USER-FRIENDLY TAIL BOUNDS FOR SUMS OF RANDOM MATRICES

JOEL A. TROPP

Technical Report No. 2010-01
April 2010

APPLIED & COMPUTATIONAL MATHEMATICS
CALIFORNIA INSTITUTE OF TECHNOLOGY
mail code 217-50 · pasadena, ca 91125

USER-FRIENDLY TAIL BOUNDS
FOR SUMS OF RANDOM MATRICES

J. A. TROPP

ABSTRACT. This work presents probability inequalities for sums of independent, random, self-adjoint matrices. The results frame simple, easily verifiable hypotheses on the summands, and they yield strong conclusions about the large-deviation behavior of the maximum eigenvalue of the sum. Tail bounds for the norm of a sum of rectangular matrices follow as an immediate corollary, and similar techniques yield 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 ease of use, diversity of application, and strength of conclusion that have made the scalar inequalities so valuable.

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.

- For numerical applications, it is important to obtain detailed quantitative information about random matrices of finite order. Asymptotic theory has limited value.
- Many problems require explicit large deviation bounds for the extreme eigenvalues of a random matrix. In other cases, we are concerned not with the eigenvalue spectrum but rather with the *action* of a random operator on some class of vectors or matrices.
- In numerical analysis, it is essential to compute effective constants to ensure that an algorithm is provably correct in practice.
- We often 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 signs multiplied by a discrete Fourier transform matrix.
- Other problems involve a sparse matrix sampled from a fixed matrix or a random submatrix drawn from a fixed matrix. These applications lead to random matrices whose distribution varies by coordinate, in contrast to the classical ensembles of random matrices that have i.i.d. entries or i.i.d. columns.

We have encountered these issues in a wide range of problems from computational mathematics: smoothed analysis of Gaussian elimination [SST06]; semidefinite relaxation and rounding of quadratic maximization problems [Nem07, So09]; construction of maps for dimensionality reduction [AC09]; matrix approximation by sparsification [AM07] and by sampling submatrices [RV07];

Date: 25 April 2010. Corrected: 29 April 2010.

Key words and phrases. Discrete-time martingale, large deviation, random matrix, sum of independent random variables.

2010 *Mathematics Subject Classification.* Primary: 60B20. Secondary: 60F10, 60G50, 60G42.

JAT is with Applied and Computational Mathematics, MC 305-16, California Inst. Technology, Pasadena, CA 91125. E-mail: `jtropp@acm.caltech.edu`. Research supported by ONR award N00014-08-1-0883, DARPA award N66001-08-1-2065, and AFOSR award FA9550-09-1-0643.

analysis of sparse approximation [Tro08] and compressive sampling [CR07] problems; randomized schemes for low-rank matrix factorization [HMT09]; and analysis of algorithms for completion [Gro09, Rec09] and decomposition [CSPW09, CLMW09] of low-rank matrices. And this list is by no means comprehensive!

In these applications, the methods currently invoked to study random matrices are often cumbersome, and they require a substantial amount of practice to use effectively. These frustrations have led us to search for simpler techniques that still yield detailed quantitative information about finite random matrices.

Inspired by the work of Ahslwede–Winter [AW02] and Rudelson–Vershynin [Rud99, RV07], we study sums of independent, random, self-adjoint matrices. Our results place simple and easily verifiable hypotheses on the summands that allow us to reach strong conclusions about the large-deviation behavior of the maximum eigenvalue of the sum. These bounds can be viewed as matrix analogs of the probability inequalities associated with the names Azuma, Bennett, Bernstein, Chernoff, Hoeffding, and McDiarmid. We hope that these new matrix inequalities will offer researchers the same ease of use, diversity of application, and strength of conclusion that have made the scalar inequalities so indispensable.

1.1. **Roadmap.** The rest of the paper is organized as follows. Section 2 provides an overview of our main results and a discussion of related work. Section 3 introduces the background required for our proofs, which ranges from the elementary to the esoteric. Section 4 contains the main technical innovations. Sections 5–8 complete the proofs of the matrix probability inequalities. Section 9 describes some complementary results, including the extension to rectangular matrices. We conclude in Section 10 with some open questions.

2. Main Results and Discussion

Our goal has been to extend the most useful of the classical tail bounds to the matrix case, rather than to produce a complete catalog of matrix inequalities. This approach allows us to introduce several different techniques that are useful for making the translation from the scalar to the matrix setting. This section summarizes the main results for easy reference. Section 2.6 describes some additional theorems that may be found deeper inside the paper.

2.1. **Technical Approach.** Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, self-adjoint matrices. We wish to bound the probability

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \geq t\right\}.$$

Here and elsewhere, $\lambda_{\max}$ 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.
- We can extend these methods to matrix-valued martingales.
- We can investigate the probability that the sum satisfies other semidefinite relations.

In the matrix setting, the structure of the main argument parallels established proofs of the classical inequalities. See [McD98, Lug09] for accessible surveys in the scalar setting. First, we describe a suitable generalization of Bernstein's argument, which is sometimes known as the *Laplace transform method*. In the matrix setting, this approach yields the bound

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \geq t\right\} \leq \inf_{\theta > 0}\left\{\mathrm{e}^{-\theta t}\,\mathrm{tr}\exp\left(\sum\nolimits_k \log \mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}_k}\right)\right\}.$$

In words, the probability of a large deviation is controlled by the "cumulant generating functions" of the random matrices. Although this inequality superficially resembles the classical Laplace transform bound for real random variables, the proof is no longer elementary. Our argument relies on a deep inequality of Lieb [Lie73, Thm. 6]. This part of the reasoning appears in Section 4.

As in the scalar case, the second stage of the development uses information about each random matrix to obtain bounds for the "cumulant generating functions." Certain classical methods extend directly to the matrix case, but they usually require additional care. Other proofs do not generalize at all, and we have to identify alternative approaches. Sections 5–8 present these arguments.

Let us emphasize that many of the ideas in this work have appeared in the literature. The primary precedent is the important paper of Ahlswede and Winter [AW02], which develops a matrix analog of the Laplace transform method; see also [Gro09, Rec09]. We have been influenced strongly by Rudelson and Vershynin's approach [Rud99, RV07] to random matrices via the noncommutative Khintchine inequality [LP86, Buc01]. Finally, the recent work of Oliveira [Oli10b] persuaded us that it might be possible to combine the best qualities of these two approaches.

2.2. **Rademacher and Gaussian Series.** For motivation, we begin with the simplest example of a sum of independent random variables: a series with real coefficients modulated by random signs. This discussion illustrates some new phenomena that arise when we try to translate scalar tail bounds to the matrix setting.

Consider a finite sequence $\{a_k\}$ of real numbers and a finite sequence $\{\varepsilon_k\}$ of independent Rademacher variables[1]. A classical result, due to Bernstein, shows that

$$\mathbb{P}\left\{\sum\nolimits_k \varepsilon_k a_k \geq t\right\} \leq \mathrm{e}^{-t^2/2\sigma^2} \quad \text{where } \sigma^2 = \sum\nolimits_k a_k^2. \tag{2.1}$$

In words, a real Rademacher series exhibits normal concentration with variance equal to the sum of the squared coefficients. The central limit theorem guarantees that there are Rademacher series where this estimate is essentially sharp.

What is the correct generalization of (2.1) to random matrices? The approach of Ahlswede and Winter [AW02] suggests the bound

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \varepsilon_k \boldsymbol{A}_k\right) \geq t\right\} \leq d \cdot \mathrm{e}^{-t^2/2\sigma^2} \quad \text{where } \sigma^2 = \sum\nolimits_k \left\|\boldsymbol{A}_k^2\right\|. \tag{2.2}$$

The symbol $\|\cdot\|$ denotes the usual norm for operators on a Hilbert space, which returns the largest singular value of its argument. Although the statement (2.2) identifies a plausible generalization for the variance, this result can be improved dramatically in most cases. Indeed, a matrix Rademacher series satisfies a fundamentally stronger tail bound.

Theorem 2.1 (Matrix Rademacher and Gaussian Series)**.** *Consider a finite sequence $\{\boldsymbol{A}_k\}$ of fixed self-adjoint matrices with dimension d, and let $\{\varepsilon_k\}$ be a sequence of independent Rademacher variables. Compute the norm of the sum of squared coefficient matrices:*

$$\sigma^2 = \left\|\sum\nolimits_k \boldsymbol{A}_k^2\right\|. \tag{2.3}$$

For all $t \geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \varepsilon_k \boldsymbol{A}_k\right) \geq t\right\} \leq d \cdot \mathrm{e}^{-t^2/2\sigma^2}. \tag{2.4}$$

In particular,

$$\mathbb{P}\left\{\left\|\sum\nolimits_k \varepsilon_k \boldsymbol{A}_k\right\| \geq t\right\} \leq 2d \cdot \mathrm{e}^{-t^2/2\sigma^2}. \tag{2.5}$$

The same bounds hold when we replace $\{\varepsilon_k\}$ by a sequence of independent, standard normal random variables.

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

When the dimension $d = 1$, the bound (2.4) reduces to the classical result (2.1). Of course, one may still wonder whether the formula (2.3) for the variance is sharp and whether the dimensional dependence is necessary. Remarks 2.2, 2.3, and 2.4 demonstrate that Theorem 2.1 cannot be improved without changing its form. A casual reader may bypass this discussion without loss of continuity.

The technology required to prove Theorem 2.1 has been available for some time now. One argument applies sharp noncommutative Khintchine inequalities, [Buc01, Thm. 5] and [Buc05, Thm. 5], to bound the moment generating function of the maximum eigenvalue of the random sum. Very recently, Oliveira has developed a different approach [Oli10b, Lem. 2] using a clever variation of Ahlswede and Winter's techniques. We present our proof in Section 7.

Remark 2.2. The matrix variance σ^2 given by (2.3) is truly the correct quantity for controlling large deviations of a matrix Gaussian series. Indeed, it follows from general principles [LT91, Cor. 3.2] that

$$\lim_{t\to\infty} \frac{1}{t^2} \log \mathbb{P}\left\{ \left\| \sum\nolimits_k \gamma_k \boldsymbol{A}_k \right\| \geq t \right\} = -\frac{1}{2\sigma^2}.$$

where $\{\gamma_k\}$ is a sequence of independent, standard normal variables. By the (scalar) central limit theorem, we can construct Rademacher series that exhibit essentially the same large-deviation behavior by repeating each matrix $\boldsymbol{A}_k$ multiple times. (Of course, a finite Rademacher series is almost surely bounded!)

In contrast to a Gaussian series, a Rademacher series can have a constant operator norm. Nevertheless, the matrix variance in (2.3) always provides a lower bound for the supremal norm of the series:

$$\sigma \leq \sup_{\boldsymbol{\varepsilon}} \left\| \sum\nolimits_k \varepsilon_k \boldsymbol{A}_k \right\|.$$

This fact follows easily from the statement of the noncommutative Khintchine inequality in [Rud99, Sec. 3]. A simple example shows that the lower bound is sharp. Let $\mathbf{E}_{ij}$ be the matrix with a unit entry in the (i, j) position and zeros elsewhere, and consider the Rademacher series with coefficients $\boldsymbol{A}_k = \mathbf{E}_{kk}$ for $k = 1, 2, \dots, d$. This example also demonstrates that the bound (2.2) is fundamentally worse than Theorem 2.1.

Remark 2.3. In general, we cannot remove the factor d from the probability bound in Theorem 2.1. Consider the Gaussian series

$$\left\| \sum\nolimits_{k=1}^d \gamma_k \mathbf{E}_{kk} \right\| = \max\nolimits_k |\gamma_k| \geq \mathrm{c}\sqrt{\log d} \quad \text{with high probability.}$$

Since the variance parameter $\sigma^2 = 1$, Theorem 2.1 yields

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

We need the factor d to ensure that the probability bound does not become effective until $t \geq \sqrt{2 \log d}$. The dimensional factor is also necessary in the tail bound for Rademacher series because of the central limit theorem.

Remark 2.4. The dimensional dependence does not appear in standard bounds for Rademacher series in Banach space because they concern the deviation of *the norm of the sum* above its mean value. For example, Ledoux [Led96, Eqn. (1.9)] proves that

$$\mathbb{P}\left\{ \left\| \sum\nolimits_k \varepsilon_k \boldsymbol{A}_k \right\| \geq \mathbb{E} \left\| \sum\nolimits_k \varepsilon_k \boldsymbol{A}_k \right\| + t \right\} \leq \mathrm{e}^{-t^2/8\sigma^2}$$

where σ^2 is given by (2.3). Unfortunately, this formula provides no information about the size of the expectation. In contrast, we can always bound the expectation by integrating (2.5), although the estimate may not be sharp.

2.3. **Sums of Random Semidefinite Matrices.** Having introduced some of the issues that arise in the matrix setting, we are prepared to state matrix extensions of the classical probability inequalities.

We begin with the Chernoff bounds, which describe the upper and lower tails of a sum of nonnegative random variables. In the matrix case, the analogous results concern a sum of positive-semidefinite random matrices. The matrix Chernoff bound shows that the extreme eigenvalues of this sum exhibit the same binomial-type behavior as in the scalar setting.

Theorem 2.5 (Matrix Chernoff)**.** *Consider a finite sequence* $\{\boldsymbol{X}_k\}$ *of independent, random, positive-semidefinite matrices with dimension* d*. Suppose that*

$$\lambda_{\max}(\boldsymbol{X}_k) \le B \quad \textit{almost surely.}$$

Define bounds for the eigenvalues of the sum of the expectations:

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

Then

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \le (1-\delta)\mu_{\min}\right\} \le d \cdot \left[\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/B} \quad \textit{for } \delta \in [0,1), \textit{ and}$$

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \ge (1+\delta)\mu_{\max}\right\} \le d \cdot \left[\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{\max}/B} \quad \textit{for } \delta \ge 0.$$

The proof of Theorem 2.5 appears in Section 5. This result can be viewed as an essential improvement on the matrix Chernoff inequality established by Ahlswede and Winter [AW02, Thm. 19]. The matrix Chernoff bound is also connected with the noncommutative Rosenthal inequality [JX03, JX08].

Remark 2.6. The following standard simplifications of the bounds in Theorem 2.5 are often more useful in practice.

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \le t\mu_{\min}\right\} \le d \cdot \mathrm{e}^{-(1-t)^2\mu_{\min}/2B} \quad \text{for } t \in [0,1)$$

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \ge t\mu_{\max}\right\} \le d \cdot \left[\frac{\mathrm{e}}{t}\right]^{t\mu_{\max}/B} \quad \text{for } t \ge \mathrm{e}.$$

Remark 2.7. The factor d in the Chernoff bounds cannot be omitted because of coupon collection issues. Consider a random matrix $\boldsymbol{X}$ with the distribution

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

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

$$\lambda_{\min}\left(\sum\nolimits_{k=1}^n \boldsymbol{X}_k\right) = 0 \quad \text{with high probability unless } n \ge \mathrm{c}d\log d.$$

The dimensional factor in the lower Chernoff bound reflects this fact. Related examples show that the upper Chernoff bound must also depend on d. We have extracted this idea from [RV07, Sec. 3.5].

2.4. **Normal Concentration for Zero-Mean Sums.** Next, we extend one of Hoeffding's inequalities to the matrix setting. Here and elsewhere, we use the semidefinite order:

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

The inequality demonstrates that a sum of bounded zero-mean random matrices exhibits normal concentration, where the variance is controlled by the maximum squared ranges of the summands.

Theorem 2.8 (Matrix Hoeffding). *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, self-adjoint matrices with dimension d, and let $\{\boldsymbol{A}_k\}$ be a sequence of fixed self-adjoint matrices. Suppose that*

$$\mathbb{E}\,\boldsymbol{X}_k = \boldsymbol{0} \quad \textit{and} \quad \boldsymbol{X}_k^2 \preccurlyeq \boldsymbol{A}_k^2 \quad \textit{almost surely.}$$

Define a bound on the sum of maximum squared ranges:

$$\sigma^2 \geq \left\| \sum\nolimits_k \boldsymbol{A}_k^2 \right\|.$$

For all $t \geq 0$,

$$\mathbb{P}\left\{ \lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k \right) \geq t \right\} \leq d \cdot \mathrm{e}^{-t^2/8\sigma^2}. \tag{2.6}$$

The proof of Theorem 2.8 appears in Section 8. We establish this result as a special case of the matrix Azuma inequality. Cristofides and Markström [CM08, Thm. 9] have established a version of the matrix Hoeffding inequality using the methods of Ahlswede–Winter; their estimate for the variance is not sharp in general.

Remark 2.9. It is not clear whether the factor 1/8 in the exponent of the bound (2.6) can be sharpened to 1/2. Indeed, the standard proof of the Hoeffding inequality [McD98, Lem. 2.6] does not extend to the matrix setting. But there are several interesting cases where the correct factor is 1/2. One situation occurs when each $\boldsymbol{X}_k$ is a symmetric random variable. Another example requires the assumption that $\boldsymbol{X}_k$ commutes with $\boldsymbol{A}_k$ almost surely for each k.

We have observed that Talagrand's inductive method [Tal88, Cor. 4] and the log-Sobolev approach of Ledoux [Led96, Eqn. (1.9)] both produce normal concentration inequalities for Rademacher series in Banach space that have a factor of 1/8 in the exponent. As a result, it seems plausible to us that the constant actually does change outside the scalar setting.

2.5. **Adding Variance Information.** A sum of independent random variables may vary substantially less than the Hoeffding bound suggests. In this situation, another inequality of Bernstein shows that the sum exhibits normal concentration near its mean with variance controlled by the variance of the sum. On the other hand, the tail of the sum decays subexponentially on a scale determined by a uniform upper bound for 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 for the eigenvalues of the summands.

Theorem 2.10 (Matrix Bernstein). *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, self-adjoint matrices with dimension d. Suppose that*

$$\mathbb{E}\,\boldsymbol{X}_k = \boldsymbol{0} \quad \textit{and} \quad \|\boldsymbol{X}_k\| \leq B \quad \textit{almost surely.}$$

Define a bound on the total variance:

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

For all $t \geq 0$,

$$\mathbb{P}\left\{ \lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k \right) \geq t \right\} \leq d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Bt/3} \right).$$

In Section 6, we derive Theorem 2.10 from a matrix extension of Bennett's inequality. The matrix Bennett inequality provides Poisson-type decay for the tail, rather than the weaker exponential decay described in the Bernstein inequality. On the other hand, the detailed result is used less often in practice.

Gross [Gro09, Thm. 5] and Recht [Rec09, Thm. 3.2] have both developed matrix extensions of the Bernstein inequality using the ideas of Ahlswede–Winter [AW02]. These arguments do not produce the sharp generalization of the variance described in Theorem 2.10. Oliveira has obtained a matrix

Bernstein inequality for sums of rank-one random matrices [Oli10b, Lem. 1] that yields the correct variance, but his argument does not extend to more general cases and the constants are not sharp. The matrix Bernstein and Bennett inequalities are closely connected with the noncommutative Rosenthal inequality [JX03, JX08].

2.6. **Miscellaneous Results.** This paper contains a number of other results that we summarize here for reference. We postpone a detailed discussion.

- We can also produce generalizations of certain martingale deviation bounds, such as the Azuma inequality and the McDiarmid bounded differences inequality. See Section 8 for these results. Unfortunately, we have not been successful in generalizing the strongest martingale bounds, as we discuss in Section 10.1.
- All the inequalities we have mentioned, with exception of the matrix Chernoff bounds, have variants that hold for non-self-adjoint and rectangular matrices. This extension is achieved by means of an elegant device from operator theory, called the self-adjoint dilation of a matrix. See Sections 3.1.12 and 9.1 for the details of this approach.
- We can also study semidefinite relations more complicated than eigenvalue bounds by applying additional ideas from the work of Ahlswede and Winter [AW02]. We describe the relevant techniques in Section 9.2.

2.7. **Related Work.** The most important precedent for our work is the influential paper of Ahlswede and Winter [AW02]. This work describes a matrix version of the Laplace transform. The authors then use elegant techniques from statistical quantum mechanics to obtain a matrix extension of the Chernoff bound [AW02, Thm. 19]. We discuss their ideas in Remark 4.1 to emphasize how the current approach differs.

Several other papers use essentially the same techniques as Ahlswede–Winter to obtain matrix versions of the classical probability inequalities. Cristofides and Markström [CM08] have established a matrix version of the Hoeffding inequality. Gross [Gro09] and Recht [Rec09] both develop extensions of Bernstein's inequality, as noted in Section 2.5. We also refer the reader to Vershynin's note [Ver09], which offers a nice introduction to this circle of ideas.

Very recently, Oliveira [Oli10b] has developed an essential improvement over the Ahlswede–Winter method. He uses this idea to obtain Theorem 2.1 and some related results. Oliveira has also used his technique to obtain a matrix Freedman inequality [Oli10a], which is sharp up to the precise value of the constants.

There is another contemporary line of work that uses noncommutative moment inequalities to study random matrices. In a significant paper [Rud99], Rudelson obtains an optimal estimate for the sample complexity of approximating the covariance matrix of an isotropic distribution. His argument relies on a version of the noncommutative Khintchine moment inequality [LP86, Buc01].

Rudelson's technique has been used widely over the last ten years, and it has emerged as one of the most powerful tools available for studying discrete random matrices. Typical applications include a bound for the norm of a random submatrix [RV07, Thm. 1.8] and an analysis of the randomized Fourier transform [HMT09, App. B].

By now, there is a substantial literature on noncommutative moment inequalities more general than the noncommutative Khintchine inequality. See the article [JX05] for a reasonably accessible and comprehensive discussion. These results can be used directly to study random matrices. See [JX08], for example. This approach requires expertise in noncommutative probability theory, and it does not produce explicit constants.

3. Preliminaries

This section provides a short introduction to the background material we use in our proofs. Section 3.1 discusses matrix theory, and Section 3.2 reviews some relevant results from probability.

3.1. **Matrix Theory.** Most of these results are drawn from Bhatia's books on matrix analysis [Bha97, Bha07]. Horn and Johnson's books [HJ85, HJ94] also serve as good general references. Higham's book [Hig08] is an excellent source for information about matrix functions.

3.1.1. *Conventions.* 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 order of a matrix when it is determined by the context.

3.1.2. *Basic Matrices.* We write $\mathbf{0}$ for the zero matrix and $\mathbf{I}$ for the identity matrix. Occasionally, we add a subscript to specify the dimension, e.g., $\mathbf{I}_d$ is the $d \times d$ identity.

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

3.1.3. *Self-Adjoint Matrices and Eigenvalues.* A square matrix that satisfies $\boldsymbol{A} = \boldsymbol{A}^*$ is called *self-adjoint*, or more briefly *s.a.* We adopt Parlett's convention that letters symmetric around the vertical axis ($\boldsymbol{A}$, $\boldsymbol{H}$, …, $\boldsymbol{Y}$) represent s.a. matrices unless otherwise noted.

Each s.a. matrix $\boldsymbol{A}$ has an *eigenvalue decomposition*

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

The diagonal entries of $\boldsymbol{\Lambda}$ are called the *eigenvalues* of $\boldsymbol{A}$. The algebraic maximum and minimum eigenvalues of $\boldsymbol{A}$ are denoted by $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$.

3.1.4. *Trace.* The *trace* of a matrix, denoted by tr, is the sum of its diagonal entries. The trace of a matrix is also equal to the sum of its eigenvalues. The trace admits a Cauchy–Schwarz inequality:

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{H}) \le \left[\operatorname{tr}\big(\boldsymbol{A}^2\big)\right]^{1/2} \left[\operatorname{tr}\big(\boldsymbol{H}^2\big)\right]^{1/2} \quad \text{for s.a. } \boldsymbol{A}, \boldsymbol{H}. \tag{3.1}$$

3.1.5. *The Semidefinite Order.* An s.a. matrix $\boldsymbol{A}$ with nonnegative eigenvalues is called *positive semidefinite* (briefly, *psd*). When the eigenvalues are strictly positive, we say the matrix is *positive definite* (briefly, *pd*). An easy consequence of the definition is that

$$\lambda_{\max}(\boldsymbol{A}) \le \operatorname{tr} \boldsymbol{A} \quad \text{when } \boldsymbol{A} \text{ is psd} \tag{3.2}$$

because the trace is the sum of the eigenvalues.

The set of all psd matrices with fixed dimension forms a closed, convex cone. Therefore, we may define the *semidefinite partial order* on s.a. matrices of the same size by the rule

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \Longleftrightarrow \quad \boldsymbol{H} - \boldsymbol{A} \text{ is psd.}$$

In particular, we may write $\boldsymbol{A} \succcurlyeq \mathbf{0}$ to indicate that $\boldsymbol{A}$ is psd and $\boldsymbol{A} \succ \mathbf{0}$ to indicate that $\boldsymbol{A}$ is pd. For a diagonal matrix, $\boldsymbol{\Lambda} \succcurlyeq \mathbf{0}$ means that each entry of $\boldsymbol{\Lambda}$ is nonnegative.

The semidefinite order is preserved by conjugation:

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \Longrightarrow \quad \boldsymbol{B}^*\boldsymbol{A}\boldsymbol{B} \preccurlyeq \boldsymbol{B}^*\boldsymbol{H}\boldsymbol{B} \quad \text{for each matrix } \boldsymbol{B}. \tag{3.3}$$

We refer to (3.3) as the *conjugation rule.* A more general result is Sylvester's *inertia theorem*, which states that an s.a. matrix $\boldsymbol{A}$ and its conjugate $\boldsymbol{B}^*\boldsymbol{A}\boldsymbol{B}$ have the same number of negative eigenvalues, provided that $\boldsymbol{B}$ is nonsingular [HJ85, Thm. 4.5.8].

3.1.6. *Matrix Functions.* Let us describe the most direct method for extending functions on the reals to functions on s.a. matrices. Consider a function $f : \mathbb{R} \to \mathbb{R}$. First, extend f to a map on diagonal matrices by applying the function to each diagonal entry:

$$(f(\mathbf{\Lambda}))_{jj} = f(\mathbf{\Lambda}_{jj}) \quad \text{for each index } j.$$

We extend f to all s.a. matrices by way of the eigenvalue decomposition. If $\boldsymbol{A} = \boldsymbol{Q}\mathbf{\Lambda}\boldsymbol{Q}^*$, then

$$f(\boldsymbol{A}) = f(\boldsymbol{Q}\mathbf{\Lambda}\boldsymbol{Q}^*) = \boldsymbol{Q}f(\mathbf{\Lambda})\boldsymbol{Q}^*.$$

The *spectral mapping theorem* states that each eigenvalue of $f(\boldsymbol{A})$ has the form $f(\lambda)$, where λ is an eigenvalue of $\boldsymbol{A}$. This point is obvious from the definition.

Inequalities for real functions extend to semidefinite relationships for matrix functions:

$$f(a) \leq g(a) \quad \text{for } a \in I \quad \Longrightarrow \quad f(\boldsymbol{A}) \preccurlyeq g(\boldsymbol{A}) \quad \text{when the eigenvalues of } \boldsymbol{A} \text{ sit in } I. \tag{3.4}$$

Indeed, let us decompose $\boldsymbol{A} = \boldsymbol{Q}\mathbf{\Lambda}\boldsymbol{Q}^*$. It is immediate that $f(\mathbf{\Lambda}) \preccurlyeq g(\mathbf{\Lambda})$. Conjugate by $\boldsymbol{Q}$, as justified by (3.3), and invoke the definition of a matrix function. We sometimes refer to (3.4) as the *transfer rule*.

When a real function has a power series expansion, we can also define an s.a. matrix function via the same power series expansion:

$$f(a) = c_0 + \sum\nolimits_{j=1}^{\infty} c_j a^j \quad \Longrightarrow \quad f(\boldsymbol{A}) = c_0\mathbf{I} + \sum\nolimits_{j=1}^{\infty} c_j \boldsymbol{A}^j.$$

In this case, the two definitions of a matrix function coincide.

Beware: One must never take for granted that a standard property of a real function generalizes to the associated matrix function.

3.1.7. *Square and Square Root.* Consider an s.a. matrix $\boldsymbol{A}$ with eigenvalue decomposition $\boldsymbol{A} = \boldsymbol{Q}\mathbf{\Lambda}\boldsymbol{Q}^*$. The matrix square is, perhaps, the simplest matrix function:

$$\boldsymbol{A}^2 = \boldsymbol{Q}\mathbf{\Lambda}^2\boldsymbol{Q}^*$$

The square of an s.a. matrix is always psd. Each psd matrix has a unique psd square root:

$$\boldsymbol{A}^{1/2} = \boldsymbol{Q}\mathbf{\Lambda}^{1/2}\boldsymbol{Q}^* \quad \text{when } \boldsymbol{A} \text{ is psd.}$$

It is understood that we always extract the positive square root of a positive number.

3.1.8. *Modulus and Operator Norm.* We define the modulus of an s.a. matrix $\boldsymbol{A} = \boldsymbol{Q}\mathbf{\Lambda}\boldsymbol{Q}^*$ as

$$|\boldsymbol{A}| = (\boldsymbol{A}^2)^{1/2} = \boldsymbol{Q}\,|\mathbf{\Lambda}|\,\boldsymbol{Q}^*$$

For each rectangular matrix $\boldsymbol{B}$, the square matrices $\boldsymbol{B}\boldsymbol{B}^*$ and $\boldsymbol{B}^*\boldsymbol{B}$ are always psd. As a result, we may define the column and row moduli of a rectangular matrix as

$$|\boldsymbol{B}|_{\text{col}} = (\boldsymbol{B}\boldsymbol{B}^*)^{1/2} \quad \text{and} \quad |\boldsymbol{B}|_{\text{row}} = (\boldsymbol{B}^*\boldsymbol{B})^{1/2}. \tag{3.5}$$

For s.a. matrices, all three moduli are equal.

The *operator norm* of an s.a. matrix $\boldsymbol{A}$ is defined by the relations

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

For a general matrix $\boldsymbol{B}$, the operator norm satisfies

$$\|\boldsymbol{B}\| = \lambda_{\max}(|\boldsymbol{B}|_{\text{row}}) = \lambda_{\max}(|\boldsymbol{B}|_{\text{col}}).$$

3.1.9. *The Matrix Exponential.* We may define the matrix exponential for an s.a. matrix $\boldsymbol{A}$ via the power series

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

The exponential of an s.a. matrix is always pd because of the spectral mapping theorem.

On account of the transfer rule (3.4), the matrix exponential satisfies some simple semidefinite relations that we collect here. Since $1 + a \le \mathrm{e}^a$ for real a, we have

$$\mathbf{I} + \boldsymbol{A} \preccurlyeq \mathrm{e}^{\boldsymbol{A}} \quad \text{for each s.a. matrix } \boldsymbol{A}. \tag{3.6}$$

By comparing Taylor series, one verifies that $\cosh(a) \le \mathrm{e}^{a^2/2}$ for real a. Therefore,

$$\cosh(\boldsymbol{A}) \preccurlyeq \mathrm{e}^{\boldsymbol{A}^2/2} \quad \text{for each s.a. matrix } \boldsymbol{A}. \tag{3.7}$$

We often work with the trace of the matrix exponential

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

The trace exponential is a convex function. It is also monotone with respect to the semidefinite order:

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \Longrightarrow \quad \operatorname{tr} \mathrm{e}^{\boldsymbol{A}} \le \operatorname{tr} \mathrm{e}^{\boldsymbol{H}}. \tag{3.8}$$

See [Pet94, Sec. 2] for some discussion of these facts.

The matrix exponential does not convert sums into products, as in the scalar case, but the trace exponential has a related property that serves as a limited substitute. The Golden–Thompson inequality states that

$$\operatorname{tr} \mathrm{e}^{\boldsymbol{A}+\boldsymbol{H}} \le \operatorname{tr}\left(\mathrm{e}^{\boldsymbol{A}}\mathrm{e}^{\boldsymbol{H}}\right) \quad \text{for all s.a. } \boldsymbol{A}, \boldsymbol{H}. \tag{3.9}$$

The bound (3.9) does not extend to three matrices in any simple way.

3.1.10. *The Matrix Logarithm.* The matrix logarithm is defined as the functional inverse of the matrix exponential:

$$\log\left(\mathrm{e}^{\boldsymbol{A}}\right) = \boldsymbol{A} \quad \text{for each s.a. matrix } \boldsymbol{A}.$$

The matrix logarithm is monotone with respect to the semidefinite order.

$$\mathbf{0} \prec \boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \Longrightarrow \quad \log(\boldsymbol{A}) \preccurlyeq \log(\boldsymbol{H}). \tag{3.10}$$

3.1.11. *An Inequality of Lieb.* The key new idea in this work requires a deep inequality of Lieb from his seminal 1973 paper on convex trace functions [Lie73, Thm. 6]. Epstein provides an alternative proof of this bound in [Eps73, Sec. II], and Ruskai offers a simplified account of Epstein's argument in [Rus02, Rus05].

Theorem 3.1 (Lieb)**.** *Fix a self-adjoint matrix $\boldsymbol{A}$. The function*

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

is concave on the positive-definite cone.

Lieb used related ideas to establish a (complicated) extension of the Golden–Thompson trace inequality for three matrices [Lie73, Thm. 7].

3.1.12. *Dilations.* An extraordinarily fruitful idea from operator theory is to embed matrices within larger block matrices, called *dilations* [Pau86]. The *s.a. dilation* of a rectangular matrix $\boldsymbol{B}$ is

$$\mathscr{S}(\boldsymbol{B}) = \begin{bmatrix} \mathbf{0} & \boldsymbol{B} \\ \boldsymbol{B}^* & \mathbf{0} \end{bmatrix}. \tag{3.11}$$

The s.a. dilation is a linear map. Evidently, $\mathscr{S}(\boldsymbol{B})$ is always s.a. At the same time, the dilation retains all the information about the modulus:

$$|\mathscr{S}(\boldsymbol{B})| = \begin{bmatrix} |\boldsymbol{B}|_{\text{col}} & \mathbf{0} \\ \mathbf{0} & |\boldsymbol{B}|_{\text{row}} \end{bmatrix}. \tag{3.12}$$

It can be verified that

$$\lambda_{\max}(\mathscr{S}(\boldsymbol{B})) = \|\mathscr{S}(\boldsymbol{B})\| = \|\boldsymbol{B}\|. \tag{3.13}$$

3.2. **Probability Background.** We continue with some material from probability, focusing on connections with matrices.

3.2.1. *Conventions.* We prefer to avoid abstraction and unnecessary technical detail, so we frame the standing assumption that all random variables are sufficiently regular that we are justified in computing expectations, interchanging limits, and so forth. Furthermore, we often state that a random variable satisfies some relation and omit the qualification "almost surely."

3.2.2. *Random Matrices.* Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, and let $\mathbb{M}_{d_1 \times d_2}$ be the set of $d_1 \times d_2$ complex matrices. A *random matrix* is a measurable map

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

It is more natural to think of the entries of $\boldsymbol{Z}$ as complex random variables that may or may nor be correlated with each other. We reserve the letters $\boldsymbol{X}, \boldsymbol{Y}$ for random s.a. matrices.

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

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

for any collection of events $\{E_k\}$.

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

$$[\mathbb{E}\,\boldsymbol{Z}]_{jk} = \mathbb{E}\,Z_{jk}.$$

Under mild assumptions, expectation commutes with linear maps. In particular, expectation commutes with the s.a. dilation:

$$\mathbb{E}\,\mathscr{S}(\boldsymbol{Z}) = \mathscr{S}(\mathbb{E}\,\boldsymbol{Z}).$$

3.2.4. *Inequalities for Expectation.* Markov's inequality states that a nonnegative (real) random variable X obeys the probability estimate

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}\,X}{t}. \tag{3.14}$$

The Cauchy–Schwarz inequality for real random variables X and Y states that

$$\mathbb{E}(XY) \leq \left[\mathbb{E}(X^2)\right]^{1/2} \left[\mathbb{E}(Y^2)\right]^{1/2}. \tag{3.15}$$

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

$$\begin{aligned} \mathbb{E}\,f(\boldsymbol{Z}) &\leq f(\mathbb{E}\,\boldsymbol{Z}) \quad \text{when } f \text{ is concave, and} \\ f(\mathbb{E}\,\boldsymbol{Z}) &\leq \mathbb{E}\,f(\boldsymbol{Z}) \quad \text{when } f \text{ is convex.} \end{aligned} \tag{3.16}$$

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:

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

Finally, we note that the matrix convexity of the matrix square implies

$$(\mathbb{E}\,\boldsymbol{X})^2 \preccurlyeq \mathbb{E}\big(\boldsymbol{X}^2\big). \tag{3.17}$$

The relation (3.17) is a special case of Kadison's inequality [Bha07, Thm. 2.3.2].

4. Tail Bounds via the Laplace Transform Method

This section describes how to bring the Laplace transform method for producing probability inequalities to the matrix setting. We demonstrate that Lieb's inequality, Theorem 3.1, provides a substitute for the favorable properties of the cumulant generating function.

4.1. The Classical Case.

Let us begin with a review of the classical ideas so we can see where they break down in the matrix setting. Suppose X is a real random variable that has moments of all orders. The *moment generating function* (mgf) packages the moments into a single object:

$$M_X(\theta) = \mathbb{E}\,\mathrm{e}^{\theta X} = 1 + \sum\nolimits_{j=1}^{\infty} \frac{\theta^j\,\mathbb{E}(X^j)}{j!} \quad \text{for } \theta \in \mathbb{R}.$$

The *cumulant generating function* (cgf) is the logarithm of the mgf:

$$C_X(\theta) = \log \mathbb{E}\,\mathrm{e}^{\theta X} = \sum\nolimits_{j=1}^{\infty} \frac{\theta^j \kappa_j}{j!} \quad \text{for } \theta \in \mathbb{R}.$$

Each cumulant κ_j can be expressed as a polynomial function of the moments up to order j. In particular, the first cumulant is the mean and the second is the variance:

$$\kappa_1 = \mathbb{E}\,X \quad \text{and} \quad \kappa_2 = \mathbb{E}\big(X^2\big) - (\mathbb{E}\,X)^2.$$

The mgf and cgf are extremely useful for studying sums of independent random variables because they decompose nicely. Indeed, suppose that $Y = \sum_k X_k$ where $\{X_k\}$ is an independent family of random variables that have moments of all orders. Then the mgf of the sum satisfies

$$M_Y(\theta) = \mathbb{E}\,\mathrm{e}^{\sum_k \theta X_k} = \mathbb{E}\prod\nolimits_k \mathrm{e}^{\theta X_k} = \prod\nolimits_k \mathbb{E}\,\mathrm{e}^{\theta X_k} = \prod\nolimits_k M_{X_k}(\theta). \tag{4.1}$$

This calculation relies on the fact that the scalar exponential function converts sums to products, a property the matrix exponential does not share. An immediate consequence of (4.1) is that the cgf is additive:

$$C_Y(\theta) = \sum\nolimits_k C_{X_k}(\theta). \tag{4.2}$$

We argue that the latter property has a companion in the matrix setting.

The cgf plays a central role in the classical theory of large deviations. The famous argument of Bernstein shows that, for each $\theta > 0$,

$$\mathbb{P}\{Y \geq t\} = \mathbb{P}\left\{\mathrm{e}^{\theta Y} \geq \mathrm{e}^{-\theta t}\right\} \leq \mathrm{e}^{-\theta t} \cdot \mathbb{E}\,\mathrm{e}^{\theta Y} = \mathrm{e}^{-\theta t + C_Y(\theta)},$$

where the second relation is Markov's inequality (3.14). When Y is a sum of independent variables, the cgf decomposes, as in (4.2). Finally, we optimize the right-hand side with respect to θ to achieve

$$\mathbb{P}\{Y \geq t\} \leq \inf_{\theta > 0}\left\{\exp\left(-\theta t + \sum\nolimits_k C_{X_k}(\theta)\right)\right\}. \tag{4.3}$$

Most of the classical large deviation results follow from the formula (4.3) once we construct appropriate upper bounds for the cgfs. The simplest example is Bernstein's bound on the sum of independent random signs. Suppose that $X_k = \varepsilon_k$, where $\{\varepsilon_k\}$ is a sequence of independent Rademacher variables. Then

$$C_{X_k}(\theta) = \log \mathbb{E}\, \mathrm{e}^{\theta \varepsilon_k} = \log \cosh(\theta) \leq \theta^2/2.$$

The formula (4.3) results in the bound

$$\mathbb{P}\left\{\sum\nolimits_{k=1}^{n} \varepsilon_k \geq nt\right\} \leq \inf_{\theta>0} \left\{\mathrm{e}^{-n\theta t + n\theta^2/2}\right\} = \mathrm{e}^{-nt^2/2}.$$

4.2. **Extension to Matrices.** To apply the Laplace transform method, we need to find the correct generalization of the mgf and the cgf. Suppose that $\boldsymbol{X}$ is an s.a. matrix that has moments of all orders. For a real parameter θ, we define the matrix-valued functions

$$\boldsymbol{M}_{\boldsymbol{X}}(\theta) = \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}} \quad \text{and} \quad \boldsymbol{C}_{\boldsymbol{X}}(\theta) = \log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}}.$$

Unfortunately, these functions are difficult to work with directly because they lack most of the favorable properties of their scalar counterparts.

This work proceeds from the insight that Lieb's inequality, Theorem 3.1, offers a completely satisfactory way to generalize the additivity rule (4.2) for cgfs to the matrix setting. Suppose that $\boldsymbol{Y} = \sum_k \boldsymbol{X}_k$ for a finite sequence $\{\boldsymbol{X}_k\}$ of random s.a. matrices. In the next section, we demonstrate that

$$\operatorname{tr} \boldsymbol{M}_{\boldsymbol{Y}}(\theta) = \operatorname{tr} \exp\left(\boldsymbol{C}_{\boldsymbol{Y}}(\theta)\right) \leq \operatorname{tr} \exp\left(\sum\nolimits_k \boldsymbol{C}_{\boldsymbol{X}_k}(\theta)\right).$$

We propose that this formula is the appropriate extension of (4.2) to the matrix setting.

Remark 4.1. In their work, Ahlswede and Winter are clearly searching for the right generalization of the additivity rule (4.2) for cgfs, which is evident from [AW02, App., Sec. E] and [AW03]. Lacking an additivity result, they attempt to parallel the multiplicative property (4.1) of the mgf by using the observation that, when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent,

$$\operatorname{tr} \boldsymbol{M}_{\boldsymbol{X}+\boldsymbol{Y}}(\theta) \leq \operatorname{tr}\left[(\mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}})(\mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{Y}})\right] = \operatorname{tr}\left[\boldsymbol{M}_{\boldsymbol{X}}(\theta) \cdot \boldsymbol{M}_{\boldsymbol{Y}}(\theta)\right]. \tag{4.4}$$

The first relation is the Golden–Thompson inequality (3.9). Unfortunately, this inequality discards too much information because it separates the random matrices into two different exponentials, where they can never be reunited. Additional difficulties arise because Golden–Thompson does not extend to three matrices.

4.3. **"Subadditivity" for Matrix Cumulants.** The following theorem is our main technical result. It encapsulates the calculations used to extend the additivity rule (4.2) for cgfs to the matrix setting.

Theorem 4.2. *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, self-adjoint matrices and a sequence $\{\boldsymbol{A}_k\}$ of fixed self-adjoint matrices that satisfy the relations*

$$\log\left(\mathbb{E}\, \mathrm{e}^{\boldsymbol{X}_k}\right) \preccurlyeq \boldsymbol{A}_k.$$

Then

$$\mathbb{E} \operatorname{tr} \exp\left(\sum\nolimits_k \boldsymbol{X}_k\right) \leq \operatorname{tr} \exp\left(\sum\nolimits_k \boldsymbol{A}_k\right).$$

In particular, it suffices to assume that $\mathbb{E}\, \mathrm{e}^{\boldsymbol{X}_k} \preccurlyeq \mathrm{e}^{\boldsymbol{A}_k}$.

Proof. Let $\mathbb{E}_k$ denote the expectation conditioned on $\boldsymbol{X}_1, \dots, \boldsymbol{X}_k$. It is convenient to abbreviate the exponentials

$$\boldsymbol{Y}_k = \mathrm{e}^{\boldsymbol{X}_k}.$$

We also define the discrepancy terms

$$\boldsymbol{\Delta}_k = \boldsymbol{A}_k - \log(\mathbb{E}_{k-1} \boldsymbol{Y}_k),$$

and we observe that each $\boldsymbol{\Delta}_k$ is psd by hypothesis.

The result is a straightforward consequence of Lieb's inequality, Theorem 3.1, and the monotonicity (3.8) of the trace exponential. We detail the first step of the iterative argument.

$$\begin{aligned}\mathbb{E}\operatorname{tr}\exp\Big(\sum\nolimits_{k=1}^{n}\boldsymbol{X}_k\Big) &= \mathbb{E}_0\,\mathbb{E}_1\cdots\mathbb{E}_{n-1}\operatorname{tr}\exp\Big(\sum\nolimits_{k=1}^{n-1}\boldsymbol{X}_k+\log(\boldsymbol{Y}_n)\Big)\\ &\le \mathbb{E}_0\,\mathbb{E}_1\cdots\mathbb{E}_{n-2}\operatorname{tr}\exp\Big(\sum\nolimits_{k=1}^{n-1}\boldsymbol{X}_k+\log(\mathbb{E}_{n-1}\boldsymbol{Y}_n)\Big)\\ &= \mathbb{E}_0\,\mathbb{E}_1\cdots\mathbb{E}_{n-2}\operatorname{tr}\exp\Big(\sum\nolimits_{k=1}^{n-1}\boldsymbol{X}_k+\boldsymbol{A}_n-\boldsymbol{\Delta}_n\Big)\\ &\le \mathbb{E}\operatorname{tr}\exp\Big(\sum\nolimits_{k=1}^{n-1}\boldsymbol{X}_k+\boldsymbol{A}_n\Big).\end{aligned}$$

The first identity holds by the tower property of conditional expectation and the definition of the matrix logarithm. The second relation follows from Jensen's inequality (3.16) after we invoke Lieb's inequality, Theorem 3.1, to verify the concavity of

$$\boldsymbol{Y}_n \longmapsto \operatorname{tr}\exp\Big(\sum\nolimits_{k=1}^{n-1}\boldsymbol{X}_k+\log(\boldsymbol{Y}_n)\Big).$$

The third line recalls the definition of the discrepancy matrix. The final inequality depends on the trace monotonicity (3.8) of the matrix exponential.

For a given index m, the random matrices $\boldsymbol{X}_1,\dots,\boldsymbol{X}_{m-1}$ do not depend on $\boldsymbol{Y}_m$ and the matrices $\boldsymbol{A}_1,\dots,\boldsymbol{A}_n$ are fixed, so no obstacle prevents us from repeating this process to draw the expectations inside the exponent one by one. Indeed, Lieb's inequality, Theorem 3.1, establishes concavity of

$$\boldsymbol{Y}_m \longmapsto \operatorname{tr}\exp\Big(\sum\nolimits_{k=1}^{m-1}\boldsymbol{X}_k+\sum\nolimits_{k=m+1}^{n}\boldsymbol{A}_k+\log(\boldsymbol{Y}_m)\Big)\quad\text{for each } m=1,2,\dots,n.$$

This observation completes the main part of the argument.

The final point is to demonstrate that the hypotheses of the theorem are fulfilled by the relations $\mathbb{E}\,\mathrm{e}^{\boldsymbol{X}_k}\preccurlyeq\mathrm{e}^{\boldsymbol{A}_k}$. But this claim follows directly from the matrix monotonicity (3.10) of the matrix logarithm. □

Remark 4.3. We have structured the proof of Theorem 4.2 to emphasize that the argument generalizes to martingales. See Section 8 for some results for martingales.

4.4. **The Laplace Transform Method for Matrices.** Next, we demonstrate that we can control tail probabilities for the maximum eigenvalue of a random matrix by using the matrix mgf. This extension of the Laplace transform method essentially goes back to the work of Ahlswede and Winter [AW02]. The account here follows the same lines as the presentation in [Oli10b].

Proposition 4.4 (Laplace Transform Method)**.** *Let $\boldsymbol{X}$ be a random self-adjoint matrix. Then*

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{X})\ge t\}\le\inf_{\theta>0}\Big\{\mathrm{e}^{-\theta t}\cdot\mathbb{E}\operatorname{tr}\mathrm{e}^{\theta\boldsymbol{X}}\Big\}.$$

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

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{X})\ge t\}=\mathbb{P}\{\lambda_{\max}(\theta\boldsymbol{X})\ge\theta t\}=\mathbb{P}\Big\{\mathrm{e}^{\lambda_{\max}(\theta\boldsymbol{X})}\ge\mathrm{e}^{\theta t}\Big\}\le\mathrm{e}^{-\theta t}\cdot\mathbb{E}\,\mathrm{e}^{\lambda_{\max}(\theta\boldsymbol{X})}.$$

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 (3.14).

The exponential can be bounded as follows.

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

The identity is the spectral mapping theorem. The inequality follows from the property that the exponential of an s.a. matrix is pd and the fact (3.2) that the maximum eigenvalue of a pd matrix cannot exceeds its trace.

Combine these relations to reach

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

Since this bound holds for any positive θ, we may take the infimum to complete the proof. □

4.5. **The Main Result.** Finally, we combine Theorem 4.2, which bounds the matrix mgf of a sum of random matrices, with the Laplace transform bound of Proposition 4.4 to obtain our key result.

Theorem 4.5. *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random self-adjoint matrices and a sequence $\{\boldsymbol{A}_k(\theta)\}$ of fixed functions that take self-adjoint matrix values. Suppose that*

$$\log\big(\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}_k}\big) \preccurlyeq \boldsymbol{A}_k(\theta) \quad \text{for all } \theta \in \Theta$$

where Θ is a set of positive numbers. In particular, it suffices to assume that

$$\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq \mathrm{e}^{\boldsymbol{A}_k(\theta)} \quad \text{for all } \theta \in \Theta.$$

Then

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \geq t\right\} \leq \inf_{\theta \in \Theta}\left\{\mathrm{e}^{-\theta t} \cdot \operatorname{tr}\exp\left(\sum\nolimits_k \boldsymbol{A}_k(\theta)\right)\right\}.$$

We conclude this section with a few additional remarks on some important situations that are also covered by this theorem.

Remark 4.6. We can use Theorem 4.5 to study the *minimum* eigenvalue of a sum of random s.a. matrices because $\lambda_{\min}(\boldsymbol{X}) = -\lambda_{\max}(-\boldsymbol{X})$. As a result,

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

Of course, we need a semidefinite bound for the cgf of each $-\boldsymbol{X}_k$ to use the theorem. In Section 5, we apply this observation to develop lower Chernoff bounds.

Remark 4.7. In Section 9.1, we study the maximum singular value of a sum of random rectangular matrices by applying Theorem 4.5 to the s.a. dilation (3.11). For a finite sequence $\{\boldsymbol{Z}_k\}$ of independent, random, rectangular matrices, we have

$$\mathbb{P}\left\{\left\|\sum\nolimits_k \boldsymbol{Z}_k\right\| \geq t\right\} = \mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \mathscr{S}(\boldsymbol{Z}_k)\right) \geq t\right\}$$

on account of (3.13) and the linearity of the dilation. The theorem now requires a semidefinite bound for the cgf of each $\mathscr{S}(\boldsymbol{Z}_k)$.

5. Sums of Random Semidefinite Matrices

In this section, we establish the matrix Chernoff bounds. We begin with a semidefinite bound for the moment generating function of a random psd matrix. This argument transfers a linear upper bound for the scalar exponential to the matrix case.

Lemma 5.1 (Chernoff mgf)**.** *Suppose that $\boldsymbol{X}$ is a random psd matrix that satisfies $\lambda_{\max}(\boldsymbol{X}) \leq 1$. Then*

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

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

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

More explicitly,

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

Since the eigenvalues of $\boldsymbol{X}$ lie in the interval $[0, 1]$, the transfer rule (3.4) implies that

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

The expectation respects the semidefinite order, so

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

where the second relation is (3.6). □

Using this bound on the mgf and the Laplace transform approach, we quickly obtain both the upper and lower matrix Chernoff bound.

Proof of Theorem 2.5, Upper Bound. By homogeneity, we can take $B = 1$. Each summand satisfies the bound on the mgf given in Lemma 5.1:

$$\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq \exp\left((\mathrm{e}^{\theta} - 1)(\mathbb{E}\,\boldsymbol{X}_k)\right).$$

Invoke Theorem 4.5 to obtain

$$\begin{aligned}
\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \geq (1+\delta)\mu_{\max}\right\} &\leq \inf_{\theta>0} \left\{\mathrm{e}^{-\theta(1+\delta)\mu_{\max}} \cdot \operatorname{tr}\exp\left((\mathrm{e}^{\theta}-1)\sum\nolimits_k (\mathbb{E}\,\boldsymbol{X}_k)\right)\right\} \\
&\leq \inf_{\theta>0} \left\{\mathrm{e}^{-\theta(1+\delta)\mu_{\max}} \cdot \operatorname{tr}\exp\left((\mathrm{e}^{\theta}-1)\mu_{\max}\mathbf{I}_d\right)\right\} \\
&= d \cdot \inf_{\theta>0} \exp\left(-\theta(1+\delta)\mu_{\max} + (\mathrm{e}^{\theta}-1)\mu_{\max}\right).
\end{aligned}$$

The second inequality follows from the monotonicity (3.8) of the trace exponential. The infimal value is attained when $\theta = \log(1+\delta)$. Substitute this value into the right-hand side and simplify to complete the argument. □

Proof of Theorem 2.5, Lower Bound. The development of the lower bound is similar.

$$\begin{aligned}
\mathbb{P}\left\{\lambda_{\min}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \leq (1-\delta)\mu_{\min}\right\} &= \mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k (-\boldsymbol{X}_k)\right) \geq -(1-\delta)\mu_{\min}\right\} \\
&\leq \inf_{\theta>0} \left\{\mathrm{e}^{\theta(1-\delta)\mu_{\min}} \cdot \operatorname{tr}\exp\left((\mathrm{e}^{-\theta}-1)\sum\nolimits_k (\mathbb{E}\,\boldsymbol{X}_k)\right)\right\} \\
&\leq \inf_{\theta>0} \left\{\mathrm{e}^{\theta(1-\delta)\mu_{\min}} \cdot \operatorname{tr}\exp\left((\mathrm{e}^{-\theta}-1)\mu_{\min}\mathbf{I}_d\right)\right\} \\
&= d \cdot \inf_{\theta>0} \exp\left(\theta(1-\delta)\mu_{\min} + (\mathrm{e}^{-\theta}-1)\cdot\mu_{\min}\right).
\end{aligned}$$

The infimal value is achieved at $\theta = \log(1-\delta)$. Simplify the formula to finish up. □

6. Incorporating Variance Information

In this section, we establish a matrix version of Bennett's inequality. This result demonstrates that a sum of random matrices has normal concentration around its mean and Poisson-type decay in the tails.

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

$$\mathbb{E}\,\boldsymbol{X}_k = \mathbf{0} \quad \textit{and} \quad \|\boldsymbol{X}_k\| \leq B \quad \textit{almost surely.}$$

Define a bound on the total variance:

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

For all $t \geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \geq t\right\} \leq d \cdot \exp\left(-\frac{\sigma^2}{B^2} \cdot h\left(\frac{Bt}{\sigma^2}\right)\right)$$

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

The demonstration of Theorem 6.1 appears below. The matrix Bennett inequality also follows from a weaker set of bounds on the moments of the summands, which is clear from the proof. We obtain the matrix Bernstein inequality, Theorem 2.10, as a corollary because

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

The latter inequality follows by comparing Taylor series.

The first lemma shows how to bound the growth of moments of a random matrix using information about the variance and an almost sure bound for the maximum eigenvalue.

Lemma 6.2 (Growth of Moments)**.** *Suppose that* $\boldsymbol{X}$ *is a random s.a. matrix where* $\|\boldsymbol{X}\| \leq B$*. Then*

$$\mathbb{E}(\boldsymbol{X}^j) \preccurlyeq B^{j-2} \cdot \mathbb{E}\left(\boldsymbol{X}^2\right) \quad \textit{for } j = 2, 3, 4, \dots.$$

Proof. The hypothesis implies that $\boldsymbol{X}^j \preccurlyeq B^j \cdot \mathbf{I}$ for each positive integer j. Therefore,

$$\mathbb{E}(\boldsymbol{X}^j) = \mathbb{E}\left(\boldsymbol{X}[\boldsymbol{X}^{j-2}]\boldsymbol{X}\right) \preccurlyeq \mathbb{E}\left(\boldsymbol{X}[B^{j-2} \cdot \mathbf{I}]\boldsymbol{X}\right) = B^{j-2} \cdot \mathbb{E}\left(\boldsymbol{X}^2\right),$$

where the semidefinite relation follows from the conjugation rule (3.3). □

Under appropriate hypotheses on the growth of moments, we can develop a semidefinite bound for the matrix mgf. This argument proceeds by estimating each term in the Taylor series of the matrix exponential.

Lemma 6.3 (Bennett mgf)**.** *Suppose that* $\boldsymbol{X}$ *is a random s.a. matrix with* $\mathbb{E}\,\boldsymbol{X} = \mathbf{0}$*, and assume the moment growth bounds*

$$\mathbb{E}(\boldsymbol{X}^j) \preccurlyeq B^{j-2} \cdot \boldsymbol{V}^2 \quad \textit{for } j = 2, 3, 4, \dots.$$

Then

$$\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \exp\left(\frac{\mathrm{e}^{\theta B} - \theta B - 1}{B^2} \cdot \boldsymbol{V}^2\right) \quad \textit{for } \theta > 0.$$

Proof. The growth condition for the moments yields the bound

$$\begin{aligned}\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}} = \mathbf{I} + \mathbb{E}\,\boldsymbol{X} + \sum_{j=2}^{\infty} \frac{\theta^j\,\mathbb{E}(\boldsymbol{X}^j)}{j!} &\preccurlyeq \mathbf{I} + \frac{1}{B^2}\sum_{j=2}^{\infty}\frac{(\theta B)^j}{j!} \cdot \boldsymbol{V}^2 \\ &= \mathbf{I} + \frac{\mathrm{e}^{\theta B} - \theta B - 1}{B^2} \cdot \boldsymbol{V}^2 \preccurlyeq \exp\left(\frac{\mathrm{e}^{\theta B} - \theta B - 1}{B^2} \cdot \boldsymbol{V}^2\right).\end{aligned}$$

The last relation follows from (3.6). □

We are prepared to establish the matrix Bennett inequality.

Proof of Theorem 6.1. Invoke Lemma 6.2 and Lemma 6.3 to see that

$$\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq \exp\left(\frac{\mathrm{e}^{\theta B} - \theta B - 1}{B^2} \cdot \mathbb{E}\left(\boldsymbol{X}_k^2\right)\right).$$

Theorem 4.5 implies that

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\max}\left(\textstyle\sum_k \boldsymbol{X}_k\right) \geq t\right\} &\leq \inf_{\theta > 0}\left\{\mathrm{e}^{-\theta t} \cdot \operatorname{tr}\exp\left(\frac{\mathrm{e}^{\theta B} - \theta B - 1}{B^2} \cdot \textstyle\sum_k \mathbb{E}\left(\boldsymbol{X}_k^2\right)\right)\right\} \\ &\leq d \cdot \inf_{\theta > 0}\left\{\mathrm{e}^{-\theta t} \cdot \exp\left(\frac{\mathrm{e}^{\theta B} - \theta B - 1}{B^2} \cdot \sigma^2\right)\right\}.\end{aligned}$$

The second inequality uses the monotonicity property (3.8). The brace attains its minimal value when $\theta = B^{-1}\log(1 + Bt/\sigma^2)$. Substitute and simplify to establish the result. □

7. Rademacher and Gaussian Series

This section establishes normal concentration for Rademacher and Gaussian series; in the next section we use related considerations to derive the matrix Hoeffding inequality. The first step is to verify the bounds for the mgf of a fixed matrix scaled by a Rademacher variable or a Gaussian variable. This result essentially appears in Oliveira's work [Oli10b, Lem. 2].

Lemma 7.1 (Rademacher and Gaussian mgfs)**.** *Suppose that $\boldsymbol{A}$ is an s.a. matrix. Let ε be a Rademacher random variable, and let γ be a standard normal random variable. Then*

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

Proof. By symmetry of the Rademacher variable, we may take $\theta > 0$. By absorbing θ into $\boldsymbol{A}$, we may assume $\theta = 1$. We begin with the Rademacher case. By direct calculation,

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

where the second relation is (3.7).

Recall that the moments of a standard normal variable are

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

Therefore,

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

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

We immediately obtain the bound for Rademacher and Gaussian series.

Proof of Theorem 2.1. Let $\{\xi_k\}$ be a sequence of independent Rademacher variables or independent standard Gaussian variables. Invoke Lemma 7.1 to obtain

$$\mathbb{E}\,\mathrm{e}^{\theta\xi_k\boldsymbol{X}_k} \preccurlyeq \mathrm{e}^{\theta^2\boldsymbol{A}_k^2/2}.$$

Apply Theorem 4.5 to reach

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \geq t\right\} &\leq \inf_{\theta>0}\left\{\mathrm{e}^{-\theta t}\cdot \operatorname{tr}\exp\left(-\frac{\theta^2}{2}\sum\nolimits_k \boldsymbol{A}_k^2\right)\right\}\\ &\leq d\cdot\inf_{\theta>0}\left\{\mathrm{e}^{-\theta t}\cdot\exp\left(-\frac{\theta^2\sigma^2}{2}\right)\right\}.\end{aligned}$$

The infimum is attained at $\theta = \sigma$. □

8. Martingale Inequalities

In this section, we establish extensions of some classical martingale deviation bounds using the same approach that has served us so well. Indeed, a matrix-valued martingale admits an Azuma inequality, a bounded difference inequality, and even a Bernstein-type inequality.

We begin with some definitions. Consider a finite sequence $\{\boldsymbol{X}_k : k = 1, 2, \dots, n\}$ of random self-adjoint matrices, and write

$$\mathbb{E}_k\,\boldsymbol{W} = \mathbb{E}[\boldsymbol{W} \mid \boldsymbol{X}_1, \dots, \boldsymbol{X}_k]$$

for the expectation conditioned on the variables $\boldsymbol{X}_1, \dots, \boldsymbol{X}_k$. For consistency, $\mathbb{E}_0$ is the unconditional expectation. We say that the sequence $\{\boldsymbol{X}_k\}$ is a (self-adjoint) *matrix martingale* when

$$\mathbb{E}_{k-1}\,\boldsymbol{X}_k = \boldsymbol{X}_{k-1} \quad \text{for } k = 1, 2, \dots, n.$$

Define the martingale difference sequence

$$\boldsymbol{Y}_k = \boldsymbol{X}_k - \boldsymbol{X}_{k-1},$$

where we place the convention that $\boldsymbol{X}_0 = \mathbb{E}\,\boldsymbol{X}_1$. The difference sequence is conditionally zero mean:

$$\mathbb{E}_{k-1}\,\boldsymbol{Y}_k = \boldsymbol{0}.$$

Of course, we have transcribed these definitions directly from the scalar case.

Azuma's inequality states that the deviation of a scalar martingale is controlled by the sum of the maximum squared ranges of the difference sequence. In the matrix case, the same result holds with a matrix extension of the sum of maximum squared ranges.

Theorem 8.1 (Matrix Azuma)**.** *Let $\{\boldsymbol{X}_k\}$ be a self-adjoint matrix martingale in dimension d, and let $\{\boldsymbol{Y}_k\}$ be the associated difference sequence. Consider a sequence $\{\boldsymbol{A}_k\}$ of fixed self-adjoint matrices that satisfy*

$$\boldsymbol{Y}_k^2 \preccurlyeq \boldsymbol{A}_k^2 \quad \textit{almost surely.}$$

Define a bound on the sum of maximum squared ranges:

$$\sigma^2 \geq \left\| \sum\nolimits_k \boldsymbol{A}_k^2 \right\|.$$

Then for $t \geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{X}_n - \mathbb{E}\,\boldsymbol{X}_n) \geq t\right\} \leq d \cdot \mathrm{e}^{-t^2/8\sigma^2}.$$

The Hoeffding inequality, Theorem 2.8 is a special case of the Azuma inequality. Consider a finite sequence $\{\boldsymbol{Y}_j\}$ of independent, random, self-adjoint matrices with $\mathbb{E}\,\boldsymbol{Y}_j = \boldsymbol{0}$. Construct the sequence

$$\boldsymbol{X}_k = \sum\nolimits_{j=1}^{k} \boldsymbol{Y}_j.$$

The Hoeffding inequality is simply Azuma's inequality applied to the martingale $\{\boldsymbol{X}_k\}$.

In the scalar case, one of the most useful corollaries of Theorem 8.1 is McDiarmid's bounded differences inequality [McD98, Thm. 3.1]. This result states that a function of independent random variables exhibits normal concentration about its mean, where the variance depends on how much a change in a single variable can alter the value of the function. A version of McDiarmid's inequality holds in the matrix setting.

Corollary 8.2 (Bounded Differences)**.** *Let $\{W_k : k = 1, 2, \dots, n\}$ be an independent family of real random variables, and let $\boldsymbol{H}$ be a function that maps n real variables to a self-adjoint matrix of dimension d. Consider a sequence $\{\boldsymbol{A}_k\}$ of fixed self-adjoint matrices that satisfy*

$$\big(\boldsymbol{H}(w_1, \dots, w_k, \dots, w_n) - \boldsymbol{H}(w_1, \dots, w_k', \dots, w_n)\big)^2 \preccurlyeq \boldsymbol{A}_k^2,$$

where the numbers w_k and w_k' range over all possible values of W_k for each index k. Define a bound for the sum of maximum squared differences:

$$\sigma^2 \geq \left\| \sum\nolimits_k \boldsymbol{A}_k^2 \right\|.$$

Then for $t \geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{H}(\boldsymbol{w}) - \mathbb{E}\,\boldsymbol{H}(\boldsymbol{w})) \geq t\right\} \leq d \cdot \mathrm{e}^{-t^2/8\sigma^2}$$

where we abbreviate $\boldsymbol{w} = (W_1, \dots, W_n)$.

We can also develop a martingale analog of the matrix Bernstein inequality using an argument similar to—but easier than—the proof of the matrix Azuma inequality.

Theorem 8.3 (Matrix Bernstein: Martingale Version)**.** *Let $\{\boldsymbol{X}_k\}$ be a self-adjoint matrix martingale in dimension d, and let $\{\boldsymbol{Y}_k\}$ be the associated difference sequence. Consider a fixed sequence $\{\boldsymbol{A}_k\}$ of self-adjoint matrices that satisfy*

$$\|\boldsymbol{Y}_k\| \le B \quad \text{and} \quad \mathbb{E}_{k-1}(\boldsymbol{Y}_k^2) \preccurlyeq \boldsymbol{A}_k^2 \quad \text{almost surely.}$$

Define a bound on the sum of maximum conditional variances:

$$\sigma^2 \ge \left\| \sum\nolimits_k \boldsymbol{A}_k^2 \right\|.$$

Then

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{X}_n - \mathbb{E}\,\boldsymbol{X}_n) \ge t\right\} \le d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Bt/3}\right).$$

The proofs of the matrix Azuma and McDiarmid inequalities appear below. We omit the demonstration of the Bernstein inequality for matrix martingales. Theorems 8.1 and 8.3 are closely related to the noncommutative Burkholder–Davis–Gundy inequalities [PX97, JX03, JX08]. Unfortunately, the methods here produce somewhat weaker results than we might hope, as discussed in Section 10.1.

8.1. **Proofs.** The proof requires us to inject additional randomness into the sum of random matrices. The following lemma shows that we can affix an independent Rademacher random variable to a zero-mean random variable at a small cost.

Lemma 8.4 (Symmetrization)**.** *Let $\boldsymbol{H}$ be a fixed s.a. matrix, and let $\boldsymbol{Y}$ be a random s.a. matrix with $\mathbb{E}\,\boldsymbol{X} = 0$. Then*

$$\mathbb{E}\operatorname{tr} \mathrm{e}^{\boldsymbol{H}+\boldsymbol{Y}} \le \mathbb{E}\operatorname{tr} \mathrm{e}^{\boldsymbol{H}+2\varepsilon\boldsymbol{Y}},$$

where ε is a Rademacher variable independent from $\boldsymbol{Y}$.

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

$$\mathbb{E}\operatorname{tr} \mathrm{e}^{\boldsymbol{H}+\boldsymbol{Y}} = \mathbb{E}\operatorname{tr} \mathrm{e}^{\boldsymbol{H}+\boldsymbol{Y}-\mathbb{E}'\boldsymbol{Y}'} \le \mathbb{E}\operatorname{tr} \mathrm{e}^{\boldsymbol{H}+(\boldsymbol{Y}-\boldsymbol{Y}')} = \mathbb{E}\operatorname{tr} \mathrm{e}^{\boldsymbol{H}+\varepsilon(\boldsymbol{Y}-\boldsymbol{Y}')}.$$

We have used the convexity of the trace exponential to invoke Jensen's inequality (3.16). Since $\boldsymbol{Y} - \boldsymbol{Y}'$ is a symmetric random variable, we can affix an independent Rademacher variable to it without changing its distribution. The result of the argument depends on a short sequence of inequalities:

$$\begin{aligned}\mathbb{E}\operatorname{tr} \mathrm{e}^{\boldsymbol{H}+\boldsymbol{Y}} &\le \mathbb{E}\operatorname{tr}\left(\mathrm{e}^{\boldsymbol{H}/2+\varepsilon\boldsymbol{Y}} \cdot \mathrm{e}^{\boldsymbol{H}/2-\varepsilon\boldsymbol{Y}'}\right) \le \mathbb{E}\left[\left(\operatorname{tr}\mathrm{e}^{\boldsymbol{H}+2\varepsilon\boldsymbol{Y}}\right)^{1/2} \cdot \left(\operatorname{tr}\mathrm{e}^{\boldsymbol{H}-2\varepsilon\boldsymbol{Y}'}\right)^{1/2}\right] \\ &\le \left(\mathbb{E}\operatorname{tr}\mathrm{e}^{\boldsymbol{H}+2\varepsilon\boldsymbol{Y}}\right)^{1/2} \cdot \left(\mathbb{E}\operatorname{tr}\mathrm{e}^{\boldsymbol{H}-2\varepsilon\boldsymbol{Y}'}\right)^{1/2} = \mathbb{E}\operatorname{tr}\mathrm{e}^{\boldsymbol{H}+2\varepsilon\boldsymbol{Y}}.\end{aligned}$$

The first relation is the Golden–Thompson inequality (3.9); the second is the Cauchy–Schwarz inequality (3.1) for the trace; and the third is the Cauchy–Schwarz inequality (3.15) for real random variables. The last identity follows because the two factors are identically distributed. □

Proof of Theorem 8.1. We work with the sum of the difference sequence, which telescopes to the quantity of interest:

$$\sum\nolimits_{k=1}^{n} \boldsymbol{Y}_k = \boldsymbol{X}_n - \boldsymbol{X}_0$$

To control the mgf of the sum, we bound the cgf of each term, one after the next.

Let us detail the first step of the argument. Conditional on $\boldsymbol{X}_1, \dots, \boldsymbol{X}_{n-1}$, we may apply Lemma 8.4 with the random matrix $\boldsymbol{Y}_n$ to obtain

$$\mathbb{E}\operatorname{tr}\exp\left(\sum\nolimits_{k=1}^{n} \theta\boldsymbol{Y}_k\right) \le \mathbb{E}\operatorname{tr}\exp\left(\sum\nolimits_{k=1}^{n-1} \theta\boldsymbol{Y}_k + 2\varepsilon\theta\boldsymbol{Y}_n\right).$$

where ε is a Rademacher variable, independent from everything. At this point, we invoke the Rademacher mgf bound, Lemma 7.1, conditionally to obtain

$$\log \mathbb{E}\left[\mathrm{e}^{2\varepsilon\theta \boldsymbol{Y}_n} \;\middle|\; \boldsymbol{Y}_n\right] \preccurlyeq 2\theta^2 \boldsymbol{Y}_n^2 \preccurlyeq 2\theta^2 \boldsymbol{A}_n^2$$

Note that the first relation also depends the monotonicity (3.10) of the matrix logarithm. Now, apply Lieb's inequality, Theorem 3.1, conditionally to reach

$$\mathbb{E}\operatorname{tr}\exp\left(\sum\nolimits_{k=1}^{n} \theta \boldsymbol{Y}_k\right) \le \mathbb{E}\operatorname{tr}\exp\left(\sum\nolimits_{k=1}^{n-1} \theta \boldsymbol{Y}_k + \log \mathbb{E}\left[\mathrm{e}^{2\varepsilon\theta \boldsymbol{Y}_n} \;\middle|\; \boldsymbol{Y}_n\right]\right)$$
$$\le \mathbb{E}\operatorname{tr}\exp\left(\sum\nolimits_{k=1}^{n-1} \theta \boldsymbol{Y}_k + 2\theta^2 \boldsymbol{A}_n^2\right).$$

The last inequality depends on the fact (3.8) that the trace exponential is monotone.

For a given index m, the matrices $\boldsymbol{X}_1, \dots, \boldsymbol{X}_{m-1}$ do not depend on $\boldsymbol{A}_m, \dots, \boldsymbol{A}_n$, so we may iterate this argument to reach

$$\mathbb{E}\operatorname{tr}\exp\left(\sum\nolimits_{k=1}^{n} \theta \boldsymbol{Y}_k\right) \le \operatorname{tr}\exp\left(2\theta^2 \sum\nolimits_{k=1}^{n} \boldsymbol{A}_k^2\right) \le d \cdot \mathrm{e}^{2\theta^2\sigma^2}.$$

Introducing this bound into the Laplace transform estimate, Proposition 4.4, we reach

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{X}_n - \boldsymbol{X}_0) \ge t\right\} \le \inf_{\theta>0}\left\{\mathrm{e}^{-\theta t} \cdot \mathrm{e}^{2\theta^2\sigma^2}\right\}.$$

The infimum is achieved when $\theta = t/4\sigma^2$. Finally, observe that $\boldsymbol{X}_0 = \mathbb{E}\,\boldsymbol{X}_n$ by the martingale property. □

Finally, we show how a matrix version of McDiarmid's inequality follows from the matrix Azuma inequality.

Proof of Corollary 8.2. For this argument only, we write $\mathbb{E}_W$ for the expectation with respect to a random variable W, holding other variables fixed. Recall that $\boldsymbol{w} = (W_1, \dots, W_n)$. For $k = 0, 1, \dots, n$, consider the random matrices

$$\boldsymbol{X}_k = \mathbb{E}[\boldsymbol{H}(\boldsymbol{w}) \mid W_1, W_2, \dots, W_k] = \mathbb{E}_{W_{k+1}} \mathbb{E}_{W_{k+2}} \dots \mathbb{E}_{W_n} \boldsymbol{H}(\boldsymbol{w}).$$

The sequence $\{\boldsymbol{X}_k\}$ forms a Doob martingale. The associated difference sequence is

$$\boldsymbol{Y}_k = \boldsymbol{X}_k - \boldsymbol{X}_{k-1} = \mathbb{E}_{W_{k+1}} \mathbb{E}_{W_{k+2}} \dots \mathbb{E}_{W_n} \left(\boldsymbol{H}(\boldsymbol{w}) - \mathbb{E}_{W_k} \boldsymbol{H}(\boldsymbol{w})\right),$$

where the second identity follows from independence and Fubini's theorem.

It remains to bound the difference sequence. Let W_k' be an independent copy of W_k, and construct the random vector $\boldsymbol{w}' = (W_1, \dots, W_{k-1}, W_k', W_{k+1}, \dots, W_n)$. Since $\mathbb{E}_{W_k} \boldsymbol{H}(\boldsymbol{w}) = \mathbb{E}_{W_k'} \boldsymbol{H}(\boldsymbol{w}')$ and $\boldsymbol{w}$ does not depend on W_k', we can write

$$\boldsymbol{Y}_k = \mathbb{E}_{W_{k+1}} \mathbb{E}_{W_{k+2}} \dots \mathbb{E}_{W_n} \mathbb{E}_{W_k'} \left(\boldsymbol{H}(\boldsymbol{w}) - \boldsymbol{H}(\boldsymbol{w}')\right).$$

The vectors $\boldsymbol{w}$ and $\boldsymbol{w}'$ differ only in the kth coordinate, so that

$$\left(\boldsymbol{H}(\boldsymbol{w}) - \boldsymbol{H}(\boldsymbol{w}')\right)^2 \preccurlyeq \boldsymbol{A}_k^2$$

by definition of the bound $\boldsymbol{A}_k^2$. Finally, the matrix convexity of the square function allows us to invoke the matrix Jensen inequality (3.17) to reach

$$\boldsymbol{Y}_k^2 \preccurlyeq \mathbb{E}_{W_{k+1}} \mathbb{E}_{W_{k+2}} \dots \mathbb{E}_{W_n} \mathbb{E}_{W_k'} \left[\left(\boldsymbol{H}(\boldsymbol{w}) - \boldsymbol{H}(\boldsymbol{w}')\right)^2\right] \preccurlyeq \boldsymbol{A}_k^2.$$

To complete the proof, we apply the matrix Azuma inequality, Theorem 8.1. □

9. Complements

The methods in this paper are not limited to the problem of studying the largest eigenvalue of a random self-adjoint matrix. Indeed, we can bound fluctuations in the norm of a rectangular matrix (Section 9.1), and we can consider other semidefinite relations (Section 9.2).

9.1. **Rectangular Matrices.** With the exception of the matrix Chernoff bound, Theorem 2.5, the probability inequalities in this paper can be adapted to provide results for sums of random rectangular matrices. The method is straightforward: we simply apply each inequality to the self-adjoint dilation (3.11) of the sum. The statements for rectangular matrices parallel the results for self-adjoint matrices, but they involve both the row and column moduli (3.5) of the matrices instead of the usual modulus.

As an example, we establish the rectangular version of the matrix Bernstein inequality, Theorem 2.10. Other results follow from the same considerations.

Corollary 9.1 (Matrix Bernstein: Rectangular Version)**.** *Consider a finite sequence* $\{\boldsymbol{Z}_k\}$ *of independent, random, rectangular matrices with dimensions* $d_1 \times d_2$*. Assume that*

$$\mathbb{E}\,\boldsymbol{Z}_k = \mathbf{0} \quad \text{and} \quad \|\boldsymbol{Z}_k\| \le B \quad \text{almost surely.}$$

Define a bound on the total variance:

$$\sigma^2 \ge \max\left\{ \left\| \sum\nolimits_k \mathbb{E}\big(|\boldsymbol{Z}_k|_{\mathrm{col}}^2\big) \right\|, \left\| \sum\nolimits_k \mathbb{E}\big(|\boldsymbol{Z}_k|_{\mathrm{row}}^2\big) \right\| \right\}.$$

For all $t \ge 0$,

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

Proof. Consider the finite sequence $\{\mathscr{S}(\boldsymbol{Z}_k)\}$ of random s.a. matrices with dimension $d_1 + d_2$. Since expectation commutes with the s.a. dilation, each of these matrices is zero mean.

$$\mathbb{E}\,\mathscr{S}(\boldsymbol{Z}_k) = \mathscr{S}(\mathbb{E}\,\boldsymbol{Z}_k) = \mathscr{S}(\mathbf{0}) = \mathbf{0}.$$

The identity (3.13) yields a uniform bound on the norm of the s.a. matrices.

$$\|\mathscr{S}(\boldsymbol{Z}_k)\| = \|\boldsymbol{Z}_k\| \le B \quad \text{almost surely.}$$

To compute the variance of the sum of the s.a. matrices, we recall the connection (3.12) between the modulus of the s.a. dilation and the row/column moduli:

$$\begin{aligned} \left\| \sum\nolimits_k \mathbb{E}\big(|\mathscr{S}(\boldsymbol{Z}_k)|^2\big) \right\| &= \left\| \sum\nolimits_k \mathbb{E} \begin{bmatrix} |\boldsymbol{Z}_k|_{\mathrm{col}}^2 & \\ & |\boldsymbol{Z}_k|_{\mathrm{row}}^2 \end{bmatrix} \right\| \\ &= \left\| \begin{bmatrix} \sum_k \mathbb{E}\big(|\boldsymbol{Z}_k|_{\mathrm{col}}^2\big) & \\ & \sum_k \mathbb{E}\big(|\boldsymbol{Z}_k|_{\mathrm{row}}^2\big) \end{bmatrix} \right\| \\ &= \max\left\{ \left\| \sum\nolimits_k \mathbb{E}\big(|\boldsymbol{Z}_k|_{\mathrm{col}}^2\big) \right\|, \left\| \sum\nolimits_k \mathbb{E}\big(|\boldsymbol{Z}_k|_{\mathrm{row}}^2\big) \right\| \right\}. \end{aligned}$$

Finally, use the linearity of the s.a. dilation and the identity (3.13) to see that

$$\lambda_{\max}\left(\sum\nolimits_k \mathscr{S}(\boldsymbol{Z}_k) \right) = \lambda_{\max}\left(\mathscr{S}\left(\sum\nolimits_k \boldsymbol{Z}_k \right) \right) = \left\| \sum\nolimits_k \boldsymbol{Z}_k \right\|.$$

Invoke Theorem 2.10 to complete the argument. □

9.2. **Semidefinite Relations.** The paper [AW02] of Ahlswede and Winter actually develops a somewhat more general version of the Laplace transform method that allows them to consider semidefinite relations beyond simple eigenvalue bounds. In this section, we describe their approach.

Proposition 9.2 (Matrix Laplace Transform Method)**.** *Let* $\boldsymbol{X}$ *be a random self-adjoint matrix, and let* $\boldsymbol{T}$ *be a fixed self-adjoint matrix. Then*

$$\mathbb{P}\{\boldsymbol{X} \not\preccurlyeq \boldsymbol{T}\} \le \inf_{\boldsymbol{\Theta} \succ \mathbf{0}} \mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\boldsymbol{\Theta}(\boldsymbol{X}-\boldsymbol{T})\boldsymbol{\Theta}}$$

where the infimum ranges over all positive definite matrices.

Proof. Let $\mathbf{\Theta}$ be a pd matrix. Then

$$\mathbb{P}\{\boldsymbol{X} \not\preccurlyeq \boldsymbol{T}\} = \mathbb{P}\{\boldsymbol{X} - \boldsymbol{T} \not\preccurlyeq \mathbf{0}\} = \mathbb{P}\{\mathbf{\Theta}^*(\boldsymbol{X} - \boldsymbol{T})\mathbf{\Theta} \not\preccurlyeq \mathbf{0}\} = \mathbb{P}\{\lambda_{\max}(\mathbf{\Theta}^*(\boldsymbol{X} - \boldsymbol{T})\mathbf{\Theta}) \geq 0\}.$$

The second equality requires Sylvester's inertia theorem (Section 3.1.5). We continue along the same lines as Proposition 4.4 to reach

$$\mathbb{P}\{\boldsymbol{X} \not\preccurlyeq \boldsymbol{T}\} = \mathbb{P}\left\{\mathrm{e}^{\lambda_{\max}(\mathbf{\Theta}^*(\boldsymbol{X}-\boldsymbol{T})\mathbf{\Theta})} \geq 1\right\} \leq \mathbb{E}\,\mathrm{e}^{\lambda_{\max}(\mathbf{\Theta}^*(\boldsymbol{X}-\boldsymbol{T})\mathbf{\Theta})}$$
$$= \mathbb{E}\,\lambda_{\max}\left(\mathrm{e}^{\mathbf{\Theta}^*(\boldsymbol{X}-\boldsymbol{T})\mathbf{\Theta}}\right) \leq \mathbb{E}\operatorname{tr}\mathrm{e}^{\mathbf{\Theta}^*(\boldsymbol{X}-\boldsymbol{T})\mathbf{\Theta}}.$$

The second relation is Markov's inequality (3.14), the third is the spectral mapping theorem, and the final one is (3.2). We complete the proof by taking the infimum over all pd $\mathbf{\Theta}$. □

Proposition 9.2 is genuinely stronger than Proposition 4.4 because it allows us to optimize over a much larger set. On account of this result, we see that the most natural matrix extensions of the mgf and cgf are

$$\boldsymbol{M}_{\boldsymbol{X}}(\mathbf{\Theta}) = \mathbb{E}\,\mathrm{e}^{\mathbf{\Theta}\boldsymbol{X}\mathbf{\Theta}} \quad \text{and} \quad \boldsymbol{C}_{\boldsymbol{X}}(\mathbf{\Theta}) = \log \mathbb{E}\,\mathrm{e}^{\mathbf{\Theta}\boldsymbol{X}\mathbf{\Theta}}.$$

For a given order d, these functions map the class of d-dimensional pd matrices to the class of d-dimensional s.a. matrices. With this insight, we frame the following extension of Theorem 4.5.

Theorem 9.3. *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, self-adjoint matrices and a sequence $\{\boldsymbol{A}_k(\cdot)\}$ of fixed functions from a set Θ of positive-definite matrices into the self-adjoint matrices. Suppose that*

$$\log\left(\mathbb{E}\,\mathrm{e}^{\mathbf{\Theta}\boldsymbol{X}_k\mathbf{\Theta}}\right) \preccurlyeq \boldsymbol{A}_k(\mathbf{\Theta}) \quad \text{for } \mathbf{\Theta} \in \Theta.$$

In particular, it suffices to assume that

$$\mathbb{E}\,\mathrm{e}^{\mathbf{\Theta}\boldsymbol{X}_k\mathbf{\Theta}} \preccurlyeq \mathrm{e}^{\boldsymbol{A}_k(\mathbf{\Theta})} \quad \text{for } \mathbf{\Theta} \in \Theta.$$

Then

$$\mathbb{P}\left\{\sum\nolimits_k \boldsymbol{X}_k \not\preccurlyeq \boldsymbol{T}\right\} \leq \inf_{\mathbf{\Theta}\in\Theta} \operatorname{tr}\exp\left(-\mathbf{\Theta}\boldsymbol{T}\mathbf{\Theta} + \sum\nolimits_k \boldsymbol{A}_k(\mathbf{\Theta})\right).$$

Observe that Theorem 4.5 follows as a special case when we take $\Theta = \{\theta\mathbf{I} : \theta > 0\}$, the set of positive scalar matrices. The proof of Theorem 9.3 combines the matrix Laplace transform method, Proposition 9.2, with the result on "subadditivity" of cumulants, Theorem 4.2. We omit the details of the argument.

10. Open Questions

We close with a short discussion of a theoretical question and an applied question that have resisted our efforts.

10.1. **Full-Strength Martingale Inequalities.** The matrix martingale inequalities we developed in Section 8 fall somewhat short of the best results that are possible in this direction. Let $\{\boldsymbol{X}_k\}$ be a self-adjoint matrix martingale, and let $\{\boldsymbol{Y}_k\}$ be the associated martingale difference sequence. The noncommutative Burkholder–Davis–Gundy inequalities [PX97, JX03, JX08] demonstrate that the correct extension of the variance to the martingale setting is

$$\sigma^2 = \sup\left\|\sum\nolimits_{k=1}^n \mathbb{E}_{k-1}\left(\boldsymbol{Y}_k^2\right)\right\|.$$

In contrast, the variance in our martingale version of the Bernstein inequality, Theorem 8.3, can be written heuristically as

$$\sigma^2 = \left\|\sum\nolimits_{k=1}^n \sup \mathbb{E}_{k-1}\left(\boldsymbol{Y}_k^2\right)\right\|.$$

The latter quantity is potentially much larger.

At present, we have been unable to reach the sharper results using an approach in the same spirit as the rest of our arguments. One attractive possibility would be to extend the exponential martingale decoupling inequality [dlPG99, Cor. 6.2.5] to the matrix setting. It is this route that de la Peña and Giné take to establish the scalar Burkholder–Davis–Gundy inequalities [dlPG99, Sec. 6.5]. So far, we have not been able to surmount the obstacles that arise along this path.

After the first draft of this paper was completed, it came to our attention that Oliveira has established a matrix version of the Freedman inequality [Oli10a, Thm. 1.2], which addresses this "open" problem. Oliveira's result is sharp up to constants. We are in the process of preparing a substantial revision to this paper that incorporates Lieb's inequality into a martingale argument to obtain Freedman's inequality with sharp constants.

10.2. **Random Matrices with Independent Entries.** Finally, we mention the problem of finding a high-probability bound for the norm of a random matrix with independent entries. For some discussion of this question, see [RV10]. The simplest problem of this sort is to compute the expected norm of a $d \times d$ random sign matrix $\boldsymbol{X}$, whose entries are independent Rademacher variables. It is known that

$$\mathbb{E}\,\|\boldsymbol{X}\| \sim \sqrt{d}.$$

Straightforward applications of the ideas in this paper, however, yield the weaker estimate

$$\mathbb{E}\,\|\boldsymbol{X}\| \sim \sqrt{d \log d}.$$

It would be very interesting to identify a method for establishing an asymptotically sharp estimate using our techniques. One possible approach would be to amplify the simple bounds using tensor products. Carl and Defant use precisely this idea to remove the logarithmic factor from a deterministic eigenvalue estimate [CD00].

Acknowledgments

We would like to thank Vern Paulsen and Bernhard Bodmann for some helpful conversations related to this project. We also with to thank Roberto Oliveira and Klas Markström for providing some additional references.

References

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

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

[AW02] R. Ahslwede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Info. Theory*, 48(3):569–579, Mar. 2002.

[AW03] R. Ahslwede and A. Winter. Addendum to "Strong converse for identification via quantum channels". *IEEE Trans. Info. Theory*, 49(1):346, Jan. 2003.

[Bha97] R. Bhatia. *Matrix Analysis*. Number 169 in GTM. Springer, Berlin, 1997.

[Bha07] R. Bhatia. *Positive Definite Matrices*. Princeton Univ. Press, 2007.

[Buc01] A. Buchholz. Operator Khintchine inequality in non-commutative probability. *Math. Ann.*, 319:1–16, 2001.

[Buc05] A. Buchholz. Optimal constants in Khintchine-type inequalities for Fermions, Rademachers and q-Gaussian operators. *Bull. Pol. Acad. Sci. Math.*, 53(3):315–321, 2005.

[CD00] B. Carl and A. Defant. An elementary approach to an eigenvalue estimate for matrices. *Positivity*, 4:131–141, 2000.

[CLMW09] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? Available at `arXiv:0912.3599`, Dec. 2009.

[CM08] D. Cristofides and K. Markström. Expansion properties of random Cayley graphs and vertex transitive graphs via matrix martingales. *Random Structures Algs.*, 32(8):88–100, 2008.

[CR07] E. Candès and J. K. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007.

[CSPW09] V. Chandrasekaran, S. Sanghavi, P. A. Parillo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. Available at `arXiv:0906.2220`, Jun. 2009.

[dlPG99] V. H. de la Peña and E. Giné. *Decoupling: From Dependence to Independence*. Probability and its Applications. Springer, Berlin, 1999.

[Eps73] H. Epstein. Remarks on two theorems of E. Lieb. *Comm. Math. Phys.*, 31:317–325, 1973.

[Gro09] D. Gross. Recovering low-rank matrices from few coefficients in any basis. Available at `arXiv:0910.1879`, Oct. 2009.

[Hig08] N. J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, Philadelphia, PA, 2008.

[HJ85] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge Univ. Press, Cambridge, 1985.

[HJ94] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge Univ. Press, Cambridge, 1994.

[HMT09] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. ACM Report 2009-05, California Inst. Tech., Pasadena, CA, Sep. 2009.

[JX03] M. Junge and Q. Xu. Noncommutative Burkholder/Rosenthal inequalities. *Ann. Probab.*, 31(2):948–995, 2003.

[JX05] M. Junge and Q. Xu. On the best constants in some non-commutative martingale inequalities. *Bull. London Math. Soc.*, 37:243–253, 2005.

[JX08] M. Junge and Q. Xu. Noncommutative Burkholder/Rosenthal inequalities II: Applications. *Israel J. Math.*, 167:227–282, 2008.

[Led96] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Stat.*, 1:63–87, Jul. 1996.

[Lie73] E. H. Lieb. Convex trace functions and the Wigner–Yanase–Dyson conjecture. *Adv. Math.*, 11:267–288, 1973.

[LP86] F. Lust-Picquard. Inégalités de Khintchine dans C_p $(1 < p < \infty)$. *C. R. Math. Acad. Sci. Paris*, 303(7):289–292, 1986.

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin, 1991.

[Lug09] G. Lugosi. Concentration-of-measure inequalities. Available at `http://www.econ.upf.edu/~lugosi/anu.pdf`, 2009.

[McD98] C. McDiarmid. Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*, number 16 in Algorithms and Combinatorics, pages 195–248. Springer, 1998.

[Nem07] A. Nemirovski. Sums of random symmetric matrices and quadratic optimization under orthogonality constraints. *Math. Prog. Ser. B*, 109:283–317, 2007.

[Oli10a] R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Available at `arXiv:0911.0600`, Feb. 2010.

[Oli10b] R. I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. Available at `arXiv:1004.3821`, 2010.

[Pau86] V. I. Paulsen. *Completely Bounded Maps and Dilations*. Number 146 in Pitman Research Notes in Mathematics. Longman Scientific & Technical, New York, NY, 1986.

[Pet94] D. Petz. A survey of certain trace inequalities. In *Functional analysis and operator theory*, volume 30 of *Banach Center Publications*, pages 287–298, Warsaw, 1994. Polish Acad. Sci.

[PX97] G. Pisier and Q. Xu. Non-commutative martingale inequalities. *Comm. Math. Phys.*, 189(3):667–698, 1997.

[Rec09] B. Recht. Simpler approach to matrix completion. Available at `http://pages.cs.wisc.edu/~brecht/papers/09.Recht.ImprovedMC.pdf`, Oct. 2009.

[Rud99] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164:60–72, 1999.

[Rus02] M. B. Ruskai. Inequalities for quantum entropy: A review with conditions for equality. *J. Math. Phys.*, 43(9):4358–4375, Sep. 2002.

[Rus05] M. B. Ruskai. Erratum: Inequalities for quantum entropy: A review with conditions for equality [*J. Math. Phys.* 43, 4358 (2002)]. *J. Math. Phys.*, 46(1):0199101, 2005.

[RV07] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4):Article 21, 19 pp., Jul. 2007. (electronic).

[RV10] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: Extreme singular values. In *Proc. Intl. Cong. Mathematicians 2010*, Hyderabad, 2010. To appear. Available at `arXiv:1003.2990`.

[So09] A. M.-C. So. Moment inequalities for sums of random matrices and their applications in optimization. *Math. Prog.*, 2009. To appear. Available at `http://www.se.cuhk.edu.hk/~manchoso/papers/momineq-MPA.pdf`.

[SST06] A. Sankar, D. A. Spielman, and S.-H. Teng. Smoothed analysis of the condition numbers and growth factors of matrices. *SIAM J. Matrix Anal. Appl.*, 28(2):446–476, 2006.

[Tal88] M. Talagrand. An isoperimetric theorem on the cube and the Kintchine–Kahane inequalities. *Proc. AMS*, 104(3):905–909, Nov 1988.

[Tro08] J. A. Tropp. On the conditioning of random subdictionaries. *Appl. Comp. Harmon. Anal.*, 25:1–24, 2008.

[Ver09] R. Vershynin. A note on sums of independent random matrices after Ahlswede–Winter. Available at `http://www-personal.umich.edu/~romanv/teaching/reading-group/ahlswede-winter.pdf`, 2009.