An Introduction to Matrix Concentration Inequalities

Joel A. Tropp

24 December 2014
FnTML Draft, Revised

I

For Margot and Benjamin

Contents

Preface

In recent years, random matrices have come to play a major role in computational mathematics, but most of the classical areas of random matrix theory remain the province of experts. Over the last decade, with the advent of matrix concentration inequalities, research has advanced to the point where we can conquer many (formerly) challenging problems with a page or two of arithmetic.

My aim is to describe the most successful methods from this area along with some interesting examples that these techniques can illuminate. I hope that the results in these pages will inspire future work on applications of random matrices as well as refinements of the matrix concentration inequalities discussed herein.

I have chosen to present a coherent body of results based on a generalization of the Laplace transform method for establishing scalar concentration inequalities. In the last two years, Lester Mackey and I, together with our coauthors, have developed an alternative approach to matrix concentration using exchangeable pairs and Markov chain couplings. With some regret, I have chosen to omit this theory because the ideas seem less accessible to a broad audience of researchers. The interested reader will find pointers to these articles in the annotated bibliography.

The work described in these notes reflects the influence of many researchers. These include Rudolf Ahlswede, Rajendra Bhatia, Eric Carlen, Sourav Chatterjee, Edward Effros, Elliott Lieb, Roberto Imbuzeiro Oliveira, Dénes Petz, Gilles Pisier, Mark Rudelson, Roman Vershynin, and Andreas Winter. I have also learned a great deal from other colleagues and friends along the way.

I would like to thank some people who have helped me improve this work. Several readers informed me about errors in the initial version of this manuscript; these include Serg Bogdanov, Peter Forrester, Nikos Karampatziakis, and Guido Lagos. The anonymous reviewers tendered many useful suggestions, and they pointed out a number of errors. Sid Barman gave me feedback on the final revisions to the monograph. Last, I want to thank Léon Nijensohn for his continuing encouragement.

I gratefully acknowledge financial support from the Office of Naval Research under awards N00014-08-1-0883 and N00014-11-1002, the Air Force Office of Strategic Research under award FA9550-09-1-0643, and an Alfred P. Sloan Fellowship. Some of this research was completed at the Institute of Pure and Applied Mathematics at UCLA. I would also like to thank the California Institute of Technology and the Moore Foundation.

Joel A. Tropp
Pasadena, CA
December 2012
Revised, March 2014 and December 2014

CHAPTER 1

Introduction

Random matrix theory has grown into a vital area of probability, and it has found applications in many other fields. To motivate the results in this monograph, we begin with an overview of the connections between random matrix theory and computational mathematics. We introduce the basic ideas underlying our approach, and we state one of our main results on the behavior of random matrices. As an application, we examine the properties of the sample covariance estimator, a random matrix that arises in statistics. Afterward, we summarize the other types of results that appear in these notes, and we assess the novelties in this presentation.

1.1 Historical Origins

Random matrix theory sprang from several different sources in the first half of the 20th century.

Geometry of Numbers. Peter Forrester [For10, p. v] traces the field of random matrix theory to work of Hurwitz, who defined the invariant integral over a Lie group. Specializing this analysis to the orthogonal group, we can reinterpret this integral as the expectation of a function of a uniformly random orthogonal matrix.

Multivariate Statistics. Another early example of a random matrix appeared in the work of John Wishart [Wis28]. Wishart was studying the behavior of the sample covariance estimator for the covariance matrix of a multivariate normal random vector. He showed that the estimator, which is a random matrix, has the distribution that now bears his name. Statisticians have often used random matrices as models for multivariate data [MKB79, Mui82].

Numerical Linear Algebra. In their remarkable work [vNG47, GvN51] on computational methods for solving systems of linear equations, von Neumann and Goldstine considered a random matrix model for the floating-point errors that arise from an LU decomposition.[1] They obtained a high-probability bound for the norm of the random matrix, which they

[1] von Neumann and Goldstine invented and analyzed this algorithm before they had any digital computer on which to implement it! See [Grc11] for a historical account.

took as an estimate for the error the procedure might typically incur. Curiously, in subsequent years, numerical linear algebraists became very suspicious of probabilistic techniques, and only in recent years have randomized algorithms reappeared in this field. See the surveys [Mah11, HMT11, Woo14] for more details and references.

Nuclear Physics. In the early 1950s, physicists had reached the limits of deterministic analytical techniques for studying the energy spectra of heavy atoms undergoing slow nuclear reactions. Eugene Wigner was the first researcher to surmise that a random matrix with appropriate symmetries might serve as a suitable model for the Hamiltonian of the quantum mechanical system that describes the reaction. The eigenvalues of this random matrix model the possible energy levels of the system. See Mehta's book [Meh04, §1.1] for an account of all this.

In each area, the motivation was quite different and led to distinct sets of questions. Later, random matrices began to percolate into other fields such as graph theory (the Erdős–Rényi model [ER60] for a random graph) and number theory (as a model for the spacing of zeros of the Riemann zeta function [Mon73]).

1.2 The Modern Random Matrix

By now, random matrices are ubiquitous. They arise throughout modern mathematics and statistics, as well as in many branches of science and engineering. Random matrices have several different purposes that we may wish to distinguish. They can be used within randomized computer algorithms; they serve as models for data and for physical phenomena; and they are subjects of mathematical inquiry. This section offers a taste of these applications. Note that the ideas and references here reflect the author's interests, and they are far from comprehensive!

1.2.1 Algorithmic Applications

The striking mathematical properties of random matrices can be harnessed to develop algorithms for solving many different problems.

Computing Matrix Approximations. Random matrices can be used to develop fast algorithms for computing a truncated singular-value decomposition. In this application, we multiply a large input matrix by a smaller random matrix to extract information about the dominant singular vectors of the input matrix. The seed of this idea appears in [FKV98, DFK+99]. The survey [HMT11] explains how to implement this method in practice, while the two monographs [Mah11, Woo14] cover more theoretical aspects.

Sparsification. One way to accelerate spectral computations on large matrices is to replace the original matrix by a sparse proxy that has similar spectral properties. An elegant way to produce the sparse proxy is to zero out entries of the original matrix at random while rescaling the entries that remain. This approach was proposed in [AM01, AM07], and the papers [AKL13, KD14] contain recent innovations. Related ideas play an important role in Spielman and Teng's work [ST04] on fast algorithms for solving linear systems.

Subsampling of Data. In large-scale machine learning, one may need to subsample data randomly to reduce the computational costs of fitting a model. For instance, we can combine

random sampling with the Nyström decomposition to obtain a randomized approximation of a kernel matrix. This method was introduced by Williams & Seeger [WS01]. The paper [DM05] provides the first theoretical analysis, and the survey [GM14] contains more complete results.

Dimension Reduction. A basic template in the theory of algorithms invokes randomized projection to reduce the dimension of a computational problem. Many types of dimension reduction are based on properties of random matrices. The two papers [JL84, Bou85] established the mathematical foundations of this approach. The earliest applications in computer science appear in the work [LLR95]. Many contemporary variants depend on ideas from [AC09] and [CW13].

Combinatorial Optimization. One approach to solving a computationally difficult optimization problem is to relax (i.e., enlarge) the constraint set so the problem becomes tractable, to solve the relaxed problem, and then to use a randomized procedure to map the solution back to the original constraint set [BTN01, §4.3]. This technique is called *relaxation and rounding*. For hard optimization problems involving a matrix variable, the analysis of the rounding procedure often involves ideas from random matrix theory [So09, NRV13].

Compressed Sensing. When acquiring data about an object with relatively few degrees of freedom as compared with the ambient dimension, we may be able to sieve out the important information from the object by taking a small number of random measurements, where the number of measurements is comparable to the number of degrees of freedom [GGI+02, CRT06, Don06]. This observation is now referred to as *compressed sensing*. Random matrices play a central role in the design and analysis of measurement procedures. For example, see [FR13, CRPW12, ALMT14, Tro14].

1.2.2 Modeling

Random matrices also appear as models for multivariate data or multivariate phenomena. By studying the properties of these models, we may hope to understand the typical behavior of a data-analysis algorithm or a physical system.

Sparse Approximation for Random Signals. Sparse approximation has become an important problem in statistics, signal processing, machine learning and other areas. One model for a "typical" sparse signal poses the assumption that the nonzero coefficients that generate the signal are chosen at random. When analyzing methods for identifying the sparse set of coefficients, we must study the behavior of a random column submatrix drawn from the model matrix [Tro08a, Tro08b].

Demixing of Structured Signals. In data analysis, it is common to encounter a mixture of two structured signals, and the goal is to extract the two signals using prior information about the structures. A common model for this problem assumes that the signals are randomly oriented with respect to each other, which means that it is usually possible to discriminate the underlying structures. Random orthogonal matrices arise in the analysis of estimation techniques for this problem [MT14, ALMT14, MT13].

Stochastic Block Model. One probabilistic framework for describing community structure in a network assumes that each pair of individuals in the same community has a relationship with high probability, while each pair of individuals drawn from different communities has a relationship with lower probability. This is referred to as the *stochastic block model* [HLL83]. It is quite common to analyze algorithms for extracting community structure from data by positing that this model holds. See [ABH14] for a recent contribution, as well as a summary of the extensive literature.

High-Dimensional Data Analysis. More generally, random models are pervasive in the analysis of statistical estimation procedures for high-dimensional data. Random matrix theory plays a key role in this field [MKB79, Mui82, Kol11, BvdG11].

Wireless Communication. Random matrices are commonly used as models for wireless channels. See the book of Tulino and Verdú for more information [TV04].

In these examples, it is important to recognize that random models may not coincide very well with reality, but they allow us to get a sense of what might be possible in some generic cases.

1.2.3 Theoretical Aspects

Random matrices are frequently studied for their intrinsic mathematical interest. In some fields, they provide examples of striking phenomena. In other areas, they furnish counterexamples to "intuitive" conjectures. Here are a few disparate problems where random matrices play a role.

Combinatorics. An expander graph has the property that every small set of vertices has edges linking it to a large proportion of the vertices. The expansion property is closely related to the spectral behavior of the adjacency matrix of the graph. The easiest construction of an expander involves a random matrix argument [AS00, §9.2].

Numerical Analysis. For worst-case examples, the Gaussian elimination method for solving a linear system is not numerically stable. In practice, however, stability problems rarely arise. One explanation for this phenomenon is that, with high probability, a small random perturbation of any fixed matrix is well conditioned. As a consequence, it can be shown that Gaussian elimination is stable for most matrices [SST06].

High-Dimensional Geometry. Dvoretzky's Theorem states that, when N is large, the unit ball of each N-dimensional Banach space has a slice of dimension $n \approx \log N$ that is close to a Euclidean ball with dimension n. It turns out that a *random* slice of dimension n realizes this property [Mil71]. This result can be framed as a statement about spectral properties of a random matrix [Gor85].

Quantum Information Theory. Random matrices appear as counterexamples for a number of conjectures in quantum information theory. Here is one instance. In classical information theory, the total amount of information that we can transmit through a pair of channels equals the sum of the information we can send through each channel separately. It was conjectured that the same property holds for quantum channels. In fact, a pair of quantum channels can have strictly larger capacity than a single channel. This result depends on a random matrix construction [Has09]. See [HW08] for related work.

1.3 Random Matrices for the People

Historically, random matrix theory has been regarded as a very challenging field. Even now, many well-established methods are only comprehensible to researchers with significant experience, and it may take months of intensive effort to prove new results. There are a small number of classes of random matrices that have been studied so completely that we know almost everything about them. Yet, moving beyond this *terra firma*, one quickly encounters examples where classical methods are brittle.

We hope to democratize random matrix theory. These notes describe tools that deliver useful information about a wide range of random matrices. In many cases, a modest amount of straightforward arithmetic leads to strong results. The methods here should be accessible to computational scientists working in a variety of fields. Indeed, the techniques in this work have already found an extensive number of applications.

1.4 Basic Questions in Random Matrix Theory

Random matrices merit special attention because they have spectral properties that are quite different from familiar deterministic matrices. Here are some of the questions we might want to investigate.

- What is the expectation of the maximum eigenvalue of a random Hermitian matrix? What about the minimum eigenvalue?
- How is the maximum eigenvalue of a random Hermitian matrix distributed? What is the probability that it takes values substantially different from its mean? What about the minimum eigenvalue?
- What is the expected spectral norm of a random matrix? What is the probability that the norm takes a value substantially different from its mean?
- What about the other eigenvalues or singular values? Can we say something about the "typical" spectrum of a random matrix?
- Can we say anything about the eigenvectors or singular vectors? For instance, is each one distributed almost uniformly on the sphere?
- We can also ask questions about the operator norm of a random matrix acting as a map between two normed linear spaces. In this case, the geometry of the domain and codomain play a role.

In this work, we focus on the first three questions above. We study the expectation of the extreme eigenvalues of a random Hermitian matrix, and we attempt to provide bounds on the probability that they take an unusual value. As an application of these results, we can control the expected spectral norm of a general matrix and bound the probability of a large deviation. These are the most relevant problems in many (but not all!) applications. The remaining questions are also important, but we will not touch on them here. We recommend the book [Tao12] for a friendly introduction to other branches of random matrix theory.

1.5 Random Matrices as Independent Sums

Our approach to random matrices depends on a fundamental principle:

> **In applications, it is common that a random matrix can be expressed as a sum of independent random matrices.**

The examples that appear in these notes should provide ample evidence for this claim. For now, let us describe a specific problem that will serve as an illustration throughout the Introduction. We hope this example is complicated enough to be interesting but simple enough to elucidate the main points.

1.5.1 Example: The Sample Covariance Estimator

Let $\boldsymbol{x} = (X_1, \dots, X_p)$ be a complex random vector with zero mean: $\mathbb{E}\,\boldsymbol{x} = \mathbf{0}$. The *covariance matrix* $\boldsymbol{A}$ of the random vector $\boldsymbol{x}$ is the positive-semidefinite matrix

$$\boldsymbol{A} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^*) = \sum_{j,k=1}^{p} \mathbb{E}\left(X_j X_k^*\right) \mathbf{E}_{jk} \tag{1.5.1}$$

The star * refers to the conjugate transpose operation, and the standard basis matrix $\mathbf{E}_{jk}$ has a one in the (j,k) position and zeros elsewhere. In other words, the (j,k) entry of the sample covariance matrix $\boldsymbol{A}$ records the covariance between the jth and kth entry of the vector $\boldsymbol{x}$.

One basic problem in statistical practice is to estimate the covariance matrix from data. Imagine that we have access to n independent samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, each distributed the same way as $\boldsymbol{x}$. The *sample covariance estimator* $\boldsymbol{Y}$ is the random matrix

$$\boldsymbol{Y} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{x}_k \boldsymbol{x}_k^*. \tag{1.5.2}$$

The random matrix $\boldsymbol{Y}$ is an unbiased estimator[2] for the sample covariance matrix: $\mathbb{E}\,\boldsymbol{Y} = \boldsymbol{A}$. Observe that the sample covariance estimator $\boldsymbol{Y}$ fits neatly into our paradigm:

> **The sample covariance estimator can be expressed as a sum of independent random matrices.**

This is precisely the type of decomposition that allows us to apply the tools in these notes.

1.6 Exponential Concentration Inequalities for Matrices

An important challenge in probability theory is to study the probability that a real random variable Z takes a value substantially different from its mean. That is, we seek a bound of the form

$$\mathbb{P}\{|Z - \mathbb{E}\,Z| \geq t\} \leq \underline{\quad ??? \quad} \tag{1.6.1}$$

[2] The formula (1.5.2) supposes that the random vector $\boldsymbol{x}$ is known to have zero mean. Otherwise, we have to make an adjustment to incorporate an estimate for the sample mean.

for a positive parameter t. When Z is expressed as a sum of independent random variables, the literature contains many tools for addressing this problem. See [BLM13] for an overview.

For a random matrix $\boldsymbol{Z}$, a variant of (1.6.1) is the question of whether $\boldsymbol{Z}$ deviates substantially from its mean value. We might frame this question as

$$\mathbb{P}\{\|\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}\| \geq t\} \leq \underline{\quad ??? \quad}. \tag{1.6.2}$$

Here and elsewhere, $\|\cdot\|$ denotes the spectral norm of a matrix. As noted, it is frequently possible to decompose $\boldsymbol{Z}$ as a sum of independent random matrices. We might even dream that established methods for studying the scalar concentration problem (1.6.1) extend to (1.6.2).

1.6.1 The Bernstein Inequality

To explain what kind of results we have in mind, let us return to the scalar problem (1.6.1). First, to simplify formulas, we assume that the real random variable Z has zero mean: $\mathbb{E}Z = 0$. If not, we can simply center the random variable by subtracting its mean. Second, and more restrictively, we suppose that Z can be expressed as a sum of independent, real random variables.

To control Z, we rely on two types of information: global properties of the sum (such as its mean and variance) and local properties of the summands (such as their maximum fluctuation). These pieces of data are usually easy to obtain. Together, they determine how well Z concentrates around zero, its mean value.

Theorem 1.6.1 (Bernstein Inequality)**.** *Let $S_1, \dots, S_n$ be independent, centered, real random variables, and assume that each one is uniformly bounded:*

$$\mathbb{E}S_k = 0 \quad \text{and} \quad |S_k| \leq L \quad \text{for each } k = 1, \dots, n.$$

Introduce the sum $Z = \sum_{k=1}^{n} S_k$, and let $v(Z)$ denote the variance of the sum:

$$v(Z) = \mathbb{E}Z^2 = \sum_{k=1}^{n} \mathbb{E}S_k^2.$$

Then

$$\mathbb{P}\{|Z| \geq t\} \leq 2\exp\left(\frac{-t^2/2}{v(Z) + Lt/3}\right) \quad \text{for all } t \geq 0.$$

See [BLM13, §2.8] for a proof of this result. We refer to Theorem 1.6.1 as an *exponential concentration inequality* because it yields exponentially decaying bounds on the probability that Z deviates substantially from its mean.

1.6.2 The Matrix Bernstein Inequality

What is truly astonishing is that the scalar Bernstein inequality, Theorem 1.6.1, lifts directly to matrices. Let us emphasize this remarkable fact:

> **There are exponential concentration inequalities for the spectral norm of a sum of independent random matrices.**

As a consequence, once we decompose a random matrix as an independent sum, we can harness global properties (such as the mean and the variance) and local properties (such as a uniform bound on the summands) to obtain detailed information about the norm of the sum. As in the scalar case, it is usually easy to acquire the input data for the inequality. But the output of the inequality is highly nontrivial.

To illustrate these claims, we will state one of the major results from this monograph. This theorem is a matrix extension of Bernstein's inequality that was developed independently in the two papers [Oli10a, Tro11c]. After presenting the result, we give some more details about its interpretation. In the next section, we apply this result to study the covariance estimation problem.

Theorem 1.6.2 (Matrix Bernstein)**.** *Let $\boldsymbol{S}_1, \dots, \boldsymbol{S}_n$ be independent, centered random matrices with common dimension $d_1 \times d_2$, and assume that each one is uniformly bounded*

$$\mathbb{E}\,\boldsymbol{S}_k = \mathbf{0} \quad \text{and} \quad \|\boldsymbol{S}_k\| \le L \quad \text{for each } k = 1, \dots, n.$$

Introduce the sum

$$\boldsymbol{Z} = \sum_{k=1}^{n} \boldsymbol{S}_k, \tag{1.6.3}$$

and let $v(\boldsymbol{Z})$ denote the matrix variance statistic of the sum:

$$\begin{aligned} v(\boldsymbol{Z}) &= \max\left\{ \|\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*)\|, \|\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})\| \right\} \\ &= \max\left\{ \left\| \sum_{k=1}^{n} \mathbb{E}\left(\boldsymbol{S}_k \boldsymbol{S}_k^*\right) \right\|, \left\| \sum_{k=1}^{n} \mathbb{E}\left(\boldsymbol{S}_k^* \boldsymbol{S}_k\right) \right\| \right\}. \end{aligned} \tag{1.6.4}$$

Then

$$\mathbb{P}\{\|\boldsymbol{Z}\| \ge t\} \le (d_1 + d_2) \cdot \exp\left(\frac{-t^2/2}{v(\boldsymbol{Z}) + Lt/3} \right) \quad \text{for all } t \ge 0. \tag{1.6.5}$$

Furthermore,

$$\mathbb{E}\,\|\boldsymbol{Z}\| \le \sqrt{2v(\boldsymbol{Z}) \log(d_1 + d_2)} + \frac{1}{3} L \log(d_1 + d_2). \tag{1.6.6}$$

The proof of this result appears in Chapter 6.

To appreciate what Theorem 1.6.2 means, it is valuable to make a direct comparison with the scalar version, Theorem 1.6.1. In both cases, we express the object of interest as an independent sum, and we instate a uniform bound on the summands. There are three salient changes:

- The variance $v(\boldsymbol{Z})$ in the result for matrices can be interpreted as the magnitude of the expected squared deviation of $\boldsymbol{Z}$ from its mean. The formula reflects the fact that a general matrix $\boldsymbol{B}$ has *two* different squares $\boldsymbol{B}\boldsymbol{B}^*$ and $\boldsymbol{B}^*\boldsymbol{B}$. For an Hermitian matrix, the two squares coincide.

- The tail bound has a dimensional factor $d_1 + d_2$ that depends on the size of the matrix. This factor reduces to two in the scalar setting. In the matrix case, it limits the range of t where the tail bound is informative.

- We have included a bound for $\mathbb{E}\,\|\boldsymbol{Z}\|$. This estimate is not particularly interesting in the scalar setting, but it is usually quite challenging to prove results of this type for matrices. In fact, the expectation bound is often more useful than the tail bound.

The latter point deserves amplification:

> **The expectation bound** (1.6.6) **is the most important aspect of the matrix Bernstein inequality.**

For further discussion of this result, turn to Chapter 6. Chapters 4 and 7 contain related results and interpretations.

1.6.3 Example: The Sample Covariance Estimator

We will apply the matrix Bernstein inequality, Theorem 1.6.2, to measure how well the sample covariance estimator approximates the true covariance matrix. As before, let $\boldsymbol{x}$ be a zero-mean random vector with dimension p. Introduce the $p \times p$ covariance matrix $\boldsymbol{A} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^*)$. Suppose we have n independent samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ with the same distribution as $\boldsymbol{x}$. Form the $p \times p$ sample covariance estimator

$$\boldsymbol{Y} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{x}_k \boldsymbol{x}_k^*.$$

Our goal is to study how the spectral-norm distance $\|\boldsymbol{Y} - \boldsymbol{A}\|$ between the sample covariance and the true covariance depends on the number n of samples.

For simplicity, we will perform the analysis under the extra assumption that the ℓ_2 norm of the random vector is bounded: $\|\boldsymbol{x}\|^2 \le B$. This hypothesis can be relaxed if we apply a variant of the matrix Bernstein inequality that reflects the typical magnitude of a summand $\boldsymbol{S}_k$. One such variant appears in the formula (6.1.6).

We are in a situation where it is quite easy to see how the matrix Bernstein inequality applies. Define the random deviation $\boldsymbol{Z}$ of the estimator $\boldsymbol{Y}$ from the true covariance matrix $\boldsymbol{A}$:

$$\boldsymbol{Z} = \boldsymbol{Y} - \boldsymbol{A} = \sum_{k=1}^{n} \boldsymbol{S}_k \quad \text{where} \quad \boldsymbol{S}_k = \frac{1}{n}\big(\boldsymbol{x}_k \boldsymbol{x}_k^* - \boldsymbol{A}\big) \quad \text{for each index } k.$$

The random matrices $\boldsymbol{S}_k$ are independent, identically distributed, and centered. To apply Theorem 1.6.2, we need to find a uniform bound L for the summands, and we need to control the matrix variance statistic $v(\boldsymbol{Z})$.

First, let us develop a uniform bound on the spectral norm of each summand. We may calculate that

$$\|\boldsymbol{S}_k\| = \frac{1}{n} \|\boldsymbol{x}_k \boldsymbol{x}_k^* - \boldsymbol{A}\| \le \frac{1}{n}\big(\|\boldsymbol{x}_k \boldsymbol{x}_k^*\| + \|\boldsymbol{A}\|\big) \le \frac{2B}{n}.$$

The first relation is the triangle inequality. The second follows from the assumption that $\boldsymbol{x}$ is bounded and the observation that

$$\|\boldsymbol{A}\| = \|\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^*)\| \le \mathbb{E}\,\|\boldsymbol{x}\boldsymbol{x}^*\| = \mathbb{E}\,\|\boldsymbol{x}\|^2 \le B.$$

This expression depends on Jensen's inequality and the hypothesis that $\boldsymbol{x}$ is bounded.

Second, we need to bound the matrix variance statistic $v(\boldsymbol{Z})$ defined in (1.6.4). The matrix $\boldsymbol{Z}$ is Hermitian, so the two squares in this formula coincide with each other:

$$v(\boldsymbol{Z}) = \big\|\mathbb{E}\,\boldsymbol{Z}^2\big\| = \left\| \sum_{k=1}^{n} \mathbb{E}\,\boldsymbol{S}_k^2 \right\|.$$

We need to determine the variance of each summand. By direct calculation,

$$\begin{aligned}\mathbb{E}\,\boldsymbol{S}_k^2 = \frac{1}{n^2}\,\mathbb{E}\left(\boldsymbol{x}_k\boldsymbol{x}_k^* - \boldsymbol{A}\right)^2 &= \frac{1}{n^2}\,\mathbb{E}\left[\,\|\boldsymbol{x}_k\|^2\cdot\boldsymbol{x}_k\boldsymbol{x}_k^* - \left(\boldsymbol{x}_k\boldsymbol{x}_k^*\right)\boldsymbol{A} - \boldsymbol{A}\left(\boldsymbol{x}_k\boldsymbol{x}_k^*\right) + \boldsymbol{A}^2\right] \\ &\preccurlyeq \frac{1}{n^2}\left[B\cdot\mathbb{E}\left(\boldsymbol{x}_k\boldsymbol{x}_k^*\right) - \boldsymbol{A}^2 - \boldsymbol{A}^2 + \boldsymbol{A}^2\right] \\ &\preccurlyeq \frac{B}{n^2}\cdot\boldsymbol{A}\end{aligned}$$

The expression $\boldsymbol{H} \preccurlyeq \boldsymbol{T}$ means that $\boldsymbol{T} - \boldsymbol{H}$ is positive semidefinite. We used the norm bound for the random vector $\boldsymbol{x}$ and the fact that expectation preserves the semidefinite order. In the last step, we dropped the negative-semidefinite term $-\boldsymbol{A}^2$. Summing this relation over k, we reach

$$\mathbf{0} \preccurlyeq \sum_{k=1}^{n}\mathbb{E}\,\boldsymbol{S}_k^2 \preccurlyeq \frac{B}{n}\cdot\boldsymbol{A}.$$

The matrix is positive-semidefinite because it is a sum of squares of Hermitian matrices. Extract the spectral norm to arrive at

$$v(\boldsymbol{Z}) = \left\|\sum_{k=1}^{n}\mathbb{E}\,\boldsymbol{S}_k^2\right\| \le \frac{B\,\|\boldsymbol{A}\|}{n}.$$

We have now collected the information we need to analyze the sample covariance estimator.

We can invoke the estimate (1.6.6) from the matrix Bernstein inequality, Theorem 1.6.2, with the uniform bound $L = 2B/n$ and the variance bound $v(\boldsymbol{Z}) \le B\,\|\boldsymbol{A}\|/n$. We attain

$$\mathbb{E}\,\|\boldsymbol{Y}-\boldsymbol{A}\| = \mathbb{E}\,\|\boldsymbol{Z}\| \le \sqrt{\frac{2B\,\|\boldsymbol{A}\|\log(2p)}{n}} + \frac{2B\log(2p)}{3n}.$$

In other words, the error in approximating the sample covariance matrix is not too large when we have a sufficient number of samples. If we wish to obtain a relative error on the order of ε, we may take

$$n \ge \frac{2B\log(2p)}{\varepsilon^2\,\|\boldsymbol{A}\|}.$$

This selection yields

$$\mathbb{E}\,\|\boldsymbol{Y}-\boldsymbol{A}\| \le \left(\varepsilon+\varepsilon^2\right)\cdot\|\boldsymbol{A}\|.$$

It is often the case that $B = \mathrm{Const}\cdot p$, so we discover that $n = \mathrm{Const}\cdot\varepsilon^{-2}p\log p$ samples are sufficient for the sample covariance estimator to provide a relatively accurate estimate of the true covariance matrix $\boldsymbol{A}$. This bound is qualitatively sharp for worst-case distributions.

The analysis in this section applies to many other examples. We encapsulate the argument in Corollary 6.2.1, which we use to study several more problems.

1.6.4 History of this Example

Covariance estimation may be the earliest application of matrix concentration tools in random matrix theory. Rudelson [Rud99], building on a suggestion of Pisier, showed how to use the noncommutative Khintchine inequality [LP86, LPP91, Buc01, Buc05] to obtain essentially optimal bounds on the sample covariance estimator of a bounded random vector. The tutorial [Ver12] of Roman Vershynin offers an overview of this problem as well as many results and references. The analysis of the sample covariance matrix here is adapted from the technical [GT14]. It leads to a result similar with the one Rudelson obtained in [Rud99].

1.6.5 Optimality of the Matrix Bernstein Inequality

Theorem 1.6.2 can be sharpened very little because it applies to every random matrix $\boldsymbol{Z}$ of the form (1.6.3). Let us say a few words about optimality now, postponing the details to §6.1.2.

Suppose that $\boldsymbol{Z}$ is a random matrix of the form (1.6.3). To make the comparison simpler, we also insist that each summand $\boldsymbol{S}_k$ is a *symmetric* random variable; that is, $\boldsymbol{S}_k$ and $-\boldsymbol{S}_k$ have the same distribution for each index k. Introduce the quantity

$$L_\star^2 = \mathbb{E} \max_k \|\boldsymbol{S}_k\|^2 .$$

In §6.1.2, we will argue that these assumptions imply

$$\begin{aligned} \text{const} \cdot \big[v(\boldsymbol{Z}) \;+\; L_\star^2 \big] \quad &\leq \quad \mathbb{E}\|\boldsymbol{Z}\|^2 \\ &\leq \quad \text{Const} \cdot \big[v(\boldsymbol{Z}) \log(d_1 + d_2) \;+\; L_\star^2 \log^2(d_1 + d_2) \big]. \end{aligned} \tag{1.6.7}$$

In other words, the scale of $\mathbb{E}\|\boldsymbol{Z}\|^2$ must depend on the matrix variance statistic $v(\boldsymbol{Z})$ and the average upper bound $L_\star^2$ for the summands. The quantity $L = \sup \|\boldsymbol{S}_k\|$ that appears in the matrix Bernstein inequality always exceeds $L_\star$, sometimes by a large margin, but they capture the same type of information.

The significant difference between the lower and upper bound in (1.6.7) comes from the dimensional factor $\log(d_1 + d_2)$. There are random matrices $\boldsymbol{Z}$ for which the lower bound gives a more accurate reflection of $\mathbb{E}\|\boldsymbol{Z}\|^2$, but there are also many random matrices where the upper bound describes the behavior correctly. At present, there is no method known for distinguishing between these two extremes under the model (1.6.3) for the random matrix.

The tail bound (1.6.5) provides a useful tool in practice, but it is not necessarily the best way to collect information about large deviation probabilities. To obtain more precise results, we recommend using the expectation bound (1.6.6) to control $\mathbb{E}\|\boldsymbol{Z}\|$ and then applying scalar concentration inequalities to estimate $\mathbb{P}\{\|\boldsymbol{Z}\| \geq \mathbb{E}\|\boldsymbol{Z}\| + t\}$. The book [BLM13] offers a good treatment of the methods that are available for establishing scalar concentration.

1.7 The Arsenal of Results

The Bernstein inequality is probably the most familiar exponential tail bound for a sum of independent random variables, but there are many more. It turns out that essentially all of these scalar results admit extensions that hold for random matrices. In fact, many of the established techniques for scalar concentration have analogs in the matrix setting.

1.7.1 What's Here...

This monograph focuses on a few key exponential concentration inequalities for a sum of independent random matrices, and it describes some specific applications of these results.

Matrix Gaussian Series. A matrix Gaussian series is a random matrix that can be expressed as a sum of fixed matrices, each weighted by an independent standard normal random variable. This formulation includes a surprising number of examples. The most important are undoubtedly Wigner matrices and rectangular Gaussian matrices. Another interesting case is a Toeplitz matrix with Gaussian entries. The analysis of matrix Gaussian series appears in Chapter 4.

Matrix Rademacher Series. A matrix Rademacher series is a random matrix that can be written as a sum of fixed matrices, each weighted by an independent Rademacher random variable.[3] This construction includes things like random sign matrices, as well as a fixed matrix whose entries are modulated by random signs. There are also interesting examples that arise in combinatorial optimization. We treat these problems in Chapter 4.

Matrix Chernoff Bounds. The matrix Chernoff bounds apply to a random matrix that can be decomposed as a sum of independent, random positive-semidefinite matrices whose maximum eigenvalues are subject to a uniform bound. These results allow us to obtain information about the norm of a random submatrix drawn from a fixed matrix. They are also appropriate for studying the Laplacian matrix of a random graph. See Chapter 5.

Matrix Bernstein Bounds. The matrix Bernstein inequality concerns a random matrix that can be expressed as a sum of independent, centered random matrices that admit a uniform spectral-norm bound. This result has many applications, including the analysis of randomized algorithms for matrix sparsification and matrix multiplication. It can also be used to study the random features paradigm for approximating a kernel matrix. Chapter 6 contains this material.

Intrinsic Dimension Bounds. Some matrix concentration inequalities can be improved when the random matrix has limited spectral content in most dimensions. In this situation, we may be able to obtain bounds that do not depend on the ambient dimension. See Chapter 7 for details.

We have chosen to present these results because they are illustrative, and they have already found concrete applications.

1.7.2 What's Not Here...

The program of extending scalar concentration results to the matrix setting has been quite fruitful, and there are many useful results beyond the ones that we detail. Let us mention some of the other tools that are available. For further information, see the annotated bibliography.

First, there are additional exponential concentration inequalities for a sum of independent random matrices. All of the following results can be established within the framework of this monograph.

- **Matrix Hoeffding.** This result concerns a sum of independent random matrices whose squares are subject to semidefinite upper bounds [Tro11c, §7].
- **Matrix Bennett.** This estimate sharpens the tail bound from the matrix Bernstein inequality [Tro11c, §6].
- **Matrix Bernstein, Unbounded Case.** The matrix Bernstein inequality extends to the case where the moments of the summands grow at a controlled rate. See [Tro11c, §6] or [Kol11].
- **Matrix Bernstein, Nonnegative Summands.** The lower tail of the Bernstein inequality can be improved when the summands are positive semidefinite [Mau03]; this result extends to the matrix setting. By a different argument, the dimensional factor can be removed from this bound for a class of interesting examples [Oli13, Thm. 3.1].

[3] A *Rademacher random variable* takes the two values ± 1 with equal probability.

The approach in this monograph can be adapted to obtain exponential concentration for matrix-valued martingales. Here are a few results from this category:

- **Matrix Azuma.** This is the martingale version of the matrix Hoeffding bound [Tro11c, §7].
- **Matrix Bounded Differences**. The matrix Azuma inequality gives bounds for the spectral norm of a matrix-valued function of independent random variables [Tro11c, §7].
- **Matrix Freedman.** This result can be viewed as the martingale extension of the matrix Bernstein inequality [Oli10a, Tro11a].

The technical report [Tro11b] explains how to extend other bounds for a sum of independent random matrices to the martingale setting.

Polynomial moment inequalities provide bounds for the expected trace of a power of a random matrix. Moment inequalities for a sum of independent random matrices can provide useful information when the summands have heavy tails or else a uniform bound does not reflect the typical size of the summands.

- **Matrix Khintchine.** The matrix Khintchine inequality is the polynomial version of the exponential bounds for matrix Gaussian series and matrix Rademacher series. This result is presented in (4.7.1). See the papers [LP86, Buc01, Buc05] or [MJC+14, Cor. 7.3] for proofs.
- **Matrix Moment Inequalities.** The matrix Chernoff inequality admits a polynomial variant; the simplest form appears in (5.1.9). The matrix Bernstein inequality also has a polynomial variant, stated in (6.1.6). These bounds are drawn from [CGT12a, App.].

The methods that lead to polynomial moment inequalities differ substantially from the techniques in this monograph, so we cannot include the proofs here. The annotated bibliography includes references to the large literature on moment inequalities for random matrices.

Recently, Lester Mackey and the author, in collaboration with Daniel Paulin and several other researchers [MJC+14, PMT14], have developed another framework for establishing matrix concentration. This approach extends a scalar argument, introduced by Chatterjee [Cha05, Cha07], that depends on exchangeable pairs and Markov chain couplings. The *method of exchangeable pairs* delivers both exponential concentration inequalities and polynomial moment inequalities for random matrices, and it can reproduce many of the bounds mentioned above. It also leads to new results:

- **Polynomial Efron–Stein Inequality for Matrices.** This bound is a matrix version of the polynomial Efron–Stein inequality [BBLM05, Thm. 1]. It controls the polynomial moments of a centered random matrix that is a function of independent random variables [PMT14, Thm. 4.2].
- **Exponential Efron–Stein Inequality for Matrices.** This bound is the matrix extension of the exponential Efron–Stein inequality [BLM03, Thm. 1]. It leads to exponential concentration inequalities for a centered random matrix constructed from independent random variables [PMT14, Thm. 4.3].

Another significant advantage is that the method of exchangeable pairs can sometimes handle random matrices built from dependent random variables. Although the simplest version of the exchangeable pairs argument is more elementary than the approach in this monograph, it takes a lot of effort to establish the more useful inequalities. With some regret, we have chosen not to include this material because the method and results are accessible to a narrower audience.

Finally, we remark that the modified logarithmic Sobolev inequalities of [BLM03, BBLM05] also extend to the matrix setting [CT14]. Unfortunately, the matrix variants do not seem to be as useful as the scalar results.

1.8 About This Monograph

This monograph is intended for graduate students and researchers in computational mathematics who want to learn some modern techniques for analyzing random matrices. The preparation required is minimal. We assume familiarity with calculus, applied linear algebra, the basic theory of normed spaces, and classical probability theory up through the elementary concentration inequalities (such as Markov and Bernstein). Beyond the basics, which can be gleaned from any good textbook, we include all the required background in Chapter 2.

The material here is based primarily on the paper "User-Friendly Tail Bounds for Sums of Random Matrices" by the present author [Tro11c]. There are several significant revisions to this earlier work:

Examples and Applications. Many of the papers on matrix concentration give limited information about how the results can be used to solve problems of interest. A major part of these notes consists of worked examples and applications that indicate how matrix concentration inequalities apply to practical questions.

Expectation Bounds. This work collects bounds for the expected value of the spectral norm of a random matrix and bounds for the expectation of the smallest and largest eigenvalues of a random symmetric matrix. Some of these useful results have appeared piecemeal in the literature [CGT12a, MJC+14], but they have not been included in a unified presentation.

Optimality. We explain why each matrix concentration inequality is (nearly) optimal. This presentation includes examples to show that each term in each bound is necessary to describe some particular phenomenon.

Intrinsic Dimension Bounds. Over the last few years, there have been some refinements to the basic matrix concentration bounds that improve the dependence on dimension [HKZ12, Min11]. We describe a new framework that allows us to prove these results with ease.

Lieb's Theorem. The matrix concentration inequalities in this monograph depend on a deep theorem [Lie73, Thm. 6] from matrix analysis due to Elliott Lieb. We provide a complete proof of this result, along with all the background required to understand the argument.

Annotated Bibliography. We have included a list of the major works on matrix concentration, including a short summary of the main contributions of these papers. We hope this catalog will be a valuable guide for further reading.

The organization of the notes is straightforward. Chapter 2 contains background material that is needed for the analysis. Chapter 3 describes the framework for developing exponential

concentration inequalities for matrices. Chapter 4 presents the first set of results and examples, concerning matrix Gaussian and Rademacher series. Chapter 5 introduces the matrix Chernoff bounds and their applications, and Chapter 6 expands on our discussion of the matrix Bernstein inequality. Chapter 7 shows how to sharpen some of the results so that they depend on an intrinsic dimension parameter. Chapter 8 contains the proof of Lieb's theorem. We conclude with resources on matrix concentration and a bibliography.

To make the presentation smoother, we have not followed all of the conventions for scholarly articles in journals. In particular, almost all the citations appear in the notes at the end of each chapter. Our aim has been to explain the ideas as clearly as possible, rather than to interrupt the narrative with an elaborate genealogy of results.

CHAPTER 2

Matrix Functions & Probability with Matrices

We begin the main development with a short overview of the background material that is required to understand the proofs and, to a lesser extent, the statements of matrix concentration inequalities. We have been careful to provide cross-references to these foundational results, so most readers will be able to proceed directly to the main theoretical development in Chapter 3 or the discussion of specific random matrix inequalities in Chapters 4, 5, and 6.

Overview

Section 2.1 covers material from matrix theory concerning the behavior of matrix functions. Section 2.2 reviews relevant results from probability, especially the parts involving matrices.

2.1 Matrix Theory Background

Let us begin with the results we require from the field of matrix analysis.

2.1.1 Conventions

We write $\mathbb{R}$ and $\mathbb{C}$ for the real and complex fields. A *matrix* is a finite, two-dimensional array of complex numbers. Many parts of the discussion do not depend on the size of a matrix, so we specify dimensions only when it really matters. Readers who wish to think about real-valued matrices will find that none of the results require any essential modification in this setting.

2.1.2 Spaces of Vectors

The symbol $\mathbb{C}^d$ denotes the complex linear space consisting of d-dimensional column vectors with complex entries, equipped with the usual componentwise addition and multiplication by a

complex scalar. We endow this space with the standard ℓ_2 inner product

$$\langle \boldsymbol{x},\ \boldsymbol{y}\rangle = \boldsymbol{x}^*\boldsymbol{y} = \sum_{i=1}^{d} x_i^* y_i \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^d.$$

The symbol * denotes the complex conjugate of a number, as well as the conjugate transpose of a vector or matrix. The inner product induces the ℓ_2 norm:

$$\|\boldsymbol{x}\|^2 = \langle \boldsymbol{x},\ \boldsymbol{x}\rangle = \sum_{i=1}^{d} |x_i|^2 \quad \text{for all } \boldsymbol{x} \in \mathbb{C}^d. \tag{2.1.1}$$

Similarly, the real linear space $\mathbb{R}^d$ consists of d-dimensional column vectors with real entries, equipped with the usual componentwise addition and multiplication by a real scalar. The inner product and ℓ_2 norm on $\mathbb{R}^d$ are defined by the same relations as for $\mathbb{C}^d$.

2.1.3 Spaces of Matrices

We write $\mathbb{M}^{d_1\times d_2}$ for the complex linear space consisting of $d_1 \times d_2$ matrices with complex entries, equipped with the usual componentwise addition and multiplication by a complex scalar. It is convenient to identify $\mathbb{C}^d$ with the space $\mathbb{M}^{d\times 1}$. We write $\mathbb{M}_d$ for the algebra of $d \times d$ square, complex matrices. The term "algebra" just means that we can multiply two matrices in $\mathbb{M}_d$ to obtain another matrix in $\mathbb{M}_d$.

2.1.4 Topology & Convergence

We can endow the space of matrices with the Frobenius norm:

$$\|\boldsymbol{B}\|_{\mathrm{F}}^2 = \sum_{j=1}^{d_1}\sum_{k=1}^{d_2} |b_{jk}|^2 \quad \text{for } \boldsymbol{B} \in \mathbb{M}^{d_1\times d_2}. \tag{2.1.2}$$

Observe that the Frobenius norm on $\mathbb{M}^{d\times 1}$ coincides with the ℓ_2 norm (2.1.1) on $\mathbb{C}^d$.

The Frobenius norm induces a norm topology on the space of matrices. In particular, given a sequence $\{\boldsymbol{B}_n : n = 1, 2, 3, \dots\} \subset \mathbb{M}^{d_1\times d_2}$, the symbol

$$\boldsymbol{B}_n \to \boldsymbol{B} \quad \text{means that} \quad \|\boldsymbol{B}_n - \boldsymbol{B}\|_{\mathrm{F}} \to 0 \quad \text{as } n \to \infty.$$

Open and closed sets are also defined with respect to the Frobenius-norm topology. Every other norm topology on $\mathbb{M}^{d_1\times d_2}$ induces the same notions of convergence and open sets. We use the same topology for the normed linear spaces $\mathbb{C}^d$ and $\mathbb{M}_d$.

2.1.5 Basic Vectors and Matrices

We write $\mathbf{0}$ for the zero vector or the zero matrix, while $\mathbf{I}$ denotes the identity matrix. Occasionally, we add a subscript to specify the dimension. For instance, $\mathbf{I}_d$ is the $d \times d$ identity.

The standard basis for the linear space $\mathbb{C}^d$ consists of *standard basis vectors*. The standard basis vector $\mathbf{e}_k$ is a column vector with a one in position k and zeros elsewhere. We also write $\mathbf{e}$ for the column vector whose entries all equal one. There is a related notation for the standard basis of $\mathbb{M}^{d_1\times d_2}$. We write $\mathbf{E}_{jk}$ for the standard basis matrix with a one in position (j, k) and zeros

elsewhere. The dimension of a standard basis vector and a standard basis matrix is typically determined by the context.

A square matrix $\boldsymbol{Q}$ that satisfies $\boldsymbol{Q}\boldsymbol{Q}^* = \mathbf{I} = \boldsymbol{Q}^*\boldsymbol{Q}$ is called a *unitary matrix*. We reserve the letter $\boldsymbol{Q}$ for a unitary matrix. Readers who prefer the real setting may prefer to regard $\boldsymbol{Q}$ as an orthogonal matrix.

2.1.6 Hermitian Matrices and Eigenvalues

An *Hermitian matrix* $\boldsymbol{A}$ is a square matrix that satisfies $\boldsymbol{A} = \boldsymbol{A}^*$. A useful intuition from operator theory is that Hermitian matrices are analogous with real numbers, while general square matrices are analogous with complex numbers.

We write $\mathbb{H}_d$ for the collection of $d \times d$ Hermitian matrices. The set $\mathbb{H}_d$ is a linear space over the real field. That is, we can add Hermitian matrices and multiply them by real numbers. The space $\mathbb{H}_d$ inherits the Frobenius-norm topology from $\mathbb{M}_d$. We adopt Parlett's convention [Par98] that bold Latin and Greek letters that are symmetric around the vertical axis ($\boldsymbol{A}$, $\boldsymbol{H}$, ..., $\boldsymbol{Y}$; $\boldsymbol{\Delta}$, $\boldsymbol{\Theta}$, ..., $\boldsymbol{\Omega}$) always represent Hermitian matrices.

Each Hermitian matrix $\boldsymbol{A} \in \mathbb{H}_d$ has an *eigenvalue decomposition*

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^* \quad \text{where } \boldsymbol{Q} \in \mathbb{M}_d \text{ is unitary and } \boldsymbol{\Lambda} \in \mathbb{H}_d \text{ is diagonal.} \tag{2.1.3}$$

The diagonal entries of $\boldsymbol{\Lambda}$ are real numbers, which are referred to as the *eigenvalues* of $\boldsymbol{A}$. The unitary matrix $\boldsymbol{Q}$ in the eigenvalue decomposition is not determined completely, but the list of eigenvalues is unique modulo permutations. The eigenvalues of an Hermitian matrix are often referred to as its *spectrum*.

We denote the algebraic minimum and maximum eigenvalues of an Hermitian matrix $\boldsymbol{A}$ by $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$. The extreme eigenvalue maps are positive homogeneous:

$$\lambda_{\min}(\alpha\boldsymbol{A}) = \alpha\lambda_{\min}(\boldsymbol{A}) \quad \text{and} \quad \lambda_{\max}(\alpha\boldsymbol{A}) = \alpha\lambda_{\max}(\boldsymbol{A}) \quad \text{for } \alpha \geq 0. \tag{2.1.4}$$

There is an important relationship between minimum and maximum eigenvalues:

$$\lambda_{\min}(-\boldsymbol{A}) = -\lambda_{\max}(\boldsymbol{A}). \tag{2.1.5}$$

The fact (2.1.5) warns us that we must be careful passing scalars through an eigenvalue map.

This work rarely requires any eigenvalues of an Hermitian matrix aside from the minimum and maximum. When they do arise, we usually order the other eigenvalues in the weakly decreasing sense:

$$\lambda_1(\boldsymbol{A}) \geq \lambda_2(\boldsymbol{A}) \geq \cdots \geq \lambda_d(\boldsymbol{A}) \quad \text{for } \boldsymbol{A} \in \mathbb{H}_d.$$

On occasion, it is more natural to arrange eigenvalues in the weakly increasing sense:

$$\lambda_1^{\uparrow}(\boldsymbol{A}) \leq \lambda_2^{\uparrow}(\boldsymbol{A}) \leq \cdots \leq \lambda_d^{\uparrow}(\boldsymbol{A}) \quad \text{for } \boldsymbol{A} \in \mathbb{H}_d.$$

To prevent confusion, we will accompany this notation with a reminder.

Readers who prefer the real setting may read "symmetric" in place of "Hermitian." In this case, the eigenvalue decomposition involves an orthogonal matrix $\boldsymbol{Q}$. Note, however, that the term "symmetric" has a different meaning when applied to random variables!

2.1.7 The Trace of a Square Matrix

The *trace* of a square matrix, denoted by tr, is the sum of its diagonal entries.

$$\operatorname{tr} \boldsymbol{B} = \sum_{j=1}^{d} b_{jj} \quad \text{for } \boldsymbol{B} \in \mathbb{M}_d. \tag{2.1.6}$$

The trace is unitarily invariant:

$$\operatorname{tr} \boldsymbol{B} = \operatorname{tr}(\boldsymbol{Q}\boldsymbol{B}\boldsymbol{Q}^*) \quad \text{for each } \boldsymbol{B} \in \mathbb{M}_d \text{ and each unitary } \boldsymbol{Q} \in \mathbb{M}_d. \tag{2.1.7}$$

In particular, the existence of an eigenvalue decomposition (2.1.3) shows that the trace of an Hermitian matrix equals the sum of its eigenvalues.[1]

Another valuable relation connects the trace with the Frobenius norm:

$$\|\boldsymbol{C}\|_{\mathrm{F}}^2 = \operatorname{tr}(\boldsymbol{C}\boldsymbol{C}^*) = \operatorname{tr}(\boldsymbol{C}^*\boldsymbol{C}). \quad \text{for all } \boldsymbol{C} \in \mathbb{M}^{d_1 \times d_2}. \tag{2.1.8}$$

This expression follows from the definitions (2.1.2) and (2.1.6) and a short calculation.

2.1.8 The Semidefinite Partial Order

A matrix $\boldsymbol{A} \in \mathbb{H}_d$ is *positive semidefinite* when it satisfies

$$\boldsymbol{u}^* \boldsymbol{A} \boldsymbol{u} \geq 0 \quad \text{for each vector } \boldsymbol{u} \in \mathbb{C}^d. \tag{2.1.9}$$

Equivalently, a matrix $\boldsymbol{A}$ is positive semidefinite when it is Hermitian and its eigenvalues are all nonnegative. Similarly, we say that $\boldsymbol{A} \in \mathbb{H}_d$ is *positive definite* when

$$\boldsymbol{u}^* \boldsymbol{A} \boldsymbol{u} > 0 \quad \text{for each nonzero vector } \boldsymbol{u} \in \mathbb{C}^d. \tag{2.1.10}$$

Equivalently, $\boldsymbol{A}$ is positive definite when it is Hermitian and its eigenvalues are all positive.

Positive-semidefinite and positive-definite matrices play a special role in matrix theory, analogous with the role of nonnegative and positive numbers in real analysis. In particular, observe that the square of an Hermitian matrix is always positive semidefinite. The square of a nonsingular Hermitian matrix is always positive definite.

The family of positive-semidefinite matrices in $\mathbb{H}_d$ forms a closed convex cone.[2] This geometric fact follows easily from the definition (2.1.9). Indeed, for each vector $\boldsymbol{u} \in \mathbb{C}^d$, the condition

$$\{\boldsymbol{A} \in \mathbb{H}_d : \boldsymbol{u}^* \boldsymbol{A} \boldsymbol{u} \geq 0\}$$

describes a closed halfspace in $\mathbb{H}_d$. As a consequence, the family of positive-semidefinite matrices in $\mathbb{H}_d$ is an intersection of closed halfspaces. Therefore, it is a closed convex set. To see why this convex set is a cone, just note that

$$\boldsymbol{A} \text{ positive semidefinite} \quad \text{implies} \quad \alpha \boldsymbol{A} \text{ is positive semidefinite for } \alpha \geq 0.$$

[1] This fact also holds true for a general square matrix.

[2] A *convex cone* is a subset C of a linear space that is closed under conic combinations. That is, $\tau_1 \boldsymbol{x}_1 + \tau_2 \boldsymbol{x}_2 \in C$ for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in C$ and all $\tau_1, \tau_2 > 0$. Equivalently, C is a set that is both convex and positively homogeneous.

Beginning from (2.1.10), similar considerations show that the family of positive-definite matrices in $\mathbb{H}_d$ forms an (open) convex cone.

We may now define the *semidefinite partial order* $\preccurlyeq$ on the real-linear space $\mathbb{H}_d$ using the rule

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \text{if and only if} \quad \boldsymbol{H} - \boldsymbol{A} \text{ is positive semidefinite.} \tag{2.1.11}$$

In particular, we write $\boldsymbol{A} \succcurlyeq \mathbf{0}$ to indicate that $\boldsymbol{A}$ is positive semidefinite and $\boldsymbol{A} \succ \mathbf{0}$ to indicate that $\boldsymbol{A}$ is positive definite. For a diagonal matrix $\boldsymbol{\Lambda}$, the expression $\boldsymbol{\Lambda} \succcurlyeq \mathbf{0}$ means that each entry of $\boldsymbol{\Lambda}$ is nonnegative.

The semidefinite order is preserved by conjugation, a simple fact whose importance cannot be overstated.

Proposition 2.1.1 (Conjugation Rule)**.** *Let $\boldsymbol{A}$ and $\boldsymbol{H}$ be Hermitian matrices of the same dimension, and let $\boldsymbol{B}$ be a general matrix with compatible dimensions. Then*

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \textit{implies} \quad \boldsymbol{B}\boldsymbol{A}\boldsymbol{B}^* \preccurlyeq \boldsymbol{B}\boldsymbol{H}\boldsymbol{B}^*. \tag{2.1.12}$$

Finally, we remark that the trace of a positive-semidefinite matrix is at least as large as its maximum eigenvalue:

$$\lambda_{\max}(\boldsymbol{A}) \le \operatorname{tr} \boldsymbol{A} \quad \text{when } \boldsymbol{A} \text{ is positive semidefinite.} \tag{2.1.13}$$

This property follows from the definition of a positive-semidefinite matrix and the fact that the trace of $\boldsymbol{A}$ equals the sum of the eigenvalues.

2.1.9 Standard Matrix Functions

Let us describe the most direct method for extending a function on the real numbers to a function on Hermitian matrices. The basic idea is to apply the function to each eigenvalue of the matrix to construct a new matrix.

Definition 2.1.2 (Standard Matrix Function)**.** *Let $f : I \to \mathbb{R}$ where I is an interval of the real line. Consider a matrix $\boldsymbol{A} \in \mathbb{H}_d$ whose eigenvalues are contained in I. Define the matrix $f(\boldsymbol{A}) \in \mathbb{H}_d$ using an eigenvalue decomposition of $\boldsymbol{A}$:*

$$f(\boldsymbol{A}) = \boldsymbol{Q} \begin{bmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_d) \end{bmatrix} \boldsymbol{Q}^* \quad \textit{where} \quad \boldsymbol{A} = \boldsymbol{Q} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} \boldsymbol{Q}^*.$$

In particular, we can apply f to a real diagonal matrix by applying the function to each diagonal entry.

It can be verified that the definition of $f(\boldsymbol{A})$ does not depend on which eigenvalue decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^*$ that we choose. Any matrix function that arises in this fashion is called a standard matrix function.

To confirm that this definition is sensible, consider the power function $f(t) = t^q$ for a natural number q. When $\boldsymbol{A}$ is Hermitian, the power function $f(\boldsymbol{A}) = \boldsymbol{A}^q$, where $\boldsymbol{A}^q$ is the q-fold product of $\boldsymbol{A}$.

For an Hermitian matrix $\boldsymbol{A}$, whenever we write the power function $\boldsymbol{A}^q$ or the exponential $\mathrm{e}^{\boldsymbol{A}}$ or the logarithm $\log \boldsymbol{A}$, we are always referring to a standard matrix function. Note that we only

define the matrix logarithm for positive-definite matrices, and non-integer powers are only valid for positive-semidefinite matrices.

The following result is an immediate, but important, consequence of the definition of a standard matrix function.

Proposition 2.1.3 (Spectral Mapping Theorem)**.** *Let $f : I \to \mathbb{R}$ be a function on an interval I of the real line, and let $\boldsymbol{A}$ be an Hermitian matrix whose eigenvalues are contained in I. If λ is an eigenvalue of $\boldsymbol{A}$, then $f(\lambda)$ is an eigenvalue of $f(\boldsymbol{A})$.*

When a real function has a power series expansion, we can also represent the standard matrix function with the same power series expansion. Indeed, suppose that $f : I \to \mathbb{R}$ is defined on an interval I of the real line, and assume that the eigenvalues of $\boldsymbol{A}$ are contained in I. Then

$$f(a) = c_0 + \sum_{q=1}^{\infty} c_q a^q \quad \text{for } a \in I \quad \text{implies} \quad f(\boldsymbol{A}) = c_0 \mathbf{I} + \sum_{q=1}^{\infty} c_q \boldsymbol{A}^q.$$

This formula can be verified using an eigenvalue decomposition of $\boldsymbol{A}$ and the definition of a standard matrix function.

2.1.10 The Transfer Rule

In most cases, the "obvious" generalization of an inequality for real-valued functions fails to hold in the semidefinite order. Nevertheless, there is one class of inequalities for real functions that extends to give semidefinite relationships for standard matrix functions.

Proposition 2.1.4 (Transfer Rule)**.** *Let f and g be real-valued functions defined on an interval I of the real line, and let $\boldsymbol{A}$ be an Hermitian matrix whose eigenvalues are contained in I. Then*

$$f(a) \le g(a) \quad \textit{for each } a \in I \quad \textit{implies} \quad f(\boldsymbol{A}) \preccurlyeq g(\boldsymbol{A}). \tag{2.1.14}$$

Proof. Decompose $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^*$. It is immediate that $f(\boldsymbol{\Lambda}) \preccurlyeq g(\boldsymbol{\Lambda})$. The Conjugation Rule (2.1.12) allows us to conjugate this relation by $\boldsymbol{Q}$. Finally, we invoke Definition 2.1.2, of a standard matrix function, to complete the argument. □

2.1.11 The Matrix Exponential

For any Hermitian matrix $\boldsymbol{A}$, we can introduce the matrix exponential $\mathrm{e}^{\boldsymbol{A}}$ using Definition 2.1.2. Equivalently, we can use a power series expansion:

$$\mathrm{e}^{\boldsymbol{A}} = \exp(\boldsymbol{A}) = \mathbf{I} + \sum_{q=1}^{\infty} \frac{\boldsymbol{A}^q}{q!}. \tag{2.1.15}$$

The Spectral Mapping Theorem, Proposition 2.1.3, implies that the exponential of an Hermitian matrix is always positive definite.

We often work with the trace of the matrix exponential:

$$\operatorname{tr}\exp : \boldsymbol{A} \longmapsto \operatorname{tr} \mathrm{e}^{\boldsymbol{A}}.$$

This function has a monotonicity property that we use extensively. For Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{H}$ with the same dimension,

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \text{implies} \quad \operatorname{tr} \mathrm{e}^{\boldsymbol{A}} \le \operatorname{tr} \mathrm{e}^{\boldsymbol{H}}. \tag{2.1.16}$$

We establish this result in §8.3.2.

2.1.12 The Matrix Logarithm

We can define the matrix logarithm as a standard matrix function. The matrix logarithm is also the functional inverse of the matrix exponential:

$$\log\bigl(\mathrm{e}^{\boldsymbol{A}}\bigr) = \boldsymbol{A} \quad \text{for each Hermitian matrix } \boldsymbol{A}. \tag{2.1.17}$$

A valuable fact about the matrix logarithm is that it preserves the semidefinite order. For positive-definite matrices $\boldsymbol{A}$ and $\boldsymbol{H}$ with the same dimension,

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \text{implies} \quad \log \boldsymbol{A} \preccurlyeq \log \boldsymbol{H}. \tag{2.1.18}$$

We establish this result in §8.4.4. Let us stress that the matrix exponential *does not* have any operator monotonicity property analogous with (2.1.18)!

2.1.13 Singular Values of Rectangular Matrices

A general matrix does not have an eigenvalue decomposition, but it admits a different representation that is just as useful. Every $d_1 \times d_2$ matrix $\boldsymbol{B}$ has a *singular value decomposition*

$$\boldsymbol{B} = \boldsymbol{Q}_1 \boldsymbol{\Sigma} \boldsymbol{Q}_2^* \quad \text{where } \boldsymbol{Q}_1 \text{ and } \boldsymbol{Q}_2 \text{ are unitary and } \boldsymbol{\Sigma} \text{ is nonnegative diagonal.} \tag{2.1.19}$$

The unitary matrices $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ have dimensions $d_1 \times d_1$ and $d_2 \times d_2$, respectively. The inner matrix $\boldsymbol{\Sigma}$ has dimension $d_1 \times d_2$, and we use the term diagonal in the sense that only the diagonal entries $(\boldsymbol{\Sigma})_{jj}$ may be nonzero.

The diagonal entries of $\boldsymbol{\Sigma}$ are called the *singular values* of $\boldsymbol{B}$, and they are denoted as $\sigma_j(\boldsymbol{B})$. The singular values are determined completely modulo permutations, and it is conventional to arrange them in weakly decreasing order:

$$\sigma_1(\boldsymbol{B}) \geq \sigma_2(\boldsymbol{B}) \geq \cdots \geq \sigma_{\min\{d_1,\, d_2\}}(\boldsymbol{B}).$$

There is an important relationship between singular values and eigenvalues. A general matrix has two squares associated with it, $\boldsymbol{B}\boldsymbol{B}^*$ and $\boldsymbol{B}^*\boldsymbol{B}$, both of which are positive semidefinite. We can use a singular value decomposition of $\boldsymbol{B}$ to construct eigenvalue decompositions of the two squares:

$$\boldsymbol{B}\boldsymbol{B}^* = \boldsymbol{Q}_1(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*)\boldsymbol{Q}_1^* \quad \text{and} \quad \boldsymbol{B}^*\boldsymbol{B} = \boldsymbol{Q}_2(\boldsymbol{\Sigma}^*\boldsymbol{\Sigma})\boldsymbol{Q}_2^* \tag{2.1.20}$$

The two squares of $\boldsymbol{\Sigma}$ are square, diagonal matrices with nonnegative entries. Conversely, we can always extract a singular value decomposition from the eigenvalue decompositions of the two squares.

We can write the Frobenius norm of a matrix in terms of the singular values:

$$\|\boldsymbol{B}\|_{\mathrm{F}}^2 = \sum_{j=1}^{\min\{d_1,d_2\}} \sigma_j(\boldsymbol{B})^2 \quad \text{for } \boldsymbol{B} \in \mathbb{M}^{d_1 \times d_2}. \tag{2.1.21}$$

This expression follows from the expression (2.1.8) for the Frobenius norm, the property (2.1.20) of the singular value decomposition, and the unitary invariance (2.1.7) of the trace.

2.1.14 The Spectral Norm

The *spectral norm* of an Hermitian matrix $\boldsymbol{A}$ is defined by the relation

$$\|\boldsymbol{A}\| = \max\{\lambda_{\max}(\boldsymbol{A}),\ -\lambda_{\min}(\boldsymbol{A})\}. \tag{2.1.22}$$

For a general matrix $\boldsymbol{B}$, the spectral norm is defined to be the largest singular value:

$$\|\boldsymbol{B}\| = \sigma_1(\boldsymbol{B}). \tag{2.1.23}$$

These two definitions are consistent for Hermitian matrices because of (2.1.20). When applied to a row vector or a column vector, the spectral norm coincides with the ℓ_2 norm (2.1.1).

We will often need the fact that

$$\|\boldsymbol{B}\|^2 = \|\boldsymbol{B}\boldsymbol{B}^*\| = \|\boldsymbol{B}^*\boldsymbol{B}\|. \tag{2.1.24}$$

This identity also follows from (2.1.20).

2.1.15 The Stable Rank

In several of the applications, we need an analytic measure of the collinearity of the rows and columns of a matrix called the *stable rank*. For a general matrix $\boldsymbol{B}$, the stable rank is defined as

$$\operatorname{srank}(\boldsymbol{B}) = \frac{\|\boldsymbol{B}\|_{\mathrm{F}}^2}{\|\boldsymbol{B}\|^2}. \tag{2.1.25}$$

The stable rank is a lower bound for the algebraic rank:

$$1 \le \operatorname{srank}(\boldsymbol{B}) \le \operatorname{rank}(\boldsymbol{B}).$$

This point follows when we use (2.1.21) and (2.1.23) to express the two norms in terms of the singular values of $\boldsymbol{B}$. In contrast to the algebraic rank, the stable rank is a continuous function of the matrix, so it is more suitable for numerical applications.

2.1.16 Dilations

An extraordinarily fruitful idea from operator theory is to embed matrices within larger block matrices, called *dilations*. Dilations have an almost magical power. In this work, we will use dilations to extend matrix concentration inequalities from Hermitian matrices to general matrices.

Definition 2.1.5 (Hermitian Dilation)**.** *The* Hermitian dilation

$$\mathscr{H} : \mathbb{M}^{d_1 \times d_2} \longrightarrow \mathbb{H}_{d_1+d_2}$$

is the map from a general matrix to an Hermitian matrix defined by

$$\mathscr{H}(\boldsymbol{B}) = \begin{bmatrix} \mathbf{0} & \boldsymbol{B} \\ \boldsymbol{B}^* & \mathbf{0} \end{bmatrix}. \tag{2.1.26}$$

It is clear that the Hermitian dilation is a real-linear map. Furthermore, the dilation retains important spectral information. To see why, note that the square of the dilation satisfies

$$\mathscr{H}(\boldsymbol{B})^2 = \begin{bmatrix} \boldsymbol{B}\boldsymbol{B}^* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B}^*\boldsymbol{B} \end{bmatrix}. \tag{2.1.27}$$

We discover that the squared eigenvalues of $\mathscr{H}(\boldsymbol{B})$ coincide with the squared singular values of $\boldsymbol{B}$, along with an appropriate number of zeros. As a consequence, $\|\mathscr{H}(\boldsymbol{B})\| = \|\boldsymbol{B}\|$. Moreover,

$$\lambda_{\max}(\mathscr{H}(\boldsymbol{B})) = \|\mathscr{H}(\boldsymbol{B})\| = \|\boldsymbol{B}\|. \tag{2.1.28}$$

We will invoke the identity (2.1.28) repeatedly.

One way to justify the first relation in (2.1.28) is to introduce the first columns $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ of the unitary matrices $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ that appear in the singular value decomposition $\boldsymbol{B} = \boldsymbol{Q}_1\boldsymbol{\Sigma}\boldsymbol{Q}_2^*$. Then we may calculate that

$$\|\boldsymbol{B}\| = \operatorname{Re}(\boldsymbol{u}_1^*\boldsymbol{B}\boldsymbol{u}_2) = \frac{1}{2}\begin{bmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\end{bmatrix}^*\begin{bmatrix}\mathbf{0} & \boldsymbol{B}\\ \boldsymbol{B}^* & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\end{bmatrix} \le \lambda_{\max}(\mathscr{H}(\boldsymbol{B})) \le \|\mathscr{H}(\boldsymbol{B})\| = \|\boldsymbol{B}\|.$$

Indeed, the spectral norm of $\boldsymbol{B}$ equals its largest singular value $\sigma_1(\boldsymbol{B})$, which coincides with $\boldsymbol{u}_1^*\boldsymbol{B}\boldsymbol{u}_2$ by construction of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. The second identity relies on a direct calculation. The first inequality follows from the variational representation of the maximum eigenvalue as a Rayleigh quotient; this fact can also be derived as a consequence of (2.1.3). The second inequality depends on the definition (2.1.22) of the spectral norm of an Hermitian matrix.

2.1.17 Other Matrix Norms

There are a number of other matrix norms that arise sporadically in this work. The *Schatten 1-norm* of a matrix can be defined as the sum of its singular values:

$$\|\boldsymbol{B}\|_{S_1} = \sum_{j=1}^{\min\{d_1,d_2\}} \sigma_j(\boldsymbol{B}) \quad \text{for } \boldsymbol{B} \in \mathbb{M}^{d_1\times d_2}. \tag{2.1.29}$$

The entrywise ℓ_1 norm of a matrix is defined as

$$\|\boldsymbol{B}\|_{\ell_1} = \sum_{j=1}^{d_1}\sum_{k=1}^{d_2} |b_{jk}| \quad \text{for } \boldsymbol{B} \in \mathbb{M}^{d_1\times d_2}. \tag{2.1.30}$$

We always have the relation

$$\|\boldsymbol{B}\|_{\ell_1} \le \sqrt{d_1 d_2}\,\|\boldsymbol{B}\|_{\mathrm{F}} \quad \text{for } \boldsymbol{B} \in \mathbb{M}^{d_1\times d_2} \tag{2.1.31}$$

because of the Cauchy–Schwarz inequality.

2.2 Probability with Matrices

We continue with some material from probability, focusing on connections with matrices.

2.2.1 Conventions

We prefer to avoid abstraction and unnecessary technical detail, so we frame the standing assumption that all random variables are sufficiently regular that we are justified in computing expectations, interchanging limits, and so forth. The manipulations we perform are valid if we assume that all random variables are bounded, but the results hold in broader circumstances if we instate appropriate regularity conditions.

Since the expectation operator is linear, we typically do not use parentheses with it. We instate the convention that powers and products take precedence over the expectation operator. In particular,

$$\mathbb{E}\, X^q = \mathbb{E}(X^q).$$

This position helps us reduce the clutter of parentheses. We sometimes include extra delimiters when it is helpful for clarity.

2.2.2 Some Scalar Random Variables

We use consistent notation for some of the basic scalar random variables.

Standard normal variables. We reserve the letter γ for a NORMAL$(0,1)$ random variable. That is, γ is a real Gaussian with mean zero and variance one.

Rademacher random variables. We reserve the letter ϱ for a random variable that takes the two values ± 1 with equal probability.

Bernoulli random variables. A BERNOULLI(p) random variable takes the value one with probability p and the value zero with probability $1-p$, where $p \in [0,1]$. We use the letters δ and ξ for Bernoulli random variables.

2.2.3 Random Matrices

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. A *random matrix* $\boldsymbol{Z}$ is a measurable map

$$\boldsymbol{Z} : \Omega \longrightarrow \mathbb{M}^{d_1 \times d_2}.$$

It is more natural to think of the entries of $\boldsymbol{Z}$ as complex random variables that may or may not be correlated with each other. We reserve the letters $\boldsymbol{X}$ and $\boldsymbol{Y}$ for random Hermitian matrices, while the letter $\boldsymbol{Z}$ denotes a general random matrix.

A finite sequence $\{\boldsymbol{Z}_k\}$ of random matrices is *independent* when

$$\mathbb{P}\{\boldsymbol{Z}_k \in F_k \text{ for each } k\} = \prod\nolimits_k \mathbb{P}\{\boldsymbol{Z}_k \in F_k\}$$

for every collection $\{F_k\}$ of Borel subsets of $\mathbb{M}^{d_1 \times d_2}$.

2.2.4 Expectation

The *expectation* of a random matrix $\boldsymbol{Z} = [Z_{jk}]$ is simply the matrix formed by taking the componentwise expectation. That is,

$$(\mathbb{E}\, \boldsymbol{Z})_{jk} = \mathbb{E}\, Z_{jk}.$$

Under mild assumptions, expectation commutes with linear and real-linear maps. Indeed, expectation commutes with multiplication by a fixed matrix:

$$\mathbb{E}(\boldsymbol{B}\boldsymbol{Z}) = \boldsymbol{B}\,(\mathbb{E}\,\boldsymbol{Z}) \quad \text{and} \quad \mathbb{E}(\boldsymbol{Z}\boldsymbol{B}) = (\mathbb{E}\,\boldsymbol{Z})\,\boldsymbol{B}.$$

In particular, the product rule for the expectation of independent random variables extends to matrices:

$$\mathbb{E}(\boldsymbol{S}\boldsymbol{Z}) = (\mathbb{E}\,\boldsymbol{S})(\mathbb{E}\,\boldsymbol{Z}) \quad \text{when } \boldsymbol{S} \text{ and } \boldsymbol{Z} \text{ are independent.}$$

We use these identities liberally, without any further comment.

2.2.5 Inequalities for Expectation

Markov's inequality states that a nonnegative (real) random variable X obeys the probability bound

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}\,X}{t} \quad \text{for } t > 0. \tag{2.2.1}$$

The Markov inequality is a central tool for establishing concentration inequalities.

Jensen's inequality describes how averaging interacts with convexity. Let $\boldsymbol{Z}$ be a random matrix, and let h be a real-valued function on matrices. Then

$$\begin{aligned} \mathbb{E}\,h(\boldsymbol{Z}) &\leq h(\mathbb{E}\,\boldsymbol{Z}) \quad \text{when } h \text{ is concave, and} \\ \mathbb{E}\,h(\boldsymbol{Z}) &\geq h(\mathbb{E}\,\boldsymbol{Z}) \quad \text{when } h \text{ is convex.} \end{aligned} \tag{2.2.2}$$

The family of positive-semidefinite matrices in $\mathbb{H}_d$ forms a convex cone, and the expectation of a random matrix can be viewed as a convex combination. Therefore, expectation preserves the semidefinite order:

$$\boldsymbol{X} \preccurlyeq \boldsymbol{Y} \quad \text{implies} \quad \mathbb{E}\,\boldsymbol{X} \preccurlyeq \mathbb{E}\,\boldsymbol{Y}.$$

We use this result many times without direct reference.

2.2.6 The Variance of a Random Hermitian Matrix

The variance of a real random variable Y is defined as the expected squared deviation from the mean:

$$\operatorname{Var}(Y) = \mathbb{E}\,(Y - \mathbb{E}\,Y)^2$$

There are a number of natural extensions of this concept in the matrix setting that play a role in our theory.

Suppose that $\boldsymbol{Y}$ is a random Hermitian matrix. We can define a matrix-valued variance:

$$\mathbf{Var}(\boldsymbol{Y}) = \mathbb{E}\,(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y})^2 = \mathbb{E}\,\boldsymbol{Y}^2 - (\mathbb{E}\,\boldsymbol{Y})^2. \tag{2.2.3}$$

The matrix $\mathbf{Var}(\boldsymbol{Y})$ is always positive semidefinite. We can interpret the (j,k) entry of this matrix as the covariance between the jth and kth columns of $\boldsymbol{Y}$:

$$(\mathbf{Var}(\boldsymbol{Y}))_{jk} = \mathbb{E}\left[(\boldsymbol{y}_{:j} - \mathbb{E}\,\boldsymbol{y}_{:j})^*(\boldsymbol{y}_{:k} - \mathbb{E}\,\boldsymbol{y}_{:k})\right],$$

where we have written $\boldsymbol{y}_{:j}$ for the jth column of $\boldsymbol{Y}$.

The matrix-valued variance contains a lot of information about the fluctuations of the random matrix. We can summarize $\mathbf{Var}(\boldsymbol{Y})$ using a single number $v(\boldsymbol{Y})$, which we call the *matrix variance statistic*:

$$v(\boldsymbol{Y}) = \|\mathbf{Var}(\boldsymbol{Y})\| = \|\mathbb{E}(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y})^2\|. \tag{2.2.4}$$

To understand what this quantity means, one may wish to rewrite it as

$$v(\boldsymbol{Y}) = \sup_{\|\boldsymbol{u}\|=1} \mathbb{E}\,\|(\boldsymbol{Y}\boldsymbol{u}) - \mathbb{E}(\boldsymbol{Y}\boldsymbol{u})\|^2.$$

Roughly speaking, the matrix variance statistic describes the maximum variance of $\boldsymbol{Y}\boldsymbol{u}$ for any unit vector $\boldsymbol{u}$.

2.2.7 The Variance of a Sum of Independent, Random Hermitian Matrices

The matrix-valued variance interacts beautifully with a sum of independent random matrices. Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random Hermitian matrices with common dimension d. Introduce the sum $\boldsymbol{Y} = \sum_k \boldsymbol{X}_k$. Then

$$\begin{aligned}\mathbf{Var}(\boldsymbol{Y}) = \mathbf{Var}\left(\textstyle\sum_k \boldsymbol{X}_k\right) &= \mathbb{E}\left(\textstyle\sum_k (\boldsymbol{X}_k - \mathbb{E}\,\boldsymbol{X}_k)\right)^2 \\ &= \textstyle\sum_{j,k} \mathbb{E}\left[(\boldsymbol{X}_j - \mathbb{E}\,\boldsymbol{X}_j)(\boldsymbol{X}_k - \mathbb{E}\,\boldsymbol{X}_k)\right] \\ &= \textstyle\sum_k \mathbb{E}\,(\boldsymbol{X}_k - \mathbb{E}\,\boldsymbol{X}_k)^2 \\ &= \textstyle\sum_k \mathbf{Var}(\boldsymbol{X}_k).\end{aligned} \tag{2.2.5}$$

This identity matches the familiar result for the variance of a sum of independent scalar random variables. It follows that the matrix variance statistic satisfies

$$v(\boldsymbol{Y}) = \left\|\textstyle\sum_k \mathbf{Var}(\boldsymbol{X}_k)\right\|. \tag{2.2.6}$$

The fact that the sum remains inside the norm is very important. Indeed, the best general inequalities between $v(\boldsymbol{Y})$ and the matrix variance statistics $v(\boldsymbol{X}_k)$ of the summands are

$$v(\boldsymbol{Y}) \le \textstyle\sum_k v(\boldsymbol{X}_k) \le d \cdot v(\boldsymbol{Y}).$$

These relations can be improved in some special cases. For example, when the matrices $\boldsymbol{X}_k$ are identically distributed, the left-hand inequality becomes an identity.

2.2.8 The Variance of a Rectangular Random Matrix

We will often work with non-Hermitian random matrices. In this case, we need to account for the fact that a general matrix has *two* different squares. Suppose that $\boldsymbol{Z}$ is a random matrix with dimension $d_1 \times d_2$. Define

$$\begin{aligned}\mathbf{Var}_1(\boldsymbol{Z}) &= \mathbb{E}\left[(\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z})(\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z})^*\right], \quad \text{and} \\ \mathbf{Var}_2(\boldsymbol{Z}) &= \mathbb{E}\left[(\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z})^*(\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z})\right].\end{aligned} \tag{2.2.7}$$

The matrix $\mathbf{Var}_1(\boldsymbol{Z})$ is a positive-semidefinite matrix with dimension $d_1 \times d_1$, and it describes the fluctuation of the rows of $\boldsymbol{Z}$. The matrix $\mathbf{Var}_2(\boldsymbol{Z})$ is a positive-semidefinite matrix with dimension $d_2 \times d_2$, and it reflects the fluctuation of the columns of $\boldsymbol{Z}$. For an Hermitian random matrix $\boldsymbol{Y}$,

$$\mathbf{Var}(\boldsymbol{Y}) = \mathbf{Var}_1(\boldsymbol{Y}) = \mathbf{Var}_2(\boldsymbol{Y}).$$

In other words, the two variances coincide in the Hermitian setting.

As before, it is valuable to reduce these matrix-valued variances to a single scalar parameter. We define the matrix variance statistic of a general random matrix $\boldsymbol{Z}$ as

$$v(\boldsymbol{Z}) = \max\bigl\{ \|\mathbf{Var}_1(\boldsymbol{Z})\|, \|\mathbf{Var}_2(\boldsymbol{Z})\| \bigr\}. \tag{2.2.8}$$

When $\boldsymbol{Z}$ is Hermitian, the definition (2.2.8) coincides with the original definition (2.2.4).

To promote a deeper appreciation for the formula (2.2.8), let us explain how it arises from the Hermitian dilation (2.1.26). By direct calculation,

$$\begin{aligned} \mathbf{Var}(\mathscr{H}(\boldsymbol{Z})) &= \mathbb{E}\begin{bmatrix} \mathbf{0} & (\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z}) \\ (\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z})^* & \mathbf{0} \end{bmatrix}^2 \\ &= \mathbb{E}\begin{bmatrix} (\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z})(\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z})^* & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z})^*(\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z}) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Var}_1(\boldsymbol{Z}) & \mathbf{0} \\ \mathbf{0} & \mathbf{Var}_2(\boldsymbol{Z}) \end{bmatrix}. \end{aligned} \tag{2.2.9}$$

The first identity is the definition (2.2.3) of the matrix-valued variance. The second line follows from the formula (2.1.27) for the square of the dilation. The last identity depends on the definition (2.2.7) of the two matrix-valued variances. Therefore, using the definitions (2.2.4) and (2.2.8) of the matrix variance statistics,

$$v(\mathscr{H}(\boldsymbol{Z})) = \|\mathbf{Var}(\mathscr{H}(\boldsymbol{Z})\| = \max\bigl\{ \|\mathbf{Var}_1(\boldsymbol{Z})\|, \|\mathbf{Var}_2(\boldsymbol{Z})\| \bigr\} = v(\boldsymbol{Z}). \tag{2.2.10}$$

The second identity holds because the spectral norm of a block-diagonal matrix is the maximum norm achieved by one of the diagonal blocks.

2.2.9 The Variance of a Sum of Independent Random Matrices

As in the Hermitian case, the matrix-valued variances interact nicely with an independent sum. Consider a finite sequence $\{\boldsymbol{S}_k\}$ of independent random matrices with the same dimension. Form the sum $\boldsymbol{Z} = \sum_k \boldsymbol{S}_k$. Repeating the calculation leading up to (2.2.6), we find that

$$\mathbf{Var}_1(\boldsymbol{Z}) = \sum\nolimits_k \mathbf{Var}_1(\boldsymbol{S}_k) \quad \text{and} \quad \mathbf{Var}_2(\boldsymbol{Z}) = \sum\nolimits_k \mathbf{Var}_2(\boldsymbol{S}_k).$$

In summary, the matrix variance statistic of an independent sum satisfies

$$v(\boldsymbol{Z}) = \max\Bigl\{ \Bigl\| \sum\nolimits_k \mathbf{Var}_1(\boldsymbol{S}_k) \Bigr\|, \Bigl\| \sum\nolimits_k \mathbf{Var}_2(\boldsymbol{S}_k) \Bigr\| \Bigr\}. \tag{2.2.11}$$

This formula arises time after time.

2.3 Notes

Everything in this chapter is firmly established. We have culled the results that are relevant to our discussion. Let us give some additional references for readers who would like more information.

2.3.1 Matrix Analysis

Our treatment of matrix analysis is drawn from Bhatia's excellent books on matrix analysis [Bha97, Bha07]. The two books [HJ13, HJ94] of Horn & Johnson also serve as good general references. Higham's work [Hig08] is a generous source of information about matrix functions. Other valuable resources include Carlen's lecture notes [Car10], the book of Petz [Pet11], and the book of Hiai & Petz [HP14].

2.3.2 Probability with Matrices

The classic introduction to probability is the two-volume treatise [Fel68, Fel71] of Feller. The book [GS01] of Grimmett & Stirzaker offers a good treatment of probability theory and random processes at an intermediate level. For a more theoretical presentation, consider the book [Shi96] of Shiryaev.

There are too many books on random matrix theory for us to include a comprehensive list; here is a selection that the author finds useful. Tao's book [Tao12] gives a friendly introduction to some of the major aspects of classical and modern random matrix theory. The lecture notes [Kem13] of Kemp are also extremely readable. The survey of Vershynin [Ver12] provides a good summary of techniques from asymptotic convex geometry that are relevant to random matrix theory. The works of Mardia, Kent, & Bibby [MKB79] and Muirhead [Mui82] present classical results on random matrices that are particularly useful in statistics, while Bai & Silverstein [BS10] contains a comprehensive modern treatment. Nica and Speicher [NS06] offer an entrée to the beautiful field of free probability. Mehta's treatise [Meh04] was the first book on random matrix theory available, and it remains solid.

CHAPTER 3

The Matrix Laplace Transform Method

This chapter contains the core part of the analysis that ultimately delivers matrix concentration inequalities. Readers who are only interested in the concentration inequalities themselves or the example applications may wish to move on to Chapters 4, 5, and 6.

In the scalar setting, the Laplace transform method provides a simple but powerful way to develop concentration inequalities for a sum of independent random variables. This technique is sometimes referred to as the "Bernstein trick" or "Chernoff bounding." For a primer, we recommend [BLM13, Chap. 2].

In the matrix setting, there is a very satisfactory extension of this argument that allows us to prove concentration inequalities for a sum of independent random matrices. As in the scalar case, the matrix Laplace transform method is both easy to use and incredibly useful. In contrast to the scalar case, the arguments that lead to matrix concentration are no longer elementary. The purpose of this chapter is to install the framework we need to support these results. Fortunately, in practical applications, all of the technical difficulty remains invisible.

Overview

We first define matrix analogs of the moment generating function and the cumulant generating function, which pack up information about the fluctuations of a random Hermitian matrix. Section 3.2 explains how we can use the matrix mgf to obtain probability inequalities for the maximum eigenvalue of a random Hermitian matrix. The next task is to develop a bound for the mgf of a sum of independent random matrices using information about the summands. In §3.3, we discuss the challenges that arise; §3.4 presents the ideas we need to overcome these obstacles. Section 3.5 establishes that the classical result on additivity of cumulants has a companion in the matrix setting. This result allows us to develop a collection of abstract probability inequalities in §3.6 that we can specialize to obtain matrix Chernoff bounds, matrix Bernstein bounds, and so forth.

3.1 Matrix Moments and Cumulants

At the heart of the Laplace transform method are the moment generating function (mgf) and the cumulant generating function (cgf) of a random variable. We begin by presenting matrix versions of the mgf and cgf.

Definition 3.1.1 (Matrix Mgf and Cgf)**.** *Let $\boldsymbol{X}$ be a random Hermitian matrix. The* matrix moment generating function $\boldsymbol{M}_{\boldsymbol{X}}$ *and the* matrix cumulant generating function $\boldsymbol{\Xi}_{\boldsymbol{X}}$ *are given by*

$$\boldsymbol{M}_{\boldsymbol{X}}(\theta) = \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}} \quad \text{and} \quad \boldsymbol{\Xi}_{\boldsymbol{X}}(\theta) = \log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}} \quad \text{for } \theta \in \mathbb{R}. \tag{3.1.1}$$

Note that the expectations may not exist for all values of θ.

The matrix mgf $\boldsymbol{M}_{\boldsymbol{X}}$ and matrix cgf $\boldsymbol{\Xi}_{\boldsymbol{X}}$ contain information about how much the random matrix $\boldsymbol{X}$ varies. We aim to exploit the data encoded in these functions to control the eigenvalues.

Let us take a moment to expand on Definition 3.1.1; this discussion is not important for subsequent developments. Observe that the matrix mgf and cgf have formal power series expansions:

$$\boldsymbol{M}_{\boldsymbol{X}}(\theta) = \mathbf{I} + \sum_{q=1}^{\infty} \frac{\theta^q}{q!} (\mathbb{E}\, \boldsymbol{X}^q) \quad \text{and} \quad \boldsymbol{\Xi}_{\boldsymbol{X}}(\theta) = \sum_{q=1}^{\infty} \frac{\theta^q}{q!} \boldsymbol{\Psi}_q.$$

We call the coefficients $\mathbb{E}\, \boldsymbol{X}^q$ *matrix moments*, and we refer to $\boldsymbol{\Psi}_q$ as a *matrix cumulant*. The matrix cumulant $\boldsymbol{\Psi}_q$ has a formal expression as a (noncommutative) polynomial in the matrix moments up to order q. In particular, the first cumulant is the mean and the second cumulant is the variance:

$$\boldsymbol{\Psi}_1 = \mathbb{E}\, \boldsymbol{X} \quad \text{and} \quad \boldsymbol{\Psi}_2 = \mathbb{E}\, \boldsymbol{X}^2 - (\mathbb{E}\, \boldsymbol{X})^2 = \mathbf{Var}(\boldsymbol{X})$$

The matrix variance was introduced in (2.2.3). Higher-order cumulants are harder to write down and interpret.

3.2 The Matrix Laplace Transform Method

In the scalar setting, the Laplace transform method allows us to obtain tail bounds for a random variable in terms of its mgf. The starting point for our theory is the observation that a similar result holds in the matrix setting.

Proposition 3.2.1 (Tail Bounds for Eigenvalues)**.** *Let $\boldsymbol{Y}$ be a random Hermitian matrix. For all $t \in \mathbb{R}$,*

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \geq t\} \leq \inf_{\theta > 0} \mathrm{e}^{-\theta t}\, \mathbb{E}\, \mathrm{tr}\, \mathrm{e}^{\theta \boldsymbol{Y}}, \quad \text{and} \tag{3.2.1}$$

$$\mathbb{P}\{\lambda_{\min}(\boldsymbol{Y}) \leq t\} \leq \inf_{\theta < 0} \mathrm{e}^{-\theta t}\, \mathbb{E}\, \mathrm{tr}\, \mathrm{e}^{\theta \boldsymbol{Y}}. \tag{3.2.2}$$

In words, we can control the tail probabilities of the extreme eigenvalues of a random matrix by producing a bound for the *trace* of the matrix mgf. The proof of this fact parallels the classical argument, but there is a twist.

Proof. We begin with (3.2.1). Fix a positive number θ, and observe that

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \geq t\} = \mathbb{P}\left\{\mathrm{e}^{\theta \lambda_{\max}(\boldsymbol{Y})} \geq \mathrm{e}^{\theta t}\right\} \leq \mathrm{e}^{-\theta t}\, \mathbb{E}\, \mathrm{e}^{\theta \lambda_{\max}(\boldsymbol{Y})} = \mathrm{e}^{-\theta t}\, \mathbb{E}\, \mathrm{e}^{\lambda_{\max}(\theta \boldsymbol{Y})}$$

The first identity holds because $a \mapsto \mathrm{e}^{\theta a}$ is a monotone increasing function, so the event does not change under the mapping. The second relation is Markov's inequality (2.2.1). The last holds because the maximum eigenvalue is a positive-homogeneous map, as stated in (2.1.4). To control the exponential, note that

$$\mathrm{e}^{\lambda_{\max}(\theta \boldsymbol{Y})} = \lambda_{\max}\big(\mathrm{e}^{\theta \boldsymbol{Y}}\big) \le \operatorname{tr} \mathrm{e}^{\theta \boldsymbol{Y}}. \tag{3.2.3}$$

The first identity depends on the Spectral Mapping Theorem, Proposition 2.1.3, and the fact that the exponential function is increasing. The inequality follows because the exponential of an Hermitian matrix is positive definite, and (2.1.13) shows that the maximum eigenvalue of a positive-definite matrix is dominated by the trace. Combine the latter two displays to reach

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} \le \mathrm{e}^{-\theta t}\, \mathbb{E} \operatorname{tr} \mathrm{e}^{\theta \boldsymbol{Y}}.$$

This inequality is valid for any positive θ, so we may take an infimum to achieve the tightest possible bound.

To prove (3.2.2), we use a similar approach. Fix a negative number θ, and calculate that

$$\mathbb{P}\{\lambda_{\min}(\boldsymbol{Y}) \le t\} = \mathbb{P}\left\{\mathrm{e}^{\theta \lambda_{\min}(\boldsymbol{Y})} \ge \mathrm{e}^{\theta t}\right\} \le \mathrm{e}^{-\theta t}\, \mathbb{E}\, \mathrm{e}^{\theta \lambda_{\min}(\boldsymbol{Y})} = \mathrm{e}^{-\theta t}\, \mathbb{E}\, \mathrm{e}^{\lambda_{\max}(\theta \boldsymbol{Y})}.$$

The function $a \mapsto \mathrm{e}^{\theta a}$ reverses the inequality in the event because it is monotone decreasing. The last identity depends on the relationship (2.1.5) between minimum and maximum eigenvalues. Finally, we introduce the inequality (3.2.3) for the trace exponential and minimize over negative values of θ. □

In the proof of Proposition 3.2.1, it may seem crude to bound the maximum eigenvalue by the trace. In fact, our overall approach leads to matrix concentration inequalities that are sharp for specific examples (see the discussion in §§4.1.2, 5.1.2, and 6.1.2), so we must conclude that the loss in this bound is sometimes inevitable. At the same time, this maneuver allows us to exploit some amazing convexity properties of the trace exponential.

We can adapt the proof of Proposition 3.2.1 to obtain bounds for the expectation of the maximum eigenvalue of a random Hermitian matrix. This argument does not have a perfect analog in the scalar setting.

Proposition 3.2.2 (Expectation Bounds for Eigenvalues)**.** *Let $\boldsymbol{Y}$ be a random Hermitian matrix. Then*

$$\mathbb{E}\, \lambda_{\max}(\boldsymbol{Y}) \le \inf_{\theta > 0} \frac{1}{\theta} \log \mathbb{E} \operatorname{tr} \mathrm{e}^{\theta \boldsymbol{Y}}, \quad \textit{and} \tag{3.2.4}$$

$$\mathbb{E}\, \lambda_{\min}(\boldsymbol{Y}) \ge \sup_{\theta < 0} \frac{1}{\theta} \log \mathbb{E} \operatorname{tr} \mathrm{e}^{\theta \boldsymbol{Y}}. \tag{3.2.5}$$

Proof. We establish the bound (3.2.4); the proof of (3.2.5) is quite similar. Fix a positive number θ, and calculate that

$$\mathbb{E}\, \lambda_{\max}(\boldsymbol{Y}) = \frac{1}{\theta}\, \mathbb{E} \log \mathrm{e}^{\lambda_{\max}(\theta \boldsymbol{Y})} \le \frac{1}{\theta} \log \mathbb{E}\, \mathrm{e}^{\lambda_{\max}(\theta \boldsymbol{Y})} = \frac{1}{\theta} \log \mathbb{E}\, \lambda_{\max}\big(\mathrm{e}^{\theta \boldsymbol{Y}}\big) \le \frac{1}{\theta} \log \mathbb{E} \operatorname{tr} \mathrm{e}^{\theta \boldsymbol{Y}}.$$

The first identity holds because the maximum eigenvalue is a positive-homogeneous map, as stated in (2.1.4). The second relation is Jensen's inequality. The third follows when we use the Spectral Mapping Theorem, Proposition 2.1.3, to draw the eigenvalue map through the exponential. The final inequality depends on the fact (2.1.13) that the trace of a positive-definite matrix dominates the maximum eigenvalue. □

3.3 The Failure of the Matrix Mgf

We would like the use the Laplace transform bounds from Section 3.2 to study a sum of independent random matrices. In the scalar setting, the Laplace transform method is effective for studying an independent sum because the mgf and the cgf decompose. In the matrix case, the situation is more subtle, and the goal of this section is to indicate where things go awry.

Consider an independent sequence $\{X_k\}$ of real random variables. The mgf of the sum satisfies a multiplication rule:

$$M_{(\sum_k X_k)}(\theta) = \mathbb{E}\exp\left(\sum_k \theta X_k\right) = \mathbb{E}\prod_k \mathrm{e}^{\theta X_k} = \prod_k \mathbb{E}\,\mathrm{e}^{\theta X_k} = \prod_k M_{X_k}(\theta). \tag{3.3.1}$$

The first identity is the definition of an mgf. The second relation holds because the exponential map converts a sum of real scalars to a product, and the third relation requires the independence of the random variables. The last identity, again, is the definition.

At first, we might imagine that a similar relationship holds for the matrix mgf. Consider an independent sequence $\{\boldsymbol{X}_k\}$ of random Hermitian matrices. Perhaps,

$$\boldsymbol{M}_{(\sum_k \boldsymbol{X}_k)}(\theta) \quad \stackrel{?}{=} \quad \prod_k \boldsymbol{M}_{\boldsymbol{X}_k}(\theta). \tag{3.3.2}$$

Unfortunately, this hope shatters when we subject it to interrogation.

It is not hard to find the reason that (3.3.2) fails. The identity (3.3.1) depends on the fact that the scalar exponential converts a sum into a product. In contrast, for Hermitian matrices,

$$\mathrm{e}^{\boldsymbol{A}+\boldsymbol{H}} \neq \mathrm{e}^{\boldsymbol{A}}\mathrm{e}^{\boldsymbol{H}} \quad \text{unless } \boldsymbol{A} \text{ and } \boldsymbol{H} \text{ commute.}$$

If we introduce the trace, the situation improves somewhat:

$$\operatorname{tr}\mathrm{e}^{\boldsymbol{A}+\boldsymbol{H}} \leq \operatorname{tr}\mathrm{e}^{\boldsymbol{A}}\mathrm{e}^{\boldsymbol{H}} \quad \text{for all Hermitian } \boldsymbol{A} \text{ and } \boldsymbol{H}. \tag{3.3.3}$$

The result (3.3.3) is known as the Golden–Thompson inequality, a famous theorem from statistical physics. Unfortunately, the analogous bound may fail for three matrices:

$$\operatorname{tr}\mathrm{e}^{\boldsymbol{A}+\boldsymbol{H}+\boldsymbol{T}} \not\leq \operatorname{tr}\mathrm{e}^{\boldsymbol{A}}\mathrm{e}^{\boldsymbol{H}}\mathrm{e}^{\boldsymbol{T}} \quad \text{for certain Hermitian } \boldsymbol{A}, \boldsymbol{H}, \text{ and } \boldsymbol{T}.$$

It seems that we have reached an impasse.

What if we consider the cgf instead? The cgf of a sum of independent real random variables satisfies an addition rule:

$$\Xi_{(\sum_k X_k)}(\theta) = \log \mathbb{E}\exp\left(\sum_k \theta X_k\right) = \log \prod_k \mathbb{E}\,\mathrm{e}^{\theta X_k} = \sum_k \Xi_{X_k}(\theta). \tag{3.3.4}$$

The relation (3.3.4) follows when we extract the logarithm of the multiplication rule (3.3.1). This result looks like a more promising candidate for generalization because a sum of Hermitian matrices remains Hermitian. We might hope that

$$\boldsymbol{\Xi}_{(\sum_k \boldsymbol{X}_k)}(\theta) \quad \stackrel{?}{=} \quad \sum_k \boldsymbol{\Xi}_{\boldsymbol{X}_k}(\theta).$$

As stated, this putative identity also fails. Nevertheless, the addition rule (3.3.4) admits a very satisfactory extension to matrices. In contrast with the scalar case, the proof involves much deeper considerations.

3.4 A Theorem of Lieb

To find the appropriate generalization of the addition rule for cgfs, we turn to the literature on matrix analysis. Here, we discover a famous result of Elliott Lieb on the convexity properties of the trace exponential function.

Theorem 3.4.1 (Lieb)**.** *Fix an Hermitian matrix $\boldsymbol{H}$ with dimension d. The function*

$$\boldsymbol{A} \longmapsto \operatorname{tr} \exp(\boldsymbol{H} + \log \boldsymbol{A})$$

is a concave map on the convex cone of $d \times d$ positive-definite matrices.

In the scalar case, the analogous function $a \mapsto \exp(h + \log a)$ is linear, so this result describes a new type of phenomenon that emerges when we move to the matrix setting. We present a complete proof of Theorem 3.4.1 in Chapter 8.

For now, let us focus on the consequences of this remarkable result. Lieb's Theorem is valuable to us because the Laplace transform bounds from Section 3.2 involve the trace exponential function. To highlight the connection, we rephrase Theorem 3.4.1 in probabilistic terms.

Corollary 3.4.2. *Let $\boldsymbol{H}$ be a fixed Hermitian matrix, and let $\boldsymbol{X}$ be a random Hermitian matrix of the same dimension. Then*

$$\mathbb{E} \operatorname{tr} \exp(\boldsymbol{H} + \boldsymbol{X}) \le \operatorname{tr} \exp\bigl(\boldsymbol{H} + \log \mathbb{E}\, \mathrm{e}^{\boldsymbol{X}}\bigr).$$

Proof. Introduce the random matrix $\boldsymbol{Y} = \mathrm{e}^{\boldsymbol{X}}$. Then

$$\begin{aligned} \mathbb{E} \operatorname{tr} \exp(\boldsymbol{H} + \boldsymbol{X}) &= \mathbb{E} \operatorname{tr} \exp(\boldsymbol{H} + \log \boldsymbol{Y}) \\ &\le \operatorname{tr} \exp(\boldsymbol{H} + \log \mathbb{E}\, \boldsymbol{Y}) = \operatorname{tr} \exp\bigl(\boldsymbol{H} + \log \mathbb{E}\, \mathrm{e}^{\boldsymbol{X}}\bigr). \end{aligned}$$

The first identity follows from the interpretation (2.1.17) of the matrix logarithm as the functional inverse of the matrix exponential. Theorem 3.4.1 shows that the trace function is concave in $\boldsymbol{Y}$, so Jensen's inequality (2.2.2) allows us to draw the expectation inside the function. □

3.5 Subadditivity of the Matrix Cgf

We are now prepared to generalize the addition rule (3.3.4) for scalar cgfs to the matrix setting. The following result is fundamental to our approach to random matrices.

Lemma 3.5.1 (Subadditivity of Matrix Cgfs)**.** *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, Hermitian matrices of the same dimension. Then*

$$\mathbb{E} \operatorname{tr} \exp\bigl(\textstyle\sum_k \theta \boldsymbol{X}_k\bigr) \le \operatorname{tr} \exp\Bigl(\textstyle\sum_k \log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}_k}\Bigr) \quad \text{for } \theta \in \mathbb{R}. \tag{3.5.1}$$

Equivalently,

$$\operatorname{tr} \exp\bigl(\boldsymbol{\Xi}_{(\sum_k \boldsymbol{X}_k)}(\theta)\bigr) \le \operatorname{tr} \exp\bigl(\textstyle\sum_k \boldsymbol{\Xi}_{\boldsymbol{X}_k}(\theta)\bigr) \quad \text{for } \theta \in \mathbb{R}. \tag{3.5.2}$$

The parallel between the additivity rule (3.3.4) and the subadditivity rule (3.5.2) is striking. With our level of preparation, it is easy to prove this result. We just apply the bound from Corollary 3.4.2 repeatedly.

Proof. Without loss of generality, we assume that $\theta = 1$ by absorbing the parameter into the random matrices. Let $\mathbb{E}_k$ denote the expectation with respect to $\boldsymbol{X}_k$, the remaining random matrices held fixed. Abbreviate

$$\boldsymbol{\Xi}_k = \log \mathbb{E}_k \mathrm{e}^{\boldsymbol{X}_k} = \log \mathbb{E} \mathrm{e}^{\boldsymbol{X}_k}.$$

We may calculate that

$$\begin{aligned}
\mathbb{E} \operatorname{tr} \exp\left(\textstyle\sum_{k=1}^{n} \boldsymbol{X}_k\right) &= \mathbb{E}\mathbb{E}_n \operatorname{tr} \exp\left(\textstyle\sum_{k=1}^{n-1} \boldsymbol{X}_k + \boldsymbol{X}_n\right) \\
&\le \mathbb{E} \operatorname{tr} \exp\left(\textstyle\sum_{k=1}^{n-1} \boldsymbol{X}_k + \log \mathbb{E}_n \mathrm{e}^{\boldsymbol{X}_n}\right) \\
&= \mathbb{E}\mathbb{E}_{n-1} \operatorname{tr} \exp\left(\textstyle\sum_{k=1}^{n-2} \boldsymbol{X}_k + \boldsymbol{X}_{n-1} + \boldsymbol{\Xi}_n\right) \\
&\le \mathbb{E}\mathbb{E}_{n-2} \operatorname{tr} \exp\left(\textstyle\sum_{k=1}^{n-2} \boldsymbol{X}_k + \boldsymbol{\Xi}_{n-1} + \boldsymbol{\Xi}_n\right) \\
\cdots \quad &\le \operatorname{tr} \exp\left(\textstyle\sum_{k=1}^{n} \boldsymbol{\Xi}_k\right).
\end{aligned}$$

We can introduce iterated expectations because of the tower property of conditional expectation. To bound the expectation $\mathbb{E}_m$ for an index $m = 1, 2, 3, \ldots, n$, we invoke Corollary 3.4.2 with the fixed matrix $\boldsymbol{H}$ equal to

$$\boldsymbol{H}_m = \sum_{k=1}^{m-1} \boldsymbol{X}_k + \sum_{k=m+1}^{n} \boldsymbol{\Xi}_k.$$

This argument is legitimate because $\boldsymbol{H}_m$ is independent from $\boldsymbol{X}_m$.

The formulation (3.5.2) follows from (3.5.1) when we substitute the expression (3.1.1) for the matrix cgf and make some algebraic simplifications. □

3.6 Master Bounds for Sums of Independent Random Matrices

Finally, we can present some general results on the behavior of a sum of independent random matrices. At this stage, we simply combine the Laplace transform bounds with the subadditivity of the matrix cgf to obtain abstract inequalities. Later, we will harness properties of the summands to develop more concrete estimates that apply to specific examples of interest.

Theorem 3.6.1 (Master Bounds for a Sum of Independent Random Matrices)**.** *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, Hermitian matrices of the same size. Then*

$$\mathbb{E} \lambda_{\max}\left(\textstyle\sum_k \boldsymbol{X}_k\right) \le \inf_{\theta > 0} \frac{1}{\theta} \log \operatorname{tr} \exp\left(\textstyle\sum_k \log \mathbb{E} \mathrm{e}^{\theta \boldsymbol{X}_k}\right), \quad \textit{and} \tag{3.6.1}$$

$$\mathbb{E} \lambda_{\min}\left(\textstyle\sum_k \boldsymbol{X}_k\right) \ge \sup_{\theta < 0} \frac{1}{\theta} \log \operatorname{tr} \exp\left(\textstyle\sum_k \log \mathbb{E} \mathrm{e}^{\theta \boldsymbol{X}_k}\right). \tag{3.6.2}$$

Furthermore, for all $t \in \mathbb{R}$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\textstyle\sum_k \boldsymbol{X}_k\right) \ge t\right\} \le \inf_{\theta > 0} \mathrm{e}^{-\theta t} \operatorname{tr} \exp\left(\textstyle\sum_k \log \mathbb{E} \mathrm{e}^{\theta \boldsymbol{X}_k}\right), \quad \textit{and} \tag{3.6.3}$$

$$\mathbb{P}\left\{\lambda_{\min}\left(\textstyle\sum_k \boldsymbol{X}_k\right) \le t\right\} \le \inf_{\theta < 0} \mathrm{e}^{-\theta t} \operatorname{tr} \exp\left(\textstyle\sum_k \log \mathbb{E} \mathrm{e}^{\theta \boldsymbol{X}_k}\right). \tag{3.6.4}$$

Proof. Substitute the subadditivity rule for matrix cgfs, Lemma 3.5.1, into the two matrix Laplace transform results, Proposition 3.2.1 and Proposition 3.2.2. □

In this chapter, we have focused on probability inequalities for the extreme eigenvalues of a sum of independent random matrices. Nevertheless, these results also give information about the spectral norm of a sum of independent, random, rectangular matrices because we can apply them to the Hermitian dilation (2.1.26) of the sum. Instead of presenting a general theorem, we find it more natural to extend individual results to the non-Hermitian case.

3.7 Notes

This section includes some historical discussion about the results we have described in this chapter, along with citations for the results that we have established.

3.7.1 The Matrix Laplace Transform Method

The idea of lifting the "Bernstein trick" to the matrix setting is due to two researchers in quantum information theory, Rudolf Ahlswede and Andreas Winter, who were working on a problem concerning transmission of information through a quantum channel [AW02]. Their paper contains a version of the matrix Laplace transform result, Proposition 3.2.1, along with a substantial number of related foundational ideas. Their work is one of the major inspirations for the tools that are described in these notes.

The statement of Proposition 3.2.1 and the proof that we present appear in the paper [Oli10b] of Roberto Oliveira. The subsequent result on expectations, Proposition 3.2.2, first appeared in the paper [CGT12a].

3.7.2 Subadditivity of Cumulants

The major impediment to applying the matrix Laplace transform method is the need to produce a bound for the trace of the matrix moment generating function (the trace mgf). This is where all the technical difficulty in the argument resides.

Ahlswede & Winter [AW02, App.] proposed an approach for bounding the trace mgf of an independent sum, based on a repeated application of the Golden–Thompson inequality (3.3.3). Their argument leads to a cumulant bound of the form

$$\mathbb{E}\,\mathrm{tr}\exp\bigl(\textstyle\sum_k \boldsymbol{X}_k\bigr) \le d\cdot\exp\bigl(\textstyle\sum_k \lambda_{\max}\bigl(\log \mathbb{E}\,\mathrm{e}^{\boldsymbol{X}_k}\bigr)\bigr) \tag{3.7.1}$$

when the random Hermitian matrices $\boldsymbol{X}_k$ have dimension d. In other words, Ahlswede & Winter bound the cumulant of a sum in terms of the sum of the *maximum eigenvalues* of the cumulants. There are cases where the bound (3.7.1) is equivalent with Lemma 3.5.1. For example, the estimates coincide when each matrix $\boldsymbol{X}_k$ is identically distributed. In general, however, the estimate (3.7.1) leads to fundamentally weaker results than our bound from Lemma 3.5.1. In the worst case, the approach of Ahlswede & Winter may produce an unnecessary factor of the dimension d in the *exponent*. See [Tro11c, §§3.7, 4.8] for details.

The first major technical advance beyond the original argument of Ahlswede & Winter appeared in a paper [Oli10a] of Oliveira. He developed a more effective way to deploy the Golden–Thompson inequality, and he used this technique to establish a matrix version of Freedman's inequality [Fre75]. In the scalar setting, Freedman's inequality extends the Bernstein concentration inequality to martingales; Oliveira obtained the analogous extension of Bernstein's inequality for matrix-valued martingales. When specialized to independent sums, his result is quite

similar with the matrix Bernstein inequality, Theorem 1.6.2, apart from the precise values of the constants. Oliveira's method, however, does not seem to deliver the full spectrum of matrix concentration inequalities that we discuss in these notes.

The approach here, based on Lieb's Theorem, was introduced in the article [Tro11c] by the author of these notes. This paper was apparently the first to recognize that Lieb's Theorem has probabilistic content, as stated in Corollary 3.4.2. This idea leads to Lemma 3.5.1, on the subadditivity of cumulants, along with the master tail bounds from Theorem 3.6.1. Note that the two articles [Oli10a, Tro11c] are independent works.

For a detailed discussion of the benefits of Lieb's Theorem over the Golden–Thompson inequality, see [Tro11c, §4]. In summary, to get the sharpest concentration results for random matrices, Lieb's Theorem appears to be indispensible. The approach of Ahlswede & Winter seems intrinsically weaker. Oliveira's argument has certain advantages, however, in that it extends from matrices to the fully noncommutative setting [JZ12].

Subsequent research on the underpinnings of the matrix Laplace transform method has led to a martingale version of the subadditivity of cumulants [Tro11a, Tro11b]; these works also depend on Lieb's Theorem. The technical report [GT14] shows how to use a related result, called the Lieb–Seiringer Theorem [LS05], to obtain upper and lower tail bounds for all eigenvalues of a sum of independent random Hermitian matrices.

3.7.3 Noncommutative Moment Inequalities

There is a closely related, and much older, line of research on noncommutative moment inequalities. These results provide information about the expected trace of a power of a sum of independent random matrices. The matrix Laplace transform method, as encapsulated in Theorem 3.6.1, gives analogous bounds for the exponential moments.

Research on noncommutative moment inequalities dates to an important paper [LP86] of Françoise Lust-Piquard, which contains an operator extension of the Khintchine inequality. Her result, now called the *noncommutative Khintchine inequality*, controls the trace moments of a sum of fixed matrices, each modulated by an independent Rademacher random variable; see Section 4.7.2 for more details.

In recent years, researchers have generalized many other moment inequalities for a sum of scalar random variables to matrices (and beyond). For instance, the Rosenthal–Pinelis inequality for a sum of independent zero-mean random variables admits a matrix version [JZ13, MJC+14, CGT12a]. We present a variant of the latter result below in (6.1.6). See the paper [JX05] for a good overview of some other noncommutative moment inequalities.

Finally, and tangentially, we mention that a different notion of matrix moments and cumulants plays a central role in the theory of free probability [NS06].

3.7.4 Quantum Statistical Mechanics

A curious feature of the theory of matrix concentration inequalities is that the most powerful tools come from the mathematical theory of quantum statistical mechanics. This field studies the bulk statistical properties of interacting quantum systems, and it would seem quite distant from the field of random matrix theory. The connection between these two areas has emerged because of research on quantum information theory, which studies how information can be encoded, operated upon, and transmitted via quantum mechanical systems.

The Golden–Thompson inequality is a major result from quantum statistical mechanics. Bhatia's book [Bha97, Sec. IX.3] contains a detailed treatment of this result from the perspective of matrix theory. For an account with more physical content, see the book of Thirring [Thi02]. The fact that the Golden–Thompson inequality fails for three matrices can be obtained from simple examples, such as combinations of Pauli spin matrices [Bha97, Exer. IX.8.4].

Lieb's Theorem [Lie73, Thm. 6] was first established in an important paper of Elliott Lieb on the convexity of trace functions. His main goal was to establish concavity properties for a function that measures the amount of information in a quantum system. See the notes in Chapter 8 for a more detailed discussion.

CHAPTER 4

Matrix Gaussian Series & Matrix Rademacher Series

In this chapter, we present our first set of matrix concentration inequalities. These results provide spectral information about a sum of fixed matrices, each modulated by an independent scalar random variable. This type of formulation is surprisingly versatile, and it captures a range of interesting examples. Our main goal, however, is to introduce matrix concentration in the simplest setting possible.

To be more precise about our scope, let us introduce the concept of a matrix Gaussian series. Consider a finite sequence $\{\boldsymbol{B}_k\}$ of fixed matrices with the same dimension, along with a finite sequence $\{\gamma_k\}$ of independent standard normal random variables. We will study the spectral norm of the random matrix

$$\boldsymbol{Z} = \sum\nolimits_k \gamma_k \boldsymbol{B}_k.$$

This expression looks abstract, but it has concrete modeling power. For example, we can express a Gaussian Wigner matrix, one of the classical random matrices, in this fashion. But the real value of this approach is that we can use matrix Gaussian series to represent many kinds of random matrices built from Gaussian random variables. This technique allows us to attack problems that classical methods do not handle gracefully. For instance, we can easily study a Toeplitz matrix with Gaussian entries.

Similar ideas allow us to treat a *matrix Rademacher series*, a sum of fixed matrices modulated by random signs. (Recall that a *Rademacher random variable* takes the values ± 1 with equal probability.) The results in this case are almost identical with the results for matrix Gaussian series, but they allow us to consider new problems. As an example, we can study the expected spectral norm of a fixed real matrix after flipping the signs of the entries at random.

Overview

In §4.1, we begin with an overview of our results for matrix Gaussian series; very similar results also hold for matrix Rademacher series. Afterward, we discuss the accuracy of the theoretical

bounds. The subsequent sections, §§4.2–4.4, describe what the matrix concentration inequalities tell us about some classical and not-so-classical examples of random matrices. Section 4.5 includes an overview of a more substantial application in combinatorial optimization. The final part §4.6 contains detailed proofs of the bounds. We conclude with bibliographical notes.

4.1 A Norm Bound for Random Series with Matrix Coefficients

Consider a finite sequence $\{b_k\}$ of real numbers and a finite sequence $\{\gamma_k\}$ of independent standard normal random variables. Form the random series $Z = \sum_k \gamma_k b_k$. A routine invocation of the scalar Laplace transform method demonstrates that

$$\mathbb{P}\{|Z| \geq t\} \leq 2\exp\left(\frac{-t^2}{2v}\right) \quad \text{where } v = \operatorname{Var}(Z) = \sum\nolimits_k b_k^2. \tag{4.1.1}$$

It turns out that the inequality (4.1.1) extends directly to the matrix setting.

Theorem 4.1.1 (Matrix Gaussian & Rademacher Series)**.** *Consider a finite sequence $\{\boldsymbol{B}_k\}$ of fixed complex matrices with dimension $d_1 \times d_2$, and let $\{\gamma_k\}$ be a finite sequence of independent standard normal variables. Introduce the matrix Gaussian series*

$$\boldsymbol{Z} = \sum\nolimits_k \gamma_k \boldsymbol{B}_k. \tag{4.1.2}$$

Let $v(\boldsymbol{Z})$ be the matrix variance statistic of the sum:

$$v(\boldsymbol{Z}) = \max\{\|\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*)\|, \|\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})\|\} \tag{4.1.3}$$

$$= \max\left\{\left\|\sum\nolimits_k \boldsymbol{B}_k\boldsymbol{B}_k^*\right\|, \left\|\sum\nolimits_k \boldsymbol{B}_k^*\boldsymbol{B}_k\right\|\right\}. \tag{4.1.4}$$

Then

$$\mathbb{E}\|\boldsymbol{Z}\| \leq \sqrt{2v(\boldsymbol{Z})\log(d_1+d_2)}. \tag{4.1.5}$$

Furthermore, for all $t \geq 0$,

$$\mathbb{P}\{\|\boldsymbol{Z}\| \geq t\} \leq (d_1+d_2)\exp\left(\frac{-t^2}{2v(\boldsymbol{Z})}\right). \tag{4.1.6}$$

The same bounds hold when we replace $\{\gamma_k\}$ by a finite sequence $\{\varepsilon_k\}$ of independent Rademacher random variables.

The proof of Theorem 4.1.1 appears below in §4.6.

4.1.1 Discussion

Let us take a moment to discuss the content of Theorem 4.1.1. The main message is that the expectation of $\|\boldsymbol{Z}\|$ is controlled by the matrix variance statistic $v(\boldsymbol{Z})$. Furthermore, $\|\boldsymbol{Z}\|$ has a subgaussian tail whose decay rate depends on $v(\boldsymbol{Z})$.

The matrix variance statistic $v(\boldsymbol{Z})$ defined in (4.1.3) specializes the general formulation (2.2.8). The second expression (4.1.4) follows from the additivity property (2.2.11) for the variance of an independent sum. When the summands are Hermitian, observe that the two terms in the maximum coincide. The formulas (4.1.3) and (4.1.4) are a direct extension of the variance that arises in the scalar bound (4.1.1).

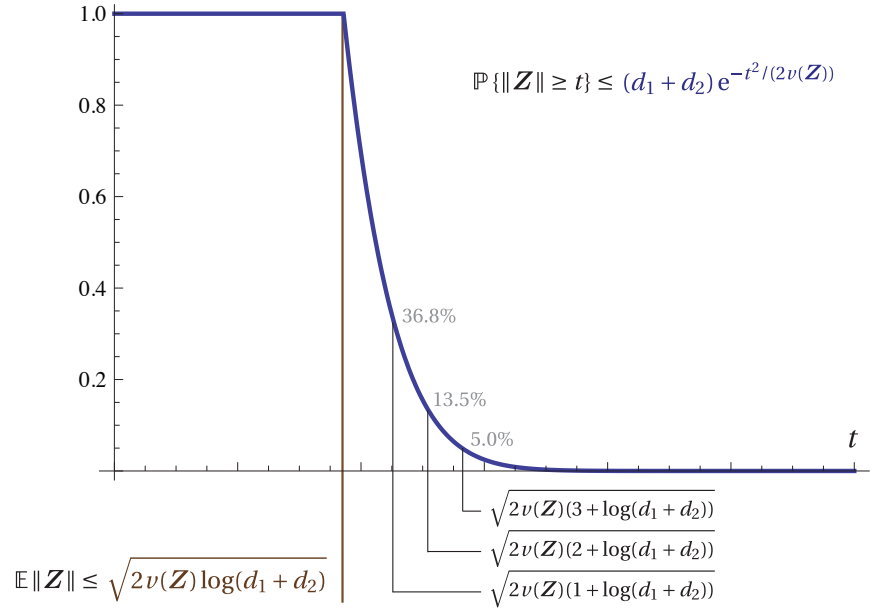

Figure 4.1: **Schematic of tail bound for matrix Gaussian series.** Consider a matrix Gaussian series $\boldsymbol{Z}$ with dimension $d_1 \times d_2$. The tail probability $\mathbb{P}\{\|\boldsymbol{Z}\| \geq t\}$ admits the upper bound $(d_1+d_2)\exp(-t^2/(2v(\boldsymbol{Z})))$, marked as a dark blue curve. This estimate provides no information below the level $t = \sqrt{2v(\boldsymbol{Z})\log(d_1+d_2)}$. This value, the dark red vertical line, coincides with the upper bound (4.1.5) for $\mathbb{E}\|\boldsymbol{Z}\|$. As t increases beyond this point, the tail probability decreases at a subgaussian rate with variance on the order of $v(\boldsymbol{Z})$.

As compared with (4.1.1), a new feature of the bound (4.1.6) is the dimensional factor d_1+d_2. When $d_1 = d_2 = 1$, the matrix bound reduces to the scalar result (4.1.1). In this case, at least, we have lost nothing by lifting the Laplace transform method to matrices. The behavior of the matrix tail bound (4.1.6) is more subtle than the behavior of the scalar tail bound (4.1.1). See Figure 4.1 for an illustration.

4.1.2 Optimality of the Bounds for Matrix Gaussian Series

One may wonder whether Theorem 4.1.1 provides accurate information about the behavior of a matrix Gaussian series. The answer turns out to be complicated. Here is the executive summary: the expectation bound (4.1.5) is always quite good, but the tail bound (4.1.6) is sometimes quite bad. The rest of this section expands on these claims.

The Expectation Bound

Let $\boldsymbol{Z}$ be a matrix Gaussian series of the form (4.1.2). We will argue that

$$v(\boldsymbol{Z}) \quad \leq \quad \mathbb{E}\|\boldsymbol{Z}\|^2 \quad \leq \quad 2v(\boldsymbol{Z})(1+\log(d_1+d_2)). \tag{4.1.7}$$

In other words, the matrix variance $v(\boldsymbol{Z})$ is roughly the correct scale for $\|\boldsymbol{Z}\|^2$. This pair of estimates is a significant achievement because it is quite challenging to compute the norm of a matrix Gaussian series in general. Indeed, the literature contains very few examples where explicit estimates are available, especially if one desires reasonable constants.

We begin with the lower bound in (4.1.7), which is elementary. Indeed, since the spectral norm is convex, Jensen's inequality ensures that

$$\mathbb{E}\,\|\boldsymbol{Z}\|^2 = \mathbb{E}\max\left\{\|\boldsymbol{Z}\boldsymbol{Z}^*\|,\ \|\boldsymbol{Z}^*\boldsymbol{Z}\|\right\} \geq \max\left\{\|\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*)\|,\ \|\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})\|\right\} = v(\boldsymbol{Z}).$$

The first identity follows from (2.1.24), and the last is the definition (2.2.8) of the matrix variance.

The upper bound in (4.1.7) is a consequence of the tail bound (4.1.6):

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{Z}\|^2 &= \int_0^\infty 2t\,\mathbb{P}\{\|\boldsymbol{Z}\| \geq t\}\,\mathrm{d}t \\ &\leq \int_0^E 2t\,\mathrm{d}t + 2(d_1+d_2)\int_E^\infty t\,\mathrm{e}^{-t^2/(2v(\boldsymbol{Z}))}\,\mathrm{d}t = E^2 + 2v(\boldsymbol{Z})\,(d_1+d_2)\,\mathrm{e}^{-E^2/(2v(\boldsymbol{Z}))}.\end{aligned}$$

In the first step, rewrite the expectation using integration by parts, and then split the integral at a positive number E. In the first term, we bound the probability by one, while the second term results from the tail bound (4.1.6). Afterward, we compute the integrals explicitly. Finally, select $E^2 = 2v(\boldsymbol{Z})\log(d_1+d_2)$ to complete the proof of (4.1.7).

About the Dimensional Factor

At this point, one may ask whether it is possible to improve either side of the inequality (4.1.7). The answer is negative unless we have additional information about the Gaussian series beyond the matrix variance statistic $v(\boldsymbol{Z})$.

Indeed, for arbitrarily large dimensions d_1 and d_2, we can exhibit a matrix Gaussian series where the left-hand inequality in (4.1.7) is correct. That is, $\mathbb{E}\,\|\boldsymbol{Z}\|^2 \approx v(\boldsymbol{Z})$ with no additional dependence on the dimensions d_1 or d_2. One such example appears below in §4.2.2.

At the same time, for arbitrarily large dimensions d_1 and d_2, we can construct a matrix Gaussian series where the right-hand inequality in (4.1.7) is correct. That is, $\mathbb{E}\,\|\boldsymbol{Z}\|^2 \approx v(\boldsymbol{Z})\log(d_1+d_2)$. See §4.4 for an example.

We can offer a rough intuition about how these two situations differ from each other. The presence or absence of the dimensional factor $\log(d_1+d_2)$ depends on how much the coefficients $\boldsymbol{B}_k$ in the matrix Gaussian series $\boldsymbol{Z}$ commute with each other. More commutativity leads to a logarithm, while less commutativity can sometimes result in cancelations that obliterate the logarithm. It remains a major open question to find a simple quantity, computable from the coefficients $\boldsymbol{B}_k$, that decides whether $\mathbb{E}\,\|\boldsymbol{Z}\|^2$ contains a dimensional factor or not.

In Chapter 7, we will describe a technique that allows us to moderate the dimensional factor in (4.1.7) for some types of matrix series. But we cannot remove the dimensional factor entirely with current technology.

The Tail Bound

What about the tail bound (4.1.6) for the norm of the Gaussian series? Here, our results are less impressive. It turns out that the large-deviation behavior of the spectral norm of a matrix Gaussian series $\boldsymbol{Z}$ is controlled by a statistic $v_\star(\boldsymbol{Z})$ called the *weak variance*:

$$v_\star(\boldsymbol{Z}) = \sup_{\|\boldsymbol{u}\|=\|\boldsymbol{w}\|=1} \mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{Z}\boldsymbol{w}|^2 = \sup_{\|\boldsymbol{u}\|=\|\boldsymbol{w}\|=1} \sum\nolimits_k |\boldsymbol{u}^*\boldsymbol{B}_k\boldsymbol{w}|^2.$$

The best general inequalities between the matrix variance statistic and the weak variance are

$$v_\star(\boldsymbol{Z}) \quad \le \quad v(\boldsymbol{Z}) \quad \le \quad \min\{d_1, d_2\} \cdot v_\star(\boldsymbol{Z})$$

There are examples of matrix Gaussian series that saturate the lower or the upper inequality.

The classical concentration inequality [BLM13, Thm. 5.6] for a function of independent Gaussian random variables implies that

$$\mathbb{P}\{\|\boldsymbol{Z}\| \ge \mathbb{E}\|\boldsymbol{Z}\| + t\} \le \mathrm{e}^{-t^2/(2v_\star(\boldsymbol{Z}))}. \tag{4.1.8}$$

Let us emphasize that the bound (4.1.8) provides no information about $\mathbb{E}\|\boldsymbol{Z}\|$; it only tells us about the probability that $\|\boldsymbol{Z}\|$ is larger than its mean.

Together, the last two displays indicate that the exponent in the tail bound (4.1.6) is sometimes too big by a factor $\min\{d_1, d_2\}$. Therefore, a direct application of Theorem 4.1.1 can badly overestimate the tail probability $\mathbb{P}\{\|\boldsymbol{Z}\| > t\}$ when the level t is large. Fortunately, this problem is less pronounced with the matrix Chernoff inequalities of Chapter 5 and the matrix Bernstein inequalities of Chapter 6.

Expectations and Tails

When studying concentration of random variables, it is quite common that we need to use one method to assess the expected value of the random variable and a separate technique to determine the probability of a large deviation.

> **The primary value of matrix concentration inequalities inheres in the estimates that they provide for the expectation of the spectral norm (or maximum eigenvalue or minimum eigenvalue) of a random matrix.**

In many cases, matrix concentration bounds provide reasonable information about the tail decay, but there are other situations where the tail bounds are feeble. In this event, we recommend applying a scalar concentration inequality to control the tails.

4.2 Example: Some Gaussian Matrices

Let us try out our methods on two types of Gaussian matrices that have been studied extensively in the classical literature on random matrix theory. In these cases, precise information about the spectral distribution is available, which provides a benchmark for assessing our results. We find that bounds based on Theorem 4.1.1 lead to very reasonable estimates, but they are not sharp. The advantage of our approach is that it applies to every example, whereas we are making comparisons with specialized techniques that only illuminate individual cases. Similar conclusions hold for matrices with independent Rademacher entries.

4.2.1 Gaussian Wigner Matrices

We begin with a family of Gaussian Wigner matrices. A $d \times d$ matrix $\boldsymbol{W}_d$ from this ensemble is real-symmetric with a zero diagonal; the entries above the diagonal are independent normal

variables with mean zero and variance one:

$$\boldsymbol{W}_d = \begin{bmatrix} 0 & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1d} \\ \gamma_{12} & 0 & \gamma_{23} & \cdots & \gamma_{2d} \\ \gamma_{13} & \gamma_{23} & 0 & & \gamma_{3d} \\ \vdots & \vdots & & \ddots & \vdots \\ \gamma_{1d} & \gamma_{2d} & \cdots & \gamma_{d-1,d} & 0 \end{bmatrix}$$

where $\{\gamma_{jk} : 1 \le j < k \le d\}$ is an independent family of standard normal variables. We can represent this matrix compactly as a Gaussian series:

$$\boldsymbol{W}_d = \sum_{1 \le j < k \le d} \gamma_{jk} (\mathbf{E}_{jk} + \mathbf{E}_{kj}). \tag{4.2.1}$$

The norm of a Wigner matrix satisfies

$$\frac{1}{\sqrt{d}} \|\boldsymbol{W}_d\| \longrightarrow 2 \quad \text{as } d \to \infty, \text{ almost surely.} \tag{4.2.2}$$

For example, see [BS10, Thm. 5.1]. To make (4.2.2) precise, we assume that $\{\boldsymbol{W}_d\}$ is an independent sequence of Gaussian Wigner matrices, indexed by the dimension d.

Theorem 4.6.1 provides a simple way to bound the norm of a Gaussian Wigner matrix. We just need to compute the matrix variance statistic $v(\boldsymbol{W}_d)$. The formula (4.1.4) for $v(\boldsymbol{W}_d)$ asks us to form the sum of the squared coefficients from the representation (4.2.1):

$$\sum_{1 \le j < k \le d} (\mathbf{E}_{jk} + \mathbf{E}_{kj})^2 = \sum_{1 \le j < k \le d} (\mathbf{E}_{jj} + \mathbf{E}_{kk}) = (d-1)\,\mathbf{I}_d.$$

Since the terms in (4.2.1) are Hermitian, we have only one sum of squares to consider. We have also used the facts that $\mathbf{E}_{jk}\mathbf{E}_{kj} = \mathbf{E}_{jj}$ while $\mathbf{E}_{jk}\mathbf{E}_{jk} = \mathbf{0}$ because of the condition $j < k$ in the limits of summation. We see that

$$v(\boldsymbol{W}_d) = \left\| \sum_{1 \le j < k \le d} (\mathbf{E}_{jk} + \mathbf{E}_{kj})^2 \right\| = \|(d-1)\,\mathbf{I}_d\| = d - 1.$$

The bound (4.1.5) for the expectation of the norm gives

$$\mathbb{E}\,\|\boldsymbol{W}_d\| \le \sqrt{2(d-1)\log(2d)}. \tag{4.2.3}$$

In conclusion, our techniques overestimate $\|\boldsymbol{W}_d\|$ by a factor of about $\sqrt{0.5 \log d}$. The result (4.2.3) is not perfect, but it only takes two lines of work. In contrast, the classical result (4.2.2) depends on a long moment calculation that involves challenging combinatorial arguments.

4.2.2 Rectangular Gaussian Matrices

Next, we consider a $d_1 \times d_2$ rectangular matrix with independent standard normal entries:

$$\boldsymbol{G} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1d_2} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \cdots & \gamma_{2d_2} \\ \vdots & \vdots & & \ddots & \vdots \\ \gamma_{d_1 1} & \gamma_{d_1 2} & \gamma_{d_1 3} & \cdots & \gamma_{d_1 d_2} \end{bmatrix}$$

where $\{\gamma_{jk}\}$ is an independent family of standard normal variables. We can express this matrix efficiently using a Gaussian series:

$$\boldsymbol{G} = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \gamma_{jk} \mathbf{E}_{jk}, \tag{4.2.4}$$

There is an elegant estimate [DS02, Thm. 2.13] for the norm of this matrix:

$$\mathbb{E}\,\|\boldsymbol{G}\| \le \sqrt{d_1} + \sqrt{d_2}. \tag{4.2.5}$$

The inequality (4.2.5) is sharp when d_1 and d_2 tend to infinity while the ratio $d_1/d_2 \to$ const. See [BS10, Thm. 5.8] for details.

Theorem 4.1.1 yields another bound on the expected norm of the matrix $\boldsymbol{G}$. In order to compute the matrix variance statistic $v(\boldsymbol{G})$, we calculate the sums of the squared coefficients from the representation (4.2.4):

$$\sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \mathbf{E}_{jk} \mathbf{E}_{jk}^* = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \mathbf{E}_{jj} = d_2 \mathbf{I}_{d_1}, \quad \text{and}$$
$$\sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \mathbf{E}_{jk}^* \mathbf{E}_{jk} = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \mathbf{E}_{kk} = d_1 \mathbf{I}_{d_2}.$$

The matrix variance statistic (4.1.3) satisfies

$$v(\boldsymbol{G}) = \max\left\{ \|d_2 \mathbf{I}_{d_1}\|, \|d_1 \mathbf{I}_{d_2}\| \right\} = \max\{d_1, d_2\}.$$

We conclude that

$$\mathbb{E}\,\|\boldsymbol{G}\| \le \sqrt{2 \max\{d_1, d_2\} \log(d_1 + d_2)}. \tag{4.2.6}$$

The leading term is roughly correct because

$$\sqrt{d_1} + \sqrt{d_2} \le 2\sqrt{\max\{d_1, d_2\}} \le 2\left(\sqrt{d_1} + \sqrt{d_2}\right).$$

The logarithmic factor in (4.2.6) does not belong, but it is rather small in comparison with the leading terms. Once again, we have produced a reasonable result with a short argument based on general principles.

4.3 Example: Matrices with Randomly Signed Entries

Next, we turn to an example that is superficially similar with the matrix discussed in §4.2.2 but is less understood. Consider a fixed $d_1 \times d_2$ matrix $\boldsymbol{B}$ with real entries, and let $\{\varrho_{jk}\}$ be an independent family of Rademacher random variables. Consider the $d_1 \times d_2$ random matrix

$$\boldsymbol{B}_\pm = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \varrho_{jk} b_{jk} \mathbf{E}_{jk}$$

In other words, we obtain the random matrix $\boldsymbol{B}_\pm$ by randomly flipping the sign of each entry of $\boldsymbol{B}$. The expected norm of this matrix satisfies the bound

$$\mathbb{E}\,\|\boldsymbol{B}_\pm\| \le \text{Const} \cdot v^{1/2} \cdot \log^{1/4} \min\{d_1, d_2\}, \tag{4.3.1}$$

where the leading factor $v^{1/2}$ satisfies

$$v = \max\left\{\max_j \|\boldsymbol{b}_{j:}\|^2, \max_k \|\boldsymbol{b}_{:k}\|^2\right\}. \tag{4.3.2}$$

We have written $\boldsymbol{b}_{j:}$ for the jth row of $\boldsymbol{B}$ and $\boldsymbol{b}_{:k}$ for the kth column of $\boldsymbol{B}$. In other words, the expected norm of a matrix with randomly signed entries is comparable with the maximum ℓ_2 norm achieved by any row or column. There are cases where the bound (4.3.1) admits a matching lower bound. These results appear in [Seg00, Thms. 3.1, 3.2] and [BV14, Cor. 4.7].

Theorem 4.1.1 leads to a quick proof of a slightly weaker result. We simply need to compute the matrix variance statistic $v(\boldsymbol{B}_\pm)$. To that end, note that

$$\sum_{j=1}^{d_1}\sum_{k=1}^{d_2}(b_{jk}\mathbf{E}_{jk})(b_{jk}\mathbf{E}_{jk})^* = \sum_{j=1}^{d_1}\left(\sum_{k=1}^{d_2}|b_{jk}|^2\right)\mathbf{E}_{jj} = \begin{bmatrix}\|\boldsymbol{b}_{1:}\|^2 & & \\ & \ddots & \\ & & \|\boldsymbol{b}_{d_1:}\|^2\end{bmatrix}.$$

Similarly,

$$\sum_{j=1}^{d_1}\sum_{k=1}^{d_2}(b_{jk}\mathbf{E}_{jk})^*(b_{jk}\mathbf{E}_{jk}) = \sum_{k=1}^{d_2}\left(\sum_{j=1}^{d_1}|b_{jk}|^2\right)\mathbf{E}_{kk} = \begin{bmatrix}\|\boldsymbol{b}_{:1}\|^2 & & \\ & \ddots & \\ & & \|\boldsymbol{b}_{:d_2}\|^2\end{bmatrix}.$$

Therefore, using the formula (4.1.4), we find that

$$\begin{aligned} v(\boldsymbol{B}_\pm) &= \max\left\{\left\|\sum_{j=1}^{d_1}\sum_{k=1}^{d_2}(b_{jk}\mathbf{E}_{jk})(b_{jk}\mathbf{E}_{jk})^*\right\|, \left\|\sum_{j=1}^{d_1}\sum_{k=1}^{d_2}(b_{jk}\mathbf{E}_{jk})^*(b_{jk}\mathbf{E}_{jk})\right\|\right\} \\ &= \max\left\{\max_j \|\boldsymbol{b}_{j:}\|^2, \max_k \|\boldsymbol{b}_{:k}\|^2\right\}. \end{aligned}$$

We see that $v(\boldsymbol{B}_\pm)$ coincides with v, the leading term (4.3.2) in the established estimate (4.3.1)! Now, Theorem 4.1.1 delivers the bound

$$\mathbb{E}\|\boldsymbol{B}_\pm\| \le \sqrt{2v(\boldsymbol{B}_\pm)\log(d_1+d_2)}. \tag{4.3.3}$$

Observe that the estimate (4.3.3) for the norm matches the correct bound (4.3.1) up to the logarithmic factor. Yet again, we obtain a result that is respectably close to the optimal one, even though it is not quite sharp.

The main advantage of using results like Theorem 4.1.1 to analyze this random matrix is that we can obtain a good result with a minimal amount of arithmetic. The analysis that leads to (4.3.1) involves a specialized combinatorial argument.

4.4 Example: Gaussian Toeplitz Matrices

Matrix concentration inequalities offer an effective tool for analyzing random matrices whose dependency structures are more complicated than those of the classical ensembles. In this section, we consider Gaussian Toeplitz matrices, which have applications in signal processing.

We construct an (unsymmetric) $d \times d$ Gaussian Toeplitz matrix $\mathbf{\Gamma}_d$ by populating the first row and first column of the matrix with independent standard normal variables; the entries along

each diagonal of the matrix take the same value:

$$\mathbf{\Gamma}_d = \begin{bmatrix} \gamma_0 & \gamma_1 & & \cdots & & \gamma_{d-1} \\ \gamma_{-1} & \gamma_0 & \gamma_1 & & & \\ & \gamma_{-1} & \gamma_0 & \gamma_1 & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \\ & & & \gamma_{-1} & \gamma_0 & \gamma_1 \\ \gamma_{-(d-1)} & & \cdots & & \gamma_{-1} & \gamma_0 \end{bmatrix}$$

where $\{\gamma_k\}$ is an independent family of standard normal variables. As usual, we represent the Gaussian Toeplitz matrix as a matrix Gaussian series:

$$\mathbf{\Gamma}_d = \gamma_0 \mathbf{I} + \sum_{k=1}^{d-1} \gamma_k \boldsymbol{C}^k + \sum_{k=1}^{d-1} \gamma_{-k} (\boldsymbol{C}^k)^*, \tag{4.4.1}$$

where $\boldsymbol{C} \in \mathbb{M}_d$ denotes the shift-up operator acting on d-dimensional column vectors:

$$\boldsymbol{C} = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix}.$$

It follows that $\boldsymbol{C}^k$ shifts a vector up by k places, introducing zeros at the bottom, while $(\boldsymbol{C}^k)^*$ shifts a vector down by k places, introducing zeros at the top.

We can analyze this example quickly using Theorem 4.1.1. First, note that

$$(\boldsymbol{C}^k)(\boldsymbol{C}^k)^* = \sum_{j=1}^{d-k} \mathbf{E}_{jj} \quad \text{and} \quad (\boldsymbol{C}^k)^*(\boldsymbol{C}^k) = \sum_{j=k+1}^{d} \mathbf{E}_{jj}.$$

To obtain the matrix variance statistic (4.1.4), we calculate the sum of the squares of the coefficient matrices that appear in (4.4.1). In this instance, the two terms in the variance are the same. We find that

$$\begin{aligned} \mathbf{I}^2 + \sum_{k=1}^{d-1} (\boldsymbol{C}^k)(\boldsymbol{C}^k)^* + \sum_{k=1}^{d-1} (\boldsymbol{C}^k)^*(\boldsymbol{C}^k) &= \mathbf{I} + \sum_{k=1}^{d-1} \left[\sum_{j=1}^{d-k} \mathbf{E}_{jj} + \sum_{j=k+1}^{d} \mathbf{E}_{jj} \right] \\ &= \sum_{j=1}^{d} \left[1 + \sum_{k=1}^{d-j} 1 + \sum_{k=1}^{j-1} 1 \right] \mathbf{E}_{jj} = \sum_{j=1}^{d} (1 + (d-j) + (j-1)) \mathbf{E}_{jj} = d\,\mathbf{I}_d. \end{aligned} \tag{4.4.2}$$

In the second line, we (carefully) switch the order of summation and rewrite the identity matrix as a sum of diagonal standard basis matrices. We reach

$$v(\mathbf{\Gamma}_d) = \| d\,\mathbf{I}_d \| = d.$$

An application of Theorem 4.1.1 leads us to conclude that

$$\mathbb{E} \| \mathbf{\Gamma}_d \| \le \sqrt{2d \log(2d)}. \tag{4.4.3}$$

It turns out that the inequality (4.4.3) is correct up to the precise value of the constant, which does not seem to be known. Nevertheless, the limiting value is available for the top eigenvalue of a (scaled) symmetric Toeplitz matrix whose first row contains independent standard normal variables [SV13, Thm. 1]. From this result, we may conclude that

$$0.8288 \quad \le \quad \frac{\mathbb{E}\,\|\mathbf{\Gamma}_d\|}{\sqrt{2d\log(2d)}} \quad \le \quad 1 \quad \text{as } d \to \infty.$$

Here, we take $\{\mathbf{\Gamma}_d\}$ to be a sequence of unsymmetric Gaussian Toeplitz matrices, indexed by the ambient dimension d. Our simple argument gives the right scaling for this problem, and our estimate for the constant lies within 21% of the optimal value!

4.5 Application: Rounding for the MaxQP Relaxation

Our final application involves a more substantial question from combinatorial optimization. One of the methods that has been proposed for solving a certain optimization problem leads to a matrix Rademacher series, and the analysis of this method requires the spectral norm bounds from Theorem 4.1.1. A detailed treatment would take us too far afield, so we just sketch the context and indicate how the random matrix arises.

There are many types of optimization problems that are computationally difficult to solve exactly. One approach to solving these problems is to enlarge the constraint set in such a way that the problem becomes tractable, a process called "relaxation." After solving the relaxed problem, we can use a randomized "rounding" procedure to map the solution back to the constraint set for the original problem. If we can perform the rounding step without changing the value of the objective function substantially, then the rounded solution is also a decent solution to the original optimization problem.

One difficult class of optimization problems has a matrix decision variable, and it requires us to maximize a quadratic form in the matrix variable subject to a set of convex quadratic constraints and a spectral norm constraint [Nem07]. This problem is referred to as MaxQP. The desired solution $\boldsymbol{B}$ to this problem is a $d_1 \times d_2$ matrix. The solution needs to satisfy several different requirements, but we focus on the condition that $\|\boldsymbol{B}\| \le 1$.

There is a natural relaxation of the MaxQP problem. When we solve the relaxation, we obtain a family $\{\boldsymbol{B}_k : k = 1, 2, \dots, n\}$ of $d_1 \times d_2$ matrices that satisfy the constraints

$$\sum_{k=1}^{n} \boldsymbol{B}_k \boldsymbol{B}_k^* \preccurlyeq \mathbf{I}_{d_1} \quad \text{and} \quad \sum_{k=1}^{n} \boldsymbol{B}_k^* \boldsymbol{B}_k \preccurlyeq \mathbf{I}_{d_2}. \tag{4.5.1}$$

In fact, these two bounds are part of the specification of the relaxed problem. To round the family of matrices back to a solution of the original problem, we form the random matrix

$$\boldsymbol{Z} = \alpha \sum_{k=1}^{n} \varrho_k \boldsymbol{B}_k,$$

where $\{\varrho_k\}$ is an independent family of Rademacher random variables. The scaling factor $\alpha > 0$ can be adjusted to guarantee that the norm constraint $\|\boldsymbol{Z}\| \le 1$ holds with high probability.

What is the expected norm of $\boldsymbol{Z}$? Theorem 4.1.1 yields

$$\mathbb{E}\,\|\boldsymbol{Z}\| \le \sqrt{2v(\boldsymbol{Z})\log(d_1 + d_2)}.$$

Here, the matrix variance statistic satisfies

$$v(\boldsymbol{Z}) = \alpha^2 \max\left\{ \left\| \sum_{k=1}^{n} \boldsymbol{B}_k \boldsymbol{B}_k^* \right\|, \left\| \sum_{k=1}^{n} \boldsymbol{B}_k^* \boldsymbol{B}_k \right\| \right\} \le \alpha^2,$$

owing to the constraint (4.5.1) on the matrices $\boldsymbol{B}_1, \dots, \boldsymbol{B}_n$. It follows that the scaling parameter α should satisfy

$$\alpha^2 = \frac{1}{2\log(d_1 + d_2)}$$

to ensure that $\mathbb{E}\,\|\boldsymbol{Z}\| \le 1$. For this choice of α, the rounded solution $\boldsymbol{Z}$ obeys the spectral norm constraint on average. By using the tail bound (4.1.6), we can even obtain high-probability estimates for the norm of the rounded solution $\boldsymbol{Z}$.

The important fact here is that the scaling parameter α is usually small as compared with the other parameters of the problem (d_1, d_2, n, and so forth). Therefore, the scaling does not have a massive effect on the value of the objective function. Ultimately, this approach leads to a technique for solving the MaxQP problem that produces a feasible point whose objective value is within a factor of $\sqrt{2\log(d_1 + d_2)}$ of the maximum objective value possible.

4.6 Analysis of Matrix Gaussian & Rademacher Series

We began this chapter with a concentration inequality, Theorem 4.1.1, for the norm of a matrix Gaussian series, and we have explored a number of different applications of this result. This section contains a proof of this theorem.

4.6.1 Random Series with Hermitian Coefficients

As the development in Chapter 3 suggests, random Hermitian matrices provide the natural setting for establishing matrix concentration inequalities. Therefore, we begin our treatment with a detailed statement of the matrix concentration inequality for a Gaussian series with Hermitian matrix coefficients.

Theorem 4.6.1 (Matrix Gaussian & Rademacher Series: The Hermitian Case)**.** *Consider a finite sequence $\{\boldsymbol{A}_k\}$ of fixed Hermitian matrices with dimension d, and let $\{\gamma_k\}$ be a finite sequence of independent standard normal variables. Introduce the matrix Gaussian series*

$$\boldsymbol{Y} = \textstyle\sum_k \gamma_k \boldsymbol{A}_k.$$

Let $v(\boldsymbol{Y})$ be the matrix variance statistic of the sum:

$$v(\boldsymbol{Y}) = \|\mathbb{E}\,\boldsymbol{Y}^2\| = \left\|\textstyle\sum_k \boldsymbol{A}_k^2\right\|. \tag{4.6.1}$$

Then

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \le \sqrt{2v(\boldsymbol{Y})\log d}. \tag{4.6.2}$$

Furthermore, for all $t \ge 0$,

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} \le d \exp\left(\frac{-t^2}{2v(\boldsymbol{Y})}\right). \tag{4.6.3}$$

The same bounds hold when we replace $\{\gamma_k\}$ by a finite sequence of independent Rademacher random variables.

The proof of this result occupies the rest of the section.

4.6.2 Discussion

Before we proceed to the analysis, let us take a moment to compare Theorem 4.6.1 with the result for general matrix series, Theorem 4.1.1.

First, we consider the matrix variance statistic $v(\boldsymbol{Y})$ defined in (4.6.1). Since $\boldsymbol{Y}$ has zero mean, this definition coincides with the general formula (2.2.4). The second expression, in terms of the coefficient matrices, follows from the additivity property (2.2.6) for the variance of a sum of independent, random Hermitian matrices.

Next, bounds for the minimum eigenvalue $\lambda_{\min}(\boldsymbol{Y})$ follow from the results for the maximum eigenvalue because $-\boldsymbol{Y}$ has the same distribution as $\boldsymbol{Y}$. Therefore,

$$\mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}) = \mathbb{E}\,\lambda_{\min}(-\boldsymbol{Y}) = -\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \geq -\sqrt{2v(\boldsymbol{Y})\log d}. \tag{4.6.4}$$

The second identity holds because of the relationship (2.1.5) between minimum and maximum eigenvalues. Similar considerations lead to a lower tail bound for the minimum eigenvalue:

$$\mathbb{P}\{\lambda_{\min}(\boldsymbol{Y}) \leq -t\} \leq d\,\exp\left(\frac{-t^2}{2v(\boldsymbol{Y})}\right) \quad \text{for } t \geq 0. \tag{4.6.5}$$

This result follows directly from the upper tail bound (4.6.3).

This observation points to the most important difference between the Hermitian case and the general case. Indeed, Theorem 4.6.1 concerns the extreme eigenvalues of the random series $\boldsymbol{Y}$ instead of the norm. This change amounts to producing one-sided tail bounds instead of two-sided tail bounds. For Gaussian and Rademacher series, this improvement is not really useful, but there are random Hermitian matrices whose minimum and maximum eigenvalues exhibit different types of behavior. For these problems, it can be extremely valuable to examine the two tails separately. See Chapter 5 and 6 for some results of this type.

4.6.3 Analysis for Hermitian Gaussian Series

We continue with the proof that matrix Gaussian series exhibit the behavior described in Theorem 4.6.1. Afterward, we show how to adapt the argument to address matrix Rademacher series. Our main tool is Theorem 3.6.1, the set of master bounds for independent sums. To use this result, we must identify the cgf of a fixed matrix modulated by a Gaussian random variable.

Lemma 4.6.2 (Gaussian × Matrix: Mgf and Cgf)**.** *Suppose that $\boldsymbol{A}$ is a fixed Hermitian matrix, and let γ be a standard normal random variable. Then*

$$\mathbb{E}\,\mathrm{e}^{\gamma\theta\boldsymbol{A}} = \mathrm{e}^{\theta^2\boldsymbol{A}^2/2} \quad \textit{and} \quad \log\mathbb{E}\,\mathrm{e}^{\gamma\theta\boldsymbol{A}} = \frac{\theta^2}{2}\boldsymbol{A}^2 \quad \textit{for } \theta \in \mathbb{R}.$$

Proof. We may assume $\theta = 1$ by absorbing θ into the matrix $\boldsymbol{A}$. It is well known that the moments of a standard normal variable satisfy

$$\mathbb{E}\left(\gamma^{2q+1}\right) = 0 \quad \text{and} \quad \mathbb{E}\left(\gamma^{2q}\right) = \frac{(2q)!}{2^q\,q!} \quad \text{for } q = 0, 1, 2, \ldots.$$

The formula for the odd moments holds because a standard normal variable is symmetric. One way to establish the formula for the even moments is to use integration by parts to obtain a recursion for the $(2q)$th moment in terms of the $(2q-2)$th moment.

Therefore, the matrix mgf satisfies

$$\mathbb{E}\,\mathrm{e}^{\gamma \boldsymbol{A}} = \mathbf{I} + \sum_{q=1}^{\infty} \frac{\mathbb{E}\,(\gamma^{2q})}{(2q)!}\boldsymbol{A}^{2q} = \mathbf{I} + \sum_{q=1}^{\infty} \frac{1}{q!}\big(\boldsymbol{A}^2/2\big)^q = \mathrm{e}^{\boldsymbol{A}^2/2}.$$

The first identity holds because the odd terms vanish from the series representation (2.1.15) of the matrix exponential when we take the expectation. To compute the cgf, we extract the logarithm of the mgf and recall (2.1.17), which states that the matrix logarithm is the functional inverse of the matrix exponential. □

We quickly reach results on the maximum eigenvalue of a matrix Gaussian series with Hermitian coefficients.

Proof of Theorem 4.6.1: Gaussian Case. Consider a finite sequence $\{\boldsymbol{A}_k\}$ of Hermitian matrices with dimension d, and let $\{\gamma_k\}$ be a finite sequence of independent standard normal variables. Define the matrix Gaussian series

$$\boldsymbol{Y} = \textstyle\sum_k \gamma_k \boldsymbol{A}_k.$$

We begin with the upper bound (4.6.2) for $\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y})$. The master expectation bound (3.6.1) from Theorem 3.6.1 implies that

$$\begin{aligned}
\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) &\le \inf_{\theta>0} \frac{1}{\theta} \log \operatorname{tr} \exp\Big(\textstyle\sum_k \log \mathbb{E}\,\mathrm{e}^{\gamma_k \theta \boldsymbol{A}_k}\Big) \\
&= \inf_{\theta>0} \frac{1}{\theta} \log \operatorname{tr} \exp\left(\frac{\theta^2}{2} \textstyle\sum_k \boldsymbol{A}_k^2\right) \\
&\le \inf_{\theta>0} \frac{1}{\theta} \log \left[d\, \lambda_{\max}\left(\exp\left(\frac{\theta^2}{2} \textstyle\sum_k \boldsymbol{A}_k^2\right)\right)\right] \\
&= \inf_{\theta>0} \frac{1}{\theta} \log \left[d \exp\left(\frac{\theta^2}{2} \lambda_{\max}\big(\textstyle\sum_k \boldsymbol{A}_k^2\big)\right)\right] \\
&= \inf_{\theta>0} \frac{1}{\theta} \left[\log d + \frac{\theta^2 v(\boldsymbol{Y})}{2}\right]
\end{aligned}$$

The second line follows when we introduce the cgf from Lemma 4.6.2. To reach the third inequality, we bound the trace by the dimension times the maximum eigenvalue. The fourth line is the Spectral Mapping Theorem, Proposition 2.1.3. Use the formula (4.6.1) to identify the matrix variance statistic $v(\boldsymbol{Y})$ in the exponent. The infimum is attained at $\theta = \sqrt{2v(\boldsymbol{Y})^{-1}\log d}$. This choice leads to (4.6.2).

Next, we turn to the proof of the upper tail bound (4.6.3) for $\lambda_{\max}(\boldsymbol{Y})$. Invoke the master tail bound (3.6.3) from Theorem 3.6.1, and calculate that

$$\begin{aligned}
\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} &\le \inf_{\theta>0} \mathrm{e}^{-\theta t} \operatorname{tr} \exp\Big(\textstyle\sum_k \log \mathbb{E}\,\mathrm{e}^{\gamma_k \theta \boldsymbol{A}_k}\Big) \\
&= \inf_{\theta>0} \mathrm{e}^{-\theta t} \operatorname{tr} \exp\left(\frac{\theta^2}{2} \textstyle\sum_k \boldsymbol{A}_k^2\right) \\
&\le \inf_{\theta>0} \mathrm{e}^{-\theta t} \cdot d \exp\left(\frac{\theta^2}{2} \lambda_{\max}\big(\textstyle\sum_k \boldsymbol{A}_k^2\big)\right)
\end{aligned}$$

$$= d \inf_{\theta > 0} \mathrm{e}^{-\theta t + \theta^2 v(\boldsymbol{Y})/2}.$$

The steps here are the same as in the previous calculation. The infimum is achieved at $\theta = t/v(\boldsymbol{Y})$, which yields (4.6.3). □

4.6.4 Analysis for Hermitian Rademacher Series

The inequalities for matrix Rademacher series involve arguments closely related to the proofs for matrix Gaussian series, but we require one additional piece of reasoning to obtain the simplest results. First, let us compute bounds for the matrix mgf and cgf of a Hermitian matrix modulated by a Rademacher random variable.

Lemma 4.6.3 (Rademacher × Matrix: Mgf and Cgf)**.** *Suppose that $\boldsymbol{A}$ is a fixed Hermitian matrix, and let ϱ be a Rademacher random variable. Then*

$$\mathbb{E}\,\mathrm{e}^{\varrho\theta\boldsymbol{A}} \preccurlyeq \mathrm{e}^{\theta^2\boldsymbol{A}^2/2} \quad \text{and} \quad \log \mathbb{E}\,\mathrm{e}^{\varrho\theta\boldsymbol{A}} \preccurlyeq \frac{\theta^2}{2}\boldsymbol{A}^2 \quad \text{for } \theta \in \mathbb{R}.$$

Proof. First, we establish a scalar inequality. Comparing Taylor series,

$$\cosh(a) = \sum_{q=0}^{\infty} \frac{a^{2q}}{(2q)!} \le \sum_{q=0}^{\infty} \frac{a^{2q}}{2^q q!} = \mathrm{e}^{a^2/2} \quad \text{for } a \in \mathbb{R}. \tag{4.6.6}$$

The inequality holds because $(2q)! \ge (2q)(2q-2)\cdots(4)(2) = 2^q q!$.

To compute the matrix mgf, we may assume $\theta = 1$. By direct calculation,

$$\mathbb{E}\,\mathrm{e}^{\varrho\boldsymbol{A}} = \tfrac{1}{2}\mathrm{e}^{\boldsymbol{A}} + \tfrac{1}{2}\mathrm{e}^{-\boldsymbol{A}} = \cosh(\boldsymbol{A}) \preccurlyeq \mathrm{e}^{\boldsymbol{A}^2/2}.$$

The semidefinite bound follows when we apply the Transfer Rule (2.1.14) to the inequality (4.6.6).

To determine the matrix cgf, observe that

$$\log \mathbb{E}\,\mathrm{e}^{\varrho\boldsymbol{A}} = \log\cosh(\boldsymbol{A}) \preccurlyeq \tfrac{1}{2}\boldsymbol{A}^2.$$

The semidefinite bound follows when we apply the Transfer Rule (2.1.14) to the scalar inequality $\log\cosh(a) \le a^2/2$ for $a \in \mathbb{R}$, which is a consequence of (4.6.6). □

We are prepared to develop some probability inequalities for the maximum eigenvalue of a Rademacher series with Hermitian coefficients.

Proof of Theorem 4.6.1: Rademacher Case. Consider a finite sequence $\{\boldsymbol{A}_k\}$ of Hermitian matrices, and let $\{\varrho_k\}$ be a finite sequence of independent Rademacher variables. Define the matrix Rademacher series

$$\boldsymbol{Y} = \textstyle\sum_k \varrho_k \boldsymbol{A}_k.$$

The bounds for the extreme eigenvalues of $\boldsymbol{Y}$ follow from an argument almost identical with the proof in the Gaussian case. The only point that requires justification is the inequality

$$\operatorname{tr}\exp\left(\textstyle\sum_k \log \mathbb{E}\,\mathrm{e}^{\varrho_k\theta\boldsymbol{A}_k}\right) \le \operatorname{tr}\exp\left(\frac{\theta^2}{2}\textstyle\sum_k \boldsymbol{A}_k^2\right).$$

To obtain this result, we introduce the semidefinite bound, Lemma 4.6.3, for the Rademacher cgf into the trace exponential. The left-hand side increases after this substitution because of the fact (2.1.16) that the trace exponential function is monotone with respect to the semidefinite order. □

4.6.5 Analysis of Matrix Series with Rectangular Coefficients

Finally, we consider a series with non-Hermitian matrix coefficients modulated by independent Gaussian or Rademacher random variables. The bounds for the norm of a rectangular series follow instantly from the bounds for the norm of an Hermitian series because of a formal device. We simply apply the Hermitian results to the Hermitian dilation (2.1.26) of the series.

Proof of Theorem 4.1.1. Consider a finite sequence $\{\boldsymbol{B}_k\}$ of $d_1 \times d_2$ complex matrices, and let $\{\zeta_k\}$ be a finite sequence of independent random variables, either standard normal or Rademacher.

Recall from Definition 2.1.5 that the Hermitian dilation is the map

$$\mathscr{H} : \boldsymbol{B} \longmapsto \begin{bmatrix} \mathbf{0} & \boldsymbol{B} \\ \boldsymbol{B}^* & \mathbf{0} \end{bmatrix}.$$

This leads us to form the two series

$$\boldsymbol{Z} = \sum\nolimits_k \zeta_k \boldsymbol{B}_k \quad \text{and} \quad \boldsymbol{Y} = \mathscr{H}(\boldsymbol{Z}) = \sum\nolimits_k \zeta_k \mathscr{H}(\boldsymbol{B}_k).$$

The second expression for $\boldsymbol{Y}$ holds because the Hermitian dilation is real-linear. Since we have written $\boldsymbol{Y}$ as a matrix series with Hermitian coefficients, we may analyze it using Theorem 4.6.1. We just need to express the conclusions in terms of the random matrix $\boldsymbol{Z}$.

First, we employ the fact (2.1.28) that the Hermitian dilation preserves spectral information:

$$\|\boldsymbol{Z}\| = \lambda_{\max}(\mathscr{H}(\boldsymbol{Z})) = \lambda_{\max}(\boldsymbol{Y}).$$

Therefore, bounds on $\lambda_{\max}(\boldsymbol{Y})$ deliver bounds on $\|\boldsymbol{Z}\|$. In view of the calculation (2.2.10) for the variance statistic of a dilation, we have

$$v(\boldsymbol{Y}) = v(\mathscr{H}(\boldsymbol{Z})) = v(\boldsymbol{Z}).$$

Recall that the matrix variance statistic $v(\boldsymbol{Z})$ defined in (4.1.3) coincides with the general definition from (2.2.8). Now, invoke Theorem 4.6.1 to obtain Theorem 4.1.1. □

4.7 Notes

We give an overview of research related to matrix Gaussian series, along with references for the specific random matrices that we have analyzed.

4.7.1 Matrix Gaussian and Rademacher Series

The main results, Theorem 4.1.1 and Theorem 4.6.1, have an interesting history. In the precise form presented here, these two statements first appeared in [Tro11c], but we can trace them back more than two decades.

In his work [Oli10b, Thm. 1], Oliveira established the mgf bounds presented in Lemma 4.6.2 and Lemma 4.6.3. He also developed an ingenious improvement on the arguments of Ahlswede & Winter [AW02, App.], and he obtained a bound similar with Theorem 4.6.1. The constants in Oliveira's result are worse, but the dependence on the dimension is better because it depends on the number of summands. We do not believe that the approach Ahlswede & Winter describe in [AW02] can deliver any of these results.

Recently, there have been some minor improvements to the dimensional factor that appears in Theorem 4.6.1. We discuss these results and give citations in Chapter 7.

4.7.2 The Noncommutative Khintchine Inequality

Our theory about matrix Rademacher and Gaussian series should be compared with a classic result, called the *noncommutative Khintchine inequality*, that was originally due to Lust-Piquard [LP86]; see also the follow-up work [LPP91]. In its simplest form, this inequality concerns a matrix Rademacher series with Hermitian coefficients:

$$\boldsymbol{Y} = \sum\nolimits_k \varrho_k \boldsymbol{A}_k$$

The noncommutative Khintchine inequality states that

$$\mathbb{E}\operatorname{tr}\left[\boldsymbol{Y}^{2q}\right] \le C_{2q}\operatorname{tr}\left[\left(\mathbb{E}\,\boldsymbol{Y}^2\right)^q\right] \quad \text{for } q = 1,2,3,\dots. \tag{4.7.1}$$

The minimum value of the constant $C_{2q} = (2q)!/(2^q\, q!)$ was obtained in the two papers [Buc01, Buc05]. Traditional proofs of the noncommutative Khintchine inequality are quite involved, but there is now an elementary argument available [MJC+14, Cor. 7.3].

Theorem 4.6.1 is the exponential moment analog of the polynomial moment bound (4.7.1). The polynomial moment inequality is somewhat stronger than the exponential moment inequality. Nevertheless, the exponential results are often more useful in practice. For a more thorough exploration of the relationships between Theorem 4.6.1 and noncommutative moment inequalities, such as (4.7.1), see the discussion in [Tro11c, §4].

4.7.3 Application to Random Matrices

It has also been known for a long time that results such as Theorem 4.6.1 and inequality (4.7.1) can be used to study random matrices.

We believe that the geometric functional analysis literature contains the earliest applications of matrix concentration results to analyze random matrices. In a well-known paper [Rud99], Mark Rudelson—acting on a suggestion of Gilles Pisier—showed how to use the noncommutative Khintchine inequality (4.7.1) to study covariance estimation. This work led to a significant amount of activity in which researchers used variants of Rudelson's argument to prove other types of results. See, for example, the paper [RV07]. This approach is powerful, but it tends to require some effort to use.

In parallel, other researchers in noncommutative probability theory also came to recognize the power of noncommutative moment inequalities in random matrix theory. The paper [JX08] contains a specific example. Unfortunately, this literature is technically formidable, which makes it difficult for outsiders to appreciate its achievements.

The work [AW02] of Ahlswede & Winter led to the first "packaged" matrix concentration inequalities of the type that we describe in these lecture notes. For the first few years after this work, most of the applications concerned quantum information theory and random graph theory. The paper [Gro11] introduced the method of Ahlswede & Winter to researchers in mathematical signal processing and statistics, and it served to popularize matrix concentration bounds.

At this point, the available matrix concentration inequalities were still significantly suboptimal. The main advances, in [Oli10a, Tro11c], led to optimal matrix concentration results of the kind that we present in these lecture notes. These results allow researchers to obtain reasonably accurate analyses of a wide variety of random matrices with very little effort.

4.7.4 Wigner and Marčenko–Pastur

Wigner matrices first emerged in the literature on nuclear physics, where they were used to model the Hamiltonians of reactions involving heavy atoms [Meh04, §1.1]. Wigner [Wig55] showed that the limiting spectral distribution of a certain type of Wigner matrix follows the semicircle law. See the book [Tao12, §2.4] of Tao for an overview and the book [BS10, Chap. 2] of Bai & Silverstein for a complete treatment. The Bai–Yin law [BY93] states that, up to scaling, the maximum eigenvalue of a Wigner matrix converges almost surely to two. See [Tao12, §2.3] or [BS10, Chap. 5] for more information. The analysis of the Gaussian Wigner matrix that we present here, using Theorem 4.6.1, is drawn from [Tro11c, §4].

The first rigorous work on a rectangular Gaussian matrix is due to Marčenko & Pastur [MP67], who established that the limiting distribution of the squared singular values follows a distribution that now bears their names. The Bai–Yin law [BY93] gives an almost-sure limit for the largest singular value of a rectangular Gaussian matrix. The expectation bound (4.2.5) appears in a survey article [DS02] by Davidson & Szarek. The latter result is ultimately derived from a comparison theorem for Gaussian processes due to Férnique [Fer75] and amplified by Gordon [Gor85]. Our approach, using Theorem 4.1.1, is based on [Tro11c, §4].

4.7.5 Randomly Signed Matrices

Matrices with randomly signed entries have not received much attention in the literature. The result (4.3.1) is due to Yoav Seginer [Seg00]. There is also a well-known paper [Lat05] by Rafał Latała that provides a bound for the expected norm of a Gaussian matrix whose entries have nonuniform variance. Riemer & Schütt [RS13] have extended the earlier results. The very recent paper [BV14] of Afonso Bandeira and Ramon Van Handel contains an elegant new proof of Seginer's result based on a general theorem for random matrices with independent entries. The analysis here, using Theorem 4.1.1, is drawn from [Tro11c, §4].

4.7.6 Gaussian Toeplitz Matrices

Research on random Toeplitz matrices is surprisingly recent, but there are now a number of papers available. Bryc, Dembo, & Jiang obtained the limiting spectral distribution of a symmetric Toeplitz matrix based on independent and identically distributed (iid) random variables [BDJ06]. Later, Mark Meckes established the first bound for the expected norm of a random Toeplitz matrix based on iid random variables [Mec07]. More recently, Sen & Virág computed the limiting value of the expected norm of a random, symmetric Toeplitz matrix whose entries have identical second-order statistics [SV13]. See the latter paper for additional references. The analysis here, based on Theorem 4.1.1, is new. Our lower bound for the value of $\mathbb{E}\|\mathbf{\Gamma}_d\|$ follows from the results of Sen & Virág. We are not aware of any analysis for a random Toeplitz matrix whose entries have different variances, but this type of result would follow from a simple modification of the argument in §4.4.

4.7.7 Relaxation and Rounding of MaxQP

The idea of using semidefinite relaxation and rounding to solve the MaxQP problem is due to Arkadi Nemirovski [Nem07]. He obtained nontrivial results on the performance of his method using some matrix moment calculations, but he was unable to reach the sharpest possible bound.

Anthony So [So09] pointed out that matrix moment inequalities imply an optimal result; he also showed that matrix concentration inequalities have applications to robust optimization. The presentation here, using Theorem 4.1.1, is essentially equivalent with the approach in [So09], but we have achieved slightly better bounds for the constants.

CHAPTER 5

A Sum of Random Positive-Semidefinite Matrices

This chapter presents matrix concentration inequalities that are analogous with the classical Chernoff bounds. In the matrix setting, Chernoff-type inequalities allow us to control the extreme eigenvalues of a sum of independent, random, positive-semidefinite matrices.

More formally, we consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random Hermitian matrices that satisfy

$$0 \le \lambda_{\min}(\boldsymbol{X}_k) \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le L \quad \text{for each index } k.$$

Introduce the sum $\boldsymbol{Y} = \sum_k \boldsymbol{X}_k$. Our goal is to study the expectation and tail behavior of $\lambda_{\max}(\boldsymbol{Y})$ and $\lambda_{\min}(\boldsymbol{Y})$. Bounds on the maximum eigenvalue $\lambda_{\max}(\boldsymbol{Y})$ give us information about the norm of the matrix $\boldsymbol{Y}$, a measure of how much the action of the matrix can dilate a vector. Bounds for the minimum eigenvalue $\lambda_{\min}(\boldsymbol{Y})$ tell us when the matrix $\boldsymbol{Y}$ is nonsingular; they also provide evidence about the norm of the inverse $\boldsymbol{Y}^{-1}$, when it exists.

The matrix Chernoff inequalities are quite powerful, and they have numerous applications. We demonstrate the relevance of this theory by considering two examples. First, we show how to study the norm of a random submatrix drawn from a fixed matrix, and we explain how to check when the random submatrix has full rank. Second, we develop an analysis to determine when a random graph is likely to be connected. These two problems are closely related to basic questions in statistics and in combinatorics.

In contrast, the matrix Bernstein inequalities, appearing in Chapter 6, describe how much a random matrix deviates from its mean value. As such, the matrix Bernstein bounds are more suitable than the matrix Chernoff bounds for problems that concern matrix approximations. Matrix Bernstein inequalities are also more appropriate when the variance $v(\boldsymbol{Y})$ is small in comparison with the upper bound L on the summands.

Overview

Section 5.1 presents the main results on the expectations and the tails of the extreme eigenvalues of a sum of independent, random, positive-semidefinite matrices. Section 5.2 explains how the

matrix Chernoff bounds provide spectral information about a random submatrix drawn from a fixed matrix. In §5.3, we use the matrix Chernoff bounds to study when a random graph is connected. Afterward, in §5.4 we explain how to prove the main results.

5.1 The Matrix Chernoff Inequalities

In the scalar setting, the Chernoff inequalities describe the behavior of a sum of independent, nonnegative random variables that are subject to a uniform upper bound. These results are often applied to study the number Y of successes in a sequence of independent—but not necessarily identical—Bernoulli trials with small probabilities of success. In this case, the Chernoff bounds show that Y behaves like a Poisson random variable. The random variable Y concentrates near the expected number of successes. Its lower tail has Gaussian decay, while its upper tail drops off faster than that of an exponential random variable. See [BLM13, §2.2] for more background.

In the matrix setting, we encounter similar phenomena when we consider a sum of independent, random, positive-semidefinite matrices whose eigenvalues meet a uniform upper bound. This behavior emerges from the next theorem, which closely parallels the scalar Chernoff theorem.

Theorem 5.1.1 (Matrix Chernoff)**.** *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, Hermitian matrices with common dimension d. Assume that*

$$0 \le \lambda_{\min}(\boldsymbol{X}_k) \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le L \quad \text{for each index } k.$$

Introduce the random matrix

$$\boldsymbol{Y} = \sum\nolimits_k \boldsymbol{X}_k.$$

Define the minimum eigenvalue $\mu_{\min}$ and maximum eigenvalue $\mu_{\max}$ of the expectation $\mathbb{E}\,\boldsymbol{Y}$:

$$\mu_{\min} = \lambda_{\min}(\mathbb{E}\,\boldsymbol{Y}) = \lambda_{\min}\left(\sum\nolimits_k \mathbb{E}\,\boldsymbol{X}_k\right), \quad \text{and} \tag{5.1.1}$$

$$\mu_{\max} = \lambda_{\max}(\mathbb{E}\,\boldsymbol{Y}) = \lambda_{\max}\left(\sum\nolimits_k \mathbb{E}\,\boldsymbol{X}_k\right). \tag{5.1.2}$$

Then, for $\theta > 0$,

$$\mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}) \ge \frac{1-\mathrm{e}^{-\theta}}{\theta}\,\mu_{\min} - \frac{1}{\theta}\,L\log d, \quad \text{and} \tag{5.1.3}$$

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \le \frac{\mathrm{e}^{\theta}-1}{\theta}\,\mu_{\max} + \frac{1}{\theta}\,L\log d. \tag{5.1.4}$$

Furthermore,

$$\mathbb{P}\left\{\lambda_{\min}(\boldsymbol{Y}) \le (1-\varepsilon)\mu_{\min}\right\} \le d\left[\frac{\mathrm{e}^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right]^{\mu_{\min}/L} \quad \text{for } \varepsilon \in [0,1), \quad \text{and} \tag{5.1.5}$$

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \ge (1+\varepsilon)\mu_{\max}\right\} \le d\left[\frac{\mathrm{e}^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right]^{\mu_{\max}/L} \quad \text{for } \varepsilon \ge 0. \tag{5.1.6}$$

The proof of Theorem 5.1.1 appears below in §5.4.

5.1.1 Discussion

Let us consider some facets of Theorem 5.1.1.

Aspects of the Matrix Chernoff Inequality

In many situations, it is easier to work with streamlined versions of the expectation bounds:

$$\mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}) \geq 0.63\,\mu_{\min} - L\log d, \quad \text{and} \tag{5.1.7}$$

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \leq 1.72\,\mu_{\max} + L\log d. \tag{5.1.8}$$

We obtain these results by selecting $\theta = 1$ in both (5.1.3) and (5.1.4) and evaluating the numerical constants.

These simplifications also help to clarify the meaning of Theorem 5.1.1. On average, $\lambda_{\min}(\boldsymbol{Y})$ is not much smaller than $\lambda_{\min}(\mathbb{E}\,\boldsymbol{Y})$, minus a fluctuation term that reflects the maximum size L of a summand and the ambient dimension d. Similarly, the average value of $\lambda_{\max}(\boldsymbol{Y})$ is close to $\lambda_{\max}(\mathbb{E}\,\boldsymbol{Y})$, plus the same fluctuation term.

We can also weaken the tail bounds (5.1.5) and (5.1.6) to reach

$$\mathbb{P}\left\{\lambda_{\min}(\boldsymbol{Y}) \leq t\mu_{\min}\right\} \leq d\,\mathrm{e}^{-(1-t)^2\mu_{\min}/2L} \quad \text{for } t \in [0,1), \text{ and}$$

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \geq t\mu_{\max}\right\} \leq d\left(\frac{\mathrm{e}}{t}\right)^{t\mu_{\max}/L} \quad \text{for } t \geq \mathrm{e}.$$

The first bound shows that the lower tail of $\lambda_{\min}(\boldsymbol{Y})$ decays at a subgaussian rate with variance $L/\mu_{\min}$. The second bound manifests that the upper tail of $\lambda_{\max}(\boldsymbol{Y})$ decays faster than that of an exponential random variable with mean $L/\mu_{\max}$. This is the same type of prediction we receive from the scalar Chernoff inequalities.

As with other matrix concentration results, the tail bounds (5.1.5) and (5.1.6) can overestimate the actual tail probabilities for the extreme eigenvalues of $\boldsymbol{Y}$, especially at large deviations from the mean. The value of the matrix Chernoff theorem derives from the estimates (5.1.3) and (5.1.4) for the expectation of the minimum and maximum eigenvalue of $\boldsymbol{Y}$. Scalar concentration inequalities may provide better estimates for tail probabilities.

Related Results

We can moderate the dimensional factor d in the bounds for $\lambda_{\max}(\boldsymbol{Y})$ from Theorem 5.1.1 when the random matrix $\boldsymbol{Y}$ has limited spectral content in most directions. We take up this analysis in Chapter 7.

Next, let us present an important refinement [CGT12a, Thm. A.1] of the bound (5.1.8) that can be very useful in practice:

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \leq 2\mu_{\max} + 8\mathrm{e}\left(\mathbb{E}\max_k \lambda_{\max}(\boldsymbol{X}_k)\right)\log d. \tag{5.1.9}$$

This estimate may be regarded as a matrix version of Rosenthal's inequality [Ros70]. Observe that the uniform bound L appearing in (5.1.8) always exceeds the large parenthesis on the right-hand side of (5.1.9). Therefore, the estimate (5.1.9) is valuable when the summands are unbounded and, especially, when they have heavy tails. See the notes at the end of the chapter for more information.

5.1.2 Optimality of the Matrix Chernoff Bounds

In this section, we explore how well bounds such as Theorem 5.1.1 and inequality (5.1.9) describe the behavior of a random matrix $\boldsymbol{Y}$ formed as a sum of independent, random positive-semidefinite matrices.

The Upper Chernoff Bounds

We will demonstrate that both terms in the matrix Rosenthal inequality (5.1.9) are necessary. More precisely,

$$\begin{aligned} \text{const} \cdot \big[\mu_{\max} + \mathbb{E} \max_k \lambda_{\max}(\boldsymbol{X}_k)\big] &\le \mathbb{E}\, \lambda_{\max}(\boldsymbol{Y}) \\ &\le \text{Const} \cdot \big[\mu_{\max} + \big(\mathbb{E} \max_k \lambda_{\max}(\boldsymbol{X}_k)\big) \log d\big]. \end{aligned} \tag{5.1.10}$$

Therefore, we have identified appropriate parameters for bounding $\mathbb{E}\, \lambda_{\max}(\boldsymbol{Y})$, although the constants and the logarithm may not be sharp in every case.

The appearance of $\mu_{\max}$ on the left-hand side of (5.1.10) is a consequence of Jensen's inequality. Indeed, the maximum eigenvalue is convex, so

$$\mathbb{E}\, \lambda_{\max}(\boldsymbol{Y}) \ge \lambda_{\max}(\mathbb{E}\, \boldsymbol{Y}) = \mu_{\max}.$$

To justify the other term, apply the fact that the summands $\boldsymbol{X}_k$ are positive semidefinite to conclude that

$$\mathbb{E}\, \lambda_{\max}(\boldsymbol{Y}) = \mathbb{E}\, \lambda_{\max}\big(\textstyle\sum_k \boldsymbol{X}_k\big) \ge \mathbb{E} \max_k \lambda_{\max}(\boldsymbol{X}_k).$$

We have used the fact that $\lambda_{\max}(\boldsymbol{A} + \boldsymbol{H}) \ge \lambda_{\max}(\boldsymbol{A})$ whenever $\boldsymbol{H}$ is positive semidefinite. Average the last two displays to develop the left-hand side of (5.1.10). The right-hand side of (5.1.10) is obviously just (5.1.9).

A simple example suffices to show that the logarithm cannot always be removed from the second term in (5.1.8) or from (5.1.9). For each natural number n, consider the $d \times d$ random matrix

$$\boldsymbol{Y}_n = \sum_{i=1}^{n} \sum_{k=1}^{d} \delta_{ik}^{(n)}\, \mathbf{E}_{kk}$$

where $\{\delta_{ik}^{(n)}\}$ is an independent family of $\textsc{bernoulli}(n^{-1})$ random variables and $\mathbf{E}_{kk}$ is the $d \times d$ matrix with a one in the (k, k) entry an zeros elsewhere. An easy application of (5.1.8) delivers

$$\lambda_{\max}(\boldsymbol{Y}_n) \le 1.72 + \log d.$$

Using the Poisson limit of a binomial random variable and the Skorokhod representation, we can construct an independent family $\{Q_k\}$ of $\textsc{poisson}(1)$ random variables for which

$$\boldsymbol{Y}_n \to \sum_{k=1}^{d} Q_k\, \mathbf{E}_{kk} \quad \text{almost surely as } n \to \infty.$$

It follows that

$$\mathbb{E}\, \lambda_{\max}(\boldsymbol{Y}_n) \to \mathbb{E} \max_k Q_k \approx \text{const} \cdot \frac{\log d}{\log\log d} \quad \text{as } n \to \infty.$$

Therefore, the logarithm on the second term in (5.1.8) cannot be reduced by a factor larger than the iterated logarithm $\log\log d$. This modest loss comes from approximations we make when developing the estimate for the mean. The tail bound (5.1.6) accurately predicts the order of $\lambda_{\max}(\boldsymbol{Y}_n)$ in this example.

The latter example depends on the commutativity of the summands and the infinite divisibility of the Poisson distribution, so it may seem rather special. Nevertheless, the logarithm really does belong in many (but not all!) examples that arise in practice. In particular, it is necessary in the application to random submatrices in §5.2.

The Lower Chernoff Bounds

The upper expectation bound (5.1.4) is quite satisfactory, but the situation is murkier for the lower expectation bound (5.1.3). The mean term appears naturally in the lower bound:

$$\mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}) \le \lambda_{\min}(\mathbb{E}\,\boldsymbol{Y}) = \mu_{\min}.$$

This estimate is a consequence of Jensen's inequality and the concavity of the minimum eigenvalue. On the other hand, it is not clear what the correct form of the second term in (5.1.3) should be for a general sum of random positive-semidefinite matrices.

Nevertheless, a simple example demonstrates that the lower Chernoff bound (5.1.3) is numerically sharp in some situations. Let $\boldsymbol{X}$ be a $d \times d$ random positive-semidefinite matrix that satisfies

$$\boldsymbol{X} = d\,\mathbf{E}_{ii} \quad \text{with probability } d^{-1} \text{ for each index } i = 1, \dots, d.$$

It is clear that $\mathbb{E}\,\boldsymbol{X} = \mathbf{I}_d$. Form the random matrix

$$\boldsymbol{Y}_n = \sum_{k=1}^{n} \boldsymbol{X}_k \quad \text{where each } \boldsymbol{X}_k \text{ is an independent copy of } \boldsymbol{X}.$$

The lower Chernoff bound (5.1.3) implies that

$$\mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}_n) \ge \frac{1 - \mathrm{e}^{-\theta}}{\theta} \cdot n - \frac{1}{\theta} d \log d.$$

The parameter $\theta > 0$ is at our disposal. This analysis predicts that $\mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}_n) > 0$ precisely when $n > d \log d$.

On the other hand, $\lambda_{\max}(\boldsymbol{Y}_n) > 0$ if and only if each diagonal matrix $d\,\mathbf{E}_{ii}$ appears at least once among the summands $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. To determine the probability that this event occurs, notice that this question is an instance of the coupon collector problem [MR95, §3.6]. The probability of collecting all d coupons within n draws undergoes a phase transition from about zero to about one at $n = d \log d$. By refining this argument [Tro11d], we can verify that both lower Chernoff bounds (5.1.3) and (5.1.5) provide a numerically sharp lower bound for the value of n where the phase transition occurs. In other words, the lower matrix Chernoff bounds are themselves sharp.

5.2 Example: A Random Submatrix of a Fixed Matrix

The matrix Chernoff inequality can be used to bound the extreme singular values of a random submatrix drawn from a fixed matrix. Theorem 5.1.1 might not seem suitable for this purpose because it deals with eigenvalues, but we can connect the method with the problem via a simple transformation. The results in this section have found applications in randomized linear algebra, sparse approximation, machine learning, and other fields. See the notes at the end of the chapter for some additional discussion and references.

5.2.1 A Random Column Submatrix

Let $\boldsymbol{B}$ be a fixed $d \times n$ matrix, and let $\boldsymbol{b}_{:k}$ denote the kth column of this matrix. The matrix can be expressed as the sum of its columns:

$$\boldsymbol{B} = \sum_{k=1}^{n} \boldsymbol{b}_{:k} \mathbf{e}_k^*.$$

The symbol $\mathbf{e}_k$ refers to the standard basis (column) vector with a one in the kth component and zeros elsewhere; the length of the vector $\mathbf{e}_k$ is determined by context.

We consider a simple model for a random column submatrix. Let $\{\delta_k\}$ be an independent sequence of BERNOULLI(p/n) random variables. Define the random matrix

$$\boldsymbol{Z} = \sum_{k=1}^{n} \delta_k \boldsymbol{b}_{:k} \mathbf{e}_k^*.$$

That is, we include each column independently with probability p/n, which means that there are typically about p nonzero columns in the matrix. We do not remove the other columns; we just zero them out.

In this section, we will obtain bounds on the expectation of the extreme singular values $\sigma_1(\boldsymbol{Z})$ and $\sigma_d(\boldsymbol{Z})$ of the $d \times n$ random matrix $\boldsymbol{Z}$. More precisely,

$$\begin{aligned} \mathbb{E}\,\sigma_1(\boldsymbol{Z})^2 &\le 1.72 \cdot \frac{p}{n} \cdot \sigma_1(\boldsymbol{B})^2 + (\log d) \cdot \max_k \|\boldsymbol{b}_{:k}\|^2, \quad \text{and} \\ \mathbb{E}\,\sigma_d(\boldsymbol{Z})^2 &\ge 0.63 \cdot \frac{p}{n} \cdot \sigma_d(\boldsymbol{B})^2 - (\log d) \cdot \max_k \|\boldsymbol{b}_{:k}\|^2. \end{aligned} \tag{5.2.1}$$

That is, the random submatrix $\boldsymbol{Z}$ gets its "fair share" of the squared singular values of the original matrix $\boldsymbol{B}$. There is a fluctuation term that depends on largest norm of a column of $\boldsymbol{B}$ and the logarithm of the number d of rows in $\boldsymbol{B}$. This result is very useful because a positive lower bound on $\sigma_d(\boldsymbol{Z})$ ensures that the rows of the random submatrix $\boldsymbol{Z}$ are linearly independent.

The Analysis

To study the singular values of $\boldsymbol{Z}$, it is convenient to define a $d \times d$ random, positive-semidefinite matrix

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{Z}^* = \sum_{j,k=1}^{n} \delta_j \delta_k (\boldsymbol{b}_{:j} \mathbf{e}_j^*)(\mathbf{e}_k \boldsymbol{b}_{:k}^*) = \sum_{k=1}^{n} \delta_k \boldsymbol{b}_{:k} \boldsymbol{b}_{:k}^*.$$

Note that $\delta_k^2 = \delta_k$ because δ_k only takes the values zero and one. The eigenvalues of $\boldsymbol{Y}$ determine the singular values of $\boldsymbol{Z}$, and vice versa. In particular,

$$\lambda_{\max}(\boldsymbol{Y}) = \lambda_{\max}(\boldsymbol{Z}\boldsymbol{Z}^*) = \sigma_1(\boldsymbol{Z})^2 \quad \text{and} \quad \lambda_{\min}(\boldsymbol{Y}) = \lambda_{\min}(\boldsymbol{Z}\boldsymbol{Z}^*) = \sigma_d(\boldsymbol{Z})^2,$$

where we arrange the singular values of $\boldsymbol{Z}$ in weakly decreasing order $\sigma_1(\boldsymbol{Z}) \ge \cdots \ge \sigma_d(\boldsymbol{Z})$.

The matrix Chernoff inequality provides bounds for the expectations of the eigenvalues of $\boldsymbol{Y}$. To apply the result, first calculate

$$\mathbb{E}\,\boldsymbol{Y} = \sum_{k=1}^{n} (\mathbb{E}\,\delta_k)\, \boldsymbol{b}_{:k} \boldsymbol{b}_{:k}^* = \frac{p}{n} \sum_{k=1}^{n} \boldsymbol{b}_{:k} \boldsymbol{b}_{:k}^* = \frac{p}{n} \cdot \boldsymbol{B}\boldsymbol{B}^*,$$

so that

$$\mu_{\max} = \lambda_{\max}(\mathbb{E}\,\boldsymbol{Y}) = \frac{p}{n}\sigma_1(\boldsymbol{B})^2 \quad \text{and} \quad \mu_{\min} = \lambda_{\min}(\mathbb{E}\,\boldsymbol{Y}) = \frac{p}{n}\sigma_d(\boldsymbol{B})^2.$$

Define $L = \max_k \|\boldsymbol{b}_{:k}\|^2$, and observe that $\|\delta_k \boldsymbol{b}_{:k} \boldsymbol{b}_{:k}^*\| \le L$ for each index k. The simplified matrix Chernoff bounds (5.1.7) and (5.1.8) now deliver the result (5.2.1).

5.2.2 A Random Row and Column Submatrix

Next, we consider a model for a random set of rows and columns drawn from a fixed $d \times n$ matrix $\boldsymbol{B}$. In this case, it is helpful to use matrix notation to represent the extraction of a submatrix. Define independent random projectors

$$\boldsymbol{P} = \operatorname{diag}(\delta_1, \dots, \delta_d) \quad \text{and} \quad \boldsymbol{R} = \operatorname{diag}(\xi_1, \dots, \xi_n)$$

where $\{\delta_k\}$ is an independent family of BERNOULLI(p/d) random variables and $\{\xi_k\}$ is an independent family of BERNOULLI(r/n) random variables. Then

$$\boldsymbol{Z} = \boldsymbol{PBR}$$

is a random submatrix of $\boldsymbol{B}$ with about p nonzero rows and r nonzero columns.

In this section, we will show that

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{Z}\|^2 \le 3 \cdot \frac{p}{d} \cdot \frac{r}{n} \cdot \|\boldsymbol{B}\|^2 + 2 \cdot \frac{p \log d}{d} \cdot \max_k \|\boldsymbol{b}_{:k}\|^2 \\ + 2 \cdot \frac{r \log n}{n} \cdot \max_j \|\boldsymbol{b}_{j:}\|^2 + (\log d)(\log n) \cdot \max_{j,k} |b_{jk}|^2 . \quad (5.2.2)\end{aligned}$$

The notations $\boldsymbol{b}_{j:}$ and $\boldsymbol{b}_{:k}$ refer to the jth row and kth column of the matrix $\boldsymbol{B}$, while b_{jk} is the (j,k) entry of the matrix. In other words, the random submatrix $\boldsymbol{Z}$ gets its share of the total squared norm of the matrix $\boldsymbol{B}$. The fluctuation terms reflect the maximum row norm and the maximum column norm of $\boldsymbol{B}$, as well as the size of the largest entry. There is also a weak dependence on the ambient dimensions d and n.

The Analysis

The argument has much in common with the calculations for a random column submatrix, but we need to do some extra work to handle the interaction between the random row sampling and the random column sampling.

To begin, we express the squared norm $\|\boldsymbol{Z}\|^2$ in terms of the maximum eigenvalue of a random positive-semidefinite matrix:

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{Z}\|^2 &= \mathbb{E}\,\lambda_{\max}((\boldsymbol{PBR})(\boldsymbol{PBR})^*) \\ &= \mathbb{E}\,\lambda_{\max}((\boldsymbol{PB})\boldsymbol{R}(\boldsymbol{PB})^*) = \mathbb{E}\left[\mathbb{E}\left[\lambda_{\max}\left(\sum_{k=1}^{n} \xi_k\, (\boldsymbol{PB})_{:k}(\boldsymbol{PB})_{:k}^*\right) \,\middle|\, \boldsymbol{P}\right]\right]\end{aligned}$$

We have used the fact that $\boldsymbol{RR}^* = \boldsymbol{R}$, and the notation $(\boldsymbol{PB})_{:k}$ refers to the kth column of the matrix $\boldsymbol{PB}$. Observe that the random positive-semidefinite matrix on the right-hand side has dimension d. Invoking the matrix Chernoff inequality (5.1.8), conditional on the choice of $\boldsymbol{P}$, we obtain

$$\mathbb{E}\,\|\boldsymbol{Z}\|^2 \le 1.72 \cdot \frac{r}{n} \cdot \mathbb{E}\,\lambda_{\max}((\boldsymbol{PB})(\boldsymbol{PB})^*) + (\log d) \cdot \mathbb{E} \max_k \|(\boldsymbol{PB})_{:k}\|^2 . \quad (5.2.3)$$

The required calculation is analogous with the one in the Section 5.2.1, so we omit the details. To reach a deterministic bound, we still have two more expectations to control.

Next, we examine the term in (5.2.3) that involves the maximum eigenvalue:

$$\mathbb{E}\,\lambda_{\max}((\boldsymbol{PB})(\boldsymbol{PB})^*) = \mathbb{E}\,\lambda_{\max}(\boldsymbol{B}^*\boldsymbol{PB}) = \mathbb{E}\,\lambda_{\max}\left(\sum_{j=1}^{d} \delta_j\, \boldsymbol{b}_{j:}^*\boldsymbol{b}_{j:}\right).$$

The first identity holds because $\lambda_{\max}(\boldsymbol{C}\boldsymbol{C}^*) = \lambda_{\max}(\boldsymbol{C}^*\boldsymbol{C})$ for any matrix $\boldsymbol{C}$, and $\boldsymbol{P}\boldsymbol{P}^* = \boldsymbol{P}$. Observe that the random positive-semidefinite matrix on the right-hand side has dimension n, and apply the matrix Chernoff inequality (5.1.8) again to reach

$$\mathbb{E}\,\lambda_{\max}((\boldsymbol{P}\boldsymbol{B})(\boldsymbol{P}\boldsymbol{B})^*) \le 1.72\cdot\frac{p}{d}\cdot\lambda_{\max}(\boldsymbol{B}^*\boldsymbol{B}) + (\log n)\cdot\max_j \|\boldsymbol{b}_{j:}\|^2. \tag{5.2.4}$$

Recall that $\lambda_{\max}(\boldsymbol{B}^*\boldsymbol{B}) = \|\boldsymbol{B}\|^2$ to simplify this expression slightly.

Last, we develop a bound on the maximum column norm in (5.2.3). This result also follows from the matrix Chernoff inequality, but we need to do a little work to see why. There are more direct proofs, but this approach is closer in spirit to the rest of our proof.

We are going to treat the maximum column norm as the maximum eigenvalue of a sum of independent, random diagonal matrices. Observe that

$$\|(\boldsymbol{P}\boldsymbol{B})_{:k}\|^2 = \sum_{j=1}^{d} \delta_j\,|b_{jk}|^2 \quad \text{for each } k = 1,\dots,n.$$

Using this representation, we see that

$$\begin{aligned}\max_k \|(\boldsymbol{P}\boldsymbol{B})_{:k}\|^2 &= \lambda_{\max}\begin{bmatrix}\sum_{j=1}^d \delta_j\,|b_{j1}|^2 & & \\ & \ddots & \\ & & \sum_{j=1}^d \delta_j\,|b_{jn}|^2\end{bmatrix}\\ &= \lambda_{\max}\left(\sum_{j=1}^{d}\delta_j\,\operatorname{diag}\bigl(|b_{j1}|^2,\dots,|b_{jn}|^2\bigr)\right).\end{aligned}$$

To activate the matrix Chernoff bound, we need to compute the two parameters that appear in (5.1.8). First, the uniform upper bound L satisfies

$$L = \max_j \lambda_{\max}\bigl(\operatorname{diag}\bigl(|b_{j1}|^2,\dots,|b_{jn}|^2\bigr)\bigr) = \max_j \max_k |b_{jk}|^2.$$

Second, to compute $\mu_{\max}$, note that

$$\begin{aligned}\mathbb{E}\sum_{j=1}^{d}\delta_j\operatorname{diag}\bigl(|b_{j1}|^2,\dots,|b_{jn}|^2\bigr) &= \frac{p}{d}\cdot\operatorname{diag}\left(\sum_{j=1}^{d}|b_{j1}|^2,\dots,\sum_{j=1}^{d}|b_{jn}|^2\right)\\ &= \frac{p}{d}\cdot\operatorname{diag}\bigl(\|\boldsymbol{b}_{:1}\|^2,\dots,\|\boldsymbol{b}_{:n}\|^2\bigr).\end{aligned}$$

Take the maximum eigenvalue of this expression to reach

$$\mu_{\max} = \frac{p}{d}\cdot\max_k \|\boldsymbol{b}_{:k}\|^2.$$

Therefore, the matrix Chernoff inequality implies

$$\mathbb{E}\max_k \|(\boldsymbol{P}\boldsymbol{B})_{:k}\|^2 \le 1.72\cdot\frac{p}{d}\cdot\max_k\|\boldsymbol{b}_{:k}\|^2 + (\log n)\cdot\max_{j,k}|b_{jk}|^2. \tag{5.2.5}$$

On average, the maximum squared column norm of a random submatrix $\boldsymbol{P}\boldsymbol{B}$ with approximately p nonzero rows gets its share p/d of the maximum squared column norm of $\boldsymbol{B}$, plus a fluctuation term that depends on the magnitude of the largest entry of $\boldsymbol{B}$ and the logarithm of the number n of columns.

Combine the three bounds (5.2.3), (5.2.4), and (5.2.5) to reach the result (5.2.2). We have simplified numerical constants to make the expression more compact.

5.3 Application: When is an Erdős–Rényi Graph Connected?

Random graph theory concerns probabilistic models for the interactions between pairs of objects. One basic question about a random graph is to ask whether there is a path connecting every pair of vertices or whether there are vertices segregated in different parts of the graph. It is possible to address this problem by studying the eigenvalues of random matrices, a challenge that we take up in this section.

5.3.1 Background on Graph Theory

Recall that an *undirected graph* is a pair $G = (V, E)$. The elements of the set V are called *vertices*. The set E is a collection of unordered pairs $\{u, v\}$ of distinct vertices, called *edges*. We say that the graph has an edge between vertices u and v in V if the pair $\{u, v\}$ appears in E. For simplicity, we assume that the vertex set $V = \{1, \dots, n\}$. The *degree* $\deg(k)$ of the vertex k is the number of edges in E that include the vertex k.

There are several natural matrices associated with an undirected graph. The *adjacency matrix* of the graph G is an $n \times n$ symmetric matrix $\boldsymbol{A}$ whose entries indicate which edges are present:

$$a_{jk} = \begin{cases} 1, & \{j, k\} \in E \\ 0, & \{j, k\} \notin E. \end{cases}$$

We have assumed that edges connect distinct vertices, so the diagonal entries of the matrix $\boldsymbol{A}$ equal zero. Next, define a diagonal matrix $\boldsymbol{D} = \operatorname{diag}(\deg(1), \dots, \deg(n))$ whose entries list the degrees of the vertices. The *Laplacian* $\boldsymbol{\Delta}$ and *normalized Laplacian* $\boldsymbol{H}$ of the graph are the matrices

$$\boldsymbol{\Delta} = \boldsymbol{D} - \boldsymbol{A} \quad \text{and} \quad \boldsymbol{H} = \boldsymbol{D}^{-1/2} \boldsymbol{\Delta} \boldsymbol{D}^{-1/2}.$$

We place the convention that $\boldsymbol{D}^{-1/2}(k, k) = 0$ when $\deg(k) = 0$. The Laplacian matrix $\boldsymbol{\Delta}$ is always positive semidefinite. The vector $\mathbf{e} \in \mathbb{R}^n$ of ones is always an eigenvector of $\boldsymbol{\Delta}$ with eigenvalue zero.

These matrices and their spectral properties play a dominant role in modern graph theory. For example, the graph G is connected if and only if the second-smallest eigenvalue of $\boldsymbol{\Delta}$ is strictly positive. The second smallest eigenvalue of $\boldsymbol{H}$ controls the rate at which a random walk on the graph G converges to the stationary distribution (under appropriate assumptions). See the book [GR01] for more information about these connections.

5.3.2 The Model of Erdős & Rényi

The simplest possible example of a random graph is the independent model $G(n, p)$ of Erdős and Rényi [ER60]. The number n is the number of vertices in the graph, and $p \in (0, 1)$ is the probability that two vertices are connected. More precisely, here is how to construct a random graph in $G(n, p)$. Between each pair of distinct vertices, we place an edge independently at random with probability p. In other words, the adjacency matrix takes the form

$$a_{jk} = \begin{cases} \xi_{jk}, & 1 \le j < k \le n \\ \xi_{kj}, & 1 \le k < j \le n \\ 0, & j = k. \end{cases} \tag{5.3.1}$$

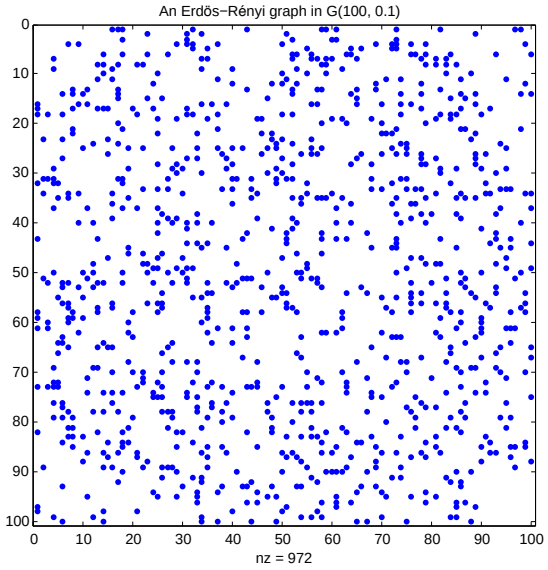

Figure 5.1: **The adjacency matrix of an Erdős–Rényi graph.** This figure shows the pattern of nonzero entries in the adjacency matrix $\boldsymbol{A}$ of a random graph drawn from $G(100, 0.1)$. Out of a possible 4,950 edges, there are 486 edges present. A basic question is whether the graph is connected. The graph is *dis*connected if and only if there is a permutation of the vertices so that the adjacency matrix is block diagonal. This property is reflected in the second-smallest eigenvalue of the Laplacian matrix $\boldsymbol{\Delta}$.

The family $\{\xi_{jk} : 1 \le j < k \le n\}$ consists of mutually independent BERNOULLI(p) random variables. Figure 5.3.2 shows one realization of the adjacency matrix of an Erdős–Rényi graph.

Let us explain how to represent the adjacency matrix and Laplacian matrix of an Erdős–Rényi graph as a sum of independent random matrices. The adjacency matrix $\boldsymbol{A}$ of a random graph in $G(n, p)$ can be written as

$$\boldsymbol{A} = \sum_{1 \le j < k \le n} \xi_{jk} (\mathbf{E}_{jk} + \mathbf{E}_{kj}). \tag{5.3.2}$$

This expression is a straightforward translation of the definition (5.3.1) into matrix form. Similarly, the Laplacian matrix $\boldsymbol{\Delta}$ of the random graph can be expressed as

$$\boldsymbol{\Delta} = \sum_{1 \le j < k \le n} \xi_{jk} (\mathbf{E}_{jj} + \mathbf{E}_{kk} - \mathbf{E}_{jk} - \mathbf{E}_{kj}). \tag{5.3.3}$$

To verify the formula (5.3.3), observe that the presence of an edge between the vertices j and k increases the degree of j and k by one. Therefore, when $\xi_{jk} = 1$, we augment the (j, j) and (k, k) entries of $\boldsymbol{\Delta}$ to reflect the change in degree, and we mark the (j, k) and (k, j) entries with -1 to reflect the presence of the edge between j and k.

5.3.3 Connectivity of an Erdős–Rényi Graph

We will obtain a near-optimal bound for the range of parameters where an Erdős–Rényi graph $G(n,p)$ is likely to be connected. We can accomplish this goal by showing that the second smallest eigenvalue of the $n \times n$ random Laplacian matrix $\mathbf{\Delta} = \boldsymbol{D} - \boldsymbol{A}$ is strictly positive. We will solve the problem by using the matrix Chernoff inequality to study the second-smallest eigenvalue of the random Laplacian $\mathbf{\Delta}$.

We need to form a random matrix $\boldsymbol{Y}$ that consists of independent positive-semidefinite terms and whose minimum eigenvalue coincides with the second-smallest eigenvalue of $\mathbf{\Delta}$. Our approach is to compress the matrix $\boldsymbol{Y}$ to the orthogonal complement of the vector $\mathbf{e}$ of ones. To that end, we introduce an $(n-1) \times n$ partial isometry $\boldsymbol{R}$ that satisfies

$$\boldsymbol{R}\boldsymbol{R}^* = \mathbf{I}_{n-1} \quad \text{and} \quad \boldsymbol{R}\mathbf{e} = \mathbf{0}. \tag{5.3.4}$$

Now, consider the $(n-1) \times (n-1)$ random matrix

$$\boldsymbol{Y} = \boldsymbol{R}\mathbf{\Delta}\boldsymbol{R}^* = \sum_{1 \le j < k \le n} \xi_{jk} \cdot \boldsymbol{R}(\mathbf{E}_{jj} + \mathbf{E}_{kk} - \mathbf{E}_{jk} - \mathbf{E}_{kj})\boldsymbol{R}^*. \tag{5.3.5}$$

Recall that $\{\xi_{jk}\}$ is an independent family of BERNOULLI(p) random variables, so the summands are mutually independent. The Conjugation Rule (2.1.12) ensures that each summand remains positive semidefinite. Furthermore, the Courant–Fischer theorem implies that the minimum eigenvalue of $\boldsymbol{Y}$ coincides with the second-smallest eigenvalue of $\mathbf{\Delta}$ because the smallest eigenvalue of $\mathbf{\Delta}$ has eigenvector $\mathbf{e}$.

To apply the matrix Chernoff inequality, we show that $L = 2$ is an upper bound for the eigenvalues of each summand in (5.3.4). We have

$$\|\xi_{jk} \cdot \boldsymbol{R}(\mathbf{E}_{jj} + \mathbf{E}_{kk} - \mathbf{E}_{jk} - \mathbf{E}_{kj})\boldsymbol{R}^*\| \le |\xi_{jk}| \cdot \|\boldsymbol{R}\| \cdot \|\mathbf{E}_{jj} + \mathbf{E}_{kk} - \mathbf{E}_{jk} - \mathbf{E}_{kj}\| \cdot \|\boldsymbol{R}^*\| \le 2.$$

The first bound follows from the submultiplicativity of the spectral norm. To obtain the second bound, note that ξ_{jk} takes 0–1 values. The matrix $\boldsymbol{R}$ is a partial isometry so its norm equals one. Finally, a direct calculation shows that $\boldsymbol{T} = \mathbf{E}_{jj} + \mathbf{E}_{kk} - \mathbf{E}_{jk} - \mathbf{E}_{kj}$ satisfies the polynomial $\boldsymbol{T}^2 = 2\boldsymbol{T}$, so each eigenvalue of $\boldsymbol{T}$ must equal zero or two.

Next, we compute the expectation of the matrix $\boldsymbol{Y}$.

$$\mathbb{E}\,\boldsymbol{Y} = p \cdot \boldsymbol{R}\left[\sum_{1 \le j < k \le n} (\mathbf{E}_{jj} + \mathbf{E}_{kk} - \mathbf{E}_{jk} - \mathbf{E}_{kj})\right]\boldsymbol{R}^* = p \cdot \boldsymbol{R}\big[(n-1)\,\mathbf{I}_n - (\mathbf{e}\mathbf{e}^* - \mathbf{I}_n)\big]\boldsymbol{R}^* = pn \cdot \mathbf{I}_{n-1}.$$

The first identity follows when we apply linearity of expectation to (5.3.5) and then use linearity of matrix multiplication to draw the sum inside the conjugation by $\boldsymbol{R}$. The term $(n-1)\,\mathbf{I}_n$ emerges when we sum the diagonal matrices. The term $\mathbf{e}\mathbf{e}^* - \mathbf{I}_n$ comes from the off-diagonal matrix units, once we note that the matrix $\mathbf{e}\mathbf{e}^*$ has one in each component. The last identity holds because of the properties of $\boldsymbol{R}$ displayed in (5.3.4). We conclude that

$$\lambda_{\min}(\mathbb{E}\,\boldsymbol{Y}) = pn.$$

This is all the information we need.

To arrive at a probability inequality for the second-smallest eigenvalue $\lambda_2^{\uparrow}(\mathbf{\Delta})$ of the matrix $\mathbf{\Delta}$, we apply the tail bound (5.1.5) to the matrix $\boldsymbol{Y}$. We obtain, for $t \in (0,1)$,

$$\mathbb{P}\left\{\lambda_2^{\uparrow}(\mathbf{\Delta}) \le t \cdot pn\right\} = \mathbb{P}\left\{\lambda_{\min}(\boldsymbol{Y}) \le t \cdot pn\right\} \le (n-1)\left[\frac{\mathrm{e}^{t-1}}{t^t}\right]^{pn/2}.$$

To appreciate what this means, we may think about the situation where $t \to 0$. Then the bracket tends to e^{-1}, and we see that the second-smallest eigenvalue of $\mathbf{\Delta}$ is unlikely to be zero when $\log(n-1) - pn/2 < 0$. Rearranging this expression, we obtain a sufficient condition

$$p > \frac{2\log(n-1)}{n}$$

for an Erdős–Rényi graph $G(n,p)$ to be connected with high probability as $n \to \infty$. This bound is quite close to the optimal result, which lacks the factor two on the right-hand side. It is possible to make this reasoning more precise, but it does not seem worth the fuss.

5.4 Proof of the Matrix Chernoff Inequalities

The first step toward the matrix Chernoff inequalities is to develop an appropriate semidefinite bound for the mgf and cgf of a random positive-semidefinite matrix. The method for establishing this result mimics the proof in the scalar case: we simply bound the exponential with a linear function.

Lemma 5.4.1 (Matrix Chernoff: Mgf and Cgf Bound)**.** *Suppose that $\boldsymbol{X}$ is a random matrix that satisfies $0 \le \lambda_{\min}(\boldsymbol{X})$ and $\lambda_{\max}(\boldsymbol{X}) \le L$. Then*

$$\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \exp\left(\frac{\mathrm{e}^{\theta L}-1}{L} \cdot \mathbb{E}\boldsymbol{X}\right) \quad \textit{and} \quad \log \mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \frac{\mathrm{e}^{\theta L}-1}{L} \cdot \mathbb{E}\boldsymbol{X} \quad \textit{for } \theta \in \mathbb{R}.$$

Proof. Consider the function $f(x) = \mathrm{e}^{\theta x}$. Since f is convex, its graph lies below the chord connecting any two points on the graph. In particular,

$$f(x) \le f(0) + \frac{f(L) - f(0)}{L} \cdot x \quad \text{for } x \in [0, L].$$

In detail,

$$\mathrm{e}^{\theta x} \le 1 + \frac{\mathrm{e}^{\theta L}-1}{L} \cdot x \quad \text{for } x \in [0, L].$$

By assumption, each eigenvalue of $\boldsymbol{X}$ lies in the interval $[0, L]$. Thus, the Transfer Rule (2.1.14) implies that

$$\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \mathbf{I} + \frac{\mathrm{e}^{\theta L}-1}{L} \cdot \boldsymbol{X}.$$

Expectation respects the semidefinite order, so

$$\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \mathbf{I} + \frac{\mathrm{e}^{\theta L}-1}{L} \cdot \mathbb{E}\boldsymbol{X} \preccurlyeq \exp\left(\frac{\mathrm{e}^{\theta L}-1}{L} \cdot \mathbb{E}\boldsymbol{X}\right).$$

The second relation is a consequence of the fact that $\mathbf{I} + \boldsymbol{A} \preccurlyeq \mathrm{e}^{\boldsymbol{A}}$ for every matrix $\boldsymbol{A}$, which we obtain by applying the Transfer Rule (2.1.14) to the inequality $1 + a \le \mathrm{e}^{a}$, valid for all $a \in \mathbb{R}$.

To obtain the semidefinite bound for the cgf, we simply take the logarithm of the semidefinite bound for the mgf. This operation preserves the semidefinite order because of the property (2.1.18) that the logarithm is operator monotone. □

We break the proof of the matrix inequality into two pieces. First, we establish the bounds on the maximum eigenvalue, which are slightly easier. Afterward, we develop the bounds on the minimum eigenvalue.

Proof of Theorem 5.1.1, Maximum Eigenvalue Bounds. Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random Hermitian matrices with common dimension d. Assume that

$$0 \le \lambda_{\min}(\boldsymbol{X}_k) \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le L \quad \text{for each index } k.$$

The cgf bound, Lemma 5.4.1, states that

$$\log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq g(\theta) \cdot \mathbb{E}\, \boldsymbol{X}_k \quad \text{where} \quad g(\theta) = \frac{\mathrm{e}^{\theta L} - 1}{L} \quad \text{for } \theta > 0. \tag{5.4.1}$$

We begin with the upper bound (5.1.4) for $\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y})$. Using the fact (2.1.16) that the trace of the exponential function is monotone with respect to the semidefinite order, we substitute these cgf bounds into the master inequality (3.6.1) for the expectation of the maximum eigenvalue to reach

$$\begin{aligned}
\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) &\le \inf_{\theta>0} \frac{1}{\theta} \log \operatorname{tr} \exp\Big(g(\theta) \textstyle\sum_k \mathbb{E}\, \boldsymbol{X}_k \Big) \\
&\le \inf_{\theta>0} \frac{1}{\theta} \log \left[d\, \lambda_{\max}\big(\exp\big(g(\theta) \cdot \mathbb{E}\, \boldsymbol{Y} \big) \big) \right] \\
&= \inf_{\theta>0} \frac{1}{\theta} \log \left[d\, \exp\big(\lambda_{\max}\big(g(\theta) \cdot \mathbb{E}\, \boldsymbol{Y} \big) \big) \right] \\
&= \inf_{\theta>0} \frac{1}{\theta} \log \left[d\, \exp\big(g(\theta) \cdot \lambda_{\max}(\mathbb{E}\, \boldsymbol{Y}) \big) \right] \\
&= \inf_{\theta>0} \frac{1}{\theta} \left[\log d + g(\theta) \cdot \mu_{\max} \right].
\end{aligned}$$

In the second line, we use the fact that the matrix exponential is positive definite to bound the trace by d times the maximum eigenvalue; we have also identified the sum as $\mathbb{E}\,\boldsymbol{Y}$. The third line follows from the Spectral Mapping Theorem, Proposition 2.1.3. Next, we use the fact (2.1.4) that the maximum eigenvalue is a positive-homogeneous map, which depends on the observation that $g(\theta) > 0$ for $\theta > 0$. Finally, we identify the statistic $\mu_{\max}$ defined in (5.1.2). The infimum does not admit a closed form, but we can obtain the expression (5.1.4) by making the change of variables $\theta \mapsto \theta / L$.

Next, we turn to the upper bound (5.1.6) for the upper tail of the maximum eigenvalue. Substitute the cgf bounds (5.4.1) into the master inequality (3.6.3) to reach

$$\begin{aligned}
\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} &\le \inf_{\theta>0} \mathrm{e}^{-\theta t} \operatorname{tr} \exp\Big(g(\theta) \textstyle\sum_k \mathbb{E}\, \boldsymbol{X}_k \Big) \\
&\le \inf_{\theta>0} \mathrm{e}^{-\theta t} \cdot d\, \exp\big(g(\theta) \cdot \mu_{\max} \big).
\end{aligned}$$

The steps here are identical with the previous argument. To complete the proof, make the change of variables $t \mapsto (1+\varepsilon)\mu_{\max}$. Then the infimum is achieved at $\theta = L^{-1} \log(1+\varepsilon)$, which leads to the tail bound (5.1.6). □

The lower bounds follow from a related argument that is slightly more delicate.

Proof of Theorem 5.1.1, Minimum Eigenvalue Bounds. Once again, consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random Hermitian matrices with dimension d. Assume that

$$0 \le \lambda_{\min}(\boldsymbol{X}_k) \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le L \quad \text{for each index } k.$$

The cgf bound, Lemma 5.4.1, states that

$$\log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq g(\theta) \cdot \mathbb{E}\, \boldsymbol{X}_k \quad \text{where} \quad g(\theta) = \frac{\mathrm{e}^{\theta L} - 1}{L} \quad \text{for } \theta < 0. \tag{5.4.2}$$

Note that $g(\theta) < 0$ for $\theta < 0$, which alters a number of the steps in the argument.

We commence with the lower bound (5.1.3) for $\mathbb{E}\,\lambda_{\min}(\boldsymbol{Y})$. As stated in (2.1.16), the trace exponential function is monotone with respect to the semidefinite order, so the master inequality (3.6.2) for the minimum eigenvalue delivers

$$\begin{aligned}
\mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}) &\ge \sup_{\theta<0} \frac{1}{\theta} \log \operatorname{tr} \exp\Big(g(\theta) \textstyle\sum_k \mathbb{E}\, \boldsymbol{X}_k\Big) \\
&\ge \sup_{\theta<0} \frac{1}{\theta} \log \left[d\, \lambda_{\max}\left(\exp\left(g(\theta) \cdot \mathbb{E}\, \boldsymbol{Y}\right)\right)\right] \\
&= \sup_{\theta<0} \frac{1}{\theta} \log \left[d\, \exp\left(\lambda_{\max}\left(g(\theta) \cdot \mathbb{E}\, \boldsymbol{Y}\right)\right)\right] \\
&= \sup_{\theta<0} \frac{1}{\theta} \log \left[d\, \exp\left(g(\theta) \cdot \lambda_{\min}(\mathbb{E}\, \boldsymbol{Y})\right)\right] \\
&= \sup_{\theta<0} \frac{1}{\theta} \left[\log d + g(\theta) \cdot \mu_{\min}\right].
\end{aligned}$$

Most of the steps are the same as in the proof of the upper bound (5.1.4), so we focus on the differences. Since the factor θ^{-1} in the first and second lines is negative, upper bounds on the trace reduce the value of the expression. We move to the fourth line by invoking the property $\lambda_{\max}(\alpha \boldsymbol{A}) = \alpha \lambda_{\min}(\boldsymbol{A})$ for $\alpha < 0$, which follows from (2.1.4) and (2.1.5). This piece of algebra depends on the fact that $g(\theta) < 0$ when $\theta < 0$. To obtain the result (5.1.3), we change variables: $\theta \mapsto -\theta / L$.

Finally, we establish the bound (5.1.5) for the lower tail of the minimum eigenvalue. Introduce the cgf bounds (5.4.2) into the master inequality (3.6.4) to reach

$$\begin{aligned}
\mathbb{P}\{\lambda_{\min}(\boldsymbol{Y}) \le t\} &\le \inf_{\theta<0} \mathrm{e}^{-\theta t} \operatorname{tr} \exp\Big(g(\theta) \textstyle\sum_k \mathbb{E}\, \boldsymbol{X}_k\Big) \\
&\le \inf_{\theta<0} \mathrm{e}^{-\theta t} \cdot d\, \exp\left(g(\theta) \cdot \mu_{\min}\right).
\end{aligned}$$

The justifications here match those in with the previous argument. Finally, we make the change of variables $t \mapsto (1-\varepsilon)\mu_{\min}$. The infimum is attained at $\theta = L^{-1} \log(1-\varepsilon)$, which yields the tail bound (5.1.5). □

5.5 Notes

As usual, we continue with an overview of background references and related work.

5.5.1 Matrix Chernoff Inequalities

Scalar Chernoff inequalities date to the paper [Che52, Thm. 1] by Herman Chernoff. The original result provides probability bounds for the number of successes in a sequence of independent but non-identical Bernoulli trials. Chernoff's proof combines the scalar Laplace transform method with refined bounds on the mgf of a Bernoulli random variable. It is very common to encounter simplified versions of Chernoff's result, such as [Lug09, Exer. 8] or [MR95, §4.1].

In their paper [AW02], Ahlswede & Winter developed a matrix version of the Chernoff inequality. The matrix mgf bound, Lemma 5.4.1, essentially appears in their work. Ahlswede & Winter focus on the case of independent and identically distributed random matrices, in which case their results are roughly equivalent with Theorem 5.1.1. For the general case, their approach leads to matrix expectation statistics of the form

$$\mu_{\min}^{\mathrm{AW}} = \sum\nolimits_k \lambda_{\min}(\mathbb{E}\, \boldsymbol{X}_k) \quad \text{and} \quad \mu_{\max}^{\mathrm{AW}} = \sum\nolimits_k \lambda_{\max}(\mathbb{E}\, \boldsymbol{X}_k).$$

It is clear that their $\mu_{\min}^{\mathrm{AW}}$ may be substantially smaller than the quantity $\mu_{\min}$ we defined in Theorem 5.1.1. Similarly, their $\mu_{\max}^{\mathrm{AW}}$ may be substantially larger than the quantity $\mu_{\max}$ that drives the upper Chernoff bounds.

The tail bounds from Theorem 5.1.1 are drawn from [Tro11c, §5], but the expectation bounds we present are new. The technical report [GT14] extends the matrix Chernoff inequality to provide upper and lower tail bounds for all eigenvalues of a sum of random, positive-semidefinite matrices. Chapter 7 contains a slight improvement of the bounds for the maximum eigenvalue in Theorem 5.1.1.

Let us mention a few other results that are related to the matrix Chernoff inequality. First, Theorem 5.1.1 has a lovely information-theoretic formulation where the tail bounds are stated in terms of an information divergence. To establish this result, we must restructure the proof and eliminate some of the approximations. See [AW02, Thm. 19] or [Tro11c, Thm. 5.1].

Second, the problem of bounding the minimum eigenvalue of a sum of random, positive-semidefinite matrices has a special character. The reason, roughly, is that a sum of independent, nonnegative random variables cannot easily take the value zero. A closely related phenomenon holds in the matrix setting, and it is possible to develop estimates that exploit this observation. See [Oli13, Thm. 3.1] and [KM13, Thm. 1.3] for two wildly different approaches.

5.5.2 The Matrix Rosenthal Inequality

The matrix Rosenthal inequality (5.1.9) is one of the earliest matrix concentration bounds. In his paper [Rud99], Rudelson used the noncommutative Khintchine inequality (4.7.1) to establish a specialization of (5.1.9) to rank-one summands. A refinement appears in [RV07], and explicit constants were first derived in [Tro08c]. We believe that the paper [CGT12a] contains the first complete statement of the moment bound (5.1.9) for general positive-semidefinite summands; see also the work [MZ11]. The constants in [CGT12a, Thm. A.1], and hence in (5.1.9), can be improved slightly by using the sharp version of the noncommutative Khintchine inequality from [Buc01, Buc05]. Let us stress that all of these results follow from easy variations of Rudelson's argument.

The work [MJC+14, Cor. 7.4] provides a self-contained and completely elementary proof of a matrix Rosenthal inequality that is closely related to (5.1.9). This result depends on different principles from the works mentioned in the last paragraph.

5.5.3 Random Submatrices

The problem of studying a random submatrix drawn from a fixed matrix has a long history. An early example is the paving problem from operator theory, which asks for a maximal well-conditioned set of columns (or a well-conditioned submatrix) inside a fixed matrix. Random selection provides a natural way to approach this question. The papers of Bourgain & Tzafriri [BT87, BT91] and Kashin & Tzafriri [KT94] study random paving using sophisticated tools from functional analysis. See the paper [NT14] for a summary of research on randomized methods for constructing pavings. Very recently, Adam Marcus, Dan Spielman, & Nikhil Srivastava [MSS14] have solved the paving problem completely.

Later, Rudelson and Vershynin [RV07] showed that the noncommutative Khintchine inequality provides a clean way to bound the norm of a random column submatrix (or a random row and column submatrix) drawn from a fixed matrix. Their ideas have found many applications in the mathematical signal processing literature. For example, the paper [Tro08a] uses similar techniques to analyze the perfomance of ℓ_1 minimization for recovering a random sparse signal. The same methods support the paper [Tro08c], which contains a modern proof of the random paving result [BT91, Thm. 2.1] of Bourgain & Tzafriri.

The article [Tro11d] contains the observation that the matrix Chernoff inequality is an ideal tool for studying random submatrices. It applies this technique to study a random matrix that arises in numerical linear algebra [HMT11], and it achieves an optimal estimate for the minimum singular value of the random matrix that arises in this setting. Our analysis of a random column submatrix is based on this work. The analysis of a random row and column submatrix is new. The paper [CD12], by Chrétien and Darses, uses matrix Chernoff bounds in a more sophisticated way to develop tail bounds for the norm of a random row and column submatrix.

5.5.4 Random Graphs

The analysis of random graphs and random hypergraphs appeared as one of the earliest applications of matrix concentration inequalities [AW02]. Christofides and Markström developed a matrix Hoeffding inequality to aid in this purpose [CM08]. Later, Oliveira wrote two papers [Oli10a, Oli11] on random graph theory based on matrix concentration. We recommend these works for further information.

To analyze the random graph Laplacian, we compressed the Laplacian to a subspace so that the minimum eigenvalue of the compression coincides with the second-smallest eigenvalue of the original Laplacian. This device can be extended to obtain tail bounds for all the eigenvalues of a sum of independent random matrices. See the technical report [GT14] for a development of this idea.

CHAPTER 6

A Sum of Bounded Random Matrices

In this chapter, we describe matrix concentration inequalities that generalize the classical Bernstein bound. The matrix Bernstein inequalities concern a random matrix formed as a sum of independent, random matrices that are bounded in spectral norm. The results allow us to study how much this type of random matrix deviates from its mean value in the spectral norm.

Formally, we consider an finite sequence $\{\boldsymbol{S}_k\}$ of random matrices of the same dimension. Assume that the matrices satisfy the conditions

$$\mathbb{E}\,\boldsymbol{S}_k = \mathbf{0} \quad \text{and} \quad \|\boldsymbol{S}_k\| \le L \quad \text{for each index } k.$$

Form the sum $\boldsymbol{Z} = \sum_k \boldsymbol{S}_k$. The matrix Bernstein inequality controls the expectation and tail behavior of $\|\boldsymbol{Z}\|$ in terms of the matrix variance statistic $v(\boldsymbol{Z})$ and the uniform bound L.

The matrix Bernstein inequality is a powerful tool with a huge number of applications. In these pages, we can only give a coarse indication of how researchers have used this result, so we have chosen to focus on problems that use random sampling to approximate a specified matrix. This model applies to the sample covariance matrix in the introduction. In this chapter, we outline several additional examples. First, we consider the technique of randomized sparsification, in which we replace a dense matrix with a sparse proxy that has similar spectral behavior. Second, we explain how to develop a randomized algorithm for approximate matrix multiplication, and we establish an error bound for this method. Third, we develop an analysis of random features, a method for approximating kernel matrices that has become popular in contemporary machine learning.

As these examples suggest, the matrix Bernstein inequality is very effective for studying randomized approximations of a given matrix. Nevertheless, when the matrix Chernoff inequality, Theorem 5.1.1, happens to apply to a problem, it often delivers better results.

Overview

Section 6.1 describes the matrix Bernstein inequality. Section 6.2 explains how to use the Bernstein inequality to study randomized methods for matrix approximation. In §§6.3, 6.4, and 6.5, we apply the latter result to three matrix approximation problems. We conclude with the proof of the matrix Bernstein inequality in §6.6.

6.1 A Sum of Bounded Random Matrices

In the scalar setting, the label "Bernstein inequality" applies to a very large number of concentration results. Most of these bounds have extensions to matrices. For simplicity, we focus on the most famous of the scalar results, a tail bound for the sum Z of independent, zero-mean random variables that are subject to a uniform bound. In this case, the Bernstein inequality shows that Z concentrates around zero. The tails of Z make a transition from subgaussian decay at moderate deviations to subexponential decay at large deviations. See [BLM13, §2.7] for more information about Bernstein's inequality.

In analogy, the simplest matrix Bernstein inequality concerns a sum of independent, zero-mean random matrices whose norms are bounded above. The theorem demonstrates that the norm of the sum acts much like the scalar random variable Z that we discussed in the last paragraph.

Theorem 6.1.1 (Matrix Bernstein)**.** *Consider a finite sequence $\{\boldsymbol{S}_k\}$ of independent, random matrices with common dimension $d_1 \times d_2$. Assume that*

$$\mathbb{E}\,\boldsymbol{S}_k = \mathbf{0} \quad \text{and} \quad \|\boldsymbol{S}_k\| \le L \quad \text{for each index } k.$$

Introduce the random matrix

$$\boldsymbol{Z} = \textstyle\sum_k \boldsymbol{S}_k.$$

Let $v(\boldsymbol{Z})$ be the matrix variance statistic of the sum:

$$v(\boldsymbol{Z}) = \max\left\{ \|\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*)\|,\ \|\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})\| \right\} \tag{6.1.1}$$

$$= \max\left\{ \left\|\textstyle\sum_k \mathbb{E}\left(\boldsymbol{S}_k\boldsymbol{S}_k^*\right)\right\|,\ \left\|\textstyle\sum_k \mathbb{E}\left(\boldsymbol{S}_k^*\boldsymbol{S}_k\right)\right\| \right\}. \tag{6.1.2}$$

Then

$$\mathbb{E}\,\|\boldsymbol{Z}\| \le \sqrt{2v(\boldsymbol{Z})\log(d_1+d_2)} + \frac{1}{3}L\log(d_1+d_2). \tag{6.1.3}$$

Furthermore, for all $t \ge 0$,

$$\mathbb{P}\{\|\boldsymbol{Z}\| \ge t\} \le (d_1+d_2)\exp\left(\frac{-t^2/2}{v(\boldsymbol{Z}) + Lt/3}\right). \tag{6.1.4}$$

The proof of Theorem 6.1.1 appears in §6.6.

6.1.1 Discussion

Let us spend a few moments to discuss the matrix Bernstein inequality, Theorem 6.1.1, its consequences, and some of the improvements that are available.

Aspects of the Matrix Bernstein Inequality

First, observe that the matrix variance statistic $v(\boldsymbol{Z})$ appearing in (6.1.1) coincides with the general definition (2.2.8) because $\boldsymbol{Z}$ has zero mean. To reach (6.1.2), we have used the additivity law (2.2.11) for an independent sum to express the matrix variance statistic in terms of the summands. Observe that, when the summands $\boldsymbol{S}_k$ are Hermitian, the two terms in the maximum coincide.

The expectation bound (6.1.3) shows that $\mathbb{E}\|\boldsymbol{Z}\|$ is on the same scale as the root $\sqrt{v(\boldsymbol{Z})}$ of the matrix variance statistic and the upper bound L for the summands; there is also a weak dependence on the ambient dimension d. In general, all three of these features are necessary. Nevertheless, the bound may not be very tight for particular examples. See Section 6.1.2 for some evidence.

Next, let us explain how to interpret the tail bound (6.1.4). The main difference between this result and the scalar Bernstein bound is the appearance of the dimensional factor $d_1 + d_2$, which reduces the range of t where the inequality is informative. To get a better idea of what this result means, it is helpful to make a further estimate:

$$\mathbb{P}\{\|\boldsymbol{Z}\| \geq t\} \leq \begin{cases} (d_1+d_2)\cdot \mathrm{e}^{-3t^2/(8v(\boldsymbol{Z}))}, & t \leq v(\boldsymbol{Z})/L \\ (d_1+d_2)\cdot \mathrm{e}^{-3t/(8L)}, & t \geq v(\boldsymbol{Z})/L. \end{cases} \tag{6.1.5}$$

In other words, for moderate values of t, the tail probability decays as fast as the tail of a Gaussian random variable whose variance is comparable with $v(\boldsymbol{Z})$. For larger values of t, the tail probability decays at least as fast as that of an exponential random variable whose mean is comparable with L. As usual, we insert a warning that the tail behavior reported by the matrix Bernstein inequality can overestimate the actual tail behavior.

Last, it is helpful to remember that the matrix Bernstein inequality extends to a sum of *uncentered* random matrices. In this case, the result describes the spectral-norm deviation of the random sum from its mean value. For reference, we include the statement here.

Corollary 6.1.2 (Matrix Bernstein: Uncentered Summands)**.** *Consider a finite sequence $\{\boldsymbol{S}_k\}$ of independent random matrices with common dimension $d_1 \times d_2$. Assume that each matrix has uniformly bounded deviation from its mean:*

$$\|\boldsymbol{S}_k - \mathbb{E}\boldsymbol{S}_k\| \leq L \quad \text{for each index } k.$$

Introduce the sum

$$\boldsymbol{Z} = \textstyle\sum_k \boldsymbol{S}_k,$$

and let $v(\boldsymbol{Z})$ denote the matrix variance statistic of the sum:

$$\begin{aligned} v(\boldsymbol{Z}) &= \max\left\{\left\|\mathbb{E}\left[(\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z})(\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z})^*\right]\right\|, \left\|\mathbb{E}\left[(\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z})^*(\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z})\right]\right\|\right\} \\ &= \max\left\{\left\|\textstyle\sum_k \mathbb{E}\left[(\boldsymbol{S}_k-\mathbb{E}\boldsymbol{S}_k)(\boldsymbol{S}_k-\mathbb{E}\boldsymbol{S}_k)^*\right]\right\|, \left\|\textstyle\sum_k \mathbb{E}\left[(\boldsymbol{S}_k-\mathbb{E}\boldsymbol{S}_k)^*(\boldsymbol{S}_k-\mathbb{E}\boldsymbol{S}_k)\right]\right\|\right\}. \end{aligned}$$

Then

$$\mathbb{E}\|\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z}\| \leq \sqrt{2v(\boldsymbol{Z})\log(d_1+d_2)} + \frac{1}{3}L\log(d_1+d_2).$$

Furthermore, for all $t \geq 0$,

$$\mathbb{P}\{\|\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z}\| \geq t\} \leq (d_1+d_2)\cdot\exp\left(\frac{-t^2/2}{v(\boldsymbol{Z})+Lt/3}\right).$$

This result follows as an immediate corollary of Theorem 6.1.1.

Related Results

The bounds in Theorem 6.1.1 are stated in terms of the ambient dimensions d_1 and d_2 of the random matrix $\boldsymbol{Z}$. The dependence on the ambient dimension is not completely natural. For example, consider embedding the random matrix $\boldsymbol{Z}$ into the top corner of a much larger matrix which is zero everywhere else. It turns out that we can achieve results that reflect only the "intrinsic dimension" of $\boldsymbol{Z}$. We turn to this analysis in Chapter 7.

In addition, there are many circumstances where the uniform upper bound L that appears in (6.1.3) does not accurately reflect the tail behavior of the random matrix. For instance, the summands themselves may have very heavy tails. In such emergencies, the following expectation bound [CGT12a, Thm. A.1] can be a lifesaver.

$$\left(\mathbb{E}\,\|\boldsymbol{Z}\|^2\right)^{1/2} \le \sqrt{2e\nu(\boldsymbol{Z})\log(d_1+d_2)} + 4e\left(\mathbb{E}\max_k \|\boldsymbol{S}_k\|^2\right)^{1/2}\log(d_1+d_2). \tag{6.1.6}$$

This result is a matrix formulation of the Rosenthal–Pinelis inequality [Pin94, Thm. 4.1].

Finally, let us reiterate that there are other types of matrix Bernstein inequalities. For example, we can sharpen the tail bound (6.1.4) to obtain a matrix Bennett inequality. We can also relax the boundedness assumption to a weaker hypothesis on the growth of the moments of each summand $\boldsymbol{S}_k$. In the Hermitian setting, the result can also discriminate the behavior of the upper and lower tails, which is a consequence of Theorem 6.6.1 below. See the notes at the end of this chapter and the annotated bibliography for more information.

6.1.2 Optimality of the Matrix Bernstein Inequality

To use the matrix Bernstein inequality, Theorem 6.1.1, and its relatives with intelligence, one must appreciate their strengths and weaknesses. We will focus on the matrix Rosenthal–Pinelis inequality (6.1.6). Nevertheless, similar insights are relevant to the estimate (6.1.3).

The Expectation Bound

Let us present lower bounds to demonstrate that the matrix Rosenthal–Pinelis inequality (6.1.6) requires both terms that appear. First, the quantity $\nu(\boldsymbol{Z})$ cannot be omitted because Jensen's inequality implies that

$$\mathbb{E}\,\|\boldsymbol{Z}\|^2 = \mathbb{E}\max\left\{\|\boldsymbol{Z}\boldsymbol{Z}^*\|,\ \|\boldsymbol{Z}^*\boldsymbol{Z}\|\right\} \ge \max\left\{\|\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*)\|,\ \|\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})\|\right\} = \nu(\boldsymbol{Z}).$$

Under a natural hypothesis, the second term on the right-hand side of (6.1.6) also is essential. Suppose that each summand $\boldsymbol{S}_k$ is a symmetric random variable; that is, $\boldsymbol{S}_k$ and $-\boldsymbol{S}_k$ have the same distribution. In this case, an involved argument [LT91, Prop. 6.10] leads to the bound

$$\mathbb{E}\,\|\boldsymbol{Z}\|^2 \ge \text{const}\cdot\mathbb{E}\max_k\|\boldsymbol{S}_k\|^2. \tag{6.1.7}$$

There are examples where the right-hand side of (6.1.7) is comparable with the uniform upper bound L on the summands, but this is not always so.

In summary, when the summands $\boldsymbol{S}_k$ are symmetric, we have matching estimates

$$\begin{aligned}\text{const}\cdot\left[\sqrt{\nu(\boldsymbol{Z})} + \left(\mathbb{E}\max_k\|\boldsymbol{S}_k\|^2\right)^{1/2}\right] &\le \left(\mathbb{E}\,\|\boldsymbol{Z}\|^2\right)^{1/2} \\ &\le \text{Const}\cdot\left[\sqrt{\nu(\boldsymbol{Z})\log(d_1+d_2)} + \left(\mathbb{E}\max_k\|\boldsymbol{S}_k\|^2\right)^{1/2}\log(d_1+d_2)\right].\end{aligned}$$

We see that the bound (6.1.6) must include some version of each term that appears, but the logarithms are not always necessary.

Examples where the Logarithms Appear

First, let us show that the variance term in (6.1.6) must contain a logarithm. For each natural number n, consider the $d \times d$ random matrix $\boldsymbol{Z}$ of the form

$$\boldsymbol{Z}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{k=1}^{d} \varrho_{ik} \mathbf{E}_{kk}$$

where $\{\varrho_{ik}\}$ is an independent family of Rademacher random variables. An easy application of the bound (6.1.6) implies that

$$\mathbb{E} \|\boldsymbol{Z}_n\| \le \text{Const} \cdot \left(\sqrt{\log(2d)} + \frac{1}{\sqrt{n}} \log(2d) \right) \to \text{Const} \cdot \sqrt{\log(2d)} \quad \text{as } n \to \infty.$$

Using the central limit theorem and the Skorokhod representation, we can construct an independent family $\{\gamma_k\}$ of standard normal random variables for which

$$\boldsymbol{Z}_n \to \sum_{k=1}^{d} \gamma_k \mathbf{E}_{kk} \quad \text{almost surely as } n \to \infty.$$

But this fact ensures that

$$\mathbb{E} \|\boldsymbol{Z}_n\| \to \mathbb{E} \left\| \sum_{k=1}^{d} \gamma_k \mathbf{E}_{kk} \right\| = \mathbb{E} \max_k |\gamma_k| \approx \sqrt{2 \log d} \quad \text{as } n \to \infty.$$

Therefore, we cannot remove the logarithm from the variance term in (6.1.6).

Next, let us justify the logarithm on the norm of the summands in (6.1.6). For each natural number n, consider a $d \times d$ random matrix $\boldsymbol{Z}$ of the form

$$\boldsymbol{Z}_n = \sum_{i=1}^{n} \sum_{k=1}^{d} \left(\delta_{ik}^{(n)} - n^{-1} \right) \mathbf{E}_{kk}$$

where $\{\delta_{ik}^{(n)}\}$ is an independent family of $\textsc{bernoulli}(n^{-1})$ random variables. The matrix Rosenthal–Pinelis inequality (6.1.6) ensures that

$$\mathbb{E} \|\boldsymbol{Z}_n\| \le \text{Const} \cdot \left(\sqrt{\log(2d)} + \log(2d) \right).$$

Using the Poisson limit of a binomial random variable and the Skorohod representation, we can construct an independent family $\{Q_k\}$ of $\textsc{poisson}(1)$ random variables for which

$$\boldsymbol{Z}_n \to \sum_{k=1}^{d} (Q_k - 1) \mathbf{E}_{kk} \quad \text{almost surely as } n \to \infty.$$

Therefore,

$$\mathbb{E} \|\boldsymbol{Z}_n\| \to \mathbb{E} \left\| \sum_{k=1}^{d} (Q_k - 1) \mathbf{E}_{kk} \right\| = \mathbb{E} \max_k |Q_k - 1| \approx \text{const} \cdot \frac{\log d}{\log \log d} \quad \text{as } n \to \infty.$$

In short, the bound we derived from (6.1.6) requires the logarithm on the second term, but it is suboptimal by a loglog factor. The upper matrix Chernoff inequality (5.1.6) correctly predicts the appearance of the iterated logarithm in this example, as does the matrix Bennett inequality.

The last two examples rely heavily on the commutativity of the summands as well as the infinite divisibility of the normal and Poisson distributions. As a consequence, it may appear that the logarithms only appear in very special contexts. In fact, many (but not all!) examples that arise in practice do require the logarithms that appear in the matrix Bernstein inequality. It is a subject of ongoing research to obtain a simple criterion for deciding when the logarithms belong.

6.2 Example: Matrix Approximation by Random Sampling

In applied mathematics, we often need to approximate a complicated target object by a more structured object. In some situations, we can solve this problem using a beautiful probabilistic approach called *empirical approximation*. The basic idea is to construct a "simple" random object whose expectation equals the target. We obtain the approximation by averaging several independent copies of the simple random object. As the number of terms in this average increases, the approximation becomes more complex, but it represents the target more faithfully. The challenge is to quantify this tradeoff.

In particular, we often encounter problems where we need to approximate a matrix by a more structured matrix. For example, we may wish to find a sparse matrix that is close to a given matrix, or we may need to construct a low-rank matrix that is close to a given matrix. Empirical approximation provides a mechanism for obtaining these approximations. The matrix Bernstein inequality offers a natural tool for assessing the quality of the randomized approximation.

This section develops a general framework for empirical approximation of matrices. Subsequent sections explain how this technique applies to specific examples from the fields of randomized linear algebra and machine learning.

6.2.1 Setup

Let $\boldsymbol{B}$ be a target matrix that we hope to approximate by a more structured matrix. To that end, let us represent the target as a sum of "simple" matrices:

$$\boldsymbol{B} = \sum_{i=1}^{N} \boldsymbol{B}_i. \tag{6.2.1}$$

The idea is to identify summands with desirable properties that we want our approximation to inherit. The examples in this chapter depend on decompositions of the form (6.2.1).

Along with the decomposition (6.2.1), we need a set of sampling probabilities:

$$\sum_{i=1}^{N} p_i = 1 \quad \text{and} \quad p_i > 0 \quad \text{for } i = 1, \dots, N. \tag{6.2.2}$$

We want to ascribe larger probabilities to "more important" summands. Quantifying what "important" means is the most difficult aspect of randomized matrix approximation. Choosing the right sampling distribution for a specific problem requires insight and ingenuity.

Given the data (6.2.1) and (6.2.2), we may construct a "simple" random matrix $\boldsymbol{R}$ by sampling:

$$\boldsymbol{R} = p_i^{-1} \boldsymbol{B}_i \quad \text{with probability } p_i. \tag{6.2.3}$$

This construction ensures that $\boldsymbol{R}$ is an unbiased estimator of the target: $\mathbb{E}\,\boldsymbol{R} = \boldsymbol{B}$. Even so, the random matrix $\boldsymbol{R}$ offers a poor approximation of the target $\boldsymbol{B}$ because it has a lot more structure.

To improve the quality of the approximation, we average n independent copies of the random matrix $\boldsymbol{R}$. We obtain an estimator of the form

$$\bar{\boldsymbol{R}}_n = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{R}_k \quad \text{where each } \boldsymbol{R}_k \text{ is an independent copy of } \boldsymbol{R}.$$

By linearity of expectation, this estimator is also unbiased: $\mathbb{E}\,\bar{\boldsymbol{R}}_n = \boldsymbol{B}$. The approximation $\bar{\boldsymbol{R}}_n$ remains structured when the number n of terms in the approximation is small as compared with the number N of terms in the decomposition (6.2.1).

Our goal is to quantify the approximation error as a function of the complexity n of the approximation:

$$\mathbb{E}\,\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\| \le \operatorname{error}(n).$$

There is a tension between the total number n of terms in the approximation and the error $\operatorname{error}(n)$ the approximation incurs. In applications, it is essential to achieve the right balance.

6.2.2 Error Estimate for Matrix Sampling Estimators

We can obtain an error estimate for the approximation scheme described in Section 6.2.1 as an immediate corollary of the matrix Bernstein inequality, Theorem 6.1.1.

Corollary 6.2.1 (Matrix Approximation by Random Sampling)**.** *Let $\boldsymbol{B}$ be a fixed $d_1 \times d_2$ matrix. Construct a $d_1 \times d_2$ random matrix $\boldsymbol{R}$ that satisfies*

$$\mathbb{E}\,\boldsymbol{R} = \boldsymbol{B} \quad \text{and} \quad \|\boldsymbol{R}\| \le L.$$

Compute the per-sample second moment:

$$m_2(\boldsymbol{R}) = \max\left\{\|\mathbb{E}(\boldsymbol{R}\boldsymbol{R}^*)\|,\ \|\mathbb{E}(\boldsymbol{R}^*\boldsymbol{R})\|\right\}. \tag{6.2.4}$$

Form the matrix sampling estimator

$$\bar{\boldsymbol{R}}_n = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{R}_k \quad \text{where each } \boldsymbol{R}_k \text{ is an independent copy of } \boldsymbol{R}.$$

Then the estimator satisfies

$$\mathbb{E}\,\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\| \le \sqrt{\frac{2m_2(\boldsymbol{R})\log(d_1+d_2)}{n}} + \frac{2L\log(d_1+d_2)}{3n}. \tag{6.2.5}$$

Furthermore, for all $t \ge 0$,

$$\mathbb{P}\left\{\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\| \ge t\right\} \le (d_1+d_2)\exp\left(\frac{-nt^2/2}{m_2(\boldsymbol{R}) + 2Lt/3}\right). \tag{6.2.6}$$

Proof. Since $\boldsymbol{R}$ is an unbiased estimator of the target matrix $\boldsymbol{B}$, we can write

$$\boldsymbol{Z} = \bar{\boldsymbol{R}}_n - \boldsymbol{B} = \frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{R}_k - \mathbb{E}\,\boldsymbol{R}) = \sum_{k=1}^{n}\boldsymbol{S}_k.$$

We have defined the summands $\boldsymbol{S}_k = n^{-1}(\boldsymbol{R}_k - \mathbb{E}\,\boldsymbol{R})$. These random matrices form an independent and identically distributed family, and each $\boldsymbol{S}_k$ has mean zero.

Now, each of the summands is subject to an upper bound:

$$\|\boldsymbol{S}_k\| \le \frac{1}{n}\left(\|\boldsymbol{R}_k\| + \|\mathbb{E}\,\boldsymbol{R}\|\right) \le \frac{1}{n}\left(\|\boldsymbol{R}_k\| + \mathbb{E}\,\|\boldsymbol{R}\|\right) \le \frac{2L}{n}.$$

The first relation is the triangle inequality; the second is Jensen's inequality. The last estimate follows from our assumption that $\|\boldsymbol{R}\| \le L$.

To control the matrix variance statistic $v(\boldsymbol{Z})$, first note that

$$v(\boldsymbol{Z}) = \max\left\{\left\|\sum_{k=1}^{n}\mathbb{E}\left(\boldsymbol{S}_k\boldsymbol{S}_k^*\right)\right\|, \left\|\sum_{k=1}^{n}\mathbb{E}\left(\boldsymbol{S}_k^*\boldsymbol{S}_k\right)\right\|\right\} = n\cdot\max\left\{\|\mathbb{E}(\boldsymbol{S}_1\boldsymbol{S}_1^*)\|,\ \|\mathbb{E}(\boldsymbol{S}_1^*\boldsymbol{S}_1)\|\right\}.$$

The first identity follows from the expression (6.1.2) for the matrix variance statistic, and the second holds because the summands $\boldsymbol{S}_k$ are identically distributed. We may calculate that

$$\begin{aligned}\boldsymbol{0} \preccurlyeq \mathbb{E}(\boldsymbol{S}_1\boldsymbol{S}_1^*) &= n^{-2}\,\mathbb{E}\left[(\boldsymbol{R}-\mathbb{E}\,\boldsymbol{R})(\boldsymbol{R}-\mathbb{E}\,\boldsymbol{R})^*\right]\\ &= n^{-2}\left[\mathbb{E}(\boldsymbol{R}\boldsymbol{R}^*) - (\mathbb{E}\,\boldsymbol{R})(\mathbb{E}\,\boldsymbol{R})^*\right] \preccurlyeq n^{-2}\,\mathbb{E}(\boldsymbol{R}\boldsymbol{R}^*).\end{aligned}$$

The first relation holds because the expectation of the random positive-semidefinite matrix $\boldsymbol{S}_1\boldsymbol{S}_1^*$ is positive semidefinite. The first identity follows from the definition of $\boldsymbol{S}_1$ and the fact that $\boldsymbol{R}_1$ has the same distribution as $\boldsymbol{R}$. The second identity is a direct calculation. The last relation holds because $(\mathbb{E}\,\boldsymbol{R})(\mathbb{E}\,\boldsymbol{R})^*$ is positive semidefinite. As a consequence,

$$\|\mathbb{E}(\boldsymbol{S}_1\boldsymbol{S}_1^*)\| \le \frac{1}{n^2}\,\|\mathbb{E}(\boldsymbol{R}\boldsymbol{R}^*)\|.$$

Likewise,

$$\|\mathbb{E}(\boldsymbol{S}_1^*\boldsymbol{S}_1)\| \le \frac{1}{n^2}\,\|\mathbb{E}(\boldsymbol{R}^*\boldsymbol{R})\|.$$

In summary,

$$v(\boldsymbol{Z}) \le \frac{1}{n}\max\left\{\|\mathbb{E}(\boldsymbol{R}\boldsymbol{R}^*)\|,\ \|\mathbb{E}(\boldsymbol{R}^*\boldsymbol{R})\|\right\} = \frac{m_2(\boldsymbol{R})}{n}.$$

The last line follows from the definition (6.2.4) of $m_2(\boldsymbol{R})$.

We are prepared to apply the matrix Bernstein inequality, Theorem 6.1.1, to the random matrix $\boldsymbol{Z} = \sum_k \boldsymbol{S}_k$. This operation results in the statement of the corollary. □

6.2.3 Discussion

One of the most common applications of the matrix Bernstein inequality is to analyze empirical matrix approximations. As a consequence, Corollary 6.2.1 is one of the most useful forms of the matrix Bernstein inequality. Let us discuss some of the important aspects of this result.

Understanding the Bound on the Approximation Error

First, let us examine how many samples n suffice to bring the approximation error bound in Corollary 6.2.1 below a specified positive tolerance ε. Examining inequality (7.3.5), we find that

$$n \ge \frac{2m_2(\boldsymbol{R})\log(d_1+d_2)}{\varepsilon^2} + \frac{2L\log(d_1+d_2)}{3\varepsilon} \quad\text{implies}\quad \mathbb{E}\,\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\| \le 2\varepsilon. \tag{6.2.7}$$

Roughly, the number n of samples should be on the scale of the per-sample second moment $m_2(\boldsymbol{R})$ and the uniform upper bound L.

The bound (6.2.7) also reveals an unfortunate aspect of empirical matrix approximation. To make the tolerance ε small, the number n of samples must increase proportional with ε^{-2}. In other words, it takes many samples to achieve a highly accurate approximation. We cannot avoid this phenomenon, which ultimately is a consequence of the central limit theorem.

On a more positive note, it is quite valuable that the error bounds (7.3.5) and (7.3.6) involve the spectral norm. This type of estimate simultaneously controls the error in every linear function of the approximation:

$$\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\| \le \varepsilon \quad \text{implies} \quad \left|\operatorname{tr}(\bar{\boldsymbol{R}}_n \boldsymbol{C}) - \operatorname{tr}(\boldsymbol{B}\boldsymbol{C})\right| \le \varepsilon \quad \text{when } \|\boldsymbol{C}\|_{S_1} \le 1.$$

The Schatten 1-norm $\|\cdot\|_{S_1}$ is defined in (2.1.29). These bounds also control the error in each singular value $\sigma_j(\bar{\boldsymbol{R}}_n)$ of the approximation:

$$\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\| \le \varepsilon \quad \text{implies} \quad \left|\sigma_j(\bar{\boldsymbol{R}}_n) - \sigma_j(\boldsymbol{B})\right| \le \varepsilon \quad \text{for each } j = 1, 2, 3, \dots, \min\{d_1, d_2\}.$$

When there is a gap between two singular values of $\boldsymbol{B}$, we can also obtain bounds for the discrepancy between the associated singular vectors of $\bar{\boldsymbol{R}}_n$ and $\boldsymbol{B}$ using perturbation theory.

To construct a good sampling estimator $\boldsymbol{R}$, we ought to control both $m_2(\boldsymbol{R})$ and L. In practice, this demands considerable creativity. This observation hints at the possibility of achieving a bias–variance tradeoff when approximating $\boldsymbol{B}$. To do so, we can drop all of the "unimportant" terms in the representation (6.2.1), i.e., those whose sampling probabilities are small. Then we construct a random approximation $\boldsymbol{R}$ only for the "important" terms that remain. Properly executed, this process may decrease both the per-sample second moment $m_2(\boldsymbol{R})$ and the upper bound L. The idea is analogous with shrinkage in statistical estimation.

A General Sampling Model

Corollary 6.2.1 extends beyond the sampling model based on the finite expansion (6.2.1). Indeed, we can consider a more general decomposition of the target matrix $\boldsymbol{B}$:

$$\boldsymbol{B} = \int_\Omega \boldsymbol{B}(\omega)\, \mathrm{d}\mu(\omega)$$

where μ is a probability measure on a sample space Ω. As before, the idea is to represent the target matrix $\boldsymbol{B}$ as an average of "simple" matrices $\boldsymbol{B}(\omega)$. The main difference is that the family of simple matrices may now be infinite. In this setting, we construct the random approximation $\boldsymbol{R}$ so that

$$\mathbb{P}\{\boldsymbol{R} \in E\} = \mu\{\omega : \boldsymbol{B}(\omega) \in E\} \quad \text{for } E \subset \mathbb{M}^{d_1 \times d_2}$$

In particular, it follows that

$$\mathbb{E}\,\boldsymbol{R} = \boldsymbol{B} \quad \text{and} \quad \|\boldsymbol{R}\| \le \sup_{\omega \in \Omega} \|\boldsymbol{B}(\omega)\|.$$

As we will discuss, this abstraction is important for applications in machine learning.

Suboptimality of Sampling Estimators

Another fundamental point about sampling estimators is that they are usually suboptimal. In other words, the matrix sampling estimator may incur an error substantially worse than the error in the best structured approximation of the target matrix.

To see why, let us consider a simple form of low-rank approximation by random sampling. The method here does not have practical value, but it highlights the reason that sampling estimators usually do not achieve ideal results. Suppose that $\boldsymbol{B}$ has singular value decomposition

$$\boldsymbol{B} = \sum_{i=1}^{N} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^* \quad \text{where} \quad \sum_{i=1}^{N} \sigma_i = 1 \quad \text{and} \quad N = \min\{d_1, d_2\}.$$

Given the SVD, we can construct a random rank-one approximation $\boldsymbol{R}$ of the form

$$\boldsymbol{R} = \boldsymbol{u}_i \boldsymbol{v}_i^* \quad \text{with probability} \quad \sigma_i.$$

Per Corollary 6.2.1, the error in the associated sampling estimator $\bar{\boldsymbol{R}}_n$ of $\boldsymbol{B}$ satisfies

$$\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\| \le \sqrt{\frac{2\log(d_1+d_2)}{n}} + \frac{2\log(d_1+d_2)}{n}$$

On the other hand, a best rank-n approximation of $\boldsymbol{B}$ takes the form $\boldsymbol{B}_n = \sum_{j=1}^{n} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^*$, and it incurs error

$$\|\boldsymbol{B}_n - \boldsymbol{B}\| = \sigma_{n+1} \le \frac{1}{n+1}.$$

The second relation is Markov's inequality, which provides an accurate estimate only when the singular values $\sigma_1, \dots, \sigma_{n+1}$ are comparable. In that case, the sampling estimator arrives within a logarithmic factor of the optimal error. But there are many matrices whose singular values decay quickly, so that $\sigma_{n+1} \ll (n+1)^{-1}$. In the latter situation, the error in the sampling estimator is much worse than the optimal error.

Warning: Frobenius-Norm Bounds

We often encounter papers that develop Frobenius-norm error bounds for matrix approximations, perhaps because the analysis is more elementary. But one must recognize that Frobenius-norm error bounds are not acceptable in most cases of practical interest:

Frobenius-norm error bounds are typically vacuous.

In particular, this phenomenon occurs in data analysis whenever we try to approximate a matrix that contains white or pink noise.

To illustrate this point, let us consider the ubiquitous problem of approximating a low-rank matrix corrupted by additive white Gaussian noise:

$$\boldsymbol{B} = \boldsymbol{x}\boldsymbol{x}^* + \alpha \boldsymbol{E} \in \mathbb{M}_d. \quad \text{where} \quad \|\boldsymbol{x}\|^2 = 1. \tag{6.2.8}$$

The desired approximation of the matrix $\boldsymbol{B}$ is the rank-one matrix $\boldsymbol{B}_{\text{opt}} = \boldsymbol{x}\boldsymbol{x}^*$. For modeling purposes, we assume that $\boldsymbol{E}$ has independent $\text{NORMAL}(0, d^{-1})$ entries. As a consequence,

$$\|\boldsymbol{E}\| \approx 2 \quad \text{and} \quad \|\boldsymbol{E}\|_{\text{F}} \approx \sqrt{d}.$$

Now, the spectral-norm error in the desired approximation satisfies

$$\|\boldsymbol{B}_{\text{opt}} - \boldsymbol{B}\| = \alpha \|\boldsymbol{E}\| \approx 2\alpha.$$

On the other hand, the Frobenius-norm error in the desired approximation satisfies

$$\|\boldsymbol{B}_{\text{opt}} - \boldsymbol{B}\|_{\text{F}} = \alpha \|\boldsymbol{E}\|_{\text{F}} \approx \alpha\sqrt{d}.$$

We see that the Frobenius-norm error can be quite large, even when we find the required approximation.

Here is another way to look at the same fact. Suppose we construct an approximation $\widehat{\boldsymbol{B}}$ of the matrix $\boldsymbol{B}$ from (6.2.8) whose Frobenius-norm error is comparable with the optimal error:

$$\|\widehat{\boldsymbol{B}} - \boldsymbol{B}\|_{\text{F}} \leq \varepsilon\sqrt{d}.$$

There is no reason for the approximation $\widehat{\boldsymbol{B}}$ to have any relationship with the desired approximation $\boldsymbol{B}_{\text{opt}}$. For example, the approximation $\widehat{\boldsymbol{B}} = \alpha\boldsymbol{E}$ satisfies this error bound with $\varepsilon = d^{-1/2}$ even though $\widehat{\boldsymbol{B}}$ consists only of noise.

6.3 Application: Randomized Sparsification of a Matrix

Many tasks in data analysis involve large, dense matrices that contain a lot of redundant information. For example, an experiment that tabulates many variables about a large number of subjects typically results in a low-rank data matrix because subjects are often similar with each other. Many questions that we pose about these data matrices can be addressed by spectral computations. In particular, factor analysis involves a singular value decomposition.

When the data matrix is approximately low rank, it has fewer degrees of freedom than its ambient dimension. Therefore, we can construct a simpler approximation that still captures most of the information in the matrix. One method for finding this approximation is to replace the dense target matrix by a sparse matrix that is close in spectral-norm distance. An elegant way to identify this sparse proxy is to randomly select a small number of entries from the original matrix to retain. This is a type of empirical approximation.

Sparsification has several potential advantages. First, it is considerably less expensive to store a sparse matrix than a dense matrix. Second, many algorithms for spectral computation operate more efficiently on sparse matrices.

In this section, we examine a very recent approach to randomized sparsification due to Kundu & Drineas [KD14]. The analysis is an immediate consequence of Corollary 6.2.1. See the notes at the end of the chapter for history and references.

6.3.1 Problem Formulation & Randomized Algorithm

Let $\boldsymbol{B}$ be a fixed $d_1 \times d_2$ complex matrix. The sparsification problem requires us to find a sparse matrix $\widehat{B}$ that has small distance from $\boldsymbol{B}$ with respect to the spectral norm. We can achieve this goal using an empirical approximation strategy.

First, let us express the target matrix as a sum of its entries:

$$\boldsymbol{B} = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} b_{ij}\, \mathbf{E}_{ij}.$$

Introduce sampling probabilities

$$p_{ij} = \frac{1}{2}\left[\frac{|b_{ij}|^2}{\|\boldsymbol{B}\|_{\mathrm{F}}^2} + \frac{|b_{ij}|}{\|\boldsymbol{B}\|_{\ell_1}}\right] \quad \text{for } i = 1,\dots,d_1 \text{ and } j = 1,\dots,d_2. \tag{6.3.1}$$

The Frobenius norm is defined in (2.1.2), and the entrywise ℓ_1 norm is defined in (2.1.30). It is easy to check that the numbers p_{ij} form a probability distribution. Let us emphasize that the non-obvious form of the distribution (6.3.1) represents a decade of research.

Now, we introduce a $d_1 \times d_2$ random matrix $\boldsymbol{R}$ that has exactly one nonzero entry:

$$\boldsymbol{R} = \frac{1}{p_{ij}} \cdot b_{ij}\mathbf{E}_{ij} \quad \text{with probability } p_{ij}.$$

We use the convention that $0/0 = 0$ so that we do not need to treat zero entries separately. It is immediate that

$$\mathbb{E}\,\boldsymbol{R} = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} \frac{1}{p_{ij}} \cdot b_{ij}\mathbf{E}_{ij} \cdot p_{ij} = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} b_{ij}\mathbf{E}_{ij} = \boldsymbol{B}.$$

Therefore, $\boldsymbol{R}$ is an unbiased estimate of $\boldsymbol{B}$.

Although the expectation of $\boldsymbol{R}$ is correct, its variance is quite high. Indeed, $\boldsymbol{R}$ has only one nonzero entry, while $\boldsymbol{B}$ typically has many nonzero entries. To reduce the variance, we combine several independent copies of the simple estimator:

$$\bar{\boldsymbol{R}}_n = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{R}_k \quad \text{where each } \boldsymbol{R}_k \text{ is an independent copy of } \boldsymbol{R}.$$

By linearity of expectation, $\mathbb{E}\,\bar{\boldsymbol{R}}_n = \boldsymbol{B}$. Therefore, the matrix $\bar{\boldsymbol{R}}_n$ has at most n nonzero entries, and its also provides an unbiased estimate of the target. The challenge is to quantify the error $\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\|$ as a function of the sparsity level n.

6.3.2 Performance of Randomized Sparsification

The randomized sparsification method is clearly a type of empirical approximation, so we can use Corollary 6.2.1 to perform the analysis. We will establish the following error bound.

$$\mathbb{E}\,\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\| \le \sqrt{\frac{4\,\|\boldsymbol{B}\|_{\mathrm{F}}^2 \cdot \max\{d_1,d_2\}\log(d_1+d_2)}{n}} + \frac{4\,\|\boldsymbol{B}\|_{\ell_1}\log(d_1+d_2)}{3n}. \tag{6.3.2}$$

The short proof of (6.3.2) appears below in Section 6.3.3.

Let us explain the content of the estimate (6.3.2). First, the bound (2.1.31) allows us to replace the ℓ_1 norm by the Frobenius norm:

$$\|\boldsymbol{B}\|_{\ell_1} \le \sqrt{d_1 d_2} \cdot \|\boldsymbol{B}\|_{\mathrm{F}} \le \max\{d_1,\, d_2\} \cdot \|\boldsymbol{B}\|_{\mathrm{F}}.$$

Placing the error (6.3.2) on a relative scale, we see that

$$\frac{\mathbb{E}\,\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\|}{\|\boldsymbol{B}\|} \le \frac{\|\boldsymbol{B}\|_{\mathrm{F}}}{\|\boldsymbol{B}\|} \cdot \left[\sqrt{\frac{4\max\{d_1,\, d_2\}\log(d_1+d_2)}{n}} + \frac{4\max\{d_1,\, d_2\}\log(d_1+d_2)}{3n}\right]$$

The stable rank srank($\boldsymbol{B}$), defined in (2.1.25), emerges naturally as a quantity of interest.

Now, suppose that the sparsity level n satisfies

$$n \geq \varepsilon^{-2} \cdot \operatorname{srank}(\boldsymbol{B}) \cdot \max\{d_1,\, d_2\} \log(d_1 + d_2)$$

where the tolerance $\varepsilon \in (0, 1]$. We determine that

$$\frac{\mathbb{E}\,\|\tilde{\boldsymbol{R}}_n - \boldsymbol{B}\|}{\|\boldsymbol{B}\|} \leq 2\varepsilon + \frac{4}{3} \cdot \frac{\varepsilon^2}{\sqrt{\operatorname{srank}(\boldsymbol{B})}}.$$

Since the stable rank always exceeds one and we have assumed that $\varepsilon \leq 1$, this estimate implies that

$$\frac{\mathbb{E}\,\|\tilde{\boldsymbol{R}}_n - \boldsymbol{B}\|}{\|\boldsymbol{B}\|} \leq 4\varepsilon.$$

We discover that it is possible to replace the matrix $\boldsymbol{B}$ by a matrix with at most n nonzero entries while achieving a small relative error in the spectral norm. When $\operatorname{srank}(\boldsymbol{B}) \ll \min\{d_1,\, d_2\}$, we can achieve a dramatic reduction in the number of nonzero entries needed to carry the spectral information in the matrix $\boldsymbol{B}$.

6.3.3 Analysis of Randomized Sparsification

Let us proceed with the analysis of randomized sparsification. To apply Corollary 6.2.1, we need to obtain bounds for the per-sample variance $m_2(\boldsymbol{R})$ and the uniform upper bound L. The key to both calculations is to obtain appropriate *lower* bounds on the sampling probabilities p_{ij}. Indeed,

$$p_{ij} \geq \frac{1}{2} \cdot \frac{|b_{ij}|}{\|\boldsymbol{B}\|_{\ell_1}} \quad \text{and} \quad p_{ij} \geq \frac{1}{2} \cdot \frac{|b_{ij}|^2}{\|\boldsymbol{B}\|_{\mathrm{F}}^2}. \tag{6.3.3}$$

Each estimate follows by neglecting one term in (6.3.3).

First, we turn to the uniform bound on the random matrix $\boldsymbol{R}$. We have

$$\|\boldsymbol{R}\| \leq \max_{ij} \|p_{ij}^{-1} b_{ij} \mathbf{E}_{ij}\| = \max_{ij} \frac{1}{p_{ij}} \cdot |b_{ij}| \leq 2\,\|\boldsymbol{B}\|_{\ell_1}.$$

The last inequality depends on the first bound in (6.3.3). Therefore, we may take $L = 2\,\|\boldsymbol{B}\|_{\ell_1}$.

Second, we turn to the computation of the per-sample second moment $m_2(\boldsymbol{R})$. We have

$$\begin{aligned}
\mathbb{E}(\boldsymbol{R}\boldsymbol{R}^*) &= \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{1}{p_{ij}^2} \cdot (b_{ij}\mathbf{E}_{ij})(b_{ij}\mathbf{E}_{ij})^* p_{ij} \\
&= \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{|b_{ij}|^2}{p_{ij}} \cdot \mathbf{E}_{ii} \\
&\preccurlyeq 2\,\|\boldsymbol{B}\|_{\mathrm{F}}^2 \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \mathbf{E}_{ii} = 2d_2\,\|\boldsymbol{B}\|_{\mathrm{F}}^2 \cdot \mathbf{I}_{d_1}.
\end{aligned}$$

The semidefinite inequality holds because each matrix $|b_{ij}|^2 \mathbf{E}_{ii}$ is positive semidefinite and because of the second bound in (6.3.3). Similarly,

$$\mathbb{E}(\boldsymbol{R}^*\boldsymbol{R}) \preccurlyeq 2d_1\,\|\boldsymbol{B}\|_{\mathrm{F}}^2 \cdot \mathbf{I}_{d_2}.$$

In summary,

$$m_2(\boldsymbol{R}) = \max\left\{ \|\mathbb{E}(\boldsymbol{R}\boldsymbol{R}^*)\|, \|\mathbb{E}(\boldsymbol{R}^*\boldsymbol{R})\| \right\} \le 2\max\{d_1, d_2\}.$$

This is the required estimate for the per-sample second moment.

Finally, to reach the advertised error bound (6.3.2), we invoke Corollary 6.2.1 with the parameters $L = \|\boldsymbol{B}\|_{\ell_1}$ and $m_2(\boldsymbol{R}) \le 2\max\{d_1, d_2\}$.

6.4 Application: Randomized Matrix Multiplication

Numerical linear algebra (NLA) is a well-established and important part of computer science. Some of the basic problems in this area include multiplying matrices, solving linear systems, computing eigenvalues and eigenvectors, and solving linear least-squares problems. Historically, the NLA community has focused on developing highly accurate deterministic methods that require as few floating-point operations as possible. Unfortunately, contemporary applications can strain standard NLA methods because problems have continued to become larger. Furthermore, on modern computer architectures, computational costs depend heavily on communication and other resources that the standard algorithms do not manage very well.

In response to these challenges, researchers have started to develop randomized algorithms for core problems in NLA. In contrast to the classical algorithms, these new methods make random choices during execution to achieve computational efficiencies. These randomized algorithms can also be useful for large problems or for modern computer architectures. On the other hand, randomized methods can fail with some probability, and in some cases they are less accurate than their classical competitors.

Matrix concentration inequalities are one of the key tools used to design and analyze randomized algorithms for NLA problems. In this section, we will describe a randomized method for matrix multiplication developed by Magen & Zouzias [MZ11, Zou13]. We will analyze this algorithm using Corollary 6.2.1. Turn to the notes at the end of the chapter for more information about the history.

6.4.1 Problem Formulation & Randomized Algorithm

One of the basic tasks in numerical linear algebra is to multiply two matrices with compatible dimensions. Suppose that $\boldsymbol{B}$ is a $d_1 \times N$ complex matrix and that $\boldsymbol{C}$ is an $N \times d_2$ complex matrix, and we wish to compute the product $\boldsymbol{B}\boldsymbol{C}$. The straightforward algorithm forms the product entry by entry:

$$(\boldsymbol{B}\boldsymbol{C})_{ik} = \sum_{j=1}^{N} b_{ij} c_{jk} \quad \text{for each } i = 1, \dots, d_1 \text{ and } k = 1, \dots, d_2. \tag{6.4.1}$$

This approach takes $\mathrm{O}(N \cdot d_1 d_2)$ arithmetic operations. There are algorithms, such as Strassen's divide-and-conquer method, that can reduce the cost, but these approaches are not considered practical for most applications.

Suppose that the inner dimension N is substantially larger than the outer dimensions d_1 and d_2. In this setting, both matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ are rank-deficient, so the columns of $\boldsymbol{B}$ contain a lot of linear dependencies, as do the rows of $\boldsymbol{C}$. As a consequence, a random sample of columns from $\boldsymbol{B}$ (or rows from $\boldsymbol{C}$) can be used as a proxy for the full matrix. Formally, the key to this approach

is to view the matrix product as a sum of outer products:

$$\boldsymbol{BC} = \sum_{j=1}^{N} \boldsymbol{b}_{:j}\boldsymbol{c}_{j:}. \tag{6.4.2}$$

As usual, $\boldsymbol{b}_{:j}$ denotes the jth column of $\boldsymbol{B}$, while $\boldsymbol{c}_{j:}$ denotes the jth row of $\boldsymbol{C}$. We can approximate this sum using the empirical method.

To develop an algorithm, the first step is to construct a simple random matrix $\boldsymbol{R}$ that provides an unbiased estimate for the matrix product. To that end, we pick a random index and form a rank-one matrix from the associated columns of $\boldsymbol{B}$ and row of $\boldsymbol{C}$. More precisely, define

$$p_j = \frac{\|\boldsymbol{b}_{:j}\|^2 + \|\boldsymbol{c}_{j:}\|^2}{\|\boldsymbol{B}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2} \quad \text{for } j = 1, 2, 3, \dots, N. \tag{6.4.3}$$

The Frobenius norm is defined in (2.1.2). Using the properties of the norms, we can easily check that $(p_1, p_2, p_3, \dots, p_N)$ forms a bonafide probability distribution. The cost of computing these probabilities is at most $O(N \cdot (d_1 + d_2))$ arithmetic operations, which is much smaller than the cost of forming the product $\boldsymbol{BC}$ when d_1 and d_2 are large.

We now define a $d_1 \times d_2$ random matrix $\boldsymbol{R}$ by the expression

$$\boldsymbol{R} = \frac{1}{p_j} \cdot \boldsymbol{b}_{:j}\boldsymbol{c}_{j:} \quad \text{with probability } p_j.$$

We use the convention that $0/0 = 0$ so we do not have to treat zero rows and columns separately. It is straightforward to compute the expectation of $\boldsymbol{R}$:

$$\mathbb{E}\,\boldsymbol{R} = \sum_{j=1}^{N} \frac{1}{p_j} \cdot \boldsymbol{b}_{:j}\boldsymbol{c}_{j:} \cdot p_j = \sum_{j=1}^{N} \boldsymbol{b}_{:j}\boldsymbol{c}_{j:} = \boldsymbol{BC}.$$

As required, $\boldsymbol{R}$ is an unbiased estimator for the product $\boldsymbol{BC}$.

Although the expectation of $\boldsymbol{R}$ is correct, its variance is quite high. Indeed, $\boldsymbol{R}$ has rank one, while the rank of $\boldsymbol{BC}$ is usually larger! To reduce the variance, we combine several independent copies of the simple estimator:

$$\bar{\boldsymbol{R}}_n = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{R}_k \quad \text{where} \quad \text{each } \boldsymbol{R}_k \text{ is an independent copy of } \boldsymbol{R}. \tag{6.4.4}$$

By linearity of expectation, $\mathbb{E}\,\bar{\boldsymbol{R}}_n = \boldsymbol{BC}$, so we imagine that $\bar{\boldsymbol{R}}_n$ approximates the product well.

To see whether this heuristic holds true, we need to understand how the error $\mathbb{E}\,\|\bar{\boldsymbol{R}}_n - \boldsymbol{BC}\|$ depends on the number n of samples. It costs $O(n \cdot d_1 d_2)$ floating-point operations to determine all the entries of $\bar{\boldsymbol{R}}_n$. Therefore, when the number n of samples is much smaller than the inner dimension N of the matrices, we can achieve significant economies over the naïve matrix multiplication algorithm.

In fact, it requires no computation beyond sampling the row/column indices to express $\bar{\boldsymbol{R}}_n$ in the form (6.4.4). This approach gives an inexpensive way to represent the product approximately.

6.4.2 Performance of Randomized Matrix Multiplication

To simplify our presentation, we will assume that both matrices have been scaled so that their spectral norms are equal to one:

$$\|\boldsymbol{B}\| = \|\boldsymbol{C}\| = 1.$$

It is relatively inexpensive to compute the spectral norm of a matrix accurately, so this preprocessing step is reasonable.

Let $\mathsf{asr} = \frac{1}{2}(\mathsf{srank}(\boldsymbol{B}) + \mathsf{srank}(\boldsymbol{C}))$ be the average stable rank of the two factors; see (2.1.25) for the definition of the stable rank. In §6.4.3, we will prove that

$$\mathbb{E}\,\|\bar{\boldsymbol{R}}_n - \boldsymbol{BC}\| \le \sqrt{\frac{4 \cdot \mathsf{asr} \cdot \log(d_1 + d_2)}{n}} + \frac{2 \cdot \mathsf{asr} \cdot \log(d_1 + d_2)}{3n}. \tag{6.4.5}$$

To appreciate what this estimate means, suppose that the number n of samples satisfies

$$n \ge \varepsilon^{-2} \cdot \mathsf{asr} \cdot \log(d_1 + d_2)$$

where ε is a positive tolerance. Then we obtain a relative error bound for the randomized matrix multiplication method

$$\frac{\mathbb{E}\,\|\bar{\boldsymbol{R}}_n - \boldsymbol{BC}\|}{\|\boldsymbol{B}\|\,\|\boldsymbol{C}\|} \le 2\varepsilon + \frac{2}{3}\varepsilon^2.$$

This expression depends on the normalization of $\boldsymbol{B}$ and $\boldsymbol{C}$. The computational cost of forming the approximation is

$$O(\varepsilon^{-2} \cdot \mathsf{asr} \cdot d_1 d_2 \log(d_1 + d_2)) \quad \text{arithmetic operations.}$$

In other words, when the average stable rank asr is substantially smaller than the inner dimension N of the two matrices $\boldsymbol{B}$ and $\boldsymbol{C}$, the random estimate $\bar{\boldsymbol{R}}_n$ for the product $\boldsymbol{BC}$ achieves a small error relative to the scale of the factors.

6.4.3 Analysis of Randomized Matrix Multiplication

The randomized matrix multiplication method is just a specific example of empirical approximation, and the error bound (6.4.5) is an immediate consequence of Corollary 6.2.1.

To pursue this approach, we need to establish a uniform bound on the norm of the estimator $\boldsymbol{R}$ for the product. Observe that

$$\|\boldsymbol{R}\| \le \max_j \|p_j^{-1} \boldsymbol{b}_{:j} \boldsymbol{c}_{j:}\| = \max_j \frac{\|\boldsymbol{b}_{:j}\|\,\|\boldsymbol{c}_{j:}\|}{p_j}.$$

To obtain a bound, recall the value (6.4.3) of the probability p_j, and invoke the inequality between geometric and arithmetic means:

$$\|\boldsymbol{R}\| \le \big(\|\boldsymbol{B}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2\big) \cdot \max_j \frac{\|\boldsymbol{b}_{:j}\|\,\|\boldsymbol{c}_{j:}\|}{\|\boldsymbol{b}_{:j}\|^2 + \|\boldsymbol{c}_{j:}\|^2} \le \frac{1}{2}\big(\|\boldsymbol{B}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2\big).$$

Since the matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ have unit spectral norm, we can express this inequality in terms of the average stable rank:

$$\|\boldsymbol{R}\| \le \frac{1}{2}\big(\mathsf{srank}(\boldsymbol{B}) + \mathsf{srank}(\boldsymbol{C})\big) = \mathsf{asr}.$$

This is the exactly kind of bound that we need.

Next, we need an estimate for the per-sample second moment $m_2(\boldsymbol{R})$. By direct calculation,

$$\begin{aligned}\mathbb{E}(\boldsymbol{R}\boldsymbol{R}^*) &= \sum_{j=1}^{N} \frac{1}{p_j^2} \cdot (\boldsymbol{b}_{:j}\boldsymbol{c}_{j:})(\boldsymbol{b}_{:j}\boldsymbol{c}_{j:})^* \cdot p_j \\ &= \left(\|\boldsymbol{B}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2\right) \cdot \sum_{j=1}^{n} \frac{\|\boldsymbol{c}_{j:}\|^2}{\|\boldsymbol{b}_{:j}\|^2 + \|\boldsymbol{c}_{j:}\|^2} \cdot \boldsymbol{b}_{:j}\boldsymbol{b}_{:j}^* \\ &\preccurlyeq \left(\|\boldsymbol{B}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2\right) \cdot \boldsymbol{B}\boldsymbol{B}^*.\end{aligned}$$

The semidefinite relation holds because each fraction lies between zero and one, and each matrix $\boldsymbol{b}_{:j}\boldsymbol{b}_{:j}^*$ is positive semidefinite. Therefore, increasing the fraction to one only increases in the matrix in the semidefinite order. Similarly,

$$\mathbb{E}(\boldsymbol{R}\boldsymbol{R}^*) \preccurlyeq \left(\|\boldsymbol{B}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2\right) \cdot \boldsymbol{C}^*\boldsymbol{C}.$$

In summary,

$$\begin{aligned}m_2(\boldsymbol{R}) &= \max\left\{\|\mathbb{E}(\boldsymbol{R}\boldsymbol{R}^*)\|, \|\mathbb{E}(\boldsymbol{R}^*\boldsymbol{R})\|\right. \\ &\le \left(\|\boldsymbol{B}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2\right) \cdot \max\left\{\|\boldsymbol{B}\boldsymbol{B}^*\|, \|\boldsymbol{C}^*\boldsymbol{C}\|\right\} \\ &= \left(\|\boldsymbol{B}\|_{\mathrm{F}}^2 + \|\boldsymbol{C}\|_{\mathrm{F}}^2\right) \\ &= 2 \cdot \mathsf{asr}.\end{aligned}$$

The penultimate line depends on the identity (2.1.24) and our assumption that both matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ have norm one.

Finally, to reach the stated estimate (6.4.5), we apply Corollary 6.2.1 with the parameters $L = \mathsf{asr}$ and $m_2(\boldsymbol{R}) \le 2 \cdot \mathsf{asr}$.

6.5 Application: Random Features

As a final application of empirical matrix approximation, let us discuss a contemporary idea from machine learning called *random features*. Although this technique may appear more sophisticated than randomized sparsification or randomized matrix multiplication, it depends on exactly the same principles. Random feature maps were proposed by Ali Rahimi and Ben Recht [RR07]. The analysis in this section is due to David Lopez-Paz et al. [LPSS+14].

6.5.1 Kernel Matrices

Let $\mathscr{X}$ be a set. We think about the elements of the set $\mathscr{X}$ as (potential) observations that we would like to use to perform learning and inference tasks. Let us introduce a bounded measure Φ of similarity between pairs of points in the set:

$$\Phi : \mathscr{X} \times \mathscr{X} \to [-1, +1].$$

The similarity measure Φ is often called a *kernel*. We assume that the kernel returns the value $+1$ when its arguments are identical, and it returns smaller values when its arguments are dissimilar. We also assume that the kernel is symmetric; that is, $\Phi(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\boldsymbol{y}, \boldsymbol{x})$ for all arguments $\boldsymbol{x}, \boldsymbol{y} \in \mathscr{X}$.

A simple example of a kernel is the angular similarity between a pair of points in a Euclidean space:

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) = \frac{2}{\pi} \arcsin \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \, \|\boldsymbol{y}\|} = 1 - \frac{2\measuredangle(\boldsymbol{x}, \boldsymbol{y})}{\pi} \quad \text{for } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d. \tag{6.5.1}$$

We write $\measuredangle(\cdot, \cdot)$ for the planar angle between two vectors, measured in radians. As usual, we instate the convention that $0/0 = 0$. See Figure 6.1 for an illustration.

Suppose that $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathscr{X}$ are observations. The kernel matrix $\boldsymbol{G} = [g_{ij}] \in \mathbb{M}_N$ just tabulates the values of the kernel function for each pair of data points:

$$g_{ij} = \Phi(\boldsymbol{x}_i, \boldsymbol{x}_j) \quad \text{for } i, j = 1, \dots, N.$$

It may be helpful to think about the kernel matrix $\boldsymbol{G}$ as a generalization of the Gram matrix of a family of points in a Euclidean space. We say that the kernel Φ is *positive definite* if the kernel matrix $\boldsymbol{G}$ is positive semidefinite for any choice of observations $\{\boldsymbol{x}_i\} \subset \mathscr{X}$. We will be concerned only with positive-definite kernels in this discussion.

In the Euclidean setting, there are statistical learning methods that only require the inner product between each pair of observations. These algorithms can be extended to the kernel setting by replacing each inner product with a kernel evaluation. As a consequence, kernel matrices can be used for classification, regression, and feature selection. In these applications, kernels are advantageous because they work outside the Euclidean domain, and they allow task-specific measures of similarity. This idea, sometimes called the *kernel trick*, is one of the major insights in modern machine learning.

A significant challenge for algorithms based on kernels is that the kernel matrix is big. Indeed, $\boldsymbol{G}$ contains $O(N^2)$ entries, where N is the number of data points. Furthermore, the cost of constructing the kernel matrix is $O(dN^2)$ where d is the number of parameters required to specify a point in the universe $\mathscr{X}$.

Nevertheless, there is an opportunity. Large data sets tend to be redundant, so the kernel matrix also tends to be redundant. This manifests in the kernel matrix being close to a low-rank matrix. As a consequence, we may try to replace the kernel matrix by a low-rank proxy. For some similarity measures, we can accomplish this task using empirical approximation.

6.5.2 Random Features and Low-Rank Approximation of the Kernel Matrix

In certain cases, a positive-definite kernel can be written as an expectation, and we can take advantage of this representation to construct an empirical approximation of the kernel matrix. Let us begin with the general construction, and then we will present a few examples in Section 6.5.3.

Let $\mathscr{W}$ be a sample space equipped with a sigma-algebra and a probability measure μ. Introduce a bounded *feature map*:

$$\psi : \mathscr{X} \times \mathscr{W} \to [-b, +b] \quad \text{where } b \geq 0.$$

Consider a random variable $\boldsymbol{w}$ taking values in $\mathscr{W}$ and distributed according to the measure μ. We assume that this random variable satisfies the *reproducing property*

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{w}} \left[\psi(\boldsymbol{x}; \boldsymbol{w}) \cdot \psi(\boldsymbol{y}; \boldsymbol{w}) \right] \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathscr{X}. \tag{6.5.2}$$

The pair $(\psi, \boldsymbol{w})$ is called a *random feature map* for the kernel Φ.

We want to approximate the kernel matrix with a set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subset \mathscr{X}$ of observations. To do so, we draw a random vector $\boldsymbol{w} \in \mathscr{W}$ distributed according to μ. Form a random vector $\boldsymbol{z} \in \mathbb{R}^N$ by applying the feature map to each data point with the *same* choice of the random vector $\boldsymbol{w}$. That is,

$$\boldsymbol{z} = \begin{bmatrix} z_1 \\ \vdots \\ z_N \end{bmatrix} = \begin{bmatrix} \psi(\boldsymbol{x}_1; \boldsymbol{w}) \\ \vdots \\ \psi(\boldsymbol{x}_N; \boldsymbol{w}) \end{bmatrix}.$$

The vector $\boldsymbol{z}$ is sometimes called a *random feature*. By the reproducing property (6.5.2) for the random feature map,

$$g_{ij} = \Phi(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathbb{E}_{\boldsymbol{w}}\left[\psi(\boldsymbol{x}_i; \boldsymbol{w}) \cdot \psi(\boldsymbol{x}_j; \boldsymbol{w})\right] = \mathbb{E}_{\boldsymbol{w}}\left[z_i \cdot z_j\right] \quad \text{for } i, j = 1, 2, 3, \dots, N.$$

We can write this relation in matrix form as $\boldsymbol{G} = \mathbb{E}(\boldsymbol{z}\boldsymbol{z}^*)$. Therefore, the random matrix $\boldsymbol{R} = \boldsymbol{z}\boldsymbol{z}^*$ is an unbiased rank-one estimator for the kernel matrix $\boldsymbol{G}$. This representation demonstrates that random feature maps, as defined here, only exist for positive-definite kernels.

As usual, we construct a better empirical approximation of the kernel matrix $\boldsymbol{G}$ by averaging several realizations of the simple estimator $\boldsymbol{R}$:

$$\bar{\boldsymbol{R}}_n = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{R}_k \quad \text{where each } \boldsymbol{R}_k \text{ is an independent copy of } \boldsymbol{R}. \tag{6.5.3}$$

In other words, we are using n independent random features $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$ to approximate the kernel matrix. The question is how many random features are needed before our estimator is accurate.

6.5.3 Examples of Random Feature Maps

Before we continue with the analysis, let us describe some random feature maps. This discussion is tangential to our theme of matrix concentration, but it is valuable to understand why random feature maps exist.

First, let us consider the angular similarity (6.5.1) defined on $\mathbb{R}^d$. We can construct a random feature map using a classical result from plane geometry. If we draw $\boldsymbol{w}$ uniformly from the unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, then

$$\Phi(\boldsymbol{x}; \boldsymbol{y}) = 1 - \frac{2\measuredangle(\boldsymbol{x}, \boldsymbol{y})}{\pi} = \mathbb{E}_{\boldsymbol{w}}\left[\operatorname{sgn}\langle \boldsymbol{x},\ \boldsymbol{w}\rangle \cdot \operatorname{sgn}\langle \boldsymbol{y},\ \boldsymbol{w}\rangle\right] \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathscr{X}. \tag{6.5.4}$$

The easy proof of this relation should be visible from the diagram in Figure 6.1. In light of the formula (6.5.4), we set $\mathscr{W} = \mathbb{S}^{d-1}$ with the uniform measure, and we define the feature map

$$\psi(\boldsymbol{x}; \boldsymbol{w}) = \operatorname{sgn}\langle \boldsymbol{x},\ \boldsymbol{w}\rangle.$$

The reproducing property (6.5.2) follows immediately from (6.5.4). Therefore, the pair $(\psi, \boldsymbol{w})$ is a random feature map for the angular similarity kernel.

Next, let us describe an important class of kernels that can be expressed using random feature maps. A kernel on $\mathbb{R}^d$ is *translation invariant* if there is a function $\varphi : \mathbb{R}^d \to \mathbb{R}$ for which

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) = \varphi(\boldsymbol{x} - \boldsymbol{y}) \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

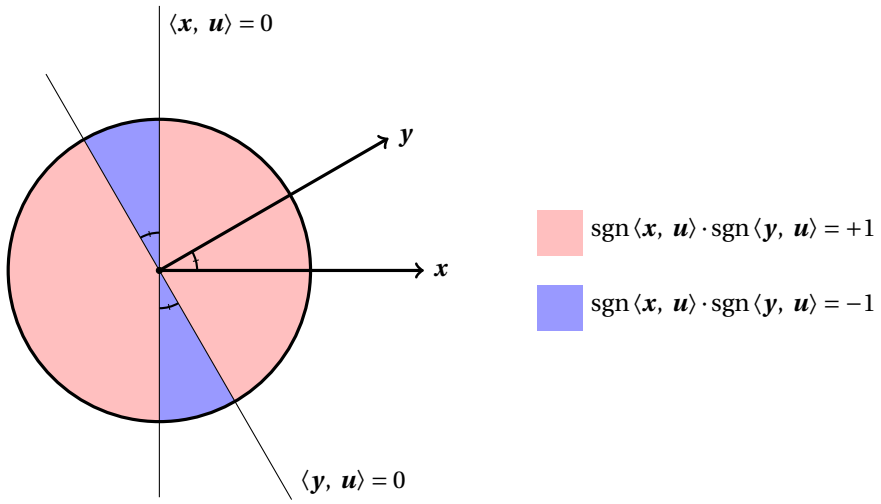

Figure 6.1: **The angular similarity between two vectors.** Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be nonzero vectors in $\mathbb{R}^2$ with angle $\angle(\boldsymbol{x}, \boldsymbol{y})$. The red region contains the directions $\boldsymbol{u}$ where the product $\operatorname{sgn}\langle \boldsymbol{x}, \boldsymbol{u}\rangle \cdot \operatorname{sgn}\langle \boldsymbol{y}, \boldsymbol{u}\rangle$ equals $+1$, and the blue region contains the directions $\boldsymbol{u}$ where the same product equals -1. The blue region subtends a total angle of $2\angle(\boldsymbol{x}, \boldsymbol{y})$, and the red region subtends a total angle of $2\pi - 2\angle(\boldsymbol{x}, \boldsymbol{y})$.

Bôchner's Theorem, a classical result from harmonic analysis, gives a representation for each continuous, positive-definite, translation-invariant kernel:

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) = \varphi(\boldsymbol{x} - \boldsymbol{y}) = c \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\langle \boldsymbol{x},\, \boldsymbol{w}\rangle} \cdot \mathrm{e}^{-\mathrm{i}\langle \boldsymbol{y},\, \boldsymbol{w}\rangle} \, \mathrm{d}\mu(\boldsymbol{w}) \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d. \tag{6.5.5}$$

In this expression, the positive scale factor c and the probability measure μ depend only on the function φ. The formula (6.5.5) yields a (complex-valued) random feature map:

$$\psi_{\mathbb{C}}(\boldsymbol{x}; \boldsymbol{w}) = \sqrt{c}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{x},\, \boldsymbol{w}\rangle} \quad \text{where } \boldsymbol{w} \text{ has distribution } \mu \text{ on } \mathbb{R}^d.$$

This map satisfies a complex variant of the reproducing property (6.5.2):

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{w}} \left[\psi_{\mathbb{C}}(\boldsymbol{x}; \boldsymbol{w}) \cdot \psi_{\mathbb{C}}(\boldsymbol{y}; \boldsymbol{w})^* \right] \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d,$$

where we have written * for complex conjugation.

With a little more work, we can construct a real-valued random feature map. Recall that the kernel Φ is symmetric, so the complex exponentials in (6.5.5) can be written in terms of cosines. This observation leads to the random feature map

$$\psi(\boldsymbol{x}; \boldsymbol{w}, U) = \sqrt{2c} \cos\big(\langle \boldsymbol{x},\, \boldsymbol{w}\rangle + U\big) \quad \text{where} \quad \boldsymbol{w} \sim \mu \text{ and } U \sim \textsc{uniform}[0, 2\pi]. \tag{6.5.6}$$

To verify that $(\psi, (\boldsymbol{w}, U))$ reproduces the kernel Φ, as required by (6.5.2), we just make a short calculation using the angle-sum formula for the cosine.

We conclude this section with the most important example of a random feature map from the class we have just described. Consider the Gaussian radial basis function kernel:

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{e}^{-\alpha\|\boldsymbol{x}-\boldsymbol{y}\|^2/2} \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

The positive parameter α reflects how close two points must be before they are regarded as "similar." For the Gaussian kernel, Bôchner's Theorem (6.5.5) holds with the scaling factor $c = 1$ and the probability measure $\mu = \text{NORMAL}(\mathbf{0}, \alpha\,\mathbf{I}_d)$. In summary, we define

$$\psi(\boldsymbol{x}; \boldsymbol{w}, U) = \sqrt{2}\cos\big(\langle \boldsymbol{x},\ \boldsymbol{w}\rangle + U\big) \quad \text{where} \quad \boldsymbol{w} \sim \text{NORMAL}(\mathbf{0}, \alpha\,\mathbf{I}_d) \text{ and } U \sim \text{UNIFORM}[0, 2\pi].$$

This random feature map reproduces the Gaussian radial basis function kernel.

6.5.4 Performance of the Random Feature Approximation

We will demonstrate that the approximation $\bar{\boldsymbol{R}}_n$ of the $N \times N$ kernel matrix $\boldsymbol{G}$ using n random features, constructed in (6.5.3), leads to an estimate of the form

$$\mathbb{E}\,\|\bar{\boldsymbol{R}}_n - \boldsymbol{G}\| \le \sqrt{\frac{2bN\,\|\boldsymbol{G}\|\log(2N)}{n}} + \frac{2bN\log(2N)}{3n}. \tag{6.5.7}$$

In this expression, b is the uniform bound on the magnitude of the feature map ψ. The short proof of (6.5.7) appears in §6.5.5.

To clarify what this result means, we introduce the *intrinsic dimension* of the $N \times N$ kernel matrix $\boldsymbol{G}$:

$$\operatorname{intdim}(\boldsymbol{G}) = \operatorname{srank}(\boldsymbol{G}^{1/2}) = \frac{\operatorname{tr}\boldsymbol{G}}{\|\boldsymbol{G}\|} = \frac{N}{\|\boldsymbol{G}\|}.$$

The stable rank is defined in Section 2.1.15. We have used the assumption that the similarity measure is positive definite to justify the computation of the square root of the kernel matrix, and $\operatorname{tr}\boldsymbol{G} = N$ because of the requirement that $\Phi(\boldsymbol{x}, \boldsymbol{x}) = +1$ for all $\boldsymbol{x} \in \mathscr{X}$. See §7.1 for further discussion of the intrinsic dimension

Now, assume that the number n of random features satisfies the bound

$$n \ge 2b\varepsilon^{-2} \cdot \operatorname{intdim}(\boldsymbol{G}) \cdot \log(2N),$$

In view of (6.5.7), the relative error in the empirical approximation of the kernel matrix satisfies

$$\frac{\mathbb{E}\,\|\bar{\boldsymbol{R}}_n - \boldsymbol{G}\|}{\|\boldsymbol{G}\|} \le \varepsilon + \varepsilon^{-2}.$$

We learn that the randomized approximation of the kernel matrix $\boldsymbol{G}$ is accurate when its intrinsic dimension is much smaller than the number of data points. That is, $\operatorname{intdim}(\boldsymbol{G}) \ll N$.

6.5.5 Analysis of the Random Feature Approximation

The analysis of random features is based on Corollary 6.2.1. To apply this result, we need the per-sample second-moment $m_2(\boldsymbol{R})$ and the uniform upper bound L. Both are easy to come by.

First, observe that

$$\|\boldsymbol{R}\| = \|\boldsymbol{z}\boldsymbol{z}^*\| = \|\boldsymbol{z}\|^2 \le bN$$

Recall that b is the uniform bound on the feature map ψ, and N is the number of components in the random feature vector $\boldsymbol{z}$.

Second, we calculate that

$$\mathbb{E}\boldsymbol{R}^2 = \mathbb{E}\left(\|\boldsymbol{z}\|^2 \boldsymbol{z}\boldsymbol{z}^*\right) \preccurlyeq bN \cdot \mathbb{E}(\boldsymbol{z}\boldsymbol{z}^*) = bN \cdot \boldsymbol{G}.$$

Each random matrix $\boldsymbol{z}\boldsymbol{z}^*$ is positive semidefinite, so we can introduce the upper bound $\|\boldsymbol{z}\|^2 \le bN$. The last identity holds because $\boldsymbol{R}$ is an unbiased estimator of the kernel matrix $\boldsymbol{G}$. It follows that

$$m_2(\boldsymbol{R}) = \|\mathbb{E}\boldsymbol{R}^2\| \le bN \cdot \|\boldsymbol{G}\|.$$

This is our bound for the per-sample second moment.

Finally, we invoke Corollary 6.2.1 with parameters $L = bN$ and $m_2(\boldsymbol{R}) \le bN\|\boldsymbol{G}\|$ to arrive at the estimate (6.5.7).

6.6 Proof of the Matrix Bernstein Inequality

Now, let us turn to the proof of the matrix Bernstein inequality, Theorem 6.1.1. This result is a corollary of a matrix concentration inequality for a sum of bounded random Hermitian matrices. We begin with a statement and discussion of the Hermitian result, and then we explain how the general result follows.

6.6.1 A Sum of Bounded Random Hermitian Matrices

The first result is a Bernstein inequality for a sum of independent, random Hermitian matrices whose eigenvalues are bounded above.

Theorem 6.6.1 (Matrix Bernstein: Hermitian Case)**.** *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, Hermitian matrices with dimension d. Assume that*

$$\mathbb{E}\boldsymbol{X}_k = \boldsymbol{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le L \quad \text{for each index } k.$$

Introduce the random matrix

$$\boldsymbol{Y} = \sum\nolimits_k \boldsymbol{X}_k.$$

Let $v(\boldsymbol{Y})$ be the matrix variance statistic of the sum:

$$v(\boldsymbol{Y}) = \|\mathbb{E}\boldsymbol{Y}^2\| = \left\|\sum\nolimits_k \mathbb{E}\boldsymbol{X}_k^2\right\|. \tag{6.6.1}$$

Then

$$\mathbb{E}\lambda_{\max}(\boldsymbol{Y}) \le \sqrt{2v(\boldsymbol{Y})\log d} + \frac{1}{3}L\log d. \tag{6.6.2}$$

Furthermore, for all $t \ge 0$.

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} \le d \cdot \exp\left(\frac{-t^2/2}{v(\boldsymbol{Y}) + Lt/3}\right). \tag{6.6.3}$$

The proof of Theorem 6.6.1 appears below in §6.6.

6.6.2 Discussion

Theorem 6.6.1 also yields information about the minimum eigenvalue of an independent sum of d-dimensional Hermitian matrices. Suppose that the independent random matrices satisfy

$$\mathbb{E}\boldsymbol{X}_k = \mathbf{0} \quad \text{and} \quad \lambda_{\min}(\boldsymbol{X}_k) \geq -\underline{L} \quad \text{for each index } k.$$

Applying the expectation bound (6.6.2) to $-\boldsymbol{Y}$, we obtain

$$\mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}) \geq -\sqrt{2v(\boldsymbol{Y})\log d} - \frac{1}{3}\underline{L}\log d. \tag{6.6.4}$$

We can use (6.6.3) to develop a tail bound. For $t \geq 0$,

$$\mathbb{P}\{\lambda_{\min}(\boldsymbol{Y}) \leq -t\} \leq d \cdot \exp\left(\frac{-t^2/2}{v(\boldsymbol{Y}) + \underline{L}t/3}\right).$$

Let us emphasize that the bounds for $\lambda_{\max}(\boldsymbol{Y})$ and $\lambda_{\min}(\boldsymbol{Y})$ may diverge because the two parameters L and $\underline{L}$ can take sharply different values. This fact indicates that the maximum eigenvalue bound in Theorem 6.6.1 is a less strict assumption than the spectral norm bound in Theorem 6.1.1.

6.6.3 Bounds for the Matrix Mgf and Cgf

In establishing the matrix Bernstein inequality, the main challenge is to obtain an appropriate bound for the matrix mgf and cgf of a zero-mean random matrix whose norm satisfies a uniform bound. We do not present the sharpest estimate possible, but rather the one that leads most directly to the useful results stated in Theorem 6.6.1.

Lemma 6.6.2 (Matrix Bernstein: Mgf and Cgf Bound)**.** *Suppose that $\boldsymbol{X}$ is a random Hermitian matrix that satisfies*

$$\mathbb{E}\boldsymbol{X} = \mathbf{0} \quad \textit{and} \quad \lambda_{\max}(\boldsymbol{X}) \leq L.$$

Then, for $0 < \theta < 3/L$,

$$\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \exp\left(\frac{\theta^2/2}{1-\theta L/3} \cdot \mathbb{E}\boldsymbol{X}^2\right) \quad \textit{and} \quad \log \mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \frac{\theta^2/2}{1-\theta L/3} \cdot \mathbb{E}\boldsymbol{X}^2.$$

Proof. Fix the parameter $\theta > 0$. In the exponential $\mathrm{e}^{\theta \boldsymbol{X}}$, we would like to expose the random matrix $\boldsymbol{X}$ and its square $\boldsymbol{X}^2$ so that we can exploit information about the mean and variance. To that end, we write

$$\mathrm{e}^{\theta \boldsymbol{X}} = \mathbf{I} + \theta\boldsymbol{X} + (\mathrm{e}^{\theta \boldsymbol{X}} - \theta\boldsymbol{X} - \mathbf{I}) = \mathbf{I} + \theta\boldsymbol{X} + \boldsymbol{X} \cdot f(\boldsymbol{X}) \cdot \boldsymbol{X}, \tag{6.6.5}$$

where f is a function on the real line:

$$f(x) = \frac{\mathrm{e}^{\theta x} - \theta x - 1}{x^2} \quad \text{for } x \neq 0 \quad \text{and} \quad f(0) = \frac{\theta^2}{2}.$$

The function f is increasing because its derivative is positive. Therefore, $f(x) \leq f(L)$ when $x \leq L$. By assumption, the eigenvalues of $\boldsymbol{X}$ do not exceed L, so the Transfer Rule (2.1.14) implies that

$$f(\boldsymbol{X}) \preccurlyeq f(L) \cdot \mathbf{I}. \tag{6.6.6}$$

The Conjugation Rule (2.1.12) allows us to introduce the relation (6.6.6) into our expansion (6.6.5) of the matrix exponential:

$$\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \mathbf{I} + \theta \boldsymbol{X} + \boldsymbol{X}(f(L) \cdot \mathbf{I})\boldsymbol{X} = \mathbf{I} + \theta \boldsymbol{X} + f(L) \cdot \boldsymbol{X}^2.$$

This relation is the basis for our matrix mgf bound.

To obtain the desired result, we develop a further estimate for $f(L)$. This argument involves a clever application of Taylor series:

$$f(L) = \frac{\mathrm{e}^{\theta L} - \theta L - 1}{L^2} = \frac{1}{L^2} \sum_{q=2}^{\infty} \frac{(\theta L)^q}{q!} \le \frac{\theta^2}{2} \sum_{q=2}^{\infty} \frac{(\theta L)^{q-2}}{3^{q-2}} = \frac{\theta^2/2}{1 - \theta L/3}.$$

The second expression is simply the Taylor expansion of the fraction, viewed as a function of θ. We obtain the inequality by factoring out $(\theta L)^2/2$ from each term in the series and invoking the bound $q! \ge 2 \cdot 3^{q-2}$, valid for each $q = 2, 3, 4, \dots$. Sum the geometric series to obtain the final identity.

To complete the proof of the mgf bound, we combine the last two displays:

$$\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \mathbf{I} + \theta \boldsymbol{X} + \frac{\theta^2/2}{1 - \theta L/3} \cdot \boldsymbol{X}^2.$$

This estimate is valid because $\boldsymbol{X}^2$ is positive semidefinite. Expectation preserves the semidefinite order, so

$$\mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \mathbf{I} + \frac{\theta^2/2}{1 - \theta L/3} \cdot \mathbb{E}\, \boldsymbol{X}^2 \preccurlyeq \exp\left(\frac{\theta^2/2}{1 - \theta L/3} \cdot \mathbb{E}\, \boldsymbol{X}^2 \right).$$

We have used the assumption that $\boldsymbol{X}$ has zero mean. The second semidefinite relation follows when we apply the Transfer Rule (2.1.14) to the inequality $1 + a \le \mathrm{e}^a$, which holds for $a \in \mathbb{R}$.

To obtain the semidefinite bound for the cgf, we extract the logarithm of the mgf bound using the fact (2.1.18) that the logarithm is operator monotone. □

6.6.4 Proof of the Hermitian Case

We are prepared to establish the matrix Bernstein inequalities for random Hermitian matrices.

Proof of Theorem 6.6.1. Consider a finite sequence $\{\boldsymbol{X}_k\}$ of random Hermitian matrices with dimension d. Assume that

$$\mathbb{E}\, \boldsymbol{X}_k = \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le L \quad \text{for each index } k.$$

The matrix Bernstein cgf bound, Lemma 6.6.2, provides that

$$\log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq g(\theta) \cdot \mathbb{E}\, \boldsymbol{X}_k^2 \quad \text{where} \quad g(\theta) = \frac{\theta^2/2}{1 - \theta L/3} \quad \text{for } 0 < \theta < 3/L. \tag{6.6.7}$$

Introduce the sum $\boldsymbol{Y} = \sum_k \boldsymbol{X}_k$.

We begin with the bound (6.6.2) for the expectation $\mathbb{E}\, \lambda_{\max}(\boldsymbol{Y})$. Invoke the master inequality, relation (3.6.1) in Theorem 3.6.1, to find that

$$\begin{aligned} \mathbb{E}\, \lambda_{\max}(\boldsymbol{Y}) &\le \inf_{\theta > 0} \frac{1}{\theta} \log \operatorname{tr} \exp\left(\textstyle\sum_k \log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}_k} \right) \\ &\le \inf_{0 < \theta < 3/L} \frac{1}{\theta} \log \operatorname{tr} \exp\left(g(\theta) \textstyle\sum_k \mathbb{E}\, \boldsymbol{X}_k^2 \right) \\ &= \inf_{0 < \theta < 3/L} \frac{1}{\theta} \log \operatorname{tr} \exp\left(g(\theta) \cdot \mathbb{E}\, \boldsymbol{Y}^2 \right). \end{aligned}$$

As usual, to move from the first to the second line, we invoke the fact (2.1.16) that the trace exponential is monotone to introduce the semidefinite bound (6.6.7) for the cgf. Then we use the additivity rule (2.2.5) for the variance of an independent sum to identify $\mathbb{E}\boldsymbol{Y}^2$. The rest of the argument glides along a well-oiled track:

$$\begin{aligned}
\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) &\le \inf_{0<\theta<3/L} \frac{1}{\theta}\log\left[d\,\lambda_{\max}\left(\exp\left(g(\theta)\cdot\mathbb{E}\,\boldsymbol{Y}^2\right)\right)\right] \\
&= \inf_{0<\theta<3/L} \frac{1}{\theta}\log\left[d\exp\left(g(\theta)\cdot\lambda_{\max}\left(\mathbb{E}\,\boldsymbol{Y}^2\right)\right)\right] \\
&\le \inf_{0<\theta<3/L} \frac{1}{\theta}\log\left[d\exp\left(g(\theta)\cdot v(\boldsymbol{Y})\right)\right] \\
&= \inf_{0<\theta<3/L} \left[\frac{\log d}{\theta} + \frac{\theta/2}{1-\theta L/3}\cdot v(\boldsymbol{Y})\right].
\end{aligned}$$

In the first inequality, we bound the trace of the exponential by the dimension d times the maximum eigenvalue. The next line follows from the Spectral Mapping Theorem, Proposition 2.1.3. In the third line, we identify the matrix variance statistic $v(\boldsymbol{Y})$ from (6.6.1). Afterward, we extract the logarithm and simplify. Finally, we compute the infimum to complete the proof of (6.6.2). For reference, the optimal argument is

$$\theta = \frac{6L\log d + 9\sqrt{2v(\boldsymbol{Y})\log d}}{2L^2 t + 9(\boldsymbol{Y}) + 6L\sqrt{2(\boldsymbol{Y})\log d}}.$$

We recommend using a computer algebra system to confirm this point.

Next, we develop the tail bound (6.6.3) for $\lambda_{\max}(\boldsymbol{Y})$. Owing to the master tail inequality (3.6.3), we have

$$\begin{aligned}
\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} &\le \inf_{\theta>0} \mathrm{e}^{-\theta t}\,\mathrm{tr}\exp\left(\textstyle\sum_k \log\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}_k}\right) \\
&\le \inf_{0<\theta<3/L} \mathrm{e}^{-\theta t}\,\mathrm{tr}\exp\left(g(\theta)\textstyle\sum_k \mathbb{E}\,\boldsymbol{X}_k^2\right) \\
&\le \inf_{0<\theta<3/L} d\,\mathrm{e}^{-\theta t}\exp\left(g(\theta)\cdot v(\boldsymbol{Y})\right).
\end{aligned}$$

The justifications are the same as before. The exact value of the infimum is messy, so we proceed with the inspired choice $\theta = t/(v(\boldsymbol{Y}) + Lt/3)$, which results in the elegant bound (6.6.3). □

6.6.5 Proof of the General Case

Finally, we explain how to derive Theorem 6.1.1, for general matrices, from Theorem 6.6.1. This result follows immediately when we apply the matrix Bernstein bounds for Hermitian matrices to the Hermitian dilation of a sum of general matrices.

Proof of Theorem 6.1.1. Consider a finite sequence $\{\boldsymbol{S}_k\}$ of $d_1 \times d_2$ random matrices, and assume that

$$\mathbb{E}\,\boldsymbol{S}_k = \mathbf{0} \quad\text{and}\quad \|\boldsymbol{S}_k\| \le L \quad\text{for each index } k.$$

We define the two random matrices

$$\boldsymbol{Z} = \textstyle\sum_k \boldsymbol{S}_k \quad\text{and}\quad \boldsymbol{Y} = \mathscr{H}(\boldsymbol{Z}) = \textstyle\sum_k \mathscr{H}(\boldsymbol{S}_k)$$

where $\mathscr{H}$ is the Hermitian dilation (2.1.26). The second expression for $\boldsymbol{Y}$ follows from the property that the dilation is a real-linear map.

We will apply Theorem 6.6.1 to analyze $\|\boldsymbol{Z}\|$. First, recall the fact (2.1.28) that

$$\|\boldsymbol{Z}\| = \lambda_{\max}(\mathscr{H}(\boldsymbol{Z})) = \lambda_{\max}(\boldsymbol{Y}).$$

Next, we express the variance (6.6.1) of the random Hermitian matrix $\boldsymbol{Y}$ in terms of the general matrix $\boldsymbol{Z}$. Indeed, the calculation (2.2.10) of the variance statistic of a dilation shows that

$$v(\boldsymbol{Y}) = v(\mathscr{H}(\boldsymbol{Y})) = v(\boldsymbol{Z}).$$

Recall that the matrix variance statistic $v(\boldsymbol{Z})$ defined in (6.1.1) coincides with the general definition from (2.2.8). Finally, we invoke Theorem 6.6.1 to establish Theorem 6.1.1. □

6.7 Notes

The literature contains a wide variety of Bernstein-type inequalities in the scalar case, and the matrix case is no different. The applications of the matrix Bernstein inequality are also numerous. We only give a brief summary here.

6.7.1 Matrix Bernstein Inequalities

David Gross [Gro11] and Ben Recht [Rec11] used the approach of Ahlswede & Winter [AW02] to develop two different versions of the matrix Bernstein inequality. These papers helped to popularize the use matrix concentration inequalities in mathematical signal processing and statistics. Nevertheless, their results involve a suboptimal variance parameter of the form

$$v_{\mathrm{AW}}(\boldsymbol{Y}) = \sum\nolimits_k \left\|\mathbb{E}\,\boldsymbol{X}_k^2\right\|.$$

This parameter can be significantly larger than the matrix variance statistic (6.6.1) that appears in Theorem 6.6.1. They do coincide in some special cases, such as when the summands are independent and identically distributed.

Oliveira [Oli10a] established the first version of the matrix Bernstein inequality that yields the correct matrix variance statistic (6.6.1). He accomplished this task with an elegant application of the Golden–Thompson inequality (3.3.3). His method even gives a result, called the matrix Freedman inequality, that holds for matrix-valued martingales. His bound is roughly equivalent with Theorem 6.6.1, up to the precise value of the constants.

The matrix Bernstein inequality we have stated here, Theorem 6.6.1, first appeared in the paper [Tro11c, §6] by the author of these notes. The bounds for the expectation are new. The argument is based on Lieb's Theorem, and it also delivers a matrix Bennett inequality. This paper also describes how to establish matrix Bernstein inequalities for sums of unbounded random matrices, given some control over the matrix moments.

The research in [Tro11c] is independent from Oliveira's work [Oli10a], although Oliveira's paper motivated the subsequent article [Tro11a] and the technical report [Tro11b], which explain how to use Lieb's Theorem to study matrix martingales. The technical report [GT14] develops a Bernstein inequality for interior eigenvalues using the Lieb–Seiringer Theorem [LS05].

For more versions of the matrix Bernstein inequality, see Vladimir Koltchinskii's lecture notes from Saint-Flour [Kol11]. In Chapter 7, we present another extension of the matrix Bernstein inequality that involves a smaller dimensional parameter.

6.7.2 The Matrix Rosenthal–Pinelis Inequality

The matrix Rosenthal–Pinelis inequality (6.1.6) is a close cousin of the matrix Rosenthal inequality (5.1.9). Both results are derived from the noncommutative Khintchine inequality (4.7.1) using the same pattern of argument [CGT12a, Thm. A.1]. We believe that [CGT12a] is the first paper to recognize and state the result (6.1.6), even though it is similar in spirit with the work in [Rud99]. A self-contained, elementary proof of a related matrix Rosenthal-Pinelis inequality appears in [MJC+14, Cor. 7.4].

Versions of the matrix Rosenthal–Pinelis inequality first appeared in the literature [JX03] on noncommutative martingales, where they were called *noncommutative Burkholder inequalities*. For an application to random matrices, see the follow-up work [JX08] by the same authors. Subsequent papers [JZ12, JZ13] contain related noncommutative martingale inequalities inspired by the research in [Oli10b, Tro11c].

6.7.3 Empirical Approximation

Matrix approximation by random sampling is a special case of a general method that Bernard Maurey developed to compute entropy numbers of convex hulls. Let us give a short presentation of the original context, along with references to some other applications.

Empirical Bounds for Covering Numbers

Suppose that X is a Banach space. Consider the convex hull $E = \operatorname{conv}\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_N\}$ of a set of N points in X, and assume that $\|\boldsymbol{e}_k\| \le L$. We would like to give an upper bound for the number of balls of radius ε it takes to cover this set.

Fix a point $\boldsymbol{u} \in E$, and express $\boldsymbol{u}$ as a convex combination:

$$\boldsymbol{u} = \sum_{i=1}^{N} p_i \boldsymbol{e}_i \quad \text{where} \quad \sum_{i=1}^{N} p_i = 1 \quad \text{and} \quad p_i \ge 0.$$

Let $\boldsymbol{x}$ be the random vector in X that takes value $\boldsymbol{e}_k$ with probability p_k. We can approximate the point $\boldsymbol{u}$ as an average $\bar{\boldsymbol{x}} = n^{-1} \sum_{k=1}^{n} \boldsymbol{x}_k$ of independent copies $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ of the random vector $\boldsymbol{x}$. Then

$$\mathbb{E}\, \|\bar{\boldsymbol{x}}_n - \boldsymbol{u}\|_X = \frac{1}{n} \mathbb{E}\, \Big\| \sum\nolimits_{k=1}^{n} (\boldsymbol{x}_k - \mathbb{E}\,\boldsymbol{x}) \Big\|_X \le \frac{2}{n} \mathbb{E}\, \Big\| \sum\nolimits_{k=1}^{n} \varrho_k \boldsymbol{x}_k \Big\|_X \le \frac{2}{n} \Big(\mathbb{E}\, \Big\| \sum\nolimits_{k=1}^{n} \varrho_k \boldsymbol{x}_k \Big\|_X^2 \Big)^{1/2}.$$

The family $\{\varrho_k\}$ consists of independent Rademacher random variables. The first inequality depends on the symmetrization procedure [LT91, Lem. 6.3], and the second is Hölder's. In certain Banach spaces, a Khintchine-type inequality holds:

$$\mathbb{E}\, \|\bar{\boldsymbol{x}}_n - \boldsymbol{u}\|_X \le \frac{2T_2(X)}{n} \Big(\sum\nolimits_{k=1}^{n} \mathbb{E}\, \|\boldsymbol{x}_k\|_X^2 \Big)^{1/2} \le \frac{2T_2(X)L}{\sqrt{n}}.$$

The last inequality depends on the uniform bound $\|\boldsymbol{e}_k\| \le L$. This estimate controls the expected error in approximating an arbitrary point in E by randomized sampling.

The number $T_2(X)$ is called the *type two constant* of the Banach space X, and it can be estimated in many concrete instances; see [LT91, Chap. 9] or [Pis89, Chap. 11]. For our purposes,

the most relevant example is the Banach space $\mathbb{M}^{d_1 \times d_2}$ consisting of $d_1 \times d_2$ matrices equipped with the spectral norm. Its type two constant satisfies

$$T_2\big(\mathbb{M}^{d_1 \times d_2}\big) \le \text{Const} \cdot \sqrt{\log(d_1 + d_2)}.$$

This result follows from work of Tomczak–Jaegermann [TJ74, Thm. 3.1(ii)]. In fact, the space $\mathbb{M}^{d_1 \times d_2}$ enjoys an even stronger property with respect to averages, namely the noncommutative Khintchine inequality (4.7.1).

Now, suppose that the number n of samples in our empirical approximation $\bar{\boldsymbol{x}}_n$ of the point $\boldsymbol{u} \in E$ satisfies

$$n \ge \left(\frac{2T_2(X)L}{\varepsilon} \right)^2.$$

Then the probabilistic method ensures that there is a some collection of $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ of points drawn with repetition from the set $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_N\}$ that satisfies

$$\big\| \big(n^{-1} \textstyle\sum_{k=1}^n \boldsymbol{u}_k\big) - \boldsymbol{u} \big\|_X \le \varepsilon.$$

There are at most N^n different ways to select the points $\boldsymbol{u}_k$. It follows that we can cover the convex hull $E = \operatorname{conv}\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_N\}$ in X with at most N^n norm balls of radius ε.

History and Applications of Empirical Approximation

Maurey did not publish his ideas, and the method was first broadcast in a paper of Pisier [Pis81, Lem. 1]. Another early reference is the work of Carl [Car85, Lem. 1]. More recently, this covering argument has been used to study the restricted isomorphism behavior of a random set of rows drawn from a discrete Fourier transform matrix [RV06].

By now, empirical approximation has appeared in a wide range of applied contexts, although many papers do not recognize the provenance of the method. Let us mention some examples in machine learning. Empirical approximation has been used to study what functions can be approximated by neural networks [Bar93, LBW96]. The same idea appears in papers on sparse modeling, such as [SSS08], and it supports the method of random features [RR07]. Empirical approximation also stands at the core of a recent algorithm for constructing approximate Nash equilibria [Bar14].

It is difficult to identify the earliest work in computational mathematics that invoked the empirical method to approximate matrices. The paper of Achlioptas & McSherry [AM01] on randomized sparsification is one possible candidate.

Corollary 6.2.1, which we use to perform the analysis of matrix approximation by sampling, does not require the full power of the matrix Bernstein inequality, Theorem 6.1.1. Indeed, Corollary 6.2.1 can be derived from the weaker methods of Ahlswede & Winter [AW02]; for example, see the papers [Gro11, Rec11].

6.7.4 Randomized Sparsification

The idea of using randomized sparsification to accelerate spectral computations appears in a paper of Achlioptas & McSherry [AM01, AM07]. d'Asprémont [d'A11] proposed to use sparsification to accelerate algorithms for semidefinite programming. The paper [AKL13] by Achlioptas, Karnin, & Liberty recommends sparsification as a mechanism for data compression.

After the initial paper [AM01], several other researchers developed sampling schemes for randomized sparsification [AHK06, GT09]. Later, Drineas & Zouzias [DZ11] pointed out that matrix concentration inequalities can be used to analyze this type of algorithm. The paper [AKL13] refined this analysis to obtain sharper bounds. The simple analysis here is drawn from a recent note by Kundu & Drineas [KD14].

6.7.5 Randomized Matrix Multiplication

The idea of using random sampling to accelerate matrix multiplication appeared in nascent form in a paper of Frieze, Kannan, & Vempala [FKV98]. The paper [DK01] of Drineas & Kannan develops this idea in full generality, and the article [DKM06] of Drineas, Kannan, & Mahoney contains a more detailed treatment. Subsequently, Tamás Sarlós obtained a significant improvement in the performance of this algorithm [Sar06]. Rudelson & Vershynin [RV07] obtained the first error bound for approximate matrix multiplication with respect to the spectral norm. The analysis that we presented is adapted from the dissertation [Zou13] of Tassos Zouzias, which refines an earlier treatment by Magen & Zouzias [MZ11]. See the monographs of Mahoney [Mah11] and Woodruff [Woo14] for a more extensive discussion.

6.7.6 Random Features

Our discussion of kernel methods is adapted from the book [SS98]. The papers [RR07, RR08] of Ali Rahimi and Ben Recht proposed the idea of using random features to summarize data for large-scale kernel machines. The construction (6.5.6) of a random feature map for a translation-invariant, positive-definite kernel appears in their work. This approach has received a significant amount of attention over the last few years, and there has been a lot of subsequent development. For example, the paper [KK12] of Kar & Karnick shows how to construct random features for inner-product kernels, and the paper [HXGD14] of Hamid et al. develops random features for polynomial kernels. Our analysis of random features using the matrix Bernstein inequality is drawn from the recent article [LPSS+14] of Lopez-Paz et al. The presentation here is adapted from the author's tutorial on randomized matrix approximation, given at ICML 2014 in Beijing. We recommend the two papers [HXGD14, LPSS+14] for an up-to-date bibliography.

CHAPTER 7

Results Involving the Intrinsic Dimension

A minor shortcoming of our matrix concentration results is the dependence on the ambient dimension of the matrix. In this chapter, we show how to obtain a dependence on an intrinsic dimension parameter, which occasionally is much smaller than the ambient dimension. In many cases, intrinsic dimension bounds offer only a modest improvement. Nevertheless, there are examples where the benefits are significant enough that we can obtain nontrivial results for infinite-dimensional random matrices.

In this chapter, present a version of the matrix Chernoff inequality that involves an intrinsic dimension parameter. We also describe a version of the matrix Bernstein inequality that involves an intrinsic dimension parameter. The intrinsic Bernstein result usually improves on Theorem 6.1.1. These results depend on a new argument that distills ideas from a paper [Min11] of Stanislav Minsker. We omit intrinsic dimension bounds for matrix series, which the reader may wish to develop as an exercise.

To give a sense of what these new results accomplish, we revisit some of the examples from earlier chapters. We apply the intrinsic Chernoff bound to study a random column submatrix of a fixed matrix. We also reconsider the randomized matrix multiplication algorithm in light of the intrinsic Bernstein bound. In each case, the intrinsic dimension parameters have an attractive interpretation in terms of the problem data.

Overview

We begin our development in §7.1 with the definition of the intrinsic dimension of a matrix. In §7.2, we present the intrinsic Chernoff bound and some of its consequences. In §7.3, we describe the intrinsic Bernstein inequality and its applications. Afterward, we describe the new ingredients that are required in the proofs. Section 7.4 explains how to extend the matrix Laplace transform method beyond the exponential function, and §7.5 describes a simple but powerful lemma that allows us to obtain the dependence on the intrinsic dimension. Section 7.6 contains

the proof of the intrinsic Chernoff bound, and §7.7 develops the proof of the intrinsic Bernstein bound.

7.1 The Intrinsic Dimension of a Matrix

Some types of random matrices are concentrated in a small number of dimensions, while they have little content in other dimensions. So far, our bounds do not account for the difference. We need to introduce a more refined notion of dimension that will help us to discriminate among these examples.

Definition 7.1.1 (Intrinsic Dimension)**.** *For a positive-semidefinite matrix $\boldsymbol{A}$, the* intrinsic dimension *is the quantity*

$$\operatorname{intdim}(\boldsymbol{A}) = \frac{\operatorname{tr} \boldsymbol{A}}{\|\boldsymbol{A}\|}.$$

We interpret the intrinsic dimension as a measure of the number of dimensions where $\boldsymbol{A}$ has significant spectral content.

Let us make a few observations that support this view. By expressing the trace and the norm in terms of the eigenvalues, we can verify that

$$1 \le \operatorname{intdim}(\boldsymbol{A}) \le \operatorname{rank}(\boldsymbol{A}) \le \dim(\boldsymbol{A}).$$

The first inequality is attained precisely when $\boldsymbol{A}$ has rank one, while the second inequality is attained precisely when $\boldsymbol{A}$ is a multiple of the identity. The intrinsic dimension is 0-homogeneous, so it is insensitive to changes in the scale of the matrix $\boldsymbol{A}$. The intrinsic dimension is *not* monotone with respect to the semidefinite order. Indeed, we can drive the intrinsic dimension to one by increasing one eigenvalue of $\boldsymbol{A}$ substantially.

7.2 Matrix Chernoff with Intrinsic Dimension

Let us present an extension of the matrix Chernoff inequality. This result controls the maximum eigenvalue of a sum of random, positive-semidefinite matrices in terms of the intrinsic dimension of the expectation of the sum.

Theorem 7.2.1 (Matrix Chernoff: Intrinsic Dimension)**.** *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of random, Hermitian matrices of the same size, and assume that*

$$0 \le \lambda_{\min}(\boldsymbol{X}_k) \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le L \quad \text{for each index } k.$$

Introduce the random matrix

$$\boldsymbol{Y} = \sum\nolimits_k \boldsymbol{X}_k.$$

Suppose that we have a semidefinite upper bound $\boldsymbol{M}$ for the expectation $\mathbb{E}\,\boldsymbol{Y}$:

$$\boldsymbol{M} \succcurlyeq \mathbb{E}\,\boldsymbol{Y} = \sum\nolimits_k \mathbb{E}\,\boldsymbol{X}_k.$$

Define an intrinsic dimension bound and a mean bound:

$$d = \operatorname{intdim}(\boldsymbol{M}) \quad \text{and} \quad \mu_{\max} = \lambda_{\max}(\boldsymbol{M}).$$

Then, for $\theta > 0$,

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \le \frac{\mathrm{e}^{\theta}-1}{\theta}\cdot\mu_{\max} + \frac{1}{\theta}\cdot L\log(2d). \tag{7.2.1}$$

Furthermore,

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \ge (1+\varepsilon)\mu_{\max}\right\} \le 2d\cdot\left[\frac{\mathrm{e}^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right]^{\mu_{\max}/L} \quad \text{for } \varepsilon \ge L/\mu_{\max}. \tag{7.2.2}$$

The proof of this result appears below in §7.6.

7.2.1 Discussion

Theorem 7.2.1 is almost identical with the parts of the basic matrix Chernoff inequality that concern the maximum eigenvalue $\lambda_{\max}(\boldsymbol{Y})$. Let us call attention to the differences. The key advantage is that the current result depends on the intrinsic dimension of the matrix $\boldsymbol{M}$ instead of the ambient dimension. When the eigenvalues of $\boldsymbol{M}$ decay, the improvement can be dramatic. We do suffer a small cost in the extra factor of two, and the tail bound is restricted to a smaller range of the parameter ε. Neither of these limitations is particularly significant.

We have chosen to frame the result in terms of the upper bound $\boldsymbol{M}$ because it can be challenging to calculate the mean $\mathbb{E}\,\boldsymbol{Y}$ exactly. The statement here allows us to draw conclusions directly from the upper bound $\boldsymbol{M}$. These estimates do not follow formally from a result stated for $\mathbb{E}\,\boldsymbol{Y}$ because the intrinsic dimension is not monotone with respect to the semidefinite order.

A shortcoming of Theorem 7.2.1 is that it does not provide any information about $\lambda_{\min}(\boldsymbol{Y})$. Curiously, the approach we use to prove the result just does not work for the minimum eigenvalue.

7.2.2 Example: A Random Column Submatrix

To demonstrate the value of Theorem 7.2.1, let us return to one of the problems we studied in §5.2. We can now develop a refined estimate for the expected norm of a random column submatrix drawn from a fixed matrix.

In this example, we consider a fixed $m \times n$ matrix $\boldsymbol{B}$, and we let $\{\delta_k\}$ be an independent family of BERNOULLI(p/n) random variables. We form the random submatrix

$$\boldsymbol{Z} = \textstyle\sum_k \delta_k \boldsymbol{b}_{:k}\mathbf{e}_k^*$$

where $\boldsymbol{b}_{:k}$ is the kth column of $\boldsymbol{B}$. This random submatrix contains an average of p nonzero columns from $\boldsymbol{B}$. To study the norm of $\boldsymbol{Z}$, we consider the positive-semidefinite random matrix

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{Z}^* = \sum_{k=1}^{n} \delta_k \boldsymbol{b}_{:k}\boldsymbol{b}_{:k}^*.$$

This time, we invoke Theorem 7.2.1 to obtain a new estimate for the maximum eigenvalue of $\boldsymbol{Y}$.

We need a semidefinite bound $\boldsymbol{M}$ for the mean $\mathbb{E}\,\boldsymbol{Y}$ of the random matrix. In this case, the exact value is available:

$$\boldsymbol{M} = \mathbb{E}\,\boldsymbol{Y} = \frac{p}{n}\boldsymbol{B}\boldsymbol{B}^*.$$

We can easily calculate the intrinsic dimension of this matrix:

$$d = \operatorname{intdim}(\boldsymbol{M}) = \operatorname{intdim}\left(\frac{p}{n}\boldsymbol{B}\boldsymbol{B}^*\right) = \operatorname{intdim}(\boldsymbol{B}\boldsymbol{B}^*) = \frac{\operatorname{tr}(\boldsymbol{B}\boldsymbol{B}^*)}{\|\boldsymbol{B}\boldsymbol{B}^*\|} = \frac{\|\boldsymbol{B}\|_{\mathrm{F}}^2}{\|\boldsymbol{B}\|^2} = \operatorname{srank}(\boldsymbol{B}).$$

The second identity holds because the intrinsic dimension is scale invariant. The last relation is simply the definition (2.1.25) of the stable rank. The maximum eigenvalue of $\boldsymbol{M}$ verifies

$$\lambda_{\max}(\boldsymbol{M}) = \frac{p}{n}\lambda_{\max}(\boldsymbol{B}\boldsymbol{B}^*) = \frac{p}{n}\|\boldsymbol{B}\|^2.$$

The maximum norm L of any term in the sum $\boldsymbol{Y}$ satisfies $L = \max_k \|\boldsymbol{b}_{:k}\|^2$.

We may now apply the intrinsic Chernoff inequality. The expectation bound (7.2.1) with $\theta = 1$ delivers

$$\mathbb{E}\,\|\boldsymbol{Z}\|^2 = \mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \le 1.72\cdot\frac{p}{n}\cdot\|\boldsymbol{B}\|^2 + \log(2\operatorname{srank}(\boldsymbol{B}))\cdot\max_k\|\boldsymbol{b}_{:k}\|^2.$$

In the earlier analysis, we obtained a similar bound (5.2.1). The new result depends on the logarithm of the stable rank instead of $\log m$, the logarithm of the number of rows of $\boldsymbol{B}$. When the stable rank of $\boldsymbol{B}$ is small—meaning that many rows are almost collinear—then the revised estimate can result in a substantial improvement.

7.3 Matrix Bernstein with Intrinsic Dimension

Next, we present an extension of the matrix Bernstein inequality. These results provide tail bounds for an independent sum of bounded random matrices that depend on the intrinsic dimension of the variance. This theorem is essentially due to Stanislav Minsker.

Theorem 7.3.1 (Intrinsic Matrix Bernstein)**.** *Consider a finite sequence $\{\boldsymbol{S}_k\}$ of random complex matrices with the same size, and assume that*

$$\mathbb{E}\,\boldsymbol{S}_k = \boldsymbol{0} \quad \text{and} \quad \|\boldsymbol{S}_k\| \le L.$$

Introduce the random matrix

$$\boldsymbol{Z} = \textstyle\sum_k \boldsymbol{S}_k.$$

Let $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ be semidefinite upper bounds for the matrix-valued variances $\mathbf{Var}_1(\boldsymbol{Z})$ and $\mathbf{Var}_2(\boldsymbol{Z})$:

$$\begin{aligned}\boldsymbol{V}_1 &\succcurlyeq \mathbf{Var}_1(\boldsymbol{Z}) = \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*) = \textstyle\sum_k \mathbb{E}\left(\boldsymbol{S}_k\boldsymbol{S}_k^*\right), \quad \text{and}\\ \boldsymbol{V}_2 &\succcurlyeq \mathbf{Var}_2(\boldsymbol{Z}) = \mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z}) = \textstyle\sum_k \mathbb{E}\left(\boldsymbol{S}_k^*\boldsymbol{S}_k\right).\end{aligned}$$

Define an intrinsic dimension bound and a variance bound

$$d = \operatorname{intdim}\begin{bmatrix}\boldsymbol{V}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{V}_2\end{bmatrix} \quad \text{and} \quad v = \max\left\{\|\boldsymbol{V}_1\|, \|\boldsymbol{V}_2\|\right\}. \tag{7.3.1}$$

Then, for $t \ge \sqrt{v} + L/3$,

$$\mathbb{P}\{\|\boldsymbol{Z}\| \ge t\} \le 4d\exp\left(\frac{-t^2/2}{v + Lt/3}\right). \tag{7.3.2}$$

The proof of this result appears below in §7.7.

7.3.1 Discussion

Theorem 7.3.1 is quite similar to Theorem 6.1.1, so we focus on the differences. Although the statement of Theorem 7.3.1 may seem circumspect, it is important to present the result in terms of upper bounds $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ for the matrix-valued variances. Indeed, it can be challenging to calculate the matrix-valued variances exactly. The fact that the intrinsic dimension is not monotone interferes with our ability to use a simpler result.

Note that the tail bound (7.3.2) now depends on the intrinsic dimension of the block-diagonal matrix $\operatorname{diag}(\boldsymbol{V}_1, \boldsymbol{V}_2)$. This intrinsic dimension quantity never exceeds the total of the two side lengths of the random matrix $\boldsymbol{Z}$. As a consequence, the new tail bound always has a better dimensional dependence than the earlier result. The costs of this improvement are small: We pay an extra factor of four in the probability bound, and we must restrict our attention to a more limited range of the parameter t. Neither of these changes is significant.

The result does not contain an explicit estimate for $\mathbb{E}\,\|\boldsymbol{Z}\|$, but we can obtain such a bound by integrating the tail inequality (7.3.2). This estimate is similar with the earlier bound (6.1.3), but it depends on the intrinsic dimension instead of the ambient dimension.

Corollary 7.3.2 (Intrinsic Matrix Bernstein: Expectation Bound)**.** *Instate the notation and hypotheses of Theorem 7.3.1. Then*

$$\mathbb{E}\,\|\boldsymbol{Z}\| \le \text{Const} \cdot \left(\sqrt{v \log(1+d)} + L \log(1+d) \right). \tag{7.3.3}$$

See §7.7.4 for the proof.

Next, let us have a closer look at the intrinsic dimension quantity defined in (7.3.1).

$$d = \frac{\operatorname{tr} \boldsymbol{V}_1 + \operatorname{tr} \boldsymbol{V}_2}{\max\{\|\boldsymbol{V}_1\|, \|\boldsymbol{V}_2\|\}}.$$

We can make a further bound on the denominator to obtain an estimate in terms of the intrinsic dimensions of the two blocks:

$$\min\{\operatorname{intdim}(\boldsymbol{V}_1), \operatorname{intdim}(\boldsymbol{V}_2)\} \quad \le \quad d \quad \le \quad \operatorname{intdim}(\boldsymbol{V}_1) + \operatorname{intdim}(\boldsymbol{V}_2). \tag{7.3.4}$$

This bound reflects a curious phenomenon: the intrinsic dimension parameter d is not necessarily comparable with the larger of $\operatorname{intdim}(\boldsymbol{V}_1)$ or $\operatorname{intdim}(\boldsymbol{V}_2)$.

The other commentary about the original matrix Bernstein inequality, Theorem 6.1.1, also applies to the intrinsic dimension result. For example, we can adapt the result to a sum of uncentered, independent, random, bounded matrices. In addition, the theorem becomes somewhat simpler for a Hermitian random matrix because there is only one matrix-valued variance to deal with. The modifications required in these cases are straightforward.

7.3.2 Example: Matrix Approximation by Random Sampling

We can apply the intrinsic Bernstein inequality to study the behavior of randomized methods for matrix approximation. The following result is an immediate consequence of Theorem 7.3.1 and Corollary 7.3.2.

Corollary 7.3.3 (Matrix Approximation by Random Sampling: Intrinsic Dimension Bounds)**.** *Let $\boldsymbol{B}$ be a fixed $d_1 \times d_2$ matrix. Construct a $d_1 \times d_2$ random matrix $\boldsymbol{R}$ that satisfies*

$$\mathbb{E}\,\boldsymbol{R} = \boldsymbol{B} \quad \text{and} \quad \|\boldsymbol{R}\| \le L.$$

Let $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ be semidefinite upper bounds for the expected squares:

$$\boldsymbol{M}_1 \succcurlyeq \mathbb{E}(\boldsymbol{R}\boldsymbol{R}^*) \quad \text{and} \quad \boldsymbol{M}_2 \succcurlyeq \mathbb{E}(\boldsymbol{R}^*\boldsymbol{R}).$$

Define the quantities

$$d = \operatorname{intdim}\begin{bmatrix} \boldsymbol{M}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_2 \end{bmatrix} \quad \text{and} \quad m = \max\bigl\{\|\boldsymbol{M}_1\|, \|\boldsymbol{M}_2\|\bigr\}.$$

Form the matrix sampling estimator

$$\bar{\boldsymbol{R}}_n = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{R}_k \quad \text{where each } \boldsymbol{R}_k \text{ is an independent copy of } \boldsymbol{R}.$$

Then the estimator satisfies

$$\mathbb{E}\,\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\| \le \text{Const} \cdot \left[\sqrt{\frac{m\log(1+d)}{n}} + \frac{L\log(1+d)}{n}\right]. \tag{7.3.5}$$

Furthermore, for all $t \ge \sqrt{m} + L/3$,

$$\mathbb{P}\{\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\| \ge t\} \le 4d\exp\left(\frac{-nt^2/2}{m + 2Lt/3}\right). \tag{7.3.6}$$

The proof is similar with that of Corollary 6.2.1, so we omit the details.

7.3.3 Application: Randomized Matrix Multiplication

We will apply Corollary 7.3.3 to study the randomized matrix multiplication algorithm from §6.4. This method results in a small, but very appealing, improvement in the number of samples that are required. This argument is essentially due to Tassos Zouzias [Zou13].

Our goal is to approximate the product of a $d_1 \times N$ matrix $\boldsymbol{B}$ and an $N \times d_2$ matrix $\boldsymbol{C}$. We assume that both matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ have unit spectral norm. The results are stated in terms of the average stable rank

$$\mathsf{asr} = \tfrac{1}{2}(\operatorname{srank}(\boldsymbol{B}) + \operatorname{srank}(\boldsymbol{C})).$$

The stable rank was introduced in (2.1.25). To approximate the product $\boldsymbol{B}\boldsymbol{C}$, we constructed a simple random matrix $\boldsymbol{R}$ whose mean $\mathbb{E}\,\boldsymbol{R} = \boldsymbol{B}\boldsymbol{C}$, and then we formed the estimator

$$\bar{\boldsymbol{R}}_n = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{R}_k \quad \text{where each } \boldsymbol{R}_k \text{ is an independent copy of } \boldsymbol{R}.$$

The challenge is to bound the error $\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\boldsymbol{C}\|$.

To do so, let us refer back to our calculations from §6.4. We find that

$$\begin{aligned} \|\boldsymbol{R}\| &\le \mathsf{asr}, \\ \mathbb{E}(\boldsymbol{R}\boldsymbol{R}^*) &\preccurlyeq 2\cdot\mathsf{asr}\cdot\boldsymbol{B}\boldsymbol{B}^*, \quad \text{and} \\ \mathbb{E}(\boldsymbol{R}^*\boldsymbol{R}) &\preccurlyeq 2\cdot\mathsf{asr}\cdot\boldsymbol{C}^*\boldsymbol{C}. \end{aligned}$$

Starting from this point, we can quickly improve on our earlier analysis by incorporating the intrinsic dimension bounds.

It is natural to set $\boldsymbol{M}_1 = 2 \cdot \mathsf{asr} \cdot \boldsymbol{B}\boldsymbol{B}^*$ and $\boldsymbol{M}_2 = 2 \cdot \mathsf{asr} \cdot \boldsymbol{C}^*\boldsymbol{C}$. We may now bound the intrinsic dimension parameter

$$\begin{aligned} d &= \operatorname{intdim} \begin{bmatrix} \boldsymbol{M}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_2 \end{bmatrix} \le \operatorname{intdim}(\boldsymbol{M}_1) + \operatorname{intdim}(\boldsymbol{M}_2) \\ &= \frac{\operatorname{tr}(\boldsymbol{B}\boldsymbol{B}^*)}{\|\boldsymbol{B}\boldsymbol{B}^*\|} + \frac{\operatorname{tr}(\boldsymbol{C}^*\boldsymbol{C})}{\|\boldsymbol{C}^*\boldsymbol{C}\|} = \frac{\|\boldsymbol{B}\|_{\mathrm{F}}^2}{\|\boldsymbol{B}\|^2} + \frac{\|\boldsymbol{C}\|_{\mathrm{F}}^2}{\|\boldsymbol{C}\|^2} \\ &= \operatorname{srank}(\boldsymbol{B}) + \operatorname{srank}(\boldsymbol{C}) = 2 \cdot \mathsf{asr}. \end{aligned}$$

The first inequality follows from (7.3.4), and the second is Definition 7.1.1, of the intrinsic dimension. The third relation depends on the norm identities (2.1.8) and (2.1.24). Finally, we identify the stable ranks of $\boldsymbol{B}$ and $\boldsymbol{C}$ and the average stable rank. The calculation of the quantity m proceeds from the same considerations as in §6.4. Thus,

$$m = \max\left\{ \|\boldsymbol{M}_1\|, \|\boldsymbol{M}_2\| \right\} = 2 \cdot \mathsf{asr}.$$

This is all the information we need to collect.

Corollary 7.3.3 now implies that

$$\mathbb{E}\,\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\boldsymbol{C}\| \le \text{Const} \cdot \left(\sqrt{\frac{\mathsf{asr} \cdot \log(1+\mathsf{asr})}{n}} + \frac{\mathsf{asr} \cdot \log(1+\mathsf{asr})}{n} \right).$$

In other words, if the number n of samples satisfies

$$n \ge \varepsilon^{-2} \cdot \mathsf{asr} \cdot \log(1+\mathsf{asr}),$$

then the error satisfies

$$\mathbb{E}\,\|\bar{\boldsymbol{R}}_n - \boldsymbol{B}\boldsymbol{C}\| \le \text{Const} \cdot \left(\varepsilon + \varepsilon^2\right).$$

In the original analysis from §6.4, our estimate for the number n of samples contained the term $\log(d_1 + d_2)$ instead of $\log(1 + \mathsf{asr})$. We have replaced the dependence on the ambient dimension of the product $\boldsymbol{B}\boldsymbol{C}$ by a measure of the stable rank of the two factors. When the average stable rank is small in comparison with the dimension of the product, the analysis based on the intrinsic dimension offers an improvement in the bound on the number of samples required to approximate the product.

7.4 Revisiting the Matrix Laplace Transform Bound

Let us proceed with the proofs of the matrix concentration inequalities based on intrinsic dimension. The challenge is to identify and remedy the weak points in the arguments from Chapter 3.

After some reflection, we can trace the dependence on the ambient dimension in our earlier results to the proof of Proposition 3.2.1. In the original argument, we used an exponential function to transform the tail event before applying Markov's inequality. This approach leads to trouble for the simple reason that the exponential function does not pass through the origin, which gives undue weight to eigenvalues that are close to zero.

We can resolve this problem by using other types of maps to transform the tail event. The functions we have in mind are adjusted versions of the exponential. In particular, for fixed $\theta > 0$, we can consider

$$\psi_1(t) = \max\{0,\ \mathrm{e}^{\theta t} - 1\} \quad \text{and} \quad \psi_2(t) = \mathrm{e}^{\theta t} - \theta t - 1.$$

Both functions are nonnegative and convex, and they are nondecreasing on the positive real line. In each case, $\psi_i(0) = 0$. At the same time, the presence of the exponential function allows us to exploit our bounds for the trace mgf.

Proposition 7.4.1 (Generalized Matrix Laplace Transform Bound)**.** *Let $\boldsymbol{Y}$ be a random Hermitian matrix. Let $\psi : \mathbb{R} \to \mathbb{R}_+$ be a nonnegative function that is nondecreasing on $[0,\infty)$. For each $t \geq 0$,*

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \geq t\} \leq \frac{1}{\psi(t)}\,\mathbb{E}\operatorname{tr}\psi(\boldsymbol{Y}).$$

Proof. The proof follows the same lines as the proof of Proposition 3.2.1, but it requires some additional finesse. Since ψ is nondecreasing on $[0,\infty)$, the bound $a \geq t$ implies that $\psi(a) \geq \psi(t)$. As a consequence,

$$\lambda_{\max}(\boldsymbol{Y}) \geq t \quad \text{implies} \quad \lambda_{\max}(\psi(\boldsymbol{Y})) \geq \psi(t).$$

Indeed, on the tail event $\lambda_{\max}(\boldsymbol{Y}) \geq t$, we must have $\psi(\lambda_{\max}(\boldsymbol{Y})) \geq \psi(t)$. The Spectral Mapping Theorem, Proposition 2.1.3, indicates that the number $\psi(\lambda_{\max}(\boldsymbol{Y}))$ is one of the eigenvalues of the matrix $\psi(\boldsymbol{Y})$, so we determine that $\lambda_{\max}(\psi(\boldsymbol{Y}))$ also exceeds $\psi(t)$.

Returning to the tail probability, we discover that

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \geq t\} \leq \mathbb{P}\left\{\lambda_{\max}(\psi(\boldsymbol{Y})) \geq \psi(t)\right\} \leq \frac{1}{\psi(t)}\,\mathbb{E}\,\lambda_{\max}(\psi(\boldsymbol{Y})).$$

The second bound is Markov's inequality (2.2.1), which is valid because ψ is nonnegative. Finally,

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \geq t\} \leq \frac{1}{\psi(t)}\,\mathbb{E}\operatorname{tr}\psi(\boldsymbol{Y}).$$

The inequality holds because of the fact (2.1.13) that the trace of $\psi(\boldsymbol{Y})$, a positive-semidefinite matrix, must be at least as large as its maximum eigenvalue. □

7.5 The Intrinsic Dimension Lemma

The other new ingredient is a simple observation that allows us to control a trace function applied to a positive-semidefinite matrix in terms of the intrinsic dimension of the matrix.

Lemma 7.5.1 (Intrinsic Dimension)**.** *Let φ be a convex function on the interval $[0,\infty)$, and assume that $\varphi(0) = 0$. For any positive-semidefinite matrix $\boldsymbol{A}$, it holds that*

$$\operatorname{tr}\varphi(\boldsymbol{A}) \leq \operatorname{intdim}(\boldsymbol{A}) \cdot \varphi(\|\boldsymbol{A}\|).$$

Proof. Since the function $a \mapsto \varphi(a)$ is convex on the interval $[0, L]$, it is bounded above by the chord connecting the graph at the endpoints. That is, for $a \in [0, L]$,

$$\varphi(a) \leq \left(1 - \frac{a}{L}\right) \cdot \varphi(0) + \frac{a}{L} \cdot \varphi(L) = \frac{a}{L} \cdot \varphi(L).$$

The eigenvalues of $\boldsymbol{A}$ fall in the interval $[0, L]$, where $L = \|\boldsymbol{A}\|$. As an immediate consequence of the Transfer Rule (2.1.14), we find that

$$\operatorname{tr} \varphi(\boldsymbol{A}) \le \frac{\operatorname{tr} \boldsymbol{A}}{\|\boldsymbol{A}\|} \cdot \varphi(\|\boldsymbol{A}\|).$$

Identify the intrinsic dimension of $\boldsymbol{A}$ to complete the argument. □

7.6 Proof of the Intrinsic Chernoff Bound

With these results at hand, we are prepared to prove our first intrinsic dimension result, which extends the matrix Chernoff inequality.

Proof of Theorem 7.2.1. Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random Hermitian matrices with

$$0 \le \lambda_{\min}(\boldsymbol{X}_k) \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le L \quad \text{for each index } k.$$

Introduce the sum

$$\boldsymbol{Y} = \textstyle\sum_k \boldsymbol{X}_k.$$

The challenge is to establish bounds for $\lambda_{\max}(\boldsymbol{Y})$ that depend on the intrinsic dimension of a matrix $\boldsymbol{M}$ that satisfies $\boldsymbol{M} \succcurlyeq \mathbb{E}\,\boldsymbol{Y}$. We begin the argument with the proof of the tail bound (7.2.2). Afterward, we show how to extract the expectation bound (7.2.1).

Fix a number $\theta > 0$, and define the function $\psi(t) = \max\{0,\ \mathrm{e}^{\theta t} - 1\}$ for $t \in \mathbb{R}$. For $t \ge 0$, the general version of the matrix Laplace transform bound, Proposition 7.4.1, states that

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} \le \frac{1}{\psi(t)}\,\mathbb{E}\operatorname{tr}\psi(\boldsymbol{Y}) = \frac{1}{\mathrm{e}^{\theta t} - 1}\,\mathbb{E}\operatorname{tr}\bigl(\mathrm{e}^{\theta \boldsymbol{Y}} - \mathbf{I}\bigr). \tag{7.6.1}$$

We have exploited the fact that $\boldsymbol{Y}$ is positive semidefinite and the assumption that $t \ge 0$. The presence of the identity matrix on the right-hand side allows us to draw stronger conclusions than we could before.

Let us study the expected trace term on the right-hand side of (7.6.1). As in the proof of the original matrix Chernoff bound, Theorem 5.1.1, we have the estimate

$$\mathbb{E}\operatorname{tr}\mathrm{e}^{\theta \boldsymbol{Y}} \le \operatorname{tr}\exp\bigl(g(\theta) \cdot \mathbb{E}\,\boldsymbol{Y}\bigr) \quad \text{where} \quad g(\theta) = \frac{\mathrm{e}^{\theta L} - 1}{L}.$$

Introduce the function $\varphi(a) = \mathrm{e}^a - 1$, and observe that

$$\mathbb{E}\operatorname{tr}\bigl(\mathrm{e}^{\theta \boldsymbol{Y}} - \mathbf{I}\bigr) \le \operatorname{tr}\bigl(\mathrm{e}^{g(\theta)\cdot \mathbb{E}\,\boldsymbol{Y}} - \mathbf{I}\bigr) \le \operatorname{tr}\bigl(\mathrm{e}^{g(\theta)\cdot \boldsymbol{M}} - \mathbf{I}\bigr) = \operatorname{tr}\varphi(g(\theta)\cdot \boldsymbol{M}).$$

The second inequality follows from the assumption that $\mathbb{E}\,\boldsymbol{Y} \preccurlyeq \boldsymbol{M}$ and the monotonicity (2.1.16) of the trace exponential. Now, apply the intrinsic dimension bound, Lemma 7.5.1, to reach

$$\mathbb{E}\operatorname{tr}\bigl(\mathrm{e}^{\theta \boldsymbol{Y}} - \mathbf{I}\bigr) \le \operatorname{intdim}(\boldsymbol{M}) \cdot \varphi\bigl(g(\theta)\,\|\boldsymbol{M}\|\bigr).$$

We have used the fact that the intrinsic dimension does not depend on the scaling factor $g(\theta)$. Recalling the notation $d = \operatorname{intdim}(\boldsymbol{M})$ and $\mu_{\max} = \|\boldsymbol{M}\|$, we continue the calculation:

$$\mathbb{E}\operatorname{tr}\bigl(\mathrm{e}^{\theta \boldsymbol{Y}} - \mathbf{I}\bigr) \le d \cdot \varphi\bigl(g(\theta) \cdot \mu_{\max}\bigr) \le d \cdot \mathrm{e}^{g(\theta)\cdot \mu_{\max}}. \tag{7.6.2}$$

We have invoked the trivial inequality $\varphi(a) \le \mathrm{e}^a$, which holds for $a \in \mathbb{R}$.

Next, introduce the bound (7.6.2) on the expected trace into the probability bound (7.6.1) to obtain

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} \le d \cdot \frac{\mathrm{e}^{\theta t}}{\mathrm{e}^{\theta t} - 1} \cdot \mathrm{e}^{-\theta t + g(\theta)\cdot\mu_{\max}} \le d \cdot \left(1 + \frac{1}{\theta t}\right) \cdot \mathrm{e}^{-\theta t + g(\theta)\cdot\mu_{\max}}. \tag{7.6.3}$$

To control the fraction, we have observed that

$$\frac{\mathrm{e}^a}{\mathrm{e}^a - 1} = 1 + \frac{1}{\mathrm{e}^a - 1} \le 1 + \frac{1}{a} \quad \text{for } a \ge 0.$$

We obtain the latter inequality by replacing the convex function $a \mapsto \mathrm{e}^a - 1$ with its tangent line at $a = 0$.

In the estimate (7.6.3), we make the change of variables $t \mapsto (1+\varepsilon)\mu_{\max}$. The bound is valid for all $\theta > 0$, so we can select $\theta = L^{-1}\log(1+\varepsilon)$ to minimize the exponential. Altogether, these steps lead to the estimate

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \ge (1+\varepsilon)\mu_{\max}\right\} \le d \cdot \left(1 + \frac{L/\mu_{\max}}{(1+\varepsilon)\log(1+\varepsilon)}\right) \cdot \left[\frac{\mathrm{e}^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right]^{\mu_{\max}/L}. \tag{7.6.4}$$

Now, instate the assumption that $\varepsilon \ge L/\mu_{\max}$. The function $a \mapsto (1+a)\log(1+a)$ is convex when $a \ge -1$, so we can bound it below using its tangent at $\varepsilon = 0$. Thus,

$$(1+\varepsilon)\log(1+\varepsilon) \ge \varepsilon \ge \frac{L}{\mu_{\max}}.$$

It follows that the parenthesis in (7.6.4) is bounded by two, which yields the conclusion (7.2.2).

Now, we turn to the expectation bound (7.2.1). Observe that the functional inverse of ψ is the increasing concave function

$$\psi^{-1}(u) = \frac{1}{\theta}\log(1+u) \quad \text{for } u \ge 0.$$

Since $\boldsymbol{Y}$ is a positive-semidefinite matrix, we can calculate that

$$\begin{aligned}\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) = \mathbb{E}\,\psi^{-1}(\psi(\lambda_{\max}(\boldsymbol{Y}))) &\le \psi^{-1}(\mathbb{E}\,\psi(\lambda_{\max}(\boldsymbol{Y}))) \\ &= \psi^{-1}(\mathbb{E}\,\lambda_{\max}(\psi(\boldsymbol{Y}))) \le \psi^{-1}(\mathbb{E}\,\operatorname{tr}\psi(\boldsymbol{Y})).\end{aligned} \tag{7.6.5}$$

The second relation is Jensen's inequality (2.2.2), which is valid because ψ^{-1} is concave. The third relation follows from the Spectral Mapping Theorem, Proposition 2.1.3, because the function ψ is increasing. We can bound the maximum eigenvalue by the trace because $\psi(\boldsymbol{Y})$ is positive semidefinite and ψ^{-1} is an increasing function.

Now, substitute the bound (7.6.2) into the last display (7.6.5) to reach

$$\begin{aligned}\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \le \psi^{-1}(d\cdot\exp(g(\theta)\cdot\mu_{\max})) &= \frac{1}{\theta}\log\bigl(1 + d\cdot\mathrm{e}^{g(\theta)\cdot\mu_{\max}}\bigr) \\ &\le \frac{1}{\theta}\log\bigl(2d\cdot\mathrm{e}^{g(\theta)\cdot\mu_{\max}}\bigr) = \frac{1}{\theta}\bigl(\log(2d) + g(\theta)\cdot\mu_{\max}\bigr).\end{aligned}$$

The first inequality again requires the property that ψ^{-1} is increasing. The second inequality follows because $1 \le d\cdot\mathrm{e}^{g(\theta)\cdot\mu_{\max}}$, which is a consequence of the fact that the intrinsic dimension exceeds one and the exponent is nonnegative. To complete the argument, introduce the definition of $g(\theta)$, and make the change of variables $\theta \mapsto \theta/L$. These steps yield (7.2.1). □

7.7 Proof of the Intrinsic Bernstein Bounds

In this section, we present the arguments that lead up to the intrinsic Bernstein bounds. That is, we develop tail inequalities for an independent sum of bounded random matrices that depend on the intrinsic dimension of the variance.

7.7.1 The Hermitian Case

As usual, Hermitian matrices provide the natural setting for matrix concentration. We begin with an explicit statement and proof of a bound for the Hermitian case.

Theorem 7.7.1 (Matrix Bernstein: Hermitian Case with Intrinsic Dimension)**.** *Consider a finite sequence* $\{\boldsymbol{X}_k\}$ *of random Hermitian matrices of the same size, and assume that*

$$\mathbb{E}\,\boldsymbol{X}_k = \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le L \quad \text{for each index } k.$$

Introduce the random matrix

$$\boldsymbol{Y} = \sum\nolimits_k \boldsymbol{X}_k.$$

Let $\boldsymbol{V}$ *be a semidefinite upper bound for the matrix-valued variance* $\mathbf{Var}(\boldsymbol{Y})$*:*

$$\boldsymbol{V} \succcurlyeq \mathbf{Var}(\boldsymbol{Y}) = \mathbb{E}\,\boldsymbol{Y}^2 = \sum\nolimits_k \mathbb{E}\,\boldsymbol{X}_k^2.$$

Define the intrinsic dimension bound and variance bound

$$d = \operatorname{intdim}(\boldsymbol{V}) \quad \text{and} \quad v = \|\boldsymbol{V}\|.$$

Then, for $t \ge \sqrt{v} + L/3$*,*

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} \le 4d \cdot \exp\left(\frac{-t^2/2}{v + Lt/3}\right). \tag{7.7.1}$$

The proof of this result appears in the next section.

7.7.2 Proof of the Hermitian Case

We commence with the results for an independent sum of random Hermitian matrices whose eigenvalues are subject to an upper bound.

Proof of Theorem 7.7.1. Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, Hermitian matrices with

$$\mathbb{E}\,\boldsymbol{X}_k = \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le L \quad \text{for each index } k.$$

Introduce the random matrix

$$\boldsymbol{Y} = \sum\nolimits_k \boldsymbol{X}_k.$$

Our goal is to obtain a tail bound for $\lambda_{\max}(\boldsymbol{Y})$ that reflects the intrinsic dimension of a matrix $\boldsymbol{V}$ that satisfies $\boldsymbol{V} \succcurlyeq \mathbf{Var}(\boldsymbol{Y})$.

Fix a number $\theta > 0$, and define the function $\psi(t) = \mathrm{e}^{\theta t} - \theta t - 1$ for $t \in \mathbb{R}$. The general version of the matrix Laplace transform bound, Proposition 7.4.1, implies that

$$\begin{aligned} \mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \geq t\} &\leq \frac{1}{\psi(t)} \mathbb{E}\operatorname{tr}\psi(\boldsymbol{Y}) \\ &= \frac{1}{\psi(t)} \mathbb{E}\operatorname{tr}\bigl(\mathrm{e}^{\theta \boldsymbol{Y}} - \theta \boldsymbol{Y} - \mathbf{I}\bigr) \\ &= \frac{1}{\mathrm{e}^{\theta t} - \theta t - 1} \mathbb{E}\operatorname{tr}\bigl(\mathrm{e}^{\theta \boldsymbol{Y}} - \mathbf{I}\bigr). \end{aligned} \tag{7.7.2}$$

The last identity holds because the random matrix $\boldsymbol{Y}$ has zero mean.

Let us focus on the expected trace on the right-hand side of (7.7.2). Examining the proof of the original matrix Bernstein bound for Hermitian matrices, Theorem 6.6.1, we see that

$$\mathbb{E}\operatorname{tr}\mathrm{e}^{\theta \boldsymbol{Y}} \leq \operatorname{tr}\exp\bigl(g(\theta) \cdot \mathbb{E}\,\boldsymbol{Y}^2\bigr) \quad \text{where} \quad g(\theta) = \exp\left(\frac{\theta^2/2}{1 - \theta L/3}\right).$$

Introduce the function $\varphi(a) = \mathrm{e}^a - 1$, and observe that

$$\mathbb{E}\operatorname{tr}\bigl(\mathrm{e}^{\theta \boldsymbol{Y}} - \mathbf{I}\bigr) \leq \operatorname{tr}\bigl(\mathrm{e}^{g(\theta) \cdot \mathbb{E}\,\boldsymbol{Y}^2} - \mathbf{I}\bigr) \leq \operatorname{tr}\bigl(\mathrm{e}^{g(\theta) \cdot \boldsymbol{V}} - \mathbf{I}\bigr) = \operatorname{tr}\varphi\bigl(g(\theta) \cdot \boldsymbol{V}\bigr).$$

The second inequality depends on the assumption that $\mathbb{E}\,\boldsymbol{Y}^2 = \mathbf{Var}(\boldsymbol{Y}) \preccurlyeq \boldsymbol{V}$ and the monotonicity property (2.1.16) of the trace exponential. Apply the intrinsic dimension bound, Lemma 7.5.1, to reach

$$\mathbb{E}\operatorname{tr}\bigl(\mathrm{e}^{\theta \boldsymbol{Y}} - \mathbf{I}\bigr) \leq \operatorname{intdim}(\boldsymbol{V}) \cdot \varphi\bigl(g(\theta) \cdot \|\boldsymbol{V}\|\bigr) = d \cdot \varphi\bigl(g(\theta) \cdot v\bigr) \leq d \cdot \mathrm{e}^{g(\theta) \cdot v}. \tag{7.7.3}$$

We have used the fact that the intrinsic dimension is scale invariant. Then we identified $d = \operatorname{intdim}(\boldsymbol{V})$ and $v = \|\boldsymbol{V}\|$. The last inequality depends on the trivial estimate $\varphi(a) \leq \mathrm{e}^a$, valid for all $a \in \mathbb{R}$.

Substitute the bound (7.7.3) into the probability inequality (7.7.2) to discover that

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \geq t\} \leq d \cdot \frac{\mathrm{e}^{\theta t}}{\mathrm{e}^{\theta t} - \theta t - 1} \cdot \mathrm{e}^{-\theta t + g(\theta) \cdot v} \leq d \cdot \left(1 + \frac{3}{\theta^2 t^2}\right) \cdot \mathrm{e}^{-\theta t + g(\theta) \cdot v}. \tag{7.7.4}$$

This estimate holds for any positive value of θ. To control the fraction, we have observed that

$$\frac{\mathrm{e}^a}{\mathrm{e}^a - a - 1} = 1 + \frac{1 + a}{\mathrm{e}^a - a - 1} \leq 1 + \frac{3}{a^2} \quad \text{for all } a \geq 0.$$

The inequality above is a consequence of the numerical fact

$$\frac{\mathrm{e}^a - a - 1}{a^2} - \frac{1 + a}{3} > 0 \quad \text{for all } a \in \mathbb{R}.$$

Indeed, the left-hand side of the latter expression defines a convex function of a, whose minimal value, attained near $a \approx 1.30$, is strictly positive.

In the tail bound (7.7.4), we select $\theta = t/(v + Lt/3)$ to reach

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \geq t\} \leq d \cdot \left(1 + 3 \cdot \frac{(v + Lt/3)^2}{t^4}\right) \cdot \exp\left(\frac{-t^2/2}{v + Lt/3}\right).$$

This probability inequality is typically vacuous when $t^2 < v + Lt/3$, so we may as well limit our attention to the case where $t^2 \geq v + Lt/3$. Under this assumption, the parenthesis is bounded by four, which gives the tail bound (7.7.1). We can simplify the restriction on t by solving the quadratic inequality to obtain the sufficient condition

$$t \geq \frac{1}{2}\left[\frac{L}{3} + \sqrt{\frac{L^2}{9} + 4v}\right].$$

We develop an upper bound for the right-hand side of this inequality as follows.

$$\frac{1}{2}\left[\frac{L}{3} + \sqrt{\frac{L^2}{9} + 4v}\right] = \frac{L}{6}\left[1 + \sqrt{1 + \frac{36v}{L^2}}\right] \leq \frac{L}{6}\left[1 + 1 + \frac{6\sqrt{v}}{L}\right] = \sqrt{v} + \frac{L}{3}.$$

We have used the numerical fact $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \geq 0$. Therefore, the tail bound (7.7.1) is valid when $t \geq \sqrt{v} + L/3$. □

7.7.3 Proof of the Rectangular Case

Finally, we present the proof of the intrinsic Bernstein inequality, Theorem 7.3.1, for general random matrices.

Proof of Theorem 7.3.1. Suppose that $\{\boldsymbol{S}_k\}$ is a finite sequence of independent random matrices that satisfy

$$\mathbb{E}\,\boldsymbol{S}_k = \mathbf{0} \quad \text{and} \quad \|\boldsymbol{S}_k\| \leq L \quad \text{for each index } k.$$

Form the sum $\boldsymbol{Z} = \sum_k \boldsymbol{S}_k$. As in the proof of Theorem 6.1.1, we derive the result by applying Theorem 7.7.1 to the Hermitian dilation $\boldsymbol{Y} = \mathscr{H}(\boldsymbol{Z})$. The only new point that requires attention is the modification to the intrinsic dimension and variance terms.

Recall the calculation of the variance of the dilation from (2.2.9):

$$\mathbb{E}\,\boldsymbol{Y}^2 = \mathbb{E}\,\mathscr{H}(\boldsymbol{Z})^2 = \begin{bmatrix} \mathbf{Var}_1(\boldsymbol{Z}) & \mathbf{0} \\ \mathbf{0} & \mathbf{Var}_2(\boldsymbol{Z}) \end{bmatrix} \preccurlyeq \begin{bmatrix} \boldsymbol{V}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{V}_2 \end{bmatrix} = \boldsymbol{V}.$$

The semidefinite inequality follows from our assumptions on $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$. Therefore, the intrinsic dimension quantity in Theorem 7.7.1 induces the definition in the general case:

$$d = \operatorname{intdim}(\boldsymbol{V}) = \operatorname{intdim}\begin{bmatrix} \boldsymbol{V}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{V}_2 \end{bmatrix}$$

Similarly,

$$v = \|\boldsymbol{V}\| = \max\left\{\|\boldsymbol{V}_1\|, \|\boldsymbol{V}_2\|\right\}.$$

This point completes the argument. □

7.7.4 Proof of the Intrinsic Bernstein Expectation Bound

Finally, let us establish the expectation bound, Corollary 7.3.2, that accompanies Theorem 7.3.1.

Proof of Corollary 7.3.2. Fix a number $\mu \geq \sqrt{v}$. We may rewrite the expectation of $\|\boldsymbol{Z}\|$ as an integral:

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{Z}\| &= \int_0^\infty \mathbb{P}\{\|\boldsymbol{Z}\| \geq t\}\,\mathrm{d}t \leq \mu + 4d\int_\mu^\infty \exp\left(\frac{-t^2/2}{v + Lt/3}\right)\mathrm{d}t \\ &\leq \mu + 4d\int_\mu^\infty \mathrm{e}^{-t^2/(2v)}\,\mathrm{d}t + 4d\int_\mu^\infty \mathrm{e}^{-3t/(2L)}\,\mathrm{d}t \\ &\leq \mu + 4d\sqrt{v}\,\mathrm{e}^{-\mu^2/(2v)} + \frac{8}{3}dL\,\mathrm{e}^{-3\mu/(2L)}.\end{aligned}$$

To obtain the first inequality, we split the integral at μ, and we bound the probability by one on the domain of integration $[0, \mu]$. The second inequality holds because

$$\exp\left(\frac{-t^2/2}{v + Lt/3}\right) \leq \max\left\{\mathrm{e}^{-t^2/(2v)},\ \mathrm{e}^{-3t/(2L)}\right\} \leq \mathrm{e}^{-t^2/(2v)} + \mathrm{e}^{-3t/(2L)}.$$

We controlled the Gaussian integral by inserting the factor $\sqrt{v}\,(t/v) \geq 1$ into the integrand:

$$\int_\mu^\infty \mathrm{e}^{-t^2/(2v)}\,\mathrm{d}t \leq \sqrt{v}\int_\mu^\infty (t/v)\,\mathrm{e}^{-t^2/(2v)}\,\mathrm{d}t = \sqrt{v}\,\mathrm{e}^{-\mu^2/(2v)}.$$

To complete the argument, select $\mu = \sqrt{2v\log(1+d)} + \frac{2}{3}L\log(1+d)$ to reach

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{Z}\| &\leq \mu + 4d\sqrt{v}\,\mathrm{e}^{-(2v\log(1+d))/(2v)} + \frac{8}{3}dL\,\mathrm{e}^{-3((2/3)L\log(1+d))/(2L)} \\ &\leq \sqrt{2v\log(1+d)} + \frac{2}{3}L\log(1+d) + 4\sqrt{v} + \frac{8}{3}L.\end{aligned}$$

The stated bound (7.3.3) follows after we combine terms and agglomerate constants. □

7.8 Notes

At present, there are two different ways to improve the dimensional factor that appears in matrix concentration inequalities.

First, there is a sequence of matrix concentration results where the dimensional parameter is bounded by the maximum rank of the random matrix. The first bound of this type is due to Rudelson [Rud99]. Oliveira's results in [Oli10b] also exhibit this reduced dimensional dependence. A subsequent paper [MZ11] by Magen & Zouzias contains a related argument that gives similar results. We do not discuss this class of bounds here.

The idea that the dimensional factor should depend on metric properties of the random matrix appears in a paper of Hsu, Kakade, & Zhang [HKZ12]. They obtain a bound that is similar with Theorem 7.7.1. Unfortunately, their argument is complicated, and the results it delivers are suboptimal.

Theorem 7.7.1 is essentially due to Stanislav Minsker [Min11]. His approach leads to somewhat sharper bounds than the approach in the paper of Hsu, Kakade, & Zhang, and his method is easier to understand.

These notes contain another approach to intrinsic dimension bounds. The intrinsic Chernoff bounds that emerge from our framework are new. The proof of the intrinsic Bernstein bound, Theorem 7.7.1, can be interpreted as a distillation of Minsker's argument. Indeed, many of the specific calculations already appear in Minsker's paper. We have obtained constants that are marginally better.

CHAPTER 8

A Proof of Lieb's Theorem

Our approach to random matrices depends on some sophisticated ideas that are not usually presented in linear algebra courses. This chapter contains a complete derivation of the results that undergird our matrix concentration inequalities. We begin with a short argument that explains how Lieb's Theorem follows from deep facts about a function called the matrix relative entropy. The balance of the chapter is devoted to an analysis of the matrix relative entropy. Along the way, we establish the core properties of the trace exponential function and the matrix logarithm. This discussion may serve as an introduction to the advanced techniques of matrix analysis.

8.1 Lieb's Theorem

In his 1973 paper on trace functions, Lieb established an important concavity theorem [Lie73, Thm. 6] for the trace exponential function. As we saw in Chapter 3, this result animates all of our matrix concentration inequalities.

Theorem 8.1.1 (Lieb)**.** *Let $\boldsymbol{H}$ be a fixed Hermitian matrix with dimension d. The map*

$$\boldsymbol{A} \longmapsto \operatorname{tr} \exp\bigl(\boldsymbol{H} + \log \boldsymbol{A}\bigr) \tag{8.1.1}$$

is concave on the convex cone of $d \times d$ positive-definite Hermitian matrices.

Section 8.1 contains an overview of the proof of Theorem 8.1.1. First, we state the background material that we require, and then we show how the theorem follows. Some of the supporting results are major theorems in their own right, and the details of their proofs will consume the rest of the chapter.

8.1.1 Conventions

The symbol $\mathbb{R}_{++}$ refers to the set of positive real numbers. We remind the reader of our convention that bold capital letters that are symmetric about the vertical axis ($\boldsymbol{A}, \boldsymbol{H}, \boldsymbol{M}, \boldsymbol{T}, \boldsymbol{U}, \boldsymbol{\Psi}$) always refer to Hermitian matrices. We reserve the letter $\mathbf{I}$ for the identity matrix, while the letter $\boldsymbol{Q}$ always refers to a unitary matrix. Other bold capital letters ($\boldsymbol{B}, \boldsymbol{K}, \boldsymbol{L}$) denote rectangular matrices.

Unless stated otherwise, the results in this chapter hold for all matrices whose dimensions are compatible. For example, any result that involves a sum $\boldsymbol{A} + \boldsymbol{H}$ includes the implicit constraint that the two matrices are the same size.

Throughout this chapter, we assume that the parameter $\tau \in [0, 1]$, and we use the shorthand $\bar{\tau} = 1 - \tau$ to make formulas involving convex combinations more legible.

8.1.2 Matrix Relative Entropy

The proof of Lieb's Theorem depends on the properties of a bivariate function called the *matrix relative entropy*.

Definition 8.1.2 (Matrix Relative Entropy)**.** *Let $\boldsymbol{A}$ and $\boldsymbol{H}$ be positive-definite matrices of the same size. The entropy of $\boldsymbol{A}$ relative to $\boldsymbol{H}$ is*

$$\mathrm{D}(\boldsymbol{A}; \boldsymbol{H}) = \operatorname{tr}\big[\boldsymbol{A}(\log \boldsymbol{A} - \log \boldsymbol{H}) - (\boldsymbol{A} - \boldsymbol{H})\big].$$

The relative entropy can be viewed as a measure of the difference between the matrix $\boldsymbol{A}$ and the matrix $\boldsymbol{H}$, but it is not a metric. Related functions arise in quantum statistical mechanics and quantum information theory.

We need two facts about the matrix relative entropy.

Proposition 8.1.3 (Matrix Relative Entropy is Nonnegative)**.** *For positive-definite matrices $\boldsymbol{A}$ and $\boldsymbol{H}$ of the same size, the matrix relative entropy* $\mathrm{D}(\boldsymbol{A}; \boldsymbol{H}) \geq 0$.

Proposition 8.1.3 is easy to prove; see Section 8.3.5 for the short argument.

Theorem 8.1.4 (The Matrix Relative Entropy is Convex)**.** *The map $(\boldsymbol{A}, \boldsymbol{H}) \mapsto \mathrm{D}(\boldsymbol{A}; \boldsymbol{H})$ is convex. That is, for positive-definite $\boldsymbol{A}_i$ and $\boldsymbol{H}_i$ of the same size,*

$$\mathrm{D}\big(\tau \boldsymbol{A}_1 + \bar{\tau} \boldsymbol{A}_2; \tau \boldsymbol{H}_1 + \bar{\tau} \boldsymbol{H}_2\big) \leq \tau \cdot \mathrm{D}\big(\boldsymbol{A}_1; \boldsymbol{H}_1\big) + \bar{\tau} \cdot \mathrm{D}\big(\boldsymbol{A}_2; \boldsymbol{H}_2\big) \quad \text{for } \tau \in [0, 1].$$

Theorem 8.1.4 is one of the crown jewels of matrix analysis. The supporting material for this result occupies the bulk of this chapter; the argument culminates in Section 8.8.

8.1.3 Partial Maximization

We also require a basic fact from convex analysis which states that partial maximization of a concave function produces a concave function. We include the simple proof.

Fact 8.1.5 (Partial Maximization)**.** *Let f be a concave function of two variables. Then the function $y \mapsto \sup_x f(x; y)$ obtained by partial maximization is concave.*

Proof. Fix $\varepsilon > 0$. For each pair of points y_1 and y_2, there are points x_1 and x_2 that satisfy

$$f(x_1; y_1) \geq \sup_x f(x; y_1) - \varepsilon \quad \text{and} \quad f(x_2; y_2) \geq \sup_x f(x; y_2) - \varepsilon.$$

For each $\tau \in [0, 1]$, the concavity of f implies that

$$\begin{aligned} \sup_x f(x;\, \tau y_1 + \bar{\tau} y_2) &\geq f(\tau x_1 + \bar{\tau} x_2;\, \tau y_1 + \bar{\tau} y_2) \\ &\geq \tau \cdot f(x_1; y_1) + \bar{\tau} \cdot f(x_2; y_2) \\ &\geq \tau \cdot \sup_x f(x; y_1) + \bar{\tau} \cdot \sup_x f(x; y_2) - \varepsilon. \end{aligned}$$

Take the limit as $\varepsilon \downarrow 0$ to see that the partial supremum is a concave function of y. □

8.1.4 A Proof of Lieb's Theorem

Taking the results about the matrix relative entropy for granted, it is not hard to prove Lieb's Theorem. We begin with a variational representation of the trace, which restates the fact that matrix relative entropy is nonnegative.

Lemma 8.1.6 (Variational Formula for Trace)**.** *Let $\boldsymbol{M}$ be a positive-definite matrix. Then*

$$\operatorname{tr} \boldsymbol{M} = \sup_{\boldsymbol{T} > \mathbf{0}} \operatorname{tr}\big[\boldsymbol{T} \log \boldsymbol{M} - \boldsymbol{T} \log \boldsymbol{T} + \boldsymbol{T}\big].$$

Proof. Proposition (8.1.3) states that $\mathrm{D}(\boldsymbol{T}; \boldsymbol{M}) \geq 0$. Introduce the definition of the matrix relative entropy, and rearrange to reach

$$\operatorname{tr} \boldsymbol{M} \geq \operatorname{tr}\big[\boldsymbol{T} \log \boldsymbol{M} - \boldsymbol{T} \log \boldsymbol{T} + \boldsymbol{T}\big].$$

When $\boldsymbol{T} = \boldsymbol{M}$, both sides are equal, which yields the advertised identity. □

To establish Lieb's Theorem, we use the variational formula to represent the trace exponential. Then we use the partial maximization result to condense the desired concavity property from the convexity of the matrix relative entropy.

Proof of Theorem 8.1.1. In the variational formula, Lemma 8.1.6, select $\boldsymbol{M} = \exp(\boldsymbol{H} + \log \boldsymbol{A})$ to obtain

$$\operatorname{tr} \exp(\boldsymbol{H} + \log \boldsymbol{A}) = \sup_{\boldsymbol{T} > \mathbf{0}} \operatorname{tr}\big[\boldsymbol{T}(\boldsymbol{H} + \log \boldsymbol{A}) - \boldsymbol{T} \log \boldsymbol{T} + \boldsymbol{T}\big]$$

The latter expression can be written compactly using the matrix relative entropy:

$$\operatorname{tr} \exp(\boldsymbol{H} + \log \boldsymbol{A}) = \sup_{\boldsymbol{T} > \mathbf{0}} \big[\operatorname{tr}(\boldsymbol{T}\boldsymbol{H}) + \operatorname{tr} \boldsymbol{A} - \mathrm{D}(\boldsymbol{T}; \boldsymbol{A})\big] \tag{8.1.2}$$

For each Hermitian matrix $\boldsymbol{H}$, the bracket is a concave function of the pair $(\boldsymbol{T}, \boldsymbol{A})$ because of Theorem 8.1.4. We see that the right-hand side of (8.1.2) is the partial maximum of a concave function, and Fact 8.1.5 ensures that this expression defines a concave function of $\boldsymbol{A}$. This observation establishes the theorem. □

8.2 Analysis of the Relative Entropy for Vectors

Many deep theorems about matrices have analogies for vectors. This observation is valuable because we can usually adapt an analysis from the vector setting to establish the parallel result for matrices. In the matrix setting, however, it may be necessary to install a significant amount of extra machinery. If we keep the simpler structure of the vector argument in mind, we can avoid being crushed in the gears.

8.2.1 The Relative Entropy for Vectors

The goal of §8.2 is to introduce the relative entropy function for positive vectors and to derive some key properties of this function. Later we will analyze the matrix relative entropy by emulating these arguments.

Definition 8.2.1 (Relative Entropy)**.** *Let $\boldsymbol{a}$ and $\boldsymbol{h}$ be positive vectors of the same size. The entropy of $\boldsymbol{a}$ relative to $\boldsymbol{h}$ is defined as*

$$\mathrm{D}(\boldsymbol{a};\boldsymbol{h}) = \sum\nolimits_k \big[a_k(\log a_k - \log h_k) - (a_k - h_k)\big].$$

A variant of the relative entropy arises in information theory and statistics as a measure of the discrepancy between two probability distributions on a finite set. We will show that the relative entropy is nonnegative and convex.

It may seem abusive to recycle the notation for the relative entropy on matrices. To justify this decision, we observe that

$$\mathrm{D}(\boldsymbol{a};\boldsymbol{h}) = \mathrm{D}\big(\operatorname{diag}(\boldsymbol{a});\, \operatorname{diag}(\boldsymbol{h})\big)$$

where $\operatorname{diag}(\cdot)$ maps a vector to a diagonal matrix in the natural way. In other words, the vector relative entropy is a special case of the matrix relative entropy. Ultimately, the vector case is easier to understand because diagonal matrices commute.

8.2.2 Relative Entropy is Nonnegative

As we have noted, the relative entropy measures the difference between two positive vectors. This interpretation is supported by the fact that the relative entropy is nonnegative.

Proposition 8.2.2 (Relative Entropy is Nonnegative)**.** *For positive vectors $\boldsymbol{a}$ and $\boldsymbol{h}$ of the same size, the relative entropy* $\mathrm{D}(\boldsymbol{a};\boldsymbol{h}) \geq 0$.

Proof. Let $f : \mathbb{R}_{++} \to \mathbb{R}$ be a differentiable convex function on the positive real line. The function f lies above its tangent lines, so

$$f(a) \geq f(h) + f'(h)\cdot(a-h) \quad \text{for positive } a \text{ and } h.$$

Instantiate this result for the convex function $f(a) = a\log a - a$, and rearrange to obtain the numerical inequality

$$a(\log a - \log h) - (a-h) \geq 0 \quad \text{for positive } a \text{ and } h.$$

Sum this expression over the components of the vectors $\boldsymbol{a}$ and $\boldsymbol{h}$ to complete the argument. □

Proposition 8.1.3 states that the matrix relative entropy satisfies the same nonnegativity property as the vector relative entropy. The argument for matrices relies on the same ideas as Proposition 8.2.2, and it is hardly more difficult. See §8.3.5 for the details.

8.2.3 The Perspective Transformation

Our next goal is to prove that the relative entropy is a convex function. To establish this claim, we use an elegant technique from convex analysis. The approach depends on the *perspective transformation*, a method for constructing a bivariate convex function from a univariate convex function.

Definition 8.2.3 (Perspective Transformation)**.** *Let $f : \mathbb{R}_{++} \to \mathbb{R}$ be a convex function on the positive real line. The perspective ψ_f of the function f is defined as*

$$\psi_f : \mathbb{R}_{++} \times \mathbb{R}_{++} \to \mathbb{R} \quad \textit{where} \quad \psi_f(a;h) = a\cdot f(h/a).$$

The perspective transformation has an interesting geometric interpretation. If we trace the ray from the origin $(0,0,0)$ in $\mathbb{R}^3$ through the point $(a, h, \psi_f(a;h))$, it pierces the plane $(1,\cdot,\cdot)$ at the point $(1, h/a, f(h/a))$. Equivalently, for each positive a, the epigraph of f is the "shadow" of the epigraph of $\psi_f(a,\cdot)$ on the plane $(1,\cdot,\cdot)$.

The key fact is that the perspective of a convex function is convex. This point follows from the geometric reasoning in the last paragraph; we also include an analytic proof.

Fact 8.2.4 (Perspectives are Convex)**.** *Let $f : \mathbb{R}_{++} \to \mathbb{R}$ be a convex function. Then the perspective ψ_f is convex. That is, for positive numbers a_i and h_i,*

$$\psi_f(\tau a_1 + \bar{\tau} a_2;\ \tau h_1 + \bar{\tau} h_2) \le \tau \cdot \psi_f(a_1; h_1) + \bar{\tau} \cdot \psi_f(a_2; h_2) \quad \text{for } \tau \in [0,1].$$

Proof. Fix two pairs (a_1, h_1) and (a_2, h_2) of positive numbers and an interpolation parameter $\tau \in [0,1]$. Form the convex combinations

$$a = \tau a_1 + \bar{\tau} a_2 \quad \text{and} \quad h = \tau h_1 + \bar{\tau} h_2.$$

We need to bound the perspective $\psi_f(a;h)$ as the convex combination of its values at $\psi_f(a_1;h_1)$ and $\psi_f(a_2;h_2)$. The trick is to introduce another pair of interpolation parameters:

$$s = \frac{\tau a_1}{a} \quad \text{and} \quad \bar{s} = \frac{\bar{\tau} a_2}{a}.$$

By construction, $s \in [0,1]$ and $\bar{s} = 1 - s$. We quickly determine that

$$\begin{aligned}
\psi_f(a;h) &= a \cdot f(h/a) \\
&= a \cdot f(\tau h_1/a + \bar{\tau} h_2/a) \\
&= a \cdot f(s \cdot h_1/a_1 + \bar{s} \cdot h_2/a_2) \\
&\le a\left[s \cdot f(h_1/a_1) + \bar{s} \cdot f(h_2/a_2)\right] \\
&= \tau \cdot a_1 \cdot f(h_1/a_1) + \bar{\tau} \cdot a_2 \cdot f(h_2/a_2) \\
&= \tau \cdot \psi_f(a_1;h_1) + \bar{\tau} \cdot \psi_f(a_2;h_2).
\end{aligned}$$

To obtain the second identity, we write h as a convex combination. The third identity follows from the definitions of s and $\bar{s}$. The inequality depends on the fact that f is convex. Afterward, we invoke the definitions of s and $\bar{s}$ again. We conclude that ψ_f is convex. □

When we study standard matrix functions, it is sometimes necessary to replace a convexity assumption by a stricter property called operator convexity. There is a remarkable extension of the perspective transform that constructs a bivariate matrix function from an operator convex function. The matrix perspective has a powerful convexity property analogous with the result in Fact 8.2.4. The analysis of the matrix perspective depends on a far-reaching generalization of the Jensen inequality for operator convex functions. We develop these ideas in §§8.4.5, 8.5, and 8.6.

8.2.4 The Relative Entropy is Convex

To establish that the relative entropy is convex, we simply need to represent it as the perspective of a convex function.

Proposition 8.2.5 (Relative Entropy is Convex)**.** *The map* $(\boldsymbol{a}, \boldsymbol{h}) \mapsto \mathrm{D}(\boldsymbol{a}; \boldsymbol{h})$ *is convex. That is, for positive vectors* $\boldsymbol{a}_i$ *and* $\boldsymbol{h}_i$ *of the same size,*

$$\mathrm{D}(\tau \boldsymbol{a}_1 + \bar{\tau} \boldsymbol{a}_2; \tau \boldsymbol{h}_1 + \bar{\tau} \boldsymbol{h}_2) \le \tau \cdot \mathrm{D}(\boldsymbol{a}_1; \boldsymbol{h}_1) + \bar{\tau} \cdot \mathrm{D}(\boldsymbol{a}_2; \boldsymbol{h}_2) \quad \text{for } \tau \in [0, 1].$$

Proof. Consider the convex function $f(a) = a - 1 - \log a$, defined on the positive real line. By direct calculation, the perspective transformation satisfies

$$\psi_f(a; h) = a(\log a - \log h) - (a - h) \quad \text{for positive } a \text{ and } h.$$

Fact 8.2.4 states that ψ_f is a convex function. For positive vectors $\boldsymbol{a}$ and $\boldsymbol{h}$, we can express the relative entropy as

$$\mathrm{D}(\boldsymbol{a}; \boldsymbol{h}) = \textstyle\sum_k \psi_f(a_k; h_k),$$

It follows that the relative entropy is convex. □

Similarly, we can express the matrix relative entropy using the matrix perspective transformation. The analysis for matrices is substantially more involved. But, as we will see in §8.8, the argument ultimately follows the same pattern as the proof of Proposition 8.2.5.

8.3 Elementary Trace Inequalities

It is time to begin our investigation into the properties of matrix functions. This section contains some simple inequalities for the trace of a matrix function that we can establish by manipulating eigenvalues and eigenvalue decompositions. These techniques are adequate to explain why the matrix relative entropy is nonnegative. In contrast, we will need more subtle arguments to study the convexity properties of the matrix relative entropy.

8.3.1 Trace Functions

We can construct a real-valued function on Hermitian matrices by composing the trace with a standard matrix function. This type of map is called a *trace function*.

Definition 8.3.1 (Trace function)**.** *Let* $f : I \to \mathbb{R}$ *be a function on an interval* I *of the real line, and let* $\boldsymbol{A}$ *be an Hermitian matrix whose eigenvalues are contained in* I*. We define the* trace function $\operatorname{tr} f$ *by the rule*

$$\operatorname{tr} f(\boldsymbol{A}) = \textstyle\sum_i f(\lambda_i(\boldsymbol{A})),$$

where $\lambda_i(\boldsymbol{A})$ *denotes the* i*th largest eigenvalue of* $\boldsymbol{A}$*. This formula gives the same result as composing the trace with the standard matrix function* f*.*

Our first goal is to demonstrate that a trace function $\operatorname{tr} f$ inherits a monotonicity property from the underlying scalar function f.

8.3.2 Monotone Trace Functions

Let us demonstrate that the trace of a weakly increasing scalar function induces a trace function that preserves the semidefinite order. To that end, recall that the relation $\boldsymbol{A} \preccurlyeq \boldsymbol{H}$ implies that each eigenvalue of $\boldsymbol{A}$ is dominated by the corresponding eigenvalue of $\boldsymbol{H}$.

Fact 8.3.2 (Semidefinite Order implies Eigenvalue Order)**.** *For Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{H}$,*

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \text{implies} \quad \lambda_i(\boldsymbol{A}) \le \lambda_i(\boldsymbol{H}) \quad \text{for each index } i.$$

Proof. This result follows instantly from the Courant–Fischer Theorem:

$$\lambda_i(\boldsymbol{A}) = \max_{\dim L = i} \min_{\boldsymbol{u} \in L} \frac{\boldsymbol{u}^* \boldsymbol{A} \boldsymbol{u}}{\boldsymbol{u}^* \boldsymbol{u}} \le \max_{\dim L = i} \min_{\boldsymbol{u} \in L} \frac{\boldsymbol{u}^* \boldsymbol{H} \boldsymbol{u}}{\boldsymbol{u}^* \boldsymbol{u}} = \lambda_i(\boldsymbol{H}).$$

The maximum ranges over all i-dimensional linear subspaces L in the domain of $\boldsymbol{A}$, and we use the convention that $0/0 = 0$. The inequality follows from the definition (2.1.11) of the semidefinite order $\preccurlyeq$. □

With this fact at hand, the claim follows quickly.

Proposition 8.3.3 (Monotone Trace Functions)**.** *Let $f : I \to \mathbb{R}$ be a weakly increasing function on an interval I of the real line, and let $\boldsymbol{A}$ and $\boldsymbol{H}$ be Hermitian matrices whose eigenvalues are contained in I. Then*

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \text{implies} \quad \operatorname{tr} f(\boldsymbol{A}) \le \operatorname{tr} f(\boldsymbol{H}).$$

Proof. In view of Fact 8.3.2,

$$\operatorname{tr} f(\boldsymbol{A}) = \sum\nolimits_i f(\lambda_i(\boldsymbol{A})) \le \sum\nolimits_i f(\lambda_i(\boldsymbol{H})) = \operatorname{tr} f(\boldsymbol{H}).$$

The inequality depends on the assumption that f is weakly increasing. □

Our approach to matrix concentration relies on a special case of Proposition 8.3.3.

Example 8.3.4 (Trace Exponential is Monotone)**.** *The trace exponential map is monotone:*

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \text{implies} \quad \operatorname{tr} \mathrm{e}^{\boldsymbol{A}} \le \operatorname{tr} \mathrm{e}^{\boldsymbol{H}}$$

for all Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{H}$.

8.3.3 Eigenvalue Decompositions, Redux

Before we continue, let us introduce a style for writing eigenvalue decompositions that will make the next argument more transparent. Each $d \times d$ Hermitian matrix $\boldsymbol{A}$ can be expressed as

$$\boldsymbol{A} = \sum_{i=1}^{d} \lambda_i \, \boldsymbol{u}_i \boldsymbol{u}_i^* .$$

The eigenvalues $\lambda_1 \ge \cdots \ge \lambda_d$ of $\boldsymbol{A}$ are real numbers, listed in weakly decreasing order. The family $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_d\}$ of eigenvectors of $\boldsymbol{A}$ forms an orthonormal basis for $\mathbb{C}^d$ with respect to the standard inner product.

8.3.4 A Trace Inequality for Bivariate Functions

In general, it is challenging to study functions of two or more matrices because the eigenvectors can interact in complicated ways. Nevertheless, there is one type of relation that always transfers from the scalar setting to the matrix setting.

Proposition 8.3.5 (Generalized Klein Inequality)**.** *Let $f_i : I \to \mathbb{R}$ and $g_i : I \to \mathbb{R}$ be functions on an interval I of the real line, and suppose that*

$$\sum\nolimits_i f_i(a)\, g_i(h) \geq 0 \quad \text{for all } a, h \in I.$$

If $\boldsymbol{A}$ and $\boldsymbol{H}$ are Hermitian matrices whose eigenvalues are contained in I, then

$$\sum\nolimits_i \operatorname{tr}\big[f_i(\boldsymbol{A})\, g_i(\boldsymbol{H})\big] \geq 0.$$

Proof. Consider eigenvalue decompositions $\boldsymbol{A} = \sum_j \lambda_j \boldsymbol{u}_j \boldsymbol{u}_j^*$ and $\boldsymbol{H} = \sum_k \mu_k \boldsymbol{v}_k \boldsymbol{v}_k^*$. Then

$$\begin{aligned}\sum\nolimits_i \operatorname{tr}\big[f_i(\boldsymbol{A})\, g_i(\boldsymbol{H})\big] &= \sum\nolimits_i \operatorname{tr}\Big[\big(\textstyle\sum_j f_i(\lambda_j)\boldsymbol{u}_j\boldsymbol{u}_j^*\big)\big(\sum_k g_i(\mu_k)\,\boldsymbol{v}_k\boldsymbol{v}_k^*\big)\Big] \\ &= \sum\nolimits_{j,k} \big[\textstyle\sum_i f_i(\lambda_j)\, g_i(\mu_k)\big] \cdot |\langle \boldsymbol{u}_j,\ \boldsymbol{v}_k\rangle|^2 \geq 0.\end{aligned}$$

We use the definition of a standard matrix function, we apply linearity of the trace to reorder the sums, and we identify the trace as a squared inner product. The inequality follows from our assumption on the scalar functions. □

8.3.5 The Matrix Relative Entropy is Nonnegative

Using the generalized Klein inequality, it is easy to prove Proposition 8.1.3, which states that the matrix relative entropy is nonnegative. The argument echoes the analysis in Proposition 8.2.2 for the vector case.

Proof of Proposition 8.1.3. Suppose that $f : \mathbb{R}_{++} \to \mathbb{R}$ is a differentiable, convex function on the positive real line. Since f is convex, the graph of f lies above its tangents:

$$f(a) \geq f(h) + f'(h)(a - h) \quad \text{for positive } a \text{ and } h.$$

Using the generalized Klein inequality, Proposition 8.3.5, we can lift this relation to matrices:

$$\operatorname{tr} f(\boldsymbol{A}) \geq \operatorname{tr}\big[f(\boldsymbol{H}) + f'(\boldsymbol{H})(\boldsymbol{A} - \boldsymbol{H})\big] \quad \text{for all positive-definite } \boldsymbol{A} \text{ and } \boldsymbol{H}.$$

This formula is sometimes called the (ungeneralized) Klein inequality.

Instantiate the latter result for the function $f(a) = a \log a - a$, and rearrange to see that

$$\mathrm{D}(\boldsymbol{A}; \boldsymbol{H}) = \operatorname{tr}\big[\boldsymbol{A}(\log \boldsymbol{A} - \log \boldsymbol{H}) - (\boldsymbol{A} - \boldsymbol{H})\big] \geq 0 \quad \text{for all positive-definite } \boldsymbol{A} \text{ and } \boldsymbol{H}.$$

In other words, the matrix relative entropy is nonnegative. □

8.4 The Logarithm of a Matrix

In this section, we commence our journey toward the proof that the matrix relative entropy is convex. The proof of Proposition 8.2.5 indicates that the convexity of the logarithm plays an important role in the convexity of the vector relative entropy. As a first step, we will demonstrate that the matrix logarithm has a striking convexity property with respect to the semidefinite order. Along the way, we will also develop a monotonicity property of the matrix logarithm.

8.4.1 An Integral Representation of the Logarithm

Initially, we defined the logarithm of a $d \times d$ positive-definite matrix $\boldsymbol{A}$ using an eigenvalue decomposition:

$$\log \boldsymbol{A} = \boldsymbol{Q} \begin{bmatrix} \log \lambda_1 & & \\ & \ddots & \\ & & \log \lambda_d \end{bmatrix} \boldsymbol{Q}^* \quad \text{where} \quad \boldsymbol{A} = \boldsymbol{Q} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} \boldsymbol{Q}^*.$$

To study how the matrix logarithm interacts with the semidefinite order, we will work with an alternative presentation based on an integral formula.

Proposition 8.4.1 (Integral Representation of the Logarithm)**.** *The logarithm of a positive number a is given by the integral*

$$\log a = \int_0^\infty \left[\frac{1}{1+u} - \frac{1}{a+u} \right] \mathrm{d}u.$$

Similarly, the logarithm of a positive-definite matrix $\boldsymbol{A}$ is given by the integral

$$\log \boldsymbol{A} = \int_0^\infty \left[(1+u)^{-1} \mathbf{I} - (\boldsymbol{A} + u\mathbf{I})^{-1} \right] \mathrm{d}u.$$

Proof. To verify the scalar formula, we simply use the definition of the improper integral:

$$\begin{aligned} \int_0^\infty \left[\frac{1}{1+u} - \frac{1}{a+u} \right] \mathrm{d}u &= \lim_{L \to \infty} \int_0^L \left[\frac{1}{1+u} - \frac{1}{a+u} \right] \mathrm{d}u \\ &= \lim_{L \to \infty} \big[\log(1+u) - \log(a+u) \big]_{u=0}^{L} \\ &= \log a + \lim_{L \to \infty} \log\left(\frac{1+L}{a+L} \right) = \log a. \end{aligned}$$

We obtain the matrix formula by applying the scalar formula to each eigenvalue of $\boldsymbol{A}$ and then expressing the result in terms of the original matrix. □

The integral formula from Proposition 8.4.1 is powerful because it expresses the logarithm in terms of the matrix inverse, which is much easier to analyze. Although it may seem that we have pulled this representation from thin air, the approach is motivated by a wonderful theory of matrix functions initiated by Löwner in the 1930s.

8.4.2 Operator Monotone Functions

Our next goal is to study the monotonicity properties of the matrix logarithm. To frame this discussion properly, we need to introduce an abstract definition.

Definition 8.4.2 (Operator Monotone Function)**.** *Let $f : I \to \mathbb{R}$ be a function on an interval I of the real line. The function f is* operator monotone *on I when*

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \textit{implies} \quad f(\boldsymbol{A}) \preccurlyeq f(\boldsymbol{H})$$

for all Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{H}$ whose eigenvalues are contained in I.

Let us state some basic facts about operator monotone functions. Many of these points follow easily from the definition.

- When $\beta \geq 0$, the weakly increasing affine function $t \mapsto \alpha + \beta t$ is operator monotone on each interval I of the real line.
- The quadratic function $t \mapsto t^2$ is **not** operator monotone on the positive real line.
- The exponential map $t \mapsto \mathrm{e}^t$ is **not** operator monotone on the real line.
- When $\alpha \geq 0$ and f is operator monotone on I, the function αf is operator monotone on I.
- If f and g are operator monotone on an interval I, then $f + g$ is operator monotone on I.

These properties imply that the operator monotone functions form a convex cone. It also warns us that the class of operator monotone functions is somewhat smaller than the class of weakly increasing functions.

8.4.3 The Negative Inverse is Operator Monotone

Fortunately, interesting operator monotone functions do exist. Let us present an important example related to the matrix inverse.

Proposition 8.4.3 (Negative Inverse is Operator Monotone)**.** *For each number $u \geq 0$, the function $a \mapsto -(a+u)^{-1}$ is operator monotone on the positive real line. That is, for positive-definite matrices $\boldsymbol{A}$ and $\boldsymbol{H}$,*

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \text{implies} \quad -(\boldsymbol{A} + u\mathbf{I})^{-1} \preccurlyeq -(\boldsymbol{H} + u\mathbf{I})^{-1}.$$

Proof. Define the matrices $\boldsymbol{A}_u = \boldsymbol{A} + u\mathbf{I}$ and $\boldsymbol{H}_u = \boldsymbol{H} + u\mathbf{I}$. The semidefinite relation $\boldsymbol{A} \preccurlyeq \boldsymbol{H}$ implies that $\boldsymbol{A}_u \preccurlyeq \boldsymbol{H}_u$. Apply the Conjugation Rule (2.1.12) to see that

$$\mathbf{0} \prec \boldsymbol{H}_u^{-1/2} \boldsymbol{A}_u \boldsymbol{H}_u^{-1/2} \preccurlyeq \mathbf{I}.$$

When a positive-definite matrix has eigenvalues bounded above by one, its inverse has eigenvalues bounded below by one. Therefore,

$$\mathbf{I} \preccurlyeq \left(\boldsymbol{H}_u^{-1/2} \boldsymbol{A}_u \boldsymbol{H}_u^{-1/2}\right)^{-1} = \boldsymbol{H}_u^{1/2} \boldsymbol{A}_u^{-1} \boldsymbol{H}_u^{1/2}.$$

Another application of the Conjugation Rule (2.1.12) delivers the inequality $\boldsymbol{H}_u^{-1} \preccurlyeq \boldsymbol{A}_u^{-1}$. Finally, we negate this semidefinite relation, which reverses its direction. □

8.4.4 The Logarithm is Operator Monotone

Now, we are prepared to demonstrate that the logarithm is an operator monotone function. The argument combines the integral representation from Proposition 8.4.1 with the monotonicity of the inverse map from Proposition 8.4.3.

Proposition 8.4.4 (Logarithm is Operator Monotone)**.** *The logarithm is an operator monotone function on the positive real line. That is, for positive-definite matrices $\boldsymbol{A}$ and $\boldsymbol{H}$,*

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \text{implies} \quad \log \boldsymbol{A} \preccurlyeq \log \boldsymbol{H}.$$

Proof. For each $u \geq 0$, Proposition 8.4.3 demonstrates that

$$(1+u)^{-1}\mathbf{I} - (\boldsymbol{A}+u\mathbf{I})^{-1} \preccurlyeq (1+u)^{-1}\mathbf{I} - (\boldsymbol{H}+u\mathbf{I})^{-1}.$$

The integral representation of the logarithm, Proposition 8.4.1, allows us to calculate that

$$\log \boldsymbol{A} = \int_0^\infty \left[(1+u)^{-1}\mathbf{I} - (\boldsymbol{A}+u\mathbf{I})^{-1}\right] \mathrm{d}u \preccurlyeq \int_0^\infty \left[(1+u)^{-1}\mathbf{I} - (\boldsymbol{H}+u\mathbf{I})^{-1}\right] \mathrm{d}u = \log \boldsymbol{H}.$$

We have used the fact that the semidefinite order is preserved by integration against a positive measure. □

8.4.5 Operator Convex Functions

Next, let us investigate the convexity properties of the matrix logarithm. As before, we start with an abstract definition.

Definition 8.4.5 (Operator Convex Function)**.** *Let $f : I \to \mathbb{R}$ be a function on an interval I of the real line. The function f is* operator convex *on I when*

$$f(\tau \boldsymbol{A} + \bar{\tau}\boldsymbol{H}) \preccurlyeq \tau \cdot f(\boldsymbol{A}) + \bar{\tau} \cdot f(\boldsymbol{H}) \quad \text{for all } \tau \in [0,1]$$

and for all Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{H}$ whose eigenvalues are contained in I. A function $g : I \to \mathbb{R}$ is operator concave *when $-g$ is operator convex on I.*

We continue with some important facts about operator convex functions. Most of these claims can be derived easily.

- When $\gamma \geq 0$, the quadratic function $t \mapsto \alpha + \beta t + \gamma t^2$ is operator convex on the real line.
- The exponential map $t \mapsto \mathrm{e}^t$ is **not** operator convex on the real line.
- When $\alpha \geq 0$ and f is operator convex on I, the function αf is operator convex in I.
- If f and g are operator convex on I, then $f + g$ is operator convex on I.

The operator monotone functions form a convex cone. We also learn that the family of operator convex functions is somewhat smaller than the family of convex functions.

8.4.6 The Inverse is Operator Convex

The inverse provides a very important example of an operator convex function.

Proposition 8.4.6 (Inverse is Operator Convex)**.** *For each $u \geq 0$, the function $a \mapsto (a+u)^{-1}$ is operator convex on the positive real line. That is, for positive-definite matrices $\boldsymbol{A}$ and $\boldsymbol{H}$,*

$$(\tau \boldsymbol{A} + \bar{\tau}\boldsymbol{H} + u\mathbf{I})^{-1} \preccurlyeq \tau \cdot (\boldsymbol{A}+u\mathbf{I})^{-1} + \bar{\tau} \cdot (\boldsymbol{H}+u\mathbf{I})^{-1} \quad \text{for } \tau \in [0,1].$$

To establish Proposition 8.4.6, we use an argument based on the Schur complement lemma. For completeness, let us state and prove this important fact.

Fact 8.4.7 (Schur Complements). *Suppose that $\boldsymbol{T}$ is a positive-definite matrix. Then*

$$\mathbf{0} \preccurlyeq \begin{bmatrix} \boldsymbol{T} & \boldsymbol{B} \\ \boldsymbol{B}^* & \boldsymbol{M} \end{bmatrix} \quad \text{if and only if} \quad \boldsymbol{B}^* \boldsymbol{T}^{-1} \boldsymbol{B} \preccurlyeq \boldsymbol{M}. \tag{8.4.1}$$

Proof of Fact 8.4.7. To see why this is true, just calculate that

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\boldsymbol{B}^* \boldsymbol{T}^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{T} & \boldsymbol{B} \\ \boldsymbol{B}^* & \boldsymbol{M} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\boldsymbol{T}^{-1}\boldsymbol{B} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{T} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{M} - \boldsymbol{B}^* \boldsymbol{T}^{-1} \boldsymbol{B} \end{bmatrix}$$

In essence, we are performing block Gaussian elimination to bring the original matrix into block-diagonal form. Now, the Conjugation Rule (2.1.12) ensures that the central matrix on the left is positive semidefinite together with the matrix on the right. From this equivalence, we extract the result (8.4.1). □

We continue with the proof that the inverse is operator convex.

Proof of Proposition 8.4.6. The Schur complement lemma, Fact 8.4.7, provides that

$$\mathbf{0} \preccurlyeq \begin{bmatrix} \boldsymbol{T} & \mathbf{I} \\ \mathbf{I} & \boldsymbol{T}^{-1} \end{bmatrix} \quad \text{whenever } \boldsymbol{T} \text{ is positive definite.}$$

Applying this observation to the positive-definite matrices $\boldsymbol{A} + u\mathbf{I}$ and $\boldsymbol{H} + u\mathbf{I}$, we see that

$$\begin{aligned} \mathbf{0} &\preccurlyeq \tau \cdot \begin{bmatrix} \boldsymbol{A} + u\mathbf{I} & \mathbf{I} \\ \mathbf{I} & (\boldsymbol{A} + u\mathbf{I})^{-1} \end{bmatrix} + \bar{\tau} \cdot \begin{bmatrix} \boldsymbol{H} + u\mathbf{I} & \mathbf{I} \\ \mathbf{I} & (\boldsymbol{H} + u\mathbf{I})^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \tau \boldsymbol{A} + \bar{\tau} \boldsymbol{H} + u\mathbf{I} & \mathbf{I} \\ \mathbf{I} & \tau \cdot (\boldsymbol{A} + u\mathbf{I})^{-1} + \bar{\tau} \cdot (\boldsymbol{H} + u\mathbf{I})^{-1} \end{bmatrix}. \end{aligned}$$

Since the top-left block of the latter matrix is positive definite, another application of Fact 8.4.7 delivers the relation

$$(\tau \boldsymbol{A} + \bar{\tau} \boldsymbol{H} + u\mathbf{I})^{-1} \preccurlyeq \tau \cdot (\boldsymbol{A} + u\mathbf{I})^{-1} + \bar{\tau} \cdot (\boldsymbol{H} + u\mathbf{I})^{-1}.$$

This is the advertised conclusion. □

8.4.7 The Logarithm is Operator Concave

We are finally prepared to verify that the logarithm is operator concave. The argument is based on the integral representation from Proposition 8.4.4 and the convexity of the inverse map from Proposition 8.4.6.

Proposition 8.4.8 (Logarithm is Operator Concave). *The logarithm is operator concave on the positive real line. That is, for positive-definite matrices $\boldsymbol{A}$ and $\boldsymbol{H}$,*

$$\tau \cdot \log \boldsymbol{A} + \bar{\tau} \cdot \log \boldsymbol{H} \preccurlyeq \log(\tau \boldsymbol{A} + \bar{\tau} \boldsymbol{H}) \quad \text{for } \tau \in [0, 1].$$

Proof. For each $u \geq 0$, Proposition 8.4.6 demonstrates that

$$-\tau \cdot (\boldsymbol{A} + u\mathbf{I})^{-1} - \bar{\tau} \cdot (\boldsymbol{H} + u\mathbf{I})^{-1} \preccurlyeq -(\tau \boldsymbol{A} + \bar{\tau} \boldsymbol{H} + u\mathbf{I})^{-1}.$$

Invoke the integral representation of the logarithm from Proposition 8.4.1 to see that

$$\begin{aligned}\tau \cdot \log \boldsymbol{A} + \bar{\tau} \cdot \log \boldsymbol{H} &= \tau \cdot \int_0^\infty \left[(1+u)^{-1}\mathbf{I} - (\boldsymbol{A}+u\mathbf{I})^{-1}\right] \mathrm{d}u + \bar{\tau} \cdot \int_0^\infty \left[(1+u)^{-1}\mathbf{I} - (\boldsymbol{H}+u\mathbf{I})^{-1}\right] \mathrm{d}u \\ &= \int_0^\infty \left[(1+u)^{-1}\mathbf{I} - \left(\tau \cdot (\boldsymbol{A}+u\mathbf{I})^{-1} + \bar{\tau} \cdot (\boldsymbol{H}+u\mathbf{I})^{-1}\right)\right] \mathrm{d}u \\ &\preccurlyeq \int_0^\infty \left[(1+u)^{-1}\mathbf{I} - (\tau\boldsymbol{A} + \bar{\tau}\boldsymbol{H} + u\mathbf{I})^{-1}\right] \mathrm{d}u = \log(\tau\boldsymbol{A} + \bar{\tau}\boldsymbol{H}).\end{aligned}$$

Once again, we have used the fact that integration preserves the semidefinite order. □

8.5 The Operator Jensen Inequality

Convexity is a statement about how a function interacts with averages. By definition, a function $f : I \to \mathbb{R}$ is convex when

$$f(\tau a + \bar{\tau} h) \le \tau \cdot f(a) + \bar{\tau} \cdot f(h) \quad \text{for all } \tau \in [0,1] \text{ and all } a, h \in I. \tag{8.5.1}$$

The convexity inequality (8.5.1) automatically extends from an average involving two terms to an arbitrary average. This is the content of Jensen's inequality.

Definition 8.4.5, of an operator convex function $f : I \to \mathbb{R}$, is similar in spirit:

$$f(\tau\boldsymbol{A} + \bar{\tau}\boldsymbol{H}) \preccurlyeq \tau \cdot f(\boldsymbol{A}) + \bar{\tau} \cdot f(\boldsymbol{H}) \quad \text{for all } \tau \in [0,1] \tag{8.5.2}$$

and all Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{H}$ whose eigenvalues are contained in I. Surprisingly, the semidefinite relation (8.5.2) automatically extends to a large family of matrix averaging operations. This remarkable property is called the *operator Jensen inequality*.

8.5.1 Matrix Convex Combinations

In a vector space, convex combinations provide a natural method of averaging. But matrices have a richer structure, so we can consider a more general class of averages.

Definition 8.5.1 (Matrix Convex Combination)**.** *Let $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ be Hermitian matrices. Consider a decomposition of the identity of the form*

$$\boldsymbol{K}_1^* \boldsymbol{K}_1 + \boldsymbol{K}_2^* \boldsymbol{K}_2 = \mathbf{I}.$$

Then the Hermitian matrix

$$\boldsymbol{K}_1^* \boldsymbol{A}_1 \boldsymbol{K}_1 + \boldsymbol{K}_2^* \boldsymbol{A}_2 \boldsymbol{K}_2 \tag{8.5.3}$$

is called a matrix convex combination *of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$.*

To see why it is reasonable to call (8.5.3) an averaging operation on Hermitian matrices, let us note a few of its properties.

- Definition 8.5.1 encompasses scalar convex combinations because we can take $\boldsymbol{K}_1 = \tau^{1/2}\mathbf{I}$ and $\boldsymbol{K}_2 = \bar{\tau}^{1/2}\mathbf{I}$.
- The matrix convex combination preserves the identity matrix: $\boldsymbol{K}_1^*\mathbf{I}\boldsymbol{K}_1 + \boldsymbol{K}_2^*\mathbf{I}\boldsymbol{K}_2 = \mathbf{I}$.

- The matrix convex combination preserves positivity:

$$\boldsymbol{K}_1^* \boldsymbol{A}_1 \boldsymbol{K}_1 + \boldsymbol{K}_2^* \boldsymbol{A}_2 \boldsymbol{K}_2 \succcurlyeq \mathbf{0} \quad \text{for all positive-semidefinite } \boldsymbol{A}_1 \text{ and } \boldsymbol{A}_2.$$

- If the eigenvalues of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are contained in an interval I, then the eigenvalues of the matrix convex combination (8.5.3) are also contained in I.

We will encounter a concrete example of a matrix convex combination later when we prove Theorem 8.6.2.

8.5.2 Jensen's Inequality for Matrix Convex Combinations

Operator convexity is a self-improving property. Even though the definition of an operator convex function only involves a scalar convex combination, it actually contains an inequality for matrix convex combinations. This is the content of the operator Jensen inequality.

Theorem 8.5.2 (Operator Jensen Inequality)**.** *Let f be an operator convex function on an interval I of the real line, and let $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ be Hermitian matrices with eigenvalues in I. Consider a decomposition of the identity*

$$\boldsymbol{K}_1^* \boldsymbol{K}_1 + \boldsymbol{K}_2^* \boldsymbol{K}_2 = \mathbf{I}. \tag{8.5.4}$$

Then

$$f\left(\boldsymbol{K}_1^* \boldsymbol{A}_1 \boldsymbol{K}_1 + \boldsymbol{K}_2^* \boldsymbol{A}_2 \boldsymbol{K}_2\right) \preccurlyeq \boldsymbol{K}_1^* f(\boldsymbol{A}_1) \boldsymbol{K}_1 + \boldsymbol{K}_2^* f(\boldsymbol{A}_2) \boldsymbol{K}_2.$$

Proof. Let us introduce a block-diagonal matrix:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}_2 \end{bmatrix} \quad \text{for which} \quad f(\boldsymbol{A}) = \begin{bmatrix} f(\boldsymbol{A}_1) & \mathbf{0} \\ \mathbf{0} & f(\boldsymbol{A}_2) \end{bmatrix}.$$

Indeed, the matrix $\boldsymbol{A}$ lies in the domain of f because its eigenvalues fall in the interval I. We can apply a standard matrix function to a block-diagonal matrix by applying the function to each block.

There are two main ingredients in the argument. The first idea is to realize the matrix convex combination of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ by conjugating the block-diagonal matrix $\boldsymbol{A}$ with an appropriate unitary matrix. To that end, let us construct a unitary matrix

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{K}_1 & \boldsymbol{L}_1 \\ \boldsymbol{K}_2 & \boldsymbol{L}_2 \end{bmatrix} \quad \text{where } \boldsymbol{Q}^* \boldsymbol{Q} = \mathbf{I} \text{ and } \boldsymbol{Q}\boldsymbol{Q}^* = \mathbf{I}.$$

To see why this is possible, note that the first block of columns is orthonormal:

$$\begin{bmatrix} \boldsymbol{K}_1 \\ \boldsymbol{K}_2 \end{bmatrix}^* \begin{bmatrix} \boldsymbol{K}_1 \\ \boldsymbol{K}_2 \end{bmatrix} = \boldsymbol{K}_1^* \boldsymbol{K}_1 + \boldsymbol{K}_2^* \boldsymbol{K}_2 = \mathbf{I}.$$

As a consequence, we can choose $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ to complete the unitary matrix $\boldsymbol{Q}$. By direct computation, we find that

$$\boldsymbol{Q}^* \boldsymbol{A} \boldsymbol{Q} = \begin{bmatrix} \boldsymbol{K}_1^* \boldsymbol{A}_1 \boldsymbol{K}_1 + \boldsymbol{K}_2^* \boldsymbol{A}_2 \boldsymbol{K}_2 & * \\ * & * \end{bmatrix} \tag{8.5.5}$$

We have omitted the precise values of the entries labeled $*$ because they do not play a role in our argument.

The second idea is to restrict the block matrix in (8.5.5) to its diagonal. To perform this maneuver, we express the diagonalizing operation as a *scalar convex combination* of two unitary conjugations, which gives us access to the operator convexity of f. Let us see how this works. Define the unitary matrix

$$\boldsymbol{U} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{bmatrix}.$$

The key observation is that, for any block matrix,

$$\frac{1}{2}\begin{bmatrix} \boldsymbol{T} & \boldsymbol{B} \\ \boldsymbol{B}^* & \boldsymbol{M} \end{bmatrix} + \frac{1}{2}\boldsymbol{U}^*\begin{bmatrix} \boldsymbol{T} & \boldsymbol{B} \\ \boldsymbol{B}^* & \boldsymbol{M} \end{bmatrix}\boldsymbol{U} = \begin{bmatrix} \boldsymbol{T} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{M} \end{bmatrix}. \tag{8.5.6}$$

Another advantage of this construction is that we can easily apply a standard matrix function to the block-diagonal matrix.

Together, these two ideas lead to a succinct proof of the operator Jensen inequality. Write $[\cdot]_{11}$ for the operation that returns the $(1,1)$ block of a block matrix. We may calculate that

$$\begin{aligned} f\left(\boldsymbol{K}_1^*\boldsymbol{A}_1\boldsymbol{K}_1 + \boldsymbol{K}_2^*\boldsymbol{A}_2\boldsymbol{K}_2\right) &= f\left([\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}]_{11}\right) \\ &= f\left(\left[\frac{1}{2}\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} + \frac{1}{2}\boldsymbol{U}^*(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q})\boldsymbol{U}\right]_{11}\right) \\ &= \left[f\left(\frac{1}{2}\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} + \frac{1}{2}(\boldsymbol{Q}\boldsymbol{U})^*\boldsymbol{A}(\boldsymbol{Q}\boldsymbol{U})\right)\right]_{11} \\ &\preccurlyeq \left[\frac{1}{2}f(\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}) + \frac{1}{2}f((\boldsymbol{Q}\boldsymbol{U})^*\boldsymbol{A}(\boldsymbol{Q}\boldsymbol{U}))\right]_{11}. \end{aligned}$$

The first identity depends on the representation (8.5.5) of the matrix convex combination as the $(1,1)$ block of $\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}$. The second line follows because the averaging operation presented in (8.5.6) does not alter the $(1,1)$ block of the matrix. In view of (8.5.6), we are looking at the $(1,1)$ block of the matrix obtained by applying f to a block-diagonal matrix. This is equivalent to applying the function f inside the $(1,1)$ block, which gives the third line. Last, the semidefinite relation follows from the operator convexity of f on the interval I.

We complete the argument by reversing the steps we have taken so far.

$$\begin{aligned} f\left(\boldsymbol{K}_1^*\boldsymbol{A}_1\boldsymbol{K}_1 + \boldsymbol{K}_2^*\boldsymbol{A}_2\boldsymbol{K}_2\right) &\preccurlyeq \left[\frac{1}{2}\boldsymbol{Q}^*f(\boldsymbol{A})\boldsymbol{Q} + \frac{1}{2}\boldsymbol{U}^*(\boldsymbol{Q}^*f(\boldsymbol{A})\boldsymbol{Q})\boldsymbol{U}\right]_{11} \\ &= \left[\boldsymbol{Q}^*f(\boldsymbol{A})\boldsymbol{Q}\right]_{11} \\ &= \boldsymbol{K}_1^*f(\boldsymbol{A}_1)\boldsymbol{K}_1 + \boldsymbol{K}_2^*f(\boldsymbol{A}_2)\boldsymbol{K}_2. \end{aligned}$$

To obtain the first relation, recall that a standard matrix function commutes with unitary conjugation. The second identity follows from the formula (8.5.6) because diagonalization preserves the $(1,1)$ block. Finally, we identify the $(1,1)$ block of $\boldsymbol{Q}^*f(\boldsymbol{A})\boldsymbol{Q}$ just as we did in (8.5.5). This step depends on the fact that the diagonal blocks of $f(\boldsymbol{A})$ are simply $f(\boldsymbol{A}_1)$ and $f(\boldsymbol{A}_2)$. □

8.6 The Matrix Perspective Transformation

To show that the vector relative entropy is convex, we represented it as the perspective of a convex function. To demonstrate that the matrix relative entropy is convex, we are going to perform a similar maneuver. This section develops an extension of the perspective transformation that applies to operator convex functions. Then we demonstrate that this matrix perspective has a strong convexity property with respect to the semidefinite order.

8.6.1 The Matrix Perspective

In the scalar setting, the perspective transformation converts a convex function into a bivariate convex function. There is a related construction that applies to an operator convex function.

Definition 8.6.1 (Matrix Perspective)**.** *Let $f : \mathbb{R}_{++} \to \mathbb{R}$ be an operator convex function, and let $\boldsymbol{A}$ and $\boldsymbol{H}$ be positive-definite matrices of the same size. Define the perspective map*

$$\boldsymbol{\Psi}_f(\boldsymbol{A}; \boldsymbol{H}) = \boldsymbol{A}^{1/2} \cdot f\big(\boldsymbol{A}^{-1/2} \boldsymbol{H} \boldsymbol{A}^{-1/2}\big) \cdot \boldsymbol{A}^{1/2}.$$

The notation $\boldsymbol{A}^{1/2}$ refers to the unique positive-definite square root of $\boldsymbol{A}$, and $\boldsymbol{A}^{-1/2}$ denotes the inverse of this square root.

The Conjugation Rule (2.1.12) ensures that all the matrices involved remain positive definite, so this definition makes sense. To see why the matrix perspective extends the scalar perspective, notice that

$$\boldsymbol{\Psi}_f(\boldsymbol{A}; \boldsymbol{H}) = \boldsymbol{A} \cdot f(\boldsymbol{H}\boldsymbol{A}^{-1}) \quad \text{when } \boldsymbol{A} \text{ and } \boldsymbol{H} \text{ commute.} \tag{8.6.1}$$

This formula is valid because commuting matrices are simultaneously diagonalizable. We will use the matrix perspective in a case where the matrices commute, but it is no harder to analyze the perspective without this assumption.

8.6.2 The Matrix Perspective is Operator Convex

The key result is that the matrix perspective is an operator convex map on a pair of positive-definite matrices. This theorem follows from the operator Jensen inequality in much the same way that Fact 8.2.4 follows from scalar convexity.

Theorem 8.6.2 (Matrix Perspective is Operator Convex)**.** *Let $f : \mathbb{R}_{++} \to \mathbb{R}$ be an operator convex function. Let $\boldsymbol{A}_i$ and $\boldsymbol{H}_i$ be positive-definite matrices of the same size. Then*

$$\boldsymbol{\Psi}_f\big(\tau \boldsymbol{A}_1 + \bar{\tau} \boldsymbol{A}_2;\ \tau \boldsymbol{H}_1 + \bar{\tau} \boldsymbol{H}_2\big) \preccurlyeq \tau \cdot \boldsymbol{\Psi}_f\big(\boldsymbol{A}_1;\ \boldsymbol{H}_1\big) + \bar{\tau} \cdot \boldsymbol{\Psi}_f\big(\boldsymbol{A}_2;\ \boldsymbol{H}_2\big) \quad \textit{for } \tau \in [0, 1].$$

Proof. Let f be an operator convex function, and let $\boldsymbol{\Psi}_f$ be its perspective transform. Fix pairs $(\boldsymbol{A}_1, \boldsymbol{H}_1)$ and $(\boldsymbol{A}_2, \boldsymbol{H}_2)$ of positive-definite matrices, and choose an interpolation parameter $\tau \in [0, 1]$. Form the scalar convex combinations

$$\boldsymbol{A} = \tau \boldsymbol{A}_1 + \bar{\tau} \boldsymbol{A}_2 \quad \text{and} \quad \boldsymbol{H} = \tau \boldsymbol{H}_1 + \bar{\tau} \boldsymbol{H}_2.$$

Our goal is to bound the perspective $\boldsymbol{\Psi}_f(\boldsymbol{A}; \boldsymbol{H})$ as a scalar convex combination of its values $\boldsymbol{\Psi}_f(\boldsymbol{A}_1; \boldsymbol{H}_1)$ and $\boldsymbol{\Psi}_f(\boldsymbol{A}_2; \boldsymbol{H}_2)$. The idea is to introduce matrix interpolation parameters:

$$\boldsymbol{K}_1 = \tau^{1/2} \boldsymbol{A}_1^{1/2} \boldsymbol{A}^{-1/2} \quad \text{and} \quad \boldsymbol{K}_2 = \bar{\tau}^{1/2} \boldsymbol{A}_2^{1/2} \boldsymbol{A}^{-1/2}.$$

Observe that these two matrices decompose the identity:

$$\boldsymbol{K}_1^* \boldsymbol{K}_1 + \boldsymbol{K}_2^* \boldsymbol{K}_2 = \tau \cdot \boldsymbol{A}^{-1/2} \boldsymbol{A}_1 \boldsymbol{A}^{-1/2} + \bar{\tau} \cdot \boldsymbol{A}^{-1/2} \boldsymbol{A}_2 \boldsymbol{A}^{-1/2} = \boldsymbol{A}^{-1/2} \boldsymbol{A} \boldsymbol{A}^{-1/2} = \mathbf{I}.$$

This construction allows us to express the perspective using a matrix convex combination, which gives us access to the operator Jensen inequality.

We calculate that

$$\begin{aligned}\boldsymbol{\Psi}_f(\boldsymbol{A};\boldsymbol{H}) &= \boldsymbol{A}^{1/2}\cdot f\big(\boldsymbol{A}^{-1/2}\boldsymbol{H}\boldsymbol{A}^{-1/2}\big)\cdot\boldsymbol{A}^{1/2}\\ &= \boldsymbol{A}^{1/2}\cdot f\big(\tau\cdot\boldsymbol{A}^{-1/2}\boldsymbol{H}_1\boldsymbol{A}^{-1/2}+\bar\tau\cdot\boldsymbol{A}^{-1/2}\boldsymbol{H}_2\boldsymbol{A}^{-1/2}\big)\cdot\boldsymbol{A}^{1/2}\\ &= \boldsymbol{A}^{1/2}\cdot f\big(\boldsymbol{K}_1^*\boldsymbol{A}_1^{-1/2}\boldsymbol{H}_1\boldsymbol{A}_1^{-1/2}\boldsymbol{K}_1+\boldsymbol{K}_2^*\boldsymbol{A}_2^{-1/2}\boldsymbol{H}_2\boldsymbol{A}_2^{-1/2}\boldsymbol{K}_2\big)\cdot\boldsymbol{A}^{1/2}.\end{aligned}$$

The first line is simply the definition of the matrix perspective. In the second line, we use the definition of $\boldsymbol{H}$ as a scalar convex combination. Third, we introduce the matrix interpolation parameters through the expressions $\tau^{1/2}\boldsymbol{A}^{-1/2}=\boldsymbol{A}_1^{1/2}\boldsymbol{K}_1$ and $\bar\tau^{1/2}\boldsymbol{A}^{-1/2}=\boldsymbol{A}_2^{1/2}\boldsymbol{K}_2$ and their conjugate transposes. To continue the calculation, we apply the operator Jensen inequality, Theorem 8.5.2, to reach

$$\begin{aligned}\boldsymbol{\Psi}_f(\boldsymbol{A};\boldsymbol{H}) &\preccurlyeq \boldsymbol{A}^{1/2}\cdot\big[\boldsymbol{K}_1^*\cdot f\big(\boldsymbol{A}_1^{-1/2}\boldsymbol{H}_1\boldsymbol{A}_1^{-1/2}\big)\cdot\boldsymbol{K}_1+\boldsymbol{K}_2^*\cdot f\big(\boldsymbol{A}_2^{-1/2}\boldsymbol{H}_2\boldsymbol{A}_2^{-1/2}\big)\cdot\boldsymbol{K}_2\big]\cdot\boldsymbol{A}^{1/2}\\ &= \tau\cdot\boldsymbol{A}_1^{1/2}\cdot f\big(\boldsymbol{A}_1^{-1/2}\boldsymbol{H}_1\boldsymbol{A}_1^{-1/2}\big)\cdot\boldsymbol{A}_1^{1/2}+\bar\tau\cdot\boldsymbol{A}_2^{1/2}\cdot f\big(\boldsymbol{A}_2^{-1/2}\boldsymbol{H}_2\boldsymbol{A}_2^{-1/2}\big)\cdot\boldsymbol{A}_2^{1/2}\\ &= \tau\cdot\boldsymbol{\Psi}_f(\boldsymbol{A}_1;\boldsymbol{H}_1)+\bar\tau\cdot\boldsymbol{\Psi}_f(\boldsymbol{A}_2;\boldsymbol{H}_2).\end{aligned}$$

We have also used the Conjugation Rule (2.1.12) to support the first relation. Finally, we recall the definitions of $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$, and we identify the two matrix perspectives. □

8.7 The Kronecker Product

The matrix relative entropy is a function of two matrices. One of the difficulties of analyzing this type of function is that the two matrix arguments do not generally commute with each other. As a consequence, the behavior of the matrix relative entropy depends on the interactions between the eigenvectors of the two matrices. To avoid this problem, we will build matrices that do commute with each other, which simplifies our task considerably.

8.7.1 The Kronecker Product

Our approach is based on an fundamental object from linear algebra. We restrict our attention to the simplest version here.

Definition 8.7.1 (Kronecker Product)**.** *Let $\boldsymbol{A}$ and $\boldsymbol{H}$ be Hermitian matrices with dimension $d\times d$. The* Kronecker product $\boldsymbol{A}\otimes\boldsymbol{H}$ *is the $d^2\times d^2$ Hermitian matrix*

$$\boldsymbol{A}\otimes\boldsymbol{H}=\begin{bmatrix}a_{11}\boldsymbol{H} & \dots & a_{1d}\boldsymbol{H}\\ \vdots & \ddots & \vdots\\ a_{d1}\boldsymbol{H} & \dots & a_{dd}\boldsymbol{H}\end{bmatrix}$$

At first sight, the definition of the Kronecker product may seem strange, but it has many delightful properties. The rest of the section develops the basic facts about this construction.

8.7.2 Linearity Properties

First of all, a Kronecker product with the zero matrix is always zero:

$$\boldsymbol{A}\otimes\boldsymbol{0}=\boldsymbol{0}\otimes\boldsymbol{0}=\boldsymbol{0}\otimes\boldsymbol{H}\quad\text{for all }\boldsymbol{A}\text{ and }\boldsymbol{H}.$$

Next, the Kronecker product is homogeneous in each factor:

$$(\alpha \boldsymbol{A}) \otimes \boldsymbol{H} = \alpha\,(\boldsymbol{A} \otimes \boldsymbol{H}) = \boldsymbol{A} \otimes (\alpha \boldsymbol{H}) \quad \text{for } \alpha \in \mathbb{R}.$$

Furthermore, the Kronecker product is additive in each coordinate:

$$(\boldsymbol{A}_1 + \boldsymbol{A}_2) \otimes \boldsymbol{H} = \boldsymbol{A}_1 \otimes \boldsymbol{H} + \boldsymbol{A}_2 \otimes \boldsymbol{H} \quad \text{and} \quad \boldsymbol{A} \otimes (\boldsymbol{H}_1 + \boldsymbol{H}_2) = \boldsymbol{A} \otimes \boldsymbol{H}_1 + \boldsymbol{A} \otimes \boldsymbol{H}_2.$$

In other words, the Kronecker product is a bilinear operation.

8.7.3 Mixed Products

The Kronecker product interacts beautifully with the usual product of matrices. By direct calculation, we obtain a simple rule for mixed products:

$$(\boldsymbol{A}_1 \otimes \boldsymbol{H}_1)(\boldsymbol{A}_2 \otimes \boldsymbol{H}_2) = (\boldsymbol{A}_1 \boldsymbol{A}_2) \otimes (\boldsymbol{H}_1 \boldsymbol{H}_2). \tag{8.7.1}$$

Since $\mathbf{I} \otimes \mathbf{I}$ is the identity matrix, the identity (8.7.1) leads to a formula for the inverse of a Kronecker product:

$$(\boldsymbol{A} \otimes \boldsymbol{H})^{-1} = \left(\boldsymbol{A}^{-1}\right) \otimes \left(\boldsymbol{H}^{-1}\right) \quad \text{when } \boldsymbol{A} \text{ and } \boldsymbol{H} \text{ are invertible.} \tag{8.7.2}$$

Another important consequence of the rule (8.7.1) is the following commutativity relation:

$$(\boldsymbol{A} \otimes \mathbf{I})(\mathbf{I} \otimes \boldsymbol{H}) = (\mathbf{I} \otimes \boldsymbol{H})(\boldsymbol{A} \otimes \mathbf{I}) \quad \text{for all Hermitian matrices } \boldsymbol{A} \text{ and } \boldsymbol{H}. \tag{8.7.3}$$

This simple fact has great importance for us.

8.7.4 The Kronecker Product of Positive Matrices

As we have noted, the Kronecker product of two Hermitian matrices is itself an Hermitian matrix. In fact, the Kronecker product preserves positivity as well.

Fact 8.7.2 (Kronecker Product Preserves Positivity)**.** *Let $\boldsymbol{A}$ and $\boldsymbol{H}$ be positive-definite matrices. Then $\boldsymbol{A} \otimes \boldsymbol{H}$ is positive definite.*

Proof. To see why, observe that

$$\boldsymbol{A} \otimes \boldsymbol{H} = \left(\boldsymbol{A}^{1/2} \otimes \boldsymbol{H}^{1/2}\right)\left(\boldsymbol{A}^{1/2} \otimes \boldsymbol{H}^{1/2}\right).$$

As usual, $\boldsymbol{A}^{1/2}$ refers to the unique positive-definite square root of the positive-definite matrix $\boldsymbol{A}$. We have expressed $\boldsymbol{A} \otimes \boldsymbol{H}$ as the square of an Hermitian matrix, so it must be a positive-semidefinite matrix. To see that it is actually positive definite, we simply apply the inversion formula (8.7.2) to discover that $\boldsymbol{A} \otimes \boldsymbol{H}$ is invertible. □

8.7.5 The Logarithm of a Kronecker Product

As we have discussed, the matrix logarithm plays a central role in our analysis. There is an elegant formula for the logarithm of a Kronecker product that will be valuable to us.

Fact 8.7.3 (Logarithm of a Kronecker Product)**.** *Let $\boldsymbol{A}$ and $\boldsymbol{H}$ be positive-definite matrices. Then*

$$\log(\boldsymbol{A} \otimes \boldsymbol{H}) = (\log \boldsymbol{A}) \otimes \mathbf{I} + \mathbf{I} \otimes (\log \boldsymbol{H}).$$

Proof. The argument is based on the fact that the matrix logarithm is the functional inverse of the matrix exponential. Since the exponential of a sum of commuting matrices equals the product of the exponentials, we have

$$\exp\big(\boldsymbol{M} \otimes \mathbf{I} + \mathbf{I} \otimes \boldsymbol{T}\big) = \exp(\boldsymbol{M} \otimes \mathbf{I}) \cdot \exp(\mathbf{I} \otimes \boldsymbol{T}).$$

This formula relies on the commutativity relation (8.7.3). Applying the power series representation of the exponential, we determine that

$$\exp(\boldsymbol{M} \otimes \mathbf{I}) = \sum_{q=0}^{\infty} \frac{1}{q!} (\boldsymbol{M} \otimes \mathbf{I})^q = \sum_{q=0}^{\infty} \frac{1}{q!} \big(\boldsymbol{M}^q\big) \otimes \mathbf{I} = \mathrm{e}^{\boldsymbol{M}} \otimes \mathbf{I}.$$

The second identity depends on the rule (8.7.1) for mixed products, and the last identity follows from the linearity of the Kronecker product. A similar calculation shows that $\exp(\mathbf{I} \otimes \boldsymbol{T}) = \mathbf{I} \otimes \mathrm{e}^{\boldsymbol{T}}$. In summary,

$$\exp\big(\boldsymbol{M} \otimes \mathbf{I} + \mathbf{I} \otimes \boldsymbol{T}\big) = \big(\mathrm{e}^{\boldsymbol{M}} \otimes \mathbf{I}\big)\big(\mathbf{I} \otimes \mathrm{e}^{\boldsymbol{T}}\big) = \mathrm{e}^{\boldsymbol{M}} \otimes \mathrm{e}^{\boldsymbol{T}}.$$

We have used the product rule (8.7.1) again. To complete the argument, simply choose $\boldsymbol{M} = \log \boldsymbol{A}$ and $\boldsymbol{T} = \log \boldsymbol{H}$ and take the logarithm of the last identity. □

8.7.6 A Linear Map

Finally, we claim that there is a linear map φ that extracts the trace of the matrix product from the Kronecker product. Let $\boldsymbol{A}$ and $\boldsymbol{H}$ be $d \times d$ Hermitian matrices. Then we define

$$\varphi(\boldsymbol{A} \otimes \boldsymbol{H}) = \operatorname{tr}(\boldsymbol{A}\boldsymbol{H}). \tag{8.7.4}$$

The map φ is linear because the Kronecker product $\boldsymbol{A} \otimes \boldsymbol{H}$ tabulates all the pairwise products of the entries of $\boldsymbol{A}$ and $\boldsymbol{H}$, and $\operatorname{tr}(\boldsymbol{A}\boldsymbol{H})$ is a sum of certain of these pairwise products. For our purposes, the key fact is that the map φ preserves the semidefinite order:

$$\textstyle\sum_i \boldsymbol{A}_i \otimes \boldsymbol{H}_i \succcurlyeq \mathbf{0} \quad \text{implies} \quad \sum_i \varphi(\boldsymbol{A}_i \otimes \boldsymbol{H}_i) \geq 0. \tag{8.7.5}$$

This formula is valid for all Hermitian matrices $\boldsymbol{A}_i$ and $\boldsymbol{H}_i$. To see why (8.7.5) holds, simply note that the map can be represented as an inner product:

$$\varphi(\boldsymbol{A} \otimes \boldsymbol{H}) = \boldsymbol{\iota}^* (\boldsymbol{A} \otimes \boldsymbol{H}) \boldsymbol{\iota} \quad \text{where} \quad \boldsymbol{\iota} := \operatorname{vec}(\mathbf{I}_d).$$

The vec operation stacks the columns of a $d \times d$ matrix on top of each other, moving from left to right, to form a column vector of length d^2.

8.8 The Matrix Relative Entropy is Convex

We are finally prepared to establish Theorem 8.1.4, which states that the matrix relative entropy is a convex function. This argument draws on almost all of the ideas we have developed over the course of this chapter.

Consider the function $f(a) = a - 1 - \log a$, defined on the positive real line. This function is operator convex because it is the sum of the affine function $a \mapsto a - 1$ and the operator convex function $a \mapsto -\log a$. The negative logarithm is operator convex because of Proposition 8.4.8.

Let $\boldsymbol{A}$ and $\boldsymbol{H}$ be positive-definite matrices. Consider the matrix perspective $\boldsymbol{\Psi}_f$ evaluated at the commuting positive-definite matrices $\boldsymbol{A} \otimes \mathbf{I}$ and $\mathbf{I} \otimes \boldsymbol{H}$:

$$\boldsymbol{\Psi}_f(\boldsymbol{A} \otimes \mathbf{I};\, \mathbf{I} \otimes \boldsymbol{H}) = (\boldsymbol{A} \otimes \mathbf{I}) \cdot f\big((\mathbf{I} \otimes \boldsymbol{H})(\boldsymbol{A} \otimes \mathbf{I})^{-1}\big) = (\boldsymbol{A} \otimes \mathbf{I}) \cdot f\big(\boldsymbol{A}^{-1} \otimes \boldsymbol{H}\big).$$

We have used the simplified definition (8.6.1) of the perspective for commuting matrices, and we have invoked the rules (8.7.1) and (8.7.2) for arithmetic with Kronecker products. Introducing the definition of the function f, we find that

$$\begin{aligned}\boldsymbol{\Psi}_f(\boldsymbol{A} \otimes \mathbf{I};\, \mathbf{I} \otimes \boldsymbol{H}) &= (\boldsymbol{A} \otimes \mathbf{I}) \cdot \big[\boldsymbol{A}^{-1} \otimes \boldsymbol{H} - \mathbf{I} \otimes \mathbf{I} - \log\big(\boldsymbol{A}^{-1} \otimes \boldsymbol{H}\big)\big] \\ &= \mathbf{I} \otimes \boldsymbol{H} - \boldsymbol{A} \otimes \mathbf{I} - (\boldsymbol{A} \otimes \mathbf{I}) \cdot \big[\big(\log \boldsymbol{A}^{-1}\big) \otimes \mathbf{I} + \mathbf{I} \otimes \big(\log \boldsymbol{H}\big)\big] \\ &= (\boldsymbol{A} \log \boldsymbol{A}) \otimes \mathbf{I} - \boldsymbol{A} \otimes (\log \boldsymbol{H}) - (\boldsymbol{A} \otimes \mathbf{I} - \mathbf{I} \otimes \boldsymbol{H})\end{aligned}$$

To reach the second line, we use more Kronecker product arithmetic, along with Fact 8.7.3, the law for calculating the logarithm of the Kronecker product. The last line depends on the property that $\log\big(\boldsymbol{A}^{-1}\big) = -\log \boldsymbol{A}$. Applying the linear map φ from (8.7.4) to both sides, we reach

$$\big(\varphi \circ \boldsymbol{\Psi}_f\big)\big(\boldsymbol{A} \otimes \mathbf{I};\, \mathbf{I} \otimes \boldsymbol{H}\big) = \operatorname{tr}\big[\boldsymbol{A} \log \boldsymbol{A} - \boldsymbol{A} \log \boldsymbol{H} - (\boldsymbol{A} - \boldsymbol{H})\big] = \mathrm{D}(\boldsymbol{A}; \boldsymbol{H}). \tag{8.8.1}$$

We have represented the matrix relative entropy in terms of a matrix perspective.

Let $\boldsymbol{A}_i$ and $\boldsymbol{H}_i$ be positive-definite matrices, and fix a parameter $\tau \in [0, 1]$. Theorem 8.6.2 tells us that the matrix perspective is operator convex:

$$\begin{aligned}\boldsymbol{\Psi}_f\big(\tau \cdot (\boldsymbol{A}_1 \otimes \mathbf{I}) + \bar{\tau} \cdot (\boldsymbol{A}_2 \otimes \mathbf{I});\, \tau \cdot (\mathbf{I} \otimes \boldsymbol{H}_1) + \bar{\tau} \cdot (\mathbf{I} \otimes \boldsymbol{H}_2)\big) \\ \preccurlyeq \tau \cdot \boldsymbol{\Psi}_f\big(\boldsymbol{A}_1 \otimes \mathbf{I};\, \mathbf{I} \otimes \boldsymbol{H}_1\big) + \bar{\tau} \cdot \boldsymbol{\Psi}_f\big(\boldsymbol{A}_2 \otimes \mathbf{I};\, \mathbf{I} \otimes \boldsymbol{H}_2\big).\end{aligned}$$

The inequality (8.7.5) states that the linear map φ preserves the semidefinite order.

$$\begin{aligned}\big(\varphi \circ \boldsymbol{\Psi}_f\big)\big(\tau \cdot (\boldsymbol{A}_1 \otimes \mathbf{I}) + \bar{\tau} \cdot (\boldsymbol{A}_2 \otimes \mathbf{I});\, \tau \cdot (\mathbf{I} \otimes \boldsymbol{H}_1) + \bar{\tau} \cdot (\mathbf{I} \otimes \boldsymbol{H}_2)\big) \\ \le \tau \cdot \big(\varphi \circ \boldsymbol{\Psi}_f\big)\big(\boldsymbol{A}_1 \otimes \mathbf{I};\, \mathbf{I} \otimes \boldsymbol{H}_1\big) + \bar{\tau} \cdot \big(\varphi \circ \boldsymbol{\Psi}_f\big)\big(\boldsymbol{A}_2 \otimes \mathbf{I};\, \mathbf{I} \otimes \boldsymbol{H}_2\big).\end{aligned}$$

Introducing the formula (8.8.1), we conclude that

$$\mathrm{D}\big(\tau \boldsymbol{A}_1 + \bar{\tau} \boldsymbol{A}_2;\, \tau \boldsymbol{H}_1 + \bar{\tau} \boldsymbol{H}_2\big) \le \tau \cdot \mathrm{D}\big(\boldsymbol{A}_1; \boldsymbol{H}_1\big) + \bar{\tau} \cdot \mathrm{D}\big(\boldsymbol{A}_2; \boldsymbol{H}_2\big).$$

The matrix relative entropy is convex.

8.9 Notes

The material in this chapter is drawn from a variety of sources, ranging from textbooks to lecture notes to contemporary research articles. The best general sources include the books on matrix analysis by Bhatia [Bha97, Bha07] and by Hiai & Petz [HP14]. We also recommend a set of notes [Car10] by Eric Carlen. More specific references appear below.

8.9.1 Lieb's Theorem

Theorem 8.1.1 is one of the major results in the important paper [Lie73] of Elliott Lieb on convex trace functions. Lieb wrote this paper to resolve a conjecture of Wigner, Yanase, & Dyson about the concavity properties of a certain measure of information in a quantum system. He was also motivated by a conjecture that quantum mechanical entropy satisfies a strong subadditivity property. The latter result states that our uncertainty about a partitioned quantum system is controlled by the uncertainty about smaller parts of the system. See Carlen's notes [Car10] for a modern presentation of these ideas.

Lieb derived Theorem 8.1.1 as a corollary of another difficult concavity theorem that he developed [Lie73, Thm. 1]. The most direct proof of Lieb's Theorem is probably Epstein's argument, which is based on methods from complex analysis [Eps73]; see Ruskai's papers [Rus02, Rus05] for a condensed version of Epstein's approach. The proof that appears in Section 8.1 is due to the author of these notes [Tro12]; this technique depends on ideas developed by Carlen & Lieb to prove some other convexity theorems [CL08, §5].

In fact, many deep convexity and concavity theorems for trace functions are equivalent with each other, in the sense that the mutual implications follow from relatively easy arguments. See [Lie73, §5] and [CL08, §5] for discussion of this point.

8.9.2 The Matrix Relative Entropy

Our definition of matrix relative entropy differs slightly from the usual definition in the literature on quantum statistical mechanics and quantum information theory because we have included an additional linear term. This alteration does not lead to substantive changes in the analysis.

The fact that matrix relative entropy is nonnegative is a classical result attributed to Klein. See [Pet94, §2] or [Car10, §2.3].

Lindblad [Lin73] is credited with the result that matrix relative entropy is convex, as stated in Theorem 8.1.4. Lindblad derived this theorem as a corollary of Lieb's results from [Lie73]. Bhatia [Bha97, Chap. IX] gives two alternative proofs, one due to Connes & Størmer [CS75] and another due to Petz [Pet86]. There is also a remarkable proof due to Ando [And79, Thm. 7].

Our approach to Theorem 8.1.4 is adapted directly from a recent paper of Effros [Eff09]. Nevertheless, many of the ideas date back to the works cited in the last paragraph.

8.9.3 The Relative Entropy for Vectors

The treatment of the relative entropy for vectors in Section 8.2 is based on two classical methods for constructing divergences. To show that the relative entropy is nonnegative, we represent it as a Bregman divergence [Brè67]. To show that the relative entropy is convex, we represent it as an f-divergence [AS66, Csi67]. Let us say a few more words about these constructions.

Suppose that f is a differentiable convex function on $\mathbb{R}^d$. Bregman considered divergences of the form

$$\mathrm{B}_f(\boldsymbol{a}; \boldsymbol{h}) := f(\boldsymbol{a}) - \big[f(\boldsymbol{h}) - \langle \nabla f(\boldsymbol{h}),\ \boldsymbol{a} - \boldsymbol{h} \rangle \big].$$

Since f is convex, the Bregman divergence B_f is always nonnegative. In the vector setting, there are two main examples of Bregman divergences. The function $f(\boldsymbol{a}) = \frac{1}{2}\|\boldsymbol{a}\|_2^2$ leads to the squared Euclidean distance, and the function $f(\boldsymbol{a}) = \sum_i (a_i \log a_i - a_i)$ leads to the vector relative entropy. Bregman divergences have many geometric properties in common with these two functions. For an introduction to Bregman divergences for matrices, see [DT07].

Suppose that $f : \mathbb{R}_{++} \to \mathbb{R}$ is a convex function. Ali & Silvey [AS66] and Csiszár [Csi67] considered divergences of the form

$$\mathrm{C}_f(\boldsymbol{a}; \boldsymbol{h}) := \sum\nolimits_i a_i \cdot f(h_i / a_i).$$

We recognize this expression as a perspective transformation, so the f-divergence C_f is always convex. The main example is based on the Shannon entropy $f(a) = a \log a$, which leads to a cousin of the vector relative entropy. The paper [RW11] contains a recent discussion of f-divergences and their applications in machine learning. Petz has studied functions related to f-divergences in the matrix setting [Pet86, Pet10].

8.9.4 Elementary Trace Inequalities

The material in Section 8.3 on trace functions is based on classical results in quantum statistical mechanics. We have drawn the arguments in this section from Petz's survey [Pet94, Sec. 2] and Carlen's lecture notes [Car10, Sec. 2.2].

8.9.5 Operator Monotone & Operator Convex Functions

The theory of operator monotone functions was initiated by Löwner [Löw34]. He developed a characterization of an operator monotone function in terms of divided differences. For a function f, the *first divided difference* is the quantity

$$f[a, h] := \begin{cases} \dfrac{f(a) - f(h)}{a - h}, & h \neq a \\ f'(a), & h = a. \end{cases}$$

Löwner proved that f is operator monotone on an interval I if and only we have the semidefinite relation

$$\begin{bmatrix} f[a_1, a_1] & \dots & f[a_1, a_d] \\ \vdots & \ddots & \vdots \\ f[a_d, a_1] & \dots & f[a_d, a_d] \end{bmatrix} \succcurlyeq \mathbf{0} \quad \text{for all } \{a_i\} \subset I \text{ and all } d \in \mathbb{N}.$$

This result is analogous with the fact that a smooth, monotone scalar function has a nonnegative derivative. Löwner also established a connection between operator monotone functions and Pick functions from the theory of complex variables. A few years later, Kraus introduced the concept of an operator convex function in [Kra36], and he developed some results that parallel Löwner's theory for operator monotone functions.

Somewhat later, Bendat & Sherman [BS55] developed characterizations of operator monotone and operator convex functions based on integral formulas. For example, f is an operator monotone function on $(0, \infty)$ if and only if it can be written in the form

$$f(t) = \alpha + \beta t + \int_0^\infty \frac{ut}{u + t} \, \mathrm{d}\rho(u) \quad \text{where} \quad \beta \geq 0 \quad \text{and} \quad \int_0^\infty \frac{u}{1 + u} \, \mathrm{d}\rho(u) < \infty.$$

Similarly, f is an operator convex function on $[0, \infty)$ if and only if it can be written in the form

$$f(t) = \alpha + \beta t + \gamma t^2 + \int_0^\infty \frac{ut^2}{u + t} \, \mathrm{d}\rho(u) \quad \text{where} \quad \gamma \geq 0 \quad \text{and} \quad \int_0^\infty \mathrm{d}\rho(u) < \infty.$$

In both cases, $\mathrm{d}\rho$ is a nonnegative measure. The integral representation of the logarithm in Proposition 8.4.1 is closely related to these formulas.

We have taken the proof that the matrix inverse is monotone from Bhatia's book [Bha97, Prop. V.1.6]. The proof that the matrix inverse is convex appears in Ando's paper [And79]. Our treatment of the matrix logarithm was motivated by a conversation with Eric Carlen at an IPAM workshop at Lake Arrowhead in December 2010.

For more information about operator monotonicity and operator convexity, we recommend Bhatia's books [Bha97, Bha07], Carlen's lecture notes [Car10], and the book of Hiai & Petz [HP14].

8.9.6 The Operator Jensen Inequality

The paper [HP82] of Hansen & Pedersen contains another treatment of operator monotone and operator convex functions. The highlight of this work is a version of the operator Jensen inequality. Theorem 8.5.2 is a refinement of this result that was established by the same authors two decades later [HP03]. Our proof of the operator Jensen inequality is drawn from Petz's book [Pet11, Thm. 8.4]; see also Carlen's lecture notes [Car10, Thm. 4.20].

8.9.7 The Matrix Perspective & the Kronecker Product

We have been unable to identify the precise source of the idea that a bivariate matrix function can be represented in terms of a matrix perspective. Two important results in this direction appear in Ando's paper [And79, Thms. 6 and 7].

$$f \text{ positive and operator concave on } (0,\infty) \quad \text{implies} \quad (\boldsymbol{A}, \boldsymbol{H}) \mapsto (\boldsymbol{A} \otimes \mathbf{I}) \cdot f\big(\boldsymbol{A}^{-1} \otimes \boldsymbol{H}\big) \quad \text{is operator concave}$$

on pairs of positive-definite matrices. Similarly,

$$f \text{ operator monotone on } (0,\infty) \quad \text{implies} \quad (\boldsymbol{A}, \boldsymbol{H}) \mapsto (\boldsymbol{A} \otimes \mathbf{I}) \cdot f\big(\boldsymbol{A} \otimes \boldsymbol{H}^{-1}\big) \quad \text{is operator convex}$$

on pairs of positive-definite matrices. Ando proves that the matrix relative entropy is convex by applying the latter result to the matrix logarithm. We believe that Ando was the first author to appreciate the value of framing results of this type in terms of the Kronecker product, and we have followed his strategy here. On the other hand, Ando's analysis is different in spirit because he relies on integral representations of operator monotone and convex functions.

In a subsequent paper [KA80], Kubo & Ando constructed operator means using a related approach. They show that

$$f \text{ positive and operator monotone on } (0,\infty) \quad \text{implies} \quad (\boldsymbol{A}, \boldsymbol{H}) \mapsto \boldsymbol{A}^{1/2} \cdot f\big(\boldsymbol{A}^{-1/2} \boldsymbol{H} \boldsymbol{A}^{-1/2}\big) \cdot \boldsymbol{A}^{1/2} \quad \text{is operator concave}$$

on pairs of positive-definite matrices. Kubo & Ando point out that particular cases of this construction appear in the work of Pusz & Woronowicz [PW75]. This is the earliest citation where we have seen the matrix perspective black-on-white.

A few years later, Petz introduced a class of quasi-entropies for matrices [Pet86]. These functions also involve a perspective-like construction, and Petz was clearly influenced by Csiszár's work on f-divergences. See [Pet10] for a contemporary treatment.

The presentation in these notes is based on a recent paper [Eff09] of Effros. He showed that convexity properties of the matrix perspective follow from the operator Jensen inequality, and he

derived the convexity of the matrix relative entropy as a consequence. Our analysis of the matrix perspective in Theorem 8.6.2 is drawn from a subsequent paper [ENG11], which removes some commutativity assumptions from Effros's argument.

The proof in §8.8 that the matrix relative entropy is convex, Theorem 8.1.4, recasts Effros's argument [Eff09, Cor. 2.2] in the language of Kronecker products. In his paper, Effros works with left- and right-multiplication operators. To appreciate the connection, simply note the identities

$$(\boldsymbol{A} \otimes \mathbf{I})\,\mathrm{vec}(\boldsymbol{M}) = \mathrm{vec}(\boldsymbol{M}\boldsymbol{A}). \quad \text{and} \quad (\mathbf{I} \otimes \boldsymbol{H})\,\mathrm{vec}(\boldsymbol{M}) = \mathrm{vec}(\boldsymbol{H}\boldsymbol{M}).$$

In other words, the matrix $\boldsymbol{A} \otimes \mathbf{I}$ can be interpreted as right-multiplication by $\boldsymbol{A}$, while the matrix $\mathbf{I} \otimes \boldsymbol{H}$ can be interpreted as left-multiplication by $\boldsymbol{H}$. (The change in sense is an unfortunate consequence of the definition of the Kronecker product.)

Matrix Concentration: Resources

This annotated bibliography describes some papers that involve matrix concentration inequalities. Right now, this presentation is heavily skewed toward theoretical results, rather than applications of matrix concentration.

Exponential Matrix Concentration Inequalities

We begin with papers that contain the most current results on matrix concentration.

- [Tro11c]. These lecture notes are based heavily on the research described in this paper. This work identifies Lieb's Theorem [Lie73, Thm. 6] as the key result that animates exponential moment bounds for random matrices. Using this technique, the paper develops the bounds for matrix Gaussian and Rademacher series, the matrix Chernoff inequalities, and several versions of the matrix Bernstein inequality. In addition, it contains a matrix Hoeffding inequality (for sums of bounded random matrices), a matrix Azuma inequality (for matrix martingales with bounded differences), and a matrix bounded difference inequality (for matrix-valued functions of independent random variables).

- [Tro12]. This note describes a simple proof of Lieb's Theorem that is based on the joint convexity of quantum relative entropy. This reduction, however, still involves a deep convexity theorem. Chapter 8 contains an explication of this paper.

- [Oli10a]. Oliveira's paper uses an ingenious argument, based on the Golden–Thompson inequality (3.3.3), to establish a matrix version of Freedman's inequality. This result is, roughly, a martingale version of Bernstein's inequality. This approach has the advantage that it extends to the fully noncommutative setting [JZ12]. Oliveira applies his results to study some problems in random graph theory.

- [Tro11a]. This paper shows that Lieb's Theorem leads to a Freedman-type inequality for matrix-valued martingales. The associated technical report [Tro11b] describes additional results for matrix-valued martingales.

- [GT14]. This article explains how to use the Lieb–Seiringer Theorem [LS05] to develop tail bounds for the interior eigenvalues of a sum of independent random matrices. It contains a Chernoff-type bound for a sum of positive-semidefinite matrices, as well as several Bernstein-type bounds for sums of bounded random matrices.

- [MJC+14]. This paper contains a strikingly different method for establishing matrix concentration inequalities. The argument is based on work of Sourav Chatterjee [Cha07] that

shows how Stein's method of exchangeable pairs [Ste72] leads to probability inequalities. This technique has two main advantages. First, it gives results for random matrices that are based on dependent random variables. As a special case, the results apply to sums of independent random matrices. Second, it delivers both exponential moment bounds and polynomial moment bounds for random matrices. Indeed, the paper describes a Bernstein-type exponential inequality and also a Rosenthal-type polynomial moment bound. Furthermore, this work contains what is arguably the simplest known proof of the noncommutative Khintchine inequality.

- [PMT14]. This paper improves on the work in [MJC+14] by extending an argument, based on Markov chains, that was developed in Chatterjee's thesis [Cha05]. This analysis leads to satisfactory matrix analogs of scalar concentration inequalities based on logarithmic Sobolev inequalities. In particular, it is possible to develop a matrix version of the exponential Efron–Stein inequality in this fashion.

- [CGT12a, CGT12b]. The primary focus of this paper is to analyze a specific type of procedure for covariance estimation. The appendix contains a new matrix moment inequality that is, roughly, the polynomial moment bound associated with the matrix Bernstein inequality.

- [Kol11]. These lecture notes use matrix concentration inequalities as a tool to study some estimation problems in statistics. They also contain some matrix Bernstein inequalities for unbounded random matrices.

- [GN]. Gross and Nesme show how to extend Hoeffding's method for analyzing sampling without replacement to the matrix setting. This result can be combined with a variety of matrix concentration inequalities.

- [Tro11d]. This paper combines the matrix Chernoff inequality, Theorem 5.1.1, with the argument from [GN] to obtain a matrix Chernoff bound for a sum of random positive-semidefinite matrices sampled without replacement from a fixed collection. The result is applied to a random matrix that plays a role in numerical linear algebra.

- [CT14]. This paper establishes logarithmic Sobolev inequalities for random matrices, and it derives some matrix concentration inequalities as a consequence. The methods in the paper have applications in quantum information theory, although the matrix concentration bounds are inferior to related results derived using Stein's method.

Bounds with Intrinsic Dimension Parameters

The following works contain matrix concentration bounds that depend on a dimension parameter that may be smaller than the ambient dimension of the matrix.

- [Oli10b]. Oliveira shows how to develop a version of Rudelson's inequality [Rud99] using a variant of the argument of Ahlswede & Winter from [AW02]. Oliveira's paper is notable because the dimensional factor is controlled by the maximum rank of the random matrix, rather than the ambient dimension.

- [MZ11]. This work contains a matrix Chernoff bound for a sum of independent positive-semidefinite random matrices where the dimensional dependence is controlled by the maximum rank of the random matrix. The approach is, essentially, the same as the argument in Rudelson's paper [Rud99]. The paper applies these results to study randomized matrix multiplication algorithms.
- [HKZ12]. This paper describes a method for proving matrix concentration inequalities where the ambient dimension is replaced by the intrinsic dimension of the matrix variance. The argument is based on an adaptation of the proof in [Tro11a]. The authors give several examples in statistics and machine learning.
- [Min11]. This work presents a more refined technique for obtaining matrix concentration inequalities that depend on the intrinsic dimension, rather than the ambient dimension. This paper motivated the results in Chapter 7.

The Method of Ahlswede & Winter

Next, we list some papers that use the ideas from the work [AW02] of Ahslwede & Winter to obtain matrix concentration inequalities. In general, these results have suboptimal parameters, but they played an important role in the development of this field.

- [AW02]. The original paper of Ahlswede & Winter describes the matrix Laplace transform method, along with a number of other foundational results. They show how to use the Golden–Thompson inequality to bound the trace of the matrix mgf, and they use this technique to prove a matrix Chernoff inequality for sums of independent and identically distributed random variables. Their main application concerns quantum information theory.
- [CM08]. Christofides and Markström develop a Hoeffding-type inequality for sums of bounded random matrices using the approach of Ahlswede & Winter. They apply this result to study random graphs.
- [Gro11]. Gross presents a matrix Bernstein inequality based on the method of Ahlswede & Winter, and he uses it to study algorithms for matrix completion.
- [Rec11]. Recht describes a different version of the matrix Bernstein inequality, which also follows from the technique of Ahlswede & Winter. His paper also concerns algorithms for matrix completion.

Noncommutative Moment Inequalities

We conclude with an overview of some major works on bounds for the polynomial moments of a noncommutative martingale. Sums of independent random matrices provide one concrete example where these results apply. The results in this literature are as strong, or stronger, than the exponential moment inequalities that we have described in these notes. Unfortunately, the proofs are typically quite abstract and difficult, and they do not usually lead to explicit constants. Recently there has been some cross-fertilization between noncommutative probability and the field of matrix concentration inequalities.

Note that "noncommutative" is not synonymous with "matrix" in that there are noncommutative von Neumann algebras much stranger than the familiar algebra of finite-dimensional matrices equipped with the operator norm.

- [TJ74]. This classic paper gives a bound for the expected trace of an even power of a matrix Rademacher series. These results are important, but they do not give the optimal bounds.
- [LP86]. This paper gives the first noncommutative Khintchine inequality, a bound for the expected trace of an even power of a matrix Rademacher series that depends on the matrix variance.
- [LPP91]. This work establishes dual versions of the noncommutative Khintchine inequality.
- [Buc01, Buc05]. These papers prove optimal noncommutative Khintchine inequalities in more general settings, and they obtain sharp constants.
- [JX03, JX08]. These papers establish noncommutative versions of the Burkholder–Davis–Gundy inequality for martingales. They also give an application of these results to random matrix theory.
- [JX05]. This paper contains an overview of noncommutative moment results, along with information about the optimal rate of growth in the constants.
- [JZ13]. This paper describes a fully noncommutative version of the Bennett inequality. The proof is based on the method of Ahlswede & Winter [AW02].
- [JZ12]. This work shows how to use Oliveira's argument [Oli10a] to obtain some results for fully noncommutative martingales.
- [MJC+14]. This work, described above, includes a section on matrix moment inequalities. This paper contains what are probably the simplest available proofs of these results.
- [CGT12a]. The appendix of this paper contains a polynomial inequality for sums of independent random matrices.

Bibliography

[ABH14] E. Abbé, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. Available at http://arXiv.org/abs/1405.3267, June 2014.

[AC09] N. Ailon and B. Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.

[AHK06] S. Arora, E. Hazan, and S. Kale. A fast random sampling algorithm for sparsifying matrices. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 272–279, 2006.

[AKL13] D. Achlioptas, Z. Karnin, and E. Liberty. Near-optimal entrywise sampling for data matrices. In *Advances in Neural Information Processing Systems 26*, 2013.

[ALMT14] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. *Inform. Inference*, 3(3):224–294, 2014. Preprint available at http://arXiv.org/abs/1303.6672.

[AM01] D. Achlioptas and F. McSherry. Fast computation of low rank matrix approximations. In *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, pages 611–618 (electronic). ACM, New York, 2001.

[AM07] D. Achlioptas and F. McSherry. Fast computation of low-rank matrix approximations. *J. Assoc. Comput. Mach.*, 54(2):Article 10, 2007. (electronic).

[And79] T. Ando. Concavity of certain maps on positive definite matrices and applications to Hadamard products. *Linear Algebra Appl.*, 26:203–241, 1979.

[AS66] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B*, 28:131–142, 1966.

[AS00] N. Alon and J. H. Spencer. *The probabilistic method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience [John Wiley & Sons], New York, second edition, 2000. With an appendix on the life and work of Paul Erdős.

[AW02] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, Mar. 2002.

[Bar93] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945, May 1993.

[Bar14] S. Barman. An approximate version of Carathéodory's theorem with applications to approximating Nash equilibria and dense bipartite subgraphs. Available at http://arXiv.org/abs/1406.2296, June 2014.

[BBLM05] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2):514–560, 2005.

[BDJ06] W. Bryc, A. Dembo, and T. Jiang. Spectral measure of large random Hankel, Markov and Toeplitz matrices. *Ann. Probab.*, 34(1):1–38, 2006.

[Bha97] R. Bhatia. *Matrix Analysis*. Number 169 in Graduate Texts in Mathematics. Springer, Berlin, 1997.

[Bha07] R. Bhatia. *Positive Definite Matrices*. Princeton Univ. Press, Princeton, NJ, 2007.

[BLM03] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *Ann. Probab.*, 31(3):1583–1614, 2003.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[Bou85] J. Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel J. Math.*, 52(1-2):46–52, 1985.

[Brè67] L. M. Brègman. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 7:620–631, 1967.

[BS55] J. Bendat and S. Sherman. Monotone and convex operator functions. *Trans. Amer. Math. Soc.*, 79:58–71, 1955.

[BS10] Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, second edition, 2010.

[BT87] J. Bourgain and L. Tzafriri. Invertibility of "large" submatrices with applications to the geometry of Banach spaces and harmonic analysis. *Israel J. Math.*, 57(2):137–224, 1987.

[BT91] J. Bourgain and L. Tzafriri. On a problem of Kadison and Singer. *J. Reine Angew. Math.*, 420:1–43, 1991.

[BTN01] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA, 2001. Analysis, algorithms, and engineering applications.

[Buc01] A. Buchholz. Operator Khintchine inequality in non-commutative probability. *Math. Ann.*, 319:1–16, 2001.

[Buc05] A. Buchholz. Optimal constants in Khintchine-type inequalities for Fermions, Rademachers and q-Gaussian operators. *Bull. Pol. Acad. Sci. Math.*, 53(3):315–321, 2005.

[BV14] A. S. Bandeira and R. Van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. Available at `http://arXiv.org/abs/1408.6185`, Aug. 2014.

[BvdG11] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[BY93] Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.*, 21(3):1275–1294, 1993.

[Car85] B. Carl. Inequalities of Bernstein–Jackson-type and the degree of compactness in Banach spaces. *Ann. Inst. Fourier (Grenoble)*, 35(3):79–118, 1985.

[Car10] E. Carlen. Trace inequalities and quantum entropy: an introductory course. In *Entropy and the quantum*, volume 529 of *Contemp. Math.*, pages 73–140. Amer. Math. Soc., Providence, RI, 2010.

[CD12] S. Chrétien and S. Darses. Invertibility of random submatrices via tail-decoupling and a matrix Chernoff inequality. *Statist. Probab. Lett.*, 82(7):1479–1487, 2012.

[CGT12a] R. Y. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: An analysis using matrix concentration inequalities. *Inform. Inference*, 1(1), 2012. doi:10.1093/imaiai/ias001.

[CGT12b] R. Y. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: An analysis using matrix concentration inequalities. ACM Report 2012-01, California Inst. Tech., Pasadena, CA, Feb. 2012.

[Cha05] S. Chatterjee. *Concentration Inequalities with Exchangeable Pairs*. ProQuest LLC, Ann Arbor, MI, 2005. Thesis (Ph.D.)–Stanford University.

[Cha07] S. Chatterjee. Stein's method for concentration inequalities. *Probab. Theory Related Fields*, 138:305–321, 2007.

[Che52] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statistics*, 23:493–507, 1952.

[CL08] E. A. Carlen and E. H. Lieb. A Minkowski type trace inequality and strong subadditivity of quantum entropy. II. Convexity and concavity. *Lett. Math. Phys.*, 83(2):107–126, 2008.

[CM08] D. Cristofides and K. Markström. Expansion properties of random Cayley graphs and vertex transitive graphs via matrix martingales. *Random Structures Algs.*, 32(8):88–100, 2008.

[CRPW12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The Convex Geometry of Linear Inverse Problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

[CRT06] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[CS75] A. Connes and E. Størmer. Entropy for automorphisms of II_1 von Neumann algebras. *Acta Math.*, 134(3-4):289–306, 1975.

[Csi67] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[CT14] R. Y. Chen and J. A. Tropp. Subadditivity of matrix φ-entropy and concentration of random matrices. *Electron. J. Probab.*, 19(27):1–30, 2014.

[CW13] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *STOC'13—Proceedings of the 2013 ACM Symposium on Theory of Computing*, pages 81–90. ACM, New York, 2013.

[d'A11] A. d'Asprémont. Subsampling algorithms for semidefinite programming. *Stoch. Syst.*, 1(2):274–305, 2011.

[DFK+99] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. In *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms (Baltimore, MD, 1999)*, pages 291–299. ACM, New York, 1999.

[DK01] P. Drineas and R. Kannan. Fast Monte Carlo algorithms for approximate matrix multiplication. In *Proc. 42nd IEEE Symp. Foundations of Computer Science (FOCS)*, pages 452–259, 2001.

[DKM06] P. Drineas, R. Kannan, and M. W. Mahoney. Fast Monte Carlo algorithms for matrices. I. Approximating matrix multiplication. *SIAM J. Comput.*, 36(1):132–157, 2006.

[DM05] P. Drineas and M. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.*, 6:2153–2175, 2005.

[Don06] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, Apr. 2006.

[DS02] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices, and Banach spaces. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook of Banach Space Geometry*, pages 317–366. Elsevier, Amsterdam, 2002.

[DT07] I. S. Dhillon and J. A. Tropp. Matrix nearness problems with Bregman divergences. *SIAM J. Matrix Anal. Appl.*, 29(4):1120–1146, 2007.

[DZ11] P. Drineas and A. Zouzias. A note on element-wise matrix sparsification via a matrix-valued Bernstein inequality. *Inform. Process. Lett.*, 111(8):385–389, 2011.

[Eff09] E. G. Effros. A matrix convexity approach to some celebrated quantum inequalities. *Proc. Natl. Acad. Sci. USA*, 106(4):1006–1008, Jan. 2009.

[ENG11] A. Ebadian, I. Nikoufar, and M. E. Gordji. Perspectives of matrix convex functions. *Proc. Natl. Acad. Sci. USA*, 108(18):7313–7314, 2011.

[Eps73] H. Epstein. Remarks on two theorems of E. Lieb. *Comm. Math. Phys.*, 31:317–325, 1973.

[ER60] P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.

[Fel68] W. Feller. *An introduction to probability theory and its applications. Vol. I.* Third edition. John Wiley & Sons, Inc., New York-London-Sydney, 1968.

[Fel71] W. Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons, Inc., New York-London-Sydney, 1971.

[Fer75] X. Fernique. Regularité des trajectoires des fonctions aléatoires gaussiennes. In *École d'Été de Probabilités de Saint-Flour, IV-1974*, pages 1–96. Lecture Notes in Math., Vol. 480. Springer, Berlin, 1975.

[FKV98] A. Frieze, R. Kannan, and S. Vempala. Fast Monte Carlo algorithms for finding low-rank approximations. In *Proc. 39th Ann. IEEE Symp. Foundations of Computer Science (FOCS)*, pages 370–378, 1998.

[For10] P. J. Forrester. *Log-gases and random matrices*, volume 34 of *London Mathematical Society Monographs Series*. Princeton University Press, Princeton, NJ, 2010.

[FR13] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

[Fre75] D. A. Freedman. On tail probabilities for martingales. *Ann. Probab.*, 3(1):100–118, Feb. 1975.

[GGI+02] A. C. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. Strauss. Near-optimal sparse Fourier representations via sampling. In *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, pages 152–161. ACM, New York, 2002.

[GM14] A. Gittens and M. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *J. Mach. Learn. Res.*, 2014. To appear. Preprint available at http://arXiv.org/abs/1303.1849.

[GN] D. Gross and V. Nesme. Note on sampling without replacing from a finite collection of matrices. Available at http://arXiv.org/abs/1001.2738.

[Gor85] Y. Gordon. Some inequalities for Gaussian processes and applications. *Israel J. Math.*, 50(4):265–289, 1985.

[GR01] C. Godsil and G. Royle. *Algebraic Graph Theory*. Number 207 in Graduate Texts in Mathematics. Springer, 2001.

[Grc11] J. F. Grcar. John von Neumann's analysis of Gaussian elimination and the origins of modern numerical analysis. *SIAM Rev.*, 53(4):607–682, 2011.

[Gro11] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 57(3):1548–1566, Mar. 2011.

[GS01] G. R. Grimmett and D. R. Stirzaker. *Probability and random processes*. Oxford University Press, New York, third edition, 2001.

[GT09] A. Gittens and J. A. Tropp. Error bounds for random matrix approximation schemes. ACM Report 2014-01, California Inst. Tech., Nov. 2009. Available at http://arXiv.org/abs/0911.4108.

[GT14] A. Gittens and J. A. Tropp. Tail bounds for all eigenvalues of a sum of random matrices. ACM Report 2014-02, California Inst. Tech., 2014. Available at http://arXiv.org/abs/1104.4513.

[GvN51] H. H. Goldstine and J. von Neumann. Numerical inverting of matrices of high order. II. *Proc. Amer. Math. Soc.*, 2:188–202, 1951.

[Has09] M. B. Hastings. Superadditivity of communication complexity using entangled inputs. *Nature Phys.*, 5:255–257, 2009.

[Hig08] N. J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008.

[HJ94] R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

[HJ13] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge Univ. Press, 2nd edition, 2013.

[HKZ12] D. Hsu, S. M. Kakade, and T. Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electron. Commun. Probab.*, 17:no. 14, 13, 2012.

[HLL83] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[HMT11] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, June 2011.

[HP82] F. Hansen and G. K. Pedersen. Jensen's inequality for operators and Löwner's theorem. *Math. Ann.*, 258(3):229–241, 1982.

[HP03] F. Hansen and G. K. Pedersen. Jensen's operator inequality. *Bull. London Math. Soc.*, 35(4):553–564, 2003.

[HP14] F. Hiai and D. Petz. *Introduction to Matrix Analysis and Applications*. Springer, Feb. 2014.

[HW08] P. Hayden and A. Winter. Counterexamples to the maximal p-norm multiplicity conjecture for all $p > 1$. *Comm. Math. Phys.*, 284(1):263–280, 2008.

[HXGD14] R. Hamid, Y. Xiao, A. Gittens, and D. DeCoste. Compact random feature maps. In *Proc. 31st Intl. Conf. Machine Learning*, Beijing, July 2014.

[JL84] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984.

[JX03] M. Junge and Q. Xu. Noncommutative Burkholder/Rosenthal inequalities. *Ann. Probab.*, 31(2):948–995, 2003.

[JX05] M. Junge and Q. Xu. On the best constants in some non-commutative martingale inequalities. *Bull. London Math. Soc.*, 37:243–253, 2005.

[JX08] M. Junge and Q. Xu. Noncommutative Burkholder/Rosenthal inequalities II: Applications. *Israel J. Math.*, 167:227–282, 2008.

[JZ12] M. Junge and Q. Zeng. Noncommutative martingale deviation and Poincaré type inequalities with applications. Available at http://arXiv.org/abs/1211.3209, Nov. 2012.

[JZ13] M. Junge and Q. Zeng. Noncommutative Bennett and Rosenthal inequalities. *Ann. Probab.*, 41(6):4287–4316, 2013.

[KA80] F. Kubo and T. Ando. Means of positive linear operators. *Math. Ann.*, 246(3):205–224, 1979/80.

[KD14] A. Kundu and P. Drineas. A note on randomized element-wise matrix sparsification. Available at http://arXiv.org/abs/1404.0320, Apr. 2014.

[Kem13] T. Kemp. Math 247a: Introduction to random matrix theory. Available at http://www.math.ucsd.edu/~tkemp/247A/247A.Notes.pdf, 2013.

[KK12] P. Kar and H. Karnick. Random feature maps for dot product kernels. In *Proc. 15th Intl. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2012.

[KM13] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. Available at http://arXiv.org/abs/1312.3580, Dec. 2013.

[Kol11] V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

[Kra36] F. Kraus. Über konvexe Matrixfunktionen. *Math. Z.*, 41(1):18–42, 1936.

[KT94] B. Kashin and L. Tzafriri. Some remarks on coordinate restriction of operators to coordinate subspaces. Insitute of Mathematics Preprint 12, Hebrew University, Jerusalem, 1993–1994.

[Lat05] R. Latała. Some estimates of norms of random matrices. *Proc. Amer. Math. Soc.*, 133(5):1273–1282, 2005.

[LBW96] W. S. Lee, P. L. Bartlett, and R. C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Inform. Theory*, 42(6):2118–2132, Nov. 1996.

[Lie73] E. H. Lieb. Convex trace functions and the Wigner–Yanase–Dyson conjecture. *Adv. Math.*, 11:267–288, 1973.

[Lin73] G. Lindblad. Entropy, information and quantum measurements. *Comm. Math. Phys.*, 33:305–322, 1973.

[LLR95] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

[Löw34] K. Löwner. Über monotone Matrixfunktionen. *Math. Z.*, 38(1):177–216, 1934.

[LP86] F. Lust-Piquard. Inégalités de Khintchine dans C_p $(1 < p < \infty)$. *C. R. Math. Acad. Sci. Paris*, 303(7):289–292, 1986.

[LPP91] F. Lust-Piquard and G. Pisier. Noncommutative Khintchine and Paley inequalities. *Ark. Mat.*, 29(2):241–260, 1991.

[LPSS+14] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schölkopf. Randomized nonlinear component analysis. In *Proc. 31st Intl. Conf. Machine Learning*, Beijing, July 2014.

[LS05] E. H. Lieb and R. Seiringer. Stronger subadditivity of entropy. *Phys. Rev. A*, 71:062329–1–9, 2005.

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin, 1991.

[Lug09] G. Lugosi. Concentration-of-measure inequalities. Available at http://www.econ.upf.edu/~lugosi/anu.pdf, 2009.

[Mah11] M. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learning*, 3(2):123–224, Feb. 2011.

[Mau03] A. Maurer. A bound on the deviation probability for sums of non-negative random variables. *JIPAM. J. Inequal. Pure Appl. Math.*, 4(1):Article 15, 6 pp. (electronic), 2003.

[Mec07] M. W. Meckes. On the spectral norm of a random Toeplitz matrix. *Electron. Comm. Probab.*, 12:315–325 (electronic), 2007.

[Meh04] M. L. Mehta. *Random matrices*, volume 142 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, third edition, 2004.

[Mil71] V. D. Milman. A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies. *Funkcional. Anal. i Priložen.*, 5(4):28–37, 1971.

[Min11] S. Minsker. Some extensions of Bernstein's inequality for self-adjoint operators. Available at http://arXiv.org/abs/1112.5448, Nov. 2011.

[MJC+14] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. Matrix concentration inequalities via the method of exchangable pairs. *Ann. Probab.*, 42(3):906–945, 2014. Preprint available at http://arXiv.org/abs/1201.6002.

[MKB79] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York-Toronto, Ont., 1979. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

[Mon73] H. L. Montgomery. The pair correlation of zeros of the zeta function. In *Analytic number theory (Proc. Sympos. Pure Math., Vol. XXIV, St. Louis Univ., St. Louis, Mo., 1972)*, pages 181–193. Amer. Math. Soc., Providence, R.I., 1973.

[MP67] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72 (114):507–536, 1967.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge Univ. Press, Cambridge, 1995.

[MSS14] A. Marcus, D. A. Spielman, and N. Srivastava. Interlacing families II: Mixed characteristic polynomials and the Kadison–Singer problem. *Ann. Math.*, June 2014. To appear. Preprint available at http://arXiv.org/abs/1306.3969.

[MT13] M. B. McCoy and J. A. Tropp. The achievable performance of convex demixing. Available at http://arXiv.org/abs/1309.7478, Sep. 2013.

[MT14] M. McCoy and J. A. Tropp. Sharp recovery thresholds for convex deconvolution, with applications. *Found. Comput. Math.*, Apr. 2014. Preprint available at http://arXiv.org/abs/1205.1580.

[Mui82] R. J. Muirhead. *Aspects of multivariate statistical theory*. John Wiley & Sons Inc., New York, 1982. Wiley Series in Probability and Mathematical Statistics.

[MZ11] A. Magen and A. Zouzias. Low rank matrix-valued Chernoff bounds and approximate matrix multiplication. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1422–1436. SIAM, Philadelphia, PA, 2011.

[Nem07] A. Nemirovski. Sums of random symmetric matrices and quadratic optimization under orthogonality constraints. *Math. Prog. Ser. B*, 109:283–317, 2007.

[NRV13] A. Naor, O. Regev, and T. Vidick. Efficient rounding for the noncommutative Grothendieck inequality (extended abstract). In *STOC'13—Proceedings of the 2013 ACM Symposium on Theory of Computing*, pages 71–80. ACM, New York, 2013.

[NS06] A. Nica and R. Speicher. *Lectures on the combinatorics of free probability*, volume 335 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2006.

[NT14] D. Needell and J. A. Tropp. Paved with good intentions: analysis of a randomized block Kaczmarz method. *Linear Algebra Appl.*, 441:199–221, 2014.

[Oli10a] R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Available at http://arXiv.org/abs/0911.0600, Feb. 2010.

[Oli10b] R. I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electron. Commun. Probab.*, 15:203–212, 2010.

[Oli11] R. I. Oliveira. The spectrum of random k-lifts of large graphs (with possibly large k). *J. Combinatorics*, 1(3/4):285–306, 2011.

[Oli13] R. I. Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. Available at http://arXiv.org/abs/1312.2903, Dec. 2013.

[Par98] B. N. Parlett. *The symmetric eigenvalue problem*, volume 20 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Corrected reprint of the 1980 original.

[Pet86] D. Petz. Quasi-entropies for finite quantum systems. *Rep. Math. Phys.*, 23(1):57–65, 1986.

[Pet94] D. Petz. A survey of certain trace inequalities. In *Functional analysis and operator theory (Warsaw, 1992)*, volume 30 of *Banach Center Publ.*, pages 287–298. Polish Acad. Sci., Warsaw, 1994.

[Pet10] D. Petz. From f-divergence to quantum quasi-entropies and their use. *Entropy*, 12(3):304–325, 2010.

[Pet11] D. Petz. Matrix analysis with some applications. Available at bolyai.cs.elte.hu/~petz/matrixbme.pdf, Feb. 2011.

[Pin94] I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab.*, 22(4):1679–1706, 1994.

[Pis81] G. Pisier. Remarques sur un résultat non publié de B. Maurey. In *Seminar on Functional Analysis, 1980–1981*, pages Exp. No. V, 13. École Polytech., Palaiseau, 1981.

[Pis89] G. Pisier. *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.

[PMT14] D. Paulin, L. Mackey, and J. A. Tropp. Efron–Stein inequalities for random matrices. Available at http://arXiv.org/abs/1408.3470, Aug. 2014.

[PW75] W. Pusz and S. L. Woronowicz. Functional calculus for sesquilinear forms and the purification map. *Rep. Mathematical Phys.*, 8(2):159–170, 1975.

[Rec11] B. Recht. A simpler approach to matrix completion. *J. Mach. Learn. Res.*, 12:3413–3430, 2011.

[Ros70] H. P. Rosenthal. On subspaces of L_p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 8:273–303, 1970.

[RR07] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184, Vancouver, Dec. 2007.

[RR08] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems 21*, 2008.

[RS13] S. Riemer and C. Schütt. On the expectation of the norm of random matrices with non-identically distributed entries. *Electron. J. Probab.*, 18:no. 29, 13, 2013.

[Rud99] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164:60–72, 1999.

[Rus02] M. B. Ruskai. Inequalities for quantum entropy: A review with conditions for equality. *J. Math. Phys.*, 43(9):4358–4375, Sep. 2002.

[Rus05] M. B. Ruskai. Erratum: Inequalities for quantum entropy: A review with conditions for equality [*J. Math. Phys.* 43, 4358 (2002)]. *J. Math. Phys.*, 46(1):0199101, 2005.

[RV06] M. Rudelson and R. Vershynin. Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. In *Proc. 40th Ann. Conf. Information Sciences and Systems (CISS)*, Mar. 2006.

[RV07] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. Assoc. Comput. Mach.*, 54(4):Article 21, 19 pp., Jul. 2007. (electronic).

[RW11] M. D. Reid and R. C. Williamson. Information, divergence and risk for binary experiments. *J. Mach. Learn. Res.*, 12:731–817, 2011.

[Sar06] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proc. 47th Ann. IEEE Symp. Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[Seg00] Y. Seginer. The expected norm of random matrices. *Combin. Probab. Comput.*, 9:149–166, 2000.

[Shi96] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1996. Translated from the first (1980) Russian edition by R. P. Boas.

[So09] A. M.-C. So. Moment inequalities for sums of random matrices and their applications in optimization. *Math. Prog. Ser. A*, Dec. 2009. (electronic).

[SS98] B. Schölkopf and S. Smola. *Learning with Kernels*. MIT Press, 1998.

[SSS08] S. Shalev-Shwartz and N. Srebro. Low ℓ_1-norm and guarantees on sparsifiability. In *ICML/COLT/UAI Sparse Optimization and Variable Selection Workshop*, July 2008.

[SST06] A. Sankar, D. A. Spielman, and S.-H. Teng. Smoothed analysis of the condition numbers and growth factors of matrices. *SIAM J. Matrix Anal. Appl.*, 28(2):446–476, 2006.

[ST04] D. A. Spielman and S.-H. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 81–90 (electronic), New York, 2004. ACM.

[Ste72] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. 6th Berkeley Symp. Math. Statist. Probab.*, Berkeley, 1972. Univ. California Press.

[SV13] A. Sen and B. Virág. The top eigenvalue of the random Toeplitz matrix and the sine kernel. *Ann. Probab.*, 41(6):4050–4079, 2013.

[Tao12] T. Tao. *Topics in random matrix theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[Thi02] W. Thirring. *Quantum mathematical physics*. Springer-Verlag, Berlin, second edition, 2002. Atoms, molecules and large systems, Translated from the 1979 and 1980 German originals by Evans M. Harrell II.

[TJ74] N. Tomczak-Jaegermann. The moduli of smoothness and convexity and the Rademacher averages of trace classes $S_p (1 \leq p < \infty)$. *Studia Math.*, 50:163–182, 1974.

[Tro08a] J. A. Tropp. On the conditioning of random subdictionaries. *Appl. Comput. Harmon. Anal.*, 25:1–24, 2008.

[Tro08b] J. A. Tropp. Norms of random submatrices and sparse approximation. *C. R. Math. Acad. Sci. Paris*, 346(23-24):1271–1274, 2008.

[Tro08c] J. A. Tropp. The random paving property for uniformly bounded matrices. *Studia Math.*, 185(1):67–82, 2008.

[Tro11a] J. A. Tropp. Freedman's inequality for matrix martingales. *Electron. Commun. Probab.*, 16:262–270, 2011.

[Tro11b] J. A. Tropp. User-friendly tail bounds for matrix martingales. ACM Report 2011-01, California Inst. Tech., Pasadena, CA, Jan. 2011.

[Tro11c] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, August 2011.

[Tro11d] J. A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Adv. Adapt. Data Anal.*, 3(1-2):115–126, 2011.

[Tro12] J. A. Tropp. From joint convexity of quantum relative entropy to a concavity theorem of Lieb. *Proc. Amer. Math. Soc.*, 140(5):1757–1760, 2012.

[Tro14] J. A. Tropp. Convex recovery of a structured signal from independent random measurements. In *Sampling Theory, a Renaissance*. Birkhäuser Verlag, 2014. To appear. Available at http://arXiv.org/abs/1405.1102.

[TV04] A. M. Tulino and S. Verdú. *Random matrix theory and wireless communications*. Number 1(1) in Foundations and Trends in Communications and Information Theory. Now Publ., 2004.

[Ver12] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

[vNG47] J. von Neumann and H. H. Goldstine. Numerical inverting of matrices of high order. *Bull. Amer. Math. Soc.*, 53:1021–1099, 1947.

[Wig55] E. P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math. (2)*, 62:548–564, 1955.

[Wis28] J. Wishart. The generalised product moment distribution in samples from a multivariate normal population. *Biometrika*, 20A(1–2):32–52, 1928.

[Woo14] D. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1–2):1–157, 2014.

[WS01] C. K. I. Williams and M. Seeger. Using the Nyström method to spped up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688, Vancouver, 2001.

[Zou13] A. Zouzias. *Randomized primitives for linear algebra and applications*. PhD thesis, Univ. Toronto, 2013.