Spectral sparsification

Spectral Sparsification: The Barrier Method and its Applications

Martin Camacho

Adviser: Jelani Nelson

An undegraduate thesis submitted to the The School of Engineering and Applied Sciences in partial fulfillment of the requirements for the joint degree of Bachelor of Arts in Computer Science and Mathematics with Honors

Harvard University
Cambridge, Massachusetts
April 1, 2014

Abstract

We survey recent literature focused on the following spectral sparsification question: Given an integer n and $\epsilon > 0$, does there exist a function $N(n, \epsilon)$ such that for every collection of $\mathbf{C}_1, \ldots, \mathbf{C}_m$ of $n \times n$ real symmetric positive semidefinite matrices whose sum is the identity, there exists a weighted subset of size $N(n, \epsilon)$ whose sum has eigenvalues lying between $1 - \epsilon$ and $1 + \epsilon$?

We present the algorithms for solving this problem given in [4, 8, 10]. These algorithms obtain $N(n, \epsilon) = O(n/\epsilon^2)$, which is optimal up to constant factors, through use of the barrier method, a proof technique involving potential functions which control the locations of the eigenvalues of a matrix under certain matrix updates.

We then survey the applications of this sparsification result and its proof techniques to graph sparsification [4, 10], low-rank matrix approximation [8], and estimating the covariance of certain distributions of random matrices [32, 26]. We end our survey by examining a multivariate generalization of the barrier method used in Marcus, Spielman, and Srivastava's recent proof [19] of the Kadison-Singer conjecture.

Acknowledgements

I am indebted to my thesis adviser, Prof. Jelani Nelson, for introducing me to this wonderful research topic and for having countless captivating and patient conversations with me explaining these topics and many, many others.

I would also like to give a heartfelt dedication to my mother and father for their lifelong support and encouragement in all aspects of my studies, as well as to the many friends who supported me during the process of writing this thesis.

Contents

1 | Spectral Sparsification and The Barrier Method

1.1 Introduction

Given an integer n and $\epsilon > 0$, does there exist a function $N(n, \epsilon)$ such that for every collection of $\mathbf{C}_1, \ldots, \mathbf{C}_m$ of $n \times n$ real symmetric positive semidefinite matrices whose sum is the identity, there exists a weighted subset of size $N(n, \epsilon)$ whose sum spectrally approximates the identity to a multiplicative factor of $(1 + \epsilon)$ – that is, whose eigenvalues lie between $1 - \epsilon$ and $1 + \epsilon$?

Remarkably, such functions $N(n, \epsilon)$ exist although they lack dependence on m, the original number of matrices in the collection. In this paper, we will concern ourselves with this type of *sparsification* result, so named since the ratio $N(n, \epsilon)/m$ can be taken to be arbitrarily small. In particular, we will focus on the following linear algebraic theorem essentially contained in the works of [4], [10], and [8], which finds that $N(n, \epsilon) = O(n/\epsilon^2)$ suffices and furthermore gives a deterministic algorithm for finding a subset of that size. This bound has optimal dependence on n and ϵ, up to constant factors [4].

Theorem 1.1.1. *Let $\epsilon > 0$, and let $\mathbf{C}_i, \mathbf{D}_i \in \mathbb{R}^{n \times n}$ for $i \in [m]$ be symmetric and positive semidefinite and suppose that*

$$\sum_i \mathbf{C}_i = \sum_i \mathbf{D}_i = \mathbf{I}_n.$$

Then there is a deterministic polynomial-time algorithm which finds scalars $s_i \geq 0$ for $i \in [m]$ such that at most $O(n/\epsilon^2)$ of the s_i are nonzero and

$$(1 - \epsilon)\mathbf{I}_n \preceq \sum_i s_i \mathbf{D}_i \text{ and } \sum_i s_i \mathbf{C}_i \preceq (1 + \epsilon)\mathbf{I}_n.$$

Note that this result is slightly different from the original question posed above. Before, we only had one collection of matrices to sparsify but could give simultaneous upper and lower bounds on the eigenvalues of the resulting sparsification; now, we are simultaneously sparsifying two collections of matrices, but can only give a lower bound for the eigenvalues of one of the sparsifications and an upper bound for the eigenvalues of the other. While the latter type of control will sometimes turn out to be useful, it is usually not needed; we will often apply this theorem to one collection of matrices by taking $\mathbf{C}_i = \mathbf{D}_i$.

In this chapter, we present the original motivation behind studying such matrix sparsifcation results, which arises from the study of graph sparsification [4, 10]. We then present the proof of Theorem 1.1.1 by the barrier method of Batson, Spielman, and Srivastava [4]. In Chapter 2, we present results of Boutsidis, Drineas, and Magdon-Ismail which use these sparsification arguments and results about approximating the SVD to obtain good spectral norm low-rank matrix approximations [8]. In Chapter 3, we present work of Srivastava-Vershynin [26] and Youssef [32], which utilize a randomization of the barrier method to derive results about the covariance of certain distributions of random matrices. Finally, in Chapter 4 we present Marcus, Spielman, and Srivastava's proof [19] of the long-standing Kadison-Singer problem, which uses a multivariate generalization of the barrier method and the methods of interlacing families and real stable polynomials.

1.1.1 Motivation: Graph Sparsification

A *sparsifier* of a graph G is a subgraph H that is structurally similar to G but may contain many fewer edges. Benczur and Karger [5] introduced *cut sparsifiers*, which preserve the weight of all of the cuts of G.

Definition 1.1.2. Let $G = (V, E, w)$ be a weighted graph, and for all $U \subseteq V$, let

$$\delta(U) = \{(u, v) \in E : u \in U, v \notin V\}$$

be the weight of the *cut* defined by U. A $(1+\epsilon)$-*cut sparsifier* of a graph G is a new weight function w' on E such that $\forall U \subseteq V$,

$$(1-\epsilon)|\delta(U)| \leq w(\delta(U)) \leq (1+\epsilon)|\delta(U)|. \tag{1.1}$$

Later, Spielman and Teng [25] introduced *spectral sparsifiers* based on the eigenvalues of the Laplacian of a graph, leading to a stronger matrix-based characterization of sparsification.

Definition 1.1.3. Let $G = (V, E, w)$ be a weighted graph, and let $n = |V|$ and $m = |E|$. The *Laplacian* $\mathbf{L}_G$ of G is given by $\mathbf{L}_G = \mathbf{B}^T\mathbf{W}\mathbf{B}$, where $\mathbf{B}$ is the $m \times n$ edge-vertex incidence matrix defined as

$$\mathbf{B}_{ev} = \begin{cases} 1, & e = (v, *) \\ -1, & e = (*, v) \\ 0, & \text{otherwise} \end{cases}$$

and $\mathbf{W}$ is the diagonal weight matrix with $\mathbf{W}_{ee} = w(e)$. The Laplacian is positive semidefinite and the number of connected components of G is equal to the multiplicity of 0 as an eigenvalue of $\mathbf{L}_G$.

Definition 1.1.4. A $(1+\epsilon)$-*spectral sparsifier* of an weighted undirected graph $G = (V, E, w)$ is a subgraph (with possibly different weight function) $H = (V, F, w')$ such that

$$(1-\epsilon)\mathbf{L}_G \preceq \mathbf{L}_H \preceq (1+\epsilon)\mathbf{L}_G. \tag{1.2}$$

Thus, spectral sparsifiers approximately preserve the eigenvalues of the Laplacian. To see that this is a stronger characterization than cut sparsification, note that the condition (1.2) implies the condition (1.1) since (1.2) is equivalent to

$$(1-\epsilon)\mathbf{x}^T\mathbf{L}_G\mathbf{x} \leq \mathbf{x}^T\mathbf{L}_H\mathbf{x} \leq (1+\epsilon)\mathbf{x}^T\mathbf{L}_G\mathbf{x} \text{ for all } \mathbf{x} \in \mathbb{R}^V \tag{1.3}$$

and (1.1) follows from this by taking $\mathbf{x}$ to be characteristic vectors of the cut sets U.

It is well-known that the family of Ramanujan graphs [17] yields good spectral sparsifiers for the complete graph K_n.

Definition 1.1.5. A *Ramanujan graph* is a d-regular graph G whose adjacency matrix has nontrivial eigenvalues (eigenvalues other than d and $-d$) of magnitude at most $2\sqrt{d-1}$. Equivalently, the non-zero eigenvalues of its Laplacian lie between $d - 2\sqrt{d-1}$ and $d + 2\sqrt{d-1}$.

Proposition 1.1.6. *Let $n, d > 0$, let H be a d-regular Ramanujan graph on n vertices, and give every edge of H weight $n/(d - 2\sqrt{d-1})$. Then H is a γ_R-spectral sparsifier of the complete graph K_n, where*

$$\gamma_R = \frac{d + 2\sqrt{d-1}}{d - 2\sqrt{d-1}}.$$

Proof. The eigenvalues of the Laplacian of K_n are 0 and n, so the proposition follows immediately from the previous definitions. □

Much research has been done to find algorithms for more general graphs G. In their paper introducing spectral sparsification, Spielman and Teng [25] gave a randomized construction for finding $(1+\epsilon)$-spectral sparsifiers with $O(n \operatorname{polylog} n/\epsilon^2)$ edges. Spielman and Srivastava [24] gave another randomized algorithm for finding $(1+\epsilon)$-spectral sparsifiers with $O(n \log n/\epsilon^2)$ edges through a concentration inequality of Rudelson [23] and resistance properties of graphs. It turns out that their result can be generalized [10] through use of the Ahlswede-Winter inequality [2], a matrix Chernoff bound, while keeping the requirement of $O(n \log n/\epsilon^2)$ edges.

Theorem 1.1.7 (Ahlswede-Winter inequality, [2])**.** *Let $\mathbf{Y} \in \mathbb{R}^{n\times n}$ be a symmetric, positive semidefinite random matrix supported in $\mathbf{0} \preceq \mathbf{Y} \preceq \mathbf{I}_n$, let $\mathbb{E}\,\mathbf{Y} = \mu\mathbf{I}_n$, let $\epsilon \in (0, 1/2)$, and let $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ be independent copies of $\mathbf{Y}$. Then*

$$\mathbb{P}\left[(1-\epsilon)\mathbf{I}_n \preceq \frac{1}{\mu m}\sum_{i=1}^{m}\mathbf{Y}_i \preceq (1+\epsilon)\mathbf{I}_n\right] \geq 1 - 2n\exp\left(-\frac{m\epsilon^2\mu}{2\ln 2}\right). \tag{1.4}$$

Corollary 1.1.8 (Theorem 18, [10])**.** *Let $\mathbf{C}_1, \ldots, \mathbf{C}_m \in \mathbb{R}^{n\times n}$ be symmetric and positive semidefinite and suppose that $\sum_i \mathbf{C}_i = \mathbf{I}_n$. Then there is a randomized algorithm which finds scalars $s_i \geq 0$ for $i \in [m]$, at most $O(n \log n/\epsilon^2)$ of which are nonzero, such that*

$$\mathbb{P}\left[(1-\epsilon)\mathbf{I}_n \preceq \sum_i s_i\mathbf{C}_i \preceq (1+\epsilon)\mathbf{I}_n\right] > 1/2.$$

Proof. Consider the discrete probability distribution which takes value $\mathbf{C}_i/\operatorname{Tr}[\mathbf{C}_i]$ with probability $\operatorname{Tr}[\mathbf{C}_i]/n$. Apply the Ahlswede-Winter inequality (Theorem 1.1.7) with $\mathbf{Y}$ sampled from this distribution, so that $\mu = 1/n$. Then for

$$m > O\left(\frac{\ln n}{\epsilon^2\mu}\right) = O(n\log n/\epsilon^2),$$

the error in (1.4) is bounded by $1/2$. □

In Section 1.3, we will see that the condition (1.2) in the definition of spectral sparsifiers is essentially a matrix sparsification condition, so that Theorem 1.1.1 directly yields graph sparsifiers of size $O(n/\epsilon^2)$.

1.2 The Barrier Method

1.2.1 Intuition: Eigenvalues under rank-one updates

Suppose that in Theorem 1.1.1 one takes $\mathbf{C}_i = \mathbf{D}_i = \mathbf{v}_i\mathbf{v}_i^T$ for some $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{R}^n$ in isotropic position; that is, $\sum_i \mathbf{v}_i\mathbf{v}_i^T = \mathbf{I}_n$. This yields the following corollary, the form of the theorem originally proved in [4].

Corollary 1.2.1. *Let $\epsilon > 0$, and let $\mathbf{v}_1, \ldots, \mathbf{v}_m$ with*

$$\sum_{i=1}^{m}\mathbf{v}_i\mathbf{v}_i^T = \mathbf{I}_n.$$

Then there is a deterministic polynomial-time algorithm which finds scalars $s_i \geq 0$ for $i \in [m]$ such that at most $O(n/\epsilon^2)$ of the s_i are nonzero and

$$(1-\epsilon)\mathbf{I}_n \preceq \sum_i s_i\mathbf{v}_i\mathbf{v}_i^T \preceq (1+\epsilon)\mathbf{I}_n.$$

To motivate the method of proof, we will look at what happens to the eigenvalues of a matrix $\mathbf{A}$ after a rank-one update of the form $\mathbf{v}\mathbf{v}^T$. We will need the well-known matrix determinant lemma, which shows that the determinant behaves nicely under such updates.

Lemma 1.2.2 (Matrix determinant lemma)**.** *Suppose that $\mathbf{A} \in \mathbb{R}^{n\times n}$ is invertible and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are column vectors. Then*

$$\det(\mathbf{A} + \mathbf{u}\mathbf{v}^T) = \det(\mathbf{A})(1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}).$$

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric and positive semidefinite with unit eigenvectors $\mathbf{u}_i$ and corresponding eigenvalues λ_i, and let $\mathbf{v} \in \mathbb{R}^n$. Since $\mathbf{A}$ is symmetric, we can eigendecompose it as $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, such that $\mathbf{u}_i$ are the columns of the orthogonal matrix $\mathbf{U}$ and $\mathbf{D}$ has λ_i on the diagonal. Then the characteristic polynomial of the rank-one update $\mathbf{A} + \mathbf{v}\mathbf{v}^T$ is equal to

$$\begin{aligned}
\chi[\mathbf{A} + \mathbf{v}\mathbf{v}^T](x) &= \det(x\mathbf{I} - \mathbf{A} - \mathbf{v}\mathbf{v}^T) \\
&= \det(x\mathbf{I} - \mathbf{A})(1 - \mathbf{v}^T(x\mathbf{I} - \mathbf{A})^{-1}\mathbf{v}) && \text{(by the matrix determinant lemma)} \\
&= \chi[\mathbf{A}](x)(1 - \mathbf{v}^T(x\mathbf{I} - \mathbf{A})^{-1}\mathbf{v}) \\
&= \chi[\mathbf{A}](x)(1 - \mathbf{v}^T\mathbf{U}(x\mathbf{I} - \mathbf{D})^{-1}\mathbf{U}^T\mathbf{v}) \\
&= \chi[\mathbf{A}](x)\left(1 - \sum_{j=1}^{n} \frac{\langle \mathbf{v}, \mathbf{u}_j\rangle^2}{x - \lambda_j}\right).
\end{aligned}$$

Suppose further that $\mathbf{v}$ is drawn uniformly at random from the set of $\mathbf{v}_i$. Then in expectation, we get that

$$\begin{aligned}
\mathbb{E}\,\chi[\mathbf{A} + \mathbf{v}\mathbf{v}^T](x) &= \chi[\mathbf{A}](x)\left(1 - \sum_{j=1}^{n} \frac{\mathbb{E}\langle \mathbf{v}, \mathbf{u}_j\rangle^2}{x - \lambda_j}\right) \\
&= \chi[\mathbf{A}](x)\left(1 - \frac{1}{m}\sum_{j=1}^{n}\sum_{i=1}^{m} \frac{\langle \mathbf{v}_i, \mathbf{u}_j\rangle^2}{x - \lambda_j}\right) \\
&= \chi[\mathbf{A}](x)\left(1 - \frac{1}{m}\sum_{j=1}^{n} \frac{\mathbf{u}_j^T(\sum_{i=1}^{m} \mathbf{v}_i\mathbf{v}_i^T)\mathbf{u}_j}{x - \lambda_j}\right) \\
&= \chi[\mathbf{A}](x)\left(1 - \frac{1}{m}\sum_{j=1}^{n} \frac{1}{x - \lambda_j}\right) \\
&= (1 - \partial_x/m)\chi[\mathbf{A}](x).
\end{aligned}$$

Beginning with $\mathbf{A} = 0$ and $\chi[\mathbf{A}](x) = x^n$ and iterating k times yields a family of *associated Laguerre polynomials*

$$p_k(x) = (1 - \partial_x/m)^k x^n, \tag{1.5}$$

whose roots are known. In particular, after $k = O(n/\epsilon^2)$ iterations, the ratio of the largest zero to the smallest zero becomes $O((1+\epsilon)/(1-\epsilon))$ [11], our desired approximation ratio. However, we have only seen how adding vectors randomly behaves *in expectation*; we need to show we can find an actual sequence consisting of scalar multiples of the vectors $\mathbf{v}_i$ from our set which mimics repeatedly adding the average vector.

Since we have two different collections $\mathbf{C}_i$ and $\mathbf{D}_i$, we will begin with matrices $\mathbf{A}^{(0)} = \mathbf{B}^{(0)} = \mathbf{0}$ and update $\mathbf{A}$ with multiples of the $\mathbf{C}_i$ and $\mathbf{B}$ with multiples of the $\mathbf{D}_i$ to form two sequences of matrices $\mathbf{A}^{(q)}$ and $\mathbf{B}^{(q)}$ for $q = 0, \ldots, T$.

In order to control the locations of the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$, we define the *upper and lower barrier potentials* as follows:

$$\Phi^u(\mathbf{A}) = \sum_{i=1}^{k} \frac{1}{u - \lambda_i(\mathbf{A})} = \mathrm{Tr}[(u\mathbf{I} - \mathbf{A})^{-1}], \qquad \Phi_\ell(\mathbf{B}) = \sum_{i=1}^{k} \frac{1}{\lambda_i(\mathbf{B}) - \ell} = \mathrm{Tr}[(\mathbf{B} - \ell\mathbf{I})^{-1}].$$

These potentials (which are equal to constant multiples of the Stieltjes transform of $\mathbf{A}$ and $\mathbf{B}$ evaluated at u and ℓ) give information about the locations of all of the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$ simultaneously. For example, if all of $\mathbf{A}$'s eigenvalues lie below u and $\Phi^u(\mathbf{A}) = D$, then no eigenvalue can be bigger than $u - 1/D$. We will control the maximum eigenvalue of $\mathbf{A}^{(q)}$ using an upper barrier u and the minimum eigenvalue of $\mathbf{B}^{(q)}$ using a lower barrier ℓ. In tandem, we will make sure these above potentials do not increase over the course of the algorithm, thus keeping the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$ safely bounded by the barriers.

Later, we will choose positive constants $u_0, \ell_0, \delta_U, \delta_L, \epsilon_U, \epsilon_L$ so that the algorithm will satisfy the following properties:

1. At the beginning of the algorithm, the upper and lower barriers are at $u = u_0$ and $\ell = \ell_0$ with initial potentials $\Phi^{u_0}(\mathbf{0}) = \epsilon_U$ and $\Phi_{\ell_0}(\mathbf{0}) = \epsilon_L$.
2. For each timestep $q = 1, \ldots, T$, there is some index $i \in [m]$ and scalar $t \geq 0$ such that
$$\mathbf{A}^{(q)} = \mathbf{A}^{(q-1)} + t\mathbf{C}_i \text{ and } \mathbf{B}^{(q)} = \mathbf{B}^{(q-1)} + t\mathbf{D}_i.$$
3. If we increment the barriers u and l by δ_U and δ_L respectively at each timestep $q = 1, \ldots, T$, neither potential increases, and no eigenvalue ever crosses a barrier:
$$\Phi^{u+\delta_U}(\mathbf{A}^{(q)}) \leq \Phi^u(\mathbf{A}^{(q-1)}) \leq \epsilon_U, \quad \Phi_{l+\delta_L}(\mathbf{B}^{(q)}) \leq \Phi_l(\mathbf{B}^{(q-1)}) \leq \epsilon_L.$$
$$\lambda_{\max}(\mathbf{A}^{(q)}) \leq u_0 + q\delta_U \text{ and } \lambda_{\min}(\mathbf{B}^{(q)}) \geq \ell_0 + q\delta_L.$$
4. The algorithm will finish after $T = O(n/\epsilon^2)$ steps, at which point
$$\frac{\lambda_{\max}(\mathbf{A}^{(T)})}{\lambda_{\min}(\mathbf{B}^{(T)})} \leq \frac{u_0 + T\delta_U}{l_0 + T\delta_L} \leq \frac{1+\epsilon}{1-\epsilon}.$$

We now introduce formulas which will allow us to compute the largest scalar multiple of a matrix $\mathbf{C}$ (resp. $\mathbf{D}$) which we can add to $\mathbf{A}$ (resp. $\mathbf{B}$) while preserving the above barrier properties. Let

$$U^{\mathbf{A}}(\mathbf{C}) \stackrel{\text{def}}{=} \frac{\mathrm{Tr}[((u+\delta_U)\mathbf{I} - \mathbf{A})^{-2}\mathbf{C}]}{\Phi^u(\mathbf{A}) - \Phi^{u+\delta_U}(\mathbf{A})} + \mathrm{Tr}[((u+\delta_U)\mathbf{I} - \mathbf{A})^{-1}\mathbf{C}],$$
$$L_{\mathbf{B}}(\mathbf{D}) \stackrel{\text{def}}{=} \frac{\mathrm{Tr}[(\mathbf{B} - (\ell+\delta_L)\mathbf{I})^{-2}\mathbf{D}]}{\Phi_{\ell+\delta_L}(\mathbf{B}) - \Phi_\ell(\mathbf{B})} - \mathrm{Tr}[(\mathbf{B} - (\ell+\delta_L)\mathbf{I})^{-1}\mathbf{D}].$$

The structure of the above quantities and the proofs of next two lemmas are inspired by the Sherman-Morrison-Woodbury formula, a description of the inverse of a matrix under certain updates.

Lemma 1.2.3 (Sherman-Morrison-Woodbury Formula, [14])**.**

$$(\mathbf{A} - \mathbf{UV})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{U}(\mathbf{I} - \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}.$$

Using this identity, Lemmas 1.2.4 and 1.2.5 explicitly determine the magnitude of feasible updates to $\mathbf{A}$ and $\mathbf{B}$ in terms of the quantities $U^{\mathbf{A}}(\mathbf{C})$ and $L_{\mathbf{B}}(\mathbf{D})$.

Lemma 1.2.4 (Upper Barrier Shift)**.** *Suppose $\lambda_{\max}(\mathbf{A}) < u$ and that $\mathbf{C} \in \mathbb{R}^{n\times n}$ is symmetric and positive semidefinite. If $t > 0$ satisfies $t^{-1} \geq U^{\mathbf{A}}(\mathbf{C})$, then*

$$\Phi^{u+\delta_U}(\mathbf{A} + t\mathbf{C}) \leq \Phi^u(\mathbf{A}) \text{ and } \lambda_{max}(\mathbf{A} + t\mathbf{C}) < u + \delta_U.$$

Proof. Let $u = u' + \delta_U$, let $\mathbf{M} = u'\mathbf{I} - \mathbf{A}$, and let $\mathbf{S} = \mathbf{C}^{1/2}$ be a symmetric square root of $\mathbf{C}$. By the Sherman-Morrison-Woodbury formula (Lemma 1.2.3), we may write

$$\begin{aligned}
\Phi^{u+\delta_U}(\mathbf{A}+t\mathbf{C}) &= \operatorname{Tr}[(u'\mathbf{I} - (\mathbf{A}+t\mathbf{S}\mathbf{S}))^{-1}] \\
&= \operatorname{Tr}[(\mathbf{M} - t\mathbf{S}\mathbf{S})^{-1}] \\
&= \operatorname{Tr}\left[\mathbf{M}^{-1} + t\mathbf{M}^{-1}\mathbf{S}(\mathbf{I} - t\mathbf{S}\mathbf{M}^{-1}\mathbf{S})^{-1}\mathbf{S}\mathbf{M}^{-1}\right] \\
&= \Phi^{u'}(\mathbf{A}) + t\operatorname{Tr}\left[\mathbf{M}^{-1}\mathbf{S}(\mathbf{I} - t\mathbf{S}\mathbf{M}^{-1}\mathbf{S})^{-1}\mathbf{S}\mathbf{M}^{-1}\right] \\
&= \Phi^{u'}(\mathbf{A}) + t\operatorname{Tr}\left[(\mathbf{I} - t\mathbf{S}\mathbf{M}^{-1}\mathbf{S})^{-1}\mathbf{S}\mathbf{M}^{-2}\mathbf{S}\right] \\
&= \Phi^{u}(\mathbf{A}) - (\Phi^{u}(\mathbf{A}) - \Phi^{u'}(\mathbf{A})) + t\operatorname{Tr}\left[(\mathbf{I} - t\mathbf{S}\mathbf{M}^{-1}\mathbf{S})^{-1}\mathbf{S}\mathbf{M}^{-2}\mathbf{S}\right].
\end{aligned}$$

Note that by assumption we have $1/t \geq U^{\mathbf{A}}(\mathbf{C}) > \operatorname{Tr}[\mathbf{S}\mathbf{M}^{-1}\mathbf{S}]$. Since $\lambda_{\max}(\mathbf{S}\mathbf{M}^{-1}\mathbf{S}) \leq \operatorname{Tr}[\mathbf{S}\mathbf{M}^{-1}\mathbf{S}]$, we have that

$$\gamma \stackrel{\text{def}}{=} \lambda_{\min}(\mathbf{I} - t\mathbf{S}\mathbf{M}^{-1}\mathbf{S}) = 1 - t\lambda_{\max}(\mathbf{S}\mathbf{M}^{-1}\mathbf{S}) > 0.$$

We thus have $\gamma\mathbf{I} \preceq \mathbf{I} - t\mathbf{S}\mathbf{M}^{-1}\mathbf{S}$, so that $0 \prec (\mathbf{I} - t\mathbf{S}\mathbf{M}^{-1}\mathbf{S})^{-1} \preceq \gamma^{-1}\mathbf{I}$, and $\operatorname{Tr}\left[(\mathbf{I} - t\mathbf{S}\mathbf{M}^{-1}\mathbf{S})^{-1}\mathbf{S}\mathbf{M}^{-2}\mathbf{S}\right] \geq t\gamma^{-1}\operatorname{Tr}[\mathbf{S}\mathbf{M}^{-2}\mathbf{S}]$. Thus, to show that $\Phi^{u+\delta_U}(\mathbf{A}+t\mathbf{C}) \leq \Phi^u(\mathbf{A})$, we must show that

$$\Phi^u(\mathbf{A}) - \Phi^{u'}(\mathbf{A}) \geq t\gamma^{-1}\operatorname{Tr}\left[\mathbf{S}\mathbf{M}^{-2}\mathbf{S}\right] = \frac{\operatorname{Tr}\left[\mathbf{S}\mathbf{M}^{-2}\mathbf{S}\right]}{t^{-1} - \lambda_{\max}(\mathbf{S}\mathbf{M}^{-1}\mathbf{S})}.$$

Since $\lambda_{\max}(\mathbf{S}\mathbf{M}^{-1}\mathbf{S}) \leq \operatorname{Tr}[\mathbf{S}\mathbf{M}^{-1}\mathbf{S}]$, it suffices to show that

$$\Phi^u(\mathbf{A}) - \Phi^{u'}(\mathbf{A}) \geq t\gamma^{-1}\operatorname{Tr}\left[\mathbf{S}\mathbf{M}^{-2}\mathbf{S}\right] = \frac{\operatorname{Tr}\left[\mathbf{S}\mathbf{M}^{-2}\mathbf{S}\right]}{t^{-1} - \operatorname{Tr}[\mathbf{S}\mathbf{M}^{-1}\mathbf{S}]},$$

which follows by substituting $t^{-1} = U^{\mathbf{A}}(\mathbf{C})$.

Now, suppose that $\lambda_{\max}(\mathbf{A}+t\mathbf{C}) \geq u'$. Then by continuity, there is some $0 < t' < t$ such that $\lambda_{\max}(\mathbf{A}+t'\mathbf{C}) = u'$, and thus $\Phi^{u'}(\mathbf{A}+t'\mathbf{C})$ is infinite. But since $1/t' \geq 1/t$, by the above we have that $\Phi^{u'}(\mathbf{A}+t'\mathbf{A}) \leq \Phi^u(\mathbf{A})$ and is thus finite, a contradiction. Thus $\lambda_{\max}(\mathbf{A}+t\mathbf{C}) < u'$. □

Lemma 1.2.5 (Lower Barrier Shift)**.** *Suppose $\ell < \lambda_{\min}(\mathbf{B}), \Phi_l(\mathbf{B}) \leq 1/\delta_L$, and that $\mathbf{D} \in \mathbb{R}^{n\times n}$ is symmetric and positive semidefinite. If $t > 0$ satisfies $t^{-1} \leq L_{\mathbf{B}}(\mathbf{D})$, then*

$$\Phi_{\ell+\delta_L}(\mathbf{B}+t\mathbf{D}) \leq \Phi_\ell(\mathbf{B}) \text{ and } \lambda_{\min}(\mathbf{B}+t\mathbf{D}) > \ell + \delta_L.$$

Proof. The proof is very similar to that of the previous lemma. Let $\ell' = \ell + \delta_L$, let $\mathbf{N} = \mathbf{B} - \ell'\mathbf{I}$, and let $\mathbf{S} = \mathbf{D}^{1/2}$ be a symmetric square root of $\mathbf{D}$. By the Sherman-Morrison formula (Lemma 1.2.3), we may write

$$\begin{aligned}
\Phi_{\ell+\delta_L}(\mathbf{B}+t\mathbf{D}) &= \operatorname{Tr}[(\mathbf{B} + t\mathbf{S}\mathbf{S} - \ell'\mathbf{I})^{-1}] \\
&= \operatorname{Tr}[(\mathbf{N} + t\mathbf{S}\mathbf{S})^{-1}] \\
&= \operatorname{Tr}[\mathbf{N}^{-1} - t\mathbf{N}^{-1}\mathbf{S}(\mathbf{I} + t\mathbf{S}\mathbf{N}^{-1}\mathbf{S})^{-1}\mathbf{S}\mathbf{N}^{-1}] \\
&= \Phi_{\ell'}(\mathbf{B}) - t\operatorname{Tr}[\mathbf{N}^{-1}\mathbf{S}(\mathbf{I} + t\mathbf{S}\mathbf{N}^{-1}\mathbf{S})^{-1}\mathbf{S}\mathbf{N}^{-1}] \\
&= \Phi_{\ell'}(\mathbf{B}) - t\operatorname{Tr}[(\mathbf{I} + t\mathbf{S}\mathbf{N}^{-1}\mathbf{S})^{-1}\mathbf{S}\mathbf{N}^{-2}\mathbf{S}] \\
&= \Phi_{\ell}(\mathbf{B}) + (\Phi_{\ell'}(\mathbf{B}) - \Phi_\ell(\mathbf{B})) - t\operatorname{Tr}[(\mathbf{I} + t\mathbf{S}\mathbf{N}^{-1}\mathbf{S})^{-1}\mathbf{S}\mathbf{N}^{-2}\mathbf{S}].
\end{aligned}$$

Thus, to show that $\Phi_{\ell'}(\mathbf{B}+t\mathbf{D}) \leq \Phi_\ell(\mathbf{B})$, it suffices to show that

$$\Phi_{\ell'}(\mathbf{B}) - \Phi_\ell(\mathbf{B}) \leq t\operatorname{Tr}[(\mathbf{I} + t\mathbf{S}\mathbf{N}^{-1}\mathbf{S})^{-1}\mathbf{S}\mathbf{N}^{-2}\mathbf{S}].$$

Let

$$\gamma \stackrel{\text{def}}{=} \lambda_{\max}(\mathbf{I} + t\mathbf{S}\mathbf{N}^{-1}\mathbf{S}) = 1 + t\lambda_{\max}(t\mathbf{S}\mathbf{N}^{-1}\mathbf{S}) > 0.$$

Then $\gamma^{-1}\mathbf{I} \prec (\mathbf{I} + t\mathbf{S}\mathbf{N}^{-1}\mathbf{S})^{-1}$, so that

$$t\,\mathrm{Tr}[(\mathbf{I} + t\mathbf{S}\mathbf{N}^{-1}\mathbf{S})^{-1}\mathbf{S}\mathbf{N}^{-2}\mathbf{S}] \geq \gamma^{-1}t\,\mathrm{Tr}[\mathbf{S}\mathbf{N}^{-2}\mathbf{S}] \geq \frac{\mathrm{Tr}[\mathbf{S}\mathbf{N}\mathbf{S}^{-1}]}{t^{-1} + \mathrm{Tr}[\mathbf{S}\mathbf{N}^{-1}\mathbf{S}]}$$

and it suffices to show that

$$\Phi_{\ell'}(\mathbf{B}) - \Phi_{\ell}(\mathbf{B}) \leq \frac{\mathrm{Tr}[\mathbf{S}\mathbf{N}^{-2}\mathbf{S}]}{t^{-1} + \mathrm{Tr}[\mathbf{S}\mathbf{N}^{-1}\mathbf{S}]},$$

which follows by substituting $t^{-1} = L_{\mathbf{B}}(\mathbf{D})$. The fact that $\lambda_{\min}(\mathbf{B} + t\mathbf{D}) > \ell + \delta_L$ follows by a similar argument to the one at the end of the previous lemma. □

The following lemma shows that at each timestep q there is a good choice of index i and scalar t such that adding $t\mathbf{C}_i$ to $\mathbf{A}^{(t)}$ and adding $t\mathbf{D}_i$ to $\mathbf{B}^{(t)}$ does not cause any eigenvalues to cross their respectively barriers. Equivalently, it will hold that this choice of i and t simultaneously achieves the bounds on $U^{\mathbf{A}}(\mathbf{C}_i)$ and $L_{\mathbf{B}}(\mathbf{D}_i)$ required by the previous two lemmas.

Lemma 1.2.6 (Both Barriers)**.** *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ be symmetric and positive semidefinite, and suppose that $\lambda_{\max}(\mathbf{A}) < u, \Phi^u(\mathbf{A}) < \epsilon_U, \lambda_{\min}(\mathbf{B}) > \ell, \Phi_\ell(\mathbf{B}) < \epsilon_L$, and $\delta_U^{-1} + \epsilon_U \leq \delta_L^{-1} - \epsilon_L$, . Then there exists an index $i \in [m]$ and $t > 0$ such that*

$$\Phi^{u+\delta_U}(\mathbf{A} + t\mathbf{C}_i) \leq \Phi^u(\mathbf{A}),\ \lambda_{\max}(\mathbf{A} + t\mathbf{C}_i) < u + \delta_U,$$

$$\Phi_{\ell+\delta_L}(\mathbf{B} + t\mathbf{D}_i) \leq \Phi_\ell(\mathbf{B}),\ \textit{and}\ \ell + \delta_L < \lambda_{\min}(\mathbf{B} + t\mathbf{D}_i).$$

Proof. We will use an averaging argument: that is, we will show that

$$\sum_{i=1}^{n} U^{\mathbf{A}}(\mathbf{C}_i) \leq \sum_{i=1}^{n} L_{\mathbf{B}}(\mathbf{D}_i). \tag{1.6}$$

Assuming this, there must be some $i \in [m]$ such that $U^{\mathbf{A}}(\mathbf{C}_i) \leq L_{\mathbf{B}}(\mathbf{D}_i)$, and applying Lemmas 1.2.4 and 1.2.5 with $t = 1/U^{\mathbf{A}}(\mathbf{C}_i)$ yields the desired results.

Now, we show (1.6). Let $u' = u + \delta_U$. We may write

$$\begin{aligned}
\sum_{i=1}^{n} U^{\mathbf{A}}(\mathbf{C}_i) &= \frac{\sum_i \mathrm{Tr}[(u'\mathbf{I} - \mathbf{A})^{-2}\mathbf{C}_i]}{\Phi^u(\mathbf{A}) - \Phi^{u'}(\mathbf{A})} + \sum_i \mathrm{Tr}[(u'\mathbf{I} - \mathbf{A})^{-1}\mathbf{C}_i] \\
&= \frac{\mathrm{Tr}[(u'\mathbf{I} - \mathbf{A})^{-2}]}{\Phi^u(\mathbf{A}) - \Phi^{u'}(\mathbf{A})} + \mathrm{Tr}[(u'\mathbf{I} - \mathbf{A})^{-1}] \qquad (\text{since } \textstyle\sum_i B_i = 1) \\
&= \frac{\sum_j (u' - \lambda_j(\mathbf{A}))^{-2}}{\sum_j [(u - \lambda_j(\mathbf{A}))^{-1} - (u' - \lambda_j(\mathbf{A}))^{-1}]} + \Phi^{u'}(\mathbf{A}) \\
&= \delta_U^{-1} \left(\frac{\sum_j (u' - \lambda_j(\mathbf{A}))^{-2}}{\sum_j [(u - \lambda_j(\mathbf{A}))^{-1}(u' - \lambda_j(\mathbf{A}))^{-1}]} \right) + \Phi^{u'}(\mathbf{A}) \\
&\leq \delta_U^{-1} + \Phi^{u'}(\mathbf{A}) \qquad (\text{since } (u - \lambda_j(\mathbf{A}))^{-1} \geq (u' - \lambda_j(\mathbf{A}))^{-1}) \\
&\leq \delta_U^{-1} + \epsilon_U.
\end{aligned}$$

We follow a similar calculation for the lower barrier:

$$\begin{aligned}\sum_{i=1}^{n} L_{\mathbf{B}}(\mathbf{D}_i) &= \frac{\sum_i \operatorname{Tr}[(\mathbf{B}-\ell'\mathbf{I})^{-2}\mathbf{D}_i]}{\Phi_{\ell'}(\mathbf{B})-\sum_i \Phi_\ell(\mathbf{B})} - \operatorname{Tr}[(\mathbf{B}-\ell'\mathbf{I})^{-1}\mathbf{D}_i] \\ &= \frac{\operatorname{Tr}[(\mathbf{B}-\ell'\mathbf{I})^{-2}]}{\Phi_{\ell'}(\mathbf{B})-\Phi_\ell(\mathbf{B})} - \operatorname{Tr}[(\mathbf{B}-\ell'\mathbf{I})^{-1}] \\ &= \frac{\sum_j (\lambda_j(\mathbf{B})-\ell')^{-2}}{\sum_j [(\lambda_j(\mathbf{B})-\ell')^{-1}-(\lambda_j(\mathbf{B})-\ell)^{-1}]} - \Phi_{\ell'}(\mathbf{B}) \\ &\geq \delta_L^{-1} - \Phi_{\ell'}(\mathbf{B}) \\ &\geq \delta_L^{-1} - \epsilon_L,\end{aligned}$$

where the last inequality follows from Lemma 1.2.7 below. Equation (1.6) then follows by our assumption that $\delta_U^{-1}+\epsilon_U \leq \delta_L^{-1}-\epsilon_L$. □

Lemma 1.2.7. *Suppose $0<\delta_L\leq\epsilon_L^{-1}\leq\lambda_i-\ell$ for all i and $0\leq\sum_i(\lambda_i-\ell)^{-1}\leq\epsilon_L$. Let $\ell'=\ell+\delta_L$. Then*

$$\frac{\sum_i(\lambda_i-\ell')^{-2}}{\sum_i(\lambda_i-\ell')^{-1}-\sum_i(\lambda_i-\ell)^{-1}} - \sum_i(\lambda_i-\ell')^{-1} \geq \delta_L^{-1} - \sum_i(\lambda_i-l)^{-1}.$$

Proof. By our first hypothesis we also have that $\lambda_i-\ell'>0$. Let $y_i=(\lambda_i-\ell')^{-1}$ and $z_i=(\lambda_i-\ell)^{-1}$. Note that $\sum_i(y_i-z_i)=\delta_L y_i z_i$. Substituting and multiplying through by the first denominator of the left hand side, we obtain that the given inequality is equivalent to

$$\sum_i y_i^2 \geq \left(\delta_L\sum_i y_i z_i\right)(\delta_L^{-1}+\delta_L\sum_i y_i z_i) = \sum_i y_i z_i + \left[\delta_L\sum_i y_i z_i\right]^2.$$

Rearranging again, we obtain

$$\delta_L\sum_i y_i^2 z_i \geq \left[\delta_L\sum_i y_i z_i\right]^2.$$

Applying Cauchy-Schwarz to the RHS, we obtain

$$\left[\delta_L\sum_i y_i z_i\right]^2 \leq \left[\delta_L\sum_i z_i\right]\left[\delta_L\sum_i y_i^2 z_i\right] \leq (\delta_L\epsilon_L)\left[\delta_L\sum_i y_i^2 z_i\right] \leq \delta_L\sum_i y_i^2 z_i,$$

where the last inequality follows from the assumption that $\delta_L\epsilon_L\leq 1$, and the claimed inequality is established. □

We are finally able to prove the main theorem.

Proof of Theorem 1.1.1. We need to choose the six constants required by the algorithm. Take

$$\delta_L = 1, \epsilon_L = \frac{\epsilon}{2}, \ell_0 = -\frac{n}{\epsilon_L}, \delta_U = \frac{2+\epsilon}{2-\epsilon}, \epsilon_U = \frac{\epsilon}{2\delta_U}, u_0 = \frac{n}{\epsilon_U}. \tag{1.7}$$

This yields $\frac{1}{\delta_U}+\epsilon_U = \frac{1}{\delta_L}-\epsilon_L$, the condition required by Lemma 1.2.6. By the previous lemmas, the algorithm will satisfy conditions 1 – 4 above. Using these constants, we obtain that after $T=4n/\epsilon^2$ iterations,

$$\frac{\lambda_{\max}(\mathbf{A}^{(T)})}{\lambda_{\min}(\mathbf{B}^{(T)})} \leq \frac{u_0+T\delta_U}{\ell_0+T\delta_L} = \frac{(2+\epsilon)^2}{(2-\epsilon)^2} \leq \frac{1+\epsilon}{1-\epsilon}.$$

We can now obtain the scalars s_i by letting s_i be the sum of all t corresponding to steps when the matrices $\mathbf{C}_i$ and $\mathbf{D}_i$ were taken as updates, so that

$$\frac{\lambda_{\max}(\sum_i s_i \mathbf{C}_i)}{\lambda_{\min}(\sum_i s_i \mathbf{D}_i)} \leq \frac{1+\epsilon}{1-\epsilon}.$$

Since the algorithm had T steps, at most $T = O(n/\epsilon^2)$ of the s_i will be nonzero. Finally, we can rescale all of the s_i so that $\lambda_{\min}(\sum_i s_i \mathbf{D}_i) = 1-\epsilon$ and $\lambda_{\max}(\sum_i s_i \mathbf{C}_i) \leq 1+\epsilon$. □

The running time of the algorithm of Theorem 1.1.1 can be analyzed as follows. At the start of each timestep, we can precompute the matrix powers $\mathbf{M}^{-1}$, etc. in time $O(n^3)$. For each $i \in [m]$, we must then calculate the functions $U^{\mathbf{A}}(\mathbf{C}_i)$ and $L_{\mathbf{B}}(\mathbf{D}_i)$: we can compute the traces in the formulae using entry-ise products which take time $O(n^2)$. Thus each iteration runs in time $O(n^3 + mn^2) = O(mn^2)$, and the total running time is $O(Tmn^2) = O(mn^3/\epsilon^2)$.

Theorem 1.1.1 has a useful generalization which drops the condition on that the sum of the matrices is the identity. We will need this version for the graph sparsification results of Section 1.3.

Corollary 1.2.8 (Dual-Set Sparsification for General Decompositions)**.** *Let $\epsilon > 0$, let $\mathbf{C}_i, \mathbf{D}_i \in \mathbb{R}^{n\times n}$ for $i \in [m]$ be symmetric and positive semidefinite, let*

$$\mathbf{C} = \sum_i \mathbf{C}_i \text{ and } \mathbf{D} = \sum_i \mathbf{D}_i.$$

Then there is a deterministic polynomial-time algorithm which finds scalars $s_i \geq 0$ for $i \in [m]$ such that at most $O(n/\epsilon^2)$ of the s_i are nonzero and

$$(1-\epsilon)\mathbf{D} \preceq \sum_i s_i \mathbf{D}_i \text{ and } \sum_i s_i \mathbf{C}_i \preceq (1+\epsilon)\mathbf{C}.$$

Proof of Corollary 1.2.8. Since invertible matrices are dense in the space of all $n \times n$ matrices, it suffices to prove the corollary when $\mathbf{C}$ and $\mathbf{D}$ are invertible. Define the functions

$$f(\mathbf{X}) = \mathbf{C}^{-1/2}\mathbf{X}\mathbf{C}^{-1/2} \text{ and } g(\mathbf{Y}) = \mathbf{D}^{-1/2}\mathbf{Y}\mathbf{D}^{-1/2},$$

and apply Theorem 1.1.1 to the collections of matrices $f(\mathbf{C}_i)$ and $g(\mathbf{D}_i)$, each of which sum to the identity, to obtain the result. □

This reduction takes time $O(mn^3)$, and does not affect the runtime of the original algorithm.

Remark. The assumption that the matrices $\mathbf{C}_i$ and $\mathbf{D}_i$ are positive semidefinite found in Theorem 1.1.1 is necessary, as shown below.

Proposition 1.2.9 (Proposition 31, [10])**.** *For every positive integer n, there exist symmetric matrices $\mathbf{C}_1, \ldots, \mathbf{C}_m \in \mathbb{R}^{n\times n}$ with $m = \Omega(n^2)$ such that $\mathbf{C} = \sum_i \mathbf{C}_i$ is positive definite and for every $\epsilon \in (0,1)$ and $y_1, \ldots, y_m \in \mathbb{R}$ such that $(1-\epsilon)\mathbf{C} \preceq \sum_i y_i \mathbf{C}_i$, all of the y_i are nonzero.*

Proof. For all $1 \leq i < j \leq n$, let $\mathbf{E}_{ij} = \mathbf{e}_i\mathbf{e}_j^T + \mathbf{e}_j\mathbf{e}_i^T$, and suppose $\mathbf{Y}$ is the matrix with all ones, which is positive semidefinite. Let the $\mathbf{C}_k$ consist of the $\mathbf{E}_{ij}$ and $2\mathbf{I}$, so that

$$\mathbf{C} = 2\mathbf{I} + \sum_{1\leq i<j\leq n} \mathbf{E}_{ij} = \mathbf{I} + \mathbf{Y},$$

which is positive definite. Now, suppose that for $\epsilon \in (0,1)$ and scalars t, z_{ij} we have that $(1-\epsilon)\mathbf{C} \preceq 2t\mathbf{I} + \sum_{1\leq i<j\leq n} z_{ij}\mathbf{E}_{ij}$. Taking traces of both sides and dividing by n yields $2(1-\epsilon)/n \leq 2t$, and multiplying both sides by $\mathbf{E}_{ij}$ yields that $2(1-\epsilon) \leq z_{ij}$ for each $1 \leq i < j \leq n$. □

1.3 Deterministic Graph Sparsification

The sparsification result of Corollary 1.2.1 yields an excellent deterministic solution to the graph sparsification problem for *any* weighted undirected graph. Batson, Spielman, and Srivastava [4] proved the following theorem:

Theorem 1.3.1. *Let $\epsilon > 0$ and let $G = (V, E, w)$ be a weighted undirected graph with $|V| = n$. Then there is a deterministic polynomial-time algorithm which finds a $(1+\epsilon)$-spectral sparsifier $H = (V, F, w')$ of G such that $|F| = O(n/\epsilon^2)$.*

Remark Their original proof used different constants in (1.7) to obtain the approximation ratio

$$1 + \epsilon = \frac{d + 1 + 2\sqrt{d}}{d + 1 - 2\sqrt{d}}$$

for some $d > 1$ and obtained sparsifiers with $\lceil d(n-1) \rceil$ edges. As such, these 'twice-Ramanujan sparsifiers' [4] have the same asymptotic approximation ratio as Ramanujan graphs and have (approximately) twice the number of edges, since Ramanujan graphs are d-regular and thus have $dn/2$ edges.

Proof of Theorem 1.3.1. For each $e = (i, j) \in E$, define the matrix

$$\mathbf{L}_e = w(e)(\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T, \tag{1.8}$$

so that $\mathbf{L}_G = \sum_{e \in E} \mathbf{L}_e$. Then applying Corollary 1.2.8 with $\mathbf{C}_e = \mathbf{D}_e = \mathbf{L}_e$ yields scalars s_e, at most $O(n/\epsilon)^2$ of which are nonzero, such that

$$(1 - \epsilon)\mathbf{L}_G \leq \sum_{e \in E} s_e \mathbf{L}_e \leq (1 - \epsilon)\mathbf{L}_G \tag{1.9}$$

Note that the quantity $\sum_{e \in E} s_e \mathbf{L}_e$ is the Laplacian $\mathbf{L}_H$ of a subgraph H of G with new weights $w'(e) = s_e w(e)$, and H is a $(1+\epsilon)$-sparsifier of G by (1.3). □

It turns out that it is still possible to find such sparsifiers even if one must also preserve cost functions or weights of colorings.

Proposition 1.3.2 ([10])**.** *Let $G = (V, E, w)$ be an undirected graph, and let $c_1, \ldots, c_k : E \to \mathbb{R}^+$ be cost functions. For any $\epsilon \in (0, 1)$, there is a deterministic polynomial-time algorithm which finds a $(1+\epsilon)$-spectral sparsifier $H = (V, F, w')$ of G such that $|F| = O((n+k)/\epsilon^2)$ and for all i,*

$$(1 - \epsilon) \sum_{e \in E} w(e)c_i(e) \leq \sum_{e \in F} w'(e)c_i(e) \leq (1 + \epsilon) \sum_{e \in E} w(e)c_i(e).$$

Proof. For each $e \in E$, recall the definition of the matrix $\mathbf{L}_e$ given in (1.8) and define the direct sum $\mathbf{B}_e = [\mathbf{L}_e \oplus w(e)c_1(e) \oplus \cdots \oplus w(e)c_k(e)]$. Then

$$\mathbf{B} = \sum_{e \in E} \mathbf{B}_e = [\mathbf{L}_G \oplus \sum_{e \in E} w(e)c_1(e) \oplus \cdots \oplus \sum_{e \in E} w(e)c_k(e)].$$

To obtain the result, apply Corollary 1.2.8 to the collections $\mathbf{C}_e = \mathbf{D}_e = \mathbf{B}_e$. □

Corollary 1.3.3 ([10])**.** *Let $G = (V, E, w)$ be an undirected graph, and let $E_1, \ldots, E_k$ be a partition (coloring) of the edges in E. For any $\epsilon \in (0, 1)$, there is a deterministic polynomial-time algorithm which finds a $(1+\epsilon)$-spectral sparsifier $H = (V, F, w')$ of G such that $|F| = O((n+k)/\epsilon^2)$ and*

$$(1 - \epsilon) \sum_{e \in E_i} w(e) \leq \sum_{e \in F \cap E_i} w'(e) \leq (1 + \epsilon) \sum_{e \in E_i} w(e).$$

Proof. Let c_i be the cost functions given by the characteristic functions of E_i, and apply the previous proposition. □

These techniques also yield spectral sparsifiers for hypergraphs. First, we define versions of the Laplacian matrix and spectral sparsifiers for hypergraphs.

Definition 1.3.4 ([10])**.** Let $\mathcal{G} = (V, \mathcal{E}, w)$ be a hypergraph. For each hyperedge $E \in \mathcal{E}$, define its Laplacian $\mathbf{L}_E$ as the Laplacian of a graph on V which forms a clique on the vertices in E and has no other edges, and then define the Laplacian of $\mathcal{G}$ by $\mathbf{L}_\mathcal{G} = \sum_{E\in\mathcal{E}} w(E)\mathbf{L}_E$.

A $(1+\epsilon)$-spectral sparsifier of $\mathcal{G}$ is a sub-hypergraph $\mathcal{H}$ satisfying

$$(1-\epsilon)\mathbf{L}_\mathcal{G} \preceq \mathbf{L}_\mathcal{H} \preceq (1+\epsilon)\mathbf{L}$$

The following analogue of Theorem 1.3.1 for hypergraphs immediately follows from the above definitions.

Proposition 1.3.5 ([10])**.** *Let $\mathcal{H} = (V, \mathcal{E}, w)$ be a hypergraph. For any $\epsilon \in (0, 1)$, there is a deterministic polynomial-time algorithm which finds a $(1+\epsilon)$-spectral sparsifier $\mathcal{G} = (V, \mathcal{F}, w')$ of $\mathcal{H}$ with $|\mathcal{F}| = O(n/\epsilon^2)$.*

Proof. Apply Corollary 1.2.8 to the matrices $w(E)\mathbf{L}_E$ for each $E \in \mathcal{E}$, which sum to $\mathbf{L}_\mathcal{G}$. □

These are just a few examples of applications taken from those in [10]; others not mentioned here include results for cut sparsifiers of hypergraphs, semidefinite programs, and Caratheodory-type theorems.

A note on lower bounds Inspired by the Alon-Boppana bound [21], which implies that if H is a d-regular *unweighted* $(1+\epsilon)$-sparsifier of the complete graph K_n, then $\epsilon \geq 4/\sqrt{d} - o(1/\sqrt{d})$ as $n, d \to \infty$, the authors of [4] make the following conjecture for weighted sparsifiers.

Conjecture 1.3.1 ([4])**.** *Let $H = (V, E, w)$ be a weighted graph with n vertices and average degree d. If H is a $(1+\epsilon)$-spectral sparsifier of the complete graph K_n, then*

$$\epsilon \geq \frac{4}{\sqrt{d}} + O\left(1/d\right).$$

They also prove the following weaker theorem instead, which still is able to yield an asymptotic lower bound of $O(n/\epsilon^2)$ for our original sparsification problem [20].

Theorem 1.3.6 (Proposition 4.2, [4])**.** *Let $H = (V, E, w)$ be a weighted graph with n vertices and average degree d. If H is a $(1+\epsilon)$-spectral sparsifier of the complete graph K_n, then*

$$\epsilon \geq \frac{2}{\sqrt{d}} - O\left(\frac{\sqrt{d}}{n}\right).$$

2 | Applications to Low-Rank Matrix Approximation

2.1 Introduction

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$, let $\text{rank}(\mathbf{A}) = \rho$, and let $k \leq \rho$ be an integer. Then we recall that the *singular value decomposition* (SVD) of $\mathbf{A}$ is given by

$$\mathbf{A} = \underbrace{\begin{pmatrix} \mathbf{U}_k & \mathbf{U}_{\rho-k} \end{pmatrix}}_{\mathbf{U}} \underbrace{\begin{pmatrix} \mathbf{\Sigma}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{\rho-k} \end{pmatrix}}_{\mathbf{\Sigma}} \underbrace{\begin{pmatrix} \mathbf{V}_k^T \\ \mathbf{V}_{\rho-k}^T \end{pmatrix}}_{\mathbf{V}^T},$$

where the matrix $\mathbf{\Sigma}$ contains the nonzero singular values $\sigma_1, \ldots, \sigma_\rho$ on the diagonal, the matrix $\mathbf{U} \in \mathbb{R}^{m \times \rho}$ contains the left singular vectors of $\mathbf{A}$, and the matrix $\mathbf{V} \in \mathbb{R}^{n \times \rho}$ contains the right singular values of $\mathbf{A}$.

It is well known that the best rank-k approximation of a matrix $\mathbf{A}$ with respect to a unitarily invariant norm is given by $\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T$, but evaluating this requires computing (parts of) the SVD, which may be too slow for certain applications. Thus, much research has been done on the column-based reconstruction problem, which concerns algorithms for efficiently finding subspaces spanned by a small set of $r \ll n$ columns of $\mathbf{A}$ which still yield good approximations in some norm.

The problem is phrased as follows. Let $\mathbf{C} \in \mathbb{R}^{m \times r}$ consist of r columns of $\mathbf{A}$ for some $r < n$. Taking $||\cdot||_N$ to be either the spectral or Frobenius norm, we define $\pi^N_{\mathbf{C},k}(\mathbf{A})$ to be the best approximation to $\mathbf{A}$ with respect to $||\cdot||_N$ that lies within the column space of $\mathbf{C}$ and that has rank at most $k \leq r$. How does the column-based reconstruction error

$$\left|\left|\mathbf{A} - \pi^N_{\mathbf{C},k}(\mathbf{A})\right|\right|_N$$

compare multiplicatively to the error of the best unconstrained rank k approximation, given by $||\mathbf{A} - \mathbf{A}_k||_N$?

In this chapter, we will follow the spectral part of the analysis of [8], which applies the sparsification results of Chapter 1 to obtain deterministic and randomized algorithms in both norms for column-based reconstruction for any $r \geq k$. For a more detailed history of the approaches to column-based reconstruction and other related formulations of low-rank approximation, as well as similar results for the Frobenius norm, we encourage the reader to consult the introduction of [8].

The main highlight will be the following theorem, which combines one of the aforementioned deterministic algorithms with a fast randomized approximation of the SVD to obtain near-optimal error in expectation while avoiding computing singular vectors.

Theorem 2.1.1 (Fast randomized spectral norm reconstruction)**.** *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ have rank ρ, and let $1 < k < \rho$ be an integer. There exists a randomized algorithm which selects $r > k$ columns of $\mathbf{A}$ and forms a matrix $\mathbf{C} \in \mathbb{R}^{m \times r}$ such that*

$$\mathbb{E}[\left|\left|\mathbf{A} - \pi^2_{\mathbf{C},k}(\mathbf{A})\right|\right|_2] \leq O(\sqrt{n/r})\, ||\mathbf{A} - \mathbf{A}_k||_2 \,.$$

The algorithm runs in $O(mnk \log(\min(m,n)/k)/\epsilon + nrk^2)$ time.

2.2 Sparsification and Low-Rank Matrix Approximation

2.2.1 Preliminaries

We will use the spectral norm $||\cdot||_2$ and the Frobenius norm $||\cdot||_F$, which are both unitarily invariant and submultiplicative: that is, for suitably sized matrices $\mathbf{A}, \mathbf{B}$, $||\mathbf{AB}|| \leq ||\mathbf{A}||||\mathbf{B}||$ under either norm. In addition, $||\mathbf{AB}||_F \leq ||\mathbf{A}||_2||\mathbf{B}||_F$ and $||\mathbf{AB}||_F \leq ||\mathbf{A}||_F||\mathbf{B}||_2$. These norms also satisfy a few properties reminiscent of the Pythagorean theorem.

Lemma 2.2.1. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$. If $\mathbf{A}^T\mathbf{B} = \mathbf{0}$ or $\mathbf{B}^T\mathbf{A} = \mathbf{0}$, then*

$$||\mathbf{A}+\mathbf{B}||_F^2 = ||\mathbf{A}||_F^2 + ||\mathbf{B}||_F^2 .$$

Proof. Suppose $\mathbf{A}^T\mathbf{B} = \mathbf{0}$. Then

$$||\mathbf{A}+\mathbf{B}||_F^2 = \mathrm{Tr}[(\mathbf{A}+\mathbf{B})(\mathbf{A}+\mathbf{B})^T] = \mathrm{Tr}[\mathbf{A}\mathbf{A}^T] + \mathrm{Tr}[\mathbf{B}\mathbf{B}^T] + \mathrm{Tr}[\mathbf{B}\mathbf{A}^T] = ||\mathbf{A}||_F^2 + ||\mathbf{B}||_F^2 ,$$

where in the last equality we used the cyclic property of trace. □

Lemma 2.2.2. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$. If $\mathbf{A}^T\mathbf{B} = \mathbf{0}$ or $\mathbf{B}^T\mathbf{A} = \mathbf{0}$, then*

$$\max(||\mathbf{A}||_2^2 , ||\mathbf{B}||_2^2) \leq ||\mathbf{A}+\mathbf{B}||_2^2 \leq ||\mathbf{A}||_2^2 + ||\mathbf{B}||_2^2 .$$

Proof. Suppose $\mathbf{A}^T\mathbf{B} = \mathbf{0}$. Let $\mathbf{u} \in \mathbf{R}^n$ be a unit vector. Then

$$||\mathbf{A}+\mathbf{B}||_2^2 = \max_{\mathbf{u}} \mathbf{u}^T(\mathbf{A}+\mathbf{B})(\mathbf{A}+\mathbf{B})^T\mathbf{u} = \max_{\mathbf{u}} \mathbf{u}^T(\mathbf{A}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T)\mathbf{u} \leq ||\mathbf{A}||_2^2 + ||\mathbf{B}||_2^2.$$

In addition,

$$\max_{\mathbf{u}} \mathbf{u}^T(\mathbf{A}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T)\mathbf{u} \geq \max\left(\max_{\mathbf{u}} \mathbf{u}^T\mathbf{A}\mathbf{A}^T\mathbf{u}, \max_{\mathbf{u}} \mathbf{u}^T\mathbf{B}\mathbf{B}^T\mathbf{u}\right) = \max(||\mathbf{A}||_2^2 , ||\mathbf{B}||_2^2).$$

□

We will also need the following inequality concerning the spectral norm of projection operators.

Lemma 2.2.3 (Theorem 2.1, [27])**.** *Let $\mathbf{P} \in \mathbb{R}^{n\times n}$ be a nonzero projection. Then*

$$||\mathbf{I}-\mathbf{P}||_2 \leq ||\mathbf{P}||_2.$$

Proof. If $\mathbf{P} = \mathbf{I}$ the result is obvious, so we can also assume that $\mathbf{I}-\mathbf{P}$ is also a nonzero projection. Let $\mathbf{u} \in \mathbb{R}^n$ be a unit vector, and let $\mathbf{x} = \mathbf{P}\mathbf{u}$ and $\mathbf{y} = (\mathbf{I}-\mathbf{P})\mathbf{u}$. If $\mathbf{x} = \mathbf{0}$, then $||(\mathbf{I}-\mathbf{P})\mathbf{u}||_2 = 1 \leq ||\mathbf{P}||_2$. If $\mathbf{y} = \mathbf{0}$, then $||(\mathbf{I}-\mathbf{P})\mathbf{u}||_2 = 0 \leq ||\mathbf{P}||_2$. Thus, assume that both $\mathbf{x}$ and $\mathbf{y}$ are nonzero. Let

$$\mathbf{w} = \frac{||\mathbf{y}||_2}{||\mathbf{x}||_2}\mathbf{x} + \frac{||\mathbf{x}||_2}{||\mathbf{y}||_2}\mathbf{y},$$

so that $||\mathbf{w}||_2 = 1$. But then

$$||\mathbf{P}||_2 \geq ||\mathbf{P}\mathbf{w}||_2 = \left|\left|\frac{||\mathbf{y}||_2}{||\mathbf{x}||_2}\mathbf{x}\right|\right|_2 = ||\mathbf{y}||_2 = ||(\mathbf{I}-\mathbf{P})\mathbf{u}||_2.$$

Thus $||\mathbf{P}||_2 \geq |(\mathbf{I}-\mathbf{P})\mathbf{u}||_2$ for every unit vector $\mathbf{u}$, and the result follows. □

We now turn to results concerning representations of $\pi^N_{\mathbf{C},k}(\mathbf{A})$. By definition, we can express it by

$$\pi^F_{\mathbf{C},k}(\mathbf{A}) = \mathbf{C}\mathbf{X}, \text{ where } \mathbf{X} = \operatorname*{argmin}_{\mathbf{Y}\in\mathbb{R}^{r\times n}:\mathrm{rank}(\mathbf{Y})\leq k} ||\mathbf{A}-\mathbf{C}\mathbf{Y}||_N . \tag{2.1}$$

The next lemma shows that $\pi^N_{\mathbf{C},k}(\mathbf{A})$ is the projection of $\mathbf{A}$ onto the column space of $\mathbf{C}\mathbf{X}$, and no other subspace of the column space of $\mathbf{C}$ can approximate $\mathbf{A}$ better.

Lemma 2.2.4. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{C} \in \mathbb{R}^{m\times r}$, and let $r > k$. Let $\mathbf{X} \in \mathbb{R}^{r\times n}$ be the matrix in (2.1), and let $\mathbf{Y} \in \mathbb{R}^{r\times n}$ be any matrix with rank $\leq k$. Then*

$$||\mathbf{A} - \mathbf{CX}||_N = ||\mathbf{A} - (\mathbf{CX})(\mathbf{CX})^+\mathbf{A}||_N \leq |\mathbf{A} - (\mathbf{CY})(\mathbf{CY})^+\mathbf{A}||_N.$$

Proof. Since $\mathbf{Y}(\mathbf{CY})^+\mathbf{A}$ has rank at most k, the second inequality follows by (2.1). Taking $\mathbf{Y} = \mathbf{X}$ and squaring, we obtain

$$||\mathbf{A} - \mathbf{CX}||_N \leq ||\mathbf{A} - (\mathbf{CX})(\mathbf{CX})^+\mathbf{A}||_N.$$

Next, we may write

$$||\mathbf{A} - \mathbf{CX}||_N^2 = ||(\mathbf{A} - (\mathbf{CX})(\mathbf{CX})^+\mathbf{A}) - (\mathbf{CX})(\mathbf{I} - (\mathbf{CX})^+\mathbf{A}))||_N^2.$$

Since $((\mathbf{CX})(\mathbf{CX})^+)^T = (\mathbf{CX})(\mathbf{CX})^+$, we have that $(\mathbf{A} - (\mathbf{CX})(\mathbf{CX})^+\mathbf{A})^T(\mathbf{CX})(\mathbf{I} - (\mathbf{CX})^+\mathbf{A})) = \mathbf{0}$, so we can apply Lemmas 2.2.1 and 2.2.2 to obtain

$$||\mathbf{A} - \mathbf{CX}||_N^2 \leq ||\mathbf{A} - (\mathbf{CX})(\mathbf{CX})^+\mathbf{A}||_N^2.$$

□

In the Frobenius norm, we can give an explicit formula for $\pi^F_{\mathbf{C},k}(\mathbf{A})$ which requires computing an SVD.

Lemma 2.2.5. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{C} \in \mathbb{R}^{m\times r}$, let $0 < k < r$ be an integer, and let $\mathbf{B} = \mathbf{O}(\mathbf{O}^T\mathbf{A})_k$, where $\mathbf{O} \in \mathbb{R}^{m\times r}$ consists of the orthonormal columns of $\mathbf{C}$. Then*

$$||\mathbf{A} - \mathbf{B}||_F^2 = \left|\left|A - \pi^F_{\mathbf{C},k}(\mathbf{A})\right|\right|_F,$$

Proof. Note that we can write

$$\left|\left|\mathbf{A} - \pi^F_{\mathbf{C},k}\right|\right|_F^2 = \left|\left|\mathbf{A} - \pi^F_{\mathbf{O},k}\right|\right|_F^2 = \min_{\mathbf{M}:\text{rank}(\mathbf{M})\leq k} ||\mathbf{A} - \mathbf{OM}||_F^2.$$

Expanding this term,

$$\begin{aligned}
||\mathbf{A} - \mathbf{OM}||_F^2 &= \left|\left|\mathbf{A} - \mathbf{OO}^T\mathbf{A} + \mathbf{O}(\mathbf{O}^T\mathbf{A} - \mathbf{M})\right|\right|_F^2 \\
&= \left|\left|\mathbf{A} - \mathbf{OO}^T\mathbf{A}\right|\right|_F^2 + \left|\left|\mathbf{O}(\mathbf{O}^T\mathbf{A} - \mathbf{M})\right|\right|_F^2 \\
&= \left|\left|\mathbf{A} - \mathbf{OO}^T\mathbf{A}\right|\right|_F^2 + \left|\left|\mathbf{O}^T\mathbf{A} - \mathbf{M}\right|\right|_F^2,
\end{aligned}$$

where the second equality follows from Lemma 2.2.1, as $(\mathbf{A} - \mathbf{OO}^T\mathbf{A})^T(\mathbf{O}(\mathbf{O}^T\mathbf{A} - \mathbf{M})) = 0$ because $\mathbf{O}$ is orthonormal. Since the above expression is minimized when $\mathbf{M} = \mathbf{O}^T\mathbf{A}$, the result follows. □

In the spectral norm, the same matrix yields a 2-approximation, which will be good enough for later results.

Lemma 2.2.6. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{C} \in \mathbb{R}^{m\times r}$, let $0 < k < r$ be an integer, and let $\mathbf{B} = \mathbf{O}(\mathbf{O}^T\mathbf{A})_k$, where $\mathbf{O} \in \mathbb{R}^{m\times r}$ consists of the orthonormal columns of $\mathbf{C}$. Then*

$$||\mathbf{A} - \mathbf{B}||_2^2 \leq 2\left|\left|\mathbf{A} - \pi^2_{\mathbf{C},k}(\mathbf{A})\right|\right|_2^2.$$

Proof. First, note that $\mathbf{OO}^T\mathbf{A}$ is the best approximation to $\mathbf{A}$ in the column space of $\mathbf{O}$, so that $||\mathbf{A} - \mathbf{OO}^T\mathbf{A}||_2^2 \leq ||\mathbf{A} - \pi^2_{\mathbf{Q},k}(\mathbf{A})||_2^2$. In addition,

$$||\mathbf{OO}^T\mathbf{A} - (\mathbf{OO}^T\mathbf{A})_k||_2^2 = \sigma_{k+1}^2(\mathbf{OO}^T\mathbf{A}) \leq \sigma_{k+1}^2(\mathbf{A}) = ||\mathbf{A} - \mathbf{A}_k||_2^2, \tag{2.2}$$

where the inequality follows since $\mathbf{OO}^T$ is a projection.

Noting that $(\mathbf{O}\mathbf{O}^T\mathbf{A})_k = \mathbf{O}(\mathbf{O}^T\mathbf{A})_k$, we now expand

$$\begin{aligned}
||\mathbf{A} - \mathbf{O}(\mathbf{O}^T\mathbf{A})_k||_2^2 &= ||\mathbf{A} - \mathbf{O}\mathbf{O}^T\mathbf{A} + \mathbf{O}(\mathbf{O}^T\mathbf{A} - (\mathbf{O}^T\mathbf{A})_k)||_2^2 \\
&\leq ||\mathbf{A} - \mathbf{O}\mathbf{O}^T\mathbf{A}||_2^2 + ||\mathbf{O}(\mathbf{O}^T\mathbf{A} - (\mathbf{O}^T\mathbf{A})_k)||_2^2 && \text{(by Lemma 2.2.2)} \\
&\leq ||\mathbf{A} - \pi^2_{\mathbf{O},k}(\mathbf{A})||_2^2 + ||\mathbf{O}(\mathbf{O}^T\mathbf{A} - (\mathbf{O}^T\mathbf{A})_k)||_2^2 && (\mathbf{O}\mathbf{O}^T\mathbf{A} \text{ is best approx. in colspan}(\mathbf{O})) \\
&\leq ||\mathbf{A} - \pi^2_{\mathbf{O},k}(\mathbf{A})||_2^2 + ||\mathbf{A} - \mathbf{A}_k||_2^2 && \text{(by (2.2))} \\
&\leq 2||\mathbf{A} - \pi^2_{\mathbf{O},k}(\mathbf{A})||_2^2 && (\mathbf{A}_k \text{ is the best rank-}k \text{ approx. to } \mathbf{A}) \\
&= 2||\mathbf{A} - \pi^2_{\mathbf{C},k}(\mathbf{A})||_2^2.
\end{aligned}$$

In applying Lemma 2.2.2 above, we used that $(\mathbf{A} - \mathbf{O}\mathbf{O}^T\mathbf{A})^T(\mathbf{O}(\mathbf{O}^T\mathbf{A} - (\mathbf{O}^T\mathbf{A})_k)) = \mathbf{0}$, which follows since $\mathbf{O}$ is orthonormal. □

2.2.2 Matrix Factorizations

For matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{X} \in \mathbb{R}^{n\times k}$, where $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, we will consider factorizations of the form

$$\mathbf{A} = \mathbf{A}\mathbf{X}\mathbf{X}^T + \mathbf{E},$$

which consist of the projection of $\mathbf{A}$ onto the column space of $\mathbf{X}$ and an orthogonal error term satisfying $\mathbf{E}\mathbf{X} = (\mathbf{A} - \mathbf{A}\mathbf{X}\mathbf{X}^T)\mathbf{X} = \mathbf{0}$. The next lemmas exhibit why we care about this family of factorizations: any such factorization of $\mathbf{A}$ yields bounds for the error obtained from approximating by $\pi^N_{\mathbf{C},k}(\mathbf{A})$. In these lemmas, the matrix $\mathbf{S} \in \mathbb{R}^{n\times r}$ will play the role of a *sampling matrix* which selects and possibly re-weights some columns of $\mathbf{A}$ to produce a matrix $\mathbf{C}$.

Lemma 2.2.7. *Consider as above $\mathbf{A} = \mathbf{B}\mathbf{X}^T + \mathbf{E}$, with $\mathbf{B} = \mathbf{A}\mathbf{X}$ and $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. Let $\mathbf{S} \in \mathbb{R}^{n\times r}$ be any matrix with $\operatorname{rank}(\mathbf{X}^T\mathbf{S}) = \operatorname{rank}(\mathbf{X}) = k$, and let $\mathbf{C} = \mathbf{A}\mathbf{S}$. Then the following inequality holds in both spectral and Frobenius norms:*

$$\left|\left|\mathbf{A} - \pi^N_{\mathbf{C},k}(\mathbf{A})\right|\right|_N^2 \leq ||\mathbf{E}||_N^2 + \left|\left|\mathbf{E}\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\right|\right|_N$$

Proof. Consider $\mathbf{Y} = \mathbf{C}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T$. Then $\operatorname{rank}(\mathbf{Y}) \leq k$ and $\mathbf{Y}$ is in the column space of $\mathbf{C}$, so that by definition

$$\left|\left|\mathbf{A} - \pi^N_{\mathbf{C},k}(\mathbf{A})\right|\right|_N \leq ||\mathbf{A} - \mathbf{Y}||_N .$$

Expanding the right hand side,

$$\begin{aligned}
\left|\left|\mathbf{A} - \mathbf{C}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T\right|\right|_N^2 &= \left|\left|\mathbf{B}\mathbf{X}^T + \mathbf{E} - (\mathbf{B}\mathbf{X}^T + \mathbf{E})\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T\right|\right|_N^2 \\
&= \left|\left|\mathbf{B}\mathbf{X}^T - \mathbf{B}\mathbf{X}^T\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T + \mathbf{E} - \mathbf{E}\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T\right|\right|_N^2 .
\end{aligned}$$

Note that $\mathbf{X}^T\mathbf{S}(\mathbf{X}^T\mathbf{S})^+ = \mathbf{I}_k$ since $\operatorname{rank}(\mathbf{X}^T\mathbf{S}) = k$, so that the first two terms cancel. Thus

$$\left|\left|\mathbf{A} - \mathbf{C}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T\right|\right|_N^2 = \left|\left|\mathbf{E} - \mathbf{E}\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T\right|\right|_N^2 .$$

Next, note that $\mathbf{E}\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T\mathbf{E}^T = \mathbf{E}\mathbf{S}(\mathbf{X}^T\mathbf{S})^+ \underbrace{\mathbf{X}^T(\mathbf{A}^T - \mathbf{X}\mathbf{X}^T\mathbf{A}^T)}_{\mathbf{X}^T\mathbf{A}^T - (\mathbf{X}^T\mathbf{X})\mathbf{X}^T\mathbf{A}^T = 0} = 0$, so that by Lemmas 2.2.1 and 2.2.2,

$$\left|\left|\mathbf{E} - \mathbf{E}\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T\right|\right|_N^2 \leq ||\mathbf{E}||_N^2 + \left|\left|\mathbf{E}\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T\right|\right|_N^2 .$$

Thus the inequality holds. □

Next, we obtain an analogue of the previous result which yields a multiplicative error bound instead of an additive one.

Lemma 2.2.8. *Consider as above* $\mathbf{A} = \mathbf{B}\mathbf{X}^T + \mathbf{E}$, *with* $\mathbf{B} = \mathbf{A}\mathbf{X}$ *and* $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. *Let* $\mathbf{S} \in \mathbb{R}^{n\times r}$ *be any matrix with* $\operatorname{rank}(\mathbf{X}^T\mathbf{S}) = \operatorname{rank}(\mathbf{X}) = k$, *and let* $\mathbf{C} = \mathbf{A}\mathbf{S}$. *Then the following inequality holds in both spectral and Frobenius norms:*

$$\left|\left|\mathbf{A} - \pi^N_{\mathbf{C},k}(\mathbf{A})\right|\right|^2_N \leq ||\mathbf{E}||^2_N \cdot \left|\left|\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\right|\right|_2$$

Proof. In the proof of the previous lemma, we showed

$$\left|\left|\mathbf{A} - \mathbf{C}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T\right|\right|^2_N = \left|\left|\mathbf{E} - \mathbf{E}\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T\right|\right|^2_N = \left|\left|\mathbf{E}(\mathbf{I} - \mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T)\right|\right|^2_N.$$

By submultiplicativity, the right hand side satisfies

$$\left|\left|\mathbf{E}(\mathbf{I} - \mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T)\right|\right|^2_N \leq ||\mathbf{E}||^2_N \cdot \left|\left|\mathbf{I} - \mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T)\right|\right|_2.$$

Next, note that $(\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T)^2 = \mathbf{S}(\mathbf{X}^T\mathbf{S})^+\underbrace{\mathbf{X}^T\mathbf{S}(\mathbf{X}^T\mathbf{S})^+}_{=\mathbf{I}_k}\mathbf{X}^T = \mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T$, so that $\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T$ is a nonzero projection. By Lemma 2.2.3 and since $\mathbf{X}^T\mathbf{X} = \mathbf{I}$,

$$\left|\left|\mathbf{I} - \mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T\right|\right|_2 \leq \left|\left|\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\mathbf{X}^T\right|\right|_2 = \left|\left|\mathbf{S}(\mathbf{X}^T\mathbf{S})^+\right|\right|_2,$$

and the claimed inequality follows. □

2.2.3 Deterministic Spectral Norm Reconstruction

We are now ready to present some preliminary deterministic algorithms for spectral column-based reconstruction. We will need the following corollary, which is yielded by different choices of constants in (1.7) and by a slight modification of Theorem 1.1.1 which allows the matrices $\mathbf{C}_i$ and $\mathbf{D}_i$ to have different (square) dimensions. We state it in the rank one case for simplicity.

Corollary 2.2.9. *Suppose* $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^k$ *and* $\mathbf{w}_1, \ldots, \mathbf{w}_n \in \mathbb{R}^d$ *satisfy* $\sum_i \mathbf{v}_i\mathbf{v}_i^T = \mathbf{I}_k$ *and* $\sum_i \mathbf{w}_i\mathbf{w}^T = \mathbf{I}_d$. *Given an integer* r *with* $k < r \leq m$, *there exist scalars* s_i, *at most* r *of which are nonzero, such that*

$$\left(1 - \sqrt{k/r}\right)^2 \mathbf{I}_k \preceq \sum_i s_i\mathbf{v}_i\mathbf{v}_i^T \text{ and } \sum_i s_i\mathbf{w}_i\mathbf{w}_i^T \preceq \left(1 + \sqrt{d/r}\right)^2 \mathbf{I}_d.$$

Proof. Note that the proof of Theorem 1.1.1 still holds even if the matrices $\mathbf{C}_i$ and $\mathbf{D}_i$ have different dimensions. The only changes will be to take constants

$$\delta_L = 1, \ell_0 = -\sqrt{rk}, \epsilon_L = -\frac{k}{\ell_0}, \delta_U = \frac{1 + \sqrt{d/r}}{1 - \sqrt{k/r}}, u_0 = \delta_U\sqrt{dr}, \epsilon_U = \frac{d}{u_0},$$

noting that $\delta_U^{-1} + \epsilon_U = \delta_L^{-1} - \epsilon_L$ as required by Lemma 1.2.6, and to run the algorithm for exactly $T = r$ steps to obtain an approximation ratio of

$$\frac{u_0 + r\delta_U}{\ell_0 + r\delta_L} = \frac{(1 + \sqrt{d/r})^2}{(1 - \sqrt{k/r})^2},$$

as required. □

Let $\mathbf{v}_i, \mathbf{w}_i$ and s_i be as in the above corollary, and suppose exactly r' of the s_i are nonzero. Using these s_i, define a sampling matrix $\mathbf{S} \in \mathbb{R}^{n\times r'}$ by including $\sqrt{s_i}\mathbf{e}_i$ as a column of $\mathbf{S}$ if and only if s_i is nonzero. Let $\mathbf{V}$ and $\mathbf{W}$ be matrices whose rows are $\mathbf{v}_i$ and $\mathbf{w}_i$, respectively. Then

$$\sum_i s_i\mathbf{v}_i\mathbf{v}_i^T = \mathbf{V}^T\mathbf{S}\mathbf{S}^T\mathbf{V} \text{ and } \sum_i s_i\mathbf{w}_i\mathbf{w}_i^T = \mathbf{W}^T\mathbf{S}\mathbf{S}^T\mathbf{W}.$$

With this notation, the above result thus implies that

$$1 - \sqrt{k/r} \leq \sigma_k(\mathbf{V}^T\mathbf{S}) \text{ and } \sigma_1(\mathbf{W}^T\mathbf{S}) \leq 1 + \sqrt{\ell/r}.$$

In the following algorithms, the sparsification result of Corollary 2.2.9 helps to control the singular values of these sampled matrices $\mathbf{V}^T\mathbf{S}$ which appear in the error terms of Lemmas 2.2.7 and 2.2.8.

Theorem 2.2.10 (Deterministic spectral norm reconstruction)**.** *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ and let $k < \rho$ be an integer. Then there exists a deterministic polynomial-time algorithm which selects $r > k$ columns of $\mathbf{A}$ and forms a matrix $\mathbf{C} \in \mathbb{R}^{m\times r}$ such that*

$$\left|\left|\mathbf{A} - \pi^2_{\mathbf{C},k}(\mathbf{A})\right|\right|_2 \leq O(\sqrt{\rho/r})\left|\left|\mathbf{A} - \mathbf{A}_k\right|\right|_2.$$

Proof. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the SVD, and let $\mathbf{S}$ be the sampling matrix obtained by applying Corollary 2.2.9 to the n orthonormal rows of $\mathbf{V}_k$ and the n orthonormal rows of $\mathbf{V}_{\rho-k}$. Then $\mathbf{S}$ satisfies

$$1 - \sqrt{k/r} \leq \sigma_k(\mathbf{V}_k^T\mathbf{S}) \text{ and } \sigma_1(\mathbf{V}_{\rho-k}^T\mathbf{S}) \leq 1 + \sqrt{(\rho-k)/r}.$$

Let $\mathbf{C} = \mathbf{A}\mathbf{S}$, so that $\mathbf{C}$ is constructed from rescaled columns of $\mathbf{A}$.

By Lemma 2.2.7 applied to the matrix $\mathbf{X} = \mathbf{V}_k$, we have that

$$\begin{aligned}
\left|\left|\mathbf{A} - \pi^2_{\mathbf{C},k}(\mathbf{A})\right|\right|_2^2 &\leq ||\mathbf{A} - \mathbf{A}_k||_2^2 + \left|\left|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}(\mathbf{V}_k^T\mathbf{S})^+\right|\right|_2^2 \\
&\leq ||\mathbf{A} - \mathbf{A}_k||_2^2 + ||(\mathbf{A} - \mathbf{A}_k)\mathbf{S}||_2^2 \cdot \left|\left|(\mathbf{V}_k^T\mathbf{S})^+\right|\right|_2^2 \\
&= ||\mathbf{A} - \mathbf{A}_k||_2^2 + \left|\left|\mathbf{U}_{\rho-k}\mathbf{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^T\mathbf{S}\right|\right|_2^2 \cdot \left|\left|(\mathbf{V}_k^T\mathbf{S})^+\right|\right|_2^2 \\
&\leq ||\mathbf{A} - \mathbf{A}_k||_2^2 + ||\mathbf{\Sigma}_{\rho-k}||_2^2 \cdot \left|\left|\mathbf{V}_{\rho-k}^T\mathbf{S}\right|\right|_2^2 \cdot \left|\left|(\mathbf{V}_k^T\mathbf{S})^+\right|\right|_2^2 \\
&= ||\mathbf{A} - \mathbf{A}_k||_2^2 + ||\mathbf{A} - \mathbf{A}_k||_2^2 \cdot \left|\left|\mathbf{V}_{\rho-k}^T\mathbf{S}\right|\right|_2^2 \cdot \left|\left|(\mathbf{V}_k^T\mathbf{S})^+\right|\right|_2^2 \\
&\leq ||\mathbf{A} - \mathbf{A}_k||_2^2 \left[1 + \left(\frac{1 + \sqrt{(\rho-k)/r}}{1 + \sqrt{k/r}}\right)^2\right]
\end{aligned}$$

Taking square roots and applying $\sqrt{1+x^2} \leq 1 + x$ yields the result. □

We can sacrifice some of this approximation accuracy for a slightly faster algorithm, which will have asymptotically similar error to the randomized algorithm of the next section. Instead of having to compute the full SVD, as in the previous algorithm, we only need the first k right singular vectors of $\mathbf{A}$.

Theorem 2.2.11 (Faster deterministic spectral norm reconstruction)**.** *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ, and let $k < \rho$ be an integer. There exists a deterministic polynomial-time algorithm which selects $r > k$ columns of $\mathbf{A}$ and forms a matrix $\mathbf{C} \in \mathbb{R}^{m\times r}$ such that*

$$\left|\left|\mathbf{A} - \pi^2_{\mathbf{C},k}(\mathbf{A})\right|\right|_2 \leq O(\sqrt{n/r})\left|\left|\mathbf{A} - \mathbf{A}_k\right|\right|_2.$$

Proof. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the SVD, and let $\mathbf{S}$ be the sampling matrix obtained by applying Corollary 2.2.9 to the n orthonormal rows of $\mathbf{V}_k$ and the n orthonormal rows of $\mathbf{I}_n$. Then $\mathbf{S}$ satisfies

$$||\mathbf{I}_n\mathbf{S}||_2 = \sigma_1(\mathbf{I}_n\mathbf{S}) \leq 1 + \sqrt{n/r}.$$

By Lemma 2.2.8, we obtain

$$\begin{aligned}
\left|\left|\mathbf{A} - \pi^2_{\mathbf{C},k}(\mathbf{A})\right|\right|_2^2 &\leq ||\mathbf{A} - \mathbf{A}_k||_2^2 \cdot \left|\left|\mathbf{S}(\mathbf{V}_k^T\mathbf{S})^+\right|\right|_2^2 \\
&= ||\mathbf{A} - \mathbf{A}_k||_2^2 \cdot \left|\left|\mathbf{I}_n\mathbf{S}(\mathbf{V}_k^T\mathbf{S})^+\right|\right|_2^2 \\
&\leq ||\mathbf{A} - \mathbf{A}_k||_2^2 \cdot ||\mathbf{I}_n\mathbf{S}||_2^2 \cdot \left|\left|(\mathbf{V}_k^T\mathbf{S})^+\right|\right|_2^2 \\
&\leq ||\mathbf{A} - \mathbf{A}_k||_2^2 \left[\frac{1 + \sqrt{n/r}}{1 - \sqrt{k/r}}\right]^2.
\end{aligned}$$

□

Note that the error in the previous theorem contains the same multiplicative factor of $O(\sqrt{n/r})$ that in Theorem 2.1.1 obtains in expectation. These factors are asymptotically optimal; we prove this by giving explicit matrices families which achieve this approximation ratio. As we take $\alpha \to 0$ in the forthcoming lemma, we require approximation error which approaches n/r.

Lemma 2.2.12. *For any $\alpha > 0$, and $k, r \geq 1$, there exists a matrix $\mathbf{A} \in \mathbb{R}^{(n+1)\times n}$ for which*

$$\left|\left|\mathbf{A} - \pi^2_{\mathbf{C},k}(\mathbf{A})\right|\right|_2 \geq \left|\left|\mathbf{A} - \mathbf{C}\mathbf{C}^+\mathbf{A}\right|\right|_2 \geq \sqrt{\frac{n+\alpha^2}{r+\alpha^2}}\left|\left|\mathbf{A} - \mathbf{A}_k\right|\right|_2,$$

where $\mathbf{C}$ is any matrix consisting of r columns of $\mathbf{A}$.

Proof. Consider the matrix $\mathbf{A} \in \mathbb{R}^{(n+1)\times n}$ given by

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & \cdots & 1 & \\ \alpha & 0 & \cdots & 0 & \\ 0 & \alpha & 0 & \cdots & 0 \\ & & & & \\ 0 & 0 & 0 & \cdots & \alpha \end{bmatrix},$$

for which $\mathbf{A}_{1i} = 1$ for $i \in [n]$, $\mathbf{A}_{i+1,i} = \alpha$, and the rest of the entries of A are zero. Then

$$\mathbf{A}^T\mathbf{A} = 1_n 1_n^T + \alpha^2\mathbf{I}_n, \quad \sigma_1^2(\mathbf{A}) = n + \alpha^2, \quad \sigma_i^2(\mathbf{A}) = \alpha^2 \text{ for } i > 1.$$

Since $k \geq 1$, we have that $||\mathbf{A} - \mathbf{A}_k||_2^2 = \sigma_{k+1}^2(\mathbf{A}) = \alpha^2$. Without loss of generality, by permuting rows of $\mathbf{A}$ we can assume that $\mathbf{C}$ consists of the first r columns of $\mathbf{A}$.

We will reconstruct $\mathbf{A}$ one column at a time. Let $\mathbf{a}_j$ be the jth column of $\mathbf{A}$. Then the reconstruction error on $\mathbf{a}_j$ is given by the minimum of $||\mathbf{a}_j - \mathbf{C}\mathbf{x}||_2^2$ over all $\mathbf{x} \in \mathbb{R}^r$. For $j \leq r$, the vector $\mathbf{e}_j$ yields zero error. For $j > r$, we can substitute $\mathbf{a}_j = \mathbf{e}_1 + \alpha\mathbf{e}_{j+1}$ to obtain

$$||\mathbf{a}_j - \mathbf{C}\mathbf{x}||_2^2 = \left|\left|\mathbf{e}_1\left(\sum_{i=1}^r x_i - 1\right) + \alpha\sum_{i=1}^r x_i\mathbf{e}_{i+1} - \mathbf{e}_{j+1}\right|\right|_2 = \left(\sum_{i=1}^r x_i - 1\right)^2 + \alpha^2\sum_{i=1}^r x_i^2 + 1.$$

The minimum of this expression must occur when all of the x_i are equal; assuming this and solving yields $x_i = 1/(r+\alpha^2)$ for all i. Let $\mathbf{X} \in \mathbb{R}^{r\times n}$ be the matrix whose first r columns are zero and which has the value $1/(r+\alpha^2)$ everywhere else. Then by Lemma 2.2.4,

$$||\mathbf{A} - \mathbf{C}\mathbf{C}^+\mathbf{A}||_2^2 = ||\mathbf{A} - \mathbf{C}\mathbf{X}||_2^2.$$

By a simple computation,

$$||\mathbf{A} - \mathbf{C}\mathbf{X}||_2^2 = \frac{n+\alpha^2}{r+\alpha^2}\alpha^2 = \frac{n+\alpha^2}{r+\alpha^2}\left|\left|\mathbf{A} - \mathbf{A}_k\right|\right|_2^2.$$

The first inequality in the lemma follows since the first quantity is the best approximation of rank k in the column space of $\mathbf{C}$, while the second approximation has no constraints on rank. □

2.3 Approximating the SVD

As a stepping stone for Theorem 2.3.1, we will prove the following randomized result, which finds a factorization of the form seen in Section 2.2.2 whose error is, in expectation, a constant multiple of the error obtained from approximating by $\mathbf{A}_k$.

Lemma 2.3.1 (Randomized approximate spectral SVD)**.** *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ, and let $1 < k < \rho$ be an integer. For $\epsilon \in (0,1)$, there exists an algorithm that computes a factorization $\mathbf{A} = \mathbf{A}\mathbf{X}\mathbf{X}^T + \mathbf{E}$ satisfying $\mathbf{X}^T\mathbf{X} = \mathbf{I}_k$ and $\mathbf{E}\mathbf{X} = \mathbf{0}$, such that*

$$\mathbb{E}[||\mathbf{E}||_2] \leq (\sqrt{2} + \epsilon)||\mathbf{A} - \mathbf{A}_k||_2.$$

Given this result, we can prove Theorem 2.1.1 by applying the same error-bounding techniques as in Theorem 2.2.11, but using the matrix $\mathbf{X}$ instead of $\mathbf{V}_k$.

Proof of 2.1.1. Let $\mathbf{S}$ be the sampling matrix obtained by applying Corollary 2.2.9 to the n orthonormal rows of the matrix $\mathbf{X}$ and the n orthonormal rows of the matrix $\mathbf{I}_n$, where $\mathbf{X}$ is obtained by running the algorithm of Lemma 2.3.1.

Following the proof of Theorem 2.2.11 with $\mathbf{A} - \mathbf{A}_k$ replaced by the error term $\mathbf{E}$ from the previous lemma and $\mathbf{V}_k$ replaced by $\mathbf{X}$, we obtain

$$||\mathbf{A} - \pi^2_{\mathbf{C},k}(\mathbf{A})||_2 \leq ||\mathbf{E}||_2 \frac{1 + \sqrt{n/r}}{1 - \sqrt{k/r}}.$$

The theorem follows from taking expectations and using the bound on $\mathbf{E}$ given in the previous lemma.

For the analysis of the runtime of this algorithm, see the proof of Theorem 1.3 in [8]. □

We will spend the rest of this section presenting the proof of Lemma 2.3.1, which parallels the analysis of the 'power scheme' algorithm of [15], which uses matrices of the form $(\mathbf{A}\mathbf{A}^T)^q\mathbf{A}\mathbf{\Pi}$ for a much smaller *standard Gaussian matrix* $\mathbf{\Pi}$ (a random matrix with entries i.i.d distributed in $\mathcal{N}(0,1)$) to obtain matrices which approximate the range of $\mathbf{A}$.

The following five lemmas are stated in [15]. These first two lemmas give bounds on the expectation of norms of Gaussian matrices and their pseudoinverses, and their proofs lie mostly out of the scope of this paper.

Lemma 2.3.2 (Proposition 10.1, [15])**.** *Let $\mathbf{X} \in \mathbb{R}^{m\times k}$, $\mathbf{Y} \in \mathbb{R}^{\ell\times n}$, and let $\mathbf{\Pi} \in \mathbb{R}^{k\times\ell}$ be a matrix with entries drawn i.i.d from $\mathcal{N}(0,1)$. Then*

$$\left(\mathbb{E}[||\mathbf{X}\mathbf{\Pi}\mathbf{Y}||_F^2]\right)^{1/2} = ||\mathbf{X}||_F||\mathbf{Y}||_F.$$

$$\mathbb{E}[||\mathbf{X}\mathbf{\Pi}\mathbf{Y}||_2] \leq ||\mathbf{X}||_2||\mathbf{Y}||_F + ||\mathbf{Y}||_2||\mathbf{X}||_F.$$

Proof. For the first inequality, we have:

$$\mathbb{E}[||\mathbf{X}\mathbf{\Pi}\mathbf{Y}||_F^2] = \sum_{a=1}^{m}\sum_{b=1}^{n}\mathbb{E}\left[\left(\sum_{i=1}^{k}\sum_{j=1}^{\ell}\mathbf{X}_{ai}\mathbf{\Pi}_{ij}\mathbf{Y}_{jb}\right)^2\right] = \sum_{a=1}^{m}\sum_{b=1}^{n}\sum_{i=1}^{k}\sum_{j=1}^{\ell}\mathbf{X}_{ai}^2\mathbf{Y}_{jb}^2 = ||\mathbf{X}||_F^2||\mathbf{Y}||_F^2.$$

For the second inequality, refer to [13]. □

Lemma 2.3.3 (Propositions A.4, A.5, A.6, [15])**.** *Let $k, p \geq 2$ be integers and let $\mathbf{\Pi} \in \mathbb{R}^{k\times(k+p)}$ be a matrix with entries drawn i.i.d. from $\mathcal{N}(0,1)$. Then*

$$\left(\mathbb{E}\left[||\mathbf{\Pi}^+||_F^2\right]\right)^{1/2} = \sqrt{k/(p-1)}$$

$$\mathbb{E}\left[||\mathbf{\Pi}^+||_2\right] \leq e(\sqrt{k+p})/p.$$

The next two lemmas deal with powers of spectral norms of projection matrices.

Lemma 2.3.4 (Proposition 8.5, [15]). *Let $\mathbf{Q} \in \mathbb{R}^{n\times n}$ be an symmetric projection, let $\mathbf{D} \in \mathbb{R}^{n\times n}$ be a nonnegative diagonal matrix, and let $t \geq 1$ be an integer. Then*

$$||\mathbf{QDQ}||_2^t \leq ||\mathbf{QD}^t\mathbf{Q}||_2^t.$$

Proof. Suppose $\mathbf{v} \in \mathbb{R}^n$ is a unit vector which satisfies

$$\mathbf{v}^T(\mathbf{QDQ})\mathbf{v} = ||\mathbf{QDQ}||_2. \tag{2.3}$$

Consider $\mathbf{w} = \mathbf{Qv}/||\mathbf{Qv}||_2$. Expanding,

$$||\mathbf{QDQ}||_2 \geq \mathbf{w}^T(\mathbf{QDQ})\mathbf{w} = \frac{\mathbf{v}^T\mathbf{Q}^T(\mathbf{QDQ})\mathbf{Qv}}{||\mathbf{Qv}||_2^2} = \frac{\mathbf{v}^T(\mathbf{QDQ})\mathbf{v}}{||\mathbf{Qv}||_2^2}.$$

For (2.3) to be an equality, we must have $||\mathbf{Qv}||_2 = 1$, so that $\mathbf{Qv} = \mathbf{v}$. Then

$$||\mathbf{QDQ}||_2^t = (\mathbf{v}^T(\mathbf{QDQ})\mathbf{v})^t = (\mathbf{v}^T\mathbf{Dv})^t = \left(\sum_j \mathbf{v}_j^2\mathbf{D}_{jj}\right)^t \leq \sum_j \mathbf{v}_j^2\mathbf{D}_{jj}^t = \mathbf{v}^T\mathbf{D}^t\mathbf{v} = (\mathbf{Qv})^T\mathbf{D}^t\mathbf{Qv} \leq ||\mathbf{QD}^t\mathbf{Q}||_2,$$

where the middle inequality follows by Jensen's inequality since $t \geq 1$ and $||\mathbf{v}||_2 = 1$. □

Lemma 2.3.5 (Proposition 8.5, [15]). *Let $\mathbf{P} \in \mathbb{R}^{n\times n}$ be an symmetric projection and let $\mathbf{A} \in \mathbb{R}^{n\times m}$. For any integer $q \geq 0$,*

$$||\mathbf{PA}||_2 \leq ||\mathbf{P}(\mathbf{AA}^T)^q\mathbf{A}||_2^{1/(2q+1)}.$$

Proof. Let $\mathbf{A} = \mathbf{U\Sigma V}^T$ be the SVD, and compute

$$||\mathbf{PA}||_2^{2(2q+1)} = ||\mathbf{PAA}^\mathbf{T}\mathbf{P}||_2^{2q+1} = ||\mathbf{PU\Sigma}^2\mathbf{U}^T\mathbf{P}||_2^{2q+1} = ||(\mathbf{U}^T\mathbf{PU})\mathbf{\Sigma}^2(\mathbf{U}^T\mathbf{PU})||_2^{2q+1},$$

where the last equality follows by unitary invariance. Note that $(\mathbf{U}^T\mathbf{PU})^2 = \mathbf{U}^T\mathbf{PU}$ is a symmetric projection, so by applying the previous lemma we obtain that

$$||(\mathbf{U}^T\mathbf{PU})\mathbf{\Sigma}^2(\mathbf{U}^T\mathbf{PU})||_2^{2q+1} \leq ||(\mathbf{U}^T\mathbf{PU})\mathbf{\Sigma}^{2(2q+1)}(\mathbf{U}^T\mathbf{PU})||_2 = ||\mathbf{PU\Sigma}^{2(2q+1)}\mathbf{U}^T\mathbf{P}||_2$$

by unitary invariance. We finish by noting that

$$||\mathbf{PU\Sigma}^{2(2q+1)}\mathbf{U}^T\mathbf{P}||_2 = ||\mathbf{P}(\mathbf{AA}^T)^{2(2q+1)}\mathbf{P}||_2 = ||\mathbf{P}(\mathbf{AA}^T)^q\mathbf{A}||_2^2.$$

□

We will also need that standard Gaussian matrices have full rank almost surely.

Lemma 2.3.6. *Let $k, \ell \geq 1$ be integers with $k \leq \ell$, and let $\mathbf{\Pi} \in \mathbb{R}^{k\times\ell}$ be a standard Gaussian matrix. Then* $\operatorname{rank}(\mathbf{\Pi}) = k$ *with probability 1.*

Proof. Let $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^\ell$ be the rows of $\mathbf{\Pi}$. For $1 \leq i \leq k$, conditioning on the Gaussian vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$, the vector space $V_i = \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1})$ has positive codimension and thus has measure zero. Since the distribution of the Gaussian vector $\mathbf{v}_i$ is absolutely continuous and independent of $\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}$, we have

$$\mathbb{P}[\mathbf{v}_i \in V_i | \mathbf{v}_1, \ldots, \mathbf{v}_{i-1}) = 0.$$

Integrating over all $\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}$, we obtain $\mathbb{P}[\mathbf{v}_i \in V_i] = 0$, so that union bounding over i yields the claim. □

The following lemma is similar to Corollary 10.10 in [15], which provides an analysis of the aforementioned 'power scheme' algorithm.

Lemma 2.3.7. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank ρ, and let $1 < k < \rho$ be an integer. Let $s \geq 2$ be an integer oversampling parameter, and let $r = k + s$. Let $\mathbf{\Pi} \in \mathbb{R}^{n\times r}$ be a matrix with entries drawn i.i.d from $\mathcal{N}(0,1)$. Let $q \geq 0$ be an integer, and let $\mathbf{B} = (\mathbf{A}\mathbf{A}^T)^q\mathbf{A}$, and let $\mathbf{Y} = \mathbf{B}\mathbf{\Pi}$. Then*

$$\mathbb{E}[||\mathbf{A} - \pi^2_{\mathbf{Y},k}(\mathbf{A})||_2] \leq \left(1 + \sqrt{k/(p-1)} + (e\sqrt{k+p}/p)\sqrt{\min(m,n)-k}\right)^{1/(2q+1)} ||\mathbf{A} - \mathbf{A}_k||_2.$$

Proof. Let $\pi^2_{\mathbf{Y},k}(\mathbf{A}) = \mathbf{Y}\mathbf{X}_A$ and $\pi^2_{\mathbf{Y},k}(\mathbf{B}) = \mathbf{Y}\mathbf{X}_B$, where $\mathbf{X}_A$ and $\mathbf{X}_B$ are the matrices in condition 2.1 for $\mathbf{A}$ and $\mathbf{B}$ respectively. Then by Lemma 2.2.4 we have that

$$||\mathbf{A} - (\mathbf{Y}\mathbf{X}_A)(\mathbf{Y}\mathbf{X}_A)^+\mathbf{A}||_2 \leq ||\mathbf{A} - (\mathbf{Y}\mathbf{X}_B)(\mathbf{Y}\mathbf{X}_B)^+\mathbf{A}||_2. \tag{2.4}$$

Write $\mathbf{A} - (\mathbf{Y}\mathbf{X}_B)(\mathbf{Y}\mathbf{X}_B)^+\mathbf{A} = (\mathbf{I} - (\mathbf{Y}\mathbf{X}_B)(\mathbf{Y}\mathbf{X}_B)^+)\mathbf{A}||_2$. Then $\mathbf{I} - (\mathbf{Y}\mathbf{X}_B)(\mathbf{Y}\mathbf{X}_B)^+$ is a symmetric projection, so that by Lemma 2.3.5, we obtain

$$||\mathbf{A}-(\mathbf{Y}\mathbf{X}_B)(\mathbf{Y}\mathbf{X}_B)^+\mathbf{A}||_2 \leq ||(\mathbf{I}-(\mathbf{Y}\mathbf{X}_B)(\mathbf{Y}\mathbf{X}_B)^+)(\mathbf{A}\mathbf{A}^T)^q\mathbf{A}||_2^{1/(2q+1)} = ||(\mathbf{I}-(\mathbf{Y}\mathbf{X}_B)(\mathbf{Y}\mathbf{X}_B)^+)\mathbf{B}||_2^{1/(2q+1)}.$$

Then combining another application of Lemma 2.2.4 with (2.4) yields

$$||\mathbf{A} - \pi^2_{\mathbf{Y},k}(\mathbf{A})||_2 \leq ||\mathbf{B} - \pi^2_{\mathbf{Y},k}(\mathbf{B})||_2^{1/(2q+1)}.$$

Taking expectations and using Holder's inequality, we get

$$\mathbb{E}[||\mathbf{A} - \pi^2_{\mathbf{Y},k}(\mathbf{A})||_2] \leq \left(\mathbb{E}[||\mathbf{B} - \pi^2_{\mathbf{Y},k}(\mathbf{B})||_2]\right)^{1/(2q+1)}.$$

Let $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the SVD of $\mathbf{B}$, and let $\rho' = \text{rank}(\mathbf{B})$. Consider

$$\mathbf{\Pi}_1 = \mathbf{V}_k^T\mathbf{\Pi} \text{ and } \mathbf{\Pi}_2 = \mathbf{V}_{\rho'-k}^T\mathbf{\Pi}.$$

Since Gaussian matrices are rotationally invariant, $\mathbf{V}^T\mathbf{\Pi}$ is also a standard Gaussian matrix. Thus, since $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ are non-intersecting submatrices of $\mathbf{V}^T\mathbf{\Pi}$, they are independent standard Gaussian matrices. In addition, by Lemma 2.3.6 $\mathbf{\Pi}_1$ has full rank k. Applying Lemma 2.2.7 and using $\sqrt{a^2+b^2} \leq a+b$, we obtain

$$\begin{aligned}\left|\left|\mathbf{B} - \pi^2_{\mathbf{Y},k}(\mathbf{B})\right|\right|_2 &\leq \left|\left|\mathbf{U}_{\rho'-k}\mathbf{\Sigma}_{\rho'-k}\mathbf{V}_{\rho'-k}^T\right|\right|_2 + \left|\left|\mathbf{U}_{\rho'-k}\mathbf{\Sigma}_{\rho'-k}\mathbf{V}_{\rho'-k}^T\Pi(\mathbf{V}_k^T\mathbf{\Pi})^+\right|\right|_2 \\ &= ||\mathbf{\Sigma}_{\rho'-k}||_2 + \left|\left|\mathbf{\Sigma}_{\rho'-k}\mathbf{\Pi}_2\mathbf{\Pi}_1^+\right|\right|_2.\end{aligned}$$

We will now take expectations with respect to $\mathbf{\Pi}_2$ and then $\mathbf{\Pi}_1$. By Lemma 2.3.2,

$$\mathop{\mathbb{E}}_{\mathbf{\Pi}_2}[\left|\left|\mathbf{\Sigma}_{\rho'-k}\mathbf{\Pi}_2\mathbf{\Pi}_1^+\right|\right|_2|\mathbf{\Pi}_1] = ||\mathbf{\Sigma}_{\rho'-k}||_2\left|\left|\mathbf{\Pi}_1^+\right|\right|_F + \left|\left|\mathbf{\Pi}_1^+\right|\right|_2||\mathbf{\Sigma}_{\rho'-k}||_F.$$

Note that $||\mathbf{\Sigma}_{\rho'-k}||_F \leq \sqrt{\min(m,n)-k}||\mathbf{\Sigma}_{\rho'-k}||_2$. Next, by Lemma 2.3.3, we have the two bounds

$$\mathbb{E}[||\mathbf{\Pi}_1^+||_F] \leq \mathbb{E}[||\mathbf{\Pi}_1^+||_F^2]^{1/2} \leq \sqrt{k/(p-1)}, \text{ and } \mathbb{E}[||\mathbf{\Pi}_1^+||_2] = e\sqrt{k+p}/p.$$

Thus

$$\mathbb{E}[\left|\left|\mathbf{B} - \pi^2_{\mathbf{Y},k}(\mathbf{B})\right|\right|_2] \leq \left(1 + \sqrt{k/(p-1)} + e\sqrt{k+p}/p\sqrt{\min(m,n)-k}\right)||\mathbf{\Sigma}_{\rho'-k}||_2.$$

Finally, expanding the definition of $\mathbf{A}$, note that $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}^{2q+1}\mathbf{V}^T$ is the SVD of A, so that

$$||\mathbf{\Sigma}_{\rho'_k}||_2 = ||\mathbf{B} - \mathbf{B}_k||_2 = ||\mathbf{A} - \mathbf{A}_k||_2^{2q+1}.$$

□

Proof of Lemma 2.3.1. Let $\mathbf{Y}$ be as in the previous lemma. It is easy to find $q \geq 0$ such that the constant of the the inequality in Lemma 2.3.7 is bounded above by $1 + \epsilon/\sqrt{2}$, so that

$$\mathbb{E}[||\mathbf{A} - \pi^2_{\mathbf{Y},k}(\mathbf{A})||_2] \leq (1 + \epsilon/\sqrt{2})||\mathbf{A} - \mathbf{A}_k||_2.$$

Let $\mathbf{O}$ be an orthonormal basis for the column space of $\mathbf{Y}$. By taking square roots in Lemma 2.2.6 we have that

$$||\mathbf{A} - \mathbf{O}(\mathbf{O}^T\mathbf{A})_k||_2 \leq \sqrt{2}||\mathbf{A} - \pi^2_{\mathbf{Y},k}(\mathbf{A})||_2.$$

Let $\mathbf{X}$ consist of the right singular vectors of the matrix $(\mathbf{O}^T\mathbf{A})_k$, and let $\mathbf{E} = \mathbf{A} - \mathbf{A}\mathbf{X}\mathbf{X}^T$. We can write $\mathbf{O}(\mathbf{O}^T\mathbf{A})_k = \mathbf{C}\mathbf{X}^T$ for some matrix $\mathbf{C}$. But since $\mathbf{X}^T$ has orthonormal columns, $\mathbf{A}\mathbf{X}\mathbf{X}^T$ is the best approximation to $\mathbf{A}$ in the column space of $\mathbf{X}$, so that

$$||\mathbf{E}||_2 = ||\mathbf{A} - \mathbf{A}\mathbf{X}\mathbf{X}^T||_2 \leq ||\mathbf{A} - \mathbf{C}\mathbf{X}^T|| \leq \sqrt{2}||\mathbf{A} - \pi^2_{\mathbf{Y},k}(\mathbf{A})||_2.$$

□

3 | Applications to Covariance Estimation

3.1 Introduction

In this chapter, we will follow the analyses of [32] and [26], which study the problem of how many times one needs to sample a random matrix before one obtains a good spectral estimator for the matrix. More concretely, let $\mathbf{X} \in \mathbb{R}^{n \times n}$ be a positive semidefinite random matrix. In order to estimate $\mathbb{E}\,\mathbf{X}$, one can use the unbiased estimator $\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i$, where the $\mathbf{X}_i$ are N independent samples of $\mathbf{X}$. If we measure the goodness of this estimator by its spectral error

$$\left|\left| \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i - \mathbb{E}\,\mathbf{X} \right|\right|_2,$$

how many of these samples do we need to get small spectral error ϵ? By dimensional reasons, one must have that $N \geq n$. However, are there general classes of distributions for which this bound is on the order of $N = O(n)$, disregarding factors of ϵ?

Note that if we take $\mathbf{X} = \mathbf{v}\mathbf{v}^T$ for a random vector $\mathbf{v}$ of finite variance valued in $\mathbb{R}^n$ such that $\mathbb{E}\,\mathbf{v} = \mathbf{0}$, then $\mathbb{E}\,\mathbf{X}$ is exactly the *covariance matrix* of $\mathbf{v}$. As such, this problem is a generalization of the problem of covariance estimation which was studied for isotropic vectors with bounded spectral norm in [23], obtaining $N = O(n \log n)$, and for sub-exponential distributions in [1], obtaining $N = O(n)$ and answering our question in the affirmative. For an overview of further results concerning covariance estimation, we refer the reader to [29].

Returning to the matrix setting, the following theorem proven in [32] holds for a much wider class of distributions which includes the sub-exponential distributions. It shows that any distribution of symmetric positive semidefinite matrices with *finite* $(2+\epsilon)$*-moments* needs only $N = O(n)$ samples, and its proof consists of a randomization of the barrier method of Chapter 1, yielding what the authors of [26] claim to be a novel proof method in random matrix theory.

Notation In what follows, for random variables X with finite pth moment, we will denote the L_p norm $(\mathbb{E}[|X|^p])^{1/p}$ by $||X||_p$. For a matrix $\mathbf{A}$, we will denote its spectral (operator) norm by $||\mathbf{A}||$, dropping the the usual subscript to avoid confusion. For a variable x, we will adopt the notation $C(x)$ to represent the value and existence of suitable constants dependent only on x.

Theorem 3.1.1. *Let $\mathbf{X} \in \mathbb{R}^{n \times n}$ be symmetric and positive semidefinite, and let $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ be its eigendecomposition, where $\mathbf{U} \in \mathbb{R}^{n \times n}$ is orthogonal and $\mathbf{D} \in \mathbb{R}^{n \times n}$ is diagonal. Suppose that $\mathbf{U}$ and $\mathbf{D}$ are independent, and that the diagonal entries $\alpha_i = \mathbf{D}_{ii}$ are independent. Suppose further that for some $p > 2$ we have for all $i \in [n]$ that*

$$\mathbb{E}\,\alpha_i = 1 \text{ and } ||\alpha_i||_p \leq C(p).$$

Then for $\epsilon \in (0,1)$ and

$$N \geq C(p)\epsilon^{-2p/(p-2)} \cdot n,$$

one has

$$\mathbb{E}\left|\left| \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i - \mathbf{I}_n \right|\right| \leq \epsilon.$$

3.2 Regularity Conditions

There is an even wider family of distributions for which $N = O(n)$ samples suffice: those which satisfy the following strong regularity condition.

Definition 3.2.1. Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ be a random matrix. For $\eta > 0$, we say that $\mathbf{X}$ satisfies the *strong regularity condition* ($\mathtt{SR}_\eta$) *with constant* $C_\eta > 0$ if

$$\mathbb{P}\left(||\mathbf{PXP}|| \geq t\right) \leq C_\eta t^{-1-\eta} \text{ for all } t > C_\eta \operatorname{rank}(\mathbf{P}), \text{ for all orthogonal projections } \mathbf{P} \in \mathbb{R}^{n\times n}.$$

We show that distributions with finite $(2+\epsilon)$ moments satisfy the strong regularity condition. We will need Rosenthal's inequality [22] for sums of symmetric random variables:

Theorem 3.2.2 (Rosenthal's inequality)**.** *Let $X_1, \ldots, X_n$ be independent symmetric random variables. Then for every $2 \leq p < \infty$,*

$$M_p \leq \left|\left|\sum X_i\right|\right|_p \leq C(p)M_p,$$

where

$$M_p = \max\left[\left(\sum ||X_i||_2^2\right)^{1/2}, \left(\sum ||X_i||_2^2\right)^{1/p}\right].$$

Lemma 3.2.3. *Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ be symmetric and positive semidefinite, and let $\mathbf{X} = \mathbf{UDU}^T$ be its eigendecomposition, where $\mathbf{U} \in \mathbb{R}^{n\times n}$ is orthogonal and $\mathbf{D} \in \mathbb{R}^{n\times n}$ is diagonal. Suppose that $\mathbf{U}$ and $\mathbf{D}$ are independent, and that the diagonal entries $\alpha_i = \mathbf{D}_{ii}$ are independent. Then if for some $p > 2$ and constant $C > 0$ we have that for all $i \in [n]$,*

$$\mathbb{E}\,\alpha_i = 1 \text{ and } ||\alpha_i||_p \leq C,$$

then $\mathbf{X}$ satisfies $\mathtt{SR}_{p/2-1}$.

Proof. We will show the following property for $\eta = p/2 - 1$:

$$\mathbb{P}\left(\operatorname{Tr}[\mathbf{PX}] \geq t\right) \leq Ct^{-1-\eta} \text{ for all } t > C_\eta \operatorname{rank}(\mathbf{P}), \text{ for all orthogonal projections } \mathbf{P} \in \mathbb{R}^{n\times n}. \tag{3.1}$$

Note that this implies $\mathtt{SR}_\eta$ since $\operatorname{Tr}[\mathbf{PX}] = \operatorname{Tr}[\mathbf{PXP}] \geq ||\mathbf{PXP}||$ by cyclic properties of trace. In addition, since the trace is invariant under changes of basis, it suffices to assume that $\mathbf{U} = \mathbf{I}_n$, so that $\mathbf{X} = \mathbf{D}$.

Now, let $\mathbf{P} \in \mathbb{R}^{n\times n}$ be an orthogonal projection of rank k, so that $\operatorname{Tr}\mathbf{P} = k$ and $|\mathbf{P}_{ii}| \leq 1$ for each i. Then we can write

$$\mathbb{P}\left(\operatorname{Tr}[\mathbf{PD}] \geq t\right) = \mathbb{P}\left(\operatorname{Tr}[\mathbf{P}(\mathbf{D}-\mathbf{I})] \geq t-k\right) \leq \mathbb{P}\left(|\operatorname{Tr}[\mathbf{P}(\mathbf{D}-\mathbf{I})]| \geq t-k\right) \leq (t-k)^{-p}\,||\operatorname{Tr}[\mathbf{P}(\mathbf{D}-\mathbf{I})]|^p]||_p^p,$$

where the last inequality follows by Markov's inequality. We can write $\operatorname{Tr}[\mathbf{P}(\mathbf{D}-\mathbf{I})] = \sum \mathbf{P}_{ii}(\alpha_i - 1)$, where the $\mathbf{P}_{ii}(\alpha_i - 1)$ are independent symmetric random variables. Applying Rosenthal's inequality, we obtain that

$$||\operatorname{Tr}[\mathbf{P}(\mathbf{D}-\mathbf{I})]|^p]||_p^p \leq \max\left[\left(\sum \mathbf{P}_{ii}^2\,\mathbb{E}\,|\alpha_i - 1|^2\right)^{p/2}, \sum |\mathbf{P}_{ii}^p|\,\mathbb{E}\,|\alpha_i - 1|^p\right].$$

Using $|\mathbf{P}_{ii}| \leq 1$ and the moment condition of the α_i, we obtain that

$$||\operatorname{Tr}[\mathbf{P}(\mathbf{D}-\mathbf{I})]|^p]||_p^p \leq C(p)k^{p/2},$$

so that $\mathbb{P}\left(\operatorname{Tr}[\mathbf{PD}] \geq t\right) \leq (t-k)^{-p}k^{p/2}$. Setting $k = Dt$ for suitable constant D, we obtain property (3.1). □

As a result, the following theorem implies Theorem 3.1.1.

Theorem 3.2.4. *Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ be a positive semidefinite random matrix with $\mathbb{E}\,\mathbf{X} = 1$. Assume that X satisfies $\mathtt{SR}_\eta$. Then, for $\epsilon \in (0,1)$ and for*

$$N \geq C(\eta)\epsilon^{-2-2/\eta} \cdot n,$$

one has

$$\mathbb{E}\left|\left|\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i - \mathbf{I}_n\right|\right| \leq \epsilon.$$

It is also possible to matrices sampled from the log-concave distributions of [1] satisfy SR; see section 8 of [32] for a reference.

The strong regularity condition is not necessary for the forthcoming statement of Proposition 3.3.1, and can be replaced with a weaker regularity condition which deals with moments of certain inner products.

Definition 3.2.5. Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ be a random matrix. For $p > 1$, we say that $\mathbf{X}$ satisfies the *weak regularity condition* ($\mathtt{WR}_p$) if

$$||\langle \mathbf{X}\mathbf{u}, \mathbf{u}\rangle||_p \leq C(p) \text{ for all unit vectors } \mathbf{u} \in \mathbb{R}^n\ .$$

The following lemma shows that SR is indeed stronger than WR.

Lemma 3.2.6. *Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ be a symmetric positive definite random matrix satisfying $\mathtt{SR}_\eta$ for some $\eta > 0$. Then $\mathbf{X}$ also satisfies $\mathtt{WR}_p$ for $p = 1 + \eta > 1$.*

Proof. Let $\mathbf{u} \in \mathbb{R}^n$ be a unit vector, and let $\mathbf{P}$ be the rank one orthogonal projection onto the span of $\mathbf{u}$. By $\mathtt{SR}_\eta$, $\mathbb{P}(||\mathbf{PXP}|| \geq t) \leq C_\eta t^{-1-\eta}$ for all $t > C_\eta$. By definition of $\mathbf{P}$, we have that $||\mathbf{PXP}|| = \langle \mathbf{X}\mathbf{u}, \mathbf{u}\rangle$, and the lemma follows by an integration of tails. □

For explanations of the optimality of these regularity conditions, see Section 1.8 of [26].

3.3 Randomizing the Barrier Method

We rely on a randomized version of the barrier method to control the upper and lower eigenvalues of the sample matrices $\mathbf{A}_N = \sum_{i=1}^{N}\mathbf{X}_i$. Since we no longer have control over the magnitude or value of our matrix updates, since each update is an independent random variable $\mathbf{X}_i$ drawn from a given distribution, we will instead take our upper and lower shifts to be random variables and bound how these shifts behave in expectation as the number of samples N increases.

We will control the the expectations of the upper and lower eigenvalues through two separate theorems. Because of this, we will not need a lemma analogous to Lemma 1.2.6. We reiterate that we only need the weaker regularity constraint WR to obtain the necessary bounds for the minimum eigenvalue, while we need the strong regularity constraint SR to control the maximum eigenvalue.

Optimality of Regularity constraints

Theorem 3.3.1 (Expectation of minimum eigenvalue)**.** *Let $\mathbf{X}_i \in \mathbb{R}^{n\times n}$ be independent positive semidefinite random matrices satisfying $\mathtt{WR}_p$ with $\mathbb{E}\,\mathbf{X}_i = \mathbf{I}_n$, and let $\epsilon \in (0,1)$. Then for*

$$N \geq C(p)\cdot n \cdot \epsilon^{-\frac{p-1}{2p-1}},$$

one has

$$\mathbb{E}\,\lambda_{\min}\left(\sum_{i=1}^{N}\mathbf{X}_i\right) \geq 1 - \epsilon.$$

Theorem 3.3.2 (Expectation of maximum eigenvalue)**.** *Let $\mathbf{X}_i \in \mathbb{R}^{n\times n}$ be independent positive semidefinite random matrices satisfying $\mathtt{SR}_\eta$ with $\mathbb{E}\,\mathbf{X}_i = \mathbf{I}_n$, and let $\epsilon \in (0,1)$. Then for*

$$N \geq C(\eta)\cdot n \cdot \epsilon^{-2-2/\eta},$$

one has

$$\mathbb{E}\,\lambda_{\max}\left(\sum_{i=1}^N \mathbf{X}_i\right) \leq 1+\epsilon.$$

We will also need the following Chernoff-type bound:

Lemma 3.3.3. *Let $p \in (1,2]$ and suppose $X_1,\ldots,X_N$ are independent (real-valued) positive random variables with $\mathbb{E}\,X_i = 1$ and the moment bound $||X_i||_p \leq C(p)$. Then*

$$\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N X_i - 1\right|\right] \leq C(p)N^{-(p-1)/p}.$$

Proof. Define centered random variables $Z_i = (X_i-1)/N$. Then we have:

$$\mathbb{E}\left|\sum Z_i\right| = \mathbb{E}\left|\sum (Z_i - \mathop{\mathbb{E}}_{Z_i'} Z_i')\right|,$$

where Z_i' is an independent copy of Z_i. By Jensen's inequality and symmetrization for ϵ_i uniform independent ± 1 Bernoullis,

$$\mathbb{E}\left|\sum Z_i\right| \leq \mathop{\mathbb{E}}_{Z_i,Z_i'}\left|\sum (Z_i - Z_i')\right| = \mathop{\mathbb{E}}_{Z_i,Z_i',\epsilon_i}\left|\sum(\epsilon_i Z_i - \epsilon_i Z_i')\right| \leq \frac{2}{N}\mathop{\mathbb{E}}_{X_i,\epsilon_i}\left|\sum \epsilon_i X_i\right|.$$

By Cauchy-Schwarz and two more applications of Jensen's inequality we have

$$\frac{2}{N}\mathop{\mathbb{E}}_{X_i,\epsilon_i}\left|\sum \epsilon_i X_i\right| \leq \frac{2}{N}\mathbb{E}\left[\sum X_i^2\right]^{1/2} \leq \frac{2}{N}\mathbb{E}\left[\sum X_i^r\right]^{1/r} \leq \frac{2}{N}\left[\sum \mathbb{E}\,X_i^r\right]^{1/r} \leq C(p)N^{-(p-1)/p}.$$

□

These three results together yield a proof of Theorem 3.2.4.

Proof of Theorem 3.2.4. Let $p = 1+\eta/2$, so that $\mathbf{X}$ satisfies $\mathtt{WR}_p$ and $\mathtt{SR}_\eta$, and choose N greater than the constraints in Theorems 3.3.1 and 3.3.2. Let $\mathbf{Y}$ be the average of the N samples $\mathbf{X}_i$, given by

$$\mathbf{Y} = \frac{1}{N}\sum_{i=1}^N \mathbf{X}_i.$$

We may define the random variables

$$Y \stackrel{\text{def}}{=} ||\mathbf{Y} - \mathbf{I}_n|| \leq ||\mathbf{Y} - (\mathrm{Tr}[\mathbf{Y}]/n)\mathbf{I}_n|| + ||(\mathrm{Tr}[\mathbf{Y}]/n)\mathbf{I}_n - \mathbf{I}_n|| \stackrel{\text{def}}{=} Y_1 + Y_2,$$

where the middle inequality follows by the triangle inequality. We will bound these terms separately. First, note that

$$Y_1 = \lambda_{\max}\left[\mathbf{Y} - (\mathrm{Tr}[\mathbf{Y}]/n)\mathbf{I}_n\right] = \max\left[\lambda_{\max}(\mathbf{Y}) - \mathrm{Tr}[\mathbf{Y}]/n, \mathrm{Tr}[\mathbf{Y}]/n - \lambda_{\min}(\mathbf{Y})\right] \leq \lambda_{\max}(\mathbf{Y}) - \lambda_{\min}(\mathbf{Y}),$$

where the last inequality follows since both terms inside the max are positive and thus their sum is an upper bound for their max. Taking expectations and using the bounds of Theorems 3.3.1 and 3.3.2 yields that $\mathbb{E}\,Y_1 \leq 2\epsilon$.

Next, define scalar random variables $X_i = \operatorname{Tr}[\mathbf{X}_i]/n$, so that

$$Y_2 = |\operatorname{Tr}(\mathbf{Y})/n - 1| = \left| \frac{1}{N} \sum_{i=1}^{N} X_i - 1 \right|.$$

Furthermore, we have that

$$||X_i||_p \leq \frac{1}{n} \sum_{j=1}^{n} ||\langle \mathbf{X}_i \mathbf{e}_j, \mathbf{e}_j \rangle||_p \leq C(\eta)$$

by $\mathtt{WR}_p$ with $p = 1 + \eta/2$, so that by Lemma 3.3.3 with parameter $p \leftarrow \min(p, 2)$ and the bound on N given by Theorem 3.3.1, we have $\mathbb{E}\, Y_2 = \epsilon$. Thus $\mathbb{E}\, Y \leq 3\epsilon$, and we are done by rescaling ϵ. □

3.3.1 The Minimum Eigenvalue

We use the notation of Chapter 2 for the barrier potential function

$$\Phi_\ell(\mathbf{A}) = \operatorname{Tr}[(\mathbf{A} - \ell \mathbf{I})^{-1}]$$

and its upper bound $\epsilon_L \geq \Phi_\ell(\mathbf{A})$. The following proposition implies Theorem 3.3.1.

Proposition 3.3.4. *Suppose $\ell < \lambda_{\min}(\mathbf{A})$ and let $\mathbf{X}$ be a positive semidefinite random matrix satisfying $\mathtt{WR}_p$ for some $p > 1$ such that $\mathbb{E}\, \mathbf{X} = \mathbf{I}_n$. Then if $\epsilon \in (0, 1)$ and*

$$\Phi_\ell(\mathbf{A}) \leq \epsilon_L = C(p) \cdot \epsilon^{p/(p-1)},$$

then there exists a shift random variable, δ, which is dependent on $\mathbf{X}$ and satisfies

$$\ell + \delta < \lambda_{\min}(\mathbf{A} + \mathbf{X}), \quad \Phi_{\ell+\delta}(\mathbf{A} + \mathbf{X}) \leq \Phi_\ell(\mathbf{A}), \quad \mathbb{E}[\delta] \geq 1 - \epsilon. \tag{3.2}$$

Proof of Theorem 3.3.1. We begin with matrix $\mathbf{A}_0 = 0$ and deterministic lower barrier $\ell_0 = -n/\epsilon_L$. For each $t \in [N]$, let $\mathbf{A}_t = \mathbf{A}_{t-1} + \mathbf{X}_t$. We apply Proposition 3.3.4 inductively: suppose that we have obtained random shifts $\delta_1, \ldots, \delta_t$ after summing samples $\mathbf{X}_1, \ldots, \mathbf{X}_t$ to obtain a matrix $\mathbf{A}_t$. Conditioning on these variables, we can apply the proposition with $\mathbf{A} = \mathbf{A}_t$ to obtain a new shift δ_{t+1} satisfying the conditions of (3.2). After N steps, we have

$$\mathop{\mathbb{E}}_{\delta_i} \sum_{i=1}^{N} \delta_i \geq N(1 - \epsilon),$$

and thus $\mathbb{E}\, \lambda_{\min}(\mathbf{A}_n/N) \geq 1/N \cdot \left(\mathbb{E}_{\delta_i} \sum_{i=1}^{N} \delta_i \right) \geq 1 - \epsilon$. □

Aside from the random shift, the conditions of this proposition are almost identical to those in the lower barrier shift of Lemma 1.2.5 in Chapter 1. However, we are not given a constant shift δ_L and must instead find a suitable shift δ. We will give an explicit formula for such a shift and then prove that this choice of shift is large enough, in the sense that $\mathbb{E}\, \delta \geq 1 - \epsilon$.

Consider the quantity

$$L'_{\mathbf{A}}(\delta, \mathbf{X}) = \frac{1}{\delta} \frac{\operatorname{Tr}[\mathbf{N}^{-2}\mathbf{X}]}{\operatorname{Tr}[\mathbf{N}^{-2}]} - \operatorname{Tr}[\mathbf{N}^{-1}\mathbf{X}] \stackrel{\text{def}}{=} \frac{1}{\delta} F_{\mathbf{A}}(\delta, \mathbf{X}) - G_{\mathbf{A}}(\delta, \mathbf{X}).$$

Then since $\delta \operatorname{Tr}[\mathbf{N}^{-2}] \leq \Phi_{\ell+\delta}(\mathbf{A}) - \Phi_\ell(\mathbf{A})$, Lemma 1.2.5 immediately implies the following corollary:

Corollary 3.3.5. *Let $\ell \in \mathbb{R}$ and $\delta > 0$. Suppose $\ell + \delta < \lambda_{\min}(\mathbf{A})$. Then if $L'_{\mathbf{A}}(\delta, \mathbf{X}) \geq 1$, then*

$$\Phi_{\ell+\delta}(\mathbf{A} + \mathbf{X}) \leq \Phi_\ell(\mathbf{A}).$$

We now enumerate some properties of the quantities $F_{\mathbf{A}}(\delta, \mathbf{X})$ and $G_{\mathbf{A}}(\delta, \mathbf{X})$ which will be useful later.

Lemma 3.3.6 (Moments of $F_{\mathbf{A}}$ and $G_{\mathbf{A}}$)**.** *Suppose $\mathbf{X} \in \mathbb{R}^{n \times n}$ satisfies $\texttt{WR}_p$ for some $p > 1$ and $\mathbb{E}\,\mathbf{X} = \mathbf{I}_n$. Then the following moment bounds hold:*

(a) $\mathbb{E}\, F_{\mathbf{A}}(0, \mathbf{X}) = 1$ *and* $\mathbb{E}[F_{\mathbf{A}}(0, \mathbf{X})^p] \le C_p$.

(b) $\mathbb{E}\, G_{\mathbf{A}}(0, \mathbf{X}) = \Phi_\ell(\mathbf{A}) \le \epsilon_L$ *and* $\mathbb{E}[G_{\mathbf{A}}(0, \mathbf{X})^p] \le C_p \epsilon_L^p$.

(c) $\mathbb{P}[F_{\mathbf{A}}(0, \mathbf{X}) \ge t] \le C_p t^{-p}$ *and* $\mathbb{P}[G_{\mathbf{A}}(0, \mathbf{X}) \ge t] \le C_p \epsilon_L^p t^{-p}$.

Proof. Since trace is linear and $\mathbb{E}\,\mathbf{X} = \mathbf{I}_n$, it is immediate that $\mathbb{E}\, F_{\mathbf{A}}(0, \mathbf{X}) = 1$ and $\mathbb{E}\, G_{\mathbf{A}}(0, \mathbf{X}) = \mathrm{Tr}[(\mathbf{A} - \ell\mathbf{I})^{-1}] = \Phi_\ell(\mathbf{A})$. Property WR implies that

$$||F_{\mathbf{A}}(0, \mathbf{X})||_p = \left|\left| \frac{\sum_{i=1}^n \langle \mathbf{X}\mathbf{u}_i, \mathbf{u}_i \rangle (\lambda_i - \ell)^{-2}}{\sum_{i=1}^n (\lambda_i - \ell)^{-2}} \right|\right|_p \le \frac{\sum_{i=1}^n ||\langle \mathbf{X}\mathbf{u}_i, \mathbf{u}_i \rangle||_p (\lambda_i - \ell)^{-2}}{\sum_{i=1}^n (\lambda_i - \ell)^{-2}} \le C_p^{1/p}.$$

$$||G_{\mathbf{A}}(0, \mathbf{X})||_p = \left|\left| \sum_{i=1}^n \frac{\langle \mathbf{X}\mathbf{u}_i, \mathbf{u}_i \rangle}{\lambda_i - \ell} \right|\right|_p \le \sum_{i=1}^n \frac{||\langle \mathbf{X}\mathbf{u}_i, \mathbf{u}_i \rangle||_p}{\lambda_i - \ell} \le \sum_{i=1}^n \frac{||\langle \mathbf{X}\mathbf{u}_i, \mathbf{u}_i \rangle||_p}{\lambda_i - \ell} \le C_p^{1/p} \Phi_\ell(\mathbf{A}) \le C_p^{1/p} \epsilon_L.$$

Part (c) of the proposition follows from the above and Markov's inequality. □

Using these properties, the next two lemmas specify a good lower shift and calculate its expectation.

Lemma 3.3.7 (Explicit lower shift)**.** *Let $t \in (0, 1)$. Let*

$$\delta = \begin{cases} (1-t)^3 F_{\mathbf{A}}(0, \mathbf{X}) & \text{if } F_{\mathbf{A}}(0, \mathbf{X}) \le t/\epsilon_L \text{ and } G_{\mathbf{A}}(0, \mathbf{X}) \le t, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\ell + \delta < \lambda_{\min}(\mathbf{A})$ and $\Phi_{\ell+\delta}(\mathbf{A} + \mathbf{X}) \le \Phi_\ell(\mathbf{A})$.

Proof. If $\delta = 0$, then the lemma is immediate. Thus, suppose that $F_{\mathbf{A}}(0, \mathbf{X}) \le t/\epsilon_U$ and $G_{\mathbf{A}}(0, \mathbf{X}) \le t$. But then $\delta = (1-t)^3 F_{\mathbf{A}}(0, \mathbf{X}) \le (1-t)^3 t/\epsilon_L < 1/\epsilon_L$, and thus $0 < 1 - \delta\epsilon_L < 1$, so that

$$\begin{aligned} \frac{1}{\delta} F_{\mathbf{A}}(\delta, \mathbf{X}) - G_{\mathbf{A}}(\delta, \mathbf{X}) &= \frac{1}{(1-t)^3 F_{\mathbf{A}}(0, \mathbf{X})} F_{\mathbf{A}}(\delta, \mathbf{X}) - G_{\mathbf{A}}(\delta, \mathbf{X}) \\ &\ge \frac{(1 - \delta\epsilon_L)^2}{(1-t)^3} - (1 - \delta\epsilon_L)^{-1} G_{\mathbf{A}}(0, B) \\ &\ge \frac{(1-t)^2}{(1-t)^3} - \frac{t}{1-t} = 1. \end{aligned}$$

□

Lemma 3.3.8 (Expectation of random lower shift δ)**.** $\mathbb{E}\,\delta \ge (1-t)^3(1 - 2C_p \epsilon_L^{p-1} t^{1-p})$.

Proof. Define a characteristic random variable as follows:

$$\chi_{\mathbf{A}}(\mathbf{X}) = \begin{cases} 0 & \text{if } F_{\mathbf{A}}(0, \mathbf{X}) \le t/\epsilon_L \text{ and } G_{\mathbf{A}}(0, \mathbf{X}) \le t, \\ 1 & \text{otherwise.} \end{cases}$$

Then $\delta = (1-t)^3 F_{\mathbf{A}}(0, \mathbf{X})(1 - \chi_{\mathbf{A}}(\mathbf{X}))$, and

$$||\chi_{\mathbf{A}}(\mathbf{X})||_q = (\mathbb{E}\,\chi_{\mathbf{A}}(\mathbf{X}))^{1/q} = (\mathbb{P}\,[F_{\mathbf{A}}(0, \mathbf{X}) > t/\epsilon_L \text{ or } G_{\mathbf{A}}(0, \mathbf{X}) > t])^{1/q} \le \left(2C_p \epsilon_L^p t^{-p}\right)^{1/q}, \tag{3.3}$$

where the last inequality follows by a union bound and Markov's inequality. Let q be such that $1/p+1/q = 1$. We then have the following series of inequalities:

$$\begin{aligned}
\mathbb{E}\,\delta &= (1-t)^3\left(\mathbb{E}\,F_{\mathbf{A}}(0,\mathbf{X}) - \mathbb{E}[F_{\mathbf{A}}(0,\mathbf{X})\cdot\chi_{\mathbf{A}}(\mathbf{X})]\right) \\
&= (1-t)^3\left(1 - \mathbb{E}[F_{\mathbf{A}}(0,\mathbf{X})\cdot\chi_{\mathbf{A}}(\mathbf{X})]\right) && \text{(By Lemma 3.3.6)} \\
&\geq (1-t)^3\left(1 - ||F_{\mathbf{A}}(0,\mathbf{X})||_p \cdot ||\chi_{\mathbf{A}}(\mathbf{X})||_q]\right) && \text{(By Holder's inequality)} \\
&\geq (1-t)^3\left(1 - ||F_{\mathbf{A}}(0,\mathbf{X})||_p \cdot \left(2C_p^{1/q}\epsilon_L^{p-1}t^{1-p}\right)\right) && \text{(By (3.3))} \\
&= (1-t)^3\left(1 - C_p^{1/p}\cdot\left(2C_p^{1/q}\epsilon_L^{p-1}t^{1-p}\right)\right) && \text{(By Lemma 3.3.6)} \\
&= (1-t)^3(1-2C_p\epsilon_L^{p-1}t^{1-p}).
\end{aligned}$$

□

We are now ready to prove Proposition 3.3.4, which completes our analysis of the minimum eigenvalue.

Proof of Proposition 3.3.4. Take $t = \epsilon/4$. Recalling the assumption $\epsilon_L = C(p)\epsilon^{p/(p-1)}$, so that $(1 - 2C(p)\epsilon_L^{p-1}t^{1-p} = 1 - O(\epsilon)$, we obtain $\mathbb{E}\,\delta \geq 1-\epsilon$ by suitable choice of constants. □

3.3.2 The Maximum Eigenvalue

Recall the upper barrier potential

$$\Phi^u(\mathbf{A}) = \mathrm{Tr}[(u\mathbf{I} - \mathbf{A})^{-1}]$$

and its upper bound $\epsilon_U \geq \Phi^u(\mathbf{A})$. The following proposition implies Theorem 3.3.2.

Proposition 3.3.9. *Suppose $\lambda_{\max}(\mathbf{A}) < u$, let $\mathbf{X}$ be a positive semidefinite random matrix satisfying $\texttt{MSR}_\eta$ for some $\eta > 0$ such that $\mathbb{E}\,\mathbf{X} = \mathbf{I}_n$, and let $\epsilon \in (0,1)$. If*

$$\Phi^u(\mathbf{A}) \leq \epsilon_U = C(\eta)\cdot\epsilon^{1+2/\eta},$$

there exists a shift random variable, δ, which is dependent on $\mathbf{X}$ and satisfies

$$\lambda_{\max}(\mathbf{A}+\mathbf{X}) < u, \quad \Phi^{u+\delta}(\mathbf{A}+\mathbf{X}) \leq \Phi^u(\mathbf{A}), \quad \mathbb{E}\,\delta \leq 1+\epsilon.$$

Proof of Theorem 3.3.2. The proof is once again inductive and is extremely similar to that of Theorem 3.3.1, so we omit it here. □

Our approach will be similar to that for the minimum eigenvalue. We recall the definition

$$U^{\mathbf{A}}(\delta,\mathbf{B}) = \frac{\mathrm{Tr}[\mathbf{M}^{-2}\mathbf{B}]}{\Phi^u(\mathbf{A}) - \Phi^{u+\delta}(\mathbf{A})} + \mathrm{Tr}[\mathbf{M}^{-1}\mathbf{B}] \stackrel{\text{def}}{=} P(\delta,\mathbf{B}) + Q(\delta,\mathbf{B}).$$

The deterministic upper barrier shift lemma, Lemma 1.2.4, immediately implies the following corollary:

Corollary 3.3.10. *Let $u \in \mathbb{R}$ and $\delta > 0$. Suppose $\lambda_{\max}(\mathbf{A}) < u$. Then if $U^{\mathbf{A}}(\delta,\mathbf{B}) \leq 1$, then*

$$\lambda_{\max}(\mathbf{A}+\mathbf{B}) < u+\delta \quad \textit{and} \quad \Phi^{u+\delta}(\mathbf{A}+\mathbf{B}) \leq \Phi^u(\mathbf{A}).$$

Note that the denominator of $P(\delta,\mathbf{B})$ is not as simple as the denominator of $F(\delta,\mathbf{B})$ above; this is because the required inequality on $U^{\mathbf{A}}(\delta,\mathbf{B})$ is an upper bound instead of a lower bound, so a similar substitution does not help in this case. Because of this, the analysis of these two terms will turn out to be more complicated. Instead of immediately finding a single feasible shift for both terms at once, we will find a shift for each of P and Q separately and then take the sum of these two shifts as our total shift.

For some choice of $\epsilon \in (0,1)$, pick another parameter $\beta = \epsilon/8$ to bound the second term $Q(\delta, \mathbf{B})$, and define the following feasible shifts:

$$\delta_P = \min\{\delta : \delta \in \mathbb{R}^+, P(\delta, \mathbf{B}) \leq 1 - \beta\} \text{ and } \delta_Q = \min\{\delta : \delta \in \mathbb{R}^+, Q(\delta, \mathbf{B}) \leq \beta\}.$$

We fix $\mathbf{A}$ and u satisfying Proposition 3.3.9 in the following results. We will now estimate these separately in Propositions 3.3.12 and 3.3.18, and choose the shift $\delta = \delta_P + \delta_Q$, which will satisfy

Lemma 3.3.11. $\mathbb{E}\,\delta = 1 + 4\beta + C(\eta)(\epsilon_U/\beta)^{\eta/2} + C(\eta)\epsilon_U^{\eta}\beta^{-1-\eta}$.

Since $\beta = \epsilon/8$, it is clear that for adequate choices of constants we obtain $\mathbb{E}\,\delta \leq 1 + \epsilon$, and thus this choice of δ will be a feasible shift in Proposition 3.3.9.

Bounding δ_P

We will prove the following bound on the expectation of δ_P.

Proposition 3.3.12. $\mathbb{E}\,\delta_P \leq 1 + 4\beta + C(\eta)(\epsilon_U/\beta)^{\eta/2}$.

Let $\mathbf{u}_i$ be the unit eigenvectors of $\mathbf{A}$, with corresponding eigenvalues λ_i. Then define

$$P(\delta, \mathbf{B}) = \frac{\sum_i \langle \mathbf{B}\mathbf{u}_i, \mathbf{u}_i\rangle (u + \delta + \lambda_i)^{-2}}{\delta \sum_i (u - \lambda_i)^{-1}(u + \delta - \lambda_i)^{-1}} \leq \frac{1}{\delta}\frac{\sum_i \langle \mathbf{B}\mathbf{u}_i, \mathbf{u}_i\rangle (u - \lambda_i)^{-1}(u + \delta - \lambda_i)^{-1}}{\sum_i (u - \lambda_i)^{-1}(u + \delta - \lambda_i)^{-1}} \stackrel{\text{def}}{=} \frac{1}{\delta} R(\delta, \mathbf{B}). \tag{3.4}$$

We will need some bounds on the moment of the quantity $R(\delta, \mathbf{B})$.

Lemma 3.3.13 (Moments of R)**.** *The quantity $R(\delta, \mathbf{B})$ satisfies the following moment bounds:*

$$\mathbf{E}R(\delta, \mathbf{B}) = 1 \text{ and } \mathbf{E}R(\delta, \mathbf{B})^p \leq C(\eta)$$

for $p < 1 + \eta$.

Proof. First, let $\zeta_i = (u - \lambda_i)^{-1}(u + \delta - \lambda_i)^{-1}$. Then

$$\mathbb{E}\, R(\delta, \mathbf{B}) = \frac{\sum_i \mathbb{E}\langle \mathbf{B}\mathbf{u}_i, \mathbf{u}_i\rangle \zeta_i}{\sum_i \zeta_i} = \frac{\sum_i ||\mathbf{u}_i||_2^2 \zeta_i}{\sum_i \zeta_i} = 1,$$

since the $\mathbf{u}_i$ are unit vectors. Next,

$$||R(\delta, \mathbf{B})||_p \leq \frac{\sum_i ||\langle \mathbf{B}\mathbf{u}_i, \mathbf{u}_i\rangle||_p \zeta_i}{\sum_i \zeta_i} \leq \frac{\sum_i C(\eta)\zeta_i}{\sum_i \zeta_i} = C(\eta),$$

where the first inequality follows by Minkowski's inequality and the second inequality follows from applying the $\mathtt{WR}_p$ condition, which is implied by $\mathtt{SR}_\eta$ since $p < 1 + \eta$. □

This allows us to bound a moment of δ_P, which will be useful later.

Lemma 3.3.14. $\mathbb{E}\,\delta_P^{1+\eta/2} \leq C(\eta)$.

Proof. We have that:

$$\mathbb{P}[\delta_P > \delta] = \mathbb{P}[P(\delta, \mathbf{B}) > 1 - \beta] \leq \mathbb{P}[R(\delta, \mathbf{B}) > \delta(1 - \beta)] \leq M_\eta(t(1 - \beta))^{-1-3\eta/4}$$

where the middle inequality follows from (3.4) and the last inequality follows by Lemma 3.3.13 and Markov's inequality. Integrating now yields

$$\begin{aligned}\mathbb{E}\,\delta_P^{1+\eta/2} &= \int_0^\infty \mathbb{P}[\delta_P^{1+\eta/2} > t]t\; dt = \int_0^\infty \mathbb{P}[\delta_P > s](1 + \eta/2)s^{\eta/2}\; ds \\ &\leq \int_0^1 (1 + \eta/2)t^{1/2} + C(\eta)(1 - \beta)^{-1-3\eta/4}\int_1^\infty t^{-1-\eta/4} \leq C(\eta),\end{aligned}$$

where in the last inequality we used that $\beta \leq 1/2$. □

It will be helpful to further split δ_P into the two random variables

$$\delta_{P,1} = \begin{cases} \delta_P, & \text{if } R(0,\mathbf{B}) \le \beta/4\epsilon_U \\ 0, & \text{otherwise} \end{cases}, \text{ and } \delta_{P,2} = \begin{cases} \delta_P, & \text{if } R(0,\mathbf{B}) > \beta/4\epsilon_U \\ 0, & \text{otherwise} \end{cases}$$

By definition, $\mathbb{E}\,\delta_P = \mathbb{E}\,\delta_{P,1} + \mathbb{E}\,\delta_{P,2}$, so we will bound the expectations of these variables separately in the following two lemmas and sum the bounds to obtain Lemma 3.3.12.

Lemma 3.3.15 (Expectation of $\delta_{P,1}$)**.** $\mathbb{E}\,\delta_{P,1} \le 1 + 4\beta$.

Proof. Suppose $R(0,\mathbf{B}) \le \beta/4\epsilon_U$, and let $z = (1+4\beta)R(0,\mathbf{B})$. Then for each i we have $(u-\lambda_i)^{-1} \le \Phi^u(\mathbf{A}) \le \epsilon_U$, so that $\epsilon_U(u-\lambda_i) \ge 1$, and therefore $u + z - \lambda_i \le (1 + z\epsilon_U)(u-\lambda_i)$. By inspection of the definition of R, we have that $R(z,\mathbf{B}) \le (1+z\epsilon_U)R(0,\mathbf{B})$, so that

$$P(z,\mathbf{B}) \le \frac{1+z\epsilon_U}{z} R(0,\mathbf{B}) \le \frac{\beta^2 + \beta/4 + 1}{1+4\beta} \le 1 - \beta.$$

But then $\delta_{P,1} \le \delta_P \le z = (1+4\beta)R(0,\mathbf{B})$. Taking expectations and using that $\mathbb{E}\,R(0,\mathbf{B}) = 1$ by Lemma 3.3.13 yields the desired bound. □

We will need to use our bounds on the moments of R and δ_P to obtain a bound for $\delta_{P,2}$.

Lemma 3.3.16 (Expectation of $\delta_{P,2}$)**.** $\mathbb{E}\,\delta_{P,2} \le C(\eta)(\epsilon_U/\beta)^{\eta/2}$.

Proof. Let $p = 1 + \eta/2$, and let q be such that $1/p + 1/q = 1$. Note that

$$(\delta_{P,2}/\delta_P)^q = \delta_{P,2}/\delta_P, \text{ and } \mathbb{E}[\delta_{P,2}/\delta_P] = \mathbb{P}[P(0,\mathbf{B}) > \beta/4\epsilon_U].$$

We now have the following chain of inequalities:

$$\begin{aligned}
\mathbb{E}\,\delta_{P,2} &= \mathbb{E}[\delta_P(\delta_{P,2}/\delta_P)] = \mathbb{E}[\delta_P^{1+\eta/2}]^{1/p}\,\mathbb{E}[\delta_{P,2}/\delta_P]^{1/q} && \text{(By Holder's inequality)} \\
&\le C(\eta)\,\mathbb{E}[\delta_{P,2}/\delta_P]^{1/q} && \text{(by Lemma 3.3.14)} \\
&= C(\eta)\,\mathbb{P}[R(0,\mathbf{B}) > \beta/4\epsilon_U]^{1/q} \\
&\le C(\eta)\left(\mathbb{E}[R(0,\mathbf{B})^{1+\eta/2}](\beta/4\epsilon_U)^{-1-\eta/2}\right)^{1/q} && \text{(By Markov's inequality)} \\
&\le C(\eta)(\epsilon_U/\beta)^{\eta/2} && \text{(By Lemma 3.3.13).}
\end{aligned}$$

□

Bounding δ_Q

Note that since $\mathbf{B}$ is symmetric and positive semidefinite it has a symmetric square root $\mathbf{S} = \mathbf{B}^{1/2}$. Once again, let $\mathbf{u}_i$ be the unit eigenvectors of $\mathbf{A}$, with corresponding eigenvalues λ_i. Let $\mathbf{C}_i = \mathbf{S}\mathbf{u}_i\mathbf{u}_i^T\mathbf{S}$, and $\mu_i = \epsilon_U(u - \lambda_i)$, so that

$$Q(\delta,\mathbf{B}) = \mathrm{Tr}[\mathbf{S}\mathbf{M}^{-1}\mathbf{S}] = \left|\left|\sum_{i=1}^n \frac{\mathbf{S}\mathbf{u}_i\mathbf{u}_i^T\mathbf{S}}{u+\delta-\lambda_i}\right|\right| = \epsilon_U \left|\left|\sum_{i=1}^n \frac{\mathbf{C}_i}{\mu_i + \epsilon_U\delta}\right|\right|.$$

Letting $\nu = \epsilon_U\delta$ in the above formula, we obtain that finding δ_Q is equivalent to finding the minimum $\nu > 0$ such that

$$\left|\left|\sum_{i=1}^n \frac{\mathbf{C}_i}{\mu_i + \nu}\right|\right| \le \beta/\epsilon_U. \tag{3.5}$$

We will call this parameter μ. We note the following simple properties of $\mathbf{C}_i$ and μ.

Lemma 3.3.17 (Properties of $\mathbf{C}_i$ and μ)**.** *The following properties hold:* $\mathbb{E}\,||\mathbf{C}_i|| = 1$, $\sum_{i=1}^n 1/\mu_i \le 1$, *and*

$$\mathbb{P}\left(\left|\left|\sum_{i=S}\mathbf{C}_i\right|\right| \ge t\right) = \mathbb{P}(||\mathbf{P}_S\mathbf{B}\mathbf{P}_S|| \ge t) \le C(\eta)t^{-1-\eta} \text{ for all } t \ge C(\eta)|S| \text{ and } |S| \subseteq [n],$$

where $\mathbf{P}_S$ *denotes the orthogonal projection onto the span of* $\{\mathbf{u}_i\}_{i\in S}$.

Proof. For the first property, we have that $\mathbb{E}\,||\mathbf{C}_i|| = \mathbb{E}\,||\mathbf{S}\mathbf{u}_i\mathbf{u}_i^T\mathbf{S}|| = \mathbb{E}\,||\mathbf{S}\mathbf{u}_i||^2 = \mathbb{E}\langle\mathbf{B}\mathbf{u}_i, \mathbf{u}_i\rangle$. The second property holds since $\sum_{i=1}^n 1/\mu_i = \epsilon_U^{-1}\Phi^u(\mathbf{A}) \le 1$. The last property follows by appealing to the facts that $\mathbf{B}$ satisfies $\mathtt{SR}_\eta$ and that $\mathrm{rank}(\mathbf{P}_S) = |S|$ for all $S \subseteq [n]$. □

Proposition 3.3.18 (Expectation of δ_Q)**.** $\mathbb{E}\,\delta_Q \le C(\eta)\epsilon_U^\eta\beta^{-1-\eta}$.

Proof. We will give a bound on $\mathbb{E}\,\mu$. Let $L = \beta/\epsilon_U$. Let $I_k = \{i : 2^k \le \mu_i < 2^{k+1}\}$ and $n_k = |I_k|$. By the previous lemma, we have that

$$\sum_{k\ge 0}\frac{n_k}{2^k} = \sum_{i=1}^n \frac{1}{\mu_i} \le 1.$$

Let $\mu' > 0$ be minimum number such that

$$\frac{1}{2^k+\mu'}\left|\left|\sum_{i\in I_k}\mathbf{C}_i\right|\right| \le \alpha_K, \text{ where } \alpha_k \stackrel{\text{def}}{=} \min\left(\frac{L}{2}\frac{n_k}{2^k}, \frac{L}{2\sigma}2^{-k\eta/(2+2\eta)}\right) \text{ and } \sigma = \sum_{k\ge 0}k^{-\eta/(2+2\eta)},$$

where these constants are chosen so that $\sum_{\alpha_K} \le L$. We can evaluate

$$\left|\left|\sum_{i=1}^n\frac{\mathbf{C}_i}{\mu_i+\mu'}\right|\right| = \left|\left|\sum_{k\ge 0}\sum_{i\in I_k}\frac{\mathbf{C}_i}{\mu_i+\mu'}\right|\right| \le \left|\left|\sum_{k\ge 0}\frac{1}{2^k+\mu'}\sum_{i\in I_k}\mathbf{C}_i\right|\right| \le \sum_{k\ge 0}\frac{1}{2^k+\mu'}\left|\left|\sum_{i\in I_k}\mathbf{C}_i\right|\right|.$$

This is bounded above by $\sum_{k\ge 0}\alpha_k \le L = \beta/\epsilon_U$ by the definition of μ'. But then by (3.5) we have $\mu \le \mu'$, so it suffices to give a bound on $\mathbb{E}\,\mu'$. We have for $t \ge 0$ that

$$\begin{aligned}
\mathbb{P}[\mu' \ge t] &\le \sum_{k\ge 0}\mathbb{P}\left[\frac{1}{2^k+t}\left|\left|\sum_{i\in I_k}\mathbf{C}_i\right|\right| > \alpha_k\right] && \text{(by a union bound)}\\
&= \sum_{k\ge 0}\mathbb{P}\left[\left|\left|\sum_{i\in I_k}\mathbf{C}_i\right|\right| > \alpha_k(2^k+t)\right]\\
&\le \sum_{k\ge 0}\frac{c}{(\alpha_k(2^k+t))^{1+\eta}} && \text{(For } c, \eta \text{ arising from Lemma 3.3.17).}
\end{aligned}$$

Since we chose α_k so that $\alpha_k \ge L/(2\sigma)2^{-k\eta/(2+2\eta)}$, we have that

$$\mathbb{P}[\mu' \ge t] \le \frac{C(\eta)}{L^{1+\eta}}\sum_{k\ge 0}\frac{1}{2^{-k\eta/2}(2^k+t)^{\eta+1}} \le \frac{C(\eta)}{L^{1+\eta}}\sum_{k\ge 0}\frac{1}{(2^k+t)^{\eta/2+1}}.$$

Integrating tails to obtain expectation, we get that

$$\mathbb{E}\,\mu' \le \frac{C(\eta)}{L^{1+\eta}}\sum_{k\ge 0}\int_0^\infty \frac{1}{(2^k+t)^{\eta/2+1}}\,dt = \frac{C(\eta)}{L^{1+\eta}}\sum_{k\ge 0}2^{-k\eta} \le \frac{C(\eta)}{L^{1+\eta}}.$$

Recalling that $\mu = \epsilon_U\delta_Q$ yields the result. □

4 | The Multivariate Barrier Method and Kadison-Singer

4.1 Introduction

The analysis in this chapter primarily follows that of Marcus-Spielman-Srivastava [19] and Tao [28]. The aim of this chapter is to prove the following linear algebraic theorem, which we will see implies an affirmative solution to the Kadison-Singer problem.

Theorem 4.1.1. *Let $\epsilon > 0$, and let $\mathbf{A}_1, \ldots, \mathbf{A}_m \in \mathbb{C}^{n\times n}$ be independent random rank one Hermitian positive semidefinite matrices taking finitely many values, such that*

$$\mathbb{E}\sum_{i=1}^{m} \mathbf{A}_i = \mathbf{I}_d \text{ and } \mathbb{E}\,\mathrm{tr}\,\mathbf{A}_i \leq \epsilon \text{ for all } i.$$

Then with positive probability, the largest root of the polynomial $\chi\left[\sum_{i=1}^m \mathbf{A}_i\right](x)$ is at most $(1+\sqrt{\epsilon})^2$. Equivalently,

$$\mathbb{P}\left[\left\|\sum_{i=1}^{m} \mathbf{A}_i\right\|_2 \leq (1+\sqrt{\epsilon})^2\right] > 0.$$

The proof will use the technique of interlacing families developed by [18] to prove the following proposition, which allows us to work with a deterministic characteristic polynomial instead of a random one.

Proposition 4.1.2. *Let $\mathbf{A}_1, \ldots, \mathbf{A}_m \in \mathbb{C}^{n\times n}$ be independent random rank one Hermitian positive semidefinite matrices taking finitely many values. Then with positive probability, the largest root of the polynomial $\chi\left[\sum_{i=1}^m \mathbf{A}_i\right](x)$ is bounded above by the largest root of the expected characteristic polynomial $\mathbb{E}\,\chi\left[\sum_{i=1}^m \mathbf{A}_i\right](x)$*

While the expected characteristic polynomial is tough to analyze as given, we will show that it has a more useful explicit formula:

Definition 4.1.3. Let $\mathbf{A}_1, \ldots, \mathbf{A}_m \in \mathbb{C}^{n\times n}$. The *mixed characteristic polynomial* of $\mathbf{A}_1, \ldots, \mathbf{A}_m$ is given by

$$\mu[\mathbf{A}_1, \ldots, \mathbf{A}_m](x) = \left(\prod_{i=1}^{m} 1 - \partial_i\right) \det\left(x\mathbf{I} + \sum_{i=1}^{m} z_i \mathbf{A}_i\right)\Bigg|_{z_1=\cdots=z_m=0}, \tag{4.1}$$

where ∂_i is shorthand for ∂_{z_i}.

Theorem 4.1.4. *Let $\mathbf{A}_1, \ldots, \mathbf{A}_m \in \mathbb{C}^{n\times n}$ be independent random Hermitian rank one positive semidefinite matrices with finite support. Then*

$$\mathbb{E}\,\chi\left[\sum_{i=1}^{m} \mathbf{A}_i\right](x) = \mu[\mathbb{E}\,\mathbf{A}_1, \ldots, \mathbb{E}\,\mathbf{A}_m](x).$$

After proving this representation, the proof will proceed as follows: first, we will show that the mixed characteristic polynomial is real rooted using the theory of real stablility, a multivariate generalization of real rootedness. Then, we will prove the following theorem, which bounds the largest root of this polynomial

by investigating how the operators $(1-\partial_i)$ affect locations of roots. Note that these operators are reminiscent of those found in (1.5). Not coincidentally, we will utilize multivariate generalizations of the barrier potentials found in Chapter 1.

Theorem 4.1.5. *Let $\mathbf{A}_1, \dots, \mathbf{A}_m \in \mathbb{C}^{n\times n}$ be Hermitian positive semidefinite matrices with $\sum_{i=1}^m \mathbf{A}_i = \mathbf{I}_d$ and $\mathrm{Tr}[\mathbf{A}_i] \leq \epsilon$ for each i. Then the largest root of $\mu[\mathbf{A}_1, \dots, \mathbf{A}_m](x)$ is at most $(1+\sqrt{\epsilon})^2$.*

Theorem 4.1.1 follows by applying Theorem 4.1.5 to the matrices $\mathbb{E}\,\mathbf{A}_i$ and subsequently applying Theorem 4.1.4 and Proposition 4.1.2.

4.1.1 The Mixed Characteristic Polynomial

In this section, we will given an elegant proof due to Tao [28] of the following deterministic version of Theorem 4.1.4, which shows that the mixed characteristic polynomial of *rank one matrices* is equal to the characteristic polynomial of their sum.

Theorem 4.1.6. *Let $\mathbf{A}_1, \dots, \mathbf{A}_m \in \mathbb{C}^{d\times d}$ be rank one matrices with sum $\mathbf{A}$. Then*

$$\chi[\mathbf{A}](x) = \mu[\mathbf{A}_1, \dots, \mathbf{A}_m](x).$$

First, we show that the determinant is affine-multilinear under rank-one updates.

Lemma 4.1.7. *Let $\mathbf{A}_1, \dots, \mathbf{A}_m, \mathbf{B} \in \mathbb{C}^{d\times d}$, and suppose that the $\mathbf{A}_i$ have rank one. The polynomial*

$$(t_1, \dots, t_m) \mapsto \det\left(\mathbf{B} + \sum_{i=1}^m t_i \mathbf{A}_i\right)$$

is affine-multilinear in the $t_1, \dots, t_m$, meaning that it is of the form

$$(t_1, \dots, t_m) \mapsto \sum_{1\leq i_1 < \cdots < i_j \leq m} a_{i_1,\dots,i_j} t_{i_1} \cdots t_{i_j}$$

for some coefficients $a_{i_1,\dots,i_j}$.

Proof. We first prove the lemma for $m = 1$. Suppose $\mathbf{A} \in \mathbb{C}^{d\times d}$ is rank one, so that $\mathbf{A} = \mathbf{u}\mathbf{v}^*$ for column vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^d$. For invertible matrices $\mathbf{B}$, we have by Sylvester's formula that

$$\det(\mathbf{B} + t\mathbf{u}\mathbf{v}^*) = \det(\mathbf{B})(1 + t\mathbf{v}^*\mathbf{B}^{-1}\mathbf{u}),$$

which is affine-linear in t. Since invertible matrices are dense in the space of all matrices, the lemma follows for all matrices $\mathbf{B}$.

If $m > 1$, to show that the polynomial is affine-linear in t_i, we can just freeze the other t_i and reduce to the case $m = 1$. □

Proof of Theorem 4.1.6. Let $p(t_1, \dots, t_m) = \det(\mathbf{B} + t_1\mathbf{A}_1 + \cdots + t_m\mathbf{A}_m)$. By the previous lemma, it is affine-multilinear, so the partial derivatives $\partial_{t_i}^k p$ vanish for $k \geq 2$. Taking the Taylor expansion to degree 1 in every t_i, we obtain

$$p(t_1, \dots, t_m) = \left(\prod_{i=1}^m (1 + t_i\partial_{z_i})\right) p(z_1, \dots, z_m)|_{z_1=\cdots=z_m=0}.$$

Setting $\mathbf{B} = x\mathbf{I}$ and $t_i = -1$ in the above yields the theorem. □

This theorem quickly implies Theorem 4.1.4, since we can expand the mixed characteristic polynomial as a linear combination of terms which are multilinear in the $\mathbf{A}_i$, which are jointly independent.

4.1.2 Interlacing and Interlacing Families

Marcus, Spielman and Srivastava [18] developed the notion of interlacing families of polynomials, which have the powerful property that at least one polynomial in the family has largest root which is at most the largest root of the sum of the polynomials in the family.

We use the following notation: if $f(x)$ is a real rooted univariate polynomial, we let $\mathrm{lc}(f)$ be the leading coefficient of f, and we let $\mathrm{lr}(f)$ be the largest (most positive) root of f. We say that f is *real rooted* if all of its coefficients and roots are real.

Definition 4.1.8. A real rooted polynomial $g(x) = (x - \alpha_i) \cdots (x - \alpha_{n-1})$ *interlaces* another real rooted polynomial $f(x) = (x - \beta_i) \cdots (x - \beta_n)$ if

$$\beta_1 \leq \alpha_1 \leq \beta_2 \leq \alpha_2 \leq \cdots \leq \alpha_{n-1} \leq \beta_n.$$

We say that $g(x)$ *strictly interlaces* $f(x)$ if all of these inequalities are strict. We say that real rooted polynomials $f_1(x), \ldots, f_k(x)$ have a *common interlacing* if there is a real rooted polynomial $g(x)$ such that $g(x)$ interlaces $f_i(x)$ for each i.

Note that the definition of common interlacing may be rephrased as a series of inequalities on the roots of the f_i, but it is often convenient to actually find a polynomial g which interlaces the f_i.

Definition 4.1.9. Let $S_1, \ldots, S_m$ be finite sets, and for every assignment $s_1, \ldots, s_m \in S_1 \times \cdots \times S_m$, suppose $f_{s_1,\ldots,s_m}(x)$ is a real rooted degree n polynomial with positive leading coefficient. For every partial assignment $s_1, \ldots, s_k \in S_1 \times \cdots \times S_k$, define

$$f_{s_1,\ldots,s_k}(x) = \sum_{\substack{s_j \in S_j \\ k+1 \leq j \leq m}} f_{s_1,\ldots,s_m}(x),$$

as well as

$$f_{\emptyset}(x) = \sum_{\substack{s_j \in S_j \\ 1 \leq j \leq m}} f_{s_1,\ldots,s_m}(x).$$

These are the sums of all the fs given by the extensions of an assignment $s_1, \ldots, s_k$.

We say that the polynomials $\{f_{s_1,\ldots,s_m}\}$ are an *interlacing family* if for all and all partial assignments $s_1, \ldots s_k \in S_1 \times \cdots \times S_k$ with $k \in [m-1]$, the polynomials

$$\{f_{s_1,\ldots,s_k,t}\}_{t \in S_{k+1}}$$

have a common interlacing.

We now present some important properties of interlacing families. The next lemma shows that the largest root of the sum of polynomials with a common interlacing bounds the largest root of at least one of them.

Lemma 4.1.10. *Let $f_1, \ldots, f_k$ be real rooted polynomials of the same degree n with positive leading coefficients, and let $F = f_1 + \cdots + f_k$. If $f_1, \ldots, f_k$ have a common interlacing, then there exists an i such that*

$$\mathrm{lr}(f_i) \leq \mathrm{lr}(F).$$

Proof. Let g be the common interlacing of the f_i. Then for each i we have that $f_i(\mathrm{lr}(g)) \leq 0$, since $\mathrm{lr}\, g$ is between the second-largest and largest roots of f_i and f_i has positive leading coefficient. Thus, $F(\mathrm{lr}(g)) \leq 0$, so that $\mathrm{lr}(F) \geq \mathrm{lr}(g)$. But then there must be some i such that $f_i(\mathrm{lr}(F)) \geq 0$, since otherwise $F(\mathrm{lr}(F)) = \sum f_i(\mathrm{lr}(F)) < 0$, a contradiction. This f_i must satisfy $\mathrm{lr}(f_i) \leq \mathrm{lr}(F)$. □

Using the inductive nature of interlacing families, the following result generalizes the previous lemma and yields the property of interlacing families mentioned at the beginning of this section.

Theorem 4.1.11. *Let $S_1, \dots, S_m$ be finite sets and let $\{f_{s_1,\dots,s_m}\}$ be an interlacing family of polynomials. Then there exists some assignment $s_1, \dots, s_m \in S_1 \times \cdots \times S_m$ such that*

$$\mathrm{lr}(f_{s_1,\dots,s_m}) \leq \mathrm{lr}(f_\emptyset).$$

Proof. Proceed by induction. By the definition of an interlacing family, the set of polynomials $\{f_{s_1}\}_{s_1 \in S_1}$ has a common interlacing, and since their sum is $f_\emptyset$, Lemma 4.1.10 yields a choice of s_1 such that $\mathrm{lr}(f_{s_1}) \leq \mathrm{lr}(f_\emptyset)$. Next, suppose that for some choice $s_1, \dots, s_k \in S_1 \times \cdots \times S_k$, $\mathrm{lr}(f_{s_1,\dots,s_k}) \leq \mathrm{lr}(f_\emptyset)$. Then since $\sum_{s_{k+1} \in S_{k+1}} f_{s_1,\dots,s_k,s_{k+1}} = f_{s_1,\dots,s_k}$, once again use Lemma 4.1.10 to choose s_{k+1} with

$$\mathrm{lr}(f_{s_1,\dots,s_k}) \leq \mathrm{lr}(f_\emptyset),$$

completing the induction. □

Finally, the following lemma proven as Proposition 1.35 in [12] states that two polynomials having a common interlacing is equivalent to a real rootedness condition on all of their convex combinations, and will be useful for proving real rootedness. We omit the proof, which consists of tedious casework.

Lemma 4.1.12. *Let f and g be monic univariate polynomials of the same degree n such that, for all constants $\lambda \in (0,1)$, the linear combination $\lambda f + (1-\lambda)g$ is real rooted. Then f and g have a common interlacing.*

4.1.3 Real Stable Polynomials

In order to prove the real rootedness criterion in Lemma 4.1.12, the authors of [18] and [19] use the theory of real stability, a generalization of real rootedness to multivariate polynomials. As we will see, real stable polynomials have useful closure and interlacing properties. Below we present results necessary for the proof of Theorem 4.1.1, including a theorem of [7] concerning the real stability of polynomials given by certain determinants; for a broader survey, see [30].

To ease notation, in what follows we let ∂_i stand for ∂_{z_i}.

Definition 4.1.13. A polynomial $p \in \mathbb{C}[z_1, \dots, z_m]$ is *stable* if whenever $\mathrm{Im}(z_i) > 0$ for all i, $p(z_1, \dots, z_m) \neq 0$. A stable polynomial $p \in \mathbb{R}[z_1, \dots, z_m]$ is said to be *real stable*. Consequently, a univariate polynomial is real stable if and only if it has real coefficients and roots.

We present two closure properties for real stable polynomials: closure under univariate restriction and closure under certain differential operators. In order to prove some of these closure properties, we will need the following theorem from complex analysis for so we can construct continuity arguments.

Theorem 4.1.14 (Hurwitz' theorem)**.** *Let D be a domain in $\mathbb{C}^n$ and suppose that $\{f_k\}$ is a sequence of nonvanishing analytic functions on D that converge to f uniformly on compact subsets of D. Then f is either nonvanishing on D or else identically zero.*

Lemma 4.1.15. *Let $p \in \mathbb{C}[z_1, \dots, z_m]$ be real stable of degree d, and let t be real. Then $p(z_1, \dots, z_{m-1}, t) \in \mathbb{C}[z_1, \dots, z_{m-1}]$ is either real stable or identically zero.*

Proof. Consider the polynomials $p_k = p(z_1, \dots, z_{m-1}, t + i2^{-t})$, which are real stable since p is and converge uniformly to $p(z_1, \dots, z_{m-1}, t)$. Applying Hurwitz' theorem on the upper half-plane of $\mathbb{C}^{m-1}$, we obtain that $p(z_1, \dots, z_{m-1}, t) \in \mathbb{C}[z_1, \dots, z_{m-1}]$ is either real stable or identically zero. □

We next establish closure under the family of operators $1 - \partial_i$.

Lemma 4.1.16. *Let $p \in \mathbb{C}[z_1, \dots, z_m]$ be real stable of degree d. Then $(1 - \partial_i)p(z)$ is also real stable.*

Proof. Fix $z_1, \ldots, z_{m-1}$, and let $g(z) = p(z_1, \ldots, z_{m-1}, z)$. It suffices to show that $p(z) - p'(z)$ is stable if $p(z)$ is stable. Factoring, $p(z) = a\prod_{i=1}^{d}(z - \alpha_i)$ for some α_i with nonpositive imaginary part, we can write

$$p(z) - p'(z) = p(z)\left(1 - \sum_{i=1}^{d}(z - \alpha_i)^{-1}\right).$$

But then if z has positive imaginary part, then $(z - \alpha_i)$ has positive imaginary part and $(z - \alpha_i)^{-1}$ has negative imaginary part for all i, so that the sum has nonzero imaginary part. Thus if $p(z)$ is stable, then $p(z) - p'(z)$ is stable, and we are done. □

In order to have a starting point for using these closure properties, we show that the term at the heart of equation (4.1) is itself a real stable polynomial.

Lemma 4.1.17 (Proposition 2.4, [7])**.** *Let $\mathbf{A}_1, \ldots, \mathbf{A}_m \in \mathbb{R}^{n\times n}$ be positive semidefinite matrices, and let $\mathbf{B} \in \mathbb{C}^{n\times n}$ be Hermitian. Then the polynomial*

$$f(z_1, \ldots, z_n) = \det\left(\sum_{i=1}^{m} z_i\mathbf{A}_i + \mathbf{B}\right)$$

is real stable or identically zero.

Proof. By a continuity argument using Hurwitz' theorem on the upper half-plane (taking limits, say, of positive definite $\mathbf{A}_{ik} \to \mathbf{A}_i$), we can reduce to the case that all of the $\mathbf{A}_i$ are positive definite. Let $z(t) = \alpha + \lambda t$ with $\alpha \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^n_+$, and $t \in \mathbb{C}$. Then $\mathbf{C} = \sum_i \lambda_i\mathbf{A}_i$ is positive definite, and thus has both an inverse and a square root. Let $\mathbf{H}$ be the Hermitian matrix given by $\mathbf{H} = \sum_i \alpha_i\mathbf{A}_i + \mathbf{B}$. Substituting, we may write

$$f(z(t)) = \det\left(\sum_{i=1}^{m}(\alpha_i + \lambda_i t)\mathbf{A}_i + \mathbf{B}\right) = \det\left(t\mathbf{C} + \mathbf{H}\right) = \det(\mathbf{C})\det(t\mathbf{I} + \mathbf{C}^{-1/2}\mathbf{H}\mathbf{C}^{-1/2}),$$

where the last equality follows by Sylvester's formula. But then $f(z(t))$ is a constant multiple of the characteristic polynomial of the Hermitian matrix $\mathbf{C}^{-1/2}\mathbf{H}\mathbf{C}^{-1/2}$, and thus must have all real zeros. Since α and λ were arbitrary, we conclude that $f(z_1, \ldots, z_n)$ is either real stable or identically zero. □

Applying the above lemmas to the representation of the mixed characteristic polynomial given in Theorem 4.1.4, we can conclude that it is real rooted.

Corollary 4.1.18. *Let $\mathbf{A}_1, \ldots, \mathbf{A}_m \in \mathbb{C}^{n\times n}$ be positive semidefinite and Hermitian. Then $\mu[\mathbf{A}_1, \ldots, \mathbf{A}_m](x)$ is real rooted.*

Proof. By taking $x\mathbf{I}$ as an extra $\mathbf{A}_i$ and setting $\mathbf{B} = 0$ in Lemma 4.1.17, we get that the multivariate polynomial

$$\det\left(x\mathbf{I} + \sum_{i=1}^{m} z_i\mathbf{A}_i\right)$$

is real stable. The closure properties of Lemmas 4.1.15 and 4.1.16 imply that $\mu[\mathbf{A}_1, \ldots, \mathbf{A}_m](x)$ is real stable; since it is univariate, it is real rooted. □

We now have all we need to prove Proposition 4.1.2. Suppose that $\mathbf{A}_i$ may take the values $\mathbf{X}_{i,j_i}$ for $j_i \in [\ell_i]$. We will show that the polynomials

$$f_{j_1,\ldots,j_k} = \left(\prod_{i=1}^{k} p_{i,j_i}\right) \cdot \mathop{\mathbb{E}}_{\mathbf{A}_{k+1},\cdots,\mathbf{A}_m} \chi\left[\sum_{i=1}^{k}\mathbf{X}_{i,j_i} + \sum_{i=k+1}^{m}\mathbf{A}_i\right](x)$$

form an interlacing family.

Proof of Proposition 4.1.2. Fix $k \in [m]$ and $j_1, \ldots, j_{k-1} \in [l_1] \times \cdots \times [l_{k-1}]$. We must show for all $s, t \in [l_k]$ that $f_{j_1,\ldots,j_{k-1},s}$ and $f_{j_1,\ldots,j_{k-1},t}$ have a common interlacing.

By Lemma 4.1.12, we only need to show that for every $\lambda \in [0,1]$,

$$p_\lambda(x) = \lambda q_{j_1,\ldots,j_{k-1},s}(x) + (1-\lambda) f_{j_1,\ldots,j_{k-1},t}(x)$$

is real rooted. Let $\mathbf{Y}_k$ be the random vector which equals $\mathbf{X}_{k,s}$ with probability λ and $\mathbf{X}_{k,t}$ with probability $1-\lambda$. We may then factor this into the expectation as:

$$p_\lambda(x) = \left(\prod_{i=1}^{k-1} p_{i,j_i}\right) \cdot \mathop{\mathbb{E}}_{\mathbf{Y}_k, \mathbf{A}_{k+1}, \cdots, \mathbf{A}_m} \chi\left[\sum_{i=1}^{k-1} \mathbf{X}_{i,j_i} + \mathbf{Y}_k + \sum_{i=k+1}^{m} \mathbf{A}_i\right](x).$$

But then by Theorem 4.1.4, $p_\lambda(x)$ is a constant multiple of a mixed characteristic polynomial, and by Corollary 4.1.18 must be real rooted.

Thus the $f_{j_1,\ldots,j_k}$ form an interlacing family. Note that $f_\emptyset = \mathbb{E}[\sum_{i=1}^m \mathbf{A}_i]$, so that the result follows by Theorem 4.1.11. □

4.2 The Multivariate Barrier Method

We will first define a multivariate polynomial whose restriction yields the mixed characteristic polynomial.

Lemma 4.2.1. *Let $\mathbf{A}_1, \ldots, \mathbf{A}_m \in \mathbb{C}^{d\times d}$ be Hermitian positive semidefinite matrices with $\sum_{i=1}^m \mathbf{A}_i = \mathbf{I}$, and define*

$$Q(y_1, \ldots, y_m) = \left(\prod_{i=1}^m 1 - \partial_{y_i}\right) \det\left(\sum_{i=1}^m y_i \mathbf{A}_i\right).$$

Then

$$\mu[\mathbf{A}_1, \ldots, \mathbf{A}_m](x) = Q(x, \ldots, x).$$

Proof. For differentiable f, we have by the chain rule that $\partial_{y_i}(f(y_i))|_{y_i = z_i + x} = \partial_{z_i} f(z_i + x)$. The lemma follows by applying this relation to equation (4.1). □

Given this lemma, it is apparent that some sort of upper bound on the roots of Q will lead to an upper bound on the roots of the mixed characteristic polynomial. More specifically, given a real stable polynomial $p \in \mathbb{C}[z_1, \ldots, z_m]$, we will say that a real vector $\mathbf{z} = (z_1, \ldots, z_m) \in \mathbb{R}^m$ is *above the roots* of p if p is positive on the real orthant

$$\{(y_1 \ldots, y_m) : y_i \geq x_i \text{ for all } i \in [m]\},$$

and we will show that $(1+\sqrt{\epsilon})^2 \cdot \mathbf{1}$ is above the roots of Q, so that the largest root of the mixed characteristic polynomial is at most $(1+\sqrt{\epsilon})^2$.

In order to control the positions of the (real) roots of Q, we will use multivariate generalizations of the barrier functions in Chapter 1.

Definition 4.2.2. Let $p \in \mathbb{C}[z_1, \ldots, z_m]$ be real stable and let $\mathbf{z}$ lie above the roots of p. Define the univariate restriction $q_{\mathbf{z},i}(t) = p(z_1, \ldots, z_{i-1}, t, z_{i+1}, \ldots, z_m)$. Then the *barrier function of p in coordinate i at z* is given by

$$\Phi_p^i(\mathbf{z}) = \partial_i(\log p(\mathbf{z})) = \frac{\partial_i p(z)}{p(z)} = \frac{q'_{\mathbf{z},i}(z_i)}{q_{\mathbf{z},i}(z_i)} = \sum_{j=1}^r \frac{1}{z_i - \lambda_j},$$

where $\lambda_1, \ldots, \lambda_r$ are the roots of $q_{\mathbf{z},i}$, which are all real by closure properties. Note that the barrier function in a certain coordinate takes the same form as the univariate barrier functions in Chapter 1.

The following is a useful 'commutation' relation.

Lemma 4.2.3. *Let $p \in \mathbb{C}[z_1, \ldots, z_m]$ be real stable. Then $\partial_j \Phi_p^i = \partial_i \Phi_p^j$.*

Proof. We have $\partial_j \Phi_p^i = \partial_j \partial_i \log p = \partial_i \partial_j \log p = \partial_i \Phi_p^j$. □

The next lemma shows that the multivariate barrier functions are are monotonic and convex in every coordinate. We make use of the following theorem of of [6], which gives a characterization of all real stable bivariate polynomials.

Lemma 4.2.4 (Essentially Corollary 6.7, [6])**.** *If $p(z_1, z_2)$ is a bivariate real stable polynomial of degree exactly d, then there exist positive semidefinite matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$ and a symmetric matrix $\mathbf{C} \in \mathbb{R}^{d \times d}$ such that*

$$p(z_1, z_2) = \pm \det(z_1 \mathbf{A} + z_2 \mathbf{B} + \mathbf{C}).$$

Furthermore, we can take $\mathbf{A} + \mathbf{B}$ to be the identity.

We will also need the well-known *Jacobi determinant formula*

$$\partial_t \det(\mathbf{X} + t\mathbf{Y}) = \operatorname{tr}[(\mathbf{X} + t\mathbf{Y})^{-1}\mathbf{Y}] \det(\mathbf{X} + t\mathbf{Y}). \tag{4.2}$$

Lemma 4.2.5. *Suppose p is real stable and $\mathbf{z}$ is above the roots of p. Then for all $i, j \leq m$ and $\delta \geq 0$, the barrier function of p in coordinate i satisfies:*

$$(-1)^k \partial_j^k \Phi_p^i(\mathbf{z}) \geq 0 \tag{4.3}$$

for $k = 0, 1, 2$. In particular, it is nonnegative, monotonic, and convex in every coordinate.

Proof. Nonnegativity follows directly frmom the assumption $\mathbf{z}$ is above the roots of p, so we focus on monotonicity and convexity.

Suppose that $i = j$, and recall the the definitions of the univariate restrictions $q_{\mathbf{z},i}(z_i) = \prod_{k=1}^r (z_i - \lambda_k)$. Then the barrier function in coordinate i is given by $\Phi_p^i(\mathbf{z}) = \sum_{k=1}^r \frac{1}{z_i - \lambda_k}$. We focus on each term of this sum. Since z_i is above the roots of p, we have that $z_i > \lambda_k$, so it is clear that term $(z_i - \lambda)^{-1}$ decreases as z_i increases. Taking its second derivative, we obtain $\partial_i^2 (z_i - \lambda_k)^{-1} = 2(z_i - \lambda_k)^{-3} > 0$, which yields convexity.

Now, suppose that $i \neq j$. Without loss of generality, by renumbering and fixing all of the other variables we can assume that p takes the form $p(z_1, z_2)$. Suppose the point (z_1, z_2) is above the roots of p. By Lemma 4.2.4, there are positive semidefinite symmetric $\mathbf{A}, \mathbf{B}$ with $\mathbf{A} + \mathbf{B} = \mathbf{I}$ and a symmetric matrix $\mathbf{C}$ such that $p(z_1, z_2) = \pm \det(z_1 \mathbf{A} + z_2 \mathbf{B} + \mathbf{C})$. The sign of $p(z_1, z_2)$ must be positive, since for sufficiently large t, $p(t, t) = \det(t\mathbf{I} + \mathbf{C})$ will be positive and there are no roots in the real orthant above (z_1, z_2).

By the Jacobi determinant formula (4.2), we have

$$\Phi_p^1(z_1, z_2) = \frac{\det(z_1 \mathbf{A} + z_2 \mathbf{B} + \mathbf{C}) \operatorname{Tr}[(z_1 \mathbf{A} + z_2 \mathbf{B} + \mathbf{C})^{-1} \mathbf{A}]}{\det(z_1 \mathbf{A} + z_2 \mathbf{B} + \mathbf{C})} = \operatorname{Tr}[(z_1 \mathbf{A} + z_2 \mathbf{B} + \mathbf{C})^{-1} \mathbf{A}].$$

Let $\mathbf{M} = z_1 \mathbf{A} + z_2 \mathbf{B} + \mathbf{C}$. Then we claim that the symmetric matrix $\mathbf{M}$ is positive definite; suppose that it is not, so that it has a nonpositive eigenvalue $-\lambda \leq 0$. But then $p(z_1 + \lambda, z_2 + \lambda) = \det(M + (-\lambda)(\mathbf{A} + \mathbf{B})) = 0$, a contradiction to the assumption that (z_1, z_2) is above the roots of p. Thus $\mathbf{M}$ is positive definite, and $\mathbf{M}$ has an invertible square root $\mathbf{M}^{1/2}$. We can expand

$$\begin{aligned}
\Phi_p^1(z_1, z_2 + \delta) &= \operatorname{Tr}[(\mathbf{M} + \delta \mathbf{B})^{-1} \mathbf{A}] = \operatorname{Tr}[\mathbf{M}^{-1/2} (\mathbf{O} + \delta \mathbf{M}^{-1/2} \mathbf{B} \mathbf{M}^{-1/2})^{-1} \mathbf{M}^{-1/2} \mathbf{A}] \\
&= \operatorname{Tr}[(\mathbf{I} + \delta \mathbf{M}^{-1/2} \mathbf{B} \mathbf{M}^{-1/2})^{-1} \mathbf{M}^{-1/2} \mathbf{A} \mathbf{M}^{-1/2}] \\
&= \operatorname{Tr}[(\mathbf{I} - \delta \cdot \mathbf{X} + \delta^2 \cdot \mathbf{X}^2 + O(\delta^3))\mathbf{Y}] \\
&= \operatorname{Tr}[\mathbf{Y}] - \delta \operatorname{Tr}[\mathbf{X}\mathbf{Y}] + \delta^2 \operatorname{Tr}[\mathbf{X}^2 \mathbf{Y}] + O(\delta^3)
\end{aligned}$$

for the positive semidefinite matrices $\mathbf{X} = \mathbf{M}^{-1/2}\mathbf{B}\mathbf{M}^{-1/2}$ and $\mathbf{Y} = \mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2}$. In the second to last equality, we expanded $(\mathbf{I}+\mathbf{X})^{-1}$ as a power series. Since the sign of the first-order term in δ is negative and the sign of the second-order term in δ is positive, we establish montonicity and convexity.

□

A stability bound of the form $\Phi_p^i(z) < 1$ combined with the monotonic ($k = 0$) properties of the barrier function in Lemma 4.2.5 lets us argue that applying operators of the form $(1 - \partial_i)$ to some real stable polynomial p preserves vectors which are above the roots of p.

Corollary 4.2.6. *Let $p \in \mathbb{C}[z_1, \ldots, z_m]$ be real stable, and suppose $\mathbf{z} \in \mathbb{R}^m$ is above the roots of p and $\Phi_p^i(\mathbf{z}) < 1$. Then $\mathbf{z}$ is above the roots of $p - \partial_i p$.*

Proof. Suppose $\mathbf{y}$ is above $\mathbf{z}$. Note that $p(\mathbf{y}) - \partial_i p(\mathbf{y})$ only vanishes if $p(\mathbf{y}) = \partial_i p(\mathbf{y})$, which is equivalent to $\Phi_p^i(\mathbf{y}) = 1$. But since $\mathbf{y}$ is above $\mathbf{z}$, by the monotonicity condition in Lemma 4.2.5 this can never happen. □

The next lemma shows that in order to maintain this stability bound in some coordinate, we only need to shift our bound on the roots of p by a small amount in that coordinate.

Lemma 4.2.7. *Let $p \in \mathbb{C}[z_1, \ldots, z_m]$ be real stable with $z \in \text{Above}_p$, and $\delta > 0$ satisfies*

$$\Phi_p^j(\mathbf{z}) \leq 1 - 1/\delta.$$

Then for all $i \in [m]$,

$$\Phi_{p-\partial_j p}^i(\mathbf{z} + \delta\mathbf{e}_j) \leq \Phi_p^i(z).$$

Proof. By the previous corollary, we have $\mathbf{z}$ is above the roots of $p - \partial_j p$, and thus so is $\mathbf{z} + \delta\mathbf{e}_j$. Expanding the operator $\Phi_{p-\partial_j p}^i$ in the domain above the roots of $p - \partial_j p$, we have that

$$\Phi_{p-\partial_j p}^i = \partial_i \log(p - \partial_j p) = \partial_i \log\left[p(1 - \Phi_p^j)\right] = \partial_i \log p + \partial_i \log(1 - \Phi_p^j) = \Phi_p^i + \frac{\partial_i \Phi_p^j}{1 - \Phi_p^j}.$$

We want to show that $\Phi_{p-\partial_1 p}^i(\mathbf{z} + \partial\mathbf{e}_1) \leq \Phi_p^i(\mathbf{z})$. Using the above, this is equivalent to

$$\Phi_p^i(\mathbf{z} + \delta\mathbf{e}_1) - \frac{\partial_1 \Phi_p^i(\mathbf{z} + \delta\mathbf{e}_1)}{1 - \Phi_p^1(\mathbf{z} + \partial\mathbf{e}_1)} \leq \Phi_p^i(\mathbf{z}).$$

By the convexity property in Lemma 4.2.5, $-\partial_j \Phi_p^i(\mathbf{z} + \delta\mathbf{e}_j) \leq \Phi_p^i(\mathbf{z}) - \Phi_p^i(\mathbf{z} + \partial\mathbf{e}_j)$, so that it is sufficient to show that

$$-\frac{\partial_j \Phi_p^i(\mathbf{z} + \delta\mathbf{e}_j)}{1 - \Phi_p^1(\mathbf{z} + \partial\mathbf{e}_j)} \leq -\delta \cdot \partial_j \Phi_p^i(\mathbf{z} + \delta\mathbf{e}_j).$$

The term $-\partial_j \Phi_p^i(\mathbf{z} + \delta\mathbf{e}_j) = -\partial_i \Phi_p^j(\mathbf{z} + \delta\mathbf{e}_j)$ by Lemma 4.2.3 and is nonnegative by Lemma 4.2.5. If it is zero, we are done, so assuming it is positive, canceling and rearranging yields that the previous inequality is equivalent to

$$1 - 1/\delta \geq \Phi_p^i(\mathbf{z} + \delta\mathbf{e}_j),$$

which is implied by the monotonicity property in Lemma 4.2.5 and the assumption that $\Phi_p^j(\mathbf{z}) \leq 1 - 1/\delta$. □

We are now able to complete the proof of Theorem 4.1.5.

Proof of Theorem 4.1.5. Set

$$p(y_1, \ldots, y_m) = \det\left(\sum_{i=1}^m y_i \mathbf{A}_i\right).$$

Let $t = \epsilon + \sqrt{\epsilon} > 0$, and note that $p(t\mathbf{1}) = \det(t\mathbf{I}_d) > 0$, so that $t\mathbf{1}$ is above the roots of p since the $\mathbf{A}_i$, are positive semidefinite. Using the Jacobi Determinant formula, write

$$\Phi_p^i(t\mathbf{1}) = \frac{\partial_i p(t\mathbf{1})}{p(t\mathbf{1})} = \mathrm{Tr}\left[\left(t\sum_{i=1}^m \mathbf{A}_i\right)^{-1}\mathbf{A}_i\right] = \mathrm{Tr}[\mathbf{A}_i]/t \leq \epsilon/(\epsilon+\sqrt{\epsilon}).$$

For each $k \in [m]$, let

$$p_k(\mathbf{y}) = \left(\prod_{i=1}^k 1 - \partial_i\right) p(\mathbf{y}).$$

Let $\mathbf{x}^{(0)} = t\mathbf{1}$, and let $\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)} + \delta\mathbf{e}_i$ for each $1 \leq i \leq m$. Proceed by induction on i. The monotonicity properties in Lemma 4.2.5 imply that $\Phi^i_{p_k}(\mathbf{x}^{(k)}) \leq \Phi^i_p(t\mathbf{1})$, so that by Corollary 4.2.6 we have that $\mathbf{x}^{(k)}$ is above the roots of p_k. In particular, we have that $\mathbf{x}^{(m)} = (t+\delta)\mathbf{1}$ is above the roots of p_m. Since $\mu[\mathbf{A}_1, \ldots, \mathbf{A}_m](x) = p_m(x\mathbf{1})$ by Lemma 4.2.1, the largest root of $\mu[\mathbf{A}_1, \ldots, \mathbf{A}_m](x)$ is at most $t + \delta = (1+\sqrt{\epsilon})^2$. □

4.3 The Kadison-Singer Problem

The original form of the Kadison-Singer problem [16] asked "whether each pure state on the algebra of bounded diagonal operators on ℓ_2 has a unique extension to a pure state on $B(\ell_2)$, the algebra of all bounded operators on ℓ_2." It was one of the core open problems in operator theory until its solution by [19], and much research had been done to show its equivalence with important results across many spheres of mathematics. For a survey covering the problem, its various formulations, and its implications, see [9].

In this section, we show how Theorem 4.1.4 implies Weaver's KS_r conjecture [31], a combinatorial form of the Kadison-Singer problem lying at the heart of discrepancy theory. The conjecture is stated as follows:

Theorem 4.3.1 (Conjecture KS_r)**.** *There exist universal constants $N \geq 2$ and $\epsilon > 0$ such that the following holds. Let $v_1, \ldots, v_n \in \mathbb{C}^k$ satisfy $||v_1||_2 \leq 1$ for all i, and suppose that*

$$\sum_i |\langle u, v_i\rangle|^2 \leq N$$

for every unit vector $u \in \mathbb{C}^k$. Then there exists a partition $X_1, \ldots, X_r$ of $\{1, \ldots, n\}$ such that

$$\sum_{i \in X_j} |\langle u, v_i\rangle|^2 \leq N - \epsilon$$

for every unit vector $u \in \mathbb{C}^k$ and all j.

Due to results of Akemann and Anderson [3] which require deep operator theory, in order to show that Conjecture KS_r implies the Kadison-Singer problem it is sufficient to show that it implies the following proposition.

Proposition 4.3.2 (Part of Theorem 1, [31])**.** *Let $\mathbf{P} \in C^{n\times n}$ be an orthogonal projection with $\max \mathbf{P}_{ii} \leq 1/N$. Then there exist diagonal projections $\mathbf{Q}_1, \ldots, \mathbf{Q}_r \in \mathbb{C}^{n\times n}$ such that $\sum \mathbf{Q}_j = \mathbf{I}_n$ and $||\mathbf{Q}_j\mathbf{P}\mathbf{Q}_j|| \leq 1 - \epsilon/N$ for all j.*

Reduction from KS_r. Suppose Conjecture KS_r holds for some r, N, and ϵ. Let $\mathbf{P} \in \mathbb{C}^{n\times n}$ be an orthogonal projection and let $\rho = \mathrm{rank}(\mathbf{P})$. Let $\mathbf{v}_i = \sqrt{N}\cdot\mathbf{P}\mathbf{e}_i$ for $i \in [n]$, so that $||\mathbf{v}_i||_2^2 = N||\mathbf{P}\mathbf{e}_i||_2^2 = N\langle\mathbf{P}\mathbf{e}_i, \mathbf{e}_i\rangle \leq N\max\mathbf{P}_{ii} \leq 1$. In addition, for any unit vector $u \in \mathrm{im}(\mathbf{P})$,

$$\sum_i |\langle \mathbf{u}, \mathbf{v}_i\rangle|^2 = \sum_i |\langle \mathbf{u}, \sqrt{N}\mathbf{P}\mathbf{e}_i\rangle|^2 = N\sum_i |\langle \mathbf{u}, \mathbf{e}_i\rangle|^2 = N.$$

Restricting to the subspace $\mathrm{im}(\mathbf{P})$ and invoking Conjecture KS_r, we obtain a partition $X_1, \dots, X_r$ of $[n]$ such that

$$\sum_{i \in X_j} |\langle u, v_i \rangle|^2 \leq N - \epsilon$$

for every unit vector $u \in \mathrm{im}(\mathbf{P})$ and all j. For $1 \leq j \leq r$, let $\mathbf{Q}_j \in \mathbb{C}^{n \times n}$ be the projection which zeroes out coordinates not in X_j: $\mathbf{Q}_j \mathbf{e}_k = 1$ if $k \in X_j$ and $\mathbf{Q}_j \mathbf{e}_k = 0$ otherwise, so that $\sum_j \mathbf{Q}_j = \mathbf{I}_n$. Then for any unit vector $\mathbf{u} \in \mathrm{im}(\mathbf{P})$, using that $\mathbf{P}$ and $\mathbf{Q}_j$ are self-adjoint,

$$||\mathbf{Q}_j \mathbf{P} \mathbf{u}||_2^2 = \sum_i |\langle \mathbf{Q}_j \mathbf{P} \mathbf{u}, \mathbf{e}_i \rangle|^2 = \sum_i |\langle \mathbf{u}, \mathbf{P} \mathbf{Q}_j \mathbf{e}_i \rangle|^2 = 1/N \sum_{i \in X_j} |\langle \mathbf{u}, \mathbf{v}_i \rangle|^2 \leq 1 - \epsilon/N.$$

Thus $||\mathbf{Q}_j \mathbf{P} \mathbf{Q}_j|| = ||\mathbf{Q}_j \mathbf{P}||^2 \leq 1 - \epsilon/N$ for all j. □

Finally, we now show that Theorem 4.1.1 implies the KS_2 conjecture. Marcus, Spielman and Srivastava prove a slightly stronger version:

Proposition 4.3.3. *Let $\mathbf{u}_1, \dots, \mathbf{u}_n$ be column vectors in $\mathbb{C}^d$ such that $\sum \mathbf{u}_i \mathbf{u}_i^* = \mathbf{I}$ and $||\mathbf{u}_i||^2 \leq L$ for all i. Then there exists a partition of $\{1, \dots, n\}$ into sets X_1 and X_2 such that*

$$\left|\left| \sum_{i \in X_j} \mathbf{u}_i \mathbf{u}_i^* \right|\right| \leq \frac{(1 + \sqrt{2L})^2}{2}.$$

Reduction to KS_2. Let $N = 18$, let $\mathbf{u}_i = \mathbf{v}_i / \sqrt{N}$ so that $\sum \mathbf{u}_i \mathbf{u}_i^* = \mathbf{I}$, and let $L = 1/N$. This yields KS_2 for $\epsilon = 2$. □

Proof. Let $\mathbf{w}_1, \dots, \mathbf{w}_n \in \mathbb{C}^{2d}$ be random column vectors such that

$$\mathbf{w}_i = \begin{bmatrix} \sqrt{2}\mathbf{u}_i \\ \mathbf{0} \end{bmatrix} \text{ or } \mathbf{w}_i = \begin{bmatrix} \mathbf{0} \\ \sqrt{2}\mathbf{u}_i \end{bmatrix} \text{ each w.p. } 1/2.$$

Note that

$$\sum_{i=1}^n \mathbb{E}\, \mathbf{w}_i \mathbf{w}_i^* = \sum_{i=1}^n \begin{bmatrix} \mathbf{u}_i \mathbf{u}_i^* & \mathbf{0} \\ \mathbf{0} & \mathbf{u}_i \mathbf{u}_i^* \end{bmatrix} = \mathbf{I}$$

and $\mathbb{E}\, ||\mathbf{w}_i||^2 = 2||\mathbf{u}_i||^2 \leq 2L$, so that the conditions of Theorem 4.1.1 apply for the matrices $\mathbf{A}_i = \mathbf{u}_i \mathbf{u}_i^*$. Applying it with $\epsilon = 2L$, we obtain that there exists a subset $X_1 \in \{1, \dots, n\}$ such that, letting $X_2 = \{1, \dots, n\} \setminus X_1$:

$$\left|\left| \sum_{i \in X_1} \begin{bmatrix} \sqrt{2}\mathbf{u}_i \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \sqrt{2}\mathbf{u}_i \\ \mathbf{0} \end{bmatrix}^* + \sum_{i \in X_2} \begin{bmatrix} \mathbf{0} \\ \sqrt{2}\mathbf{u}_i \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \sqrt{2}\mathbf{u}_i \end{bmatrix}^* \right|\right| \leq (1 + \sqrt{2L})^2.$$

Rearranging, this bounds each of the terms as follows:

$$\left|\left| \sum_{i \in X_1} \begin{bmatrix} \mathbf{u}_i \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u}_i \\ \mathbf{0} \end{bmatrix}^* \right|\right| \leq \frac{(1 + \sqrt{2L})^2}{2} \text{ and } \left|\left| \sum_{i \in X_2} \begin{bmatrix} \mathbf{0} \\ \mathbf{u}_i \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{u}_i \end{bmatrix}^* \right|\right| \leq \frac{(1 + \sqrt{2L})^2}{2}$$

Since

$$\left|\left| \sum_{i \in X_1} \begin{bmatrix} \mathbf{u}_i \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u}_i \\ \mathbf{0} \end{bmatrix}^* \right|\right| = \left|\left| \sum_{i \in X_1} \mathbf{u}_i \mathbf{u}_i^* \right|\right| \text{ and } \left|\left| \sum_{i \in X_2} \begin{bmatrix} \mathbf{0} \\ \mathbf{u}_i \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{u}_i \end{bmatrix}^* \right|\right| = \left|\left| \sum_{i \in X_2} \mathbf{u}_i \mathbf{u}_i^* \right|\right|,$$

the claim follows. □

References

[1] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23:535–561, April 2010.

[2] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inf. Theor.*, 48(3):569–579, September 2006.

[3] C.A. Akemann and J. Anderson. *Lyapunov Theorems for Operator Algebras*. Number no. 458 in American Mathematical Society: Memoirs of the American Mathematical Society. American Mathematical Society, 1991.

[4] Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *CoRR*, abs/0808.0163, 2008.

[5] Andras A. Benczur and David R. Karger. Randomized approximation schemes for cuts and flows in capacitated graphs. *CoRR*, cs.DS/0207078, 2002.

[6] J. Borcea and P. Brändén. Multivariate Polya-Schur classification problems in the Weyl algebra. *ArXiv Mathematics e-prints*, June 2006.

[7] Julius Borcea and Petter Branden. Applications of stable polynomials to mixed determinants: Johnson's conjectures, unimodality, and symmetrized fischer products. *Duke Mathematical Journal*, 2008.

[8] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal column-based matrix reconstruction. *CoRR*, abs/1103.0995, 2011.

[9] Peter G. Casazza, Matthew Fickus, Janet C. Tremain, and Eric Weber. The kadisonâĂŞsinger problem in mathematics and engineering. In *Proc. Natl. Acad. Sci. USA 103 (2006*, pages 297–356, 2006.

[10] Marcel K. de Carli Silva, Nicholas J. A. Harvey, and Cristiane M. Sato. Sparse sums of positive semidefinite matrices. *CoRR*, abs/1107.0088, 2011.

[11] H. Dette and W.J. Studden. Some new asymptotic properties for the zeros of jacobi, laguerre, and hermite polynomials. *Constructive Approximation*, 11(2):227–238, 1995.

[12] S. Fisk. Polynomials, roots, and interlacing. *ArXiv Mathematics e-prints*, December 2006.

[13] Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

[14] William W. Hager. Updating the inverse of a matrix. *SIAM Review*, 31(2):pp. 221–239, 1989.

[15] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, May 2011.

[16] Richard V. Kadison and I. M. Singer. Extensions of pure states. *American Journal of Mathematics*, 81(2):pp. 383–400, 1959.

[17] A. Lubotzky, R. Phillips, and P. Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988.

[18] A. Marcus, D. A. Spielman, and N. Srivastava. Interlacing Families I: Bipartite Ramanujan Graphs of All Degrees. *ArXiv e-prints*, April 2013.

[19] A. Marcus, D. A Spielman, and N. Srivastava. Interlacing Families II: Mixed Characteristic Polynomials and the Kadison-Singer Problem. *ArXiv e-prints*, June 2013.

[20] Assaf Naor. Sparse quadratic forms and their geometric applications. *séminaire Bourbaki*, 1033, 2011.

[21] A. Nilli. On the second eigenvalue of a graph. *Discrete Mathematics*, 91(2):207 – 210, 1991.

[22] HaskellP. Rosenthal. On the subspaces of $l_p(p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8(3):273–303, 1970.

[23] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal*, pages 60–72, 1999.

[24] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, pages 563–568, New York, NY, USA, 2008. ACM.

[25] Daniel A. Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *SIAM J. Comput.*, 40(4):981–1025, 2011.

[26] N. Srivastava and R. Vershynin. Covariance estimation for distributions with $2 + \varepsilon$ moments. *ArXiv e-prints*, June 2011.

[27] Daniel B. Szyld. The many proofs of an identity on the norm of oblique projections. *Numer. Algorithms*, pages 309–323.

[28] Terence Tao. Real stable polynomials and the kadison-singer problem, November 2013.

[29] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *CoRR*, abs/1011.3027, 2010.

[30] D. G. Wagner. Multivariate stable polynomials: theory and applications. *ArXiv e-prints*, November 2009.

[31] N. Weaver. The Kadison-Singer problem in discrepancy theory. *ArXiv Mathematics e-prints*, September 2002.

[32] Pierre Youssef. Estimating the covariance of random matrices. *Electron. J. Probab.*, 18:no. 107, 1–26, 2013.