User-Friendly Tools for Random Matrices: An Introduction

Joel A. Tropp

3 December 2012
NIPS Version

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
03 DEC 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
User-Friendly Tools for Random Matrices: An Introduction	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **California Institute of Technology,Division of Engineering and Applied Science,Pasadena,CA,91125**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**151**	

For Margot

Contents

Preface

Nota Bene: This manuscript has not yet reached its final form. In particular, I have not had the opportunity to check all the details carefully and to polish the writing so that it reflects the main points as brightly as possible. At this stage, the citations and background references still lack the precision that I try to bring to my published works. I welcome any comments or corrections that may help improve subsequent versions of these notes.

These lecture notes are designed to bring random matrix theory to the people. In recent years, random matrices have come to play a major role in computational mathematics, but most of the classical methods for studying random matrices 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.

As with any extended work, my own interests and experience necessarily govern the content. In other words, I unapologetically emphasize the projects that I have engaged in over the last five years. This slant is not intended to diminish other contributions to the study of matrix concentration inequalities and their applications. Indeed, I have been influenced strongly by the work of many researchers, including Rudolf Ahlswede, Rajendra Bhatia, Jean Bourgain, Eric Carlen, Sourav Chatterjee, Edward Effros, Elliott Lieb, Lester Mackey, Roberto 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 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

CHAPTER 1

Introduction

Random matrix theory has become a large and vital field of probability, and it has found applications in a wide variety of other areas. To motivate the results in these notes, 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 classical 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.

Multivariate Statistics. One of the earliest examples 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 [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 LU decomposition.[1] They obtained an high-probability bound for the norm of the random matrix, which they took as an estimate for the amount of 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 [HMT11].

[1] It is breathtaking that 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.

Nuclear Physics. In the early 1950s, physicists had reached the limits of deterministic analytical techniques for modeling 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, then, would model the possible energy levels of the system. See Mehta's book for an account of all this [Meh04].

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.

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 provide an efficient way to construct approximations of large matrices. For example, they 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. See the paper [HMT11] for an overview of these ideas. This approach has been very successful in practice.

Subsampling of Data. One method that has been used in large-scale machine learning is to subsample data randomly before fitting a model. For instance, we can combine random sampling with the Nyström decomposition to approximate a kernel matrix efficiently [Git11]. The success of this approach depends on the properties of a small random submatrix drawn from a large, fixed matrix.

Dimension Reduction. In theoretical computer science, a common algorithmic template involves using randomness to reduce the dimension of the problem. The paper [AC09] describes an approach to nearest neighbor computations, based on random projection of the input data, that has become very popular. Random matrix theory forms a core part of the analysis.

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 [AM07]. This idea plays an important role in Spielman and Teng's work on fast algorithms for solving linear systems [ST04].

Combinatorial Optimization. One way to solve a hard combinatorial optimization problem is to replace the intractable computation with a related optimization problem that may be more tractable [BTN01]. After solving the easier problem, we can perform a randomized operation to obtain an approximate solution to the original hard problem. For optimization problems involving matrices, random matrix theory is central to the analysis [So09].

Compressed Sensing. Random matrices appear as measurement operators in the field of compressed sensing [Don06]. When acquiring data about an object with relatively few degrees of freedom as compared with the ambient dimension, we can sieve out the important information from the object by taking a small number of random measurements, where the number of measurements is comparable too the number of degrees of freedom. This application is possible because of geometric properties of random matrices [CRPW12].

1.2.2 Modeling

Random matrices also appears as models for multivariate data or multivariate phenomena. By studying the properties of these models, we may hope to obtain an understanding of the average-case 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 involves 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 superposition 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 matrices arise in the analysis of estimation techniques for this problem [MT12].

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 [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 [AS00, §9.2].

Algorithms. For worst-case examples, the Gaussian elimination method for solving a linear system is not numerically stable. In practice, however, this is a non-issue. 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. Dvoretsky's Theorem states that, when N in large, the unit ball of every 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. This important result can be framed as a statement about spectral properties of a random matrix [Gor85].

Quantum Information Theory. Random matrices appear as examples and counterexamples for a number of conjectures in quantum information theory. We refer the reader to the papers [HW08, Has09] for details.

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 accessible to researchers with significant experience, and it takes 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 intend 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. Almost every week, we learn about a paper that uses these ideas for a novel purpose.

1.4 Basic Questions in Random Matrix Theory

Although it sounds prosaic, random matrices merit attention precisely because they are matrices. As a consequence, random matrices have spectral properties: eigenvalues and eigenvectors

in the case of square matrices, singular values and singular vectors in the case of general matrices. The most basic problems all concern these spectral properties. Here are some questions that we might ask:

- What is the expectation of the maximum eigenvalue of a random symmetric matrix? What about the minimum eigenvalue?
- How are the extreme eigenvalues of a random symmetric matrix distributed? What is the probability that they take values substantially different from the mean?
- What is the expected spectral norm of a random matrix? What is the probability that the norm takes a value substantially different from the 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 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 an important role.

In this work, we focus on the first three questions above. We study the expectation of the extreme eigenvalues of random symmetric matrices, and we attempt to provide bounds on the probability that they take an unusual value. As an application of these results, we show how to control the expected spectral norm of a general matrix and to bound the probability of a large deviation. These are the most important issues for most (but not all!) applications. We will not touch on the remaining questions.

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

1.5.1 Example: A Sample Covariance Matrix

Let $\boldsymbol{x} = (X_1, \dots, X_p)$ be a random vector with zero mean $\mathbb{E}\,\boldsymbol{x} = \mathbf{0}$. Assume that the Euclidean norm of the distribution is bounded: $\|\boldsymbol{x}\|^2 \le B$. The *covariance* of the random vector $\boldsymbol{x}$ is the positive-semidefinite matrix

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

In other words, the (j,k) entry of the sample covariance records the covariance between the jth and kth entry of the vector.

One basic problem in statistical practice is to estimate the covariance matrix from data. Imagine that we have access to n independent samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, distributed the same way as $\boldsymbol{x}$. The *sample covariance estimator* is defined as 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 for the sample covariance matrix: $\mathbb{E}\,\boldsymbol{Y} = \boldsymbol{A}$. The formula (1.5.2) supposes that the random vector $\boldsymbol{x}$ is known to have zero mean; in general, we would have to make some adjustments to incorporate an estimate for the sample mean. To emphasize,

> **The sample covariance estimator $\boldsymbol{Y}$ can be expressed as a sum of independent random matrices.**

This is precisely the type of decomposition that our tools require.

1.6 Exponential Concentration Inequalities for Matrices

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

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

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.

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 the classical 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, we return to the scalar problem (1.6.1). Suppose that we can express the real random variable Z 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 guarantee that Z concentrates sharply around its mean value.

Theorem 1.6.1 (Bernstein Inequality)**.** *Let $S_1, \dots, S_n$ be independent random variables that have bounded deviation from their mean values:*

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

Form the sum $Z = \sum_{k=1}^{n} S_k$, *and introduce a variance parameter* $\sigma^2 = \mathbb{E}\big[(Z - \mathbb{E}\,Z)^2\big]$. *Then*

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

See the survey paper [Lug09] 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. More precisely, the result implies that the probability that the sum Z exhibits a moderate deviation ($t \leq \sigma^2/R$) decays like the tail of a normal random variable with variance σ^2. The probability that the sum Z exhibits a large deviation ($t \geq \sigma^2/R$) decays like an exponential random variable with mean R.

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 this point, we state one of the major results from these notes. This theorem is a matrix extension of Bernstein's inequality that was developed independently in the two papers [Oli10a, Tro11d]. 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 random matrices with common dimension* $d_1 \times d_2$. *Assume that each matrix has bounded deviation from its mean:*

$$\|\boldsymbol{S}_k - \mathbb{E}\,\boldsymbol{S}_k\| \leq R \quad \text{for each } k = 1, \dots, n.$$

Form the sum $\boldsymbol{Z} = \sum_{k=1}^{n} \boldsymbol{S}_k$, *and introduce a variance parameter*

$$\sigma^2 = \max\left\{\left\|\mathbb{E}\big[(\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z})(\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z})^*\big]\right\|, \left\|\mathbb{E}\big[(\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z})^*(\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z})\big]\right\|\right\}.$$

Then

$$\mathbb{P}\{\|\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z}\| \geq t\} \leq (d_1 + d_2) \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right) \quad \text{for all } t \geq 0.$$

Furthermore,

$$\mathbb{E}\,\|\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z}\| \leq \sqrt{2\sigma^2 \log(d_1 + d_2)} + \frac{1}{3}R\log(d_1 + d_2).$$

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 parameter σ^2 in the result for matrices can be interpreted as the magnitude of the expected squared deviation from the mean. The formula reflects the fact that a matrix $\boldsymbol{B}$ has *two* different squares $\boldsymbol{B}\boldsymbol{B}^*$ and $\boldsymbol{B}^*\boldsymbol{B}$.
- 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 tail bound is informative.
- We have included a bound for the expected deviation $\|\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{Z}\|$. This estimate is not particularly interesting the scalar setting, but it is usually quite challenging to prove results of this type for matrices. In fact, we often find the expectation bound more useful than the tail bound.

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

1.6.3 Example: A Sample Covariance Matrix

The reader may not yet perceive why abstract matrix inequalities, such as Theorem 1.6.2, deliver information about random matrices that arise in practice. Our burden remains to show that the results are worthwhile.

We will apply the matrix Bernstein inequality, Theorem 1.6.2, to measure how well a sample covariance matrix approximates the true covariance matrix. As before, let $\boldsymbol{x}$ be a zero-mean random vector with dimension p, and assume that the Euclidean norm of the distribution is bounded: $\|\boldsymbol{x}\|^2 \le B$. The covariance matrix of the vector is $\boldsymbol{A} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^*)$. Suppose we have n independent samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ with the same distribution as $\boldsymbol{x}$. We can form the sample covariance matrix

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

Our goal is to study the spectral-norm distance $\|\boldsymbol{Y} - \boldsymbol{A}\|$ between the sample covariance and the true covariance.

To that end, let us express the error matrix as a sum of independent random matrices:

$$\boldsymbol{E} = \boldsymbol{Y} - \boldsymbol{A} = \sum_{k=1}^{n} \boldsymbol{S}_k.$$

where $\boldsymbol{S}_k = n^{-1}(\boldsymbol{x}_k \boldsymbol{x}_k^* - \boldsymbol{A})$ for each index k. To apply the matrix concentration inequality, we must bound the norm of each summand, and we must compute the variance of the matrix $\boldsymbol{E}$.

To obtain the uniform bound, observe that

$$\mathbb{E}\,\boldsymbol{S}_k = \boldsymbol{0} \quad \text{and} \quad \|\boldsymbol{S}_k\| \le \frac{2B}{n}.$$

We reach the latter inequality as follows:

$$\|\boldsymbol{S}_k\| = \frac{1}{n} \left\| \boldsymbol{x}_k \boldsymbol{x}_k^* - \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^*) \right\| \le \frac{1}{n} \left(\|\boldsymbol{x}_k\|^2 + \mathbb{E}\,\|\boldsymbol{x}\|^2 \right) \le \frac{2B}{n}.$$

The first bound follows from the triangle inequality for the spectral norm and Jensen's inequality. The second relies on the uniform bound for the norm of a random vector distributed as $\boldsymbol{x}$.

Next, we must find a bound for the matrix variance $\sigma^2(\boldsymbol{E})$. Let us calculate that

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

The expression $\boldsymbol{H} \preccurlyeq \boldsymbol{M}$ means that $\boldsymbol{M} - \boldsymbol{H}$ is positive semidefinite. This argument relies on the uniform upper bound for the norm of the random vector. From here, we quickly obtain the variance $\sigma^2(\boldsymbol{E})$:

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

The second relation depends on the fact that the summands are independent and zero mean. The inequality is valid because $\boldsymbol{0} \preccurlyeq \boldsymbol{H} \preccurlyeq \boldsymbol{M}$ implies that the norm of $\boldsymbol{M}$ exceeds that norm of $\boldsymbol{H}$.

Now, we may invoke Theorem 1.6.2 to obtain

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

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 of ε, where $\varepsilon \in (0,1]$, we may take

$$n \ge \text{Const}\cdot\frac{B\log p}{\varepsilon^2\|\boldsymbol{A}\|}.$$

This selection yields

$$\mathbb{E}\|\boldsymbol{Y} - \boldsymbol{A}\| \le \text{Const}\cdot\varepsilon\cdot\|\boldsymbol{A}\|.$$

It is often the case that $B = \text{Const}\cdot p$, so we discover that $n = \text{Const}\cdot\varepsilon^{-2}p\log p$ samples suffice to estimate the covariance matrix $\boldsymbol{A}$ accurately. This bound is qualitatively sharp for worst-case distributions.

1.6.4 History of this Example

Covariance estimation may be the earliest application of matrix concentration bounds in random matrix theory. Rudelson [Rud99] showed how to use the noncommutative Khintchine inequality [LP86, LPP91, Buc01, Buc05] to obtain essentially optimal bounds on the sample covariance estimator for a bounded random vector. The tutorial [Ver12] of Roman Vershynin provides an excellent overview of this problem as well as many results and references.

The analysis of the sample covariance matrix here is adapted from the paper [GT11]. It leads to essentially the same result as Rudelson obtained in [Rud99]. For an analysis of sparse covariance estimation using matrix concentration inequalities, see the paper [CGT12a] and the technical report [CGT12b].

1.7 The Arsenal of Results

The classical literature contains many exponential tail bounds for sums of independent random variables. Some of the best known results are the Bernstein inequality and the Chernoff inequality, but there are many more. It turns out that essentially all of these results admit extensions

that hold for random matrices. These lecture notes focus on some exponential concentration inequalities for matrices that have already found significant applications.

Matrix Gaussian Series. A matrix Gaussian series is a random matrix that can be expressed as a sum of fixed matrices weighted by independent standard normal random variables. This formulation includes a surprising number of examples. The most important are undoubtedly Wigner matrices and rectangular Gaussian matrices. Other interesting cases include a Toeplitz matrix with Gaussian entries. This material 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 weighted by independent Rademacher random variables.[2] 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 random matrices that can be decomposed as a sum of independent positive-semidefinite random matrices whose maximum eigenvalues are subject to a uniform bound. These results are appropriate for studying the Laplacian matrix of a random graph. They also allow us to obtain information about the norm of a random submatrix drawn from a fixed matrix. See Chapter 5.

Matrix Bernstein Bounds. Matrix Bernstein inequalities concern random matrices that can be expressed as a sum of independent bounded random matrices that are bounded in norm. These results have many applications, including the analysis of randomized algorithms for approximate matrix multiplication and randomized algorithms for matrix sparsification. 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.

The literature describes other exponential matrix inequalities for sums of independent random matrices. These include a matrix Bennett inequality [Tro11d, §6], matrix Bernstein inequalities for unbounded random matrices [Tro11d, §6], and a matrix Hoeffding inequality [Tro11d, §7]. These results extend to give bounds for matrix-valued martingales, such as the matrix Azuma and McDiarmid inequalities [Tro11d, §7] and the matrix Freedman inequality [Oli10a, Tro11a].

Furthermore, the paper [MJC+12] develops a very different technique that can yield matrix concentration inequalities for random matrices based on dependent random variables. The results in this work include several exponential inequalities. This approach also leads to polynomial concentration inequalities, which can be viewed as a generalization of Chebyshev's inequality. See the annotated bibliography for more information.

1.8 These Lecture Notes

These lecture notes are intended for researchers and graduate students 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

[2] A Rademacher random variable is uniformly distributed on $\{\pm 1\}$.

theory of normed spaces, and classical probability theory up through the basic concentration inequalities (such as Markov and Bernstein).

The material here is based primarily on the paper "User-Friendly Tail Bounds for Sums of Random Matrices" by the present author [Tro11d]. 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 are used in practice.

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+12], but they have not been included in a unified presentation.

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 [HKZ12b, Min11]. We describe a new framework that allows us to prove these results with ease.

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

The organization of the notes is straightforward. Chapter 2 contains background material that is needed for the proofs. 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. We conclude with resources on matrix concentration and a bibliography.

Since these are lecture notes, 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. In the current version, these notes are still not as polished and complete as we might like, and we intend to expand them in future revisions.

CHAPTER 2

Matrix Functions and 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 detailed 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.

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

2.1 Matrix Theory Background

Most of these results are drawn from Bhatia's excellent books on matrix analysis [Bha97, Bha07]. The books [HJ85, HJ94] of Horn and Johnson also serve as good general references. Higham's work [Hig08] is a generous source of information about matrix functions.

2.1.1 Conventions

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 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 Matrices

Complex matrices with fixed dimensions form a linear space because we can add them and multiply them by complex scalars. We write $\mathbb{M}_{d_1 \times d_2}$ for the linear space of $d_1 \times d_2$ matrices. In addition to the usual linear operations, we can multiply square matrices, so they form an algebra.

We write $\mathbb{M}_d$ for the algebra of $d \times d$ square, complex matrices. The set $\mathbb{H}_d$ consists of Hermitian matrices with dimension d; it is a linear space over the real field. That is, we can add Hermitian matrices and multiply them by real numbers. Multiplication by a complex scalar is verboten inside $\mathbb{H}_d$. We rarely require this notation, but it is occasionally important for clarity.

2.1.3 Basic Matrices

We write $\mathbf{0}$ for the zero matrix and $\mathbf{I}$ for 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 $\mathbb{M}_{d_1 \times d_2}$ is comprised of *unit matrices*. We write $\mathbf{E}_{jk}$ for the unit matrix with a one in position (j,k) and zeros elsewhere. We use a related notation for unit vectors. The symbol $\mathbf{e}_k$ denotes a column vector with a one in position k and zeros elsewhere. The dimensions of unit matrices and unit vectors are typically determined by the context.

A square matrix that satisfies $\boldsymbol{Q}\boldsymbol{Q}^* = \mathbf{I} = \boldsymbol{Q}^*\boldsymbol{Q}$ is called *unitary*. We reserve the symbol $\boldsymbol{Q}$ for a unitary matrix. The symbol * denotes the conjugate transpose.

Readers who prefer the real setting may prefer to regard $\boldsymbol{Q}$ as an orthogonal matrix and to interpret * as the (simple) transpose operation.

2.1.4 Hermitian Matrices and Eigenvalues

A square matrix that satisfies $\boldsymbol{A} = \boldsymbol{A}^*$ is called *Hermitian*. We adopt Parlett's convention 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 $\boldsymbol{A}$ has an *eigenvalue decomposition*

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^* \quad \text{with } \boldsymbol{Q} \text{ unitary and } \boldsymbol{\Lambda} \text{ real diagonal.} \tag{2.1.1}$$

The diagonal entries of $\boldsymbol{\Lambda}$ are called the *eigenvalues* of $\boldsymbol{A}$. The unitary matrix $\boldsymbol{Q}$ in the eigenvalue decomposition is not completely determined, 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:

$$\lambda_{\min}(\theta\boldsymbol{A}) = \theta\lambda_{\min}(\boldsymbol{A}) \quad \text{and} \quad \lambda_{\max}(\theta\boldsymbol{A}) = \theta\lambda_{\max}(\boldsymbol{A}) \quad \text{for } \theta \geq 0. \tag{2.1.2}$$

There is an important relationship between minimum and maximum eigenvalues:

$$\lambda_{\min}(-\boldsymbol{A}) = -\lambda_{\max}(\boldsymbol{A}). \tag{2.1.3}$$

The fact (2.1.3) warns us that we must be careful passing scalars through an eigenvalue map.

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 in probability!

2.1.5 The Trace of a Square Matrix

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

$$\operatorname{tr}\boldsymbol{B} = \sum_{j=1}^{d} b_{jj} \quad \text{for a } d \times d \text{ matrix } \boldsymbol{B}.$$

The trace is unitarily invariant:

$$\operatorname{tr} \boldsymbol{B} = \operatorname{tr}(\boldsymbol{Q}^* \boldsymbol{B} \boldsymbol{Q}) \quad \text{for each square matrix } \boldsymbol{B} \text{ and each unitary } \boldsymbol{Q}.$$

In particular, the existence of an eigenvalue decomposition (2.1.1) shows that the trace of an Hermitian matrix equals the sum of its eigenvalues. This fact also holds true for a general square matrix.

2.1.6 The Semidefinite Order

An Hermitian matrix $\boldsymbol{A}$ with nonnegative eigenvalues is *positive semidefinite*. When each eigenvalue is strictly positive, we say that the matrix $\boldsymbol{A}$ is *positive definite*. Positive semidefinite matrices play a special role in matrix theory, analogous to the role of nonnegative numbers in real analysis.

The set of positive-semidefinite matrices with size d forms a closed, convex cone in the real-linear space of Hermitian matrices of dimension d. Therefore, we may define the *semidefinite partial order* on Hermitian matrices of the same size by the rule

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \iff \quad \boldsymbol{H} - \boldsymbol{A} \text{ is positive semidefinite.}$$

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 fact whose importance cannot be overstated.

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

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \implies \quad \boldsymbol{B}\boldsymbol{A}\boldsymbol{B}^* \preccurlyeq \boldsymbol{B}\boldsymbol{H}\boldsymbol{B}^* \tag{2.1.4}$$

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

$$\lambda_{\max}(\boldsymbol{A}) \le \operatorname{tr} \boldsymbol{A} \quad \text{when } \boldsymbol{A} \text{ is positive semidefinite.} \tag{2.1.5}$$

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

2.1.7 Standard Matrix Functions

Let us describe the most direct method for extending a function on the reals 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. Let $\boldsymbol{A}$ be a $d \times d$ Hermitian matrix with eigenvalues in I. Define the $d \times d$ Hermitian matrix $f(\boldsymbol{A})$ via the eigenvalue decomposition of $\boldsymbol{A}$:*

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

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.

For an Hermitian matrix $\boldsymbol{A}$, when we write the power function $\boldsymbol{A}^p$ 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 $\boldsymbol{A}$ be an Hermitian matrix, and let $f : \mathbb{R} \to \mathbb{R}$. Each eigenvalue of $f(\boldsymbol{A})$ has the form $f(\lambda)$, where λ is an eigenvalue of $\boldsymbol{A}$.*

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 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*

$$f(a) \le g(a) \quad \text{for each } a \in I \quad \Longrightarrow \quad f(\boldsymbol{A}) \preccurlyeq g(\boldsymbol{A}). \tag{2.1.6}$$

Proof. Decompose $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^*$. It is immediate that $f(\boldsymbol{\Lambda}) \preccurlyeq g(\boldsymbol{\Lambda})$. The Conjugation Rule (2.1.4) allows us to conjugate this relation by $\boldsymbol{Q}$. Finally, invoke the definition of the matrix function to complete the argument. □

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 $\boldsymbol{A}$ has eigenvalues in I. Then

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

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

2.1.8 The Matrix Exponential

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

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

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 several properties that we use extensively. First, the trace exponential is monotone with respect to the semidefinite order. That is, for Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{H}$ of the same size,

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \Longrightarrow \quad \operatorname{tr} \mathrm{e}^{\boldsymbol{A}} \leq \operatorname{tr} \mathrm{e}^{\boldsymbol{H}}. \tag{2.1.7}$$

The trace exponential is also a convex function on the real-linear space of Hermitian matrices. That is, for Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{H}$ of the same size,

$$\operatorname{tr} \mathrm{e}^{\tau \boldsymbol{A} + \bar{\tau} \boldsymbol{H}} \leq \tau \cdot \operatorname{tr} \mathrm{e}^{\boldsymbol{A}} + \bar{\tau} \cdot \operatorname{tr} \mathrm{e}^{\boldsymbol{H}} \quad \text{where } \tau \in [0,1] \text{ and } \bar{\tau} = 1 - \tau.$$

In other words, the trace exponential of an average is no greater than the average value of the trace exponentials. The proofs of these two results are not particularly hard, but they fall outside the boundary of these notes. See the survey article [Pet94, Sec. 2] or the lecture notes [Car10, Sec. 2.2] for a complete demonstration.

2.1.9 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:

$$\log\bigl(\mathrm{e}^{\boldsymbol{A}}\bigr) = \boldsymbol{A} \quad \text{for each Hermitian matrix } \boldsymbol{A}. \tag{2.1.8}$$

A deep and significant fact about the matrix logarithm is that it preserves the semidefinite order. For positive-definite matrices $\boldsymbol{A}$ and $\boldsymbol{H}$ of the same size,

$$\mathbf{0} \prec \boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \Longrightarrow \quad \log(\boldsymbol{A}) \preccurlyeq \log(\boldsymbol{H}). \tag{2.1.9}$$

For a good treatment of operator monotonicity at an introductory level, see [Bha97, Chap. V]. Let us emphasize that the matrix exponential *does not* have any operator monotonicity property analogous with (2.1.9)!

2.1.10 Singular Values of General Matrices

A general matrix $\boldsymbol{B}$ 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*

$$\boldsymbol{B} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^* \quad \text{with } \boldsymbol{U}, \boldsymbol{V} \text{ unitary and } \boldsymbol{\Sigma} \text{ nonnegative diagonal.} \tag{2.1.10}$$

The unitary matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ 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 $\boldsymbol{B}$. They are determined completely modulo permutations, and it is standard 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 Hermitian. We can use a singular value decomposition of $\boldsymbol{B}$ to construct eigenvalue decompositions of the two squares:

$$\boldsymbol{B}\boldsymbol{B}^* = \boldsymbol{U}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*)\boldsymbol{U}^* \quad \text{and} \quad \boldsymbol{B}^*\boldsymbol{B} = \boldsymbol{V}(\boldsymbol{\Sigma}^*\boldsymbol{\Sigma})\boldsymbol{V}^*$$

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

2.1.11 The Spectral Norm and the Euclidean Norm

The *spectral norm* of an Hermitian matrix is defined by the relation

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

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

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

These two definitions are consistent for Hermitian matrices.

When applied to a row vector or a column vector, the spectral norm coincides with the Euclidean norm:

$$\|\boldsymbol{b}\| = \left(\sum_{k=1}^{d} |b_k|^2\right)^{1/2} \quad \text{for } \boldsymbol{b} \in \mathbb{C}^d.$$

We are certainly justified, therefore, in using the same symbol for both norms.

2.1.12 Dilations

An extraordinarily fruitful idea from operator theory is to embed matrices within larger block matrices, called *dilations* [Pau02].

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 a Hermitian matrix given by

$$\mathscr{H}(\boldsymbol{B}) = \begin{bmatrix} \mathbf{0} & \boldsymbol{B} \\ \boldsymbol{B}^* & \mathbf{0} \end{bmatrix}. \tag{2.1.11}$$

The dilation retains important spectral information. To see why, note that the square of the dilation satisfies

$$\mathscr{H}(\boldsymbol{B})^2 = \begin{bmatrix} \boldsymbol{B}\boldsymbol{B}^* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B}^*\boldsymbol{B} \end{bmatrix}. \tag{2.1.12}$$

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. Since the trace of $\mathscr{H}(\boldsymbol{B})$ is zero, its maximum eigenvalue must be nonnegative. Together, these two facts yield an important identity:

$$\lambda_{\max}(\mathscr{H}(\boldsymbol{B})) = \|\mathscr{H}(\boldsymbol{B})\| = \|\boldsymbol{B}\|. \tag{2.1.13}$$

Finally, we note that the Hermitian dilation is a real-linear map.

2.2 Probability Background

We continue with some material from probability, focusing on connections with matrices. For more details, consult any good probability text.

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. All 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.

2.2.2 Random Matrices

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, and let $\mathbb{M}_{d_1 \times d_2}$ be the set of $d_1 \times d_2$ complex matrices. 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 nor be correlated with each other. We reserve the letters $\boldsymbol{X}, \boldsymbol{Y}$ for random Hermitian matrices, and the letter $\boldsymbol{Z}$ denotes a general random matrix.

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

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

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

2.2.3 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.

2.2.4 Inequalities for Expectation

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

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

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

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

Let us emphasize that these inequalities hold for *every* real-valued function h on matrices that is concave or convex.

The expectation of a random matrix can be viewed as a convex combination, and the cone of positive-semidefinite matrices is convex. Therefore, expectation preserves the semidefinite order:

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

We use this result many times without direct reference.

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 sample applications may wish to move on to Chapters 4, 5, and 6.

The approach that we take can be viewed as a matrix extension of the Laplace transform method, sometimes referred to as the "Bernstein trick." In the scalar setting, this so-called trick is one of the most basic and successful paths to reach concentration inequalities for sums of independent random variables. It turns out that there is a very satisfactory version of this argument that applies to sums of independent random matrices. In the more general setting, however, we must invest more care and wield sharper tools to execute this technique successfully.

We first define matrix analogs of the moment generating function and the cumulant generating function, which pack up information about the growth of a random 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, and §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 specialize to obtain matrix Chernoff bounds, matrix Bernstein bounds, etc.

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.

To expand on Definition 3.1.1, let us observe that the matrix mgf and cgf have formal power series expansions:

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

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

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

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

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

The first identity holds because $a \mapsto \mathrm{e}^{\theta a}$ is a monotone increasing function, so the event doesn't change under the mapping. The second relation is Markov's inequality (2.2.1). To control the exponential, note that

$$\mathrm{e}^{\theta \lambda_{\max}(\boldsymbol{Y})} = \mathrm{e}^{\lambda_{\max}(\theta \boldsymbol{Y})} = \lambda_{\max}\left(\mathrm{e}^{\theta \boldsymbol{Y}}\right) \leq \mathrm{tr}\, \mathrm{e}^{\theta \boldsymbol{Y}}. \tag{3.2.3}$$

The first identity holds because the maximum eigenvalue map is positive homogeneous, as stated in (2.1.2). The second depends on the Spectral Mapping Theorem, Proposition 2.1.3. The inequality holds because the exponential of an Hermitian matrix is positive definite, and (2.1.5) shows that the maximum eigenvalue of a positive-definite matrix is dominated by the trace. Combine the latter two relations to reach

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

This inequality holds 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

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

The function $a \mapsto \mathrm{e}^{\theta a}$ reverses the inequality in the event because it is monotone decreasing. The third relation owes to the relationship (2.1.3) between minimum and maximum eigenvalues. Finally, introduce the inequality (3.2.3) for the trace exponential and minimize over negative θ. □

In the proof of Proposition 3.2.1, it may seem crude to bound the maximum eigenvalue by the trace. It turns out that, at most, this estimate results in a loss of a logarithmic factor. At the same time, the 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*

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

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

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

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

The first identity holds because the maximum eigenvalue map is positive homogeneous, as stated in (2.1.2). The third follows when we use the Spectral Mapping Theorem, Proposition 2.1.3 to draw the exponential inside the eigenvalue map. The inequality depends on the fact (2.1.5) 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 independent sums 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:

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

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. Note that 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:

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

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:

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

It seems that we have reached an impasse.

What if we consider the cgf instead? The cgf of a sum of independent 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*

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

is concave on the positive-definite cone in dimension d.

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. Theorem 3.4.1 is not easy to prove, so we must take it for granted.

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, let us 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 size. Then*

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

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

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

The first identity follows from the definition (2.1.8) 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.

Lemma 3.5.1 (Subadditivity of Matrix Cgfs)**.** *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, Hermitian matrices of the same size. 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. To simplify notation, we take $\theta = 1$. Let $\mathbb{E}_k$ denote the expectation with respect to $\boldsymbol{X}_k$, the remaining random matrices held fixed. Abbreviate

$$\boldsymbol{\Xi}_k := \log\bigl(\mathbb{E}_k\, \mathrm{e}^{\boldsymbol{X}_k}\bigr) = \log\bigl(\mathbb{E}\, \mathrm{e}^{\boldsymbol{X}_k}\bigr).$$

We may calculate that

$$\mathbb{E} \operatorname{tr} \exp\bigl(\textstyle\sum_{k=1}^{n} \boldsymbol{X}_k\bigr) = \mathbb{E}\, \mathbb{E}_n \operatorname{tr} \exp\Bigl(\textstyle\sum_{k=1}^{n-1} \boldsymbol{X}_k + \boldsymbol{X}_n\Bigr)$$

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

We can introduce iterated expectations because of the tower property of conditional expectation. At each step $m = 1, 2, \dots, n$, 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 equivalent formulation (3.5.2) follows from (3.5.1) when we substitute the definition (3.1.1) of the matrix cgf and make some algebraic simplifications. □

3.6 Master Bounds for Independent Sums of 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 Bound for an Independent Sum of Matrices)**.** *Consider a finite sequence* $\{\boldsymbol{X}_k\}$ *of independent, random, Hermitian matrices. Then*

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

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

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

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

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

Furthermore,

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, general matrices because we can apply them to the Hermitian dilation of the sum. Instead of presenting a general theorem, we find it more natural to extend the specific tail bounds to general matrices.

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 precise version of Proposition 3.2.1 and the proof that we present here are due to Roberto Oliveira, from his an elegant paper [Oli10b]. 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. Ahslwede and Winter [AW02, App.] proposed a different approach for bounding the trace mgf of an independent sum, based on a repeated application of the Golden–Thompson inequality (3.3.3). The Ahlswede–Winter argument leads to a cumulant bound of the form

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

In other words, they bound the cumulant of a sum in terms of the sum of *maximum eigenvalues* of the cumulants. There are cases where the bound (3.7.1) is equivalent with Lemma 3.5.1. For example, the bounds coincide when each matrix $\boldsymbol{X}_k$ is identically distributed. In general, however, the estimate (3.7.1) leads to fundamentally weaker results.

The first major technical advance beyond the original argument of Ahlswede and Winter appears in another paper [Oli10a] of Oliveira. He developed a much 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 to the matrix Bernstein inequality, Theorem 6.1.1, 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 we describe here, based on Lieb's Theorem, was developed in the paper [Tro11d]. This research recognized the probabilistic content of Lieb's Theorem, Corollary 3.4.2, and it used this idea to establish 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, Tro11d] are independent works.

For a detailed discussion of the benefits of Lieb's Theorem over the Golden–Thompson inequality, turn to [Tro11d, §4]. In summary, to get the sharpest concentration results for random

matrices, Lieb's theorem is indispensible. The Ahlswede–Winter approach seems to be 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, Tro11c]; these works also depend on Lieb's Theorem. Another paper [GT11] shows how to use a more general 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.

This research originates in an important paper [LP86] of Françoise Lust-Picquard. This article develops an extension of the Khintchine inequality for matrices. Her result concerns a sum of fixed matrices that are modulated by independent Gaussian random variables. It shows that the expected trace of an even power of this random matrix is controlled by its variance. Subsequent papers have refined the noncommutative Khintchine inequality to its optimal form [LPP91, Buc01, Buc05].

In recent years, researchers have generalized other moment inequalities for sums of scalar random variables to matrices (and beyond). For instance, the Rosenthal inequality, concerning a sum of independent zero-mean random variables, admits a matrix version [JZ11, MJC+12, CGT12a]. See the paper [JX05] for a good overview of some other noncommutative moment inequalities.

Finally, and tangentially, we mention that matrix moments and cumulants also play a central role in the theory of free probability [Spe11].

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. For a detailed treatment from the perspective of matrix theory, see Bhatia's book [Bha97, Sec. IX.3]. 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]. For an account with more physical content, see the book of Thirring [Thi02].

Lieb's Theorem [Lie73, Thm. 6] was first established in an important paper of Elliott Lieb on the convexity of trace functions. His argument is difficult. Subsequent work has led to more direct routes to the result. Epstein provides an alternative proof of Theorem 3.4.1 in [Eps73, Sec. II], and Ruskai offers a simplified account of Epstein's argument in [Rus02, Rus05]. The note [Tro11b] shows how to derive Lieb's theorem from the joint convexity of quantum relative

entropy [Lin74, Lem. 2]. The latter approach is advantageous because the joint convexity result admits several elegant, conceptual proofs [Pet86, Eff09].

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, modulated by independent scalar random variables. This type of formulation is surprisingly versatile, and it already encompasses a range of interesting examples.

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

$$\boldsymbol{Y} = \sum\nolimits_k \gamma_k \boldsymbol{A}_k.$$

As an example, we can express a Wigner matrix, one of the classical random matrices, in this fashion. The real value of this perspective, however, is that we can use matrix Gaussian series to represent many other kinds of random matrices formed from Gaussian random variables. These models allow us to attack problems that classical methods do not handle gracefully. For instance, we can study a symmetric Toeplitz matrix with Gaussian entries.

We do not need to limit our attention to the Hermitian case. This chapter also contains bounds on the spectral norm of a Gaussian series with general matrix coefficients. Remarkably, these results follow as an immediate corollary of the Hermitian theory. This theory brings rectangular matrices based on Gaussian variables within our purview.

Furthermore, 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 values in $\{\pm 1\}$ 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. For instance, we can study the expected spectral norm of a fixed real matrix after flipping the signs of the entries at random.

We begin, in §§4.1–4.2, with an overview of our results for matrix Gaussian series; very similar results also hold for matrix Rademacher series. Afterward, in §4.3, we discuss the accuracy of the

theoretical bounds. The subsequent sections, §§4.4–4.6, describe what the matrix concentration inequalities tell us about some classical and not-so-classical examples of random matrices. Section 4.7 includes an overview of a more substantial application in combinatorial optimization. The final part of the chapter, §§4.8–4.9, contains detailed proofs of the bounds. We conclude with bibliographical notes.

4.1 Series with Hermitian Matrices

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

$$\mathbb{P}\left\{\sum_k \gamma_k a_k \geq t\right\} \leq \mathrm{e}^{-t^2/2\sigma^2} \quad \text{where } \sigma^2 = \sum_k a_k^2. \tag{4.1.1}$$

This result indicates that the upper tail of a scalar Gaussian series behaves like the upper tail of a single Gaussian random variable with variance σ^2. It turns out that the inequality (4.1.1) extends directly to the matrix setting.

Theorem 4.1.1 (Matrix Gaussian and 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. Form the matrix Gaussian series*

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

Compute the variance parameter

$$\sigma^2 = \sigma^2(\boldsymbol{Y}) = \left\|\mathbb{E}\left(\boldsymbol{Y}^2\right)\right\|. \tag{4.1.2}$$

Then

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \leq \sqrt{2\sigma^2 \log d}. \tag{4.1.3}$$

Furthermore, for all $t \geq 0$,

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \geq t\} \leq d\,\mathrm{e}^{-t^2/2\sigma^2}. \tag{4.1.4}$$

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

The proof of this result appears below in §4.8.

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 the maximum eigenvalue of $\boldsymbol{Y}$ is controlled by the matrix variance σ^2. Furthermore, the maximum eigenvalue of $\boldsymbol{Y}$ has a Gaussian tail whose decay rate depends on σ^2.

We can obtain a more explicit expression for the variance (4.1.2) in terms of the coefficients in the Gaussian series. Simply compute that

$$\sigma^2(\boldsymbol{Y}) = \left\|\mathbb{E}\left(\boldsymbol{Y}^2\right)\right\| = \left\|\mathbb{E}\left(\sum\nolimits_{j,k} \gamma_j \gamma_k \boldsymbol{A}_j \boldsymbol{A}_k\right)\right\| = \left\|\sum\nolimits_k \boldsymbol{A}_k^2\right\|. \tag{4.1.5}$$

The second identity follows because $\{\gamma_k\}$ is an independent family. As in the scalar case (4.1.1), the variance is the sum of the squares of the coefficients.

A new feature of the bound (4.1.4) is the dimensional factor d. When $d = 1$, this factor vanishes, and the matrix bound coincides with the scalar result (4.1.1). When $d = 1$, the expectation bound (4.1.3) also produces a sharp result, namely $\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \le 0$. In this case, at least, we have lost nothing by lifting the Laplace transform method to matrices. In §4.3, we discuss the extent to which Theorem 4.1.1 provides accurate predictions.

Finally, the reader may be concerned about the lack of explicit inequalities for the minimum eigenvalue $\lambda_{\min}(\boldsymbol{Y})$. But these bounds are consequences of the results for the maximum eigenvalue because $-\boldsymbol{Y}$ has the same distribution as $\boldsymbol{Y}$. Therefore,

$$\mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}) = \mathbb{E}\,\lambda_{\min}(-\boldsymbol{Y}) = -\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \ge -\sqrt{2\sigma^2 \log d}. \tag{4.1.6}$$

The second identity holds because of the relationship (2.1.3) between minimum and maximum eigenvalues. Similar considerations lead to a lower tail bound for the minimum eigenvalue:

$$\mathbb{P}\{\lambda_{\min}(\boldsymbol{Y}) \le -t\} \le d\,\mathrm{e}^{-t^2/2\sigma^2} \quad \text{for } t \ge 0. \tag{4.1.7}$$

This result follows directly from the upper tail bound (4.1.4).

4.2 Series with General Matrices

Most of the inequalities in these notes can be adapted to study the spectral norm of a sum of general random matrices. Although this problem might seem to have a character different from the Hermitian case, the results for general matrices are an easy formal consequence of the theory for Hermitian matrices. Here is the extension of Theorem 4.1.1.

Corollary 4.2.1 (Matrix Gaussian and Rademacher Series: The General Case)**.** *Consider a finite sequence $\{\boldsymbol{B}_k\}$ of fixed complex matrices with dimensions $d_1 \times d_2$, and let $\{\gamma_k\}$ be a finite sequence of independent standard normal variables. Form the matrix Gaussian series*

$$\boldsymbol{Z} = \textstyle\sum_k \gamma_k \boldsymbol{B}_k.$$

Compute the variance parameter

$$\sigma^2 = \sigma^2(\boldsymbol{Z}) = \max\left\{\left\|\mathbb{E}\left(\boldsymbol{Z}\boldsymbol{Z}^*\right)\right\|,\ \left\|\mathbb{E}\left(\boldsymbol{Z}^*\boldsymbol{Z}\right)\right\|\right\}. \tag{4.2.1}$$

Then

$$\mathbb{E}\,\|\boldsymbol{Z}\| \le \sqrt{2\sigma^2 \log(d_1 + d_2)}. \tag{4.2.2}$$

Furthermore, for all $t \ge 0$,

$$\mathbb{P}\{\|\boldsymbol{Z}\| \ge t\} \le (d_1 + d_2)\,\mathrm{e}^{-t^2/2\sigma^2}. \tag{4.2.3}$$

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

The proof of Corollary 4.2.1 appears below in §4.9.

4.2.1 Discussion

The results for rectangular matrices are similar with the results in Theorem 4.1.1 for Hermitian matrices, so many of the same intuitions apply. Still, the differences deserve some comment.

The most salient change occurs in the definition (4.2.1) of the variance parameter. The variance has this particular form because a general matrix has *two* squares associated with it, and we can omit neither one. Note that, when $\boldsymbol{Z}$ is Hermitian, the general variance (4.2.1) reduces to the Hermitian variance (4.1.2), so the new definition extends the previous one.

To represent the variance in terms of the coefficient matrices, we simply calculate that

$$\begin{aligned} \sigma^2(\boldsymbol{Z}) &= \max\left\{\left\|\mathbb{E}\left(\boldsymbol{Z}\boldsymbol{Z}^*\right)\right\|, \left\|\mathbb{E}\left(\boldsymbol{Z}^*\boldsymbol{Z}\right)\right\|\right\} \\ &= \max\left\{\left\|\mathbb{E}\left(\sum\nolimits_{j,k}\gamma_j\gamma_k\boldsymbol{B}_j\boldsymbol{B}_k^*\right)\right\|, \left\|\mathbb{E}\left(\sum\nolimits_{j,k}\gamma_j\gamma_k\boldsymbol{B}_j^*\boldsymbol{B}_k\right)\right\|\right\} \\ &= \max\left\{\left\|\sum\nolimits_k \boldsymbol{B}_k\boldsymbol{B}_k^*\right\|, \left\|\sum\nolimits_k \boldsymbol{B}_k^*\boldsymbol{B}_k\right\|\right\}. \end{aligned} \tag{4.2.4}$$

The expression (4.2.4) provides a natural formulation of the "sum of squares" of a sequence of general matrices.

The dimensional factor $d_1 + d_2$ in Corollary 4.2.1 apparently differs from the factor d that appears in Theorem 4.1.1. Nevertheless, properly interpreted, the two results coincide: Observe that we must bound the maximum and minimum eigenvalues of a Hermitian Gaussian series $\boldsymbol{Y}$ to control its spectral norm. Thus,

$$\mathbb{P}\{\|\boldsymbol{Y}\| \geq t\} \leq 2d\,\mathrm{e}^{-t^2/2\sigma^2}. \tag{4.2.5}$$

This inequality follows when we apply the union bound to the upper (4.1.4) and lower (4.1.7) tail bounds. The dimensional factor $d_1 + d_2$ in Corollary 4.2.1 matches the factor $2d$ in (4.2.5). We conclude that it is appropriate for both dimensions of the general matrix to play a role.

4.3 Are the Bounds Sharp?

One may wonder whether Theorem 4.1.1 and Corollary 4.2.1 provide accurate information about the behavior of a matrix Gaussian series. The answer turns out to be complicated, so we must limit ourselves to a summary of facts.

First, we consider the bound (4.2.2) for the expectation of a Gaussian series $\boldsymbol{Z}$ taking $d_1 \times d_2$ matrix values:

$$\mathbb{E}\,\|\boldsymbol{Z}\| \leq \sqrt{2\sigma^2 \log(d_1 + d_2)},$$

where σ^2 is defined in (4.2.1). Since the upper tail of $\|\boldsymbol{Z}\|$ decays so quickly, it is easy to believe (and true!) that

$$\mathbb{E}\,\|\boldsymbol{Z}\|^2 \lessapprox 2\sigma^2 \log(d_1 + d_2).$$

On the other hand, since the spectral norm is convex, Jensen's inequality (2.2.2) shows that

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

The first identity holds because $\|\boldsymbol{Z}\|^2 = \|\boldsymbol{Z}\boldsymbol{Z}^*\| = \|\boldsymbol{Z}^*\boldsymbol{Z}\|$. The final relation depends on the calculation (4.2.4). In summary,

$$\sigma^2 \quad\leq\quad \mathbb{E}(\|\boldsymbol{Z}\|^2) \quad\lessapprox\quad 2\sigma^2 \log(d_1 + d_2). \tag{4.3.1}$$

We see that the matrix variance σ^2 defined in (4.2.1) is roughly the correct scale for $\mathbb{E}(\|\boldsymbol{Z}\|^2)$. In general, it is a challenging problem to identify the expected norm of a Gaussian series, so the estimate (4.3.1) is already a significant achievement.

At this point, we might ask whether either side of the inequality (4.3.1) can be tightened. The answer is negative, unless we have additional information beyond the variance σ^2. There are examples of matrix Gaussian series where the left-hand inequality is correct up to constant factors, while there are other examples that saturate the right-hand inequality. Later in this chapter, when we turn to applications, we will encounter both of these cases (and more). In Chapter 7, we will show how to moderate the dimensional factor, but we cannot remove it entirely using current techniques.

What about the tail bound (4.2.3) for the norm of the Gaussian series? Here, our results are less impressive. It turns out that the large-deviation behavior of a Gaussian series is controlled by a different parameter σ_*^2 called the *weak variance*. There are cases where the weak variance σ_*^2 is substantially smaller than the variance σ^2, which means that the tail bound (4.2.3) can badly overestimate the tail probability 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.

In short, the primary value of matrix concentration inequalities inheres in the estimates that they provide for the expectation of the norm (maximum eigenvalue, minimum eigenvalue) of a random matrix. In many cases, they also provide reasonable information about the tail decay, but there are other situations where the tail bounds are depressingly feeble.

4.4 Example: Some Gaussian Matrices

Let us begin by applying our tools to 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 eigenvalue distribution is available, which provides a benchmark for assessing our results. We find that bounds based on Theorem 4.1.1 and Corollary 4.2.1 lead to very reasonable estimates but they are not sharp. We can reach similar conclusions for matrices with independent Rademacher entries.

4.4.1 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 more compactly as a Gaussian series:

$$\boldsymbol{W} = \sum_{1 \le j < k \le d} \gamma_{jk} (\mathbf{E}_{jk} + \mathbf{E}_{kj}) \tag{4.4.1}$$

It is known that

$$\frac{1}{\sqrt{d}}\,\lambda_{\max}(\boldsymbol{W}_d) \longrightarrow 2 \quad \text{almost surely as } d \to \infty. \tag{4.4.2}$$

To make (4.4.2) precise, we assume that $\{\boldsymbol{W}_d\}$ is a sequence of independent Gaussian Wigner matrices, indexed by the dimension d.

Theorem 4.1.1 provides a simple way to bound the maximum eigenvalue of a Gaussian Wigner matrix. We just need to compute the variance $\sigma^2(\boldsymbol{W}_d)$. To that end, note that the sum of the squared coefficient matrices takes the form

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

We have used the fact that $\boldsymbol{E}_{jk}\boldsymbol{E}_{kj} = \mathbf{E}_{jj}$, while $\mathbf{E}_{jk}\mathbf{E}_{jk} = \mathbf{0}$ because the limits of the summation ensure that $j \ne k$. We see that

$$\sigma^2(\boldsymbol{W}_d) = \|(d-1)\,\mathbf{I}_d\| = d-1.$$

The bound (4.1.3) for the expectation of the maximum eigenvalue gives

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{W}) \le \sqrt{2(d-1)\log d}. \tag{4.4.3}$$

In conclusion, our techniques overestimate the maximum eigenvalue of $\boldsymbol{W}_d$ by a factor of approximately $\sqrt{0.5\log d}$. Our result (4.4.3) is not perfect, but it only takes two lines of work. In contrast, the classical result (4.4.2) depends on a long moment calculation that involves challenging combinatorial arguments.

4.4.2 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:

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

For this matrix, the literature contains an elegant estimate of the form

$$\mathbb{E}\,\|\boldsymbol{G}\| \le \sqrt{d_1} + \sqrt{d_2}. \tag{4.4.4}$$

The inequality (4.4.4) is saturated when d_1 and d_2 tend to infinity with the ratio d_1/d_2 fixed.

Corollary 4.2.1 yields another bound on the expected norm of the matrix $\boldsymbol{G}$. In order to compute the variance $\sigma^2(\boldsymbol{G})$, we form the sums of squared coefficients:

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

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

The matrix variance (4.2.1) is

$$\sigma^2(\boldsymbol{G}) = \max\left\{\left\|d_2\,\mathbf{I}_{d_1}\right\|,\ \left\|d_1\,\mathbf{I}_{d_2}\right\|\right\} = \max\{d_1,\ d_2\}.$$

We conclude that

$$\mathbb{E}\,\|\boldsymbol{G}\| \le \sqrt{2\max\{d_1,\ d_2\}\log(d_1+d_2)}. \tag{4.4.5}$$

The leading term is roughly correct because

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

The logarithmic factor in (4.4.5) does not belong, but it is rather small in comparison with the leading terms. Once again, we have produced a good result with a minimal amount of effort. In contrast, the proof of (4.4.4) depends on a miraculous application of a comparison theorem for Gaussian processes.

4.5 Example: Matrices with Randomly Signed Entries

Next, we turn to an example that superficially appears similar to the matrix discussed in §4.4.2 but is much less understood. Consider a fixed $d_1 \times d_2$ matrix $\boldsymbol{B}$ with real entries, and let $\{\varepsilon_{jk}\}$ be an independent family of Rademacher random variables. Consider the $d_1 \times d_2$ random matrix

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

In other words, we obtain the random matrix $\boldsymbol{B}_{\pm}$ be randomly flipping the sign of each entry of $\boldsymbol{B}$. The literature contains the following bound on the expected norm of this matrix:

$$\mathbb{E}\,\|\boldsymbol{B}_{\pm}\| \le \text{Const}\cdot\sigma\cdot\log^{1/4}(\min\{d_1,\ d_2\}), \tag{4.5.1}$$

where the leading factor

$$\sigma = \max\left\{\max_j \left\|\boldsymbol{b}_{j:}\right\|,\ \max_k \|\boldsymbol{b}_{:k}\|\right\}. \tag{4.5.2}$$

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 Euclidean norm achieved by any row or column. There are cases where the bound (4.5.1) admits a matching lower bound.

Corollary 4.2.1 leads to a quick proof of a slightly weaker result. We simply need to compute the variance $\sigma^2(\boldsymbol{B}_{\pm})$. To that end, note that

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

Similarly,

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

Therefore, the variance (4.2.1) is

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

We see that $\sigma(\boldsymbol{B}_\pm)$ coincides with σ, the leading term (4.5.2) in the established estimate (4.5.1)! Now, Corollary 4.2.1 delivers the bound

$$\mathbb{E}\|\boldsymbol{B}_\pm\| \le \sqrt{2}\cdot\sigma(\boldsymbol{B}_\pm)\cdot\log^{1/2}(d_1+d_2). \tag{4.5.3}$$

Observe that the estimate (4.5.3) for the norm matches the correct bound (4.5.1) up to a factor of $\log^{1/4}(\max\{d_1,\, d_2\})$. 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 Corollary 4.2.1 to analyze this random matrix is that we can obtain a good result with a minimal amount of arithmetic. The analysis that leads to (4.5.1) involves a long sequence of combinatorial arguments.

4.6 Example: Gaussian Toeplitz Matrices

Matrix concentration inequalities offer very effective tools for analyzing random matrices that involve dependency structures that are more complicated than 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 $\boldsymbol{R}$ 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:

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

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

$$\boldsymbol{R}_d = \gamma_0\mathbf{I} + \sum_{k=1}^{d-1}\gamma_k\boldsymbol{S}^k + \sum_{k=1}^{d-1}\gamma_{-k}(\boldsymbol{S}^k)^*, \tag{4.6.1}$$

where $\boldsymbol{S}$ is the shift-up operator acting on d-dimensional column vectors:

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

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

We can analyze this example quickly using Corollary 4.2.1. First, note that

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

To obtain the variance parameter (4.2.1), we calculate the sum of the "squares" of the coefficient matrices that appear in (4.6.1). In this instance, the two terms in the matrix variance are the same. We find that

$$\begin{aligned}\mathbf{I}^2 + \sum_{k=1}^{d-1} (\boldsymbol{S}^k)(\boldsymbol{S}^k)^* + \sum_{k=1}^{d-1} (\boldsymbol{S}^k)^*(\boldsymbol{S}^k) &= \mathbf{I} + \sum_{k=1}^{d-1} \left[\sum_{j=1}^{d-k} \mathbf{E}_{jj} + \sum_{j=k+1}^{d} \mathbf{E}_{jj} \right] \\ &= \sum_{j=1}^{d} \left[1 + \sum_{k=1}^{d-j} 1 + \sum_{k=1}^{j-1} 1 \right] \mathbf{E}_{jj} = \sum_{j=1}^{d} (1 + (d-j) + (j-1)) \, \mathbf{E}_{jj} = d\,\mathbf{I}_d. \quad (4.6.2)\end{aligned}$$

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

$$\sigma^2(\boldsymbol{R}_d) = \| d\,\mathbf{I}_d \| = d.$$

An application of Corollary 4.2.1 leads us to conclude that

$$\mathbb{E}\,\|\boldsymbol{R}_d\| \le \sqrt{2d \log(2d)}. \quad (4.6.3)$$

It turns out that the inequality (4.6.3) is correct up to the precise value of the constant, which does not seem to be known. In other words,

$$\text{const} \quad \le \quad \frac{\mathbb{E}\,\|\boldsymbol{R}_d\|}{\sqrt{d \log d}} \quad \le \quad \text{Const} \quad \text{as } d \to \infty.$$

Here, we take $\{\boldsymbol{R}_d\}$ to be a sequence of unsymmetric Gaussian Toeplitz matrices, indexed by the ambient dimension d.

4.7 Application: Rounding for the MaxQP Relaxation

Our final application involves a more substantial question in 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 Corollary 4.2.1. 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 "round" the solution to ensure that it falls in 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 involves maximizing a quadratic form subject to a set of quadratic constraints and a spectral norm constraint. This problem is referred to as MAXQP. The desired solution $\boldsymbol{Z}$ 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 $\|\boldsymbol{Z}\| \le 1$.

There is a natural relaxation of the MAXQP problem that has been studied for the last decade or so. 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

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

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

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

where $\{\varepsilon_k : k = 1, \dots, n\}$ is a family of independent Rademacher random variables. The scaling factor $\alpha > 0$ can be adjusted to guarantee that the norm constraint holds with high probability.

What is the expected norm of $\boldsymbol{Z}$? Corollary 4.2.1 yields

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

Here, the variance parameter satisfies

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

owing to the properties of 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 $\boldsymbol{Z}$ observes the spectral norm constraint on average.

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.

4.8 Proof of Bounds for Hermitian Matrix Series

We continue with the proof that matrix Gaussian series exhibit the behavior described in Theorem 4.1.1. Afterward, we show how to adapt the argument to address matrix Rademacher series.

4.8.1 Hermitian Gaussian Series

Our main tool is the 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 4.8.1 (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

$$\mathbb{E}(\gamma^{2p+1}) = 0 \quad \text{and} \quad \mathbb{E}(\gamma^{2p}) = \frac{(2p)!}{p!\,2^p} \quad \text{for } p = 0,1,2,\dots.$$

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 $(2p)$th moment in terms of the $(2p-2)$th moment.

Therefore, the matrix mgf satisfies

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

The first identity holds because the odd terms in the series vanish. To compute the cgf, we extract the logarithm of the mgf and recall (2.1.8), which states that the matrix logarithm is the functional inverse of the matrix exponential. □

The results for the maximum and minimum eigenvalues of a matrix Gaussian series follow easily.

Proof of Theorem 4.1.1: Gaussian Case. Consider a finite sequence $\{\boldsymbol{A}_k\}$ of Hermitian matrices, 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 (4.1.3) for $\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y})$. The master expectation bound, relation (3.6.1) from Theorem 3.6.1, implies that

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

The second line follows when we introduce the cgf from Lemma 4.8.1. 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. Identify the variance parameter (4.1.2) in the exponent. The infimum is attained at $\theta = \sqrt{2\sigma^{-2}\log d}$, which leads to (4.1.3).

Next, we turn to the proof of the upper tail bound (4.1.4) for $\lambda_{\max}(\boldsymbol{Y})$. Invoke the master tail bound, relation (3.6.3) from Theorem 3.6.1, and calculate that

$$\begin{aligned}\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \le t\} &\le \inf_{\theta>0} \mathrm{e}^{-\theta t} \operatorname{tr}\exp\left(\sum\nolimits_k \log \mathbb{E}\,\mathrm{e}^{\gamma_k \theta \boldsymbol{A}_k}\right) \\ &= \inf_{\theta>0} \mathrm{e}^{-\theta t} \operatorname{tr}\exp\left(\frac{\theta^2}{2}\sum\nolimits_k \boldsymbol{A}_k^2\right) \\ &\le \inf_{\theta>0} \mathrm{e}^{-\theta t}\cdot d\exp\left(\frac{\theta^2}{2}\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{A}_k^2\right)\right) = d\inf_{\theta>0}\mathrm{e}^{-\theta t+\theta^2\sigma^2/2}.\end{aligned}$$

The steps here are the same as in the previous calculation. The infimum is achieved at $\theta = t/\sigma^2$, which yields (4.1.4). □

4.8.2 Hermitian Rademacher Series

The results 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 4.8.2 (Rademacher × Matrix: Mgf and Cgf)**.** *Suppose that $\boldsymbol{A}$ is a fixed Hermitian matrix, and let ε be a Rademacher random variable. Then*

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

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

$$\cosh(a) = \sum_{p=0}^{\infty}\frac{a^{2p}}{(2p)!} \le \sum_{p=0}^{\infty}\frac{a^{2p}}{2^p p!} = \mathrm{e}^{a^2/2} \quad \text{for } a\in\mathbb{R}. \tag{4.8.1}$$

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

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

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

The semidefinite bound follows when we apply the Transfer Rule (2.1.6) to the inequality (4.8.1).

To determine the matrix cgf, observe that

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

The semidefinite bound follows when we apply the Transfer Rule (2.1.6) to the bound $\log\cosh(a) \le a^2/2$ for $a\in\mathbb{R}$, which is a consequence of (4.8.1). □

We are prepared to develop probability inequalities for the extreme eigenvalues of a Rademacher series with matrix coefficients.

Proof of Theorem 4.1.1: Rademacher Case. Consider a finite sequence $\{\boldsymbol{A}_k\}$ of Hermitian matrices, and let $\{\varepsilon_k\}$ be a finite sequence of independent standard normal variables. Define the matrix Rademacher series

$$\boldsymbol{Y} = \sum\nolimits_k \varepsilon_k \boldsymbol{A}_k.$$

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

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

To obtain this result, we introduce the semidefinite bound, Lemma 4.8.2, for the Rademacher cgf into the trace exponential. The left-hand side increases after this substitution because of the fact (2.1.7) that the trace exponential function is monotone with respect to the semidefinite order. □

4.9 Proof of Bounds for Rectangular Matrix Series

Next, we consider a series with rectangular matrix coefficients modulated by independent Gaussian or Rademacher random variables. The bounds for the norm of a rectangular series follow instantly from the the bounds for the norm of an Hermitian series because of a formal device: We simply apply the Hermitian results to the Hermitian dilation (2.1.11) of the series.

Proof of Corollary 4.2.1. Consider a finite sequence $\{\boldsymbol{B}_k\}$ of $d_1 \times d_2$ complex matrices, and let $\{\xi_k\}$ 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

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

This leads us to form the two series

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

To analyze $\|\boldsymbol{Z}\|$, we wish to invoke Theorem 4.1.1. To make this step, we apply the fact (2.1.13) 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 $\|\boldsymbol{Z}\|$. To use these results, we must express the variance (4.1.2) of the random Hermitian matrix $\boldsymbol{Y}$ in terms of the general matrix $\boldsymbol{Z}$. Observe that

$$\begin{aligned} \sigma^2(\boldsymbol{Y}) = \left\|\mathbb{E}(\boldsymbol{Y}^2)\right\| = \left\|\mathbb{E}\left(\mathscr{H}(\boldsymbol{Z})^2\right)\right\| &= \left\|\mathbb{E}\begin{bmatrix} \boldsymbol{Z}\boldsymbol{Z}^* & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}^*\boldsymbol{Z} \end{bmatrix}\right\| \\ &= \left\|\begin{bmatrix} \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*) & \boldsymbol{0} \\ \boldsymbol{0} & \mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z}) \end{bmatrix}\right\| = \max\left\{\left\|\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*)\right\|, \left\|\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})\right\|\right\} = \sigma^2(\boldsymbol{Z}). \end{aligned}$$

The third relation is the identity (2.1.12) for the square of the Hermitian dilation. The penultimate equation holds because the norm of a block-diagonal matrix is the maximum norm of any diagonal block. We obtain the formula (4.2.1) for the variance of the matrix $\boldsymbol{Z}$.

We are, finally, prepared to apply Theorem 4.1.1, whose conclusions lead to the statement of Corollary 4.2.1. □

4.10 Notes

The material in this chapter is, perhaps, more firmly established than anything else in these lecture notes. We give an overview of research related to matrix Gaussian series, along with references for the specific random matrices that we have analyzed.

4.10.1 Matrix Gaussian and Rademacher Series

The main results, Theorem 4.1.1 and Corollary 4.2.1, have an interesting history. In the precise form stated here, these two results first appeared in [Tro11d], but we can trace them back more than two decades.

In his work [Oli10b, Thm. 1], Oliveira established the mgf bounds, Lemma 4.8.1 and Lemma 4.8.2. He also developed an ingenious improvement on the arguments of Ahlswede and Winter [AW02, App.] that gives a bound similar with Theorem 4.1.1. The constants in Oliveira's result are a bit worse, but the dependence on the dimension is sometimes better. We do not believe that the original approach of Ahlswede–Winter can deliver any of these results.

It turns out that Theorem 4.1.1 is roughly comparable with the noncommutative Khintchine inequality [LP86]. The noncommutative Khintchine inequality provides a bound for the expected trace of an even power of a matrix Gaussian series (or a matrix Rademacher series) in terms of the variance of the series. The sharpest forms [LPP91, Buc01, Buc05] are slightly more powerful than Theorem 4.1.1. Unfortunately, established proofs of the noncommutative Khintchine inequality are abstract or difficult or both. Recently, the paper [MJC+12] propounded an elementary proof, based on Stein's method of exchangeable pairs [Ste72, Cha07].

For a detailed exploration of the relationships between matrix concentration inequalities and noncommutative moment inequalities, see [Tro11d, Sec. 4]. This discussion also indicates the extent to which Theorem 4.1.1 and its relatives are sharp.

Recently, there have been some minor improvements to the dimensional factor that appears in Theorem 4.1.1. We discuss these results and give citations in Chapter 7.

4.10.2 Application to Random Matrices

It has also been known for a long time that results such as Theorem 4.1.1 can be used to study random matrices.

We believe that the 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 to study a problem connected with 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 very 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. See the paper [JX08] for a specific example. Unfortunately, this literature is technically formidable, which makes it difficult for outsiders to appreciate its achievements.

The work [AW02] of Ahslwede and Winter led to first "finished" 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 Ahslwede–Winter method 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, Tro11d], led to nearly 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. New applications of these ideas now appear on a weekly basis.

4.10.3 Wigner and Marčenko–Pastur

Wigner matrices first emerged in the literature on nuclear physics, where they were used to model the Hamiltonians of heavy atoms [Meh04]. Wigner showed that the limiting spectral distribution of a Wigner matrix follows the semicircle law; see [Tao12, §2.4] for an overview of the proof. 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] for a detailed treatment. The analysis that we present here, using Theorem 4.1.1, is drawn from [Tro11d, §4].

The first analysis of a rectangular Gaussian matrix is due to Marčenko and Pastur [MP67], who established that the limiting distribution of the squared singular values follows a semicircular distribution. The Bai–Yin law [BY93] gives an almost sure limit for the largest singular value of a rectangular Gaussian matrix. The expectation bound (4.4.4) appears in a survey article [DS02] by Davidson and Szarek. The expectation bound is ultimately derived from a comparison theorem for Gaussian processes due to Férnique and amplified by Gordon [Gor85]. Our approach, using Corollary 4.2.1, is based on [Tro11d, §4].

4.10.4 Randomly Signed Matrices

Matrices with randomly signed entries have not received much attention in the literature. The result (4.5.1) 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. The analysis here, using Corollary 4.2.1, appears in [Tro11d, §4].

4.10.5 Gaussian Toeplitz Matrices

Research on random Toeplitz matrices is quite recent, but there are now a number of papers available. Bryc, Dembo, and Jiang obtained the limiting spectral distribution of a symmetric Toeplitz matrix based on 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 and Virág computed the limiting value of the expected norm of a random, symmetric Toeplitz matrix whose entries have identical second-order statistics [SV11]. See the latter paper for additional references. The analysis here, based on Corollary 4.2.1, is new.

4.10.6 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 could be used to obtain an optimal result; he also showed that matrix concentration inequalities have applications to robust optimization. The presentation here, using Corollary 4.2.1, 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 study the extreme eigenvalues of an independent sum of random, positive-semidefinite matrices. This approach is valuable for controlling the norm of a random matrix and for understanding when a random matrix is singular.

More formally, we consider independent random matrices $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ with the properties

$$\boldsymbol{X}_k \succcurlyeq \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le R \quad \text{for each } k = 1, \dots, n.$$

Form 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 $\lambda_{\min}(\boldsymbol{Y})$. Matrix Chernoff inequalities offer all of these estimates. Note that it is better to use the matrix Bernstein inequalities, from Chapter 6, to study how much a random matrix deviates from its mean.

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 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 provides 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.

Section 5.1 presents the main results on the expectations and the tails of the extreme eigenvalues of a sum of independent, positive-semidefinite random matrices. Section 1.6.3 describes the application to sample covariance estimation, while §5.2 explains how the matrix Chernoff bounds provide spectral information about a random submatrix drawn from a fixed matrix. 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, positive 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 identical—Bernoulli trials with relatively 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 below the mean, while its upper tail drops off faster than an exponential random variable.

In the matrix setting, we encounter similar phenomena when we consider a sum of independent, positive-semidefinite random 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 that satisfy*

$$\boldsymbol{X}_k \succcurlyeq \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le R.$$

Define the random matrix

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

Compute the expectation parameters:

$$\mu_{\max} = \mu_{\max}(\boldsymbol{Y}) = \lambda_{\max}(\mathbb{E}\,\boldsymbol{Y}) \quad \text{and} \quad \mu_{\min} = \mu_{\min}(\boldsymbol{Y}) = \lambda_{\min}(\mathbb{E}\,\boldsymbol{Y}). \tag{5.1.1}$$

Then, for $\theta > 0$,

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \le \frac{\mathrm{e}^{\theta} - 1}{\theta}\,\mu_{\max} + \frac{1}{\theta}\,R \log d, \tag{5.1.2}$$

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

Furthermore,

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \ge (1+\delta)\mu_{\max}\right\} \le d\left[\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{\max}/R} \quad \text{for } \delta \ge 0, \quad \text{and} \tag{5.1.4}$$

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

The proofs of Theorem 5.1.1 appears below in §5.4.

5.1.1 Discussion

First, observe that we can easily compute the matrix expectation parameters $\mu_{\max}$ and $\mu_{\min}$ in terms of the coefficient matrices:

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

This point follows from the linearity of expectation.

In many situations, it is easier to work with streamlined versions of the bounds from Theorem 5.1.1:

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \le (\mathrm{e}-1)\,\mu_{\max} + R\log d, \quad \text{and} \tag{5.1.6}$$

$$\mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}) \ge (1-\mathrm{e}^{-1})\,\mu_{\min} - R\log d. \tag{5.1.7}$$

We obtain these results by selecting $\theta = 1$ in both (5.1.3) and (5.1.2). Note that, in the scalar case $d = 1$, we can take $\theta \to 0$ to obtain a numerical constant of one in each bound.

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

We can weaken the tail bounds to reach

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \ge t\mu_{\max}\right\} \le d\left(\frac{\mathrm{e}}{t}\right)^{t\mu_{\max}/R} \quad \text{for } t \ge \mathrm{e}, \quad \text{and}$$

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

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

5.2 Example: A Random Submatrix of a Fixed Matrix

The matrix Chernoff inequality plays an important role in bounding the extreme singular values of a random submatrix drawn from a fixed matrix. Although Theorem 5.1.1 might not seem suitable for this purpose (since it deals with eigenvalues), we can connect the problem with the method via a simple transformation. The results in this section have found applications in randomized linear algebra, sparse approximation, and other fields.

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 a sum of columns:

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

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

We consider a simple model for a random column submatrix. Let $\{\eta_k\}$ be an independent sequence of Bernoulli random variables with common mean q/n. Define the random matrix

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

That is, we include each column independently with probability q/n, which means that there are typically about q 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 of the $d \times n$ matrix $\boldsymbol{Z}$. In particular,

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

That is, the random submatrix $\boldsymbol{Z}$ gets its fair share of the spectrum 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 bound on $\sigma_d(\boldsymbol{Z})$ ensures that the nonzero columns of the random submatrix $\boldsymbol{Z}$ are linearly independent, at least on average.

The Analysis

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

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

Note that $\eta_k^2 = \eta_k$ because η_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 $\sigma_1 \ge \cdots \ge \sigma_d$.

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

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

so that

$$\mu_{\max} = \frac{q}{n}\,\sigma_1(\boldsymbol{B})^2 \quad \text{and} \quad \mu_{\min} = \frac{q}{n}\,\sigma_d(\boldsymbol{B})^2.$$

Define $R = \max_k \|\boldsymbol{b}_{:k}\|^2$, and observe that $\left\|\eta_k \boldsymbol{b}_{:k}\boldsymbol{b}_{:k}^*\right\| \le R$ for each k. Theorem 5.1.1 now ensures that

$$\begin{aligned} \mathbb{E}\left(\sigma_1(\boldsymbol{Z})^2\right) = \mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) &\le \frac{(\mathrm{e}-1)q}{n}\,\sigma_1(\boldsymbol{B})^2 + R\log d, \quad \text{and} \\ \mathbb{E}\left(\sigma_d(\boldsymbol{Z})^2\right) = \mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}) &\ge \frac{(1-\mathrm{e}^{-1})q}{n}\,\sigma_d(\boldsymbol{B})^2 - R\log d. \end{aligned}$$

We have taken $\theta = 1$ in the upper (5.1.2) and lower (5.1.3) bounds for the expectation. To obtain the stated result (5.2.1), we simply introduce numerical estimates for the constants.

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. Let

$$\boldsymbol{P} = \operatorname{diag}(\eta_1, \dots, \eta_d) \quad \text{and} \quad \boldsymbol{Q} = \operatorname{diag}(\xi_1, \dots, \xi_n)$$

where $\{\eta_k\}$ is an independent family of Bernoulli random variables with common mean p/d and $\{\xi_k\}$ is an independent family of Bernoulli random variables with common mean q/n. Then

$$\boldsymbol{Z} = \boldsymbol{PBQ}$$

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

In this section, we will show that

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

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 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 $\|\boldsymbol{Z}\|^2$ in terms of the maximum eigenvalue of a random positive-semidefinite matrix:

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

We have used the facts that $\boldsymbol{P} = \boldsymbol{P}^*$ and that $\boldsymbol{QQ}^* = \boldsymbol{Q}$. Invoking the matrix Chernoff inequality (5.1.2), conditional on the choice of $\boldsymbol{P}$, we obtain

$$\mathbb{E}\left(\|\boldsymbol{Z}\|^2\right) \le \frac{(\mathrm{e}-1)q}{n}\,\mathbb{E}\,\lambda_{\max}(\boldsymbol{PBB}^*\boldsymbol{P}) + \mathbb{E}\max_k \|(\boldsymbol{PB})_k\|^2 \cdot \log d. \quad (5.2.3)$$

The notation $(\boldsymbol{PB})_k$ refers to the kth column of the matrix $\boldsymbol{PB}$. The required calculation is analogous to 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:

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{PBB}^*\boldsymbol{P}) = \mathbb{E}\,\lambda_{\max}(\boldsymbol{B}^*\boldsymbol{P}^2\boldsymbol{B}) = \mathbb{E}\,\lambda_{\max}\left(\sum_{j=1}^{d}\eta_j\,\boldsymbol{b}_{j:}^*\boldsymbol{b}_{j:}\right).$$

The first identity holds because the nonzero eigenvalues of $\boldsymbol{C}\boldsymbol{C}^*$ equal the nonzero eigenvalues of $\boldsymbol{C}^*\boldsymbol{C}$ for any matrix $\boldsymbol{C}$. Another application of the matrix Chernoff inequality (5.1.2) yields

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{P}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{P}) \le \frac{(\mathrm{e}-1)p}{d}\,\lambda_{\max}(\boldsymbol{B}^*\boldsymbol{B}) + \max_j \left\|\boldsymbol{b}_{j:}\right\|^2 \cdot \log n. \tag{5.2.4}$$

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

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

Using this representation, we see that

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

When applied to a vector, the notation $|\cdot|^2$ refers to the componentwise modulus squared. To activate the matrix Chernoff bound, we need to compute the two parameters that appear in (5.1.2). First, the upper bound parameter R satisfies

$$R = \max_j \lambda_{\max}\left(\operatorname{diag}\left(\left|\boldsymbol{b}_{j:}\right|^2\right)\right) = \max_j \max_k \left|b_{jk}\right|^2.$$

Second, to compute the upper mean parameter $\mu_{\max}$, note that

$$\mathbb{E}\sum_{j=1}^{d} \eta_j \operatorname{diag}\left(\left|\boldsymbol{b}_{j:}\right|^2\right) = \frac{p}{d}\operatorname{diag}\left(\sum_{j=1}^{d}\left|\boldsymbol{b}_{j:}\right|^2\right) = \frac{p}{d}\operatorname{diag}\left(\left\|\boldsymbol{b}_{j:}\right\|^2\right),$$

which yields

$$\mu_{\max} = \frac{p}{d}\max_j \left\|\boldsymbol{b}_{j:}\right\|^2.$$

Therefore, the matrix Chernoff inequality implies

$$\mathbb{E}\max_k \|(\boldsymbol{P}\boldsymbol{B})_k\|^2 \le \frac{(e-1)p}{d}\max_j \left\|\boldsymbol{b}_{j:}\right\|^2 + \max_{j,k}\left|b_{jk}\right|^2 \cdot \log n. \tag{5.2.5}$$

On average, the maximum column norm of a random submatrix $\boldsymbol{P}\boldsymbol{B}$ with about p nonzero rows gets its share p/d of the maximum 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 some vertices are 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 $G = (V, E)$ where V is a set of vertices and E is a set of edges connecting pairs of distinct vertices. 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 some 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* and *normalized Laplacian* of the graph are the matrices

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

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

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

5.3.2 The Model of Erdős and 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

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

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

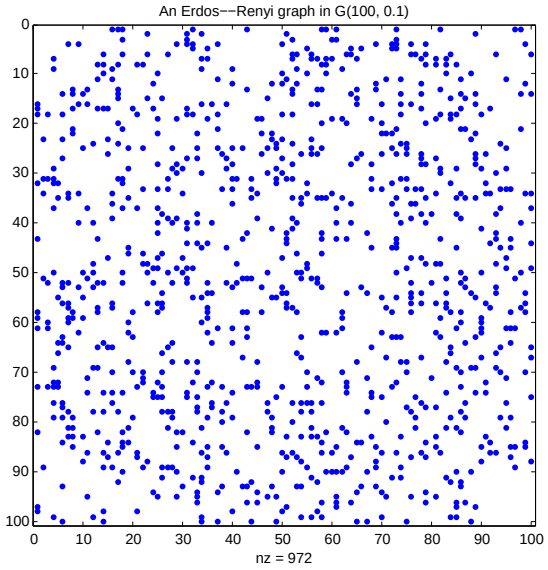


Figure 5.1: **The adjacency matrix of an Erdős–Rényi graph.** This figure shows the pattern of nonzero entries in the adjacency matrix $\boldsymbol{A}$ of a random graph drawn from $G(100, 0.1)$. Out of a possible 4,950 edges, there are 486 edges present. A basic question is whether the graph is connected.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

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 graph in $G(n, p)$ can be written as

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

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

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

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 $\delta_{jk} = 1$, we augment the (j, j) and (k, k) entries of $\boldsymbol{L}$ 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 small-

est eigenvalue of the $n \times n$ random Laplacian matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ is strictly positive. We will solve the problem by using the matrix Chernoff inequality to study the second-smallest eigenvalue of the random Laplacian $\boldsymbol{L}$.

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 $\boldsymbol{L}$. To that end, define an $(n-1) \times n$ partial unitary matrix $\boldsymbol{R}$ that restricts a vector to the orthogonal complement of the vector $\mathbf{1}$ of ones. That is, the rows of $\boldsymbol{R}$ form an orthonormal family and the null space of $\boldsymbol{R}$ is the vector $\mathbf{1}$. Now, consider the random matrix

$$\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{L}\boldsymbol{R}^* = \sum_{1 \le j < k \le n} \delta_{jk} \cdot \boldsymbol{R}(\boldsymbol{E}_{jj} + \boldsymbol{E}_{kk} - \boldsymbol{E}_{jk} - \boldsymbol{E}_{kj})\boldsymbol{R}^*. \tag{5.3.4}$$

Recall that $\{\delta_{jk}\}$ is an independent family of BERNOULLI(p) random variables, so the summands are mutually independent. The Conjugation Rule (2.1.4) ensures that each summand remains positive-semidefinite. Since $\mathbf{1}$ is an eigenvector with eigenvalue zero associated with the positive-semidefinite matrix $\boldsymbol{L}$, the minimum eigenvalue of $\boldsymbol{Y}$ coincides with the second-smallest eigenvalue of $\boldsymbol{L}$.

To apply the matrix Chernoff inequality, we need a uniform upper bound B on the eigenvalues of the summands. We have

$$\begin{aligned} B &\le \left\| \delta_{jk} \cdot \boldsymbol{R}(\boldsymbol{E}_{jj} + \boldsymbol{E}_{kk} - \boldsymbol{E}_{jk} - \boldsymbol{E}_{kj})\boldsymbol{R}^* \right\| \\ &\le \left|\delta_{jk}\right| \cdot \left\|\boldsymbol{R}\right\| \cdot \left\|\boldsymbol{E}_{jj} + \boldsymbol{E}_{kk} - \boldsymbol{E}_{jk} - \boldsymbol{E}_{kj}\right\| \cdot \left\|\boldsymbol{R}^*\right\| = 2. \end{aligned}$$

The first bound follows from the submultiplicativity of the spectral norm. To obtain the second bound, note that δ_{jk} takes 0–1 values. The matrix $\boldsymbol{R}$ is a partial isometry so its norm equals one. Finally, a direct calculation shows that $\boldsymbol{T} = \boldsymbol{E}_{jj} + \boldsymbol{E}_{kk} - \boldsymbol{E}_{jk} - \boldsymbol{E}_{kj}$ satisfies the polynomial $\boldsymbol{T}^2 = 2\boldsymbol{T}$, so the eigenvalues of $\boldsymbol{T}$ must equal zero and two.

Next, we compute the expectation of the matrix $\boldsymbol{Y}$.

$$\begin{aligned} \mathbb{E}\,\boldsymbol{Y} &= p \cdot \boldsymbol{R}\left[\sum_{1 \le j < k \le n} (\mathbf{E}_{jj} + \mathbf{E}_{kk} - \mathbf{E}_{jk} - \mathbf{E}_{kj}) \right] \boldsymbol{R}^* \\ &= p \cdot \boldsymbol{R}\big[(n-1)\,\mathbf{I}_n - (\mathbf{1}\mathbf{1}^* - \mathbf{I}_n) \big] \boldsymbol{R}^* = pn\,\mathbf{I}_{n-1}. \end{aligned}$$

The first identity follows when we apply linearity of expectation to (5.3.4) 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 $\mathbf{1}\mathbf{1}^* - \mathbf{I}_n$ comes from the off-diagonal matrix units, once we note that the matrix $\mathbf{1}\mathbf{1}^*$ has one in each component. The last identity holds because $\boldsymbol{R}$ annihilates the vector $\mathbf{1}$, while $\boldsymbol{R}\boldsymbol{R}^* = \mathbf{I}_{n-1}$. We conclude that

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

This is all the information we need.

Invoke the tail bound (5.1.5) to obtain, for $\varepsilon \in (0,1)$,

$$\mathbb{P}\left\{ \lambda_2^{\uparrow}(\boldsymbol{L}) \le \varepsilon \cdot pn \right\} = \mathbb{P}\left\{ \lambda_{\min}(\boldsymbol{Y}) \le \varepsilon \cdot pn \right\} \le (n-1) \left[\frac{\mathrm{e}^{\varepsilon - 1}}{\varepsilon^{\varepsilon}} \right]^{pn/2}.$$

To appreciate what this means, we may think about the situation where $\varepsilon \to 0$. Then the bracket tends to e^{-1}, and we see that the second smallest eigenvalue of $\boldsymbol{L}$ 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 bound 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 positive-semidefinite matrix that satisfies $\lambda_{\max}(\boldsymbol{X}) \le R$. Then*

$$\mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \exp\left(\frac{\mathrm{e}^{R\theta} - 1}{R} \cdot (\mathbb{E}\, \boldsymbol{X}) \right) \quad \textit{and} \quad \log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \frac{\mathrm{e}^{R\theta} - 1}{R} \cdot (\mathbb{E}\, \boldsymbol{X}) \quad \textit{for } \theta \in \mathbb{R}.$$

Proof. Consider the function $f(x) = \mathrm{e}^{\theta x}$. Since f is convex, its graph lies below the chord connecting two points. In particular,

$$f(x) \le f(0) + \frac{f(R) - f(0)}{R} \cdot x \quad \text{for } x \in [0, R].$$

In detail,

$$\mathrm{e}^{\theta x} \le 1 + \frac{\mathrm{e}^{R\theta} - 1}{R} \cdot x \quad \text{for } x \in [0, R].$$

By assumption, each eigenvalue of $\boldsymbol{X}$ lie in the interval $[0, R]$. Thus, the Transfer Rule (2.1.6) implies that

$$\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \mathbf{I} + \frac{\mathrm{e}^{R\theta} - 1}{R} \cdot \boldsymbol{X}.$$

Expectation respects the semidefinite order, so

$$\mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \mathbf{I} + \frac{\mathrm{e}^{R\theta} - 1}{R} \cdot (\mathbb{E}\, \boldsymbol{X}) \preccurlyeq \exp\left(\frac{\mathrm{e}^{R\theta} - 1}{R} \cdot (\mathbb{E}\, \boldsymbol{X}) \right).$$

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 (2.1.6) 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 (2.1.9) 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 that satisfy

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

The cgf bound, Lemma 5.4.1, states that

$$\log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq g(\theta)(\mathbb{E}\, \boldsymbol{X}_k) \quad \text{where} \quad g(\theta) = \frac{\mathrm{e}^{R\theta} - 1}{R} \quad \text{for } \theta > 0. \tag{5.4.1}$$

We begin with the upper bound (5.1.2) for $\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y})$. Using the fact (2.1.7) 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 maximum eigenvalue to reach

$$\begin{aligned}
\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) &\le \inf_{\theta>0} \frac{1}{\theta} \log \operatorname{tr} \exp\bigl(g(\theta) \textstyle\sum_k \mathbb{E}\, \boldsymbol{X}_k\bigr) \\
&\le \inf_{\theta>0} \frac{1}{\theta} \log \bigl[d\, \lambda_{\max}\bigl(\exp\bigl(g(\theta)(\mathbb{E}\, \boldsymbol{Y})\bigr)\bigr)\bigr] \\
&= \inf_{\theta>0} \frac{1}{\theta} \log \bigl[d\, \exp\bigl(\lambda_{\max}\bigl(g(\theta)(\mathbb{E}\, \boldsymbol{Y})\bigr)\bigr)\bigr] \\
&= \inf_{\theta>0} \frac{1}{\theta} \log \bigl[d\, \exp\bigl(g(\theta) \cdot \lambda_{\max}(\mathbb{E}\, \boldsymbol{Y})\bigr)\bigr] \\
&= \inf_{\theta>0} \frac{1}{\theta} \bigl[\log d + g(\theta) \cdot \mu_{\max}\bigr].
\end{aligned}$$

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 (2.1.2) that the maximum eigenvalue map is positive homogeneous, which depends on the observation that $g(\theta) > 0$ for $\theta > 0$. Finally, we identify the quantity $\mu_{\max}$, defined in (5.1.1). The infimum does not admit a closed form, but we can obtain the expression (5.1.2) by making the change of variables $\theta \mapsto \theta / R$.

Next, we turn to the upper bound (5.1.4) for the upper tail of the maximum eigenvalue. Substitute the cgf bounds (5.4.1) into the master inequality (3.6.3) to reach

$$\begin{aligned}
\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} &\le \inf_{\theta>0} \mathrm{e}^{-\theta t} \operatorname{tr} \exp\bigl(g(\theta) \textstyle\sum_k \mathbb{E}\, \boldsymbol{X}_k\bigr) \\
&\le \inf_{\theta>0} \mathrm{e}^{-\theta t} \cdot d\, \exp\bigl(g(\theta) \cdot \mu_{\max}\bigr).
\end{aligned}$$

The steps here are identical with the previous argument. Make the change of variables $t \mapsto (1+\delta)\mu_{\max}$. The infimum is achieved at $\theta = R^{-1} \log(1+\delta)$, which leads to the tail bound (5.1.4). □

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 that satisfy

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

The cgf bound, Lemma 5.4.1, states that

$$\log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq g(\theta) \cdot (\mathbb{E}\, \boldsymbol{X}_k) \quad \text{where} \quad g(\theta) = \frac{\mathrm{e}^{R\theta} - 1}{R} \quad \text{for } \theta < 0. \tag{5.4.2}$$

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 (2.1.7), the trace exponential function is monotone with respect to the semidefinite order, so the master inequality (3.6.2) for the minimum eigenvalue delivers

$$\begin{aligned}
\mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}) &\geq \sup_{\theta<0} \frac{1}{\theta} \log \operatorname{tr} \exp\bigl(g(\theta) \textstyle\sum_k \mathbb{E}\, \boldsymbol{X}_k\bigr) \\
&\geq \sup_{\theta<0} \frac{1}{\theta} \log \bigl[d\, \lambda_{\max}\bigl(\exp\bigl(g(\theta) \cdot (\mathbb{E}\, \boldsymbol{Y})\bigr)\bigr)\bigr] \\
&= \sup_{\theta<0} \frac{1}{\theta} \log \bigl[d\, \exp\bigl(\lambda_{\max}\bigl(g(\theta) \cdot (\mathbb{E}\, \boldsymbol{Y})\bigr)\bigr)\bigr] \\
&= \sup_{\theta<0} \frac{1}{\theta} \log \bigl[d\, \exp\bigl(g(\theta) \cdot \lambda_{\min}(\mathbb{E}\, \boldsymbol{Y})\bigr)\bigr] \\
&= \sup_{\theta<0} \frac{1}{\theta} \bigl[\log d + g(\theta) \cdot \mu_{\min}\bigr].
\end{aligned}$$

Most of the steps are the same as in the proof of the upper bound (5.1.2), 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 (2.1.2) and (2.1.3). 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: $\theta \mapsto -\theta/R$.

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

$$\begin{aligned}
\mathbb{P}\{\lambda_{\min}(\boldsymbol{Y}) \leq t\} &\leq \inf_{\theta<0} \mathrm{e}^{-\theta t} \operatorname{tr} \exp\bigl(g(\theta) \textstyle\sum_k \mathbb{E}\, \boldsymbol{X}_k\bigr) \\
&\leq \inf_{\theta<0} \mathrm{e}^{-\theta t} \cdot d\, \exp\bigl(g(\theta) \cdot \mu_{\min}\bigr).
\end{aligned}$$

The justifications here match those in with the previous argument. Make the change of variables $t \mapsto (1-\delta)\mu_{\min}$. The infimum is attained at $\theta = R^{-1} \log(1-\delta)$, 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 and 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 iid random matrices, in which case their results are comparable with Theorem 5.1.1. For the general case, their approach leads to mean parameters of the form

$$\mu_{\max}^{\mathrm{AW}} = \sum\nolimits_k \lambda_{\max}(\mathbb{E}\,\boldsymbol{X}_k) \quad \text{and} \quad \mu_{\min}^{\mathrm{AW}} = \sum\nolimits_k \lambda_{\min}(\mathbb{E}\,\boldsymbol{X}_k).$$

It is clear that these mean parameters may be substantially inferior to the mean parameters $\mu_{\max}$ and $\mu_{\min}$ that we defined in Theorem 5.1.1.

The tail bounds from Theorem 5.1.1 are drawn from [Tro11d, §5], but the expectation bounds we present are new. The paper [GT11] extends the matrix Chernoff inequality to provide upper and lower tail bounds for all eigenvalues of a sum of positive-semidefinite random matrices. Finally, Chapter 7 contains a slight improvement of the upper bounds from Theorem 5.1.1.

5.5.2 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 well-conditioned set of columns (or a well-conditioned submatrix) inside a fixed matrix. Random selection provides a natural approach to this question. The papers [BT87, BT91, KT94] study random paving using sophisticated tools from functional analysis. See the paper [NT12] for a summary of research on the paving problem.

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. See, for example, the paper [Tro08a]. The same approach led to the work [Tro08c], which contains a new proof of [BT91, Thm. 2.1].

The article [Tro11e] 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, and it achieves the optimal estimate for this problem. 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.

5.5.3 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.

The device we have used to analyze the second smallest eigenvalue of a random graph Laplacian can be extended to obtain tail bounds for all the eigenvalues of a sum of independent random matrices. See the paper [GT11] 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, bounded random matrices. The results allow us to study how much a random matrix deviates from its mean value in the spectral norm.

To be rigorous, let us suppose that $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are independent random matrices with the properties

$$\mathbb{E}\,\boldsymbol{X}_k = \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le R \quad \text{for each } k = 1, \dots, n.$$

Form the sum $\boldsymbol{Y} = \sum_k \boldsymbol{X}_k$. The matrix Bernstein inequality allows us to study the expectation and tail behavior of $\lambda_{\max}(\boldsymbol{Y})$ in terms of the variance $\mathbb{E}(\boldsymbol{Y}^2)$.

Matrix Bernstein inequalities have a much wider scope of application than the last paragraph might suggest. First, if the summands are not centered, we can subtract off the mean and use the matrix Bernstein method to obtain information about $\lambda_{\max}(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y})$. Second, we can obtain bounds for the minimum eigenvalue $\lambda_{\min}(\boldsymbol{Y})$ by applying the matrix Bernstein bounds to $-\boldsymbol{Y}$. Third, we can extend the result to study the spectral norm of a sum of independent, general random matrices that satisfy a uniform norm bound.

In these pages, we can only give a coarse indication of how researchers have used the matrix Bernstein inequality. We have selected two typical examples from the literature on randomized matrix approximation. First, we explain how to develop a randomized algorithm for approximate matrix multiplication, and we establish an error bound for this method. Second, we consider the technique of randomized sparsification, in which we replace a dense matrix with a sparse proxy that has similar spectral behavior. There are many other examples, some of which appear in the annotated bibliography.

Altogether, the matrix Bernstein inequality is a powerful tool with a huge number of applications. It is particularly effective for studying randomized approximations of a given matrix. Nevertheless, let us emphasize that, when the matrix Chernoff inequality, Theorem 5.1.1, happens to apply, it often delivers better results for a given problem.

Section 6.1 describes the Bernstein inequality for Hermitian matrices, and §6.2 presents the adaptation to general matrices. Afterward, in §§6.3–6.4, we continue with the two randomized approximation examples. We conclude with the proof of the matrix Bernstein inequalities in §6.5.

6.1 A Sum of Bounded Hermitian Matrices

In the scalar setting, there are a large number of concentration bounds that fall under the heading "Bernstein inequality." Most of these bounds have extensions to matrices. For simplicity, we focus on the most famous of them all, a tail bound for the sum Y of independent, zero-mean random variables that are subject to a uniform upper bound. In this case, the Bernstein inequality shows that the sum Y concentrates around its mean value. For moderate deviations, the sum behaves like a normal random variable with the same variance as Y. For large deviations, the sum has tails that decay at least as fast as an exponential random variable.

In analogy, the matrix Bernstein inequality concerns a sum of independent, zero-mean Hermitian matrices whose eigenvalues are bounded above. The theorem demonstrates that the maximum eigenvalue of the sum acts much like the scalar random variable Y that we discussed in the last paragraph.

Theorem 6.1.1 (Matrix Bernstein: Hermitian Case)**.** *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, Hermitian matrices with dimension d. Assume that*

$$\mathbb{E}\,\boldsymbol{X}_k = \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le R.$$

Introduce the random matrix

$$\boldsymbol{Y} = \textstyle\sum_k \boldsymbol{X}_k.$$

Compute the variance parameter

$$\sigma^2 = \sigma^2(\boldsymbol{Y}) = \left\|\mathbb{E}\left(\boldsymbol{Y}^2\right)\right\|. \tag{6.1.1}$$

Then

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \le \sqrt{2\sigma^2 \log d} + \frac{1}{3} R \log d. \tag{6.1.2}$$

Furthermore, for all $t \ge 0$.

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} \le d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right). \tag{6.1.3}$$

The proof of Theorem 6.1.1 appears below in §6.5.

6.1.1 Discussion

Let us spend a few moments to discuss this result and its implications. First, observe that we can express the variance parameter (6.1.1) in terms of the summands:

$$\sigma^2(\boldsymbol{Y}) = \left\|\mathbb{E}\left(\boldsymbol{Y}^2\right)\right\| = \left\|\mathbb{E}\left(\textstyle\sum_{j,k} \boldsymbol{X}_j \boldsymbol{X}_k\right)\right\| = \left\|\textstyle\sum_k \mathbb{E}\left(\boldsymbol{X}_k^2\right)\right\|.$$

The second relation holds because the summands are independent, and each one has zero mean. This identity parallels the scalar result that the variances of a sum of independent random variables is the sum of the variances.

The expectation bound (6.1.2) shows that the expectation of $\lambda_{\max}(\boldsymbol{Y})$ is on the same scale as the standard deviation σ and the upper bound R on the summands; there is also a weak dependence on the ambient dimension d. In general, all three of these features are necessary.

Next, let us interpret the tail bound (6.1.3). The only difference between this result and the scalar Bernstein bound is the addition of the dimensional factor d, 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:

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \geq t\} \leq \begin{cases} d \cdot \exp(-3t^2/8\sigma^2), & t \leq \sigma^2/R \\ d \cdot \exp(-3t/8R), & t \geq \sigma^2/R. \end{cases} \tag{6.1.4}$$

In other words, for moderate values of t, the tail probability decays as fast as the tail of a Gaussian random variable with variance $4\sigma^2/3$. For larger values of t, the tail probability decays at least as fast as that of an exponential random variable with mean $4R/3$.

Next, we point out that Theorem 6.1.1 also yields information about the *minimum* eigenvalue of an independent sum of d-dimensional Hermitian matrices. In this case, we must assume that

$$\mathbb{E}\boldsymbol{X}_k = \mathbf{0} \quad \text{and} \quad \lambda_{\min}(\boldsymbol{X}_k) \geq -\underline{R}.$$

Form the random matrix $\boldsymbol{Y} = \sum_k \boldsymbol{X}_k$. By applying the expectation bound (6.1.2) to $-\boldsymbol{Y}$, we obtain

$$\mathbb{E}\lambda_{\min}(\boldsymbol{Y}) \geq -\sqrt{2\sigma^2 \log d} - \frac{1}{3}\underline{R}\log d \tag{6.1.5}$$

where $\sigma^2 = \sigma^2(\boldsymbol{Y})$. We can use (6.1.3) to develop a tail bound. For $t \geq 0$,

$$\mathbb{P}\{\lambda_{\min}(\boldsymbol{Y}) \leq -t\} \leq d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + \underline{R}t/3}\right).$$

Let us emphasize that the bounds for $\lambda_{\max}(\boldsymbol{Y})$ and the bounds for $\lambda_{\min}(\boldsymbol{Y})$ may diverge because the two parameters R and $\underline{R}$ can take sharply different values.

Finally, it is important to recognize that the matrix Bernstein inequality applies just as well to uncentered matrices. Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random Hermitian matrices with dimension d. Assume that each matrix satisfies the bound

$$\lambda_{\max}(\boldsymbol{X}_k - \mathbb{E}\boldsymbol{X}_k) \leq R.$$

Introduce the sum $\boldsymbol{Y} = \sum_k \boldsymbol{X}_k$, and compute the variance parameter

$$\sigma^2 = \sigma^2(\boldsymbol{Y}) = \left\|\mathbb{E}\left((\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y})^2\right)\right\| = \left\|\sum\nolimits_k \mathbb{E}\left((\boldsymbol{X}_k - \mathbb{E}\boldsymbol{X}_k)^2\right)\right\|.$$

Then we have the expectation bound

$$\mathbb{E}\lambda_{\max}(\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y}) \leq \sqrt{2\sigma^2 \log d} + \frac{1}{3}R\log d.$$

Furthermore, for $t \geq 0$,

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y}) \geq t\} \leq d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

Similar results hold for $\lambda_{\min}(\boldsymbol{Y})$, as discussed in the previous paragraph.

There are many other types of matrix Bernstein inequalities. For example, we can sharpen the tail bound (6.1.3) 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 $\boldsymbol{X}_k$. See the notes at the end of this chapter and the annotated bibliography for more information.

6.2 A Sum of Bounded Rectangular Matrices

The matrix Bernstein inequality admits an extension to a sum of general random matrices that are subject to a uniform norm bound. This result turns out to be a formal consequence of the Hermitian result, Theorem 6.1.1, even though it may initially seem more powerful.

Corollary 6.2.1 (Matrix Bernstein: General Case)**.** *Consider a finite sequence $\{\boldsymbol{S}_k\}$ of independent, random matrices with dimension $d_1 \times d_2$. Assume that*

$$\mathbb{E}\,\boldsymbol{S}_k = \mathbf{0} \quad \text{and} \quad \|\boldsymbol{S}_k\| \le R.$$

Introduce the random matrix

$$\boldsymbol{Z} = \textstyle\sum_k \boldsymbol{S}_k.$$

Compute the variance parameter

$$\sigma^2 = \sigma^2(\boldsymbol{Z}) = \max\left\{\left\|\mathbb{E}\left(\boldsymbol{Z}\boldsymbol{Z}^*\right)\right\|, \left\|\mathbb{E}\left(\boldsymbol{Z}^*\boldsymbol{Z}\right)\right\|\right\}. \tag{6.2.1}$$

Then

$$\mathbb{E}\,\|\boldsymbol{Z}\| \le \sqrt{2\sigma^2 \log(d_1+d_2)} + \frac{1}{3}R\log(d_1+d_2). \tag{6.2.2}$$

Furthermore, for all $t \ge 0$,

$$\mathbb{P}\{\|\boldsymbol{Z}\| \ge t\} \le (d_1+d_2)\exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3}\right).$$

The proof of Corollary 6.2.1 appears in §6.5.

6.2.1 Discussion

The general case is similar with the Hermitian case, Theorem 6.1.1, in many respects. Corollary 6.2.1 also has a lot in common with Corollary 4.2.1, concerning a Gaussian series with general matrix coefficients. As a consequence, we do not indulge in an extensive commentary.

First, let us express the variance parameter (6.2.1) in terms of the summands:

$$\begin{aligned}\sigma^2(\boldsymbol{Z}) &= \max\left\{\left\|\mathbb{E}\left(\boldsymbol{Z}\boldsymbol{Z}^*\right)\right\|, \left\|\mathbb{E}\left(\boldsymbol{Z}^*\boldsymbol{Z}\right)\right\|\right\} \\ &= \max\left\{\left\|\mathbb{E}\left(\textstyle\sum_{j,k}\boldsymbol{S}_j\boldsymbol{S}_k^*\right)\right\|, \left\|\mathbb{E}\left(\textstyle\sum_{j,k}\boldsymbol{S}_j^*\boldsymbol{S}_k\right)\right\|\right\} \\ &= \max\left\{\left\|\textstyle\sum_k \mathbb{E}\left(\boldsymbol{S}_k\boldsymbol{S}_k^*\right)\right\|, \left\|\textstyle\sum_k \mathbb{E}\left(\boldsymbol{S}_k^*\boldsymbol{S}_k\right)\right\|\right\}.\end{aligned}$$

As usual, the last relation holds because the summands are independent, zero-mean random matrices.

The version of Corollary 6.2.1 for uncentered matrices is important enough that we lay out the details. Consider a finite sequence $\{\boldsymbol{S}_k\}$ of independent, random matrices with dimension $d_1 \times d_2$. Assume that each matrix satisfies the bound

$$\|\boldsymbol{S}_k - \mathbb{E}\,\boldsymbol{S}_k\| \le R.$$

Introduce the sum $\boldsymbol{Z} = \sum_k \boldsymbol{S}_k$, and compute the variance parameter

$$\begin{aligned}\sigma^2 &= \max\left\{\left\|\mathbb{E}\left((\boldsymbol{Z}-\mathbb{E}\,\boldsymbol{Z})(\boldsymbol{Z}-\mathbb{E}\,\boldsymbol{Z})^*\right)\right\|,\ \left\|\mathbb{E}\left((\boldsymbol{Z}-\mathbb{E}\,\boldsymbol{Z})^*(\boldsymbol{Z}-\mathbb{E}\,\boldsymbol{Z})\right)\right\|\right\} \\ &= \max\left\{\left\|\textstyle\sum_k \mathbb{E}\left((\boldsymbol{S}_k-\mathbb{E}\,\boldsymbol{S}_k)(\boldsymbol{S}_k-\mathbb{E}\,\boldsymbol{S}_k)^*\right)\right\|,\ \left\|\textstyle\sum_k \mathbb{E}\left((\boldsymbol{S}_k-\mathbb{E}\,\boldsymbol{S}_k)^*(\boldsymbol{S}_k-\mathbb{E}\,\boldsymbol{S}_k)\right)\right\|\right\}.\end{aligned}$$

Then we have the expectation bound

$$\mathbb{E}\,\|\boldsymbol{Z}-\mathbb{E}\,\boldsymbol{Z}\| \le \sqrt{2\sigma^2 \log(d_1+d_2)} + \frac{1}{3}R\log(d_1+d_2).$$

Furthermore, for $t \ge 0$,

$$\mathbb{P}\{\|\boldsymbol{Z}-\mathbb{E}\,\boldsymbol{Z}\| \ge t\} \le (d_1+d_2)\cdot\exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

The results in this paragraph are probably the most commonly used versions of the matrix Bernstein bounds.

6.3 Application: Randomized Sparsification of a Matrix

Many tasks in data analysis require spectral computations on large, dense matrices. Yet many spectral decomposition algorithms operate most efficiently on sparse matrices. If we can tolerate approximate results, we may be able to reduce the computational cost by replacing the original dense matrix with a sparse proxy that has a similar spectrum. An elegant way to identify the sparse proxy is to randomly zero out entries of the original matrix. In this example, we examine the performance of one such approach.

Let $\boldsymbol{B}$ be a fixed $d_1 \times d_2$ complex matrix. Write $L = \max_{jk}|b_{jk}|$ for the maximum absolute entry of the matrix. Fix a sparsification parameter $p \in (0,1)$, and consider a family of independent Bernoulli random variables:

$$\delta_{jk} \sim \text{BERNOULLI}(p_{jk}) \quad \text{where} \quad p_{jk} = \frac{p\,|b_{jk}|}{p\,|b_{jk}| + L}.$$

It is easy to verify that $0 \le p_{jk} < 1$, so this is a legitimate probability. Draw a random, sparse matrix $\boldsymbol{Z}$ with entries

$$Z_{jk} = \delta_{jk}\frac{b_{jk}}{p_{jk}} \quad \text{for } j = 1,\dots,d_1 \text{ and } k = 1,\dots,d_2.$$

In other words, we zero out small entries with high probability, and we zero out larger entries with low probability. We rescale the entries that we keep to compensate. Observe that $\mathbb{E}(Z_{jk}) = b_{jk}$, so that the random sparse matrix $\boldsymbol{Z}$ is an unbiased approximation to the original matrix $\boldsymbol{B}$.

We must assess the typical sparsity of the random matrix $\boldsymbol{Z}$, and we must bound the distance between $\boldsymbol{Z}$ and the original matrix $\boldsymbol{B}$ in the spectral norm. An elementary calculation shows that the expected number of nonzero entries in $\boldsymbol{Z}$ is at most

$$\sum_{j,k} \mathbb{E}(\delta_{jk}) = \sum_{j,k} \frac{p\,|b_{jk}|}{p\,|b_{jk}| + L} \le p \sum_{j,k} \frac{|b_{jk}|}{L} \le p \cdot d_1 d_2.$$

So the parameter p is a bound for the proportion of nonzero entries appearing in the reduced matrix. We will show that the expected approximation error satisfies

$$\frac{\mathbb{E}\,\|\boldsymbol{Z} - \boldsymbol{B}\|}{\|\boldsymbol{B}\|} \le \sqrt{\frac{2L \max\{d_1,\, d_2\} \log(d_1 + d_2)}{p\,\|\boldsymbol{B}\|^2}} + \frac{2L \log(d_1 + d_2)}{3p\,\|\boldsymbol{B}\|}. \tag{6.3.1}$$

Ignoring the logarithmic factors, we learn that it is possible to construct a sparse matrix that approximates $\boldsymbol{B}$ with a small relative error, provided that

$$L \ll \|\boldsymbol{B}\| \quad \text{and} \quad L \max\{d_1,\, d_2\} \ll \|\boldsymbol{B}\|^2$$

Matrices whose largest entries are relatively small as compared with the norm are natural candidates for this type of processing.

The Analysis

We will use the matrix Bernstein inequality to study how well the sparsification procedure preserves the spectral properties of the original matrix. For reference, we calculate the mean and variance of the entries of $\boldsymbol{Z}$:

$$\mathbb{E}\, Z_{jk} = b_{jk} \quad \text{and} \quad \mathbb{E}\,|Z_{jk} - b_{jk}|^2 = \frac{L}{p}.$$

It follows that $\mathbb{E}\,\boldsymbol{Z} = \boldsymbol{B}$. Define the error matrix $\boldsymbol{E} = \boldsymbol{Z} - \boldsymbol{B}$, and write

$$\boldsymbol{E} = \sum_{j,k} (Z_{jk} - b_{jk})\, \mathbf{E}_{jk} = \sum_{j,k} \boldsymbol{S}_{jk},$$

where the expression above defines the summands $\boldsymbol{S}_{jk}$. It is immediate that each summand satisfies $\mathbb{E}\,\boldsymbol{S}_{jk} = \mathbf{0}$ and that $\{\boldsymbol{S}_{jk}\}$ is an independent family.

To apply the matrix Bernstein inequality, we first observe that the summands satisfy a uniform bound:

$$\|\boldsymbol{S}_{jk}\| \le \frac{|b_{jk}|}{p_{jk}} \le \frac{p\,|b_{jk}| + L}{p} \le \frac{2L}{p}.$$

Determining the variance of the error matrix $\boldsymbol{E}$ takes a little more work. We have

$$\mathbb{E}\left(\boldsymbol{S}_{jk} \boldsymbol{S}_{jk}^*\right) = \left[\mathbb{E}\,|Z_{jk} - b_{jk}|^2\right] \mathbf{E}_{jk} \mathbf{E}_{kj} = \frac{L}{p}\, \mathbf{E}_{jj}.$$

It follows that

$$\left\| \sum_{j,k} \mathbb{E}\left(\boldsymbol{S}_{jk} \boldsymbol{S}_{jk}^*\right) \right\| = \frac{L}{p} \left\| d_2\, \mathbf{I}_{d_1} \right\| = \frac{d_2 L}{p}.$$

Similarly,

$$\left\| \sum_{j,k} \mathbb{E}\left(\boldsymbol{S}_{jk}^* \boldsymbol{S}_{jk} \right) \right\| = \frac{d_1 L}{p}.$$

We conclude that the variance (6.2.1) of the error matrix

$$\sigma^2(\boldsymbol{E}) = \frac{L}{p} \max\{d_1,\ d_2\}.$$

The expectation bound (6.2.2) from Corollary 6.2.1 delivers

$$\mathbb{E}\,\|\boldsymbol{E}\| \le \sqrt{\frac{2L \max\{d_1,\ d_2\} \log(d_1 + d_2)}{p}} + \frac{2L \log(d_1 + d_2)}{3p}.$$

The result (6.3.1) is a direct consequence of this inequality.

6.4 Application: Randomized Matrix Multiplication

Over the last decade, randomized algorithms have started to play an important role in numerical linear algebra. One of the basic tasks in 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 $\boldsymbol{BC}$. The straightforward algorithm forms the product entry by entry:

$$(\boldsymbol{BC})_{ik} = \sum_{j=1}^{N} b_{ij} c_{jk} \quad \text{for each } i = 1, \dots, d_1 \text{ and } k = 1, \dots, d_2. \tag{6.4.1}$$

This approach takes $\mathrm{O}(N d_1 d_2)$ 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.

In certain circumstances, we can accelerate matrix multiplication using randomized methods. The key to this approach is to view the matrix product as a sum of outer products:

$$\boldsymbol{BC} = \sum_{k=1}^{N} \boldsymbol{b}_{:k} \boldsymbol{c}_{k:}. \tag{6.4.2}$$

Next, we reinterpret this sum as the expectation of a random matrix. It takes some care to do this properly. Define a set of probabilities

$$p_j = \frac{\|\boldsymbol{b}_{:j}\| \, \|\boldsymbol{c}_{j:}\|}{\sum_{k=1}^{N} \|\boldsymbol{b}_{:k}\| \, \|\boldsymbol{c}_{k:}\|} \quad \text{for } j = 1, 2, \dots, N.$$

Now, we introduce a random matrix $\boldsymbol{R}$ with distribution

$$\boldsymbol{R} = \frac{1}{p_j} \boldsymbol{b}_{:j} \boldsymbol{c}_{j:} \quad \text{with probability } p_j \text{ for each } j = 1, \dots, N.$$

It follows that

$$\mathbb{E}\,\boldsymbol{R} = \sum_{j=1}^{N} \boldsymbol{b}_{:j} \boldsymbol{c}_{j:} = \boldsymbol{BC}.$$

Therefore, we can regard $\boldsymbol{R}$ as a randomized proxy for the product $\boldsymbol{BC}$. This estimator is unbiased, but the variance may be intolerable. To obtain a more precise estimate for the product, we can average n independent copies of $\boldsymbol{R}$:

$$\bar{\boldsymbol{R}}_n = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{R}_k$$

We must assess how large n must be for $\bar{\boldsymbol{R}}_n$ to achieve a reasonable error, and we must bound the computational cost of the resulting estimator.

It is helpful to frame the results in terms of the *stable rank* of the matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ that appear in the product.

Definition 6.4.1 (Stable Rank)**.** *The stable rank of a matrix $\boldsymbol{F}$ is defined as*

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

The Frobenius norm is defined by the relation $\|\boldsymbol{F}\|_{\mathrm{F}}^2 = \operatorname{tr}(\boldsymbol{F}\boldsymbol{F}^*)$.

The stable rank is a lower bound for the algebraic rank: $1 \le \operatorname{srank}(\boldsymbol{F}) \le \operatorname{rank}(\boldsymbol{F})$. Check these inequalities by expressing the two norms in terms of the singular values of $\boldsymbol{F}$. In contrast with the algebraic rank, the stable rank is a continuous function of the matrix, so it is more suitable for numerical applications.

We are prepared to present the main claim about randomized matrix multiplication. Fix a parameter $\varepsilon \in (0, 1]$. Suppose that the number n of samples satisfies

$$n \ge \frac{5 \cdot \operatorname{srank}(\boldsymbol{A}) \cdot \operatorname{srank}(\boldsymbol{B}) \cdot \log(d_1 + d_2)}{\varepsilon^2} \tag{6.4.3}$$

Then the randomized estimate $\bar{\boldsymbol{R}}_n$ for the product achieves a relative error of ε in the spectral norm:

$$\mathbb{E}\left\|\bar{\boldsymbol{R}}_n - \boldsymbol{BC}\right\| \le \varepsilon \|\boldsymbol{B}\| \|\boldsymbol{C}\| \tag{6.4.4}$$

To compute $\bar{\boldsymbol{R}}_n$, we need $\mathrm{O}(nd_1d_2)$ arithmetic operations. Therefore, the estimator is efficient when the number n of samples is much smaller than N, the inner dimension of the product $\boldsymbol{BC}$.

This result is natural because a matrix with low stable rank contains a lot of redundant information. As a consequence, we do not need to multiply each column of $\boldsymbol{B}$ with each row of $\boldsymbol{C}$ to get a good estimate for the product. In particular, when the outer dimensions d_1 and d_2 are much smaller than the inner dimension N, many of the terms in (6.4.2) can be omitted without a significant loss.

Remark 6.4.2. Since our goal is to illustrate the analysis of a random matrix, the algorithmic details are not especially important. Nevertheless, we should point out that the method we have described is not the most effective way to perform randomized matrix multiplication. It is better to apply a preprocessing step to ensure that the columns of $\boldsymbol{B}$ have comparable norms and that the rows of $\boldsymbol{C}$ have comparable norms. In this case, it is possible to obtain a somewhat better bound on the number of samples required.

The Analysis

To study the behavior of randomized matrix multiplication, we introduce the error matrix

$$\boldsymbol{E} = \bar{\boldsymbol{R}}_n - \boldsymbol{BC} = \frac{1}{n}\sum_{k=1}^{n} (\boldsymbol{R}_k - \boldsymbol{BC}) = \sum_{k=1}^{n} \boldsymbol{S}_k.$$

The random matrices $\boldsymbol{S}_k$ are defined by the previous expression. Observe that the summands are independent, and each has zero mean. Therefore, we can apply the matrix Bernstein inequality to study the expected norm of the error.

First, let us bound the norm of a generic summand $\boldsymbol{S} = n^{-1}(\boldsymbol{R} - \boldsymbol{BC})$. Note that

$$\|\boldsymbol{R}\| \le \max_j \frac{1}{p_j} \left\| \boldsymbol{b}_{:j} \boldsymbol{c}_{j:} \right\| = \sum_{k=1}^{N} \|\boldsymbol{b}_{:k}\| \, \|\boldsymbol{c}_{k:}\| \le \|\boldsymbol{B}\|_{\mathrm{F}} \, \|\boldsymbol{C}\|_{\mathrm{F}}$$

The last inequality is Cauchy–Schwarz. Therefore, we have the uniform bound

$$\|\boldsymbol{S}\| = \frac{1}{n} \|\boldsymbol{R} - \boldsymbol{BC}\| \le \frac{1}{n} \left(\|\boldsymbol{R}\| + \|\boldsymbol{B}\| \, \|\boldsymbol{C}\| \right) \le \frac{2}{n} \|\boldsymbol{B}\|_{\mathrm{F}} \, \|\boldsymbol{C}\|_{\mathrm{F}}.$$

Observe that the bound decreases with the number n of samples.

Next, we compute the variance of $\boldsymbol{E}$. This takes some effort. First, consider a generic summand $\boldsymbol{S}$. Form the expectation

$$\mathbb{E}\left[\boldsymbol{S}\boldsymbol{S}^*\right] = \frac{1}{n^2} \mathbb{E}\left[(\boldsymbol{R} - \boldsymbol{BC})(\boldsymbol{R} - \boldsymbol{BC})^* \right] = \frac{1}{n^2} \left[\mathbb{E}\left(\boldsymbol{R}\boldsymbol{R}^*\right) - \boldsymbol{BCC}^*\boldsymbol{B}^* \right].$$

Let us focus on the first term on the right-hand side.

$$\left\| \mathbb{E}\left(\boldsymbol{R}\boldsymbol{R}^*\right) \right\| = \left\| \sum_{j=1}^{N} \frac{1}{p_j} \left(\boldsymbol{b}_{:j}\boldsymbol{c}_{j:}\right)\left(\boldsymbol{b}_{:j}\boldsymbol{c}_{j:}\right)^* \right\| \le \sum_{j=1}^{N} \frac{1}{p_j} \left\| \boldsymbol{b}_{:j} \right\|^2 \left\| \boldsymbol{c}_{j:} \right\|^2 = \left(\sum_{j=1}^{N} \left\| \boldsymbol{b}_{:j} \right\| \left\| \boldsymbol{c}_{j:} \right\| \right)^2 \le \|\boldsymbol{B}\|_{\mathrm{F}}^2 \, \|\boldsymbol{C}\|_{\mathrm{F}}^2.$$

In combination, the last two displays yield

$$\left\| \mathbb{E}\left(\boldsymbol{S}\boldsymbol{S}^*\right) \right\| \le \frac{1}{n^2} \left[\|\boldsymbol{B}\|_{\mathrm{F}}^2 \, \|\boldsymbol{C}\|_{\mathrm{F}}^2 + \left\| \boldsymbol{BCC}^*\boldsymbol{B} \right\| \right] \le \frac{2}{n^2} \|\boldsymbol{B}\|_{\mathrm{F}}^2 \, \|\boldsymbol{C}\|_{\mathrm{F}}^2.$$

To obtain the variance of the error matrix $\boldsymbol{E}$, we calculate that

$$\left\| \mathbb{E}\left(\boldsymbol{E}\boldsymbol{E}^*\right) \right\| = \left\| \sum_{j,k=1}^{n} \mathbb{E}\left(\boldsymbol{S}_j \boldsymbol{S}_k^*\right) \right\| = \left\| \sum_{k=1}^{n} \mathbb{E}\left(\boldsymbol{S}_k \boldsymbol{S}_k^*\right) \right\| \le \frac{2}{n} \|\boldsymbol{B}\|_{\mathrm{F}}^2 \, \|\boldsymbol{C}\|_{\mathrm{F}}^2.$$

The second identity holds because the summands are independent and zero mean. The last bound follows from the triangle inequality and the calculation for a generic summand. The second component of the variance does not require any additional ideas, and we reach the bound

$$\sigma^2(\boldsymbol{E}) = \max\left\{ \left\| \mathbb{E}\left(\boldsymbol{E}\boldsymbol{E}^*\right) \right\|, \, \left\| \mathbb{E}\left(\boldsymbol{E}^*\boldsymbol{E}\right) \right\| \right\} \le \frac{2}{n} \|\boldsymbol{B}\|_{\mathrm{F}}^2 \, \|\boldsymbol{C}\|_{\mathrm{F}}^2.$$

Observe that we retain the favorable dependence on the number n of samples.

We have acquired what we need to apply the matrix Bernstein inequality. Invoke the expectation bound (6.2.2) to reach

$$\mathbb{E}\,\|\boldsymbol{E}\| \le \sqrt{\frac{4\log(d_1+d_2)}{n}}\,\|\boldsymbol{B}\|_{\mathrm{F}}\,\|\boldsymbol{C}\|_{\mathrm{F}} + \frac{2\log(d_1+d_2)}{3n}\,\|\boldsymbol{B}\|_{\mathrm{F}}\,\|\boldsymbol{C}\|_{\mathrm{F}}.$$

With our choice of n from (6.4.3), we conclude that

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{E}\| &\le \frac{4\varepsilon}{5}\,\|\boldsymbol{B}\|\,\|\boldsymbol{C}\| + \frac{2\varepsilon^2}{15}\,\frac{\|\boldsymbol{B}\|^2}{\|\boldsymbol{B}\|_{\mathrm{F}}}\,\frac{\|\boldsymbol{C}\|^2}{\|\boldsymbol{C}\|_{\mathrm{F}}}\\ &< \varepsilon\,\|\boldsymbol{B}\|\,\|\boldsymbol{C}\|.\end{aligned}$$

The last bound holds because the Frobenius norm dominates the spectral norm. This is the result (6.4.4).

6.5 Proof of the Matrix Bernstein Inequalities

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 eigenvalues satisfy 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 6.1.1.

Lemma 6.5.1 (Matrix Bernstein: Mgf and Cgf Bound)**.** *Suppose that $\boldsymbol{X}$ is a random Hermitian matrix that satisfies*

$$\mathbb{E}\,\boldsymbol{X} = \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}) \le R.$$

Then, for $0 < \theta < 3/R$,

$$\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \exp\left(\frac{\theta^2/2}{1-R\theta/3}\,\mathbb{E}\left(\boldsymbol{X}^2\right)\right) \quad \text{and} \quad \log \mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \frac{\theta^2/2}{1-R\theta/3}\,\mathbb{E}\left(\boldsymbol{X}^2\right).$$

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

$$\mathrm{e}^{\theta \boldsymbol{X}} = \mathbf{I} + \theta \boldsymbol{X} + (\mathrm{e}^{\theta \boldsymbol{X}} - \theta \boldsymbol{X} - \mathbf{I}) = \mathbf{I} + \theta \boldsymbol{X} + \boldsymbol{X} \cdot f(\boldsymbol{X}) \cdot \boldsymbol{X}, \tag{6.5.1}$$

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, $f(x) \le f(R)$ when $x \le R$. By assumption, the eigenvalues of $\boldsymbol{X}$ do not exceed R, so the Transfer Rule (2.1.6) implies that

$$f(\boldsymbol{X}) \preccurlyeq f(R) \cdot \mathbf{I}. \tag{6.5.2}$$

The Conjugation Rule (2.1.4) allows us to introduce the relation (6.5.2) into our expansion (6.5.1) of the matrix exponential:

$$\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \mathbf{I} + \theta \boldsymbol{X} + \boldsymbol{X}(f(R) \cdot \mathbf{I})\boldsymbol{X} = \mathbf{I} + \theta \boldsymbol{X} + f(R) \cdot \boldsymbol{X}^2.$$

Take the expectation of this semidefinite bound to reach

$$\mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \mathbf{I} + f(R) \cdot \mathbb{E}(\boldsymbol{X}^2). \tag{6.5.3}$$

The expression (6.5.3) provides a powerful bound for the matrix mgf. In fact, this result leads to the matrix Bennett inequality, which strengthens Theorem 6.1.1. We have chosen to present the weaker result because it is easier to apply in practice. To arrive at the mgf bound required for Theorem 6.1.1, we must keep working.

We need an inequality for the quantity $f(R)$. This argument involves a clever application of Taylor series:

$$f(R) = \frac{\mathrm{e}^{R\theta} - R\theta - 1}{R^2} = \frac{1}{R^2} \sum_{q=2}^{\infty} \frac{(R\theta)^q}{q!} \le \frac{\theta^2}{2} \sum_{q=2}^{\infty} \frac{(R\theta)^{q-2}}{3^{q-2}} = \frac{\theta^2/2}{1 - R\theta/3}. \tag{6.5.4}$$

The second expression is simply the Taylor expansion of the fraction, viewed as a function of θ. We obtain the inequality by factoring out $(R\theta)^2/2$ 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.

Introduce the inequality (6.5.4) for $f(R)$ into the semidefinite bound (6.5.3) for the matrix mgf to reach

$$\mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \mathbf{I} + \frac{\theta^2/2}{1 - R\theta/3} \mathbb{E}(\boldsymbol{X}^2) \preccurlyeq \exp\left(\frac{\theta^2/2}{1 - R\theta/3} \right).$$

The second semidefinite relation follows when we apply the Transfer Rule (2.1.6) 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 (2.1.9) that the logarithm is operator monotone. □

We are prepared to establish the matrix Bernstein inequalities for random Hermitian matrices.

Proof of Theorem 6.1.1. Consider a finite sequence $\{\boldsymbol{X}_k\}$ of random Hermitian matrices with dimension d. Assume that

$$\mathbb{E}\, \boldsymbol{X}_k = \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le R.$$

The matrix Bernstein cgf bound, Lemma 6.5.1, provides that

$$\log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq g(\theta) \cdot \mathbb{E}\left(\boldsymbol{X}_k^2\right) \quad \text{where} \quad g(\theta) = \frac{\theta^2/2}{1 - R\theta/3} \quad \text{for } 0 < \theta < 3/R. \tag{6.5.5}$$

Define the sum $\boldsymbol{Y} = \sum_k \boldsymbol{X}_k$, which it is our task to analyze.

We begin with the bound (6.1.2) 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

$$\begin{aligned} \mathbb{E}\, \lambda_{\max}(\boldsymbol{Y}) &\le \inf_{\theta > 0} \frac{1}{\theta} \log \operatorname{tr} \exp\left(\textstyle\sum_k \log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}_k} \right) \\ &\le \inf_{0 < \theta < 3/R} \frac{1}{\theta} \log \operatorname{tr} \exp\left(g(\theta) \textstyle\sum_k \mathbb{E}\left(\boldsymbol{X}_k^2\right) \right) \\ &= \inf_{0 < \theta < 3/R} \frac{1}{\theta} \log \operatorname{tr} \exp\left(g(\theta) \cdot \mathbb{E}\left(\boldsymbol{Y}^2\right) \right). \end{aligned}$$

As usual, to move from the first to the second line, we invoke the fact (2.1.7) that the trace exponential is monotone to introduce the semidefinite bound (6.5.5) for the cgf. The rest of the argument glides along a well-oiled track:

$$\begin{aligned}
\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) &\le \inf_{0<\theta<3/R} \frac{1}{\theta} \log\left[d\,\lambda_{\max}\left(\exp\left(g(\theta)\cdot\mathbb{E}\left(\boldsymbol{Y}^2\right)\right)\right)\right] \\
&= \inf_{0<\theta<3/R} \frac{1}{\theta} \log\left[d\,\exp\left(g(\theta)\cdot\lambda_{\max}\left(\mathbb{E}\left(\boldsymbol{Y}^2\right)\right)\right)\right] \\
&= \inf_{0<\theta<3/R} \frac{1}{\theta} \log\left[d\,\exp\left(g(\theta)\cdot\sigma^2\right)\right] \\
&= \inf_{0<\theta<3/R} \left[\frac{\log d}{\theta} + \frac{\theta/2}{1-R\theta/3}\cdot\sigma^2\right].
\end{aligned}$$

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 variance parameter (6.1.1). Afterward, we extract the logarithm and simplify. Finally, we minimize the expression—ideally with a computer algebra system—to complete the proof of (6.1.2).

Next, we develop the tail bound (6.1.3) for $\lambda_{\max}(\boldsymbol{Y})$. Owing to the master tail inequality (3.6.3), we have

$$\begin{aligned}
\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \ge t\right\} &\le \inf_{\theta>0} \mathrm{e}^{-\theta t}\,\mathrm{tr}\exp\left(\textstyle\sum_k \log \mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}_k}\right) \\
&\le \inf_{0<\theta<3/R} \mathrm{e}^{-\theta t}\,\mathrm{tr}\exp\left(g(\theta)\textstyle\sum_k \mathbb{E}\left(\boldsymbol{X}_k^2\right)\right) \\
&= \inf_{0<\theta<3/R} d\,\mathrm{e}^{-\theta t}\exp\left(g(\theta)\cdot\sigma^2\right).
\end{aligned}$$

The justifications are the same as before. The exact value of the infimum is messy, so we proceed with the inspired choice $\theta = t/(\sigma^2 + Rt/3)$, which results in the elegant bound (6.1.3). □

Finally, we explain how to derive Corollary 6.2.1, for general matrices, from Theorem 6.1.1. 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 Corollary 6.2.1. Consider a finite sequence $\{\boldsymbol{S}_k\}$ of $d_1 \times d_2$ random matrices, and assume that

$$\mathbb{E}\,\boldsymbol{S}_k = \mathbf{0} \quad\text{and}\quad \lambda_{\max}(\boldsymbol{S}_k) \le R$$

We define the two random matrices

$$\boldsymbol{Z} = \textstyle\sum_k \boldsymbol{S}_k \quad\text{and}\quad \boldsymbol{Y} = \mathscr{H}(\boldsymbol{Z}).$$

where $\mathscr{H}$ is the Hermitian dilation (2.1.11). We will invoke Theorem 6.1.1 to analyze $\|\boldsymbol{Z}\|$. First, recall the fact (2.1.13) that

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

Next, we express the variance (6.1.1) of the random Hermitian matrix $\boldsymbol{Y}$ in terms of the general matrix $\boldsymbol{Z}$. Indeed,

$$\sigma^2(\boldsymbol{Y}) = \left\|\mathbb{E}(\boldsymbol{Y}^2)\right\| = \left\|\mathbb{E}\left(\mathscr{H}(\boldsymbol{Z})^2\right)\right\| = \left\|\mathbb{E}\begin{bmatrix}\boldsymbol{Z}\boldsymbol{Z}^* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}^*\boldsymbol{Z}\end{bmatrix}\right\|$$

$$= \left\| \begin{bmatrix} \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*) & \mathbf{0} \\ \mathbf{0} & \mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z}) \end{bmatrix} \right\| = \max\left\{ \left\| \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*) \right\|, \left\| \mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z}) \right\| \right\} = \sigma^2(\boldsymbol{Z}).$$

The third relation is the identity (2.1.12) for the square of the Hermitian dilation. The penultimate equation holds because the norm of a block-diagonal matrix is the maximum norm of any diagonal block. We obtain the formula (4.2.1) for the variance of the matrix $\boldsymbol{Z}$. Finally, we invoke Theorem 6.1.1 to establish Corollary 6.2.1. □

6.6 Notes

There are a wide variety of Bernstein-type inequalities available 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.

6.6.1 Matrix Bernstein Inequalities

David Gross [Gro09] and Ben Recht [Rec11] used the approach of Ahlswede–Winter [AW02] to develop two different versions of the matrix Bernstein inequality. These papers played a big role in popularizing the use matrix concentration inequalities in mathematical signal processing and statistics. Nevertheless, their results involve a suboptimal variance parameter of the form

$$\sigma_{\mathrm{AW}}^2 = \sum\nolimits_k \left\| \mathbb{E}(\boldsymbol{X}_k^2) \right\|.$$

In general, this parameter is significantly larger than the variance (6.1.1) that appears in Theorem 6.1.1. 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 variance parameter (6.1.1). 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 6.1.1, up to the precise value of the constants.

The matrix Bernstein inequality we have stated here, Theorem 6.1.1, first appeared in the paper [Tro11d, §6]. The bounds for the expectation are new. The argument is based on Lieb's Theorem, and it also delivers a matrix Bennett inequality, and the split Bernstein inequality (6.1.4) discussed here. 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 [Tro11d] is independent from Oliveira's ideas [Oli10a]. Motivated by Oliveira's paper, the article [Tro11a] and the technical report [Tro11c] show how to use Lieb's Theorem to study matrix martingales. The subsequent paper [GT11] explains how to develop 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].

6.6.2 Randomized Matrix Multiplication

The idea of using random sampling to accelerate matrix multiplication appears in a paper by Drineas, Kannan, and Mahoney [DKM06]. Subsequently, Tamás Sarlós obtained a significant

improvement in the performance of this algorithm [Sar06]. The analysis we have given here is a corrected version of the argument in the work of Hsu, Kakade, and Zhang [HKZ12b]; see also [HKZ12a]. A related analysis appears in the paper of Magen and Zouzias [MZ11].

6.6.3 Randomized Sparsification

The idea of using randomized sparsification to accelerate spectral computations appears in a paper of Achlioptas and McSherry [AM07]. Drineas and Zouzias [DZ11] point out that matrix concentration inequalities can be used to analyze this type of algorithm. For further results on sparsification, see the paper [GT].

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 is sometimes 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.

We present a version of the matrix Chernoff inequality for an independent sum of bounded, positive-semidefinite random matrices that involves an intrinsic dimension parameter. This result is interesting, but it is not entirely satisfactory because it lacks a bound for the minimum eigenvalue. We also describe a version of the matrix Bernstein inequality for an independent sum of bounded, zero-mean random matrices that involves an intrinsic dimension parameter. The intrinsic Bernstein result often improves on Theorem 6.1.1. 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 reconsider some of the examples from earlier chapters. We apply the intrinsic Chernoff bound to study a random column submatrix of a fixed matrix, and we use the intrinsic Bernstein bound to analyze the sample covariance estimator. In each case, the intrinsic dimension parameters have an attractive interpretation in terms of the problem data.

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 bounds and their 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 allow 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*

$$\operatorname{intdim}(\boldsymbol{A}) = \frac{\operatorname{tr} \boldsymbol{A}}{\|\boldsymbol{A}\|}.$$

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 lower inequality is attained precisely when $\boldsymbol{A}$ has rank one, while the upper inequality is attained precisely when $\boldsymbol{A}$ is a multiple of the identity. Note that the intrinsic dimension is 0-homogeneous, so it is insensitive to changes in the scale of the matrix $\boldsymbol{A}$. We interpret the intrinsic dimension as a reflection of the number of dimensions where $\boldsymbol{A}$ has significant spectral content.

7.2 Matrix Chernoff with Intrinsic Dimension

Let us begin with an extension of the matrix Chernoff inequality. We obtain bounds for the maximum eigenvalue of a sum of bounded, positive-semidefinite matrices that depend on 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 that satisfy*

$$\boldsymbol{X}_k \succcurlyeq \boldsymbol{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le R.$$

Define the random matrix

$$\boldsymbol{Y} = \textstyle\sum_k \boldsymbol{X}_k.$$

Introduce an intrinsic dimension parameter and a mean parameter:

$$d = d(\boldsymbol{Y}) = \operatorname{intdim}(\mathbb{E}\,\boldsymbol{Y}) \quad \text{and} \quad \mu_{\max} = \mu_{\max}(\boldsymbol{Y}) = \lambda_{\max}(\mathbb{E}\,\boldsymbol{Y}).$$

Then, for $\theta > 0$,

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \le \frac{\mathrm{e}^{\theta} - 1}{\theta} \cdot \mu_{\max} + \frac{1}{\theta} \cdot R \log(2d). \tag{7.2.1}$$

Furthermore,

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \ge (1+\delta)\mu_{\max}\right\} \le 2d \cdot \left[\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{\max}/R} \quad \text{for } \delta \ge 1. \tag{7.2.2}$$

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 mean $\mathbb{E}\,\boldsymbol{Y}$ instead of the ambient dimension. When the eigenvalues of $\mathbb{E}\,\boldsymbol{Y}$ 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 δ. Neither of these limitations is particularly significant.

A more serious flaw in Theorem 7.2.1 is that it does not provide any information about the *minimum* eigenvalue $\lambda_{\min}(\boldsymbol{Y})$. 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, we apply it to bound the expected norm of a random column submatrix drawn from a fixed matrix, a problem we considered in §5.2.

In this example, we began with a fixed $d \times n$ matrix $\boldsymbol{B}$, and we formed a random submatrix $\boldsymbol{Z}$ containing an average of q nonzero columns from $\boldsymbol{B}$. In the analysis, we applied the matrix Chernoff inequality to the random matrix $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{Z}^*$, which takes the form

$$\boldsymbol{Y} = \sum_{k=1}^{n} \eta_k \boldsymbol{b}_{k:} \boldsymbol{b}_{k:}^*.$$

Here, $\{\eta_k\}$ is an independent family of Bernoulli random variables with common mean q/n. We have written $\boldsymbol{b}_{k:}$ for the kth column of $\boldsymbol{B}$.

To invoke Theorem 7.2.1, we just need to compute the intrinsic dimension $d(\boldsymbol{Y}) = \operatorname{intdim}(\mathbb{E}\,\boldsymbol{Y})$. Recall that $\mathbb{E}\,\boldsymbol{Y} = (q/n)\,\boldsymbol{B}\boldsymbol{B}^*$, so that

$$d(\boldsymbol{Y}) = \operatorname{intdim}\left(\frac{q}{n}\boldsymbol{B}\boldsymbol{B}^*\right) = \operatorname{intdim}(\boldsymbol{B}\boldsymbol{B}^*) = \frac{\operatorname{tr}(\boldsymbol{B}\boldsymbol{B}^*)}{\|\boldsymbol{B}\boldsymbol{B}^*\|} = \frac{\|\boldsymbol{B}\|_{\mathrm{F}}^2}{\|\boldsymbol{B}\|^2} = \operatorname{srank}(\boldsymbol{B}).$$

The second identity holds because the intrinsic dimension is scale invariant. The last relation is simply Definition 6.4.1. Therefore, the expectation bound (7.2.1) with $\theta = 1$ delivers

$$\mathbb{E}\left(\|\boldsymbol{Z}\|^2\right) = \mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \le (\mathrm{e}-1)\cdot\mu_{\max}(\boldsymbol{Y}) + R\log(2\cdot\operatorname{srank}(\boldsymbol{B})).$$

In contrast, our previous analysis led to a logarithmic factor of $\log d$. If the matrix $\boldsymbol{B}$ has deficient stable rank—meaning that it has many rows which are almost collinear—then the new bound can result in a serious improvement.

7.3 Matrix Bernstein with Intrinsic Dimension

We continue with extensions 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.

7.3.1 The Hermitian Case

We begin with the results for an independent sum of Hermitian random matrices whose eigenvalues are bounded above.

Theorem 7.3.1 (Matrix Bernstein: Hermitian Case with Intrinsic Dimension)**.** *Consider a finite sequence* $\{\boldsymbol{X}_k\}$ *of random Hermitian matrices that satisfy*

$$\mathbb{E}\,\boldsymbol{X}_k = \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le R.$$

Define the random matrix

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

Introduce the intrinsic dimension and variance parameters

$$d = d(\boldsymbol{Y}) = \operatorname{intdim}\left(\mathbb{E}(\boldsymbol{Y}^2)\right) \quad \text{and} \quad \sigma^2 = \sigma^2(\boldsymbol{Y}) = \left\|\mathbb{E}(\boldsymbol{Y}^2)\right\|.$$

Then, for $t \ge \sigma + R/3$,

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} \le 4d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right). \tag{7.3.1}$$

The proof of this result appears below in §7.7.

Discussion

Theorem 7.3.1 is quite similar to Theorem 6.1.1, so we focus on the differences. Note that the tail bound (7.3.1) now depends on the intrinsic dimension of the variance matrix $\mathbb{E}(\boldsymbol{Y}^2)$, which is never larger than the ambient dimension. As a consequence, the tail bound is almost always sharper than the earlier result. The costs of this improvement are small: We pay an extra factor of four, and we must restrict our attention to a more limited range of the parameter t. Neither of these changes is significant.

We can obtain a bound for $\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y})$ by integrating the tail inequality (7.3.1), which gives

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \le \text{Const} \cdot \left(\sigma\sqrt{\log d} + R\log d\right).$$

It seems likely that we could adapt the argument to obtain a more direct proof of the expectation bound, along with an explicit constant.

The other commentary about the original matrix Bernstein inequality, Theorem 6.1.1, also applies to the intrinsic dimension result. Using similar arguments, we can obtain bounds for $\lambda_{\min}(\boldsymbol{Y})$, and we can adapt the result to an independent sum of uncentered, bounded, random Hermitian matrices. The modifications required in these cases are straightforward.

Finally, let us mention a subtle but important point concerning the application of Theorem 7.3.1. It is often difficult or unwieldy to compute the exact values of the parameters $d(\boldsymbol{Y})$ and $\sigma^2(\boldsymbol{Y})$. In this case, we can proceed as follows. Suppose that $\mathbb{E}(\boldsymbol{Y}^2) \preccurlyeq \boldsymbol{V}$ for some positive-semidefinite matrix $\boldsymbol{V}$. A slight modification to the proof of Theorem 6.1.1 yields the tail bound

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} \le 4 \cdot \operatorname{intdim}(\boldsymbol{V}) \cdot \exp\left(\frac{-t^2/2}{\|\boldsymbol{V}\| + Rt/3}\right) \tag{7.3.2}$$

for all $t \ge \|\boldsymbol{V}\|^{1/2} + R/3$. This version of the result is often much easier to apply.

7.3.2 The General Case

Next, we present the adaptation for an independent sum of general random matrices that are bounded in spectral norm.

Corollary 7.3.2 (Matrix Bernstein: Rectangular Case with Intrinsic Dimension)**.** *Consider a finite sequence* $\{\boldsymbol{S}_k\}$ *of random complex matrices that satisfy*

$$\mathbb{E}\,\boldsymbol{S}_k = \mathbf{0} \quad \text{and} \quad \|\boldsymbol{S}_k\| \le R.$$

Define the random matrix

$$\boldsymbol{Z} = \sum\nolimits_k \boldsymbol{S}_k.$$

Introduce the intrinsic dimension parameter

$$d = d(\boldsymbol{Z}) = \operatorname{intdim}\begin{bmatrix} \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*) & \mathbf{0} \\ \mathbf{0} & \mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z}) \end{bmatrix}. \tag{7.3.3}$$

and the variance parameter

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

Then, for $t \ge \sigma + R/3$,

$$\mathbb{P}\{\|\boldsymbol{Z}\| \ge t\} \le 4d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right). \tag{7.3.4}$$

The proof of this result appears below in §7.7.

Discussion

Corollary 7.3.2 is very similar to Theorem 7.3.1 and our earlier result, Corollary 6.2.1. As a consequence, we limit our discussion to a single point. Note that the intrinsic dimension parameter (7.3.3) is computed from a block-diagonal matrix that contains both of the squares of the matrix $\boldsymbol{Z}$. It follows that

$$d(\boldsymbol{Z}) = \frac{\mathbb{E}\operatorname{tr}(\boldsymbol{Z}\boldsymbol{Z}^*) + \mathbb{E}\operatorname{tr}(\boldsymbol{Z}^*\boldsymbol{Z})}{\max\{\|\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*)\|, \|\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})\|\}}.$$

In other words, we divide by the norm of the larger block. We can make a further bound to obtain a result in terms of the intrinsic dimensions of the two blocks:

$$d(\boldsymbol{Z}) \le \operatorname{intdim}\left(\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*)\right) + \operatorname{intdim}\left(\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})\right).$$

An interesting consequence is that the intrinsic dimension $d(\boldsymbol{Z})$ can be much smaller than the intrinsic dimension of either $\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*)$ or $\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})$.

7.3.3 Example: Sample Covariance Matrices, Redux

To demonstrate the value of the intrinsic dimension results, let us apply Theorem 7.3.1 to the sample covariance matrix example we analyzed in §1.6.3.

Consider a random vector $\boldsymbol{x}$ with zero mean, covariance $\boldsymbol{A}$, and uniform upper bound $\|\boldsymbol{x}\|^2 \le B$. The sample covariance matrix $\boldsymbol{Y} = n^{-1}\sum_{k=1}^n \boldsymbol{x}_k\boldsymbol{x}_k^*$, where $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are independent samples

from the distribution $\boldsymbol{x}$. Recall that the random matrix of interest is $\boldsymbol{E} = \boldsymbol{Y} - \boldsymbol{A}$, the discrepancy between the sample covariance matrix and the true covariance.

We expressed the error matrix $\boldsymbol{E}$ as the sum of the independent random matrices

$$\boldsymbol{S}_k = \frac{1}{n}(\boldsymbol{x}_k \boldsymbol{x}_k^* - \boldsymbol{A}).$$

The summands have the properties that $\mathbb{E}\,\boldsymbol{S}_k = \boldsymbol{0}$ and $\|\boldsymbol{S}_k\| \le 2B/n$. Moreover, $\mathbb{E}(\boldsymbol{S}_k^2) \preccurlyeq (B/n^2)\cdot\boldsymbol{A}$, so that

$$\mathbb{E}(\boldsymbol{E}^2) \preccurlyeq \frac{B}{n}\cdot\boldsymbol{A}$$

As discussed, we may substitute the semidefinite upper bound $\boldsymbol{V} = (B/n)\cdot\boldsymbol{A}$ for $\mathbb{E}(\boldsymbol{E}^2)$ when we compute the variance parameter and the intrinsic dimension parameter in Theorem 7.3.1.

Let us introduce the intrinsic dimension and variance parameters

$$\operatorname{intdim}(\boldsymbol{V}) = \frac{\operatorname{tr}\boldsymbol{A}}{\|\boldsymbol{A}\|} \quad\text{and}\quad \|\boldsymbol{V}\| = \frac{B}{n}\|\boldsymbol{A}\|.$$

We can apply the modified tail bound (7.3.2) to both $\boldsymbol{E}$ and $-\boldsymbol{E}$ to control $\lambda_{\max}(\boldsymbol{E})$ and $\lambda_{\min}(\boldsymbol{E})$. Combine these two results with the union bound to reach the spectral norm estimate

$$\mathbb{P}\{\|\boldsymbol{Y}-\boldsymbol{E}\| \ge t\} \le \frac{8\operatorname{tr}\boldsymbol{A}}{\|\boldsymbol{A}\|}\cdot\exp\left(\frac{-t^2/2}{B\|\boldsymbol{A}\|/n + 2Bt/3n}\right),$$

valid when t is sufficiently large. To achieve a relative error $\varepsilon \in (0,1]$, the number n of samples should satisfy

$$n \ge \text{Const}\cdot\frac{B\log(\operatorname{intdim}(\boldsymbol{A}))}{\varepsilon^2\|\boldsymbol{A}\|}. \tag{7.3.5}$$

In this case, we obtain a tail bound of the form

$$\mathbb{P}\{\|\boldsymbol{Y}-\boldsymbol{E}\| \ge \varepsilon\|\boldsymbol{A}\|\} \le \text{Const}\cdot\operatorname{intdim}(\boldsymbol{A})^{-\text{Const}}.$$

By increasing the number n of samples, we can increase the exponent in the tail probability.

The key observation is that the intrinsic dimension term in (7.3.5) may be much smaller than the ambient dimension of the covariance matrix $\boldsymbol{A}$. For instance, if the ordered eigenvalues of $\boldsymbol{A}$ satisfy the bounds

$$\lambda_j(\boldsymbol{A}) \le \frac{1}{j^2} \quad\text{for each } j = 1,2,3,\dots,$$

then the logarithmic factor in (7.3.5) reduces to a *constant* that is independent of the dimension of the covariance matrix $\boldsymbol{A}$!

Finally, we note that this result has an attractive interpretation: The intrinsic dimension parameter $\operatorname{intdim}(\boldsymbol{A})$ is the total variance of all the components of the random vector $\boldsymbol{x}$ divided by the maximum variance achieved by any component of $\boldsymbol{x}$.

7.4 Revisiting the Matrix Laplace Transform Bound

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 functions to transform the tail event. The functions we have in mind are adjusted versions of the exponential function. 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 $\psi(a) \geq \psi(t)$. It follows that

$$\lambda_{\max}(\boldsymbol{Y}) \geq t \quad \Longrightarrow \quad \lambda_{\max}(\psi(\boldsymbol{Y})) \geq \psi(t).$$

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 $\psi(\lambda_{\max}(\boldsymbol{Y}))$ is among the eigenvalues of $\psi(\boldsymbol{Y})$, and 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 (2.1.5) that the trace of $\psi(\boldsymbol{Y})$, a positive-semidefinite matrix, must be 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 $[0,\infty)$ with $\varphi(0) = 0$. 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 $a \mapsto \varphi(a)$ is convex on the interval $[0, R]$, it is bounded above by the chord connecting the endpoints. That is, for $a \in [0, R]$,

$$\varphi(a) \le \left(1 - \frac{a}{R}\right) \cdot \varphi(0) + \frac{a}{R} \cdot \varphi(R) = \frac{a}{R} \cdot \varphi(R).$$

The eigenvalues of $\boldsymbol{A}$ fall in the interval $[0, R]$, where $R = \|\boldsymbol{A}\|$. As an immediate consequence of the Transfer Rule (2.1.6), 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

$$\boldsymbol{X}_k \succcurlyeq \boldsymbol{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le R.$$

Introduce the sum

$$\boldsymbol{Y} = \textstyle\sum_k \boldsymbol{X}_k.$$

The challenge is to establish bounds for $\lambda_{\max}(\boldsymbol{Y})$ that depends on the intrinsic dimension of the matrix $\mathbb{E}\,\boldsymbol{Y}$. 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 $t \in \mathbb{R}$. The general version of the matrix Laplace transform bound, Proposition 7.4.1, states that

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} \le \frac{1}{\psi(t)}\, \mathbb{E} \operatorname{tr} \psi(\boldsymbol{Y}) = \frac{1}{\mathrm{e}^{\theta t} - 1}\, \mathbb{E} \operatorname{tr}\left(\mathrm{e}^{\theta \boldsymbol{Y}} - \mathbf{I}\right). \tag{7.6.1}$$

We have exploited the fact that $\boldsymbol{Y}$ is positive semidefinite and 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 our original matrix Chernoff bound, Theorem 5.1.1, we have the bound

$$\mathbb{E} \operatorname{tr} \mathrm{e}^{\theta \boldsymbol{Y}} \le \operatorname{tr} \exp\left(g(\theta)(\mathbb{E}\,\boldsymbol{Y})\right) \quad \text{where} \quad g(\theta) = \frac{\mathrm{e}^{R\theta} - 1}{R}.$$

Invoke the latter inequality, and introduce the function $\varphi(a) = \mathrm{e}^a - 1$ to see that

$$\mathbb{E} \operatorname{tr}\left(\mathrm{e}^{\theta \boldsymbol{Y}} - \mathbf{I}\right) \le \operatorname{tr} \varphi\left(g(\theta)(\mathbb{E}\,\boldsymbol{Y})\right) \le \operatorname{intdim}(\mathbb{E}\,\boldsymbol{Y}) \cdot \varphi\left(g(\theta)\,\|\mathbb{E}\,\boldsymbol{Y}\|\right).$$

The second inequality results from Lemma 7.5.1, the intrinsic dimension bound, and the fact that the intrinsic dimension does not depend on the scaling factor $g(\theta)$. Recalling the notation $d = \operatorname{intdim}(\mathbb{E}\,\boldsymbol{Y})$ and $\mu_{\max} = \|\mathbb{E}\,\boldsymbol{Y}\|$, we continue the calculation:

$$\mathbb{E} \operatorname{tr}\left(\mathrm{e}^{\theta \boldsymbol{Y}} - \mathbf{I}\right) \le d \cdot \varphi\left(g(\theta) \cdot \mu_{\max}\right) \le d \cdot \exp\left(g(\theta) \cdot \mu_{\max}\right). \tag{7.6.2}$$

We have used the trivial bound $\varphi(a) \le \mathrm{e}^a$, which holds for $a \in \mathbb{R}$.

To complete the argument, introduce the bound (7.6.2) on the expected trace into the probability bound (7.6.1) to obtain

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} \le d \cdot \frac{\mathrm{e}^{\theta t}}{\mathrm{e}^{\theta t} - 1} \cdot \mathrm{e}^{-\theta t + g(\theta)\cdot\mu_{\max}}.$$

It is convenient to make the change of variables $t \mapsto (1+\delta)\mu_{\max}$. The previous estimate is valid for all $\theta > 0$, so we can select $\theta = R^{-1}\log(1+\delta)$ to minimize the final exponential. To bound the fraction, observe 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 at $a = 0$.

Altogether, these steps lead to the estimate

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge (1+\delta)\mu_{\max}\} \le d \cdot \left(1 + \frac{R/\mu_{\max}}{(1+\delta)\log(1+\delta)}\right) \cdot \left[\frac{\mathrm{e}^\delta}{(1+\delta)^{1+\delta}}\right]^{\mu_{\max}/R}. \tag{7.6.3}$$

For random matrices, this inequality is rarely useful when $\delta < 1$, so it does little harm to place the restriction that $\delta \ge 1$. Subject to this condition, the bracket (including the exponent) exceeds one unless we also have

$$(1+\delta)\log(1+\delta) \ge \frac{R}{\mu_{\max}}.$$

Therefore, we can use the latter bound to make a numerical estimate for the parenthesis in (7.6.3), which leads to 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

$$\begin{aligned}\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) = \mathbb{E}\,\psi^{-1}(\psi(\lambda_{\max}(\boldsymbol{Y}))) &\le \psi^{-1}(\mathbb{E}\,\psi(\lambda_{\max}(\boldsymbol{Y}))) \\ &= \psi^{-1}(\mathbb{E}\,\lambda_{\max}(\psi(\boldsymbol{Y}))) \le \psi^{-1}(\mathbb{E}\,\operatorname{tr}\psi(\boldsymbol{Y})). \end{aligned} \tag{7.6.4}$$

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 (7.6.4) 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 owes to the fact that the exponent is nonnegative. To complete the argument, introduce the definition of $g(\theta)$, and make the change of variables $\theta \mapsto \theta/R$. 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

We commence with the results for an independent sum of random Hermitian matrices whose eigenvalues are subject to an upper bound.

Proof of Theorem 7.3.1. Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, Hermitian matrices with

$$\mathbb{E}\,\boldsymbol{X}_k = \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \le R.$$

Introduce the random matrix

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

It is our goal to obtain a tail bound for $\lambda_{\max}(\boldsymbol{Y})$ that reflects the intrinsic dimension of its variance $\mathbb{E}(\boldsymbol{Y}^2)$.

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

$$\begin{aligned} \mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} &\le \frac{1}{\psi(t)}\,\mathbb{E}\,\mathrm{tr}\,\psi(\boldsymbol{Y}) \\ &= \frac{1}{\psi(t)}\,\mathbb{E}\,\mathrm{tr}\big(\mathrm{e}^{\theta\boldsymbol{Y}} - \theta\boldsymbol{Y} - \mathbf{I}\big) \\ &= \frac{1}{\mathrm{e}^{\theta t} - \theta t - 1}\,\mathbb{E}\,\mathrm{tr}\big(\mathrm{e}^{\theta\boldsymbol{Y}} - \mathbf{I}\big). \end{aligned} \tag{7.7.1}$$

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 (7.7.1). Examining the proof of the original matrix Bernstein bound, Theorem 6.1.1, we recall that

$$\mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta\boldsymbol{Y}} \le \mathrm{tr}\exp\big(g(\theta)\cdot\mathbb{E}(\boldsymbol{Y}^2)\big) \quad \text{where} \quad g(\theta) = \exp\left(\frac{\theta^2/2}{1 - R\theta/3}\right).$$

Applying this inequality and introducing the function $\varphi(a) = \mathrm{e}^a - 1$, we obtain

$$\begin{aligned} \mathbb{E}\,\mathrm{tr}\big(\mathrm{e}^{\theta\boldsymbol{Y}} - \mathbf{I}\big) &\le \mathrm{tr}\big(\mathrm{e}^{g(\theta)\,\mathbb{E}(\boldsymbol{Y}^2)} - \mathbf{I}\big) \\ &= \mathrm{tr}\,\varphi\big(g(\theta)\,\mathbb{E}(\boldsymbol{Y}^2)\big) \\ &\le \mathrm{intdim}\big(\mathbb{E}(\boldsymbol{Y}^2)\big)\cdot\varphi\big(g(\theta)\,\big\|\mathbb{E}(\boldsymbol{Y}^2)\big\|\big) \end{aligned}$$

The last inequality depends on the intrinsic dimension result, Lemma 7.5.1, and the fact that the intrinsic dimension does not depend on the scaling factor $g(\theta)$. Identify the dimensional parameter $d = \mathrm{intdim}\big(\mathbb{E}(\boldsymbol{Y}^2)\big)$ and the variance parameter $\sigma^2 = \big\|\mathbb{E}(\boldsymbol{Y}^2)\big\|$. It follows that

$$\mathbb{E}\,\mathrm{tr}\big(\mathrm{e}^{\theta\boldsymbol{Y}} - \mathbf{I}\big) \le d\cdot\varphi\big(g(\theta)\cdot\sigma^2\big) \le d\cdot\exp\big(g(\theta)\cdot\sigma^2\big). \tag{7.7.2}$$

This bound depends on the obvious estimate $\varphi(a) \le \mathrm{e}^a$, valid for all $a \in \mathbb{R}$.

Substitute the bound (7.7.2) into the probability estimate (7.7.1) to reach

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \geq t\} \leq d \cdot \frac{\mathrm{e}^{\theta t}}{\mathrm{e}^{\theta t} - \theta t - 1} \cdot \mathrm{e}^{-\theta t + g(\theta)\cdot\sigma^2}.$$

This estimate holds for any positive value of θ. Choose $\theta = t/(\sigma^2 + Rt/3)$ to obtain a nice form for the final exponential. To control the fraction, we remark 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 follows from the 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.

Combine the results from the last paragraph to reach

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \geq t\} \leq d \cdot \left(1 + \frac{3(\sigma^2 + Rt/3)^2}{t^4}\right) \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

This probability inequality is typically vacuous when $t^2 < \sigma^2 + Rt/3$, so we may as well limit out attention to the case where $t^2 \geq \sigma^2 + Rt/3$. Under this assumption, the parenthesis is bounded by four, which gives the tail bound (7.3.1). We can simplify the restriction on t by solving the quadratic inequality to obtain the sufficient condition

$$t \geq \frac{1}{2}\left[\frac{R}{3} + \sqrt{\frac{R^2}{9} + 4\sigma^2}\right].$$

We develop an upper bound for the right-hand side of this inequality as follows.

$$\frac{1}{2}\left[\frac{R}{3} + \sqrt{\frac{R^2}{9} + 4\sigma^2}\right] = \frac{R}{6}\left[1 + \sqrt{1 + \frac{36\sigma^2}{R^2}}\right] \leq \frac{R}{6}\left[1 + 1 + \frac{6\sigma}{R}\right] = \sigma + \frac{R}{3}.$$

We have used the numerical fact $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \geq 0$. Therefore, the tail bound (7.3.1) is valid when $t \geq \sigma + R/3$. □

7.7.2 The General Case

Finally, we present the proof of the intrinsic Bernstein inequality for an independent sum of bounded random matrices.

Proof of Corollary 7.3.2. Suppose that $\{\boldsymbol{S}_k\}$ is a finite sequence of independent random matrices that satisfy

$$\mathbb{E}\,\boldsymbol{S}_k = \mathbf{0} \quad \text{and} \quad \|\boldsymbol{S}_k\| \leq R.$$

Form the sum $\boldsymbol{Z} = \sum_k \boldsymbol{S}_k$. As in the proof of Corollary 6.2.1, we derive the result by applying Theorem 7.3.1 to the Hermitian dilation $\boldsymbol{Y} = \mathscr{H}(\boldsymbol{Z})$. The only new point that requires attention is the definition of the intrinsic dimension of $\boldsymbol{Z}$. From the statement of Theorem 7.3.1, we have

$$d(\boldsymbol{Y}) = \operatorname{intdim}\big(\mathbb{E}(\boldsymbol{Y}^2)\big) = \operatorname{intdim}\big(\mathbb{E}\,\mathscr{H}(\boldsymbol{Z})^2\big) = \operatorname{intdim}\begin{bmatrix}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*) & \mathbf{0} \\ \mathbf{0} & \mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})\end{bmatrix}.$$

The last identity arises from the formula (2.1.12) for the square of the dilation. We determine that the appropriate definition for the intrinsic dimension parameter of $\boldsymbol{Z}$ is

$$d(\boldsymbol{Z}) = \operatorname{intdim}\begin{bmatrix}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*) & \mathbf{0} \\ \mathbf{0} & \mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})\end{bmatrix}.$$

This point completes the argument. □

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 total 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 and 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, and Zhang [HKZ12b]. They obtain a bound that is similar to Theorem 7.3.1. Unfortunately, their argument is complicated, and the results it delivers are less refined than the ones given here.

Theorem 7.3.1 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.

We present a new, general approach that delivers intrinsic dimension bounds. The intrinsic Chernoff bounds that emerge from our framework are new. The proof of the intrinsic Bernstein bound, Theorem 7.3.1, 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.

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. It favors, unapologetically, the work of the author. Additional papers may be included at a later time.

Exponential Matrix Concentration Inequalities

We begin with papers that contain the most current results on matrix concentration.

- [Tro11d]. 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.

- [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 [Tro11c] describes additional results for matrix-valued martingales.

- [GT11]. 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+12]. 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. In particular, 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.

- [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.

- [Tro11e]. 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.

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 Ahlswede–Winter argument [AW02]. This 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. The paper applies these results to study randomized matrix multiplication algorithms.

- [HKZ12b]. 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.

The Ahlswede–Winter Method

In this section, we list some papers that use the ideas from the Ahslwede–Winter paper [AW02] 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 and Winter describes the matrix Laplace transform method, along with a number of other fundamental 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 Ahlswede–Winter argument. They apply this result to study random graphs.
- [Gro11]. Gross presents a matrix Bernstein inequality based on the Ahlswede–Winter method, 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 Ahlswede–Winter technique. 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 an optimal version of the noncommutative Khintchine inequality.

- [Buc01, Buc05]. These papers prove optimal noncommutative Khintchine inequalities is more general settings.
- [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.
- [JZ11]. This paper describes a fully noncommutative version of the Bennett inequality. The proof is based on the Ahlswede–Winter method [AW02].
- [JZ12]. This work shows how to use Oliveira's argument [Oli10a] to obtain some results for fully noncommutative martingales.
- [MJC+12]. This work, described above, includes a section on matrix moment inequalities. This paper contains what are probably the simplest available proofs of these results.

Bibliography

[AC09] N. Ailon and B. Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.

[AM07] D. Achlioptas and F. McSherry. Fast computation of low-rank matrix approximations. *J. Assoc. Comput. Mach.*, 54(2):Article 10, 2007. (electronic).

[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.

[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.

[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*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2001.

[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.

[But] S. Butler. Spectral graph theory.

[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.

[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. Infer.*, 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.

[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.

[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.

[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.

[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.

[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.

[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.

[Fre75] D. A. Freedman. On tail probabilities for martingales. *Ann. Probab.*, 3(1):100–118, Feb. 1975.

[Git11] A. Gittens. The spectral norm error of the naïve Nyström extension. Available at arXiv:1110.5305, Oct. 2011.

[GN] D. Gross and V. Nesme. Note on sampling without replacing from a finite collection of matrices. Available at arXiv: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. Royce. *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.

[Gro09] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, Oct. 2009. To appear. Available at arXiv:0910.1879.

[Gro11] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 57(3):1548–1566, Mar. 2011.

[GT] A. Gittens and J. A. Tropp. Error bounds for random matrix approximation schemes. Available at arXiv:0911.4108.

[GT11] A. Gittens and J. A. Tropp. Tail bounds for all eigenvalues of a sum of random matrices. Available at arXiv:1104.4513, Apr. 2011.

[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 entangles 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.

[HJ85] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge Univ. Press, Cambridge, 1985.

[HJ94] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge Univ. Press, Cambridge, 1994.

[HKZ12a] D. Hsu, S. Kakade, and T. Zhang. Analysis of a randomized approximation scheme for matrix multiplication. Available at arXiv:1211.5414, Nov. 2012.

[HKZ12b] 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.

[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.

[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.

[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.

[JZ11] M. Junge and Q. Zheng. Noncommutative Bennett and Rosenthal inequalities. Available at arXiv:1111.1027, Nov. 2011.

[JZ12] M. Junge and Q. Zeng. Noncommutative martingale deviation and Poincaré type inequalities with applications. Available at arXiv:1211.3209, Nov. 2012.

[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].

[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.

[Lie73] E. H. Lieb. Convex trace functions and the Wigner–Yanase–Dyson conjecture. *Adv. Math.*, 11:267–288, 1973.

[Lin74] G. Lindblad. Expectations and entropy inequalities for finite quantum systems. *Comm. Math. Phys.*, 39:111–119, 1974.

[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.

[LS05] E. H. Lieb and R. Seiringer. Stronger subadditivity of entropy. *Phys. Rev. A*, 71:062329–1–9, 2005.

[Lug09] G. Lugosi. Concentration-of-measure inequalities. Available at http://www.econ.upf.edu/~lugosi/anu.pdf, 2009.

[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.

[Min11] S. Minsker. Some extensions of Bernstein's inequality for self-adjoint operators. Available at arXiv:1112.5448, Nov. 2011.

[MJC+12] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. Matrix concentration inequalities via the method of exchangable pairs. Available at arXiv:1201.6002, Jan. 2012.

[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.

[MT12] M. McCoy and J. A. Tropp. Sharp recovery thresholds for convex deconvolution, with applications. Available at arXiv:1205.1580, May 2012.

[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, Philadelphia, PA, 2011. SIAM.

[Nem07] A. Nemirovski. Sums of random symmetric matrices and quadratic optimization under orthogonality constraints. *Math. Prog. Ser. B*, 109:283–317, 2007.

[NT12] D. Needell and J. A. Tropp. Paved with good intentions: Analysis of a randomized block Kaczmarz method. Available at arXiv:1208.3805, Aug. 2012.

[Oli10a] R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Available at arXiv: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.

[Pau02] V. I. Paulsen. *Completely Bounded Maps and Operator Algebras*. Number 78 in Cambridge Studies in Advanced Mathematics. Cambridge Univ. Press, Cambridge, 2002.

[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*, volume 30 of *Banach Center Publications*, pages 287–298, Warsaw, 1994. Polish Acad. Sci.

[Rec11] B. Recht. A simpler approach to matrix completion. *J. Mach. Learn. Res.*, pages 3413–3430, Dec. 2011.

[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.

[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).

[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.

[So09] A. M.-C. So. Moment inequalities for sums of random matrices and their applications in optimization. *Math. Prog. Ser. A*, Dec. 2009. (electronic).

[Spe11] R. Speicher. Free probability theory. In *The Oxford handbook of random matrix theory*, pages 452–470. Oxford Univ. Press, Oxford, 2011.

[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.

[SV11] A. Sen and B. Virág. The top eigenvalue of the random Toeplitz matrix and the Sine kernel. Available at arXiv:1109.5494., Sep. 2011.

[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. From the joint convexity of quantum relative entropy to a concavity theorem of Lieb. *Proc. Amer. Math. Soc.*, 2011. To appear. Available at arXiv:1101.1070.

[Tro11c] J. A. Tropp. User-friendly tail bounds for matrix martingales. ACM Report 2011-01, California Inst. Tech., Pasadena, CA, Jan. 2011.

[Tro11d] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, August 2011.

[Tro11e] 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.

[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.

[Wis28] J. Wishart. The generalised product moment distribution in samples from a multivariate normal population. *Biometrika*, 20A(1–2):32–52, 1928.

User-Friendly Tools for Random Matrices

Joel A. Tropp

Computing + Mathematical Sciences
California Institute of Technology
jtropp@cms.caltech.edu

Research supported by ONR, AFOSR, NSF, DARPA, Sloan, and Moore.

Download the Notes:

tinyurl.com/bocrqhe

[URL] `http://users.cms.caltech.edu/~jtropp/notes/Tro12-User-Friendly-Tools-NIPS.pdf`

Random Matrices in the Mist

Random Matrices in Statistics

- Covariance estimation for the multivariate normal distribution

John Wishart

3. *Multi-variate Distribution. Use of Quadratic co-ordinates.*

A comparison of equation (8) with the corresponding results (1) and (2) for uni-variate and bi-variate sampling, respectively, indicates the form the general result may be expected to take. In fact, we have for the simultaneous distribution in random samples of the n variances (squared standard deviations) and the $\frac{n(n-1)}{2}$ product moment coefficients the following expression:

$$dp = \frac{\begin{vmatrix} A_{11} & A_{12} \dots A_{1n} \\ A_{21} & A_{22} \dots A_{2n} \\ \vdots & \vdots \quad \vdots \\ A_{n1} & A_{n2} \dots A_{nn} \end{vmatrix}^{\frac{N-1}{2}}}{(\sqrt{\pi})^{\frac{1}{2}n(n-1)}\,\Gamma\left(\frac{N-1}{2}\right)\Gamma\left(\frac{N-2}{2}\right)\dots\Gamma\left(\frac{N-n}{2}\right)} \times e^{-A_{11}a_{11}-A_{22}a_{22}-\dots-A_{nn}a_{nn}-2A_{12}a_{12}-2A_{13}a_{13}-\dots-2A_{n-1\,n}a_{n-1\,n}} \times \begin{vmatrix} a_{11} & a_{12} \dots a_{1n} \\ a_{21} & a_{22} \dots a_{2n} \\ \vdots & \vdots \quad \vdots \\ a_{n1} & a_{n2} \dots a_{nn} \end{vmatrix}^{\frac{N-n-2}{2}} da_{11}da_{12}\dots da_{nn} \quad \dots\dots\dots\dots\dots\dots(9),$$

where $a_{pq} = s_p s_q r_{pq}$, and $A_{pq} = \frac{N}{2\sigma_p\sigma_q} \cdot \frac{\Delta_{pq}}{\Delta}$, Δ being the determinant

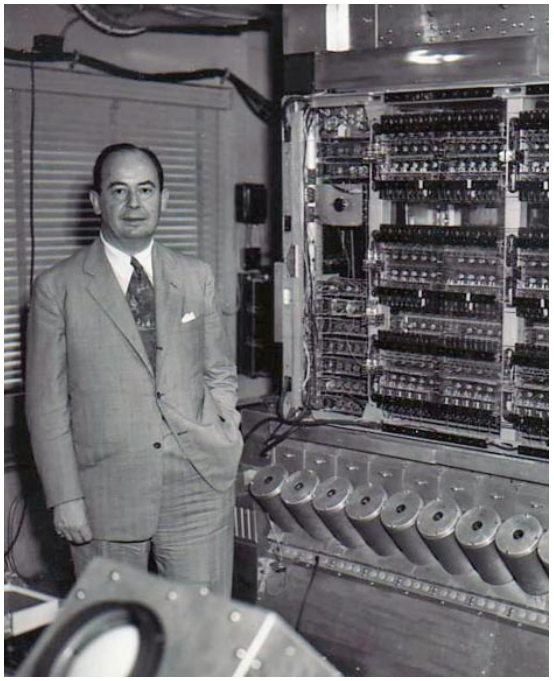
$$|\rho_{pq}|, p, q = 1, 2, 3, \dots n,$$

and Δ_{pq} the minor of ρ_{pq} in Δ.

[Refs] Wishart, *Biometrika* 1928. Photo from `apprendre-math.info`.

Random Matrices in Numerical Linear Algebra

- Model for floating-point errors in LU decomposition



John von Neumann

now combining (8.6) and (8.7) we obtain our desired result:

$$\text{Prob}\ (\lambda > 2\sigma^2 rn) < \frac{(rn)^{n-1/2} e^{-rn} \pi^{1/2} e^n \cdot 2^{n-2}}{\pi n^{n-1}(r-1)n} \tag{8.8}$$
$$= \left(\frac{2r}{e^{r-1}}\right)^n \times \frac{1}{4(r-1)(r\pi n)^{1/2}}.$$

We sum up in the following theorem:

(8.9) The probability that the upper bound $|A|$ of the matrix A of (8.1) exceeds $2.72\sigma n^{1/2}$ is less than $.027 \times 2^{-n} n^{-1/2}$, that is, with probability greater than 99% the upper bound of A is less than $2.72\sigma n^{1/2}$ for $n = 2, 3, \cdots$.

This follows at once by taking $r = 3.70$.

[Refs] von Neumann and Goldstine, *Bull. AMS* 1947 and *Proc. AMS* 1951. Photo ©IAS Archive.

Random Matrices in Nuclear Physics

- Model for the Hamiltonian of a heavy atom in a slow nuclear reaction

Eugene Wigner

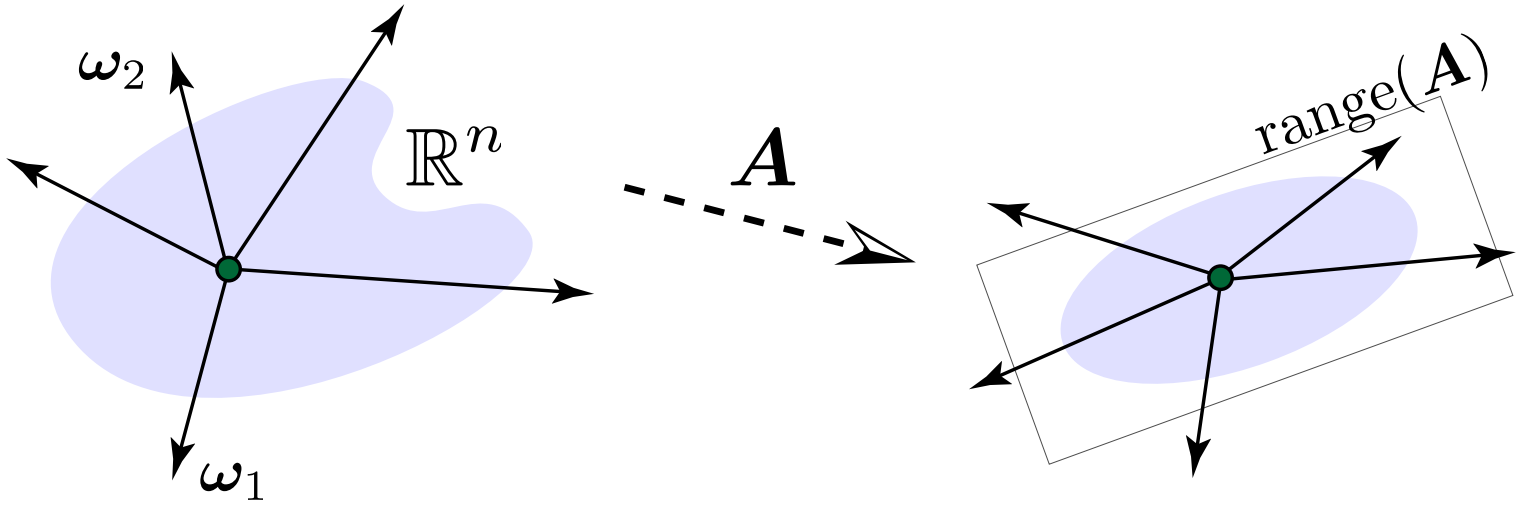
Random sign symmetric matrix

The matrices to be considered are $2N + 1$ dimensional real symmetric matrices; N is a very large number. The diagonal elements of these matrices are zero, the non diagonal elements $v_{ik} = v_{ki} = \pm v$ have all the same absolute value but random signs. There are $\mathfrak{N} = 2^{N(2N+1)}$ such matrices. We shall calculate, after an introductory remark, the averages of $(H^r)_{00}$ and hence the strength function $S'(x) = \sigma(x)$. This has, in the present case, a second interpretation: it also gives the density of the characteristic values of these matrices. This will be shown first.

[Refs] Wigner, *Ann. Math* 1955. Photo from Nobel Foundation.

Modern Applications

Randomized Linear Algebra



Input: An $m \times n$ matrix $\boldsymbol{A}$, a target rank k, an oversampling parameter p

Output: An $m \times (k + p)$ matrix $\boldsymbol{Q}$ with orthonormal columns

1. Draw an $n \times (k + p)$ random matrix $\boldsymbol{\Omega}$
2. Form the matrix product $\boldsymbol{Y} = \boldsymbol{A\Omega}$
3. Construct an orthonormal basis $\boldsymbol{Q}$ for the range of $\boldsymbol{Y}$

[Ref] Halko–Martinsson–T, *SIAM Rev.* 2011.

Other Algorithmic Applications

- **Sparsification.** Accelerate spectral calculation by randomly zeroing entries in a matrix.
- **Subsampling.** Accelerate construction of kernels by randomly subsampling data.
- **Dimension Reduction.** Accelerate nearest neighbor calculations by random projection to a lower dimension.
- **Relaxation & Rounding.** Approximate solution of maximization problems with matrix variables.

[Refs] Achlioptas–McSherry 2001 and 2007, Spielman–Teng 2004; Williams–Seeger 2001, Drineas–Mahoney 2006, Gittens 2011; Indyk–Motwani 1998, Ailon–Chazelle 2006; Nemirovski 2007, So 2009...

Random Matrices as Models

- **High-Dimensional Data Analysis.** Random matrices are used to model multivariate data.

- **Wireless Communications.** Random matrices serve as models for wireless channels.

- **Demixing Signals.** Random model for incoherence when separating two structured signals.

[Refs] Bühlmann and van de Geer 2011, Koltchinskii 2011; Tulino–Verdú 2004; McCoy–T 2011.

Theoretical Applications

- **Algorithms.** Smoothed analysis of Gaussian elimination.
- **Combinatorics.** Random constructions of expander graphs.
- **High-Dimensional Geometry.** Structure of random slices of convex bodies.
- **Quantum Information Theory.** (Counter)examples to conjectures about quantum channel capacity.

[Refs] Sankar–Spielman–Teng 2006; Pinsker 1973; Gordon 1985; Hayden–Winter 2008, Hastings 2009.

Random Matrices: My Way

The Conventional Wisdom

"Random Matrices are Tough!"

[Refs] youtube.com/watch?v=NO0cvqT1tAE, most monographs on RMT.

Principle A

"But...

In many applications, a random matrix can be decomposed as a sum of independent random matrices:

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

Principle B

and

There are exponential concentration inequalities for the spectral norm of a sum of independent random matrices:

$$\mathbb{P}\{\|\boldsymbol{Z}\| \geq t\} \leq \exp(\quad \cdots \quad)$$

!!!”

Matrix Gaussian Series

The Norm of a Matrix Gaussian Series

Theorem 1. [Oliveira 2010, T 2010] **Suppose**

- $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \dots$ *are fixed matrices with dimension* $d_1 \times d_2$, *and*
- $\gamma_1, \gamma_2, \gamma_3, \dots$ *are independent standard normal RVs.*

Define $d := d_1 + d_2$ *and the variance parameter*

$$\sigma^2 := \max\left\{ \left\| \sum\nolimits_k \boldsymbol{B}_k \boldsymbol{B}_k^* \right\|, \ \left\| \sum\nolimits_k \boldsymbol{B}_k^* \boldsymbol{B}_k \right\| \right\}.$$

Then

$$\mathbb{P}\left\{ \left\| \sum\nolimits_k \gamma_k \boldsymbol{B}_k \right\| \geq t \right\} \leq d \cdot \mathrm{e}^{-t^2/2\sigma^2}.$$

[Refs] Tomczak–Jaegerman 1974, Lust-Picquard 1986, Lust-Picquard–Pisier 1991, Rudelson 1999, Buchholz 2001 and 2005, Oliveira 2010, T 2011. **Notes: Cor. 4.2.1, page 33.**

The Norm of a Matrix Gaussian Series

Theorem 2. [Oliveira 2010, T 2010] **Suppose**

- $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \dots$ *are fixed matrices with dimension* $d_1 \times d_2$, *and*
- $\gamma_1, \gamma_2, \gamma_3, \dots$ *are independent standard normal RVs.*

Define $d := d_1 + d_2$ *and the variance parameter*

$$\sigma^2 := \max\left\{ \left\| \sum\nolimits_k \boldsymbol{B}_k \boldsymbol{B}_k^* \right\|, \ \left\| \sum\nolimits_k \boldsymbol{B}_k^* \boldsymbol{B}_k \right\| \right\}.$$

Then

$$\mathbb{E} \left\| \sum\nolimits_k \gamma_k \boldsymbol{B}_k \right\| \leq \sqrt{2\sigma^2 \log d}.$$

[Refs] Tomczak–Jaegerman 1974, Lust-Picquard 1986, Lust-Picquard–Pisier 1991, Rudelson 1999, Buchholz 2001 and 2005, Oliveira 2010, T 2011. **Notes: Cor. 4.2.1, page 33.**

The Variance Parameter

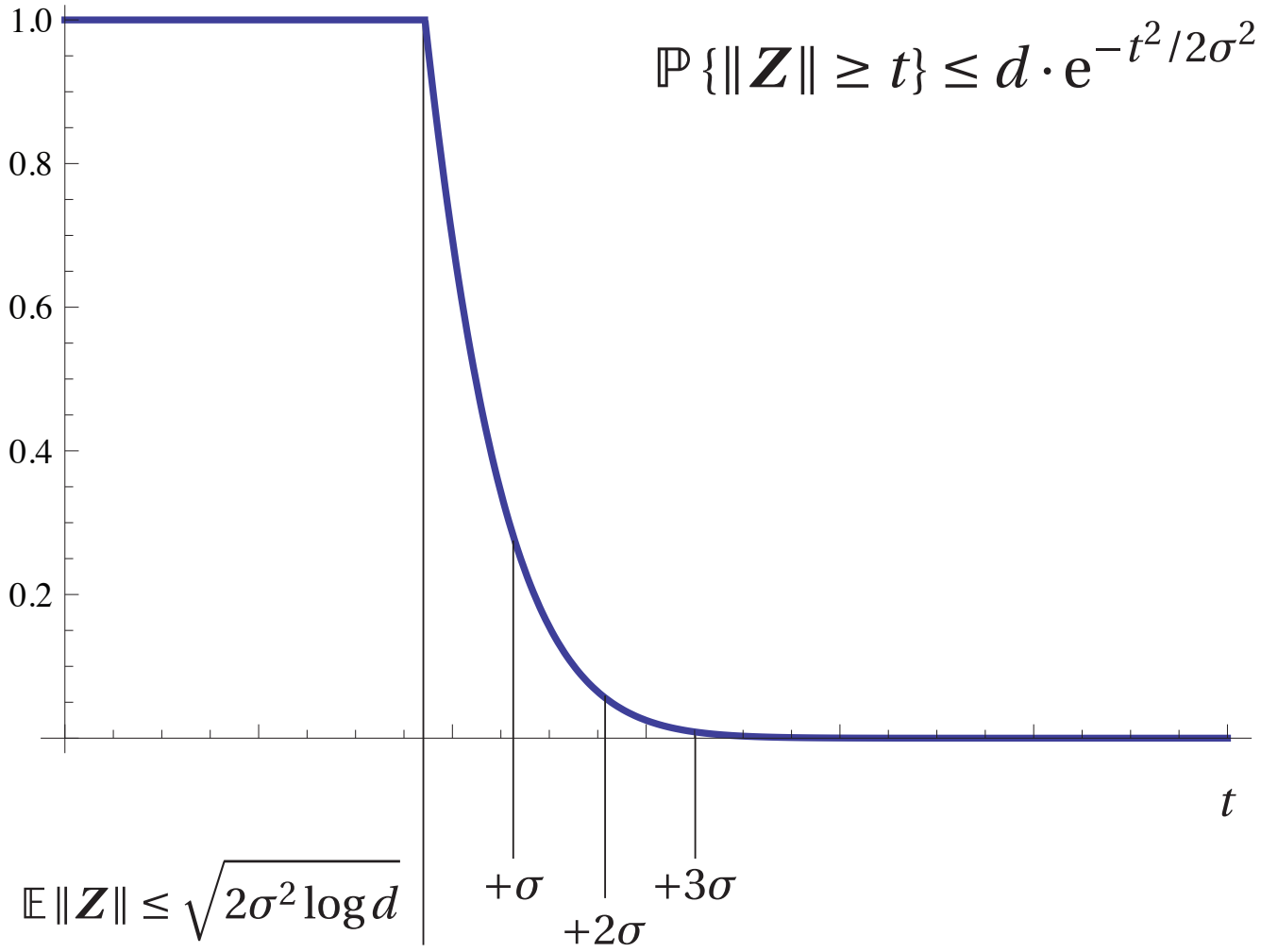
- Define the matrix Gaussian series $\boldsymbol{Z} = \sum_{k=1}^{n} \gamma_k \boldsymbol{B}_k$
- The variance parameter $\sigma^2(\boldsymbol{Z})$ derives from the "mean square of $\boldsymbol{Z}$"
- But a general matrix has *two* different squares!

$$\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*) = \sum_{j=1}^{n}\sum_{k=1}^{n} \mathbb{E}(\gamma_j\gamma_k)\boldsymbol{B}_j\boldsymbol{B}_k^* = \sum_{k=1}^{n} \boldsymbol{B}_k\boldsymbol{B}_k^*$$

$$\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z}) = \sum_{j=1}^{n}\sum_{k=1}^{n} \mathbb{E}(\gamma_j\gamma_k)\boldsymbol{B}_j^*\boldsymbol{B}_k = \sum_{k=1}^{n} \boldsymbol{B}_k^*\boldsymbol{B}_k$$

- Variance parameter $\sigma^2(\boldsymbol{Z}) = \max\{\|\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^*)\|\,,\ \|\mathbb{E}(\boldsymbol{Z}^*\boldsymbol{Z})\|\}$.

Schematic of Gaussian Series Tail Bound



Warmup: A Wigner Matrix

- Let $\{\gamma_{jk} : 1 \leq j < k \leq n\}$ be independent standard normal variables
- A Gaussian Wigner matrix:

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

- **Problem:** What is $\mathbb{E}\,\|\boldsymbol{W}\|$?

Notes: §4.4.1, page 35.

The Wigner Matrix, *qua* Gaussian Series

- Express the Wigner matrix as a Gaussian series:

$$\boldsymbol{W} = \sum_{1 \le j < k \le n} \gamma_{jk} (\mathbf{E}_{jk} + \mathbf{E}_{kj})$$

- The symbol $\mathbf{E}_{jk}$ denotes the $n \times n$ matrix unit

$$\mathbf{E}_{jk} = \begin{bmatrix} & & \\ & 1 & \\ & & \end{bmatrix} \begin{matrix} \\ \leftarrow\ j \\ \\ \end{matrix}$$

$$\begin{matrix} \\ \uparrow \\ k \end{matrix}$$

Norm Bound for the Wigner Matrix

- Need to compute the variance parameter $\sigma^2(\boldsymbol{W})$
- Summands are symmetric, so both matrix squares are the same:

$$\sum_{1 \le j < k \le n} (\mathbf{E}_{jk} + \mathbf{E}_{kj})^2 = \sum_{1 \le j < k \le n} (\mathbf{E}_{jk}\mathbf{E}_{jk} + \mathbf{E}_{jk}\mathbf{E}_{kj} + \mathbf{E}_{kj}\mathbf{E}_{jk} + \mathbf{E}_{kj}\mathbf{E}_{kj})$$
$$= \sum_{1 \le j < k \le n} (\mathbf{0} + \mathbf{E}_{jj} + \mathbf{E}_{kk} + \mathbf{0}) = (n-1)\,\mathbf{I}_n$$

- Thus, the variance $\sigma^2(\boldsymbol{W}) = \|(n-1)\,\mathbf{I}_n\| = n - 1$.
- **Conclusion:** $\mathbb{E}\,\|\boldsymbol{W}\| \le \sqrt{2(n-1)\log(2n)}$
- **Optimal:** $\mathbb{E}\,\|\boldsymbol{W}\| \sim 2\sqrt{n}$

[Refs] Wigner 1955, Davidson–Szarek 2002, Tao 2012.

Example: A Gaussian Toeplitz Matrix

- Let $\{\gamma_k\}$ be independent standard normal variables
- An unsymmetric Gaussian Toeplitz matrix:

$$\boldsymbol{T} = \begin{bmatrix} \gamma_0 & \gamma_1 & & \cdots & & \gamma_{n-1} \\ \gamma_{-1} & \gamma_0 & \gamma_1 & & & \\ & \gamma_{-1} & \gamma_0 & \gamma_1 & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \\ & & & \gamma_{-1} & \gamma_0 & \gamma_1 \\ \gamma_{-(n-1)} & & \cdots & & \gamma_{-1} & \gamma_0 \end{bmatrix}$$

- **Problem:** What is $\mathbb{E}\,\|\boldsymbol{T}\|$?

Notes: §4.6, page 38.

The Toeplitz Matrix, *qua* Gaussian Series

- Express the unsymmetric Toeplitz matrix as a Gaussian series:

$$\boldsymbol{T} = \gamma_0 \, \mathbf{I} + \sum_{k=1}^{n-1} \gamma_k \boldsymbol{S}^k + \sum_{k=1}^{n-1} \gamma_{-k} (\boldsymbol{S}^k)^*$$

- The matrix $\boldsymbol{S}$ is the shift-up operator on n-dimensional column vectors:

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

Variance Calculation for the Toeplitz Matrix

- Note that

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

- *Both* sums of squares take the form

$$\begin{aligned}
&\mathbf{I}^2 + \sum_{k=1}^{n-1} (\boldsymbol{S}^k)(\boldsymbol{S}^k)^* + \sum_{k=1}^{n-1} (\boldsymbol{S}^k)^*(\boldsymbol{S}^k) \\
&\quad = \mathbf{I} + \sum_{k=1}^{n-1} \left[\sum_{j=1}^{n-k} \mathbf{E}_{jj} + \sum_{j=k+1}^{n} \mathbf{E}_{jj} \right] = \sum_{j=1}^{n} \left[1 + \sum_{k=1}^{n-j} 1 + \sum_{k=1}^{j-1} 1 \right] \mathbf{E}_{jj} \\
&\quad = \sum_{j=1}^{n} (1 + (n-j) + (j-1)) \, \mathbf{E}_{jj} = n \, \mathbf{I}_n.
\end{aligned}$$

Norm Bound for the Toeplitz Matrix

- The variance parameter $\sigma^2(\boldsymbol{T}) = \|n\,\mathbf{I}_n\| = n$
- **Conclusion:** $\mathbb{E}\,\|\boldsymbol{T}\| \leq \sqrt{2n\log(2n)}$
- **Optimal:** $\mathbb{E}\,\|\boldsymbol{T}\| \sim \text{const} \cdot \sqrt{2n\log n}$
- The optimal constant is at least 0.8288...

[Refs] Bryc–Dembo–Jiang 2006, Meckes 2007, Sen–Virág 2011, T 2011.

Matrix Chernoff Inequality

The Matrix Chernoff Bound

Theorem 3. [T 2010] **Suppose**

- $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \dots$ *are random* **psd** *matrices with dimension* d*, and*
- $\lambda_{\max}(\boldsymbol{X}_k) \le R$ *for each* k*.*

Then

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \le (1-t)\cdot \mu_{\min}\right\} \le d \cdot \left[\frac{\mathrm{e}^{-t}}{(1-t)^{1-t}}\right]^{\mu_{\min}/R}$$

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \ge (1+t)\cdot \mu_{\max}\right\} \le d \cdot \left[\frac{\mathrm{e}^{t}}{(1+t)^{1+t}}\right]^{\mu_{\max}/R}$$

where $\mu_{\min} := \lambda_{\min}\left(\sum_k \mathbb{E}\,\boldsymbol{X}_k\right)$ *and* $\mu_{\max} := \lambda_{\max}\left(\sum_k \mathbb{E}\,\boldsymbol{X}_k\right)$*.*

[Refs] Ahlswede–Winter 2002, T 2011. **Notes: Thm. 5.1.1, page 48.**

The Matrix Chernoff Bound

Theorem 4. [T 2010] **Suppose**

- $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \dots$ *are random* **psd** *matrices with dimension* d*, and*
- $\lambda_{\max}(\boldsymbol{X}_k) \le R$ *for each* k*.*

Then

$$\mathbb{E}\,\lambda_{\min}\Big(\sum\nolimits_k \boldsymbol{X}_k\Big) \ge 0.6\,\mu_{\min} - R\log d$$

$$\mathbb{E}\,\lambda_{\max}\Big(\sum\nolimits_k \boldsymbol{X}_k\Big) \le 1.8\,\mu_{\max} + R\log d$$

where $\mu_{\min} := \lambda_{\min}\left(\sum_k \mathbb{E}\,\boldsymbol{X}_k\right)$ *and* $\mu_{\max} := \lambda_{\max}\left(\sum_k \mathbb{E}\,\boldsymbol{X}_k\right)$*.*

[Refs] Ahlswede–Winter 2002, T 2011. **Notes: Thm. 5.1.1, page 48.**

Example: Random Submatrices

Fixed matrix, in captivity:

$$\boldsymbol{C} = \begin{bmatrix} | & | & | & | & & | \\ \boldsymbol{c}_1 & \boldsymbol{c}_2 & \boldsymbol{c}_3 & \boldsymbol{c}_4 & \cdots & \boldsymbol{c}_n \\ | & | & | & | & & | \end{bmatrix}_{d \times n}$$

Random matrix, formed by picking random columns:

$$\boldsymbol{Z} = \begin{bmatrix} & | & | & & | \\ & \boldsymbol{c}_2 & \boldsymbol{c}_3 & \cdots & \boldsymbol{c}_n \\ & | & | & & | \end{bmatrix}_{d \times n}$$

$$\uparrow \quad \uparrow \qquad \uparrow$$

Problem: What is the expectation of $\sigma_1(\boldsymbol{Z})$? What about $\sigma_d(\boldsymbol{Z})$?

Notes: §5.2.1, page 49.

Model for Random Submatrix

- Let $\boldsymbol{C}$ be a fixed $d \times n$ matrix with columns $\boldsymbol{c}_1, \dots, \boldsymbol{c}_n$
- Let $\delta_1, \dots, \delta_n$ be independent 0–1 random variables with mean s/n
- Define $\boldsymbol{\Delta} = \operatorname{diag}(\delta_1, \dots, \delta_n)$
- Form a random submatrix $\boldsymbol{Z}$ by turning off columns from $\boldsymbol{C}$

$$\boldsymbol{Z} = \boldsymbol{C}\boldsymbol{\Delta} = \begin{bmatrix} | & | & & | \\ \boldsymbol{c}_1 & \boldsymbol{c}_2 & \dots & \boldsymbol{c}_n \\ | & | & & | \end{bmatrix}_{d \times n} \begin{bmatrix} \delta_1 & & & \\ & \delta_2 & & \\ & & \ddots & \\ & & & \delta_n \end{bmatrix}_{n \times n}$$

- Note that $\boldsymbol{Z}$ typically contains about s nonzero columns

The Random Submatrix, *qua* PSD Sum

- The largest and smallest singular values of $\boldsymbol{Z}$ satisfy

$$\sigma_1(\boldsymbol{Z})^2 = \lambda_{\max}(\boldsymbol{Z}\boldsymbol{Z}^*)$$
$$\sigma_d(\boldsymbol{Z})^2 = \lambda_{\min}(\boldsymbol{Z}\boldsymbol{Z}^*)$$

- Define the psd matrix $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{Z}^*$, and observe that

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{Z}^* = \boldsymbol{C}\boldsymbol{\Delta}^2\boldsymbol{C}^* = \boldsymbol{C}\boldsymbol{\Delta}\boldsymbol{C}^* = \sum\nolimits_{k=1}^{n} \delta_k \, \boldsymbol{c}_k\boldsymbol{c}_k^*$$

- We have expressed $\boldsymbol{Y}$ as a sum of independent psd random matrices

Preparing to Apply the Chernoff Bound

- Consider the random matrix

$$\boldsymbol{Y} = \sum\nolimits_k \delta_k \, \boldsymbol{c}_k \boldsymbol{c}_k^*$$

- The maximal eigenvalue of each summand is bounded as

$$R = \max\nolimits_k \lambda_{\max}(\delta_k \, \boldsymbol{c}_k \boldsymbol{c}_k^*) \leq \max\nolimits_k \left\| \boldsymbol{c}_k \right\|^2$$

- The expectation of the random matrix $\boldsymbol{Y}$ is

$$\mathbb{E}(\boldsymbol{Y}) = \frac{s}{n} \sum\nolimits_{k=1}^{n} \boldsymbol{c}_k \boldsymbol{c}_k^* = \frac{s}{n} \boldsymbol{C}\boldsymbol{C}^*$$

- The mean parameters satisfy

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

What the Chernoff Bound Says

Applying the Chernoff bound, we reach

$$\mathbb{E}\left[\sigma_1(\boldsymbol{Z})^2\right] = \mathbb{E}\,\lambda_{\max}(\boldsymbol{Y}) \leq 1.8 \cdot \frac{s}{n}\,\sigma_1(\boldsymbol{C})^2 + \max_k \|\boldsymbol{c}_k\|_2^2 \cdot \log d$$

$$\mathbb{E}\left[\sigma_d(\boldsymbol{Z})^2\right] = \mathbb{E}\,\lambda_{\min}(\boldsymbol{Y}) \geq 0.6 \cdot \frac{s}{n}\,\sigma_d(\boldsymbol{C})^2 - \max_k \|\boldsymbol{c}_k\|_2^2 \cdot \log d$$

- Matrix $\boldsymbol{C}$ has n columns; the random submatrix $\boldsymbol{Z}$ includes about s
- The singular value $\sigma_i(\boldsymbol{Z})^2$ inherits an s/n share of $\sigma_i(\boldsymbol{C})^2$ for $i = 1, d$
- Additive correction reflects number d of rows of $\boldsymbol{C}$, max column norm
- **[Gittens–T 2011]** Remaining singular values have similar behavior

Key Example: Unit-Norm Tight Frame

- A $d \times n$ unit-norm tight frame $\boldsymbol{C}$ satisfies

$$\boldsymbol{C}\boldsymbol{C}^* = \frac{n}{d}\,\mathbf{I}_d \quad \text{and} \quad \|\boldsymbol{c}_k\|_2^2 = 1 \quad \text{for } k = 1, 2, \dots, n$$

- Specializing the inequalities from the previous slide...

$$\mathbb{E}\left[\sigma_1(\boldsymbol{Z})^2\right] \leq 1.8 \cdot \frac{s}{d} + \log d$$

$$\mathbb{E}\left[\sigma_d(\boldsymbol{Z})^2\right] \geq 0.6 \cdot \frac{s}{d} - \log d$$

- Choose $s \geq 1.67\, d \log d$ columns for a nontrivial lower bound
- Sharp condition $s > d \log d$ also follows from matrix Chernoff bound

[Refs] Rudelson 1999, Rudelson–Vershynin 2007, T 2008, Gittens–T 2011, T 2011, Chrétien–Darses 2012.

Matrix Bernstein Inequality

The Matrix Bernstein Inequality

Theorem 5. [Oliveira 2010, T 2010] **Suppose**

- $\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3, \dots$ *are indep. random matrices with dimension* $d_1 \times d_2$,
- $\mathbb{E}\, \boldsymbol{S}_k = \mathbf{0}$ *for each* k, *and*
- $\|\boldsymbol{S}_k\| \leq R$ *for each* k.

Then

$$\mathbb{P}\left\{ \left\| \sum\nolimits_k \boldsymbol{S}_k \right\| \geq t \right\} \leq d \cdot \exp\left\{ \frac{-t^2/2}{\sigma^2 + Rt/3} \right\}$$

where $d := d_1 + d_2$ *and the variance parameter*

$$\sigma^2 := \max\left\{ \left\| \sum\nolimits_k \mathbb{E}(\boldsymbol{S}_k \boldsymbol{S}_k^*) \right\|, \ \left\| \sum\nolimits_k \mathbb{E}(\boldsymbol{S}_k^* \boldsymbol{S}_k) \right\| \right\}$$

[Refs] Gross 2010, Recht 2011, Oliveira 2010, T 2011. **Notes: Cor. 6.2.1, page 64.**

The Matrix Bernstein Inequality

Theorem 6. [Oliveira 2010, T 2010] **Suppose**

- $\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3, \dots$ *are indep. random matrices with dimension* $d_1 \times d_2$,
- $\mathbb{E}\, \boldsymbol{S}_k = \mathbf{0}$ *for each* k, *and*
- $\|\boldsymbol{S}_k\| \le R$ *for each* k.

Then

$$\mathbb{E} \left\| \sum\nolimits_k \boldsymbol{S}_k \right\| \le \sqrt{2\sigma^2 \log d} + \tfrac{1}{3} R \log d$$

where $d := d_1 + d_2$ *and the variance parameter*

$$\sigma^2 := \max \left\{ \left\| \sum\nolimits_k \mathbb{E}(\boldsymbol{S}_k \boldsymbol{S}_k^*) \right\|, \ \left\| \sum\nolimits_k \mathbb{E}(\boldsymbol{S}_k^* \boldsymbol{S}_k) \right\| \right\}$$

[Refs] Gross 2010, Recht 2011, Oliveira 2010, T 2011. **Notes: Cor. 6.2.1, page 64.**

Example: Randomized Matrix Multiplication

Product of two matrices, in captivity:

$$\boldsymbol{BC}^* = \begin{bmatrix} | & | & | & | & & | \\ \boldsymbol{b}_1 & \boldsymbol{b}_2 & \boldsymbol{b}_3 & \boldsymbol{b}_4 & \dots & \boldsymbol{b}_n \\ | & | & | & | & & | \end{bmatrix}_{d_1 \times n} \begin{bmatrix} — & \boldsymbol{c}_1^* & — \\ — & \boldsymbol{c}_2^* & — \\ — & \boldsymbol{c}_3^* & — \\ — & \boldsymbol{c}_4^* & — \\ & \vdots & \\ — & \boldsymbol{c}_n^* & — \end{bmatrix}_{n \times d_2}$$

[Idea] Approximate multiplication by random sampling

[Refs] Drineas–Mahoney–Kannan 2004, Magen–Zouzias 2010, Magdon-Ismail 2010, Hsu–Kakade–Zhang 2011 and 2012.

A Sampling Model for Tutorial Purposes

- **Assume**

$$\|\boldsymbol{b}_j\|_2 = 1 \quad \text{and} \quad \|\boldsymbol{c}_j\|_2 = 1 \quad \text{for } j = 1, 2, \dots, n$$

- Construct a random variable $\boldsymbol{S}$ whose value is a $d_1 \times d_2$ matrix:
  - Draw $J \sim \text{UNIFORM}\{1, 2, \dots, n\}$
  - Set $\boldsymbol{S} = n \cdot \boldsymbol{b}_J \boldsymbol{c}_J^*$
- The random matrix $\boldsymbol{S}$ is an unbiased estimator of the product $\boldsymbol{BC}^*$

$$\mathbb{E}\,\boldsymbol{S} = \sum\nolimits_{j=1}^{n} (n \cdot \boldsymbol{b}_j \boldsymbol{c}_j^*) \cdot \mathbb{P}\{J = j\} = \sum\nolimits_{j=1}^{n} \boldsymbol{b}_j \boldsymbol{c}_j^* = \boldsymbol{BC}^*$$

- Approximate $\boldsymbol{BC}^*$ by averaging m independent copies of $\boldsymbol{S}$

$$\boldsymbol{Z} = \frac{1}{m} \sum\nolimits_{k=1}^{m} \boldsymbol{S}_k \approx \boldsymbol{BC}^*$$

Notes: §6.4, page 67.

Preparing to Apply the Bernstein Bound I

- Let $\boldsymbol{S}_k$ be independent copies of $\boldsymbol{S}$, and consider the average

$$\boldsymbol{Z} = \frac{1}{m} \sum\nolimits_{k=1}^{m} \boldsymbol{S}_k$$

- We study the typical approximation error

$$\mathbb{E} \, \|\boldsymbol{Z} - \boldsymbol{B}\boldsymbol{C}^*\| = \frac{1}{m} \cdot \mathbb{E} \left\| \sum\nolimits_{k=1}^{m} (\boldsymbol{S}_k - \boldsymbol{B}\boldsymbol{C}^*) \right\|$$

- The summands are independent and $\mathbb{E}\, \boldsymbol{S}_k = \boldsymbol{B}\boldsymbol{C}^*$, so we *symmetrize*:

$$\mathbb{E} \, \|\boldsymbol{Z} - \boldsymbol{B}\boldsymbol{C}^*\| \leq \frac{2}{m} \cdot \mathbb{E} \left\| \sum\nolimits_{k=1}^{m} \varepsilon_k \boldsymbol{S}_k \right\|$$

  where $\{\varepsilon_k\}$ are independent Rademacher RVs, independent from $\{\boldsymbol{S}_k\}$

Preparing to Apply the Bernstein Bound II

- The norm of each summand satisfies the uniform bound

$$R = \|\varepsilon \boldsymbol{S}\| = \|\boldsymbol{S}\| = \|n \cdot (\boldsymbol{b}_J \boldsymbol{c}_J^*)\| = n \|\boldsymbol{b}_J\|_2 \|\boldsymbol{c}_J\|_2 = n$$

- Compute the variance in two stages:

$$\begin{aligned}
\mathbb{E}(\boldsymbol{S}\boldsymbol{S}^*) &= \sum\nolimits_{j=1}^{n} n^2 (\boldsymbol{b}_j \boldsymbol{c}_j^*)(\boldsymbol{b}_j \boldsymbol{c}_j^*)^* \, \mathbb{P}\{J = j\} = n \sum\nolimits_{j=1}^{n} \|\boldsymbol{c}_j\|_2^2 \, \boldsymbol{b}_j \boldsymbol{b}_j^* \\
&= n \, \boldsymbol{B}\boldsymbol{B}^* \\
\mathbb{E}(\boldsymbol{S}^*\boldsymbol{S}) &= n \, \boldsymbol{C}\boldsymbol{C}^*
\end{aligned}$$

$$\begin{aligned}
\sigma^2 &= \max\left\{ \left\| \sum\nolimits_{k=1}^{m} \mathbb{E}(\boldsymbol{S}_k \boldsymbol{S}_k^*) \right\|, \ \left\| \sum\nolimits_{k=1}^{m} \mathbb{E}(\boldsymbol{S}_k \boldsymbol{S}_k^*) \right\| \right\} \\
&= \max\left\{ \|mn \cdot \boldsymbol{B}\boldsymbol{B}^*\|, \ \|mn \cdot \boldsymbol{C}\boldsymbol{C}^*\| \right\} \\
&= mn \cdot \max\{ \|\boldsymbol{B}\|^2, \ \|\boldsymbol{C}\|^2 \}
\end{aligned}$$

What the Bernstein Bound Says

Applying the Bernstein bound, we reach

$$\begin{aligned}
\mathbb{E}\,\|\boldsymbol{Z} - \boldsymbol{B}\boldsymbol{C}^*\| &\leq \frac{2}{m}\,\mathbb{E}\left\|\sum\nolimits_{k=1}^{m} \varepsilon_k \boldsymbol{S}_k\right\| \\
&\leq \frac{2}{m}\left[\sigma\sqrt{2\log(d_1+d_2)} + \tfrac{1}{3}R\log(d_1+d_2)\right] \\
&= 2\sqrt{\frac{n\log(d_1+d_2)}{m}}\cdot\max\{\|\boldsymbol{B}\|,\|\boldsymbol{C}\|\} + \frac{2}{3}\cdot\frac{n\log(d_1+d_2)}{m}
\end{aligned}$$

[Q] What can this possibly mean? Is this bound any good at all?

Detour: The Stable Rank

- The *stable rank* of a matrix is defined as

$$\operatorname{srank}(\boldsymbol{A}) := \frac{\|\boldsymbol{A}\|_{\mathrm{F}}^2}{\|\boldsymbol{A}\|^2}$$

- In general, $1 \leq \operatorname{srank}(\boldsymbol{A}) \leq \operatorname{rank}(\boldsymbol{A})$
- When $\boldsymbol{A}$ has either n rows or n columns, $1 \leq \operatorname{srank}(\boldsymbol{A}) \leq n$
- **Assume** that $\boldsymbol{A}$ has n unit-norm columns, so that $\|\boldsymbol{A}\|_{\mathrm{F}}^2 = n$
- When all columns of $\boldsymbol{A}$ are the same, $\|\boldsymbol{A}\|^2 = n$ and $\operatorname{srank}(\boldsymbol{A}) = 1$
- When all columns of $\boldsymbol{A}$ are orthogonal, $\|\boldsymbol{A}\|^2 = 1$ and $\operatorname{srank}(\boldsymbol{A}) = n$

Randomized Matrix Multiply, Relative Error

- Define the (geometric) *mean stable rank* of the factors to be

$$s := \sqrt{\operatorname{srank}(\boldsymbol{B}) \cdot \operatorname{srank}(\boldsymbol{C})}.$$

- Converting the error bound to a relative scale, we obtain

$$\frac{\mathbb{E}\,\|\boldsymbol{Z} - \boldsymbol{B}\boldsymbol{C}^*\|}{\|\boldsymbol{B}\|\,\|\boldsymbol{C}\|} \leq 2\sqrt{\frac{s\log(d_1+d_2)}{m}} + \frac{2}{3}\cdot\frac{s\log(d_1+d_2)}{m}$$

- For relative error $\varepsilon \in (0,1)$, the number m of samples should be

$$m \geq \text{Const} \cdot \varepsilon^{-2} \cdot s\log(d_1+d_2)$$

- **The number of samples is proportional to the mean stable rank!**
- We also pay weakly for the dimension $d_1 \times d_2$ of the product $\boldsymbol{B}\boldsymbol{C}^*$

More Things in Heaven & Earth

- **[More Bounds for Eigenvalues]** There are exponential tail bounds for maximum eigenvalues, minimum eigenvalues, and eigenvalues in between...
- **[More Exponential Bounds]** There is a matrix Hoeffding inequality and a matrix Bennett inequality, plus matrix Chernoff and Bernstein for unbounded matrices...
- **[Matrix Martingales]** There is a matrix Azuma inequality, a matrix bounded difference inequality, and a matrix Freedman inequality...
- **[Dependent Sums]** Exponential tail bounds hold for some random matrices based on dependent random variables...
- **[Polynomial Bounds]** There are matrix versions of the Rosenthal inequality, the Pinelis inequality, and the Burkholder–Davis–Gundy inequality...
- **[Intrinsic Dimension]** The dimensional dependence can sometimes be weakened...
- **[The Proofs!]** And the technical arguments are amazingly pretty...

[Refs] T 2011, Gittens–T 2011, Oliveira 2010, Mackey et al. 2012, ...

To learn more...

E-mail: jtropp@cms.caltech.edu
Web: http://users.cms.caltech.edu/~jtropp

Some papers:

- "User-friendly tail bounds for sums of random matrices," *FOCM*, 2011.
- "User-friendly tail bounds for matrix martingales." Caltech ACM Report 2011-01.
- "Freedman's inequality for matrix martingales," *ECP*, 2011.
- "A comparison principle for functions of a uniformly random subspace," *PTRF*, 2011.
- "From the joint convexity of relative entropy to a concavity theorem of Lieb," *PAMS*, 2012.

- "Improved analysis of the subsampled randomized Hadamard transform," *AADA*, 2011.
- "Tail bounds for all eigenvalues of a sum of random matrices" with A. Gittens. Submitted 2011.
- "The masked sample covariance estimator" with R. Chen and A. Gittens. *I&I*, 2012.
- "Matrix concentration inequalities..." with L. Mackey et al.. Submitted 2012.
- "User-Friendly Tools for Random Matrices: An Introduction." 2012.

See also...

- Ahlswede and Winter, "Strong converse for identification via quantum channels," *Trans. IT*, 2002.
- Oliveira, "Concentration of the adjacency matrix and of the Laplacian." Submitted 2010.
- Vershynin, "Introduction to the non-asymptotic analysis of random matrices," 2011.
- Minsker, "Some extensions of Bernstein's inequality for self-adjoint operators," 2011.