

User-Friendly Tail Bounds for Sums of Random Matrices

Joel A. Tropp

Received: 16 January 2011 / Accepted: 13 June 2011 / Published online: 2 August 2011
© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract This paper presents new probability inequalities for sums of independent, random, self-adjoint matrices. These results place simple and easily verifiable hypotheses on the summands, and they deliver strong conclusions about the large-deviation behavior of the maximum eigenvalue of the sum. Tail bounds for the norm of a sum of random rectangular matrices follow as an immediate corollary. The proof techniques also yield some information about matrix-valued martingales.

In other words, this paper provides noncommutative generalizations of the classical bounds associated with the names Azuma, Bennett, Bernstein, Chernoff, Hoeffding, and McDiarmid. The matrix inequalities promise the same diversity of application, ease of use, and strength of conclusion that have made the scalar inequalities so valuable.

Keywords Discrete-time martingale · Large deviation · Probability inequality · Random matrix · Sum of independent random variables

Mathematics Subject Classification (2000) Primary 60B20 · Secondary 60F10 · 60G50 · 60G42

Communicated by Albert Cohen.

J.A. Tropp (✉)
Computing & Mathematical Sciences, MC 305-16, California Institute of Technology, Pasadena,
CA 91125, USA
e-mail: jtropp@cms.caltech.edu

1 Introduction

Random matrices have come to play a significant role in computational mathematics. This line of research has advanced by using established methods from random matrix theory, but it has also generated difficult questions that cannot be addressed without new tools. Let us summarize some of the challenges that arise in numerical applications.

- Research has extended well beyond the classical ensembles (e.g., Wishart matrices and Wigner matrices) to encompass many other classes of random matrices. For instance, it is now common to study the properties of a sparse matrix sampled from a fixed matrix or a random submatrix drawn from a fixed matrix.
- We also encounter highly structured matrices that involve a limited amount of randomness. One important example is the randomized DFT, which consists of a diagonal matrix of random signs multiplied by a discrete Fourier transform matrix.
- Questions about the spectral properties of random matrices remain fundamental, but modern problems can also involve other considerations. For example, we might need to estimate the cut norm of a random adjacency matrix, or we might want to study the action of a random operator on a class of vectors or matrices.
- Most problems in numerical mathematics concern matrices of finite dimension. Asymptotic theory is less relevant in practice.
- We often require explicit large-deviation theorems for statistics of random matrices so that we can study rates of convergence.
- Results with effective constants are essential to ensure that algorithms are provably correct.

We have encountered these issues in a wide range of problems from computational mathematics: smoothed analysis of Gaussian elimination [51]; semidefinite relaxation and rounding of quadratic maximization problems [38, 50]; construction of maps for dimensionality reduction [1]; matrix approximation by sparsification [2] and by sampling submatrices [48]; analysis of sparse approximation [53] and compressive sampling [11] algorithms; randomized schemes for low-rank matrix factorization [24]; and analysis of algorithms for completion of low-rank matrices [20, 45]. And this list is by no means comprehensive!

In most of these applications, the methods currently invoked to study random matrices require a substantial amount of practice to use effectively. Even so, the final results tend to be a little disappointing: the constants are usually poor, and the predictions are sometimes coarser than we might like. These frustrations have led us to search for simpler techniques that still yield detailed quantitative information about finite random matrices.

1.1 Technical Overview

We consider a finite sequence $\{\mathbf{X}_k\}$ of random, self-adjoint matrices with dimension d . Our goal is to harness basic properties of these matrices to bound the probability

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geq t\right\}. \quad (1.1)$$

Here and elsewhere, λ_{\max} denotes the algebraically largest eigenvalue of a self-adjoint matrix. This formulation is more general than it may appear because we can exploit the same ideas to explore several related problems:

- We can study the smallest eigenvalue of the sum.
- We can bound the largest singular value of a sum of random rectangular matrices.
- Related arguments apply to matrix martingales and other adapted sequences.

Indeed, the expression (1.1) captures the essence of many questions that arise in numerical applications of random matrix theory, including most of the research cited above.

Observe that (1.1) formally resembles the probability that a sum of real random variables exceeds a certain level. The *Laplace transform method*, attributed to Bernstein, is a particularly elegant system for producing tail bounds for sums of scalar random variables; see [34, 36] for accessible discussions. In a remarkable paper [3], Ahlswede and Winter show how to transport the Laplace transform method to the matrix setting. They establish that

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geq t\right\} \leq \inf_{\theta > 0} \left\{e^{-\theta t} \cdot \mathbb{E} \operatorname{tr} \exp\left(\sum_k \theta \mathbf{X}_k\right)\right\}. \tag{1.2}$$

In words, the probability (1.1) is controlled by a matrix version of the moment-generating function (mgf). See Proposition 3.1 for an easy proof of (1.2) that is due to Oliveira [40].

The matrix Laplace transform estimate (1.2) presents a serious technical challenge. We must control the trace of the matrix mgf

$$\mathbb{E} \operatorname{tr} \exp\left(\sum_k \theta \mathbf{X}_k\right)$$

using information about the summands $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \dots$. This estimate requires powerful tools, and it stands as the major impediment to bounding the tail probability (1.1).

The main technical contribution of the Ahlswede–Winter argument [3, Appendix] consists in their technique for computing the required bounds on the matrix mgf. We describe their method in Sect. 3.7. The following probability inequality for a matrix Gaussian series is typical of the results that emerge from their approach. Let $\{\mathbf{A}_k\}$ be a family of fixed self-adjoint matrices with dimension d , and let $\{\gamma_k\}$ be a sequence of independent standard normal variables. Then

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \gamma_k \mathbf{A}_k\right) \geq t\right\} \leq d \cdot e^{-t^2/2\sigma_{\text{AW}}^2} \quad \text{where } \sigma_{\text{AW}}^2 := \sum_k \lambda_{\max}(\mathbf{A}_k^2). \tag{1.3}$$

The Ahlswede–Winter apparatus leads to a collection of other interesting probability inequalities; see Sect. 1.3 for references. Nevertheless, tail bounds developed in this fashion, including (1.3), are usually very far from optimal. See Sect. 3.7 and Sect. 4.8 for further discussion of this point.

This paper describes a more satisfactory framework for completing the bound on the matrix mgf. The crucial new ingredient in our argument is a deep theo-

rem [29, Theorem 6] of Lieb from his seminal paper on convex trace functions. We introduce Lieb's theorem in Sect. 3.4, and we explain how to combine this result with the matrix Laplace transform technique. We use this scheme to obtain a large family of probability inequalities that are essentially sharp in a wide variety of situations.

Our approach represents a dramatic advance beyond the Ahlswede–Winter technique. For example, our method delivers the following bound for a matrix Gaussian series:

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \gamma_k \mathbf{A}_k\right) \geq t\right\} \leq d \cdot e^{-t^2/2\sigma^2} \quad \text{where } \sigma^2 := \lambda_{\max}\left(\sum_k \mathbf{A}_k^2\right). \quad (1.4)$$

The estimate (1.4) offers a fundamental advantage over (1.3) because the variance parameter σ^2 is often d times smaller than σ_{AW}^2 . Furthermore, the discussion in Sect. 4 demonstrates that the inequality (1.4) cannot be sharpened without changing its structure. This improvement is typical of results constructed from our blueprint.

1.2 Index of Inequalities

This work contains a large number of bounds for the probability (1.1). The precise form of each inequality depends on prior information about the summands. As a service to the reader, we have collected the most useful results in this section. We have also included a short qualitative discussion of each bound, along with the location in the paper where the full treatment appears.

1.2.1 Notation

The symbol \preceq denotes the semidefinite order on self-adjoint matrices. The maps λ_{\min} and λ_{\max} return the algebraically smallest and largest eigenvalue of a self-adjoint matrix. We write $\|\cdot\|$ for the spectral norm, which equals the largest singular value of a matrix.

1.2.2 Main Results for Positive Semidefinite Matrices

In classical probability theory, one of the most famous concentration results concerns the number of successes in a sequence of independent random trials. This quantity can be expressed as a sum of independent, bounded random variables. Chernoff's large-deviation theorem [9] provides explicit estimates on the probability that this type of series is greater than (or smaller than) a specified level.

In the matrix setting, the analogous theorem concerns a sum of positive semidefinite random matrices subject to a uniform eigenvalue bound. The matrix Chernoff inequality shows that the extreme eigenvalues of the matrix series have the same binomial-type behavior that occurs in the scalar case.

Theorem 1.1 (Matrix Chernoff) *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, self-adjoint matrices with dimension d . Assume that each random matrix satisfies*

$$\mathbf{X}_k \succeq \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\mathbf{X}_k) \leq R \quad \text{almost surely.}$$

Define

$$\mu_{\min} := \lambda_{\min} \left(\sum_k \mathbb{E} \mathbf{X}_k \right) \quad \text{and} \quad \mu_{\max} := \lambda_{\max} \left(\sum_k \mathbb{E} \mathbf{X}_k \right).$$

Then

$$\mathbb{P} \left\{ \lambda_{\min} \left(\sum_k \mathbf{X}_k \right) \leq (1 - \delta) \mu_{\min} \right\} \leq d \cdot \left[\frac{e^{-\delta}}{(1 - \delta)^{1-\delta}} \right]^{\mu_{\min}/R} \quad \text{for } \delta \in [0, 1], \quad \text{and}$$

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_k \mathbf{X}_k \right) \geq (1 + \delta) \mu_{\max} \right\} \leq d \cdot \left[\frac{e^{\delta}}{(1 + \delta)^{1+\delta}} \right]^{\mu_{\max}/R} \quad \text{for } \delta \geq 0.$$

Chernoff bounds are well suited to studying the spectrum of a random matrix with independent columns. For additional details and related inequalities, turn to Sect. 5.

1.2.3 Main Results for Self-Adjoint Matrices

Another basic example of concentration is provided by a sum of real numbers modulated by independent standard normal variables or, alternatively, by independent Rademacher¹ random variables. A classical result shows that this type of random series exhibits sub-Gaussian tails. When we replace the real numbers by self-adjoint random matrices, we discover that the maximum and minimum eigenvalues of the matrix sum retain this normal tail behavior.

Theorem 1.2 (Matrix Gaussian and Rademacher series) *Consider a finite sequence $\{\mathbf{A}_k\}$ of fixed, self-adjoint matrices with dimension d , and let $\{\xi_k\}$ be a finite sequence of independent standard normal or independent Rademacher random variables. Then, for all $t \geq 0$,*

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_k \xi_k \mathbf{A}_k \right) \geq t \right\} \leq d \cdot e^{-t^2/2\sigma^2} \quad \text{where } \sigma^2 := \left\| \sum_k \mathbf{A}_k^2 \right\|.$$

Theorem 1.2 was first established explicitly by Oliveira using a different method [40]. We have included the result here because it is very important and because it follows from a mechanical application of our techniques. Turn to Sect. 4 for an exhaustive discussion of matrix Gaussian series. This presentation also describes several new phenomena that arise when we translate scalar inequalities to the matrix setting.

The Hoeffding inequality is a more general result that describes a sum of independent, zero-mean random variables that are subject to upper and lower bounds; it demonstrates that this random series exhibits normal concentration. We can extend this result to the matrix setting by considering random matrices that satisfy semidefinite upper bounds. In the matrix case, the maximum and minimum eigenvalues of the sum also have sub-Gaussian behavior.

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

Theorem 1.3 (Matrix Hoeffding) *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, self-adjoint matrices with dimension d , and let $\{\mathbf{A}_k\}$ be a sequence of fixed self-adjoint matrices. Assume that each random matrix satisfies*

$$\mathbb{E} \mathbf{X}_k = \mathbf{0} \quad \text{and} \quad \mathbf{X}_k^2 \preceq \mathbf{A}_k^2 \quad \text{almost surely.}$$

Then, for all $t \geq 0$,

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_k \mathbf{X}_k \right) \geq t \right\} \leq d \cdot e^{-t^2/8\sigma^2} \quad \text{where } \sigma^2 := \left\| \sum_k \mathbf{A}_k^2 \right\|.$$

The constant $1/8$ in Theorem 1.3 can be improved when there is additional information available. See Sect. 7 for a discussion and some related results for martingales.

In fact, a sum of independent, bounded random variables may vary substantially less than the Hoeffding bound suggests. A famous inequality of Bernstein demonstrates that this type of random series exhibits normal concentration near its mean on a scale determined by the variance of the sum. On the other hand, the tail of the sum decays subexponentially on a scale controlled by a uniform upper bound on the summands. Sums of independent random matrices exhibit the same type of behavior, where the normal concentration depends on a matrix generalization of the variance and the tails are controlled by a uniform bound on the maximum eigenvalue of each summand.

Theorem 1.4 (Matrix Bernstein) *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, self-adjoint matrices with dimension d . Assume that each random matrix satisfies*

$$\mathbb{E} \mathbf{X}_k = \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\mathbf{X}_k) \leq R \quad \text{almost surely.}$$

Then, for all $t \geq 0$,

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_k \mathbf{X}_k \right) \geq t \right\} \leq d \cdot \exp \left(\frac{-t^2/2}{\sigma^2 + Rt/3} \right) \quad \text{where } \sigma^2 := \left\| \sum_k \mathbb{E}(\mathbf{X}_k^2) \right\|.$$

Independently, Oliveira has established a somewhat weaker version of Theorem 1.4 using alternative techniques [39]. The reader is probably aware that the probability literature contains a huge number of results that extend Bernstein's inequality to include other a priori information on the summands, such as bounds on the rate of moment growth. Section 6 contains additional matrix probability inequalities of this species.

1.2.4 Main Results for Rectangular Matrices

As an immediate corollary of our results for self-adjoint random matrices, we can also establish a collection of inequalities for the maximum singular value of a sum of random rectangular matrices. In each case, we extend the result to rectangular matrices

by using a device from operator theory called the *self-adjoint dilation* (Sect. 2.6). Remark 3.11 and Sect. 4.2 offer some discussion of this technique. This section presents two of the most important inequalities for sums of random rectangular matrices.

As in the self-adjoint case, the norm of a Gaussian or Rademacher series with rectangular matrix coefficients has sub-Gaussian tails. This result follows directly from Theorem 1.2; see Sect. 4.2 for a complete proof. Observe that the variance parameter changes to reflect the fact that the row and column spaces of a general matrix are independent from each other; the variance can be viewed as a noncommutative “sum of squares.”

Theorem 1.5 (Matrix Gaussian and Rademacher series: rectangular case) *Consider a finite sequence $\{\mathbf{B}_k\}$ of fixed matrices with dimension $d_1 \times d_2$, and let $\{\xi_k\}$ be a finite sequence of independent standard normal or independent Rademacher random variables. Define the variance parameter*

$$\sigma^2 := \max \left\{ \left\| \sum_k \mathbf{B}_k \mathbf{B}_k^* \right\|, \left\| \sum_k \mathbf{B}_k^* \mathbf{B}_k \right\| \right\}.$$

Then, for all $t \geq 0$,

$$\mathbb{P} \left\{ \left\| \sum_k \xi_k \mathbf{B}_k \right\| \geq t \right\} \leq (d_1 + d_2) \cdot e^{-t^2/2\sigma^2}.$$

We can also develop a rectangular version of the matrix Bernstein inequality. Notice the parallel between the variance parameter here and the variance parameter for a rectangular Gaussian series. This result is an immediate corollary of Theorem 1.4; a proof sketch appears in Remark 6.3.

Theorem 1.6 (Matrix Bernstein: rectangular case) *Consider a finite sequence $\{\mathbf{Z}_k\}$ of independent, random matrices with dimensions $d_1 \times d_2$. Assume that each random matrix satisfies*

$$\mathbb{E} \mathbf{Z}_k = \mathbf{0} \quad \text{and} \quad \|\mathbf{Z}_k\| \leq R \quad \text{almost surely.}$$

Define

$$\sigma^2 := \max \left\{ \left\| \sum_k \mathbb{E}(\mathbf{Z}_k \mathbf{Z}_k^*) \right\|, \left\| \sum_k \mathbb{E}(\mathbf{Z}_k^* \mathbf{Z}_k) \right\| \right\}.$$

Then, for all $t \geq 0$,

$$\mathbb{P} \left\{ \left\| \sum_k \mathbf{Z}_k \right\| \geq t \right\} \leq (d_1 + d_2) \cdot \exp \left(\frac{-t^2/2}{\sigma^2 + Rt/3} \right).$$

We trust that the reader can develop other probability inequalities for rectangular matrices as needed. For brevity, we have omitted further examples.

1.2.5 Inequalities for Matrix Martingales

The techniques in this paper also lead directly to some simple results for matrix martingales. This material appears in Sect. 7.

Azuma Inequality: The Azuma inequality is the martingale extension of the Hoeffding inequality.

McDiarmid Inequality: The McDiarmid bounded difference inequality concerns matrix-valued functions of a family of independent random variables. It demonstrates that the extreme eigenvalues of the matrix-valued function exhibit normal concentration.

For more refined martingale inequalities, see the papers [39, 55] and the technical report [57].

1.3 Summary of Related Work

We continue with an overview of some related work on finite-dimensional random matrices. The first group of papers relies on the matrix extension of the Laplace transform method; the second group uses noncommutative moment inequalities.

1.3.1 The Matrix Laplace Transform Method

The most important precedent for our work is the influential paper of Ahlswede and Winter [3]. They are responsible for developing the matrix version of the Laplace transform method, which shows that the tail probability (1.1) is controlled by a matrix generalization of the mgf. They describe an iterative argument, based on the Golden–Thompson inequality, (2.6) below, that allows them to provide a weak bound for the mgf of a sum of independent random matrices in terms of mgf bounds for the individual summands. In particular, they apply this technique to obtain an extension of the Chernoff inequality [3, Theorem 19].

The Ahlswede–Winter method for bounding the matrix mgf is quite general. Several other authors have exploited their technique to obtain matrix extensions of classical probability inequalities. Christofides and Markström establish a matrix version of the Azuma and Hoeffding inequalities [10]. Gross [20, Theorem 6] and Recht [45, Theorem 3.2] develop two different matrix extensions of Bernstein’s inequality. We also refer the reader to Vershynin’s note [58], which offers a self-contained introduction to the Ahlswede–Winter circle of ideas.

Results established within the Ahlswede–Winter framework are often sharp for sums of i.i.d. random matrices, but the inequalities are far less accurate when applied to other types of sums. Roughly speaking, the tail bounds have the correct shape, but the method often leads to poor estimates for the quantity that controls the scale of large deviations. For a specific example, compare the variance parameter in (1.3) with the (correct) variance parameter appearing in (1.4). All the results we have mentioned so far have this shortcoming. See Sect. 3.7 for technical details.

Very recently, Oliveira has developed two notable variations [39, 40] on the Ahlswede–Winter method for bounding the matrix mgf. These techniques can sometimes identify the correct matrix generalization of the scale parameter. In particular,

the approach in [40] can be used to prove Theorem 1.2. Oliveira has also developed a version of the matrix Bernstein inequality [39, Theorem 1.2] that is similar to Theorem 1.4; his proof involves a matrix extension of the martingale techniques from [17].

The current article was inspired by the work of Ahlswede–Winter [3] and Oliveira [40]. Our results were obtained independently from Oliveira’s paper [39].

1.3.2 Noncommutative Moment Inequalities

There is another contemporary line of research that uses noncommutative (nc) moment inequalities to study random matrices. In a significant article [46], Rudelson obtains an optimal estimate for the sample complexity of approximating the covariance matrix of a general isotropic distribution. The argument in his paper, which is due to Pisier, depends on a version of the nc Khintchine inequality [31, 32, 44].

Rudelson’s technique has been applied widely over the last ten years, and it has emerged as a valuable tool for studying discrete random matrices. For example, the method can be used to provide bounds on the norm of a random submatrix [48, Theorem 1.8] drawn from a fixed matrix. However, it seems likely that matrix probability inequalities will replace the nc Khintchine inequality for many applications because they are easier to use and often produce better results.

By now, there is a substantial literature on other nc moment inequalities. The article [26] contains a reasonably accessible and comprehensive discussion. Some of these results have been applied to the study of random matrices; see [27] for an example. As we discuss in Sect. 4.7, nc moment bounds can also be combined with the matrix Laplace transform method because they sometimes provide an alternative way to control the matrix mgf.

1.4 Roadmap

The rest of the paper is organized as follows. Section 2 introduces the background results required for our proofs. Section 3 proves the main technical results that lead to probability inequalities for sums of independent random matrices. Section 4 uses Gaussian series as a case study to illustrate the main features of matrix probability inequalities and to argue that the bounds in this paper are structurally optimal. We develop the matrix Chernoff and Bernstein inequalities in Sects. 5–6. Finally, we establish some simple martingale results in Sect. 7.

2 Algebra, Analysis, and Probability with Matrices

This section provides a short introduction to the background we require for our proofs. The proofs contain detailed cross-references to this material, so the reader may wish to proceed directly to the main thread of the argument in Sect. 3.

Most of these results can be located in Bhatia’s books on matrix analysis [4, 5]. The works of Horn and Johnson [22, 23] also serve as good general references. Higham’s book [21] is an excellent source for information about matrix functions.

2.1 Conventions on Matrices

A *matrix* is a finite, two-dimensional array of complex numbers. *In this paper, all matrices are square unless otherwise noted.* We add the qualification *rectangular* when we need to refer to a general array, which may be square or nonsquare. Many parts of the discussion do not depend on the size of a matrix, so we specify dimensions only when it matters. In particular, we usually do not state the size of a matrix when it is determined by the context.

Several abbreviations are ubiquitous. Instead of self-adjoint, we often write *s.a.* Positive semidefinite becomes *psd*, and we shorten positive definite to *pd*.

We write $\mathbf{0}$ for the zero matrix and \mathbf{I} for the identity matrix. The matrix \mathbf{E}_{ij} has a unit entry in the (i, j) position and zeros elsewhere. The symbol \mathbf{Q} is reserved for a unitary matrix. We adopt Parlett's convention [41] that bold capital letters symmetric about the vertical axis ($\mathbf{A}, \dots, \mathbf{Y}$ and $\mathbf{\Delta}, \dots, \mathbf{\Omega}$) refer to s.a. matrices.

The symbols λ_{\min} and λ_{\max} refer to the algebraic minimum and maximum eigenvalues of an s.a. matrix. We use curly inequalities to denote the semidefinite ordering: $\mathbf{A} \succcurlyeq \mathbf{0}$ means that \mathbf{A} is psd. The symbol $\|\cdot\|$ always refers to the ℓ_2 vector norm or the associated operator norm, which is called the *spectral norm* because it returns the maximum singular value of its argument.

2.2 Conventions on Probability

We prefer to avoid unnecessary abstraction and technical detail, so we frame the standing assumption that all random variables are sufficiently regular so that we are justified in computing expectations, interchanging limits, and so forth. Furthermore, we often state that a random variable satisfies some relation and omit the qualification “almost surely.” We reserve the symbols \mathbf{X}, \mathbf{Y} for random s.a. matrices.

2.3 Matrix Functions

Consider a function $f : \mathbb{R} \rightarrow \mathbb{R}$. We define a map on diagonal matrices by applying the function to each diagonal entry. We then extend f to a function on s.a. matrices using the eigenvalue decomposition:

$$f(\mathbf{A}) := \mathbf{Q} \cdot f(\mathbf{\Lambda}) \cdot \mathbf{Q}^* \quad \text{where } \mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^*. \quad (2.1)$$

The *spectral mapping theorem* states that each eigenvalue of $f(\mathbf{A})$ is equal to $f(\lambda)$ for some eigenvalue λ of \mathbf{A} . This point is obvious from our definition.

Standard inequalities for real functions typically *do not* have parallel versions that hold for the semidefinite ordering. Nevertheless, there is one type of relation for real functions that always extends to the semidefinite setting:

$$\begin{aligned} f(a) \leq g(a) \quad \text{for } a \in I &\implies f(\mathbf{A}) \preceq g(\mathbf{A}) \\ &\text{when the eigenvalues of } \mathbf{A} \text{ lie in } I. \end{aligned} \quad (2.2)$$

We sometimes refer to (2.2) as the *transfer rule*.

2.4 The Matrix Exponential

The exponential of an s.a. matrix \mathbf{A} can be defined by applying (2.1) with the function $f(x) = e^x$. Alternatively, we may use the power series expansion

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

The exponential of an s.a. matrix is always pd from the spectral mapping theorem. On account of the transfer rule (2.2), the matrix exponential satisfies some simple semidefinite relations that we collect here. For each s.a. matrix \mathbf{A} , it holds that

$$\mathbf{I} + \mathbf{A} \preceq e^{\mathbf{A}}, \quad \text{and} \tag{2.3}$$

$$\cosh(\mathbf{A}) \preceq e^{\mathbf{A}^2/2}. \tag{2.4}$$

We often work with the trace of the matrix exponential, $\text{tr exp} : \mathbf{A} \mapsto \text{tr } e^{\mathbf{A}}$. The trace exponential function is convex. It is also monotone with respect to the semidefinite order:

$$\mathbf{A} \preceq \mathbf{H} \implies \text{tr } e^{\mathbf{A}} \leq \text{tr } e^{\mathbf{H}}. \tag{2.5}$$

See [43, Sect. 2] for short proofs of these facts.

The matrix exponential *does not* convert sums into products, but the trace exponential has a related property that serves as a limited substitute. The Golden–Thompson inequality [4, Sect. IX.3] states that

$$\text{tr } e^{\mathbf{A}+\mathbf{H}} \leq \text{tr}(e^{\mathbf{A}}e^{\mathbf{H}}) \quad \text{for all s.a. } \mathbf{A}, \mathbf{H}. \tag{2.6}$$

The obvious generalization of the bound (2.6) to three matrices is false [4, Prob. IX.8.4].

2.5 The Matrix Logarithm

We define the matrix logarithm as the functional inverse of the matrix exponential:

$$\log(e^{\mathbf{A}}) := \mathbf{A} \quad \text{for each s.a. matrix } \mathbf{A}. \tag{2.7}$$

This formula determines the logarithm on the pd cone, which is adequate for our purposes.

The matrix logarithm interacts beautifully with the semidefinite order [5, Exercise 4.2.5]. Indeed, the logarithm is operator monotone:

$$\mathbf{0} \prec \mathbf{A} \preceq \mathbf{H} \implies \log(\mathbf{A}) \preceq \log(\mathbf{H}). \tag{2.8}$$

The logarithm is also operator concave:

$$\tau \log(\mathbf{A}) + (1 - \tau) \log(\mathbf{H}) \preceq \log(\tau \mathbf{A} + (1 - \tau)\mathbf{H}) \quad \text{for all pd } \mathbf{A}, \mathbf{H} \text{ and } \tau \in [0, 1]. \tag{2.9}$$

Caveat lector: Operator monotone functions and operator convex functions are depressingly rare. In particular, the matrix exponential does not belong to either class [4, Chap. V].

2.6 Dilations

An extraordinarily fruitful idea from operator theory is to embed matrices within larger block matrices, called *dilations* [42]. The *s.a. dilation* of a rectangular matrix \mathbf{B} is

$$\mathcal{S}(\mathbf{B}) := \begin{bmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{0} \end{bmatrix}. \quad (2.10)$$

Evidently, $\mathcal{S}(\mathbf{B})$ is always s.a. A short calculation yields the important identity

$$\mathcal{S}(\mathbf{B})^2 = \begin{bmatrix} \mathbf{B}\mathbf{B}^* & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^*\mathbf{B} \end{bmatrix}. \quad (2.11)$$

It can also be verified that the s.a. dilation preserves spectral information:

$$\lambda_{\max}(\mathcal{S}(\mathbf{B})) = \|\mathcal{S}(\mathbf{B})\| = \|\mathbf{B}\|. \quad (2.12)$$

We use dilations to extend results for s.a. matrices to rectangular matrices. See Remark 3.11 and Sect. 4.2 for more information about this technique.

2.7 Expectation and the Semidefinite Order

Since the expectation of a random matrix can be viewed as a convex combination and the psd cone is convex, expectation preserves the semidefinite order:

$$\mathbf{X} \preceq \mathbf{Y} \quad \text{almost surely} \quad \implies \quad \mathbb{E}\mathbf{X} \preceq \mathbb{E}\mathbf{Y}. \quad (2.13)$$

Every operator convex function admits an operator Jensen's inequality [25]. In particular, the matrix square is operator convex, which implies that

$$(\mathbb{E}\mathbf{X})^2 \preceq \mathbb{E}(\mathbf{X}^2). \quad (2.14)$$

The relation (2.14) is also a specific instance of Kadison's inequality [5, Theorem 2.3.2].

3 Tail Bounds via the Matrix Laplace Transform Method

This section develops some general probability inequalities for the maximum eigenvalue of a sum of independent random matrices. The main argument can be viewed as a matrix extension of the Laplace transform method for sums of independent real random variables. However, in the matrix setting, it requires great care to execute this technique successfully.

3.1 Matrix Moments and Cumulants

Consider a random s.a. matrix \mathbf{X} that has moments of all orders. By analogy with the classical scalar definitions, we may construct matrix extensions of the moment-generating function (mgf) and the cumulant-generating function (cgf):

$$\mathbf{M}_{\mathbf{X}}(\theta) := \mathbb{E} e^{\theta \mathbf{X}} \quad \text{and} \quad \mathbf{\Xi}_{\mathbf{X}}(\theta) := \log \mathbb{E} e^{\theta \mathbf{X}} \quad \text{for } \theta \in \mathbb{R}. \tag{3.1}$$

We admit the possibility that these expectations do not exist for all values of θ . The matrix cgf can be viewed as an *exponential mean*, a weighted average that emphasizes large deviations (with the same sign as θ). The matrix mgf and cgf have formal power series expansions:

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

The coefficients $\mathbb{E}(\mathbf{X}^p)$ are called *matrix moments*, and we refer to $\mathbf{\Psi}_p$ as a *matrix cumulant*. The matrix cumulant $\mathbf{\Psi}_p$ 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:

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

Higher-order cumulants are harder to write down and interpret.

3.2 The Laplace Transform Method for Matrices

We begin our main development with a striking idea drawn from the influential paper [3] of Ahlswede and Winter. Their work contains a matrix analog of the classical Laplace transform bound. We need the following variant, which is due to Oliveira [40].

Proposition 3.1 (The Laplace Transform Method) *Let \mathbf{Y} be a random self-adjoint matrix. For all $t \in \mathbb{R}$,*

$$\mathbb{P}\{\lambda_{\max}(\mathbf{Y}) \geq t\} \leq \inf_{\theta > 0} \{e^{-\theta t} \cdot \mathbb{E} \operatorname{tr} e^{\theta \mathbf{Y}}\}.$$

In words, we can control tail probabilities for the maximum eigenvalue of a random matrix by producing a bound for the trace of the matrix mgf defined in (3.1).

Proof Fix a positive number θ . We have the chain of relations

$$\mathbb{P}\{\lambda_{\max}(\mathbf{Y}) \geq t\} = \mathbb{P}\{\lambda_{\max}(\theta \mathbf{Y}) \geq \theta t\} = \mathbb{P}\{e^{\lambda_{\max}(\theta \mathbf{Y})} \geq e^{\theta t}\} \leq e^{-\theta t} \cdot \mathbb{E} e^{\lambda_{\max}(\theta \mathbf{Y})}.$$

The first identity uses the homogeneity of the maximum eigenvalue map, and the second relies on the monotonicity of the scalar exponential function; the third relation is Markov’s inequality. To bound the exponential, note that

$$e^{\lambda_{\max}(\theta \mathbf{Y})} = \lambda_{\max}(e^{\theta \mathbf{Y}}) \leq \operatorname{tr} e^{\theta \mathbf{Y}}.$$

The identity is the spectral mapping theorem; the inequality holds because the exponential of an s.a. matrix is pd and the maximum eigenvalue of a pd matrix is dominated by the trace. Combine the latter two relations to reach

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

This inequality holds for any positive θ , so we may take an infimum to complete the proof. □

3.3 The Failure of the Matrix mgf

In the scalar setting, the Laplace transform method is very effective for studying sums of independent random variables because the mgf decomposes. Consider an independent sequence $\{X_k\}$ of real random variables. Operating formally, we see that the (scalar) mgf of the sum satisfies a multiplication rule:

$$M_{(\sum_k X_k)}(\theta) = \mathbb{E} \exp\left(\sum_k \theta X_k\right) = \mathbb{E} \prod_k e^{\theta X_k} = \prod_k \mathbb{E} e^{\theta X_k} = \prod_k M_{X_k}(\theta). \quad (3.2)$$

This calculation relies on the fact that the scalar exponential function converts sums to products, a property the matrix exponential does not share. As a consequence, there is no immediate analog of (3.2) in the matrix setting.

Ahlsvede and Winter attempt to imitate the multiplication rule (3.2) using the following observation. When \mathbf{X}_1 and \mathbf{X}_2 are independent random matrices,

$$\operatorname{tr} \mathbf{M}_{\mathbf{X}_1 + \mathbf{X}_2}(\theta) \leq \mathbb{E} \operatorname{tr}[e^{\theta \mathbf{X}_1} e^{\theta \mathbf{X}_2}] = \operatorname{tr}[(\mathbb{E} e^{\theta \mathbf{X}_1})(\mathbb{E} e^{\theta \mathbf{X}_2})] = \operatorname{tr}[\mathbf{M}_{\mathbf{X}_1}(\theta) \cdot \mathbf{M}_{\mathbf{X}_2}(\theta)]. \quad (3.3)$$

The first relation is the Golden–Thompson trace inequality (2.6). Unfortunately, we cannot extend the bound (3.3) to include additional matrices. This cold fact suggests that the Golden–Thompson inequality may not be the natural way to proceed. In Sect. 3.7, we map out the route Ahlsvede and Winter pursue, but we continue along a different path.

3.4 A Concave Trace Function

For inspiration, we turn to the literature on matrix analysis. Some of the most beautiful and profound results in this domain concern the convexity of trace functions. We have observed that this theory has incredible implications for the study of random matrices. This paper demonstrates that a large class of matrix probability inequalities follows from a deep theorem [29, Theorem 6] of Lieb that appears in his seminal work on convex trace functions.

Theorem 3.2 (Lieb) *Fix a self-adjoint matrix \mathbf{H} . The function*

$$\mathbf{A} \mapsto \operatorname{tr} \exp(\mathbf{H} + \log(\mathbf{A}))$$

is concave on the positive definite cone.

Epstein provides an alternative proof of Theorem 3.2 in [16, Sect. II], and Ruskai offers a simplified account of Epstein’s argument in [47]. The note [56] derives Lieb’s theorem from the joint convexity of quantum relative entropy [30, Lemma 2]. The latter approach is advantageous because the joint convexity result admits several elegant, conceptual proofs, such as [15, Corollary 2.2].

We require a simple but powerful corollary of Lieb’s theorem. This result describes how expectation interacts with the trace exponential.

Corollary 3.3 *Let \mathbf{H} be a fixed self-adjoint matrix, and let \mathbf{X} be a random self-adjoint matrix. Then*

$$\mathbb{E} \operatorname{tr} \exp(\mathbf{H} + \mathbf{X}) \leq \operatorname{tr} \exp(\mathbf{H} + \log(\mathbb{E} e^{\mathbf{X}})).$$

Proof Define the random matrix $\mathbf{Y} = e^{\mathbf{X}}$, and calculate that

$$\begin{aligned} \mathbb{E} \operatorname{tr} \exp(\mathbf{H} + \mathbf{X}) &= \mathbb{E} \operatorname{tr} \exp(\mathbf{H} + \log(\mathbf{Y})) \leq \operatorname{tr} \exp(\mathbf{H} + \log(\mathbb{E} \mathbf{Y})) \\ &= \operatorname{tr} \exp(\mathbf{H} + \log(\mathbb{E} e^{\mathbf{X}})). \end{aligned}$$

The first identity follows from the definition (2.7) of the matrix logarithm because \mathbf{Y} is always pd. Lieb’s result, Theorem 3.2, ensures that the trace function is concave in \mathbf{Y} , so we may invoke Jensen’s inequality to draw the expectation inside the logarithm. □

3.5 Subadditivity of the Matrix cgf

Let us return to the problem of bounding the matrix mgf of an independent sum. Although the multiplication rule (3.2) is a dead end in the matrix case, the scalar cgf has a related property that submits to generalization. For an independent family $\{X_k\}$ of real random variables, the scalar cgf is additive:

$$\Xi_{(\sum_k X_k)}(\theta) = \log \mathbb{E} \exp\left(\sum_k \theta X_k\right) = \sum_k \log \mathbb{E} e^{\theta X_k} = \sum_k \Xi_{X_k}(\theta), \tag{3.4}$$

where the second identity follows from (3.2) when we take logarithms.

Our key insight is that Corollary 3.3 offers a completely satisfactory way to extend the addition rule (3.4) for scalar cgf’s to the matrix setting. We have the following result.

Lemma 3.4 (Subadditivity of Matrix cgf’s) *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, self-adjoint matrices. Then*

$$\mathbb{E} \operatorname{tr} \exp\left(\sum_k \theta \mathbf{X}_k\right) \leq \operatorname{tr} \exp\left(\sum_k \log \mathbb{E} e^{\theta \mathbf{X}_k}\right) \quad \text{for } \theta \in \mathbb{R}.$$

Proof It does no harm to assume $\theta = 1$. Let \mathbb{E}_k denote the expectation, conditioned on $\mathbf{X}_1, \dots, \mathbf{X}_k$. Abbreviate

$$\Xi_k := \log(\mathbb{E}_{k-1} e^{\mathbf{X}_k}) = \log(\mathbb{E} e^{\mathbf{X}_k}),$$

where the equality holds because the family $\{\mathbf{X}_k\}$ is independent. We see that

$$\begin{aligned} \mathbb{E} \operatorname{tr} \exp \left(\sum_{k=1}^n \mathbf{X}_k \right) &= \mathbb{E}_0 \cdots \mathbb{E}_{n-1} \operatorname{tr} \exp \left(\sum_{k=1}^{n-1} \mathbf{X}_k + \mathbf{X}_n \right) \\ &\leq \mathbb{E}_0 \cdots \mathbb{E}_{n-2} \operatorname{tr} \exp \left(\sum_{k=1}^{n-1} \mathbf{X}_k + \log(\mathbb{E}_{n-1} e^{\mathbf{X}_n}) \right) \\ &= \mathbb{E}_0 \cdots \mathbb{E}_{n-2} \operatorname{tr} \exp \left(\sum_{k=1}^{n-2} \mathbf{X}_k + \mathbf{X}_{n-1} + \mathbf{\Xi}_n \right) \\ &\leq \mathbb{E}_0 \cdots \mathbb{E}_{n-3} \operatorname{tr} \exp \left(\sum_{k=1}^{n-2} \mathbf{X}_k + \mathbf{\Xi}_{n-1} + \mathbf{\Xi}_n \right) \\ &\vdots \\ &\leq \operatorname{tr} \exp \left(\sum_{k=1}^n \mathbf{\Xi}_k \right). \end{aligned}$$

The first line relies on the tower property of conditional expectation. At each step $m = 1, 2, \dots, n$, we invoke Corollary 3.3 with the fixed matrix \mathbf{H} equal to

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

This act is legal because \mathbf{H}_m does not depend on \mathbf{X}_m . □

Remark 3.5 To make the parallel with the addition rule (3.4) clearer, we can rewrite the conclusion of Lemma 3.4 in the form

$$\operatorname{tr} \exp(\mathbf{\Xi}_{(\sum_k \mathbf{X}_k)}(\theta)) \leq \operatorname{tr} \exp \left(\sum_k \mathbf{\Xi}_{\mathbf{X}_k}(\theta) \right)$$

by applying the definition (3.1) of the matrix cgf.

3.6 Tail Bounds for Independent Sums

This section contains abstract tail bounds for the sum of independent random matrices. Later, we will specialize these results to some specific situations. We begin with a very general inequality, which is the progenitor of our other results.

Theorem 3.6 (Master Tail Bound for Independent Sums) *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, self-adjoint matrices. For all $t \in \mathbb{R}$,*

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_k \mathbf{X}_k \right) \geq t \right\} \leq \inf_{\theta > 0} \left\{ e^{-\theta t} \cdot \operatorname{tr} \exp \left(\sum_k \log \mathbb{E} e^{\theta \mathbf{X}_k} \right) \right\}. \tag{3.5}$$

Proof Substitute the subadditivity rule for matrix cgf’s, Lemma 3.4, into the Laplace transform bound, Proposition 3.1. \square

Our first corollary adapts Theorem 3.6 to the case that arises most often in practice. We call upon this result several times to obtain tail bounds under a variety of assumptions about the structure of the random matrices.

Corollary 3.7 *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, self-adjoint matrices with dimension d . Assume there is a function $g : (0, \infty) \rightarrow [0, \infty)$ and a sequence $\{\mathbf{A}_k\}$ of fixed self-adjoint matrices that satisfy the relations*

$$\mathbb{E} e^{\theta \mathbf{X}_k} \preceq e^{g(\theta) \cdot \mathbf{A}_k} \quad \text{for } \theta > 0. \tag{3.6}$$

Define the scale parameter

$$\rho := \lambda_{\max} \left(\sum_k \mathbf{A}_k \right).$$

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

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_k \mathbf{X}_k \right) \geq t \right\} \leq d \cdot \inf_{\theta > 0} e^{-\theta t + g(\theta) \cdot \rho}. \tag{3.7}$$

Proof The hypothesis (3.6) implies that

$$\log \mathbb{E} e^{\theta \mathbf{X}_k} \preceq g(\theta) \cdot \mathbf{A}_k \quad \text{for } \theta > 0 \tag{3.8}$$

because of the property (2.8) that the matrix logarithm is operator monotone. Recall the fact (2.5) that the trace exponential is monotone with respect to the semidefinite order. As a consequence, we can introduce each relation from the family (3.8) into the master inequality (3.5). For each $\theta > 0$, it follows that

$$\begin{aligned} \mathbb{P} \left\{ \lambda_{\max} \left(\sum_k \mathbf{X}_k \right) \geq t \right\} &\leq e^{-\theta t} \cdot \text{tr} \exp \left(g(\theta) \cdot \sum_k \mathbf{A}_k \right) \\ &\leq e^{-\theta t} \cdot d \cdot \lambda_{\max} \left(\exp \left(g(\theta) \cdot \sum_k \mathbf{A}_k \right) \right) \\ &= d \cdot e^{-\theta t} \cdot \exp \left(g(\theta) \cdot \lambda_{\max} \left(\sum_k \mathbf{A}_k \right) \right). \end{aligned}$$

The second inequality holds because the trace of a pd matrix, such as the exponential, is bounded by the dimension d times the maximum eigenvalue. The last line depends on the spectral mapping theorem and the fact that the function g is non-negative. Identify the quantity ρ , and take the infimum over positive θ to reach the conclusion (3.7). \square

Remark 3.8 An alternative expression of the result (3.7) is that

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geq t\right\} \leq d \cdot \exp\left(-\sup_{\theta>0}\{\theta t - g(\theta) \cdot \rho\}\right) = d \cdot \exp(-\rho \cdot g^*(t/\rho)).$$

In words, the exponent in the tail bound can be written in terms of the perspective transformation of the Fenchel–Legendre conjugate of the function g . This inequality parallels the upper estimate in Cramér’s classical result for large deviations [14, Theorem 2.2.3].

It is also worthwhile to state another consequence of Theorem 3.6. This bound is sometimes more useful than Corollary 3.7 because it combines the mgf’s of the random matrices together under a single logarithm.

Corollary 3.9 *Consider a sequence $\{\mathbf{X}_k : k = 1, 2, \dots, n\}$ of independent, random, self-adjoint matrices with dimension d . For all $t \in \mathbb{R}$,*

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_{k=1}^n \mathbf{X}_k\right) \geq t\right\} \leq d \cdot \inf_{\theta>0} \exp\left(-\theta t + n \cdot \log \lambda_{\max}\left(\frac{1}{n} \sum_{k=1}^n \mathbb{E} e^{\theta \mathbf{X}_k}\right)\right). \tag{3.9}$$

Proof Recall the fact (2.9) that the matrix logarithm is operator concave. For each $\theta > 0$, it follows that

$$\sum_{k=1}^n \log \mathbb{E} e^{\theta \mathbf{X}_k} = n \cdot \frac{1}{n} \sum_{k=1}^n \log \mathbb{E} e^{\theta \mathbf{X}_k} \preceq n \cdot \log\left(\frac{1}{n} \sum_{k=1}^n \mathbb{E} e^{\theta \mathbf{X}_k}\right).$$

The property (2.5) that the trace exponential is monotone allows us to introduce the latter relation into the master inequality (3.5) to obtain

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_{k=1}^n \mathbf{X}_k\right) \geq t\right\} \leq e^{-\theta t} \cdot \text{tr exp}\left(n \cdot \log\left(\frac{1}{n} \sum_{k=1}^n \mathbb{E} e^{\theta \mathbf{X}_k}\right)\right).$$

To complete the proof, we bound the trace by d times the maximum eigenvalue, and we invoke the spectral mapping theorem (twice!) to draw the maximum eigenvalue map inside the logarithm. Take the infimum over positive θ to reach (3.9). \square

We conclude this section with remarks on some other situations that we can analyze using the master tail bound, Theorem 3.6, and its corollaries.

Remark 3.10 (Minimum Eigenvalue) We can study the minimum eigenvalue of a sum of random s.a. matrices because $\lambda_{\min}(\mathbf{X}) = -\lambda_{\max}(-\mathbf{X})$. As a result,

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_k \mathbf{X}_k\right) \leq t\right\} = \mathbb{P}\left\{\lambda_{\max}\left(\sum_k -\mathbf{X}_k\right) \geq -t\right\}.$$

In Sect. 5, we apply this observation to develop lower Chernoff bounds.

Remark 3.11 (Maximum Singular Value) We can also analyze the maximum singular value of a sum of random rectangular matrices by applying these results to the s.a. dilation (2.10). For a finite sequence $\{\mathbf{Z}_k\}$ of independent, random, rectangular matrices, we have

$$\mathbb{P}\left\{\left\|\sum_k \mathbf{Z}_k\right\| \geq t\right\} = \mathbb{P}\left\{\lambda_{\max}\left(\sum_k \mathcal{S}(\mathbf{Z}_k)\right) \geq t\right\}$$

on account of (2.12) and the property that the dilation is real-linear. This device allows us to extend most of the tail bounds in this paper to rectangular matrices. See Sect. 4 for an application to Gaussian and Rademacher series.

Remark 3.12 (Martingales) It is possible to combine the proofs of Lemma 3.4 and Theorem 3.6 to obtain some simple results for matrix martingales. See the demonstration of the matrix Azuma inequality in Sect. 7 for an example of this approach. To obtain fully detailed results for martingales, one must use a fundamentally different style of argument [39, 55].

3.7 The Ahlswede–Winter Method

Ahlsweide and Winter use a different approach to bound the matrix mgf, which exploits the multiplicative bound (3.3) for the trace exponential of a sum of two independent, random, s.a. matrices. The reader may find their argument interesting.

Consider a sequence $\{\mathbf{X}_k : k = 1, 2, \dots, n\}$ of independent, random, s.a. matrices with dimension d , and let $\mathbf{Y} = \sum_k \mathbf{X}_k$. The trace inequality (3.3) implies that

$$\text{tr} \mathbf{M}_{\mathbf{Y}}(\theta) \leq \text{tr}\left[\left(\mathbb{E} e^{\sum_{k=1}^{n-1} \theta \mathbf{X}_k}\right)\left(\mathbb{E} e^{\theta \mathbf{X}_n}\right)\right] \leq \text{tr}\left(\mathbb{E} e^{\sum_{k=1}^{n-1} \theta \mathbf{X}_k}\right) \cdot \lambda_{\max}\left(\mathbb{E} e^{\theta \mathbf{X}_n}\right).$$

Iterating this procedure leads to the relation

$$\text{tr} \mathbf{M}_{\mathbf{Y}}(\theta) \leq (\text{tr} \mathbf{I}) \cdot \left[\prod_k \lambda_{\max}\left(\mathbb{E} e^{\theta \mathbf{X}_k}\right)\right] = d \cdot \exp\left(\sum_k \lambda_{\max}\left(\log \mathbb{E} e^{\theta \mathbf{X}_k}\right)\right). \tag{3.10}$$

The bound (3.10) is the key to the Ahlswede–Winter method for producing probability inequalities. As a consequence, their approach generally leads to tail bounds that depend on a scale parameter involving “the sum of eigenvalues.” See, for example, the bound (1.3) or the matrix probability inequalities presented in the papers [3, 10, 20, 45].

In contrast, our result on the subadditivity of cumulants, Lemma 3.4, implies that

$$\text{tr} \mathbf{M}_{\mathbf{Y}}(\theta) \leq d \cdot \exp\left(\lambda_{\max}\left(\sum_k \log \mathbb{E} e^{\theta \mathbf{X}_k}\right)\right). \tag{3.11}$$

Probability inequalities developed with (3.11) contain a scale parameter that involves the “eigenvalue of a sum.” See, for example, the bound (1.4). The exponent in (3.10) often exceeds the exponent in (3.11) by a factor of d , the ambient dimension, which is a serious loss. Section 4.8 describes concrete situations where this discrepancy occurs.

4 Case Study: Matrix Gaussian Series

A matrix Gaussian series stands among the simplest instances of a sum of independent random matrices. Nevertheless, this example already exhibits several new phenomena that arise when we translate scalar tail bounds to the matrix setting. Consequently, we explore this fundamental case in depth as a way to develop insights about other matrix probability inequalities.

4.1 Main Results

We begin with the scalar case. Consider a finite sequence $\{a_k\}$ of real numbers and a finite sequence $\{\gamma_k\}$ of independent standard Gaussian variables. We have the probability inequality

$$\mathbb{P}\left\{\sum_k \gamma_k a_k \geq t\right\} \leq e^{-t^2/2\sigma^2} \quad \text{where } \sigma^2 := \sum_k a_k^2. \quad (4.1)$$

This result testifies that a Gaussian series with real coefficients satisfies a normal-type tail bound where the variance is controlled by the sum of the squared coefficients. The relation (4.1) follows easily from the scalar Laplace transform method. An alternative proof proceeds using the rotational invariance of a standard normal vector along with basic estimates on the error function.

The inequality (4.1) generalizes directly to the noncommutative setting, as do many other scalar tail bounds. The matrix Laplace transform method, Proposition 3.1, delivers the following result on the tail behavior of a matrix Gaussian series.

Theorem 4.1 (Matrix Gaussian and Rademacher Series) *Consider a finite sequence $\{\mathbf{A}_k\}$ of fixed self-adjoint matrices with dimension d , and let $\{\gamma_k\}$ be a finite sequence of independent standard normal variables. Compute the variance parameter*

$$\sigma^2 := \left\| \sum_k \mathbf{A}_k^2 \right\|. \quad (4.2)$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \gamma_k \mathbf{A}_k\right) \geq t\right\} \leq d \cdot e^{-t^2/2\sigma^2}. \quad (4.3)$$

In particular,

$$\mathbb{P}\left\{\left\| \sum_k \gamma_k \mathbf{A}_k \right\| \geq t\right\} \leq 2d \cdot e^{-t^2/2\sigma^2}. \quad (4.4)$$

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

Observe that the bound (4.3) reduces to the scalar result (4.1) when the dimension $d = 1$. Of course, one may wonder whether the generalization (4.2) of the scalar

variance is sharp and whether the dimensional dependence in (4.3) is necessary. A primary objective of this section is to demonstrate that Theorem 4.1 cannot be improved without changing its form.

Most of the inequalities in this paper have variants that concern the maximum singular value of a sum of rectangular random matrices. These extensions follow immediately when we apply the s.a. results to the s.a. dilation of the sum of rectangular matrices. Here is the general version of Theorem 4.1, which serves as a model for other rectangular results.

Corollary 4.2 (Rectangular Matrix Gaussian and Rademacher Series) *Consider a finite sequence $\{\mathbf{B}_k\}$ of fixed matrices with dimension $d_1 \times d_2$, and let $\{\gamma_k\}$ be a finite sequence of independent standard normal variables. Compute the variance parameter*

$$\sigma^2 := \max \left\{ \left\| \sum_k \mathbf{B}_k \mathbf{B}_k^* \right\|, \left\| \sum_k \mathbf{B}_k^* \mathbf{B}_k \right\| \right\}.$$

Then, for all $t \geq 0$,

$$\mathbb{P} \left\{ \left\| \sum_k \gamma_k \mathbf{B}_k \right\| \geq t \right\} \leq (d_1 + d_2) \cdot e^{-t^2/2\sigma^2}.$$

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

The proofs of Theorem 4.1 and Corollary 4.2 appear below in Sect. 4.2. Unlike our other results, these two bounds are not new. One established argument, which we discuss in Sect. 4.7, involves noncommutative Khintchine inequalities. It is also possible to prove these results using Oliveira’s ideas [40].

4.2 Proofs

We continue with a short demonstration of the main results for matrix Gaussian and Rademacher series. The first step is to obtain a semidefinite bound for the mgf of a fixed matrix modulated by a Gaussian variable or a Rademacher variable. This mgf bound essentially appears in Oliveira’s work [40, Lemma 2].

Lemma 4.3 (Rademacher and Gaussian mgf’s) *Suppose that \mathbf{A} is an s.a. matrix. Let ε be a Rademacher random variable, and let γ be a standard normal random variable. Then*

$$\mathbb{E} e^{\varepsilon\theta\mathbf{A}} \preceq e^{\theta^2\mathbf{A}^2/2} \quad \text{and} \quad \mathbb{E} e^{\gamma\theta\mathbf{A}} = e^{\theta^2\mathbf{A}^2/2} \quad \text{for } \theta \in \mathbb{R}.$$

Proof Absorbing θ into \mathbf{A} , we may assume $\theta = 1$ in each case. We begin with the Rademacher mgf. By direct calculation,

$$\mathbb{E} e^{\varepsilon\mathbf{A}} = \cosh(\mathbf{A}) \preceq e^{\mathbf{A}^2/2},$$

where the second relation is (2.4).

For the Gaussian case, recall that the moments of a standard normal variable satisfy

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

Therefore,

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

The first identity holds because the odd terms in the series vanish. □

The tail bounds for s.a. matrix Gaussian and Rademacher series follow easily.

Proof of Theorem 4.1 Let $\{\xi_k\}$ be a finite sequence of independent standard normal variables or independent Rademacher variables. Invoke Lemma 4.3 to obtain

$$\mathbb{E} e^{\xi_k \theta \mathbf{A}_k} \preceq e^{g(\theta) \cdot \mathbf{A}_k^2} \quad \text{where } g(\theta) := \theta^2/2 \text{ for } \theta > 0.$$

Recall that

$$\sigma^2 = \left\| \sum_k \mathbf{A}_k^2 \right\| = \lambda_{\max} \left(\sum_k \mathbf{A}_k^2 \right).$$

Corollary 3.7 delivers

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_k \xi_k \mathbf{A}_k \right) \geq t \right\} \leq d \cdot \inf_{\theta > 0} e^{-\theta t + g(\theta) \cdot \sigma^2} = d \cdot e^{-t^2/2\sigma^2}. \tag{4.5}$$

For the record, the infimum is attained when $\theta = t/\sigma^2$.

To obtain the norm bound (4.4), recall that $\|\mathbf{Y}\| = \max\{\lambda_{\max}(\mathbf{Y}), -\lambda_{\min}(\mathbf{Y})\}$. Standard Gaussian variables and Rademacher variables are symmetric, so the inequality (4.5) implies

$$\mathbb{P} \left\{ -\lambda_{\min} \left(\sum_k \xi_k \mathbf{A}_k \right) \geq t \right\} = \mathbb{P} \left\{ \lambda_{\max} \left(\sum_k (-\xi_k) \mathbf{A}_k \right) \geq t \right\} \leq d \cdot e^{-t^2/2\sigma^2}.$$

Apply the union bound to the estimates for λ_{\max} and $-\lambda_{\min}$ to complete the proof. □

The result for a series with rectangular matrix coefficients follows immediately when we apply Theorem 4.1 to the s.a. dilation of the series.

Proof of Corollary 4.2 Let $\{\xi_k\}$ be a finite sequence of independent standard normal random variables or independent Rademacher random variables. Consider the sequence $\{\xi_k \mathcal{S}(\mathbf{B}_k)\}$ of random s.a. matrices with dimension $d_1 + d_2$. The spectral identity (2.12) ensures that

$$\left\| \sum_k \xi_k \mathbf{B}_k \right\| = \lambda_{\max} \left(\mathcal{S} \left(\sum_k \xi_k \mathbf{B}_k \right) \right) = \lambda_{\max} \left(\sum_k \xi_k \mathcal{S}(\mathbf{B}_k) \right).$$

Thus, we may invoke Theorem 4.1 to obtain a probability inequality for the norm of the series. Simply observe that the matrix variance parameter (4.2) satisfies the relation

$$\begin{aligned} \sigma^2 &= \left\| \sum_k \mathcal{S}(\mathbf{B}_k)^2 \right\| = \left\| \begin{bmatrix} \sum_k \mathbf{B}_k \mathbf{B}_k^* & \mathbf{0} \\ \mathbf{0} & \sum_k \mathbf{B}_k^* \mathbf{B}_k \end{bmatrix} \right\| \\ &= \max \left\{ \left\| \sum_k \mathbf{B}_k \mathbf{B}_k^* \right\|, \left\| \sum_k \mathbf{B}_k^* \mathbf{B}_k \right\| \right\} \end{aligned}$$

on account of the identity (2.11) for the square of the s.a. dilation. □

4.3 Application: A Gaussian Matrix with Nonuniform Variances

It may not be immediately clear why abstract probability inequalities, such as Theorem 4.1 and Corollary 4.2, deliver information about interesting random matrices that arise in practice. Let us describe a simple application that speaks to this concern.

Fix a $d_1 \times d_2$ matrix \mathbf{B} , and draw a random $d_1 \times d_2$ matrix $\mathbf{\Gamma}$ whose entries are independent standard normal variables. Let \odot denote the componentwise (i.e., Schur or Hadamard) product of matrices. Construct the random matrix $\mathbf{\Gamma} \odot \mathbf{B}$, and observe that its (j, k) component is a Gaussian variable with mean zero and variance $|b_{jk}|^2$. We claim that

$$\begin{aligned} \mathbb{P}\{\|\mathbf{\Gamma} \odot \mathbf{B}\| \geq t\} &\leq (d_1 + d_2) \cdot e^{-t^2/2\sigma^2} \\ \text{where } \sigma^2 &= \max \left\{ \max_j \|\mathbf{b}_j\|^2, \max_k \|\mathbf{b}_{\cdot k}\|^2 \right\}. \end{aligned} \tag{4.6}$$

The symbols \mathbf{b}_j and $\mathbf{b}_{\cdot k}$ represent the j th row and k th column of the matrix \mathbf{B} . An immediate consequence of (4.6) is that the median of the norm satisfies

$$\mathbb{M}(\|\mathbf{\Gamma} \odot \mathbf{B}\|) \leq \sigma \sqrt{2 \log(2(d_1 + d_2))}. \tag{4.7}$$

There are nonuniform Gaussian matrices where the estimate (4.7) for the median has the correct order and other examples where the logarithmic factor is parasitic; see Sects. 4.4–4.5 below. The reader may also wish to juxtapose (4.7) with the work of Segner [49, Theorem 3.1] and Latała [28, Theorem 1], although these results are not fully comparable.

To establish (4.6), we first decompose the matrix of interest as a Gaussian series:

$$\mathbf{\Gamma} \odot \mathbf{B} = \sum_{jk} \gamma_{jk} \cdot b_{jk} \mathbf{E}_{jk}.$$

Next, we must determine the variance parameter. Note that

$$\sum_{jk} (b_{jk} \mathbf{E}_{jk})(b_{jk} \mathbf{E}_{jk})^* = \sum_j \left(\sum_k |b_{jk}|^2 \right) \mathbf{E}_{jj} = \text{diag}(\|\mathbf{b}_1\|^2, \|\mathbf{b}_2\|^2, \dots, \|\mathbf{b}_{d_1}\|^2).$$

Similarly,

$$\sum_{jk} (b_{jk} \mathbf{E}_{jk})^* (b_{jk} \mathbf{E}_{jk}) = \sum_k \left(\sum_j |b_{jk}|^2 \right) \mathbf{E}_{kk} = \text{diag}(\|\mathbf{b}_{:1}\|^2, \|\mathbf{b}_{:2}\|^2, \dots, \|\mathbf{b}_{:d_2}\|^2).$$

Therefore,

$$\begin{aligned} \sigma^2 &= \max \left\{ \left\| \text{diag}(\|\mathbf{b}_{:1}\|^2, \|\mathbf{b}_{:2}\|^2, \dots, \|\mathbf{b}_{:d_1}\|^2) \right\|, \right. \\ &\quad \left. \left\| \text{diag}(\|\mathbf{b}_{:1}\|^2, \|\mathbf{b}_{:2}\|^2, \dots, \|\mathbf{b}_{:d_2}\|^2) \right\| \right\} \\ &= \max \left\{ \max_j \|\mathbf{b}_{j:}\|^2, \max_k \|\mathbf{b}_{:k}\|^2 \right\}. \end{aligned}$$

An application of Corollary 4.2 yields the tail bound (4.6).

4.4 Controlling the Expectation

A remarkable feature of Theorem 4.1 is that it always allows us to obtain reasonably accurate estimates for the expected norm of the s.a. Gaussian series

$$\mathbf{Y} = \sum_k \gamma_k \mathbf{A}_k. \tag{4.8}$$

To establish this point, we first compute upper and lower bounds for the second moment of $\|\mathbf{Y}\|$. With σ defined as in (4.2), Theorem 4.1 yields

$$\begin{aligned} \mathbb{E}(\|\mathbf{Y}\|^2) &= \int_0^\infty \mathbb{P}\{\|\mathbf{Y}\| > \sqrt{t}\} dt \leq 2\sigma^2 \log(2d) + 2d \int_{2\sigma^2 \log(2d)}^\infty e^{-t/2\sigma^2} dt \\ &= 2\sigma^2 \log(2ed). \end{aligned}$$

Jensen’s inequality furnishes the lower estimate:

$$\mathbb{E}(\|\mathbf{Y}\|^2) = \mathbb{E}\|\mathbf{Y}^2\| \geq \|\mathbb{E}(\mathbf{Y}^2)\| = \left\| \sum_k \mathbf{A}_k^2 \right\| = \sigma^2.$$

The (homogeneous) first and second moments of the norm of a Gaussian series are equivalent up to a universal constant [33, Corollary 3.2], so we conclude that

$$c\sigma \leq \mathbb{E}\|\mathbf{Y}\| \leq \sigma\sqrt{2\log(2ed)}. \tag{4.9}$$

This argument demonstrates that the matrix variance parameter σ^2 controls the expected norm $\mathbb{E}\|\mathbf{Y}\|$ up to a factor that depends very weakly on the dimension. A similar remark applies to the median value $\mathbb{M}(\|\mathbf{Y}\|)$.

4.5 The Dimensional Factor

In the inequality (4.9), the gap between the upper and lower bounds for $\mathbb{E}\|\mathbf{Y}\|$ arises because of the dimensional factor d in the statement (4.4). This dimensional dependence is a new feature of probability inequalities in the matrix setting. The extra term

appears in each of our main results, and it is usually possible to identify a simple case where it is necessary.

In particular, we cannot remove the factor d from the probability bound in Theorem 4.1. Observe that the norm of a diagonal Gaussian matrix is typically bounded below:

$$\left\| \sum_{k=1}^d \gamma_k \mathbf{E}_{kk} \right\| = \max_k |\gamma_k| > \sqrt{2 \log d} \quad \text{with high probability.}$$

Theorem 4.1 delivers the following tail bound for this series:

$$\mathbb{P} \left\{ \left\| \sum_{k=1}^d \gamma_k \mathbf{E}_{kk} \right\| \geq t \right\} \leq 2d \cdot e^{-t^2/2}.$$

The factor $2d$ ensures that this probability inequality does not become effective until $t \geq \sqrt{2 \log(2d)}$, as is proper.

We can also identify situations where the dimensional term produces an overestimate of the expected norm. For instance, consider a d -dimensional matrix drawn from the unnormalized Gaussian orthogonal ensemble (GOE):

$$\mathbf{W} = \sum_{1 \leq j \leq k \leq d} \gamma_{jk} (\mathbf{E}_{jk} + \mathbf{E}_{kj}).$$

The literature contains a sharp bound for the expected norm of this matrix:

$$\mathbb{E} \|\mathbf{W}\| \leq 2\sqrt{d}. \tag{4.10}$$

The result (4.10) follows from ideas of Gordon [18, 19] elaborated in [13, Theorem 2.11]. Meanwhile, integrating the tail bound (4.4) from Theorem 4.1 yields the weaker result

$$\mathbb{E} \|\mathbf{W}\| \leq \sqrt{(d + 3) \log(2ed)}. \tag{4.11}$$

The estimate (4.11) is too large by a factor of about $\sqrt{\log d}$, which is the worst possible discrepancy in view of (4.9).

Remark 4.4 (Effective Dimension) Let us stress that the *nominal* dimension of the matrices does not play a role in Theorem 4.1. If the ranges of the matrices $\mathbf{A}_1, \mathbf{A}_2, \dots$ are contained within a fixed r -dimensional subspace, we can replace the ambient dimension d with the effective dimension r . A similar remark applies to our other results.

4.6 Comparison with Concentration Inequalities

It is fruitful to think about Theorem 4.1 as a statement that the matrix Gaussian series (4.8) typically falls near its expectation *as a random matrix* when we measure the size of deviations using the operator norm:

$$\mathbb{P} \{ \|\mathbf{Y} - \mathbb{E} \mathbf{Y}\| \geq t \} \leq 2d \cdot e^{-t^2/2\sigma^2}. \tag{4.12}$$

In contrast, the classical concentration inequality [6, Theorem 1.7.6] concerns the variation of *the norm* about its mean value:

$$\mathbb{P}\{\|\mathbf{Y}\| - \mathbb{E}\|\mathbf{Y}\| \geq t\} \leq 2 \cdot e^{-t^2/2\sigma_*^2}, \tag{4.13}$$

where the scale for deviations depends on the *weak variance* parameter

$$\sigma_*^2 := \sup\left\{\sum_k |\mathbf{u}^* \mathbf{A}_k \mathbf{v}|^2 : \|\mathbf{u}\| = \|\mathbf{v}\| = 1\right\}. \tag{4.14}$$

It can be shown [33, Corollary 3.2] that the bound (4.13) is asymptotically sharp as $t \rightarrow \infty$.

Let us elaborate on the relationship between the matrix variance σ^2 defined in (4.2) and the weak variance σ_*^2 appearing in (4.14). First, note that

$$\sigma_*^2 \leq \sup_{\|\mathbf{u}\|=1} \sum_k \mathbf{u}^* \mathbf{A}_k^2 \mathbf{u} = \left\|\sum_k \mathbf{A}_k^2\right\| = \sigma^2. \tag{4.15}$$

Equality holds in (4.15) when, for example, the family $\{\mathbf{A}_k\}$ commutes. We can also establish a reverse inequality:

$$\sigma^2 = \left\|\sum_k \mathbf{A}_k \left(\sum_j \mathbf{e}_j \mathbf{e}_j^*\right) \mathbf{A}_k\right\| \leq \sum_j \sup_{\|\mathbf{u}\|=1} \sum_k |\mathbf{u}^* \mathbf{A}_k \mathbf{e}_j|^2 \leq d \cdot \sigma_*^2, \tag{4.16}$$

where $\{\mathbf{e}_j : j = 1, \dots, d\}$ is the standard basis for \mathbb{R}^d . In the worst case,² the bound (4.16) has roughly the correct order.

In summary, the matrix concentration inequality (4.12) always leads to a good estimate for the expected norm $\mathbb{E}\|\mathbf{Y}\|$. Nevertheless, the presence of the parameter σ^2 in the tail bound can lead to a significant overestimate of the probability that $\|\mathbf{Y}\|$ is large. On the other hand, the classical inequality (4.13) contains no information about the mean, but it always produces a sharp large-deviation bound. Therefore, the two results complement each other well.

4.7 Noncommutative Moment Inequalities

The matrix Laplace transform bound, Proposition 3.1, demonstrates that we can bound tail probabilities for the norm of a random series by controlling the matrix mgf. In certain special cases, it is possible to bound the matrix mgf using noncommutative (nc) moment inequalities. Let us describe how to establish Theorem 4.1 in this fashion. This material is unrelated to the main development, so the reader may skip it with impunity.

The nc Khintchine inequality provides an estimate for the expectation of the $(2p)$ th moment of the Schatten $2p$ -norm of a matrix Gaussian series [31, 32, 44].

²A worst-case example occurs with high probability when the sequence $\{\mathbf{A}_k : k = 1, \dots, d\}$ consists of independent matrices drawn from the d -dimensional GOE, but the proof seems to be complicated.

The most elementary formulation of this result states that

$$\mathbb{E} \operatorname{tr} \left(\sum_k \gamma_k \mathbf{A}_k \right)^{2p} \leq C_{2p} \cdot \operatorname{tr} \left(\sum_k \mathbf{A}_k^2 \right)^p \quad \text{for } p = 1, 2, 3, \dots \tag{4.17}$$

Buchholz [7, Theorem 5] has shown that the optimal constant in (4.17) satisfies

$$C_{2p} := \mathbb{E} |\gamma_1|^{2p} = (2p - 1)!! = \frac{(2p)!}{p! 2^p}.$$

The bound (4.17) also holds with the same constant when we replace $\{\gamma_k\}$ by a sequence of independent Rademacher variables [8, Theorem 5].

The family (4.17) of inequalities allows us to develop a short proof of the tail bound for matrix Gaussian and Rademacher series.

Alternative Proof of Theorem 4.1 Proposition 3.1 yields

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_k \gamma_k \mathbf{A}_k \right) \geq t \right\} \leq \inf_{\theta > 0} \left\{ e^{-\theta t} \cdot \mathbb{E} \operatorname{tr} \exp \left(\theta \sum_k \gamma_k \mathbf{A}_k \right) \right\}. \tag{4.18}$$

We may use (4.17) to bound the Taylor series for the matrix mgf term by term:

$$\begin{aligned} \mathbb{E} \operatorname{tr} \exp \left(\theta \sum_k \gamma_k \mathbf{A}_k \right) &= \sum_{p=0}^{\infty} \frac{\theta^{2p}}{(2p)!} \mathbb{E} \operatorname{tr} \left(\sum_k \gamma_k \mathbf{A}_k \right)^{2p} \\ &\leq \sum_{p=0}^{\infty} \frac{\theta^{2p}}{p! 2^p} \operatorname{tr} \left(\sum_k \mathbf{A}_k^2 \right)^p = \operatorname{tr} \exp \left(\frac{\theta^2}{2} \sum_k \mathbf{A}_k^2 \right). \end{aligned} \tag{4.19}$$

Substitute (4.19) into (4.18), and select $\theta = t/\sigma^2$ to complete the minimization. \square

We may regard the mgf bound (4.19) as an “exponential generating function” for the family of nc Khintchine inequalities (4.17), but—unfortunately—the nc Khintchine inequalities *do not* follow as a consequence of this mgf bound. Recall that Lieb’s result, Theorem 3.2, also delivers a proof of the inequality (4.19). This observation suggests that it might be possible to use Lieb’s theorem to prove the nc Khintchine inequalities (4.17). We regard this as a tantalizing open question.

4.8 Comparison with the Ahlswede–Winter Bound

In Sect. 3.7, we describe how Ahlswede and Winter bound the matrix mgf [3, Appendix]. It is natural to ask how inequalities developed using their approach compare with the results in this paper.

Gaussian series provide an excellent illustration of the discrepancy between the two techniques. In this case, the Ahlswede–Winter method yields the probability inequality

$$\mathbb{P} \left\{ \left\| \sum_k \gamma_k \mathbf{A}_k \right\| \geq t \right\} \leq 2d \cdot e^{-t^2/2\sigma_{AW}^2} \quad \text{where } \sigma_{AW}^2 := \sum_k \|\mathbf{A}_k^2\|. \tag{4.20}$$

The estimate (4.20) should be compared with our bound (4.4). The Ahlswede–Winter variance parameter σ_{AW}^2 always dominates the matrix variance parameter (4.2) because

$$\sigma^2 = \left\| \sum_k \mathbf{A}_k^2 \right\| \leq \sum_k \|\mathbf{A}_k^2\| = \sigma_{AW}^2.$$

The two variance parameters rarely coincide, and the best reverse inequality is

$$\sigma_{AW}^2 \leq \sum_k \text{tr} \mathbf{A}_k^2 \leq d \cdot \left\| \sum_k \mathbf{A}_k^2 \right\| = d \cdot \sigma^2.$$

This worst-case behavior is typical. For instance, consider the two Gaussian matrices presented in Sect. 4.5. The Ahlswede–Winter tail bound (4.20) provides essentially no information about the norm of either matrix.

Remark 4.5 (Moment Inequalities) There is an alternative approach to establishing the result (4.20) that parallels the method presented in Sect. 4.7. We simply bound the Taylor series of the matrix mgf term by term using an appropriate family of moment inequalities:

$$\mathbb{E} \text{tr} \left(\sum_k \gamma_k \mathbf{A}_k \right)^{2p} \leq C_{2p} \cdot \left(\sum_k [\text{tr}(\mathbf{A}_k^{2p})]^{1/p} \right)^p$$

where $C_{2p} := \frac{(2p)!}{p! 2^p}$ for $p = 1, 2, 3, \dots$

These estimates follow from a result of Tomczak–Jaegermann [52, Theorem 3.1] for Rademacher series together with the central limit theorem.

5 Sums of Random Positive Semidefinite Matrices

The classical Chernoff bounds concern the sum of independent, nonnegative, and uniformly bounded random variables. In sympathy, matrix Chernoff bounds describe the extreme eigenvalues of a sum of independent, psd random matrices whose maximum eigenvalues are subject to a uniform bound. These probability inequalities demonstrate that the upper and lower tails of the sum exhibit binomial-type behavior.

Our first result parallels the strongest versions of the scalar Chernoff inequality for the proportion of successes in a sequence of independent (but not identical) Bernoulli trials [34, Exercise 7].

Theorem 5.1 (Matrix Chernoff I) *Consider a sequence $\{\mathbf{X}_k : k = 1, 2, \dots, n\}$ of independent, random, self-adjoint matrices that satisfy*

$$\mathbf{X}_k \succcurlyeq \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\mathbf{X}_k) \leq 1 \quad \text{almost surely.}$$

Compute the minimum and maximum eigenvalues of the average expectation,

$$\bar{\mu}_{\min} := \lambda_{\min} \left(\frac{1}{n} \sum_{k=1}^n \mathbb{E} \mathbf{X}_k \right) \quad \text{and} \quad \bar{\mu}_{\max} := \lambda_{\max} \left(\frac{1}{n} \sum_{k=1}^n \mathbb{E} \mathbf{X}_k \right).$$

Then

$$\mathbb{P} \left\{ \lambda_{\min} \left(\frac{1}{n} \sum_{k=1}^n \mathbf{X}_k \right) \leq \alpha \right\} \leq d \cdot e^{-n \cdot D(\alpha \parallel \bar{\mu}_{\min})} \quad \text{for } 0 \leq \alpha \leq \bar{\mu}_{\min}, \quad \text{and}$$

$$\mathbb{P} \left\{ \lambda_{\max} \left(\frac{1}{n} \sum_{k=1}^n \mathbf{X}_k \right) \geq \alpha \right\} \leq d \cdot e^{-n \cdot D(\alpha \parallel \bar{\mu}_{\max})} \quad \text{for } \bar{\mu}_{\max} \leq \alpha \leq 1.$$

The binary information divergence $D(a \parallel u) := a(\log(a) - \log(u)) + (1 - a) \times (\log(1 - a) - \log(1 - u))$ for $a, u \in [0, 1]$.

We have found that the following weaker version of Theorem 5.1 also produces excellent results and is simpler to apply. This corollary corresponds with the usual statement of the scalar Chernoff inequalities for sums of nonnegative random variables; see [34, Exercise 8] or [37, Sect. 4.1].

Corollary 5.2 (Matrix Chernoff II) *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, self-adjoint matrices that satisfy*

$$\mathbf{X}_k \succcurlyeq \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\mathbf{X}_k) \leq R \quad \text{almost surely.}$$

Compute the minimum and maximum eigenvalues of the sum of expectations,

$$\mu_{\min} := \lambda_{\min} \left(\sum_k \mathbb{E} \mathbf{X}_k \right) \quad \text{and} \quad \mu_{\max} := \lambda_{\max} \left(\sum_k \mathbb{E} \mathbf{X}_k \right).$$

Then

$$\mathbb{P} \left\{ \lambda_{\min} \left(\sum_k \mathbf{X}_k \right) \leq (1 - \delta) \mu_{\min} \right\} \leq d \cdot \left[\frac{e^{-\delta}}{(1 - \delta)^{1-\delta}} \right]^{\mu_{\min}/R} \quad \text{for } \delta \in [0, 1], \quad \text{and}$$

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_k \mathbf{X}_k \right) \geq (1 + \delta) \mu_{\max} \right\} \leq d \cdot \left[\frac{e^{\delta}}{(1 + \delta)^{1+\delta}} \right]^{\mu_{\max}/R} \quad \text{for } \delta \geq 0.$$

The proofs of Theorem 5.1 and Corollary 5.2 appear below in Sect. 5.1. We continue this discussion with some telegraphic remarks concerning various aspects of the Chernoff bounds.

Remark 5.3 (Related Inequalities) The following standard simplification of Corollary 5.2 is useful:

$$\mathbb{P} \left\{ \lambda_{\min} \left(\sum_k \mathbf{X}_k \right) \leq t \mu_{\min} \right\} \leq d \cdot e^{-(1-t)^2 \mu_{\min}/2R} \quad \text{for } t \in [0, 1], \quad \text{and}$$

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geq t\mu_{\max}\right\} \leq d \cdot \left[\frac{e}{t}\right]^{t\mu_{\max}/R} \quad \text{for } t \geq e.$$

These inequalities manifest that the minimum eigenvalue has normal-type behavior and the maximum eigenvalue exhibits Poisson-type decay.

Remark 5.4 (Applications) Matrix Chernoff inequalities are very effective for studying random matrices with independent columns. Consider a rectangular random matrix

$$\mathbf{Z} = [\mathbf{z}_1 \quad \mathbf{z}_2 \quad \dots \quad \mathbf{z}_n],$$

where $\{\mathbf{z}_k\}$ is a family of independent random vectors in \mathbb{C}^m . The norm of \mathbf{Z} satisfies

$$\|\mathbf{Z}\|^2 = \lambda_{\max}(\mathbf{Z}\mathbf{Z}^*) = \lambda_{\max}\left(\sum_{k=1}^n \mathbf{z}_k \mathbf{z}_k^*\right).$$

Similarly, the minimum singular value s_m of the matrix satisfies

$$s_m(\mathbf{Z})^2 = \lambda_{\min}(\mathbf{Z}\mathbf{Z}^*) = \lambda_{\min}\left(\sum_{k=1}^n \mathbf{z}_k \mathbf{z}_k^*\right).$$

In each case, the summands are stochastically independent and psd, so the matrix Chernoff bounds apply. See [54] for a problem where this method is effective.

Remark 5.5 (Expectations) Corollary 5.2 produces accurate estimates for the expectation of the maximum eigenvalue:

$$\mu_{\max} \leq \mathbb{E} \lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \leq C \cdot \max\{\mu_{\max}, R \log d\}.$$

The lower bound is Jensen’s inequality; the upper bound follows from a messy—but standard—calculation. Observe that the dimensional dependence vanishes when the mean μ_{\max} is sufficiently large in comparison with the upper bound R !

Remark 5.6 (Dimensional Factor) The factor d in the Chernoff bounds cannot be omitted because of the coupon collector’s problem [37, Sect. 3.6]. Consider a d -dimensional random matrix \mathbf{X} with the distribution

$$\mathbf{X} = \mathbf{E}_{jj} \quad \text{with probability } d^{-1} \text{ for each } j = 1, 2, \dots, d.$$

If $\{\mathbf{X}_k\}$ is a sequence of independent random matrices with the same distribution as \mathbf{X} , then

$$\lambda_{\min}\left(\sum_{k=1}^n \mathbf{X}_k\right) = 0 \quad \text{with high probability unless } n > d \log d.$$

The dimensional factor in the lower Chernoff bound reflects this fact. The same example shows that the upper Chernoff bound must also exhibit a dimensional dependence. We have extracted this idea from [48, Sect. 3.5].

Remark 5.7 (Previous work) Theorem 5.1 is a considerable strengthening of the matrix Chernoff bound established by Ahlswede and Winter [3, Theorem 19]. Their proof requires the extra assumption that the summands are identically distributed, in which case their result matches Theorem 5.1.

5.1 Proofs

To establish the matrix Chernoff inequalities, we commence with a semidefinite bound for the matrix mgf of a random psd contraction.

Lemma 5.8 (Chernoff mgf) *Suppose that \mathbf{X} is a random psd matrix that satisfies $\lambda_{\max}(\mathbf{X}) \leq 1$. Then*

$$\mathbb{E} e^{\theta \mathbf{X}} \preceq \mathbf{I} + (e^\theta - 1)(\mathbb{E} \mathbf{X}) \quad \text{for } \theta \in \mathbb{R}.$$

The proof of Lemma 5.8 parallels the classical argument; the matrix adaptation is due to Ahlswede and Winter [3, Theorem 19].

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

$$f(x) \leq f(0) + [f(1) - f(0)] \cdot x \quad \text{for } x \in [0, 1].$$

More explicitly,

$$e^{\theta x} \leq 1 + (e^\theta - 1) \cdot x \quad \text{for } x \in [0, 1].$$

The eigenvalues of \mathbf{X} lie in the interval $[0, 1]$, so the transfer rule (2.2) implies that

$$e^{\theta \mathbf{X}} \preceq \mathbf{I} + (e^\theta - 1)\mathbf{X}.$$

Expectation respects the semidefinite order, so

$$\mathbb{E} e^{\theta \mathbf{X}} \preceq \mathbf{I} + (e^\theta - 1)(\mathbb{E} \mathbf{X}).$$

This is the advertised conclusion. □

We prove the upper Chernoff bounds first because the argument is slightly easier.

Proof of Theorem 5.1, Upper Bound The Chernoff mgf bound, Lemma 5.8, states that

$$\mathbb{E} e^{\theta \mathbf{X}_k} \preceq \mathbf{I} + g(\theta) \cdot (\mathbb{E} \mathbf{X}_k) \quad \text{where } g(\theta) := e^\theta - 1 \text{ for } \theta > 0.$$

As a result, Corollary 3.9 implies

$$\begin{aligned} \mathbb{P}\left\{\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geq t\right\} &\leq d \cdot \exp\left(-\theta t + n \cdot \log \lambda_{\max}\left(\frac{1}{n} \sum_k (\mathbf{I} + g(\theta) \cdot \mathbb{E} \mathbf{X}_k)\right)\right) \\ &= d \cdot \exp\left(-\theta t + n \cdot \log \lambda_{\max}\left(\mathbf{I} + g(\theta) \cdot \frac{1}{n} \sum_k \mathbb{E} \mathbf{X}_k\right)\right) \\ &= d \cdot \exp(-\theta t + n \cdot \log(1 + g(\theta) \cdot \bar{\mu}_{\max})). \end{aligned} \tag{5.1}$$

The third relation follows from basic properties of the eigenvalue map and the definition of $\bar{\mu}_{\max}$. Make the change of variables $t \mapsto n\alpha$. The right-hand side is smallest when

$$\theta = \log(\alpha/(1 - \alpha)) - \log(\bar{\mu}_{\max}/(1 - \bar{\mu}_{\max})).$$

Substitute these quantities into (5.1) to obtain the information divergence upper bound. □

Proof of Corollary 5.2, Upper Bound Assume that the summands satisfy the uniform eigenvalue bound with $R = 1$; the general result follows by rescaling. The shortest route to the weaker Chernoff upper bound starts at (5.1). The numerical inequality $\log(1 + x) \leq x$, valid for $x > -1$, implies that

$$\begin{aligned} \mathbb{P}\left\{\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geq t\right\} &\leq d \cdot \exp(-\theta t + g(\theta) \cdot n\bar{\mu}_{\max}) \\ &= d \cdot \exp(-\theta t + g(\theta) \cdot \mu_{\max}). \end{aligned}$$

Make the change of variables $t \mapsto (1 + \delta)\mu_{\max}$, and select the parameter $\theta = \log(1 + \delta)$. Simplify the resulting tail bound to complete the proof. □

The lower bounds follow from a closely related argument.

Proof of Theorem 5.1, Lower Bound We intend to apply Corollary 3.9 to the sequence $\{-\mathbf{X}_k\}$. In this case, the Chernoff mgf, Lemma 5.8, states that

$$\mathbb{E} e^{\theta(-\mathbf{X}_k)} = \mathbb{E} e^{(-\theta)\mathbf{X}_k} \preceq \mathbf{I} - g(\theta) \cdot (\mathbb{E} \mathbf{X}_k) \quad \text{where } g(\theta) := 1 - e^{-\theta} \text{ for } \theta > 0.$$

The minimum eigenvalue $\lambda_{\min}(-\mathbf{A}) = -\lambda_{\max}(\mathbf{A})$, so we can apply Corollary 3.9 as follows.

$$\begin{aligned} \mathbb{P}\left\{\lambda_{\min}\left(\sum_k \mathbf{X}_k\right) \leq t\right\} &= \mathbb{P}\left\{\lambda_{\max}\left(\sum_k (-\mathbf{X}_k)\right) \geq -t\right\} \\ &\leq d \cdot \exp\left(\theta t + n \cdot \log \lambda_{\max}\left(\frac{1}{n} \sum_k (\mathbf{I} - g(\theta) \cdot \mathbb{E} \mathbf{X}_k)\right)\right) \\ &= d \cdot \exp\left(\theta t + n \cdot \log\left(1 - g(\theta) \cdot \lambda_{\min}\left(\frac{1}{n} \sum_{k=1}^n \mathbb{E} \mathbf{X}_k\right)\right)\right) \\ &= d \cdot \exp(\theta t + n \cdot \log(1 - g(\theta) \cdot \bar{\mu}_{\min})). \end{aligned} \tag{5.2}$$

Make the substitution $t \mapsto n\alpha$. The right-hand side is minimal when

$$\theta = \log(\bar{\mu}_{\min}/(1 - \bar{\mu}_{\min})) - \log(\alpha/(1 - \alpha)).$$

These steps result in the information divergence lower bound. □

Proof of Corollary 5.2, Lower Bound As before, assume that the uniform bound $R = 1$. We obtain the weaker lower bound as a consequence of (5.2). The inequality $\log(1 + x) \leq x$ holds for $x > -1$, so we have

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_k \mathbf{X}_k\right) \leq t\right\} \leq d \cdot \exp(\theta t - g(\theta) \cdot n\bar{\mu}_{\min}) = d \cdot \exp(\theta t - g(\theta) \cdot \mu_{\min}).$$

Make the replacement $t \mapsto (1 - \delta)\mu_{\min}$, and select $\theta = -\log(1 - \delta)$ to complete the proof. □

Remark 5.9 (Alternative Proof) Corollary 5.2 can also be established directly using Corollary 3.7 instead of Corollary 3.9. In this case, we use the mgf bound

$$\mathbb{E} e^{\theta \mathbf{X}} \preceq \exp((e^\theta - 1)(\mathbb{E} \mathbf{X})) \quad \text{for } \theta \in \mathbb{R},$$

which follows instantly from Lemma 5.8 and the semidefinite relation (2.3). The remaining details mirror the arguments here.

6 Matrix Bennett and Bernstein Inequalities

In the scalar setting, Bennett and Bernstein inequalities describe the upper tail of a sum of independent, zero-mean random variables that are either bounded or subexponential. In the matrix case, the analogous results concern a sum of zero-mean random matrices.

Our first result describes the case where the maximum eigenvalue of each summand satisfies a uniform bound.

Theorem 6.1 (Matrix Bernstein: bounded case) *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, self-adjoint matrices with dimension d . Assume that*

$$\mathbb{E} \mathbf{X}_k = \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\mathbf{X}_k) \leq R \quad \text{almost surely.}$$

Compute the norm of the total variance,

$$\sigma^2 := \left\| \sum_k \mathbb{E}(\mathbf{X}_k^2) \right\|.$$

Then the following chain of inequalities holds for all $t \geq 0$:

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geq t\right\} \leq d \cdot \exp\left(-\frac{\sigma^2}{R^2} \cdot h\left(\frac{Rt}{\sigma^2}\right)\right) \tag{i}$$

$$\leq d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right) \tag{ii}$$

$$\leq \begin{cases} d \cdot \exp(-3t^2/8\sigma^2) & \text{for } t \leq \sigma^2/R; \\ d \cdot \exp(-3t/8R) & \text{for } t \geq \sigma^2/R. \end{cases} \tag{iii}$$

The function $h(u) := (1 + u) \log(1 + u) - u$ for $u \geq 0$.

Observe that Theorem 6.1 places no assumption on the minimum eigenvalues of the summands, which may be arbitrarily small. As a consequence, when we apply the result to the two sequences $\{\mathbf{X}_k\}$ and $\{-\mathbf{X}_k\}$, the parameter R may differ.

Theorem 6.1(i) can be viewed as a matrix version of the Bennett inequality [34, Theorem 5], which implies that the tail probabilities exhibit Poisson-type decay. Part (ii) parallels a well-known result [34, Theorem 6], which is perhaps the most famous among the probability inequalities attributed to Bernstein. Part (iii), which we call the split Bernstein inequality, clearly delineates between the normal behavior that occurs at moderate deviations and the slower decay that emerges in the tail.

A related inequality holds when we allow the moments of the random matrices to grow at a limited rate, which we interpret as a matrix extension of the moment behavior of a subexponential random variable [12, Lemma 4.1.9].

Theorem 6.2 (Matrix Bernstein: subexponential case) *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, self-adjoint matrices with dimension d . Assume that*

$$\mathbb{E} \mathbf{X}_k = \mathbf{0} \quad \text{and} \quad \mathbb{E}(\mathbf{X}_k^p) \preceq \frac{p!}{2} \cdot R^{p-2} \mathbf{A}_k^2 \quad \text{for } p = 2, 3, 4, \dots$$

Compute the variance parameter

$$\sigma^2 := \left\| \sum_k \mathbf{A}_k^2 \right\|.$$

Then the following chain of inequalities holds for all $t \geq 0$:

$$\mathbb{P}\left\{ \lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geq t \right\} \leq d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt}\right) \tag{i}$$

$$\leq \begin{cases} d \cdot \exp(-t^2/4\sigma^2) & \text{for } t \leq \sigma^2/R; \\ d \cdot \exp(-t/4R) & \text{for } t \geq \sigma^2/R. \end{cases} \tag{ii}$$

The hypotheses of Theorem 6.2 are not fully comparable with the hypotheses of Theorem 6.1, because Theorem 6.2 allows the random matrices to be unbounded but it also demands that we control the fluctuation of the maximum and minimum eigenvalues. The resulting tail bound is very similar to Theorem 6.1(ii). We cannot achieve a Bennett-type inequality, like Theorem 6.1(i), without stricter assumptions on the growth of moments.

The proofs of Theorems 6.1 and 6.2 appear below. We finish the discussion with an assorted collection of enriching comments.

Remark 6.3 (Rectangular Versions) The matrix Bernstein inequalities admit rectangular variants. For example, consider a sequence $\{\mathbf{Z}_k\}$ of $d_1 \times d_2$ random matrices that satisfy the assumptions

$$\mathbb{E} \mathbf{Z}_k = \mathbf{0} \quad \text{and} \quad \|\mathbf{Z}_k\| \leq R \quad \text{almost surely.}$$

We can apply Theorem 6.1 to the s.a. dilation (2.10) of the sum of these random matrices to see that the probability

$$\mathbb{P} \left\{ \left\| \sum_k \mathbf{Z}_k \right\| \geq t \right\} \leq d \cdot \exp \left(\frac{\sigma^2}{R^2} \cdot h \left(\frac{Rt}{\sigma^2} \right) \right),$$

where $d := d_1 + d_2$ and where the variance parameter

$$\sigma^2 := \max \left\{ \left\| \sum_k \mathbb{E}(\mathbf{Z}_k \mathbf{Z}_k^*) \right\|, \left\| \sum_k \mathbb{E}(\mathbf{Z}_k^* \mathbf{Z}_k) \right\| \right\}.$$

This argument leads to Theorem 1.6, stated in the Introduction. There is also a rectangular extension of Theorem 6.2, but the hypotheses are messier.

Remark 6.4 (Related Inequalities) There are too many variants of the scalar Bernstein inequality to present the matrix generalization of each one. Let us just mention a few of the possibilities.

- Theorem 6.2 can be sharpened using an idea of Rio that appears in [35, Sect. 2.2.3].
- When the random matrices exhibit moment growth of the form $\mathbb{E}(\mathbf{X}_k^p) \preceq R^{p-2} \mathbf{A}_k^2$, we recover the Poissonian tail behavior captured in Theorem 6.1(i).
- When the summands are symmetric random variables (i.e., $\mathbf{X}_k \sim -\mathbf{X}_k$), we can exploit the fact that the matrix mgf $\mathbb{E} e^{\theta \mathbf{X}_k} = \mathbb{E} \cosh(\theta \mathbf{X}_k)$ to obtain arcsinh inequalities.

Remark 6.5 (Expectations) We can use the matrix Bernstein inequality to bound the mean of the maximum eigenvalue of the random sum. For example, assume that the hypotheses of Theorem 6.1 or 6.2 are in force. Then

$$\mathbb{E} \lambda_{\max} \left(\sum_k \mathbf{X}_k \right) \leq C \cdot \max \{ \sigma \sqrt{\log d}, R \log d \}. \tag{6.1}$$

The upper bound follows by integrating Theorem 6.1(ii) or Theorem 6.2(i). Lower bounds seem to require additional assumptions.

Remark 6.6 (Previous work) Oliveira’s results are quite similar to the bounds presented here. In particular, Oliveira’s martingale inequality [39, Theorem 1.2] implies a weaker version of Theorem 6.1(ii). The main result from [40] has a similar flavor.

6.1 Proof of Theorem 6.1

The main lemma shows how to bound the mgf of a zero-mean random matrix using a bound for its largest eigenvalue.

Lemma 6.7 (Bounded Bernstein mgf) *Suppose that \mathbf{X} is a random s.a. matrix that satisfies*

$$\mathbb{E} \mathbf{X} = \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\mathbf{X}) \leq 1.$$

Then

$$\mathbb{E} e^{\theta \mathbf{X}} \preceq \exp((e^\theta - \theta - 1) \cdot \mathbb{E}(\mathbf{X}^2)) \quad \text{for } \theta > 0.$$

As usual, the proof of the mgf bound parallels a classical method, which we learned from correspondence with Yao-Liang Yu.

Proof Fix the parameter $\theta > 0$, and define a smooth function f on the real line:

$$f(x) = \frac{e^{\theta x} - \theta x - 1}{x^2} \quad \text{for } x \neq 0 \quad \text{and} \quad f(0) = \frac{\theta^2}{2}.$$

An exercise in differential calculus verifies that f is increasing. Therefore, $f(x) \leq f(1)$ when $x \leq 1$. The eigenvalues of \mathbf{X} do not exceed one, so the transfer rule (2.2) implies that

$$f(\mathbf{X}) \preceq f(\mathbf{I}) = f(1) \cdot \mathbf{I}.$$

Expanding the matrix exponential and applying the latter relation, we discover that

$$e^{\theta \mathbf{X}} = \mathbf{I} + \theta \mathbf{X} + \mathbf{X} \cdot f(\mathbf{X}) \cdot \mathbf{X} \preceq \mathbf{I} + \theta \mathbf{X} + f(1) \cdot \mathbf{X}^2.$$

To complete the proof, we take the expectation of this semidefinite bound:

$$\mathbb{E} e^{\theta \mathbf{X}} \preceq \mathbf{I} + f(1) \cdot \mathbb{E}(\mathbf{X}^2) \preceq \exp(f(1) \cdot \mathbb{E}(\mathbf{X}^2)) = \exp((e^\theta - \theta - 1) \cdot \mathbb{E}(\mathbf{X}^2)).$$

The second semidefinite relation follows from (2.3). □

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

Proof of Theorem 6.1 We assume that $R = 1$; the general result follows by a scaling argument once we note that the summands are 1-homogeneous and the variance σ^2 is 2-homogeneous.

The main challenge is to establish the Bennett inequality, part (i); the remaining bounds are consequences of simple numerical estimates. Invoke Lemma 6.7 to see that

$$\mathbb{E} e^{\theta \mathbf{X}_k} \preceq \exp(g(\theta) \cdot \mathbb{E}(\mathbf{X}_k^2)) \quad \text{where } g(\theta) := e^\theta - \theta - 1 \text{ for } \theta > 0.$$

For each $\theta > 0$, Corollary 3.7 implies that

$$\begin{aligned} \mathbb{P}\left\{\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geq t\right\} &\leq d \cdot \exp\left(-\theta t + g(\theta) \cdot \lambda_{\max}\left(\sum_k \mathbb{E}(\mathbf{X}_k^2)\right)\right) \\ &= d \cdot \exp(-\theta t + g(\theta) \cdot \sigma^2). \end{aligned}$$

The right-hand side attains its minimal value when $\theta = \log(1 + t/\sigma^2)$. Substitute and simplify to establish part (i).

The Bennett inequality (i) implies the Bernstein inequality (ii) because of the numerical bound

$$h(u) \geq \frac{u^2/2}{1 + u/3} \quad \text{for } u \geq 0.$$

The latter relation is established by comparing derivatives.

The Bernstein inequality (ii) implies the split Bernstein inequality (iii). To obtain the sub-Gaussian piece of (iii), observe that

$$\frac{1}{\sigma^2 + Rt/3} \geq \frac{1}{\sigma^2 + R(\sigma^2/R)/3} = \frac{3}{4\sigma^2} \quad \text{for } t \leq \sigma^2/R$$

because the left-hand side is a decreasing function of t for $t \geq 0$. Similarly, we obtain the subexponential piece of (iii) from the fact that

$$\frac{t}{\sigma^2 + Rt/3} \geq \frac{(\sigma^2/R)}{\sigma^2 + R(\sigma^2/R)/3} = \frac{3}{4R} \quad \text{for } t \geq \sigma^2/R,$$

which holds because the left-hand side is an increasing function of t for $t \geq 0$. □

6.2 Proof of Theorem 6.2

We begin with the appropriate estimate for the matrix mgf.

Lemma 6.8 (Subexponential Bernstein mgf) *Suppose that \mathbf{X} is a random s.a. matrix that satisfies*

$$\mathbb{E} \mathbf{X} = \mathbf{0} \quad \text{and} \quad \mathbb{E}(\mathbf{X}^p) \preceq \frac{p!}{2} \cdot \mathbf{A}^2 \quad \text{for } p = 2, 3, 4, \dots$$

Then

$$\mathbb{E} e^{\theta \mathbf{X}} \preceq \exp\left(\frac{\theta^2}{2(1-\theta)} \cdot \mathbf{A}^2\right) \quad \text{for } 0 < \theta < 1.$$

Proof The argument proceeds by estimating each term in the Taylor series of the matrix exponential. Indeed,

$$\begin{aligned} \mathbb{E} e^{\theta \mathbf{X}} &= \mathbf{I} + \theta \mathbb{E} \mathbf{X} + \sum_{p=2}^{\infty} \frac{\theta^p \mathbb{E}(\mathbf{X}^p)}{p!} \preceq \mathbf{I} + \sum_{p=2}^{\infty} \frac{\theta^p}{2} \cdot \mathbf{A}^2 \\ &= \mathbf{I} + \frac{\theta^2}{2(1-\theta)} \cdot \mathbf{A}^2 \preceq \exp\left(\frac{\theta^2}{2(1-\theta)} \cdot \mathbf{A}^2\right). \end{aligned}$$

As usual, the last relation is (2.3). □

The Bernstein inequality for subexponential random matrices is an easy consequence of the previous lemma.

Proof of Theorem 6.2 As before, we assume that $R = 1$; the general result follows by scaling. Invoke Lemma 6.8 to see that

$$\mathbb{E} e^{\theta \mathbf{X}_k} \preceq \exp(g(\theta) \cdot \mathbf{A}_k^2) \quad \text{where } g(\theta) := \frac{\theta^2}{2(1-\theta)} \text{ for } 0 < \theta < 1.$$

For each $\theta > 0$, Corollary 3.7 implies that

$$\begin{aligned} \mathbb{P}\left\{\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geq t\right\} &\leq d \cdot \exp\left(-\theta t + g(\theta) \cdot \lambda_{\max}\left(\sum_k \mathbf{A}_k^2\right)\right) \\ &= d \cdot \exp(-\theta t + g(\theta) \cdot \sigma^2). \end{aligned}$$

We select $\theta = t/(\sigma^2 + t)$. Substitute and simplify to complete part (i).

The split inequality (ii) follows from part (i) by the same argument presented in the proof of Theorem 6.1. □

7 The Matrix Hoeffding, Azuma, and McDiarmid Inequalities

In this section, we prove some simple martingale deviation bounds by modifying the approach that we have used to study sums of independent random matrices. More sophisticated martingale results require additional machinery [39, 55].

7.1 Matrix Martingales

We begin with the required definitions. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a master probability space. Consider a filtration $\{\mathcal{F}_k\}$ contained in the master sigma algebra:

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots \subset \mathcal{F}_\infty \subset \mathcal{F}.$$

Given such a filtration, we define the conditional expectation $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot | \mathcal{F}_k]$. A sequence $\{\mathbf{X}_k\}$ of random matrices is *adapted* to the filtration when each \mathbf{X}_k is

measurable with respect to \mathcal{F}_k . Loosely speaking, an adapted sequence is one where the present depends only upon the past.

An adapted sequence $\{\mathbf{Y}_k\}$ of s.a. matrices is called a *matrix martingale* when

$$\mathbb{E}_{k-1} \mathbf{Y}_k = \mathbf{Y}_{k-1} \quad \text{and} \quad \mathbb{E} \|\mathbf{Y}_k\| < \infty \quad \text{for } k = 1, 2, 3, \dots$$

We obtain a scalar martingale if we track any fixed coordinate of a matrix martingale $\{\mathbf{Y}_k\}$. Given a matrix martingale $\{\mathbf{Y}_k\}$, we can construct the *difference sequence*

$$\mathbf{X}_k := \mathbf{Y}_k - \mathbf{Y}_{k-1} \quad \text{for } k = 1, 2, 3, \dots$$

Note that the difference sequence is conditionally zero mean: $\mathbb{E}_{k-1} \mathbf{X}_k = \mathbf{0}$.

7.2 Main Results

The scalar version of Azuma’s inequality states that a scalar martingale exhibits normal concentration about its mean value, and the scale for deviations is controlled by the total maximum squared range of the difference sequence. Here is a matrix extension.

Theorem 7.1 (Matrix Azuma) *Consider a finite adapted sequence $\{\mathbf{X}_k\}$ of self-adjoint matrices in dimension d , and a fixed sequence $\{\mathbf{A}_k\}$ of self-adjoint matrices that satisfy*

$$\mathbb{E}_{k-1} \mathbf{X}_k = \mathbf{0} \quad \text{and} \quad \mathbf{X}_k^2 \preceq \mathbf{A}_k^2 \quad \text{almost surely.}$$

Compute the variance parameter

$$\sigma^2 := \left\| \sum_k \mathbf{A}_k^2 \right\|. \tag{7.1}$$

Then, for all $t \geq 0$,

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_k \mathbf{X}_k \right) \geq t \right\} \leq d \cdot e^{-t^2/8\sigma^2}. \tag{7.2}$$

Theorem 7.1 can also be phrased directly in terms of a matrix martingale.

Corollary 7.2 *Consider an s.a. matrix martingale $\{\mathbf{Y}_k : k = 1, \dots, n\}$ in dimension d , and let $\{\mathbf{X}_k\}$ be the associated difference sequence. Suppose that the difference sequence satisfies the hypotheses of Theorem 7.1, and compute the parameter σ^2 according to (7.1). Then*

$$\mathbb{P} \left\{ \lambda_{\max} (\mathbf{Y}_n - \mathbb{E} \mathbf{Y}_n) \geq t \right\} \leq d \cdot e^{-t^2/8\sigma^2}. \tag{7.3}$$

We continue with a few tangential comments.

Remark 7.3 (Rectangular Version) The matrix Azuma inequality has a rectangular version, which we obtain by applying Theorem 7.1 to the s.a. dilation (2.10) of the adapted sequence.

Remark 7.4 (Related Inequalities) There are several situations where the constant $1/8$ in the bound (7.2) can be improved to $1/2$. One case occurs when each summand \mathbf{X}_k is conditionally symmetric; see Remark 7.8. Another example requires the assumption that \mathbf{X}_k commutes almost surely with \mathbf{A}_k , which allows us to generalize the classical proof [36, Lemma 2.6] of the Azuma inequality to the matrix setting.

If we place the additional assumption that the summands are independent, Theorem 7.1 gives a matrix extension of one of Hoeffding's inequalities, which we have presented as Theorem 1.3 in the introduction.

In the scalar setting, one of the most useful corollaries of Azuma's inequality is the bounded differences inequality of McDiarmid [36, Theorem 3.1]. This result states that a function of independent random variables exhibits normal concentration about its mean, and the variance depends on how much a change in a single variable can alter the value of the function. A version of the bounded differences inequality holds in the matrix setting.

Corollary 7.5 (Matrix bounded differences) *Let $\{Z_k : k = 1, 2, \dots, n\}$ be an independent family of random variables, and let \mathbf{H} be a function that maps n variables to a self-adjoint matrix of dimension d . Consider a sequence $\{\mathbf{A}_k\}$ of fixed self-adjoint matrices that satisfy*

$$\left(\mathbf{H}(z_1, \dots, z_k, \dots, z_n) - \mathbf{H}(z_1, \dots, z'_k, \dots, z_n)\right)^2 \preceq \mathbf{A}_k^2,$$

where z_i and z'_i range over all possible values of Z_i for each index i . Compute the variance parameter

$$\sigma^2 := \left\| \sum_k \mathbf{A}_k^2 \right\|.$$

Then, for all $t \geq 0$,

$$\mathbb{P}\{\lambda_{\max}(\mathbf{H}(\mathbf{z}) - \mathbb{E}\mathbf{H}(\mathbf{z})) \geq t\} \leq d \cdot e^{-t^2/8\sigma^2},$$

where $\mathbf{z} = (Z_1, \dots, Z_n)$.

The proofs of the matrix Azuma and McDiarmid inequalities appear in the next two sections.

7.3 Proof of Theorem 7.1

The classical approach to Azuma's inequality does not seem to extend directly to the matrix setting. See [36, Lemma 2.6] for a short presentation of this argument. We use a different type of proof that is inspired by methods from probability in Banach spaces [33]. The main idea is to inject additional randomness into the sum via a symmetrization procedure.

Lemma 7.6 (Symmetrization) *Let \mathbf{H} be a fixed s.a. matrix, and let \mathbf{X} be a random s.a. matrix with $\mathbb{E} \mathbf{X} = \mathbf{0}$. Then*

$$\mathbb{E} \operatorname{tr} e^{\mathbf{H}+\mathbf{X}} \leq \mathbb{E} \operatorname{tr} e^{\mathbf{H}+2\varepsilon\mathbf{X}},$$

where ε is a Rademacher variable independent from \mathbf{X} .

Proof Construct an independent copy \mathbf{X}' of the random matrix, and let \mathbb{E}' denote integration with respect to the new variable. Since the matrix is zero mean,

$$\mathbb{E} \operatorname{tr} e^{\mathbf{H}+\mathbf{X}} = \mathbb{E} \operatorname{tr} e^{\mathbf{H}+\mathbf{X}-\mathbb{E}' \mathbf{X}'} \leq \mathbb{E} \operatorname{tr} e^{\mathbf{H}+(\mathbf{X}-\mathbf{X}')} = \mathbb{E} \operatorname{tr} e^{\mathbf{H}+\varepsilon(\mathbf{X}-\mathbf{X}')}.$$

We have used the convexity of the trace exponential to justify Jensen’s inequality. Since $\mathbf{X} - \mathbf{X}'$ is a symmetric random variable, we can modulate it by an independent Rademacher variable ε without changing its distribution. The final bound depends on a short sequence of inequalities:

$$\begin{aligned} \mathbb{E} \operatorname{tr} e^{\mathbf{H}+\mathbf{X}} &\leq \mathbb{E} \operatorname{tr} (e^{\mathbf{H}/2+\varepsilon\mathbf{X}} \cdot e^{\mathbf{H}/2-\varepsilon\mathbf{X}'}) \\ &\leq \mathbb{E} [(\operatorname{tr} e^{\mathbf{H}+2\varepsilon\mathbf{X}})^{1/2} \cdot (\operatorname{tr} e^{\mathbf{H}-2\varepsilon\mathbf{X}'})^{1/2}] \\ &\leq (\mathbb{E} \operatorname{tr} e^{\mathbf{H}+2\varepsilon\mathbf{X}})^{1/2} \cdot (\mathbb{E} \operatorname{tr} e^{\mathbf{H}-2\varepsilon\mathbf{X}'})^{1/2} = \mathbb{E} \operatorname{tr} e^{\mathbf{H}+2\varepsilon\mathbf{X}}. \end{aligned}$$

The first relation is the Golden–Thompson inequality (2.6), the second is the Cauchy–Schwarz inequality for the trace, and the third is the Cauchy–Schwarz inequality for real random variables. The last identity follows because the two factors are identically distributed. □

The other essential ingredient in the proof is a conditional bound for the matrix cgf of a symmetrized random matrix.

Lemma 7.7 (Azuma cgf) *Suppose that \mathbf{X} is a random s.a. matrix and \mathbf{A} is a fixed s.a. matrix that satisfy $\mathbf{X}^2 \preceq \mathbf{A}^2$. Let ε be a Rademacher random variable independent from \mathbf{X} . Then*

$$\log \mathbb{E}[e^{2\varepsilon\theta\mathbf{X}} \mid \mathbf{X}] \preceq 2\theta^2\mathbf{A}^2 \quad \text{for } \theta \in \mathbb{R}.$$

Proof We apply the Rademacher mgf bound, Lemma 4.3, conditionally to obtain

$$\mathbb{E}[e^{2\theta\varepsilon\mathbf{X}} \mid \mathbf{X}] \preceq e^{2\theta^2\mathbf{X}^2}.$$

The fact (2.8) that the logarithm is operator monotone implies that

$$\log \mathbb{E}[e^{2\theta\varepsilon\mathbf{X}} \mid \mathbf{X}] \preceq 2\theta^2\mathbf{X}^2 \preceq 2\theta^2\mathbf{A}^2,$$

where the second relation follows from the hypothesis on \mathbf{X} . □

We are prepared to establish the matrix Azuma inequality. The proof involves an iteration similar to the argument that implies the subadditivity of cgf’s, Lemma 3.4, for sums of independent random matrices.

Proof of Theorem 7.1 The matrix Laplace transform method, Proposition 3.1, states that

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geq t\right\} \leq \inf_{\theta > 0} \left\{e^{-\theta t} \cdot \mathbb{E} \operatorname{tr} \exp\left(\sum_k \theta \mathbf{X}_k\right)\right\}. \tag{7.4}$$

The main difficulty in the proof is to bound the matrix mgf, which we accomplish by an iterative argument that alternates between symmetrization and cumulant bounds.

Let us detail the first step of the iteration. Define the natural filtration $\mathcal{F}_k := \mathcal{F}(\mathbf{X}_1, \dots, \mathbf{X}_k)$ of the process $\{\mathbf{X}_k\}$. Then we may compute

$$\begin{aligned} \mathbb{E} \operatorname{tr} \exp\left(\sum_k \theta \mathbf{X}_k\right) &= \mathbb{E} \mathbb{E}\left[\operatorname{tr} \exp\left(\sum_{k=1}^{n-1} \theta \mathbf{X}_k + \theta \mathbf{X}_n\right) \middle| \mathcal{F}_{n-1}\right] \\ &\leq \mathbb{E} \mathbb{E}\left[\operatorname{tr} \exp\left(\sum_{k=1}^{n-1} \theta \mathbf{X}_k + 2\varepsilon \theta \mathbf{X}_n\right) \middle| \mathcal{F}_n\right] \\ &\leq \mathbb{E} \operatorname{tr} \exp\left(\sum_{k=1}^{n-1} \theta \mathbf{X}_k + \log \mathbb{E}[e^{2\varepsilon \theta \mathbf{X}_n} \mid \mathcal{F}_n]\right) \\ &\leq \mathbb{E} \operatorname{tr} \exp\left(\sum_{k=1}^{n-1} \theta \mathbf{X}_k + 2\theta^2 \mathbf{A}_n^2\right). \end{aligned}$$

The first identity is the tower property of conditional expectation. In the second line, we invoke the symmetrization method, Lemma 7.6, conditional on \mathcal{F}_{n-1} , and then we relax the conditioning on the inner expectation to the larger algebra \mathcal{F}_n . By construction, the Rademacher variable ε is independent from \mathcal{F}_n , so we can apply the concavity result, Corollary 3.3, conditional on \mathcal{F}_n . Finally, we use the fact (2.5) that the trace exponential is monotone to introduce the Azuma cgf bound, Lemma 7.7, in the last inequality.

By iteration, we achieve

$$\mathbb{E} \operatorname{tr} \exp\left(\sum_k \theta \mathbf{X}_k\right) \leq \operatorname{tr} \exp\left(2\theta^2 \sum_k \mathbf{A}_k^2\right). \tag{7.5}$$

Note that this procedure relies on the fact that the sequence $\{\mathbf{A}_k\}$ of upper bounds does not depend on the values of the random sequence $\{\mathbf{X}_k\}$. Substitute the mgf bound (7.5) into the Laplace transform bound (7.4), and observe that the infimum is achieved when $\theta = t/4\sigma^2$. □

Remark 7.8 Suppose that the sequence $\{\mathbf{X}_k\}$ is conditionally symmetric:

$$\mathbf{X}_k \sim -\mathbf{X}_k \quad \text{conditional on } \mathcal{F}_{k-1}.$$

When we execute the proof of Theorem 7.1 under this assumption, we can symmetrize each term in the sum without suffering an extra factor of two. For example,

$$\mathbb{E} \left[\operatorname{tr} \exp \left(\sum_{k=1}^{n-1} \theta \mathbf{X}_k + \theta \mathbf{X}_n \right) \middle| \mathcal{F}_{n-1} \right] = \mathbb{E} \left[\operatorname{tr} \exp \left(\sum_{k=1}^{n-1} \theta \mathbf{X}_k + \varepsilon \theta \mathbf{X}_n \right) \middle| \mathcal{F}_{n-1} \right],$$

where ε is independent of \mathcal{F}_n . The rest of the proof remains the same, but the analog of the bound (7.2) has a constant of $1/2$ instead of $1/8$ in the exponent.

7.4 Proof of Corollary 7.5

Finally, we establish the matrix version of the bounded differences inequality. The main idea in the argument is to construct the Doob martingale associated with the natural filtration of the independent random sequence. We compute semidefinite bounds for the difference sequence, and then we apply the matrix Azuma inequality to control the deviations of the martingale.

Proof of Corollary 7.5 In this argument only, we write \mathbb{E}_Z for the expectation with respect to a random variable Z , holding other variables fixed. Recall that $\mathbf{z} = (Z_1, \dots, Z_n)$. For $k = 0, 1, \dots, n$, consider the random matrices

$$\mathbf{Y}_k := \mathbb{E}[\mathbf{H}(\mathbf{z}) \mid Z_1, Z_2, \dots, Z_k] = \mathbb{E}_{Z_{k+1}} \mathbb{E}_{Z_{k+2}} \cdots \mathbb{E}_{Z_n} \mathbf{H}(\mathbf{z}).$$

The sequence $\{\mathbf{Y}_k\}$ forms a Doob martingale. The associated difference sequence is

$$\mathbf{X}_k := \mathbf{Y}_k - \mathbf{Y}_{k-1} = \mathbb{E}_{Z_{k+1}} \mathbb{E}_{Z_{k+2}} \cdots \mathbb{E}_{Z_n} (\mathbf{H}(\mathbf{z}) - \mathbb{E}_{Z_k} \mathbf{H}(\mathbf{z})),$$

where the identity follows from independence and Fubini’s theorem.

It remains to bound the difference sequence. Let Z'_k be an independent copy of Z_k , and construct the random vector $\mathbf{z}' = (Z_1, \dots, Z_{k-1}, Z'_k, Z_{k+1}, \dots, Z_n)$. Observe that $\mathbb{E}_{Z_k} \mathbf{H}(\mathbf{z}) = \mathbb{E}_{Z'_k} \mathbf{H}(\mathbf{z}')$ and that $\mathbf{H}(\mathbf{z})$ does not depend on Z'_k . Therefore, we can write

$$\mathbf{X}_k = \mathbb{E}_{Z_{k+1}} \mathbb{E}_{Z_{k+2}} \cdots \mathbb{E}_{Z_n} \mathbb{E}_{Z'_k} (\mathbf{H}(\mathbf{z}) - \mathbf{H}(\mathbf{z}')).$$

The vectors \mathbf{z} and \mathbf{z}' differ only in the k th coordinate, so that

$$(\mathbf{H}(\mathbf{z}) - \mathbf{H}(\mathbf{z}'))^2 \preceq \mathbf{A}_k^2$$

by definition of the bound \mathbf{A}_k^2 . Finally, the semidefinite Jensen inequality (2.14) for the matrix square yields

$$\mathbf{X}_k^2 \preceq \mathbb{E}_{Z_{k+1}} \mathbb{E}_{Z_{k+2}} \cdots \mathbb{E}_{Z_n} \mathbb{E}_{Z'_k} (\mathbf{H}(\mathbf{z}) - \mathbf{H}(\mathbf{z}'))^2 \preceq \mathbf{A}_k^2.$$

To complete the proof, we apply (7.3) to the martingale $\{\mathbf{Y}_k\}$. □

Acknowledgements I would like to thank Vern Paulsen and Bernhard Bodmann for some helpful conversations connected with this project. Klas Markström and David Gross provided references to related work. Ben Recht offered some useful comments on the presentation. Yao-Liang Yu proposed the argument in Lemma 6.7. Richard Chen and Alex Gittens have helped me root out (numerous) typographic errors. Finally, let me mention Roberto Oliveira's elegant work [40] on matrix probability inequalities, which originally spurred me to pursue this project.

Research supported by ONR award N00014-08-1-0883, DARPA award N66001-08-1-2065, and AFOSR award FA9550-09-1-0643.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. N. Ailon, B. Chazelle, The fast Johnson–Lindenstrauss transform and approximate nearest neighbors, *SIAM J. Comput.* **39**(1), 302–322 (2009).
2. D. Achlioptas, F. McSherry, Fast computation of low-rank matrix approximations, *J. Assoc. Comput. Mach.* **54**(2), Article 10 (2007) (electronic).
3. R. Ahlswede, A. Winter, Strong converse for identification via quantum channels, *IEEE Trans. Inf. Theory* **48**(3), 569–579 (2002).
4. R. Bhatia, *Matrix Analysis*. Graduate Texts in Mathematics, vol. 169 (Springer, Berlin, 1997), p. 10.
5. R. Bhatia, *Positive Definite Matrices* (Princeton Univ. Press, Princeton, 2007).
6. V. Bogdanov, *Gaussian Measures* (American Mathematical Society, Providence, 1998).
7. A. Buchholz, Operator Khintchine inequality in non-commutative probability, *Math. Ann.* **319**, 1–16 (2001).
8. 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).
9. H. Chernoff, A measure of the asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Ann. Math. Stat.* **23**(4), 493–507 (1952).
10. D. Cristofides, K. Markström, Expansion properties of random Cayley graphs and vertex transitive graphs via matrix martingales, *Random Struct. Algorithms* **32**(8), 88–100 (2008).
11. E. Candès, J.K. Romberg, Sparsity and incoherence in compressive sampling, *Inverse Probl.* **23**(3), 969–985 (2007).
12. V.H. de la Peña, E. Giné, *Decoupling: From Dependence to Independence*, Probability and Its Applications (Springer, Berlin, 2002).
13. K.R. Davidson, S. J. Szarek. Local operator theory, random matrices, and Banach spaces, in *Handbook of Banach Space Geometry*, ed. by W.B. Johnson, J. Lindenstrauss (Elsevier, Amsterdam, 2002), pp. 317–366.
14. A. Dembo, O. Zeitouni, *Large Deviations: Techniques and Applications*, 2nd edn. (Springer, Berlin, 1998).
15. E.G. Effros, A matrix convexity approach to some celebrated quantum inequalities, *Proc. Natl. Acad. Sci. USA* **106**(4), 1006–1008 (2009).
16. H. Epstein, Remarks on two theorems of E. Lieb, *Commun. Math. Phys.* **31**, 317–325 (1973).
17. D.A. Freedman, On tail probabilities for martingales, *Ann. Probab.* **3**(1), 100–118 (1975).
18. Y. Gordon, Some inequalities for Gaussian processes and applications, *Isr. J. Math.* **50**(4), 265–289 (1985).
19. Y. Gordon, Majorization of Gaussian processes and geometric applications, *Probab. Theory Relat. Fields* **91**(2), 251–267 (1992).
20. D. Gross, Recovering low-rank matrices from few coefficients in any basis, *IEEE Trans. Inf. Theory* **57**(3), 1548–1566 (2011).
21. N.J. Higham, *Functions of Matrices: Theory and Computation* (Society for Industrial and Applied Mathematics, Philadelphia, 2008).
22. R.A. Horn, C.R. Johnson, *Matrix Analysis* (Cambridge Univ. Press, Cambridge, 1985).
23. R.A. Horn, C.R. Johnson, *Topics in Matrix Analysis* (Cambridge Univ. Press, Cambridge, 1994).

24. N. Halko, P.-G. Martinsson, J.A. Tropp, Finding structure with randomness: stochastic algorithms for constructing approximate matrix decompositions, *SIAM Rev.* **53**(2), 217–288 (2011).
25. F. Hansen, G.K. Pedersen, Jensen’s operator inequality, *Bull. Lond. Math. Soc.* **35**, 553–564 (2003).
26. M. Junge, Q. Xu, On the best constants in some non-commutative martingale inequalities, *Bull. Lond. Math. Soc.* **37**, 243–253 (2005).
27. M. Junge, Q. Xu, Noncommutative Burkholder/Rosenthal inequalities II: Applications, *Isr. J. Math.* **167**, 227–282 (2008).
28. R. Latała, Some estimates of norms of random matrices, *Proc. Am. Math. Soc.* **133**(5), 1273–1282 (2005).
29. E.H. Lieb, Convex trace functions and the Wigner–Yanase–Dyson conjecture, *Adv. Math.* **11**, 267–288 (1973).
30. G. Lindblad, Expectations and entropy inequalities for finite quantum systems, *Commun. Math. Phys.* **39**, 111–119 (1974).
31. F. Lust-Piquard, Inégalités de Khintchine dans C_p ($1 < p < \infty$), *C. R. Math. Acad. Sci. Paris* **303**(7), 289–292 (1986).
32. F. Lust-Piquard, G. Pisier, Noncommutative Khintchine and Paley inequalities, *Ark. Mat.* **29**(2), 241–260 (1991).
33. M. Ledoux, M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes* (Springer, Berlin, 1991).
34. G. Lugosi, Concentration-of-measure inequalities (2009), Available at <http://www.econ.upf.edu/~lugosi/anu.pdf>.
35. P. Massart, *Concentration Inequalities and Model Selection: Ecole d’Eté de Probabilités de Saint-Flour XXXIII—2003*. Lecture Notes in Mathematics, vol. 1896 (Springer, Berlin, 2007).
36. C. McDiarmid, Concentration, in *Probabilistic Methods for Algorithmic Discrete Mathematics*. Algorithms and Combinatorics, vol. 16 (Springer, Berlin, 1998), pp. 195–248.
37. R. Motwani, P. Raghavan, *Randomized Algorithms* (Cambridge Univ. Press, Cambridge, 1995).
38. A. Nemirovski, Sums of random symmetric matrices and quadratic optimization under orthogonality constraints, *Math. Program., Ser. B* **109**, 283–317 (2007).
39. R.I. Oliveira, Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges (2010), [arXiv:0911.0600](https://arxiv.org/abs/0911.0600).
40. R.I. Oliveira, Sums of random Hermitian matrices and an inequality by Rudelson, *Electron. Commun. Probab.* **15**, 203–212 (2010).
41. B.N. Parlett, *The Symmetric Eigenvalue Problem*. Classics in Applied Mathematics, vol. 20 (Society for Industrial and Applied Mathematics, Philadelphia, 1987).
42. V.I. Paulsen, *Completely Bounded Maps and Operator Algebras*. Cambridge Studies in Advanced Mathematics, vol. 78 (Cambridge Univ. Press, Cambridge, 2002).
43. D. Petz, A survey of certain trace inequalities, in *Functional Analysis and Operator Theory*. Banach Center Publications, vol. 30 (Polish Acad. Sci., Warsaw, 1994), pp. 287–298.
44. G. Pisier, *Introduction to Operator Spaces* (Cambridge Univ. Press, Cambridge, 2003).
45. B. Recht, Simpler approach to matrix completion, *J. Mach. Learn. Res.* (2009, to appear). Available at <http://pages.cs.wisc.edu/brecht/papers/09.Recht.ImprovedMC.pdf>.
46. M. Rudelson, Random vectors in the isotropic position, *J. Funct. Anal.* **164**, 60–72 (1999).
47. M.B. Ruskai, Inequalities for quantum entropy: a review with conditions for equality, *J. Math. Phys.* **43**(9), 4358–4375 (2002). Erratum: *J. Math. Phys.* **46**(1), 0199101 (2005).
48. M. Rudelson, R. Vershynin, Sampling from large matrices: an approach through geometric functional analysis, *J. Assoc. Comput. Mach.* **54**(4), Article 21 (2007) (electronic) 19 pp.
49. Y. Seginer, The expected norm of random matrices, *Comb. Probab. Comput.* **9**, 149–166 (2000).
50. A.M.-C. So, Moment inequalities for sums of random matrices and their applications in optimization, *Math. Prog. Ser. A* (2009) (electronic).
51. A. Sankar, D.A. Spielman, S.-H. Teng, Smoothed analysis of the condition numbers and growth factors of matrices, *SIAM J. Matrix Anal. Appl.* **28**(2), 446–476 (2006).
52. N. Tomczak-Jaegermann, The moduli of smoothness and convexity and the Rademacher averages of trace classes S_p ($1 \leq p < \infty$), *Stud. Math.* **50**, 163–182 (1974).
53. J.A. Tropp, On the conditioning of random subdictionaries, *Appl. Comput. Harmon. Anal.* **25**, 1–24 (2008).
54. J.A. Tropp, Improved analysis of the subsampled randomized Hadamard transform. *Adv. Adapt. Data Anal.* (2010, to appear). Available at [arXiv:1011.1595](https://arxiv.org/abs/1011.1595).
55. J.A. Tropp, Freedman’s inequality for matrix martingales, *Electron. Commun. Probab.* **16**, 262–270 (2011).

56. 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](https://arxiv.org/abs/1101.1070).
57. J.A. Tropp, User-friendly tail bounds for matrix martingales, ACM Report 2011-01, California Inst. Tech., Pasadena, CA (2011).
58. R. Vershynin, A note on sums of independent random matrices after Ahlswede–Winter (2009), Available at <http://www-personal.umich.edu/~romanv/teaching/reading-group/ahlswe-de-winter.pdf>.