, (\bar{\kappa} + 1)\, \frac{\ln^2 \left(n(m-1)^{8p}\right)}{(m-1)^2}$$

for $n \geq 100$*,* $m \geq 10$*,* $p \geq 1$*.*

As previously mentioned, it is not possible to produce uniform bounds for the relative error of $\kappa^{(m)}(A,b)$, and so some dependence on $\kappa(A)$ is necessary.

5.6 Experimental Results

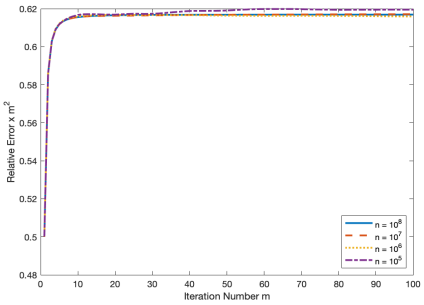
In this section, we present a number of experimental results that illustrate the error of the Lanczos method in practice. We consider:

- eigenvalues of 1D Laplacian with Dirichlet boundary conditions $\Lambda_{lap} = \{2+2\cos(i\pi/(n+1))\}_{i=1}^{n}$,
- a uniform partition of $[0,1]$, $\Lambda_{unif} = \{(n-i)/(n-1)\}_{i=1}^{n}$,
- eigenvalues from the semi-circle distribution, $\Lambda_{semi} = \{\lambda_i\}_{i=1}^{n}$, where $1/2 + \left(\lambda_i\sqrt{1-\lambda_i^2} + \arcsin\lambda_i\right)/\pi = (n-i)/(n-1)$, $i = 1, ..., n$,
- eigenvalues corresponding to Lemma 21, $\Lambda_{log} = \{1 - [(n+1-i)/n]^{\frac{\ln\ln n}{\ln n}}\}_{i=1}^{n}$.

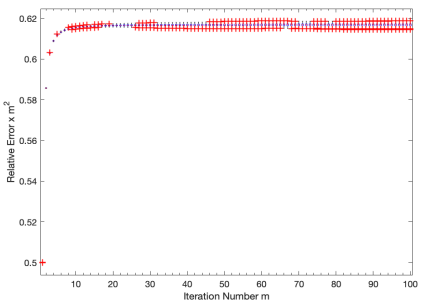
For each one of these spectra, we perform tests for dimensions $n = 10^i$, $i = 5, 6, 7, 8$. For each dimension, we generate 100 random vectors $b \sim \mathcal{U}(S^{n-1})$, and perform $m = 100$ iterations of the Lanczos method on each vector. In Figure 5-1, we report the results of these experiments. In particular, for each spectrum, we plot the empirical average relative error $(\lambda_1 - \lambda_1^{(m)})/(\lambda_1 - \lambda_n)$ for each dimension as m varies from 1 to 100. In addition, for $n = 10^8$, we present a box plot for each spectrum, illustrating the variability in relative error that each spectrum possesses.

In Figure 5-1[2], the plots of Λ_{lap}, Λ_{unif}, Λ_{semi} all illustrate an empirical average error estimate that scales with m^{-2} and has no dependence on dimension n (for n sufficiently large). This is consistent with the theoretical results of the chapter, most notably Theorem 21, as all of these spectra exhibit suitable convergence of their empirical spectral distribution, and the related integral minimization problems all have solutions of order m^{-2}. In addition, the box plots corresponding to $n = 10^8$ illustrate that the relative error for a given iteration number has a relatively small variance. For instance, all extreme values remain within a range of length less than $.01\,m^{-2}$ for Λ_{lap}, $.1\,m^{-2}$ for Λ_{unif}, and $.4\,m^{-2}$ for Λ_{semi}. This is also consistent with the convergence of Theorem 21. The empirical spectral distribution of all three spectra converge to shifted and scaled versions of Jacobi weight functions, namely, Λ_{lap} corresponds to $\omega^{-1/2,-1/2}(x)$, Λ_{unif} to $\omega^{0,0}(x)$, and Λ_{semi} to $\omega^{1/2,1/2}(x)$. The limiting value of $m^2(1-\xi(m))/2$ for each of these three cases is given by $j_{1,\alpha}^2/4$, where $j_{1,\alpha}$ is

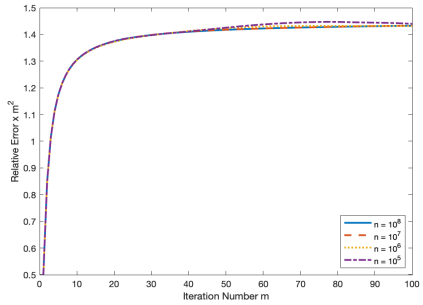
[2] In Figure 5-1, the box plots are labelled as follows. For a given m, the 25^{th} and 75^{th} percentile of the values, denoted by q_1 and q_3, are the bottom and top of the corresponding box, and the red line in the box is the median. The whiskers extend to the most extreme points in the interval $[q_1 - 1.5(q_3 - q_1), q_3 + 1.5(q_3 - q_1)]$, and outliers not in this interval correspond to the '+' symbol.



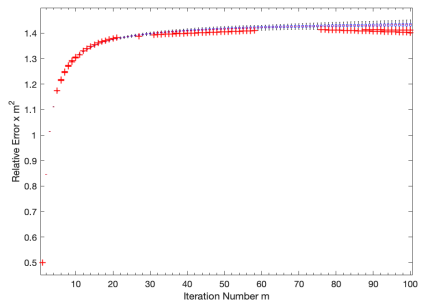
(a) Plot, Λ_{lap}



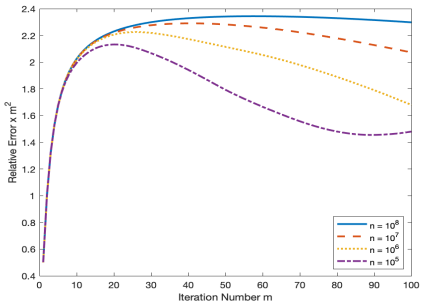
(b) Box Plot, Λ_{lap}



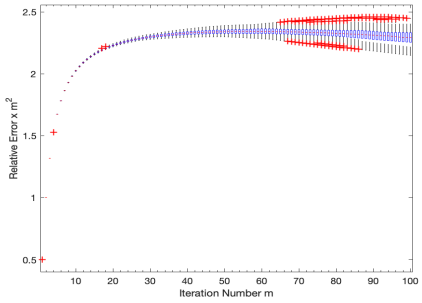
(c) Plot, Λ_{unif}



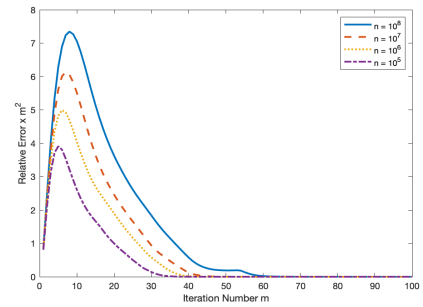
(d) Box Plot, Λ_{unif}



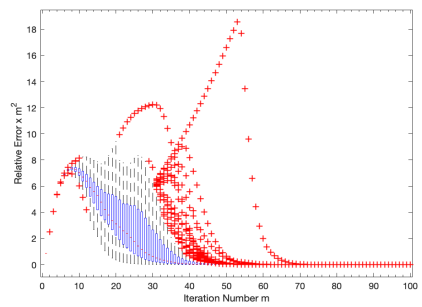
(e) Plot, Λ_{semi}



(f) Box Plot, Λ_{semi}



(g) Plot, Λ_{log}



(h) Box Plot, Λ_{log}

Figure 5-1: Plot and box plot of relative error times m^2 vs iteration number m for Λ_{lap}, Λ_{unif}, Λ_{semi}, and Λ_{log}. The plot contains curves for each dimension n tested. Each curve represents the empirical average relative error for each value of m, averaged over 100 random initializations. The box plot illustrates the variability of relative error for $n = 10^8$.

the first positive zero of the Bessel function $J_\alpha(x)$, namely, the scaled error converges to $\pi^2/16 \approx .617$ for Λ_{lap}, ≈ 1.45 for Λ_{unif}, and $\pi^2/4 \approx 2.47$ for Λ_{semi}. Two additional properties suggested by Figure 5-1 are that the variance and the dimension n required to observe asymptotic behavior both appear to increase with α. This is consistent with the theoretical results, as Lemma 21 (and Λ_{log}) results from considering $\omega^{\alpha,0}(x)$ with α as a function of n.

The plot of relative error for Λ_{log} illustrates that relative error does indeed depend on n in a tangible way, and is increasing with n in what appears to be a logarithmic fashion. For instance, when looking at the average relative error scaled by m^2, the maximum over all iteration numbers m appears to increase somewhat logarithmically (≈ 4 for $n = 10^5$, ≈ 5 for $n = 10^6$, ≈ 6 for $n = 10^7$, and ≈ 7 for $n = 10^8$). In addition, the boxplot for $n = 10^8$ illustrates that this spectrum exhibits a large degree of variability and is susceptible to extreme outliers. These numerical results support the theoretical lower bounds of Section 3, and illustrate that the asymptotic theoretical lower bounds which depend on n do occur in practical computations.

Bibliography

[1] Raja Hafiz Affandi, Emily B. Fox, Ryan P. Adams, and Benjamin Taskar. Learning the parameters of determinantal point process kernels. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1224–1232, 2014.

[2] Edoardo Amaldi, Claudio Iuliano, and Romeo Rizzi. Efficient deterministic algorithms for finding a minimum cycle basis in undirected graphs. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 397–410. Springer, 2010.

[3] Nima Anari and Thuy-Duong Vuong. Simple and near-optimal map inference for nonsymmetric dpps. *arXiv preprint arXiv:2102.05347*, 2021.

[4] Vishal Arul. personal communication.

[5] Mihai Badoiu, Kedar Dhamdhere, Anupam Gupta, Yuri Rabinovich, Harald Räcke, Ramamoorthi Ravi, and Anastasios Sidiropoulos. Approximation algorithms for low-distortion embeddings into low-dimensional spaces. In *SODA*, volume 5, pages 119–128. Citeseer, 2005.

[6] Nematollah Kayhan Batmanghelich, Gerald Quon, Alex Kulesza, Manolis Kellis, Polina Golland, and Luke Bornn. Diversifying sparsity using variational determinantal point processes. *ArXiv: 1411.6307*, 2014.

[7] Giuseppe Di Battista, Peter Eades, Roberto Tamassia, and Ioannis G Tollis. *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR, 1998.

[8] Julius Borcea, Petter Brändén, and Thomas Liggett. Negative dependence and the geometry of polynomials. *Journal of the American Mathematical Society*, 22(2):521–567, 2009.

[9] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[10] Alexei Borodin. Determinantal point processes. *arXiv preprint arXiv:0911.1153*, 2009.

[11] Alexei Borodin and Eric M Rains. Eynard–mehta theorem, schur process, and their pfaffian analogs. *Journal of statistical physics*, 121(3):291–317, 2005.

[12] Jane Breen, Alex Riasanovsky, Michael Tait, and John Urschel. Maximum spread of graphs and bipartite graphs. *In Preparation*.

[13] Andries E Brouwer and Willem H Haemers. *Spectra of graphs*. Springer Science & Business Media, 2011.

[14] Victor-Emmanuel Brunel. Learning signed determinantal point processes through the principal minor assignment problem. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[15] Victor-Emmanuel Brunel, Michel Goemans, and John Urschel. Recovering a magnitude-symmetric matrix from its principal minors. *In preparation*.

[16] Victor-Emmanuel Brunel, Ankur Moitra, Philippe Rigollet, and John Urschel. Maximum likelihood estimation of determinantal point processes. *arXiv:1701.06501*, 2017.

[17] Victor-Emmanuel Brunel, Ankur Moitra, Philippe Rigollet, and John Urschel. Rates of estimation for determinantal point processes. In *Conference on Learning Theory*, pages 343–345. PMLR, 2017.

[18] David Carlson. What are Schur complements, anyway? *Linear Algebra and its Applications*, 74:257–275, 1986.

[19] Lawrence Cayton and Sanjoy Dasgupta. Robust euclidean embedding. In *Proceedings of the 23rd international conference on machine learning*, pages 169–176, 2006.

[20] Norishige Chiba, Takao Nishizeki, Shigenobu Abe, and Takao Ozawa. A linear algorithm for embedding planar graphs using PQ-trees. *Journal of computer and system sciences*, 30(1):54–76, 1985.

[21] David M. Chickering, Dan Geiger, and David Heckerman. On finding a cycle basis with a shortest maximal cycle. *Information Processing Letters*, 54(1):55 – 58, 1995.

[22] Ali Çivril and Malik Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science*, 410(47-49):4801–4811, 2009.

[23] Ed S Coakley and Vladimir Rokhlin. A fast divide-and-conquer algorithm for computing the spectra of real symmetric tridiagonal matrices. *Applied and Computational Harmonic Analysis*, 34(3):379–414, 2013.

[24] Martin Costabel. Boundary integral operators on Lipschitz domains: elementary results. *SIAM J. Math. Anal.*, 19(3):613–626, 1988.

[25] Jack Cuzick. A strong law for weighted sums of iid random variables. *Journal of Theoretical Probability*, 8(3):625–641, 1995.

[26] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[27] Timothy A Davis and Yifan Hu. The university of florida sparse matrix collection. *ACM Transactions on Mathematical Software (TOMS)*, 38(1):1–25, 2011.

[28] Jan De Leeuw. Convergence of the majorization method for multidimensional scaling. *Journal of classification*, 5(2):163–180, 1988.

[29] Jan De Leeuw, In JR Barra, F Brodeau, G Romier, B Van Cutsem, et al. Applications of convex analysis to multidimensional scaling. In *Recent Developments in Statistics*. Citeseer, 1977.

[30] Gianna M Del Corso and Giovanni Manzini. On the randomized error of polynomial methods for eigenvector and eigenvalue estimates. *Journal of Complexity*, 13(4):419–456, 1997.

[31] Erik Demaine, Adam Hesterberg, Frederic Koehler, Jayson Lynch, and John Urschel. Multidimensional scaling: Approximation and complexity. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2568–2578. PMLR, 18–24 Jul 2021.

[32] Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 329–338. IEEE, 2010.

[33] Emeric Deutsch. On the spread of matrices and polynomials. *Linear Algebra and Its Applications*, 22:49–55, 1978.

[34] Michel Marie Deza and Monique Laurent. *Geometry of cuts and metrics*, volume 15. Springer, 2009.

[35] Peter Sheridan Dodds, Roby Muhamad, and Duncan J Watts. An experimental study of search in global social networks. *science*, 301(5634):827–829, 2003.

[36] Petros Drineas, Ilse CF Ipsen, Eugenia-Maria Kontopoulou, and Malik Magdon-Ismail. Structural convergence results for approximation of dominant subspaces from block Krylov spaces. *SIAM Journal on Matrix Analysis and Applications*, 39(2):567–586, 2018.

[37] Kathy Driver, Kerstin Jordaan, and Norbert Mbuyi. Interlacing of the zeros of Jacobi polynomials with different parameters. *Numerical Algorithms*, 49(1-4):143, 2008.

[38] Peter Eades. A heuristic for graph drawing. *Congressus numerantium*, 42:149–160, 1984.

[39] John Ellson, Emden Gansner, Lefteris Koutsofios, Stephen C North, and Gordon Woodhull. Graphviz—open source graph drawing tools. In *International Symposium on Graph Drawing*, pages 483–484. Springer, 2001.

[40] Miroslav Fiedler. Remarks on the Schur complement. *Linear Algebra Appl.*, 39:189–195, 1981.

[41] Klaus-Jürgen Förster and Knut Petras. On estimates for the weights in Gaussian quadrature in the ultraspherical case. *Mathematics of computation*, 55(191):243–264, 1990.

[42] Emilio Gagliardo. Caratterizzazioni delle tracce sulla frontiera relative ad alcune classi di funzioni in n variabili. *Rend. Sem. Mat. Univ. Padova*, 27:284–305, 1957.

[43] Emden R Gansner, Yehuda Koren, and Stephen North. Graph drawing by stress majorization. In *International Symposium on Graph Drawing*, pages 239–250. Springer, 2004.

[44] Mike Gartrell, Victor-Emmanuel Brunel, Elvis Dohmatob, and Syrine Krichene. Learning nonsymmetric determinantal point processes. *arXiv preprint arXiv:1905.12962*, 2019.

[45] Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. Low-rank factorization of determinantal point processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[46] Jennifer A Gillenwater, Alex Kulesza, Emily Fox, and Ben Taskar. Expectation-maximization for learning determinantal point processes. In *NIPS*, 2014.

[47] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2012.

[48] David A Gregory, Daniel Hershkowitz, and Stephen J Kirkland. The spread of the spectrum of a graph. *Linear Algebra and its Applications*, 332:23–35, 2001.

[49] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[50] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1988. Reprint of the 1952 edition.

[51] Joseph Douglas Horton. A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM Journal on Computing*, 16(2):358–366, 1987.

[52] Xiaozhe Hu, John C Urschel, and Ludmil T Zikatanov. On the approximation of laplacian eigenvalues in graph disaggregation. *Linear and Multilinear Algebra*, 65(9):1805–1822, 2017.

[53] Charles R Johnson, Ravinder Kumar, and Henry Wolkowicz. Lower bounds for the spread of a matrix. *Linear Algebra and Its Applications*, 71:161–173, 1985.

[54] Charles R Johnson and Michael J Tsatsomeros. Convex sets of nonsingular and P-matrices. *Linear and Multilinear Algebra*, 38(3):233–239, 1995.

[55] Tomihisa Kamada, Satoru Kawai, et al. An algorithm for drawing general undirected graphs. *Information processing letters*, 31(1):7–15, 1989.

[56] Shmuel Kaniel. Estimates for some computational techniques in linear algebra. *Mathematics of Computation*, 20(95):369–378, 1966.

[57] Michael Kaufmann and Dorothea Wagner. *Drawing graphs: methods and models*, volume 2025. Springer, 2003.

[58] Sanjeev Khanna, Madhu Sudan, Luca Trevisan, and David P Williamson. The approximability of constraint satisfaction problems. *SIAM Journal on Computing*, 30(6):1863–1920, 2001.

[59] Kevin Knudson and Evelyn Lamb. My favorite theorem, episode 23 - ingrid daubechies.

[60] Y. Koren. Drawing graphs by eigenvectors: theory and practice. *Comput. Math. Appl.*, 49(11-12):1867–1888, 2005.

[61] Yehuda Koren. On spectral graph drawing. In *Computing and combinatorics*, volume 2697 of *Lecture Notes in Comput. Sci.*, pages 496–508. Springer, Berlin, 2003.

[62] Yehuda Koren, Liran Carmel, and David Harel. Drawing huge graphs by algebraic multigrid optimization. *Multiscale Model. Simul.*, 1(4):645–673 (electronic), 2003.

[63] B. Korte and J. Vygen. *Combinatorial Optimization: Theory and Algorithms*. Springer, 2011.

[64] Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[65] Joseph B Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.

[66] Joseph B Kruskal. *Multidimensional scaling*. Number 11. Sage, 1978.

[67] Jacek Kuczyński and Henryk Woźniakowski. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM journal on matrix analysis and applications*, 13(4):1094–1122, 1992.

[68] Jacek Kuczyński and Henryk Woźniakowski. Probabilistic bounds on the extremal eigenvalues and condition number by the Lanczos algorithm. *SIAM Journal on Matrix Analysis and Applications*, 15(2):672–691, 1994.

[69] A. Kulesza. *Learning with determinantal point processes*. PhD thesis, University of Pennsylvania, 2012.

[70] Alex Kulesza and Ben Taskar. k-DPPs: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 1193–1200, 2011.

[71] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

[72] Alex Kulesza and Ben Taskar. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., Hanover, MA, USA, 2012.

[73] Donghoon Lee, Geonho Cha, Ming-Hsuan Yang, and Songhwai Oh. Individualness and determinantal point processes for pedestrian detection. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 330–346, 2016.

[74] Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast dpp sampling for nystrom with application to kernel methods. *International Conference on Machine Learning (ICML)*, 2016.

[75] Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast sampling for strongly rayleigh measures with application to determinantal point processes. *1607.03559*, 2016.

[76] Hui Lin and Jeff A. Bilmes. Learning mixtures of submodular shells with application to document summarization. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, August 14-18, 2012*, pages 479–490, 2012.

[77] J.-L. Lions and E. Magenes. *Non-homogeneous boundary value problems and applications. Vol. I.* Springer-Verlag, New York-Heidelberg, 1972. Translated from the French by P. Kenneth, Die Grundlehren der mathematischen Wissenschaften, Band 181.

[78] Chia-an Liu and Chih-wen Weng. Spectral radius of bipartite graphs. *Linear Algebra and its Applications*, 474:30–43, 2015.

[79] Raphael Loewy. Principal minors and diagonal similarity of matrices. *Linear algebra and its applications*, 78:23–64, 1986.

[80] Zelda Mariet and Suvrit Sra. Fixed-point algorithms for learning determinantal point processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2389–2397, 2015.

[81] Claire Mathieu and Warren Schudy. Yet another algorithm for dense max cut: go greedy. In *SODA*, pages 176–182, 2008.

[82] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[83] Leon Mirsky. The spread of a matrix. *Mathematika*, 3(2):127–130, 1956.

[84] Cameron Musco and Christopher Musco. Randomized block Krylov methods for stronger and faster approximate singular value decomposition. In *Advances in Neural Information Processing Systems*, pages 1396–1404, 2015.

[85] Assaf Naor. An introduction to the ribe program. *Japanese Journal of Mathematics*, 7(2):167–233, 2012.

[86] S. V. Nepomnyaschikh. Mesh theorems on traces, normalizations of function traces and their inversion. *Soviet J. Numer. Anal. Math. Modelling*, 6(3):223–242, 1991.

[87] Jindřich Nečas. *Direct methods in the theory of elliptic equations*. Springer Monographs in Mathematics. Springer, Heidelberg, 2012. Translated from the 1967 French original by Gerard Tronel and Alois Kufner, Editorial coordination and preface by Šárka Nečasová and a contribution by Christian G. Simader.

[88] Aleksandar Nikolov. Randomized rounding for the largest simplex problem. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 861–870. ACM, 2015.

[89] Aleksandar Nikolov and Mohit Singh. Maximizing determinants under partition constraints. In *STOC*, pages 192–201, 2016.

[90] Geno Nikolov. Inequalities of Duffin-Schaeffer type ii. *Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences*, 2003.

[91] Christopher Conway Paige. *The computation of eigenvalues and eigenvectors of very large sparse matrices.* PhD thesis, University of London, 1971.

[92] Beresford N Parlett, H Simon, and LM Stringer. On estimating the largest eigenvalue with the Lanczos algorithm. *Mathematics of computation*, 38(157):153–165, 1982.

[93] Jack Poulson. High-performance sampling of generic determinantal point processes. *Philosophical Transactions of the Royal Society A*, 378(2166):20190059, 2020.

[94] Patrick Rebeschini and Amin Karbasi. Fast mixing for discrete point processes. In *COLT*, pages 1480–1500, 2015.

[95] Alex W. N. Riasanocsky. spread_numeric. https://github.com/ariasanovsky/spread_numeric, commit = c75e5d1726361eba04292c46c41009ea76e401e9, 2021.

[96] Justin Rising, Alex Kulesza, and Ben Taskar. An efficient algorithm for the symmetric principal minor assignment problem. *Linear Algebra and its Applications*, 473:126 – 144, 2015.

[97] Dhruv Rohatgi, John C Urschel, and Jake Wellens. Regarding two conjectures on clique and biclique partitions. *arXiv preprint arXiv:2005.02529*, 2020.

[98] Donald J Rose, R Endre Tarjan, and George S Lueker. Algorithmic aspects of vertex elimination on graphs. *SIAM Journal on computing*, 5(2):266–283, 1976.

[99] Yousef Saad. On the rates of convergence of the Lanczos and the block-Lanczos methods. *SIAM Journal on Numerical Analysis*, 17(5):687–706, 1980.

[100] Yousef Saad. *Numerical methods for large eigenvalue problems: revised edition*, volume 66. Siam, 2011.

[101] Warren Schudy. *Approximation Schemes for Inferring Rankings and Clusterings from Pairwise Data*. PhD thesis, Brown University, 2012.

[102] Michael Ian Shamos and Dan Hoey. Geometric intersection problems. In *17th Annual Symposium on Foundations of Computer Science (sfcs 1976)*, pages 208–215. IEEE, 1976.

[103] Jie Shen, Tao Tang, and Li-Lian Wang. *Spectral methods: algorithms, analysis and applications*, volume 41. Springer Science & Business Media, 2011.

[104] Max Simchowitz, Ahmed El Alaoui, and Benjamin Recht. On the gap between strict-saddles and true convexity: An omega (log d) lower bound for eigenvector approximation. *arXiv preprint arXiv:1704.04548*, 2017.

[105] Max Simchowitz, Ahmed El Alaoui, and Benjamin Recht. Tight query complexity lower bounds for PCA via finite sample deformed Wigner law. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1249–1259, 2018.

[106] Jasper Snoek, Richard S. Zemel, and Ryan Prescott Adams. A determinantal point process latent variable model for inhibition in neural spiking data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 1932–1940, 2013.

[107] Ernst Steinitz. Polyeder und raumeinteilungen. *Encyk der Math Wiss*, 12:38–43, 1922.

[108] 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*, pages 315–323. Society for Industrial and Applied Mathematics, 2015.

[109] JW Suurballe. Disjoint paths in a network. *Networks*, 4(2):125–145, 1974.

[110] Gabor Szego. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1939.

[111] P. L. Tchebichef. Sur le rapport de deux integrales etendues aux memes valeurs de la variable. *Zapiski Akademii Nauk, vol. 44, Supplement no. 2, 1883. Oeuvres. Vol. 2, pp. 375-402.*

[112] Tamás Terpai. Proof of a conjecture of V. Nikiforov. *Combinatorica*, 31(6):739–754, 2011.

[113] Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

[114] Francesco G Tricomi. Sugli zeri dei polinomi sferici ed ultrasferici. *Annali di Matematica Pura ed Applicata*, 31(4):93–97, 1950.

[115] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

[116] W. T. Tutte. How to draw a graph. *Proc. London Math. Soc. (3)*, 13:743–767, 1963.

[117] William Thomas Tutte. How to draw a graph. *Proceedings of the London Mathematical Society*, 3(1):743–767, 1963.

[118] John Urschel. spread_numeric+. https://math.mit.edu/~urschel/spread_numeric+.zip, April, 2021. [Online].

[119] John Urschel, Victor-Emmanuel Brunel, Ankur Moitra, and Philippe Rigollet. Learning determinantal point processes with moments and cycles. In *International Conference on Machine Learning*, pages 3511–3520. PMLR, 2017.

[120] John C Urschel. On the characterization and uniqueness of centroidal voronoi tessellations. *SIAM Journal on Numerical Analysis*, 55(3):1525–1547, 2017.

[121] John C Urschel. Nodal decompositions of graphs. *Linear Algebra and its Applications*, 539:60–71, 2018.

[122] John C Urschel. Uniform error estimates for the Lanczos method. *SIAM Journal of Matrix Analysis*, To appear.

[123] John C Urschel and Jake Wellens. Testing gap k-planarity is NP-complete. *Information Processing Letters*, 169:106083, 2021.

[124] John C Urschel and Ludmil T Zikatanov. Discrete trace theorems and energy minimizing spring embeddings of planar graphs. *Linear Algebra and its Applications*, 609:73–107, 2021.

[125] Jos LM Van Dorsselaer, Michiel E Hochstenbach, and Henk A Van Der Vorst. Computing probabilistic bounds for extreme eigenvalues of symmetric matrices with the Lanczos method. *SIAM Journal on Matrix Analysis and Applications*, 22(3):837–852, 2001.

[126] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[127] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[128] H. Whitney. Non-separable and planar graphs. *Transactions of the American Mathematical Society*, 34:339–362, 1932.

[129] Junliang Wu, Pingping Zhang, and Wenshi Liao. Upper bounds for the spread of a matrix. *Linear algebra and its applications*, 437(11):2813–2822, 2012.

[130] Yangqingxiang Wu and Ludmil Zikatanov. Fourier method for approximating eigenvalues of indefinite Stekloff operator. In *International Conference on High Performance Computing in Science and Engineering*, pages 34–46. Springer, 2017.

[131] Haotian Xu and Haotian Ou. Scalable discovery of audio fingerprint motifs in broadcast streams with determinantal point process based motif clustering. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 24(5):978–989, 2016.

[132] Jin-ge Yao, Feifan Fan, Wayne Xin Zhao, Xiaojun Wan, Edward Y. Chang, and Jianguo Xiao. Tweet timeline generation with determinantal point processes. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 3080–3086, 2016.

[133] Grigory Yaroslavtsev. Going for speed: Sublinear algorithms for dense r-csps. *arXiv preprint arXiv:1407.7887*, 2014.

[134] Fuzhen Zhang, editor. *The Schur complement and its applications*, volume 4 of *Numerical Methods and Algorithms*. Springer-Verlag, New York, 2005.

[135] Jonathan X Zheng, Samraat Pawar, and Dan FM Goodman. Graph drawing by stochastic gradient descent. *IEEE transactions on visualization and computer graphics*, 25(9):2738–2748, 2018.