USER-FRIENDLY TAIL BOUNDS FOR MATRIX MARTINGALES

JOEL A. TROPP

Technical Report No. 2011-01
January 2011

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

USER-FRIENDLY TAIL BOUNDS FOR MATRIX MARTINGALES

JOEL A. TROPP

Abstract. This report presents probability inequalities for sums of adapted sequences of 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. The methods also specialize to sums of independent random matrices.

1. Main Results

This technical report is a companion to two other works, the papers "User-friendly tail bounds for sums of random matrices" [Tro10c] and "Freedman's inequality for matrix martingales" [Tro10a]. Since this report is intended as a supplement, we have removed most of the background discussion, citations to related work, and auxiliary commentary that places the research in a wider context. We recommend that the reader peruse the original papers before studying this report.

The paper [Tro10a] describes a martingale technique that leads to an extension of Freedman's inequality in the matrix setting, which is similar to the result [Oli10a, Thm. 1.2]. The purpose of this work is to show how the arguments from [Tro10a] allow us to establish the matrix probability inequalities for sums of *independent* random matrices that appear in [Tro10c]. The discussion here also contains some new probability inequalities for sums of adapted sequences of random matrices; we have removed these results from the other two papers because they are somewhat specialized.

1.1. Roadmap.

The rest of the report is organized as follows. The balance of §1 provides an overview of the main results for sums of independent random matrices. Section 2 contains the main technical ingredients for the proof. Sections 3–5 complete the proofs of the matrix probability inequalities for adapted sequences. Appendix A provides an overview of the background material that we require.

1.2. Rademacher and Gaussian Series.

Let $\|\cdot\|$ denote the usual norm for operators on a Hilbert space, which returns the largest singular value of its argument, and let $\lambda_{\max}$ denote the algebraically largest eigenvalue of a self-adjoint matrix. The extreme eigenvalues of a Rademacher series with self-adjoint matrix coefficients exhibit normal concentration.

Theorem 1.1 (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 finite sequence of independent Rademacher variables. Compute the norm of the sum of squared coefficient matrices:*

$$\sigma^2 := \left\| \sum\nolimits_k \boldsymbol{A}_k^2 \right\|. \tag{1.1}$$

Date: 25 April 2010. Revised on 15 June 2010, 10 August 2010, 14 November 2010, and 16 January 2011.

Key words and phrases. Discrete-time martingale, large deviation, probability inequality, 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.

For all $t \geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \varepsilon_k \boldsymbol{A}_k\right) \geq t\right\} \leq d \cdot \mathrm{e}^{-t^2/2\sigma^2}. \tag{1.2}$$

In particular,

$$\mathbb{P}\left\{\left\|\sum\nolimits_k \varepsilon_k \boldsymbol{A}_k\right\| \geq t\right\} \leq 2d \cdot \mathrm{e}^{-t^2/2\sigma^2}. \tag{1.3}$$

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

See [Tro10c, §4] for a detailed discussion of Theorem 1.1, which indicates that it is essentially sharp. We present the proof in §5.

1.3. **Sums of Random Semidefinite Matrices.** Chernoff bounds 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 1.2 (Matrix Chernoff)**.** *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, positive-semidefinite matrices with dimension d. Suppose that*

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

Compute the eigenvalues of the sum of the expectations:

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

Then

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

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

We establish Theorem 1.2 in §3, where it emerges as a consequence of Theorem 3.1, a Chernoff inequality for sums of adapted sequences of positive-semidefinite matrices.

1.4. **Adding Variance Information.** In the scalar case, a well-known 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 largest eigenvalue of each summand.

Theorem 1.3 (Matrix Bernstein)**.** *Consider a finite sequence $\{\boldsymbol{X}_k\}$ of independent, random, self-adjoint matrices with dimension d. Suppose that*

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

Compute the norm of the total variance:

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

For all $t \geq 0$,

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

The matrix Bernstein inequality, Theorem 1.3, follows from a more detailed result, which provides stronger Poisson-type decay for the tail. In §4, we derive these results from a martingale result.

1.5. Miscellaneous Results.

The methods in this paper deliver a number of other results:

- All of the results described in the front matter follow from more general bounds for large deviations of matrix martingales. See §3–5 for the full story.
- All the inequalities we have mentioned, with exception of the matrix Chernoff bounds, have variants that hold for rectangular matrices. The extensions follow immediately from the self-adjoint case by applying an elegant device from operator theory, called the self-adjoint dilation of a matrix [Pau86]. See [Tro10c, §4.2] for additional details.

2. Tail Bounds via Martingale Methods

This section contains the main part of the argument, which parallels Freedman's argument for producing large deviation bounds for scalar martingales [Fre75]. The material here duplicates the note [Tro10a].

2.1. Matrix Moments and Cumulants.

Consider a random s.a. matrix $\boldsymbol{X}$ that has moments of all orders. By analogy with the classical definitions for scalar random variables, we construct the matrix *moment generating function* (mgf) and *cumulant generating function* (cgf).

$$\boldsymbol{M}_{\boldsymbol{X}}(\theta) := \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}} \quad \text{and} \quad \boldsymbol{\Xi}_{\boldsymbol{X}}(\theta) := \log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}} \quad \text{for } \theta \in \mathbb{R}. \tag{2.1}$$

The mgf has a formal power series expansion that displays the raw moments of the random matrix:

$$\boldsymbol{M}_{\boldsymbol{X}}(\theta) = \mathbf{I} + \sum\nolimits_{j=1}^{\infty} \frac{\theta^j}{j!} \cdot \mathbb{E}(\boldsymbol{X}^j).$$

In the scalar setting, the cgf can be interpreted as an *exponential mean*, a weighted average of a random variable that emphasizes large (positive) deviations. The matrix cgf admits a similar intuition, and we treat it as a measure of the variability of a random matrix.

2.2. The Large Deviation Supermartingale.

In this section, we extend Freedman's martingale techniques [Fre75] to the matrix setting. The matrix cgf and Lieb's result, Theorem A.1, play a central role in this development.

We begin with a filtration $\{\mathscr{F}_k : k = 0, 1, 2, \dots\}$ of a master probability space, and we write $\mathbb{E}_k$ for the conditional expectation with respect to $\mathscr{F}_k$. Consider an adapted random process $\{\boldsymbol{X}_k : k = 1, 2, 3, \dots\}$ and a previsible random process $\{\boldsymbol{V}_k : k = 1, 2, 3, \dots\}$ whose values are s.a. matrices with dimension d. Suppose that the two processes are related through a conditional cgf bound of the form

$$\log \mathbb{E}_{k-1} \mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq g(\theta) \cdot \boldsymbol{V}_k \quad \text{almost surely for } \theta > 0. \tag{2.2}$$

The function $g : (0, \infty) \to [0, \infty]$, and—for simplicity—we do not allow this function to depend on the index k. It is convenient to define the partial sums of the original process and the partial sums of the conditional cgf bounds:

$$\boldsymbol{Y}_0 := \mathbf{0} \quad \text{and} \quad \boldsymbol{Y}_k := \sum\nolimits_{j=1}^{k} \boldsymbol{X}_j.$$

$$\boldsymbol{W}_0 := \mathbf{0} \quad \text{and} \quad \boldsymbol{W}_k := \sum\nolimits_{j=1}^{k} \boldsymbol{V}_j.$$

In almost all our examples, $\{\boldsymbol{V}_k\}$ is a sequence of psd matrices, and so $\{\boldsymbol{W}_k\}$ increases with respect to the semidefinite order. The random matrix $\boldsymbol{W}_k$ can be viewed as a measure of the total variability of the process $\{\boldsymbol{X}_k\}$ up to time k.

To continue, we fix the function g and a positive number θ. Define a real-valued function with two s.a. matrix arguments:

$$G_\theta(\boldsymbol{Y}, \boldsymbol{W}) := \operatorname{tr} \exp\big(\theta \boldsymbol{Y} - g(\theta) \cdot \boldsymbol{W}\big).$$

We use the function G_θ to construct a real-valued random process.

$$S_k := S_k(\theta) = G_\theta(\boldsymbol{Y}_k, \boldsymbol{W}_k) \quad \text{for } k = 0, 1, 2, \dots. \tag{2.3}$$

This process is an evolving measure of the discrepancy between the partial sum process $\{\boldsymbol{Y}_k\}$ and the cumulant sum process $\{\boldsymbol{W}_k\}$. The following lemma describes the key properties of this random sequence. In particular, the average discrepancy decreases with time. The proof relies on Lieb's result, Theorem A.1.

Lemma 2.1. *For each fixed $\theta > 0$, the random process $\{S_k(\theta) : k = 0, 1, 2, \dots\}$ defined in (2.3) is a positive supermartingale whose initial value $S_0 = d$.*

Proof. It is easily seen that S_k is positive because the exponential of a self-adjoint matrix is pd, and the trace of a pd matrix is positive. We obtain the initial value from a short calculation:

$$S_0 = \operatorname{tr}\exp\left(\theta\boldsymbol{Y}_0 - \boldsymbol{W}_0(\theta)\right) = \operatorname{tr}\exp(\boldsymbol{0}_d) = \operatorname{tr}\mathbf{I}_d = d.$$

To prove that the process is a supermartingale, we ascend a short chain of inequalities.

$$\begin{aligned}
\mathbb{E}_{k-1} S_k &= \mathbb{E}_{k-1} \operatorname{tr}\exp\left(\theta\boldsymbol{Y}_{k-1} - g(\theta)\cdot\boldsymbol{W}_k + \log \mathrm{e}^{\theta\boldsymbol{X}_k}\right) \\
&\le \operatorname{tr}\exp\left(\theta\boldsymbol{Y}_{k-1} - g(\theta)\cdot\boldsymbol{W}_k + \log \mathbb{E}_{k-1}\mathrm{e}^{\theta\boldsymbol{X}_k}\right) \\
&\le \operatorname{tr}\exp\left(\theta\boldsymbol{Y}_{k-1} - g(\theta)\cdot\boldsymbol{W}_k + g(\theta)\cdot\boldsymbol{V}_k\right) \\
&= \operatorname{tr}\exp\left(\theta\boldsymbol{Y}_{k-1} - g(\theta)\cdot\boldsymbol{W}_{k-1}\right) \\
&= S_{k-1}.
\end{aligned}$$

In the first step, we remove the term $\boldsymbol{X}_k$ from the partial sum $\boldsymbol{Y}_k$ and rewrite it using the definition (A.7) of the matrix logarithm. Next, we invoke Lieb's Theorem, conditional on $\mathscr{F}_{k-1}$, to verify the concavity of the function

$$\boldsymbol{A} \longmapsto \operatorname{tr}\exp\left(\theta\boldsymbol{Y}_{k-1} - g(\theta)\cdot\boldsymbol{W}_k + \log(\boldsymbol{A})\right).$$

We apply Jensen's inequality (A.9) to draw the conditional expectation inside the function. This act is legal because $\boldsymbol{Y}_{k-1}$ and $\boldsymbol{W}_k$ are both measurable with respect to $\mathscr{F}_{k-1}$. The second inequality depends on the assumption (2.2) together with the fact (A.6) that the trace of the matrix exponential is monotone. The final step recalls that $\{\boldsymbol{W}_k\}$ is the sequence of partial sums of $\{\boldsymbol{V}_k\}$. □

Finally, we present a simple inequality for the function G_θ that holds when we have control on the eigenvalues of its arguments.

Lemma 2.2. *Suppose that $\lambda_{\max}(\boldsymbol{Y}) \ge y$ and that $\lambda_{\max}(\boldsymbol{W}) \le w$. For each $\theta > 0$,*

$$G_\theta(\boldsymbol{Y}, \boldsymbol{W}) \ge \mathrm{e}^{\theta y - g(\theta) w}.$$

Proof. Recall that $g(\theta) \ge 0$. The bound results from a straightforward calculation:

$$G_\theta(\boldsymbol{Y}, \boldsymbol{W}) = \operatorname{tr}\mathrm{e}^{\theta\boldsymbol{Y} - g(\theta)\cdot\boldsymbol{W}} \ge \operatorname{tr}\mathrm{e}^{\theta\boldsymbol{Y} - g(\theta)w\mathbf{I}} \ge \lambda_{\max}\left(\mathrm{e}^{\theta\boldsymbol{Y} - g(\theta)w\mathbf{I}}\right) = \mathrm{e}^{\theta\lambda_{\max}(\boldsymbol{Y}) - g(\theta)w} \ge \mathrm{e}^{\theta y - g(\theta)w}.$$

The first inequality depends on the fact that $\boldsymbol{W} \preccurlyeq w\mathbf{I}$ and the monotonicity (A.6) of the trace exponential. The second inequality relies on the property (A.1) that the trace of a psd matrix is at least as large as its maximum eigenvalue. The third identity follows from the spectral mapping theorem and elementary properties of the maximum eigenvalue map. □

2.3. **The Main Result.** Our key theorem provides a bound on the probability that the partial sum of a matrix-valued random process is large.

Theorem 2.3. *Consider an adapted sequence $\{\boldsymbol{X}_k\}$ and a previsible sequence $\{\boldsymbol{V}_k\}$ of self-adjoint matrices with dimension d. Assume these sequences satisfy the relations*

$$\log \mathbb{E}_{k-1} \mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq g(\theta) \cdot \boldsymbol{V}_k \quad \textit{almost surely for each } \theta > 0,$$

where $g : (0, \infty) \to [0, \infty]$. Define the partial sums

$$\boldsymbol{Y}_k := \sum\nolimits_{j=1}^{k} \boldsymbol{X}_j \quad \textit{and} \quad \boldsymbol{W}_k := \sum\nolimits_{j=1}^{k} \boldsymbol{V}_j.$$

For all $t, w \in \mathbb{R}$,

$$\mathbb{P}\left\{\exists k : \lambda_{\max}(\boldsymbol{Y}_k) \geq t \quad \textit{and} \quad \lambda_{\max}(\boldsymbol{W}_k) \leq w\right\} \leq d \cdot \inf_{\theta > 0} \mathrm{e}^{-\theta t + g(\theta) w}.$$

In particular, the cumulant bound holds when

$$\mathbb{E}_{k-1} \mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq \mathrm{e}^{g(\theta) \cdot \boldsymbol{V}_k} \quad \textit{almost surely for each } \theta > 0.$$

Proof. First, note that the cgf hypothesis holds when

$$\mathbb{E}_{k-1} \mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq \mathrm{e}^{g(\theta) \cdot \boldsymbol{V}_k}$$

because of the operator monotonicity (A.8) of the logarithm.

The strategy for the main argument is identical with the stopping-time technique used by Freedman [Fre75]. Fix a positive parameter θ, which we will optimize later. Following the discussion in Section 2.2, we introduce the random process $S_k = G_\theta(\boldsymbol{Y}_k, \boldsymbol{W}_k)$. Lemma 2.1 implies that $\{S_k\}$ is a positive supermartingale with initial value d. Let us emphasize that these simple properties of the auxiliary random process distill all the essential information from the hypotheses of the theorem.

Define a stopping time κ by finding the first time instant k when the maximum eigenvalue of the partial sum process $\{\boldsymbol{Y}_k\}$ reaches the level t even though the sum of cumulant bounds has maximum eigenvalue no larger than w.

$$\kappa := \inf\{k \geq 0 : \lambda_{\max}(\boldsymbol{Y}_k) \geq t \quad \text{and} \quad \lambda_{\max}(\boldsymbol{W}_k) \leq w\}.$$

When the infimum is empty, the stopping time $\kappa = \infty$. Consider a system of exceptional events:

$$E_k := \{\lambda_{\max}(\boldsymbol{Y}_k) \geq t \quad \text{and} \quad \lambda_{\max}(\boldsymbol{W}_k) \leq w\} \quad \text{for } k = 0, 1, 2, \ldots.$$

Construct the event $E := \bigcup_{k=0}^{\infty} E_k$ that one or more of these exceptional situations takes place. The intuition behind this definition is that the partial sum $\boldsymbol{Y}_k$ is typically not large unless the process $\{\boldsymbol{X}_k\}$ has varied substantially, a situation that the bound on $\boldsymbol{W}_k$ disallows. As a result, the event E is rather unlikely.

We are prepared to estimate the probability of the exceptional event. First, note that $\kappa < \infty$ on the event E. Therefore, Lemma 2.2 provides a conditional lower bound for the process $\{S_k\}$ at the stopping time κ:

$$S_\kappa = G_\theta(\boldsymbol{Y}_\kappa, \boldsymbol{W}_\kappa) \geq \mathrm{e}^{\theta t - g(\theta) w} \quad \text{on the event } E.$$

Since $\mathbb{E}\, S_k \leq d$ for each (finite) index k,

$$d \geq \sum\nolimits_{k=1}^{\infty} \mathbb{E}[S_\kappa \,|\, \kappa = k] \cdot \mathbb{P}\{\kappa = k\} = \mathbb{E}[S_\kappa \,|\, \kappa < \infty] \geq \int_{\{\kappa < \infty\}} S_\kappa \,\mathrm{d}\mathbb{P}$$
$$\geq \int_E S_\kappa \,\mathrm{d}\mathbb{P} \geq \mathbb{P}(E) \cdot \inf\nolimits_E S_\kappa \geq \mathbb{P}(E) \cdot \mathrm{e}^{\theta t - g(\theta) w}.$$

We require the fact that S_κ is positive to justify these inequalities. Rearrange the relation to obtain

$$\mathbb{P}(E) \leq d \cdot \mathrm{e}^{-\theta t + g(\theta) w}.$$

Minimize the right-hand side with respect to θ to complete the main part of the argument. □

We often prefer to use a corollary of Theorem 2.3 that describes the sum of a finite process. This focus allows us to avoid distracting details about the convergence of infinite series.

Corollary 2.4. *Suppose the hypotheses of Theorem 2.3 are in force, and suppose the random processes are finite in length. Define*

$$\boldsymbol{Y} := \sum\nolimits_k \boldsymbol{X}_k \quad \textit{and} \quad \boldsymbol{W} := \sum\nolimits_k \boldsymbol{V}_k.$$

For all $t, w \in \mathbb{R}$,

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \geq t \quad \textit{and} \quad \lambda_{\max}(\boldsymbol{W}) \leq w\right\} \leq d \cdot \inf_{\theta > 0} \mathrm{e}^{-\theta t + g(\theta) w}.$$

3. Sums of Random Semidefinite Matrices

In this section, we establish Chernoff inequalities for the sum of an adapted sequence of random psd matrices. This result extends the Chernoff bounds for independent random matrices, Theorem 1.2, that we presented in §1.3.

Theorem 3.1 (Matrix Chernoff: Adapted Sequences)**.** *Consider a finite adapted sequence* $\{\boldsymbol{X}_k\}$ *of positive-semidefinite matrices with dimension* d*, and suppose that*

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

Define the finite series

$$\boldsymbol{Y} := \sum\nolimits_k \boldsymbol{X}_k \quad \textit{and} \quad \boldsymbol{W} := \sum\nolimits_k \mathbb{E}_{k-1} \boldsymbol{X}_k.$$

For all $\mu \geq 0$,

$$\mathbb{P}\left\{\lambda_{\min}(\boldsymbol{Y}) \leq (1-\delta)\mu \quad \textit{and} \quad \lambda_{\min}(\boldsymbol{W}) \geq \mu\right\} \leq d \cdot \left[\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu/R} \quad \textit{for } \delta \in [0,1), \textit{ and}$$

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \geq (1+\delta)\mu \quad \textit{and} \quad \lambda_{\max}(\boldsymbol{W}) \leq \mu\right\} \leq d \cdot \left[\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu/R} \quad \textit{for } \delta \geq 0.$$

The Chernoff bound for independent random matrices, Theorem 1.2, follows as an immediate corollary.

Proof of Theorem 1.2 from Theorem 3.1. In this case, we assume that $\{\boldsymbol{X}_k\}$ is an independent sequence of psd matrices. Then the matrix $\boldsymbol{W}$ is not random, so we can define the numbers

$$\mu_{\min} := \lambda_{\min}(\boldsymbol{W}) \quad \text{and} \quad \mu_{\max} := \lambda_{\max}(\boldsymbol{W}).$$

As a consequence, we can replace μ with $\mu_{\min}$ or $\mu_{\max}$, as appropriate, and remove the part of the event involving $\boldsymbol{W}$ from both probabilities in Theorem 3.1. □

3.1. Proofs.

We begin with a semidefinite bound for the mgf of a random psd matrix. This argument transfers a linear upper bound for the scalar exponential to the matrix case.

Lemma 3.2 (Chernoff mgf)**.** *Suppose that* $\boldsymbol{X}$ *is a random psd matrix that satisfies* $\lambda_{\max}(\boldsymbol{X}) \leq 1$. *Then*

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

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

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

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 (A.3) implies that

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

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 (A.4). □

We prove the upper Chernoff bound first, since the argument is slightly easier.

Proof of Theorem 3.1, Upper Bound. By homogeneity, we may assume that $\lambda_{\max}(\boldsymbol{X}_k) \leq 1$; the general case follows by re-scaling. An application of Lemma 3.2 demonstrates that

$$\mathbb{E}_{k-1}\, \mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq \mathrm{e}^{g(\theta) \cdot \mathbb{E}_{k-1} \boldsymbol{X}_k} \quad \text{where } g(\theta) = \mathrm{e}^{\theta} - 1 \text{ for } \theta > 0.$$

Corollary 2.4 provides that

$$\mathbb{P}\left\{ \lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k \right) \geq (1+\delta)\mu \quad \text{and} \quad \lambda_{\max}\left(\sum\nolimits_k \mathbb{E}_{k-1} \boldsymbol{X}_k \right) \leq \mu \right\} \leq d \cdot \inf_{\theta > 0} \mathrm{e}^{-\theta(1+\delta)\mu + g(\theta)\mu}.$$

The infimum is achieved when $\theta = \log(1+\delta)$. Substitute and simplify to complete the proof. □

The lower Chernoff bound follows from a similar argument.

Proof of Theorem 3.1, Lower Bound. As before, we may assume that $\lambda_{\max}(\boldsymbol{X}_k) \leq 1$. This time, we intend to apply Corollary 2.4 to the sequence $\{-\boldsymbol{X}_k\}$. Lemma 3.2 demonstrates that

$$\mathbb{E}_{k-1}\, \mathrm{e}^{(-\theta)\boldsymbol{X}_k} \preccurlyeq \mathrm{e}^{g(\theta) \cdot \mathbb{E}_{k-1}(-\boldsymbol{X}_k)} \quad \text{where } g(\theta) = 1 - \mathrm{e}^{-\theta} \text{ for } \theta > 0.$$

Corollary 2.4 delivers

$$\mathbb{P}\left\{ \lambda_{\max}\left(-\sum\nolimits_k \boldsymbol{X}_k \right) \geq -(1-\delta)\mu \quad \text{and} \quad \lambda_{\max}\left(-\sum\nolimits_k \mathbb{E}_{k-1} \boldsymbol{X}_k \right) \leq -\mu \right\} \leq d \cdot \inf_{\theta > 0} \mathrm{e}^{(\theta(1-\delta) - g(\theta))\mu}.$$

Since $\lambda_{\max}(-\boldsymbol{A}) = -\lambda_{\min}(\boldsymbol{A})$ for each s.a. matrix $\boldsymbol{A}$, we can draw the negation out of the eigenvalue maps and reverse the sense of the inequalities inside the probability. Finally, we observe that the infimum occurs when $\theta = -\log(1-\delta)$. □

4. Incorporating Variance Information

In this section, we establish a variant of the Freedman inequality for martingales [Fre75, Thm. (1.6)]. This inequality demonstrates that a sum of random matrices has normal concentration around its mean and Poisson-type decay in the tails.

Theorem 4.1 (Matrix Bennett: Adapted Sequences)**.** *Consider a finite adapted sequence $\{\boldsymbol{X}_k\}$ of self-adjoint matrices with dimension d that satisfy the relations*

$$\mathbb{E}_{k-1} \boldsymbol{X}_k = \mathbf{0} \quad \textit{and} \quad \lambda_{\max}(\boldsymbol{X}_k) \leq R \quad \textit{almost surely.}$$

Define the finite series

$$\boldsymbol{Y} := \sum\nolimits_k \boldsymbol{X}_k \quad \textit{and} \quad \boldsymbol{W} := \sum\nolimits_k \mathbb{E}_{k-1}\left(\boldsymbol{X}_k^2 \right).$$

For all $t \geq 0$ and $\sigma^2 > 0$,

$$\mathbb{P}\left\{ \lambda_{\max}(\boldsymbol{Y}) \geq t \quad \textit{and} \quad \lambda_{\max}(\boldsymbol{W}) \leq \sigma^2 \right\} \leq d \cdot \exp\left\{ -\frac{\sigma^2}{R^2} \cdot h\left(\frac{Rt}{\sigma^2} \right) \right\}.$$

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

We obtain a Freedman-type inequality for matrix martingales when we simplify the right-hand side of the probability bound in Theorem 4.1.

Corollary 4.2 (Matrix Freedman)**.** *Under the hypotheses of Theorem 4.1,*

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

Proof. This corollary is a direct consequence of Theorem 4.1 and the numerical inequality

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

which can be obtained by comparing derivatives. □

The Bernstein inequality, Theorem 1.3, for sums of independent random matrices follows directly from the Freedman inequality, Corollary 4.2.

Proof of Theorem 1.3 from Corollary 4.2. Indeed, when $\{\boldsymbol{X}_k\}$ is an independent family of random matrices, the matrix $\boldsymbol{W}$ is deterministic. Therefore, if the bound $\sigma^2 \geq \|\boldsymbol{W}\|$ holds, then it holds almost surely. As a result, we can remove the condition on $\boldsymbol{W}$ from the probability bound in the theorem. We can derive a matrix Bennett inequality from Theorem 4.1 in precisely the same manner. □

The proof of Theorem 4.1 appears below. Remark 4.4 shows that we can obtain the same results if we are provided with a set of bounds on the moments of the summands.

4.1. **Proofs.** The first lemma shows how to bound the mgf of a zero-mean random matrix using an almost-sure bound for its largest eigenvalue. We learned this argument from Yao-Liang Yu.

Lemma 4.3 (Bennett mgf)**.** *Suppose that $\boldsymbol{X}$ is a random s.a. matrix that satisfies*

$$\mathbb{E}\,\boldsymbol{X} = \mathbf{0} \quad \textit{and} \quad \lambda_{\max}(\boldsymbol{X}) \leq 1 \quad \textit{almost surely.}$$

Then

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

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

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

An exercise in differential calculus verifies that f is nonnegative and increasing. The matrix $\boldsymbol{X}$ has a (random) eigenvalue decomposition $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^*$ where $\boldsymbol{\Lambda} \preccurlyeq \mathbf{I}$ almost surely. We see that

$$f(\boldsymbol{X}) = \boldsymbol{Q} f(\boldsymbol{\Lambda}) \boldsymbol{Q}^* \preccurlyeq \boldsymbol{Q} \cdot f(\mathbf{I}) \cdot \boldsymbol{Q}^* = f(1) \cdot \mathbf{I}.$$

Expanding the matrix exponential and invoking the conjugation rule (A.2), we discover that

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

To complete the proof, we take the expectation of this semidefinite relation.

$$\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \mathbf{I} + f(1) \cdot \mathbb{E}(\boldsymbol{X}^2) \preccurlyeq \exp\big(f(1) \cdot \mathbb{E}(\boldsymbol{X}^2)\big)$$

The final step follows from (A.4). □

We are ready to establish the Bennett inequality for adapted sequences of random matrices.

Proof of Theorem 4.1. We assume that $R = 1$; the general result follows by re-scaling since $\boldsymbol{Y}$ is 1-homogeneous and $\boldsymbol{W}$ is 2-homogeneous. Invoke Lemma 4.3 to see that

$$\mathbb{E}_{k-1}\,\mathrm{e}^{\theta \boldsymbol{X}_k} \preccurlyeq \exp\big(g(\theta) \cdot \mathbb{E}_{k-1}\left(\boldsymbol{X}_k^2\right)\big) \quad \text{where } g(\theta) = \mathrm{e}^\theta - \theta - 1.$$

Corollary 2.4 implies that

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_k \boldsymbol{X}_k\right) \geq t \quad \text{and} \quad \lambda_{\max}\left(\sum\nolimits_k \mathbb{E}_{k-1}\left(\boldsymbol{X}_k^2\right)\right) \leq \sigma^2\right\} \leq d \cdot \inf_{\theta > 0} \mathrm{e}^{\theta t - g(\theta)\sigma^2}.$$

The infimum is achieved when $\theta = \log(1 + t/\sigma^2)$. □

Remark 4.4. We can also establish the Bennett mgf bound under appropriate hypotheses on the growth of the moments of $\boldsymbol{X}$. This argument proceeds by estimating each term in the Taylor series of the matrix exponential.

Suppose that $\boldsymbol{X}$ is a random s.a. matrix with $\mathbb{E}\,\boldsymbol{X} = \mathbf{0}$, and assume the moment growth bounds

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

We demonstrate that

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

Indeed, the growth condition for the moments yields the bound

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

As usual, the last relation follows from (A.4).

5. Rademacher and Gaussian Series

This section establishes normal concentration for Rademacher and Gaussian series with matrix coefficients. The first step is to verify the bounds for the mgf of a fixed matrix modulated by a Rademacher variable or a Gaussian variable; see also [Oli10b, Lem. 2].

Lemma 5.1 (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 absorbing θ into $\boldsymbol{A}$, we may assume $\theta = 1$ in each case. We begin with the Rademacher mgf. 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 (A.5).

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 1.1. Let $\{\xi_k\}$ be a finite sequence of independent Rademacher variables or independent standard normal variables. Invoke Lemma 5.1 to obtain

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

By assumption, $\lambda_{\max}(\sum_k \boldsymbol{A}_k^2) \le \sigma^2$ almost surely. Therefore, Corollary 2.3 yields

$$\mathbb{P}\left\{ \lambda_{\max}\left(\sum\nolimits_k \xi_k \boldsymbol{A}_k \right) \ge t \right\} \le d \cdot \inf_{\theta > 0} \mathrm{e}^{-\theta t + \theta^2\sigma^2/2}.$$

The infimum is attained at $\theta = t/\sigma^2$. □

Appendix A. Mathematical Background

This section provides a short introduction to the background material we use in our proofs. Section A.1 discusses matrix theory, and Section A.2 reviews some relevant ideas from probability.

A.1. **Matrix Theory.**

Most of these results can be located in Bhatia's books on matrix analysis [Bha97, Bha07]. The works of Horn and Johnson [HJ85, HJ94] also serve as good general references. Higham's book [Hig08] is an excellent source for information about matrix functions.

A.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 size of a matrix when it is determined by the context.

A.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.

A square matrix that satisfies $\boldsymbol{A} = \boldsymbol{A}^*$ is called *self-adjoint* (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.

A.1.3. *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

$$\lambda_{\max}(\boldsymbol{A}) \leq \operatorname{tr} \boldsymbol{A} \quad \text{when } \boldsymbol{A} \text{ is psd} \tag{A.1}$$

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:

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \Longrightarrow \quad \boldsymbol{B}^*\boldsymbol{A}\boldsymbol{B} \preccurlyeq \boldsymbol{B}^*\boldsymbol{H}\boldsymbol{B} \quad \text{for each matrix } \boldsymbol{B}. \tag{A.2}$$

We refer to (A.2) as the *conjugation rule.*

A.1.4. *Matrix Functions.* Let us describe the most direct method for lifting 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:

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

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

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

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 our definition.

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

$$f(a) \leq g(a) \quad \text{for } a \in I \quad \Longrightarrow \quad f(\boldsymbol{A}) \preccurlyeq g(\boldsymbol{A}) \quad \text{when the eigenvalues of } \boldsymbol{A} \text{ lie in } I. \tag{A.3}$$

Indeed, let us decompose $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^*$. It is immediate that $f(\boldsymbol{\Lambda}) \preccurlyeq g(\boldsymbol{\Lambda})$. Conjugate by $\boldsymbol{Q}$, as justified by (A.2), and invoke the definition of a matrix function. We sometimes refer to (A.3) as the *transfer rule.*

When a real function has a convergent 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.

A.1.5. *The Matrix Exponential.* We may define the matrix exponential of an s.a. matrix $\boldsymbol{A}$ via the power series

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

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

On account of the transfer rule (A.3), the matrix exponential satisfies some simple semidefinite relations that we collect here. Since $1 + a \leq \mathrm{e}^a$ for real a, we have

$$\mathbf{I} + \boldsymbol{A} \preccurlyeq \mathrm{e}^{\boldsymbol{A}} \quad \text{for each s.a. matrix } \boldsymbol{A}. \tag{A.4}$$

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

$$\cosh(\boldsymbol{A}) \preccurlyeq \mathrm{e}^{\boldsymbol{A}^2/2} \quad \text{for each s.a. matrix } \boldsymbol{A}. \tag{A.5}$$

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 monotone with respect to the semidefinite order:

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \Longrightarrow \quad \operatorname{tr} \mathrm{e}^{\boldsymbol{A}} \leq \operatorname{tr} \mathrm{e}^{\boldsymbol{H}}. \tag{A.6}$$

See [Pet94, Sec. 2] for a short proof of this fact.

A.1.6. *The Matrix Logarithm.* The matrix logarithm is defined as the functional inverse of the matrix exponential:

$$\log\left(\mathrm{e}^{\boldsymbol{A}}\right) := \boldsymbol{A} \quad \text{for each s.a. matrix } \boldsymbol{A}. \tag{A.7}$$

This formula determines the logarithm on the pd cone, which is adequate for our purposes. The matrix logarithm is monotone with respect to the semidefinite order.

$$\mathbf{0} \prec \boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \Longrightarrow \quad \log(\boldsymbol{A}) \preccurlyeq \log(\boldsymbol{H}). \tag{A.8}$$

A.1.7. *A Theorem of Lieb.* The central tool in this paper is a deep theorem of Lieb from his seminal 1973 work 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]. For another approach that is based on the joint convexity of quantum relative entropy [Lin74, Lem. 2], see the recent note [Tro10b].

Theorem A.1 (Lieb)**.** *Fix a self-adjoint matrix $\boldsymbol{H}$. The function*

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

is concave on the positive-definite cone.

A.2. **Probability.** We continue with some material from probability, focusing on connections with matrices. Rogers and Williams [RW00] is our main source for information about martingales.

A.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." We reserve the letters $\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}$ for random s.a. matrices.

A.2.2. *Adapted Sequences.* Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a master probability space. Consider a filtration $\{\mathscr{F}_k\}$ contained in the master sigma algebra:

$$\mathscr{F}_0 \subset \mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots \subset \mathscr{F}_\infty \subset \mathscr{F}.$$

Given such a filtration, we define the conditional expectation

$$\mathbb{E}_k[\,\cdot\,] := \mathbb{E}[\,\cdot \mid \mathscr{F}_k].$$

We say that a sequence $\{\boldsymbol{X}_k\}$ of random matrices is *adapted* to the filtration when each $\boldsymbol{X}_k$ is measurable with respect to $\mathscr{F}_k$. Loosely speaking, an adapted sequence is one where the present depends only upon the past.

We say that a sequence $\{\boldsymbol{V}_k\}$ of random matrices is *previsible* when each $\boldsymbol{V}_k$ is measurable with respect to $\mathscr{F}_{k-1}$. In particular, the sequence $\{\mathbb{E}_{k-1} \boldsymbol{X}_k\}$ of conditional expectations of an adapted sequence $\{\boldsymbol{X}_k\}$ is previsible.

A *stopping time* is a random variable $\kappa : \Omega \to \{0, 1, 2, \dots, \infty\}$ that satisfies

$$\{\kappa \le k\} \subset \mathscr{F}_k \quad \text{for } k = 0, 1, 2, \dots, \infty.$$

In words, we can determine if the stopping time has arrived from past experience.

A.2.3. *Matrix Martingales.* We say that an adapted sequence $\{\boldsymbol{Y}_k : k = 0, 1, 2, \dots\}$ of s.a. matrices is a *matrix martingale* when

$$\mathbb{E}_{k-1} \boldsymbol{Y}_k = \boldsymbol{Y}_{k-1} \quad \text{for } k = 1, 2, 3, \dots.$$

We also impose an L_1 boundedness criterion:

$$\mathbb{E}\, \|\boldsymbol{Y}_k\| < \infty \quad \text{for } k = 1, 2, 3, \dots.$$

Since all norms on a finite-dimensional space are equivalent, this condition is the same as the requirement that each coordinate of each matrix $\boldsymbol{Y}_k$ is integrable. It follows that we obtain a scalar martingale if we track any fixed coordinate of the sequence $\{\boldsymbol{Y}_k\}$.

Given a matrix martingale $\{\boldsymbol{Y}_k\}$, we construct the *difference sequence*

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

Observe that the difference sequence is conditionally zero mean: $\mathbb{E}_{k-1} \boldsymbol{X}_k = \boldsymbol{0}$. Alternatively, we may begin with an adapted sequence $\{\boldsymbol{X}_k\}$ of conditionally zero-mean random matrices and then form the partial sum process

$$\boldsymbol{Y}_0 := \boldsymbol{0} \quad \text{and} \quad \boldsymbol{Y}_k := \sum\nolimits_{j=1}^{k} \boldsymbol{X}_j.$$

It is easy to verify that $\{\boldsymbol{Y}_k\}$ is a martingale, provided that the integrability requirement holds.

A.2.4. *Inequalities for Expectation.* 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

$$\mathbb{E}\, f(\boldsymbol{Z}) \le f(\mathbb{E}\, \boldsymbol{Z}) \quad \text{when } f \text{ is concave.} \tag{A.9}$$

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, let us emphasize that each of these bounds holds when we replace the expectation $\mathbb{E}$ by the conditional expectation $\mathbb{E}_k$.

Acknowledgments

I would like to thank Vern Paulsen and Bernhard Bodmann for some helpful conversations connected with this project. Klas Markström and David Gross provided some references to related work. Roberto Oliveira introduced me to Freedman's inequality and encouraged me to apply the methods in the paper [Tro10c] to this problem. It was Oliveira's elegant work [Oli10b] on matrix probability inequalities that spurred me to pursue this project in the first place. Finally, I would like to thank Yao-Liang Yu, who pointed out an inconsistency in the proof of Theorem 2.3 and who proposed the argument in Lemma 4.3. Richard Chen and Alex Gittens have also helped me root out typographic errors.

References

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

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

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

[Fre75] D. A. Freedman. On tail probabilities for martingales. *Ann. Probab.*, 3(1):100–118, Feb. 1975.

[Hig08] N. J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008.

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

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

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

[Lin74] G. Lindblad. Expectations and entropy inequalities for finite quantum systems. *Comm. Math. Phys.*, 39:111–119, 1974.

[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. *Elect. Comm. Probab.*, 15:203–212, 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.

[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.

[RW00] L. C. G. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales. Volume I: Foundations.* Cambridge Univ. Press, Cambridge, 2nd edition, 2000.

[Tro10a] J. A. Tropp. Freedman's inequality for matrix martingales. Available at `arXiv`., June 2010.

[Tro10b] J. A. Tropp. From the joint convexity of quantum relative entropy to a concavity theorem of Lieb. Available at `arXiv:1101.1070`, Dec. 2010.

[Tro10c] J. A. Tropp. User-friendly tail bounds for sums of random matrices. Available at `arXiv:1004.4389`, Apr. 2010.