*The Annals of Probability*
2014, Vol. 42, No. 3, 906–945
DOI: 10.1214/13-AOP892


# MATRIX CONCENTRATION INEQUALITIES VIA THE METHOD OF EXCHANGEABLE PAIRS[1]

BY LESTER MACKEY[2], MICHAEL I. JORDAN, RICHARD Y. CHEN[3], BRENDAN FARRELL[3] AND JOEL A. TROPP[3]

*Stanford University, University of California, Berkeley, California Institute of Technology, California Institute of Technology and California Institute of Technology*

This paper derives exponential concentration inequalities and polynomial moment inequalities for the spectral norm of a random matrix. The analysis requires a matrix extension of the scalar concentration theory developed by Sourav Chatterjee using Stein's method of exchangeable pairs. When applied to a sum of independent random matrices, this approach yields matrix generalizations of the classical inequalities due to Hoeffding, Bernstein, Khintchine and Rosenthal. The same technique delivers bounds for sums of dependent random matrices and more general matrix-valued functions of dependent random variables.

**1. Introduction.** Matrix concentration inequalities control the fluctuations of a random matrix about its mean. At present, these results provide an effective method for studying sums of independent random matrices and matrix martingales [32, 35, 48, 49]. They have been used to streamline the analysis of structured random matrices in a range of applications, including statistical estimation [24], randomized linear algebra [10, 14], stability of least-squares approximation [12], combinatorial and robust optimization [9, 46], matrix completion [16, 30, 34, 42] and random graph theory [35]. These works compose only a small sample of the papers that rely on matrix concentration inequalities. Nevertheless, it remains common to encounter new classes of random matrices that we cannot treat with the available techniques.

The purpose of this paper is to lay the foundations of a new approach for analyzing structured random matrices. Our work is based on Chatterjee's technique for developing scalar concentration inequalities [6, 7] via Stein's method of exchangeable pairs [47]. We extend this argument to the matrix setting, where we

Received February 2012; revised February 2013.

[1]Supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under Contract/Grant number W911NF-11-1-0391.

[2]Supported by the National Defense Science and Engineering Graduate Fellowship.

[3]Supported by ONR awards N00014-08-1-0883 and N00014-11-1002, AFOSR award FA9550-09-1-0643, DARPA award N66001-08-1-2065 and a Sloan Research Fellowship.

*MSC2010 subject classifications.* Primary 60B20, 60E15; secondary 60G09, 60F10.

*Key words and phrases.* Concentration inequalities, moment inequalities, Stein's method, exchangeable pairs, random matrix, noncommutative.

use it to establish exponential concentration bounds (Theorems 4.1 and 5.1) and polynomial moment inequalities (Theorem 7.1) for the spectral norm of a random matrix.

To illustrate the power of this idea, we show that our general results imply several important concentration bounds for a sum of independent, random, Hermitian matrices [21, 29, 49]. In particular, we obtain a matrix Hoeffding inequality with optimal constants (Corollary 4.2) and a version of the matrix Bernstein inequality (Corollary 5.2). Our techniques also yield concise proofs of the matrix Khintchine inequality (Corollary 7.3) and the matrix Rosenthal inequality (Corollary 7.4).

The method of exchangeable pairs also applies to matrices constructed from dependent random variables. We offer a hint of the prospects by establishing concentration results for several other classes of random matrices. In Section 9, we consider sums of dependent matrices that satisfy a conditional zero-mean property. In Section 10, we treat a broad class of combinatorial matrix statistics. Finally, in Section 11, we analyze general matrix-valued functions that have a self-reproducing property.

1.1. *Notation and preliminaries.* The symbol $\|\cdot\|$ is reserved for the spectral norm, which returns the largest singular value of a general complex matrix.

We write $\mathbb{M}^d$ for the algebra of all $d \times d$ complex matrices. The trace and normalized trace of a square matrix are defined as

$$\operatorname{tr} \mathbf{B} := \sum_{j=1}^{d} b_{jj} \quad \text{and} \quad \bar{\operatorname{tr}}\, \mathbf{B} := \frac{1}{d} \sum_{j=1}^{d} b_{jj} \qquad \text{for } \mathbf{B} \in \mathbb{M}^d.$$

We define the linear space $\mathbb{H}^d$ of Hermitian $d \times d$ matrices. *All matrices in this paper are Hermitian unless explicitly stated otherwise.* The symbols $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ refer to the algebraic maximum and minimum eigenvalues of a matrix $\mathbf{A} \in \mathbb{H}^d$. For each interval $I \subset \mathbb{R}$, we define the set of Hermitian matrices whose eigenvalues fall in that interval,

$$\mathbb{H}^d(I) := \{\mathbf{A} \in \mathbb{H}^d : [\lambda_{\min}(\mathbf{A}), \lambda_{\max}(\mathbf{A})] \subset I\}.$$

The set $\mathbb{H}^d_+$ consists of all positive-semidefinite (psd) $d \times d$ matrices. Curly inequalities refer to the semidefinite partial order on Hermitian matrices. For example, we write $\mathbf{A} \preccurlyeq \mathbf{B}$ to signify that the matrix $\mathbf{B} - \mathbf{A}$ is psd.

We require operator convexity properties of the matrix square so often that we state them now:

$$\left(\frac{\mathbf{A} + \mathbf{B}}{2}\right)^2 \preccurlyeq \frac{\mathbf{A}^2 + \mathbf{B}^2}{2} \qquad \text{for all } \mathbf{A}, \mathbf{B} \in \mathbb{H}^d. \tag{1.1}$$

More generally, we have the operator Jensen inequality

$$(\mathbb{E}\, \mathbf{X})^2 \preccurlyeq \mathbb{E}\, \mathbf{X}^2, \tag{1.2}$$

valid for any random Hermitian matrix, provided that $\mathbb{E}\, \|\mathbf{X}\|^2 < \infty$. To verify this result, simply expand the inequality $\mathbb{E}(\mathbf{X} - \mathbb{E}\, \mathbf{X})^2 \succcurlyeq \mathbf{0}$. The operator Jensen inequality also holds for conditional expectation, again provided that $\mathbb{E}\, \|\mathbf{X}\|^2 < \infty$.

**2. Exchangeable pairs of random matrices.** Our approach to studying random matrices is based on the method of exchangeable pairs, which originates in the work of Charles Stein [47] on normal approximation for a sum of dependent random variables. In this section, we explain how some central ideas from this theory extend to matrices.

2.1. *Matrix Stein pairs*. First, we define an exchangeable pair.

DEFINITION 2.1 (Exchangeable pair). Let $Z$ and $Z'$ be random variables taking values in a Polish space $\mathcal{Z}$. We say that $(Z, Z')$ is an *exchangeable pair* if it has the same distribution as $(Z', Z)$. In particular, $Z$ and $Z'$ must share the same distribution.

We can obtain a lot of information about the fluctuations of a random matrix $\mathbf{X}$ if we can construct a good exchangeable pair $(\mathbf{X}, \mathbf{X}')$. With this motivation in mind, let us introduce a special class of exchangeable pairs.

DEFINITION 2.2 (Matrix Stein pair). Let $(Z, Z')$ be an exchangeable pair of random variables taking values in a Polish space $\mathcal{Z}$, and let $\boldsymbol{\Psi} : \mathcal{Z} \to \mathbb{H}^d$ be a measurable function. Define the random Hermitian matrices

$$\mathbf{X} := \boldsymbol{\Psi}(Z) \quad \text{and} \quad \mathbf{X}' := \boldsymbol{\Psi}(Z').$$

We say that $(\mathbf{X}, \mathbf{X}')$ is a *matrix Stein pair* if there is a constant $\alpha \in (0, 1]$ for which

$$\mathbb{E}[\mathbf{X} - \mathbf{X}' \mid Z] = \alpha \mathbf{X} \qquad \text{almost surely.} \tag{2.1}$$

The constant $\alpha$ is called the *scale factor* of the pair. When discussing a matrix Stein pair $(\mathbf{X}, \mathbf{X}')$, we always assume that $\mathbb{E}\,\|\mathbf{X}\|^2 < \infty$.

A matrix Stein pair $(\mathbf{X}, \mathbf{X}')$ has several useful properties. First, $(\mathbf{X}, \mathbf{X}')$ always forms an exchangeable pair. Second, it must be the case that $\mathbb{E}\,\mathbf{X} = \mathbf{0}$. Indeed,

$$\mathbb{E}\,\mathbf{X} = \frac{1}{\alpha}\,\mathbb{E}\big[\mathbb{E}[\mathbf{X} - \mathbf{X}' \mid Z]\big] = \frac{1}{\alpha}\,\mathbb{E}[\mathbf{X} - \mathbf{X}'] = \mathbf{0}$$

because of identity (2.1), the tower property of conditional expectation and the exchangeability of $(\mathbf{X}, \mathbf{X}')$. In Section 2.4, we construct a matrix Stein pair for a sum of centered, independent random matrices. More sophisticated examples appear in Sections 9, 10 and 11.

REMARK 2.3 (Approximate matrix Stein pairs). In the scalar setting, it is common to consider exchangeable pairs that satisfy an approximate Stein condition. For matrices, this condition reads $\mathbb{E}[\mathbf{X} - \mathbf{X}' \mid Z] = \alpha\mathbf{X} + \mathbf{R}$, where $\mathbf{R}$ is an error term. The methods in this paper extend easily to this case.

2.2. *The method of exchangeable pairs.* A well-chosen matrix Stein pair $(\mathbf{X}, \mathbf{X}')$ provides a surprisingly powerful tool for studying the random matrix $\mathbf{X}$. The technique depends on a fundamental technical lemma.

LEMMA 2.4 (Method of exchangeable pairs). *Suppose that* $(\mathbf{X}, \mathbf{X}') \in \mathbb{H}^d \times \mathbb{H}^d$ *is a matrix Stein pair with scale factor* $\alpha$. *Let* $\mathbf{F}: \mathbb{H}^d \to \mathbb{H}^d$ *be a measurable function that satisfies the regularity condition*

$$\mathbb{E}\|(\mathbf{X} - \mathbf{X}') \cdot \mathbf{F}(\mathbf{X})\| < \infty. \tag{2.2}$$

*Then*

$$\mathbb{E}[\mathbf{X} \cdot \mathbf{F}(\mathbf{X})] = \frac{1}{2\alpha} \mathbb{E}[(\mathbf{X} - \mathbf{X}')(\mathbf{F}(\mathbf{X}) - \mathbf{F}(\mathbf{X}'))]. \tag{2.3}$$

In short, the randomness in the Stein pair furnishes an alternative expression for the expected product of $\mathbf{X}$ and the function $\mathbf{F}$. Identity (2.3) is valuable because it allows us to estimate this integral using the smoothness properties of the function $\mathbf{F}$ and the discrepancy between $\mathbf{X}$ and $\mathbf{X}'$.

PROOF OF LEMMA 2.4. Suppose $(\mathbf{X}, \mathbf{X}')$ is a matrix Stein pair constructed from an auxiliary exchangeable pair $(Z, Z')$. The defining property (2.1) implies

$$\alpha \cdot \mathbb{E}[\mathbf{X} \cdot \mathbf{F}(\mathbf{X})] = \mathbb{E}[\mathbb{E}[\mathbf{X} - \mathbf{X}' \mid Z] \cdot \mathbf{F}(\mathbf{X})] = \mathbb{E}[(\mathbf{X} - \mathbf{X}')\mathbf{F}(\mathbf{X})].$$

We have used regularity condition (2.2) to invoke the pull-through property of conditional expectation. Since $(\mathbf{X}, \mathbf{X}')$ is an exchangeable pair,

$$\mathbb{E}[(\mathbf{X} - \mathbf{X}')\mathbf{F}(\mathbf{X})] = \mathbb{E}[(\mathbf{X}' - \mathbf{X})\mathbf{F}(\mathbf{X}')] = -\mathbb{E}[(\mathbf{X} - \mathbf{X}')\mathbf{F}(\mathbf{X}')].$$

Identity (2.3) follows when we average the two preceding displays. □

2.3. *The conditional variance.* To each matrix Stein pair $(\mathbf{X}, \mathbf{X}')$, we may associate a random matrix called the *conditional variance* of $\mathbf{X}$. The ultimate purpose of this paper is to argue that the spectral norm of $\mathbf{X}$ is unlikely to be large when the conditional variance is small.

DEFINITION 2.5 (Conditional variance). Suppose that $(\mathbf{X}, \mathbf{X}')$ is a matrix Stein pair, constructed from an auxiliary exchangeable pair $(Z, Z')$. The *conditional variance* is the random matrix

$$\boldsymbol{\Delta}_{\mathbf{X}} := \boldsymbol{\Delta}_{\mathbf{X}}(Z) := \frac{1}{2\alpha} \mathbb{E}[(\mathbf{X} - \mathbf{X}')^2 \mid Z], \tag{2.4}$$

where $\alpha$ is the scale factor of the pair. We may take any version of the conditional expectation in this definition.

The conditional variance $\boldsymbol{\Delta}_{\mathbf{X}}$ should be regarded as a stochastic estimate for the variance of the random matrix $\mathbf{X}$. Indeed,

$$\mathbb{E}[\boldsymbol{\Delta}_{\mathbf{X}}] = \mathbb{E}\,\mathbf{X}^2. \tag{2.5}$$

This identity follows from Lemma 2.4 with the choice $\mathbf{F}(\mathbf{X}) = \mathbf{X}$.

2.4. *Example*: *A sum of independent random matrices*. To make the definitions in this section more vivid, we describe a simple but important example of a matrix Stein pair. Consider an independent sequence $Z := (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ of random Hermitian matrices that satisfies $\mathbb{E}\,\mathbf{Y}_k = \mathbf{0}$ and $\mathbb{E}\,\|\mathbf{Y}_k\|^2 < \infty$ for each $k$. Introduce the random series

$$\mathbf{X} := \mathbf{Y}_1 + \cdots + \mathbf{Y}_n.$$

Let us explain how to build a good matrix Stein pair $(\mathbf{X}, \mathbf{X}')$. We need the exchangeable counterpart $\mathbf{X}'$ to have the same distribution as $\mathbf{X}$, but it should also be close to $\mathbf{X}$ so that we can control the conditional variance. To achieve these goals, we construct $\mathbf{X}'$ by picking a summand from $\mathbf{X}$ at random and replacing it with a fresh copy.

Formally, let $\mathbf{Y}'_k$ be an independent copy of $\mathbf{Y}_k$ for each index $k$, and draw a random index $K$ uniformly from $\{1, \dots, n\}$ and independently from everything else. Define the random sequence

$$Z' := (\mathbf{Y}_1, \dots, \mathbf{Y}_{K-1}, \mathbf{Y}'_K, \mathbf{Y}_{K+1}, \dots, \mathbf{Y}_n).$$

One can check that $(Z, Z')$ forms an exchangeable pair. The random matrix

$$\mathbf{X}' := \mathbf{Y}_1 + \cdots + \mathbf{Y}_{K-1} + \mathbf{Y}'_K + \mathbf{Y}_{K+1} + \cdots + \mathbf{Y}_n$$

is thus an exchangeable counterpart for $\mathbf{X}$. To verify that $(\mathbf{X}, \mathbf{X}')$ is a Stein pair, calculate that

$$\begin{aligned}\mathbb{E}[\mathbf{X} - \mathbf{X}' \mid Z] &= \mathbb{E}[\mathbf{Y}_K - \mathbf{Y}'_K \mid Z] \\ &= \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[\mathbf{Y}_k - \mathbf{Y}'_k \mid Z] = \frac{1}{n}\sum_{k=1}^{n} \mathbf{Y}_k = \frac{1}{n}\mathbf{X}.\end{aligned}$$

The third identity holds because $\mathbf{Y}'_k$ is a centered random matrix that is independent from $Z$. Therefore, $(\mathbf{X}, \mathbf{X}')$ is a matrix Stein pair with scale factor $\alpha = n^{-1}$.

Next, we compute the conditional variance:

$$\begin{aligned}\boldsymbol{\Delta}_{\mathbf{X}} &= \frac{n}{2} \cdot \mathbb{E}[(\mathbf{X} - \mathbf{X}')^2 \mid Z] \\ &= \frac{n}{2} \cdot \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[(\mathbf{Y}_k - \mathbf{Y}'_k)^2 \mid Z] \\ &= \frac{1}{2}\sum_{k=1}^{n} [\mathbf{Y}_k^2 - \mathbf{Y}_k(\mathbb{E}\,\mathbf{Y}'_k) - (\mathbb{E}\,\mathbf{Y}'_k)\mathbf{Y}_k + \mathbb{E}(\mathbf{Y}'_k)^2] \\ &= \frac{1}{2}\sum_{k=1}^{n} (\mathbf{Y}_k^2 + \mathbb{E}\,\mathbf{Y}_k^2).\end{aligned} \tag{2.6}$$

For the third relation, expand the square and invoke the pull-through property of conditional expectation. We may drop the conditioning because $\mathbf{Y}'_k$ is independent

from $Z$. In the last line, we apply the property that $\mathbf{Y}_k'$ has the same distribution as $\mathbf{Y}_k$.

Expression (2.6) shows that we can control the size of the conditional expectation uniformly if we can control the size of the individual summands. This example also teaches us that we may use the symmetries of the distribution of the random matrix to construct a matrix Stein pair.

**3. Exponential moments and eigenvalues of a random matrix.** Our main goal in this paper is to study the behavior of the extreme eigenvalues of a random Hermitian matrix. In Section 3.2, we describe an approach to this problem that parallels the classical Laplace transform method for scalar random variables. The adaptation to the matrix setting leads us to consider the *trace* of the moment generating function (m.g.f.) of a random matrix. After presenting this background, we explain how the method of exchangeable pairs can be used to control the growth of the trace m.g.f. This result, which appears in Section 3.5, is the key to our exponential concentration bounds for random matrices.

3.1. *Standard matrix functions.* Before entering the discussion, recall that a *standard matrix function* is obtained by applying a real function to the eigenvalues of a Hermitian matrix. Higham [17] provides an excellent treatment of this concept.

DEFINITION 3.1 (Standard matrix function). Let $f : I \to \mathbb{R}$ be a function on an interval $I$ of the real line. Suppose that $\mathbf{A} \in \mathbb{H}^d(I)$ has the eigenvalue decomposition $\mathbf{A} = \mathbf{Q} \cdot \operatorname{diag}(\lambda_1, \dots, \lambda_d) \cdot \mathbf{Q}^*$ where $\mathbf{Q}$ is a unitary matrix. Then the matrix extension $f(\mathbf{A}) := \mathbf{Q} \cdot \operatorname{diag}(f(\lambda_1), \dots, f(\lambda_d)) \cdot \mathbf{Q}^*$.

The *spectral mapping theorem* states that, if $\lambda$ is an eigenvalue of $\mathbf{A}$, then $f(\lambda)$ is an eigenvalue of $f(\mathbf{A})$. This fact follows from Definition 3.1.

When we apply a familiar scalar function to a Hermitian matrix, we are always referring to a standard matrix function. For instance, $|\mathbf{A}|$ is the matrix absolute value, $\exp(\mathbf{A})$ is the matrix exponential, and $\log(\mathbf{A})$ is the matrix logarithm. The latter is defined only for positive-definite matrices.

3.2. *The matrix Laplace transform method.* Let us introduce a matrix variant of the classical moment generating function. We learned this definition from Ahlswede–Winter [1], Appendix.

DEFINITION 3.2 (Trace m.g.f.). Let $\mathbf{X}$ be a random Hermitian matrix. The *(normalized) trace moment generating function* of $\mathbf{X}$ is defined as

$$m(\theta) := m_{\mathbf{X}}(\theta) := \mathbb{E}\,\bar{\operatorname{tr}}\, e^{\theta \mathbf{X}} \qquad \text{for } \theta \in \mathbb{R}.$$

We admit the possibility that the expectation may not exist for all $\theta$.

Ahlswede and Winter [1], Appendix, had the insight that the classical Laplace transform method could be extended to the matrix setting by replacing the classical m.g.f. with the trace m.g.f. This adaptation allows us to obtain concentration inequalities for the extreme eigenvalues of a random Hermitian matrix using methods from matrix analysis. The following proposition distills results from the papers [1, 8, 36, 49].

PROPOSITION 3.3 (Matrix Laplace transform method). *Let $\mathbf{X} \in \mathbb{H}^d$ be a random matrix with trace m.g.f. $m(\theta) := \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta \mathbf{X}}$. For each $t \in \mathbb{R}$,*

$$\mathbb{P}\{\lambda_{\max}(\mathbf{X}) \geq t\} \leq d \cdot \inf_{\theta > 0} \exp\{-\theta t + \log m(\theta)\}, \tag{3.1}$$

$$\mathbb{P}\{\lambda_{\min}(\mathbf{X}) \leq t\} \leq d \cdot \inf_{\theta < 0} \exp\{-\theta t + \log m(\theta)\}. \tag{3.2}$$

*Furthermore*,

$$\mathbb{E}\,\lambda_{\max}(\mathbf{X}) \leq \inf_{\theta > 0} \frac{1}{\theta}[\log d + \log m(\theta)], \tag{3.3}$$

$$\mathbb{E}\,\lambda_{\min}(\mathbf{X}) \geq \sup_{\theta < 0} \frac{1}{\theta}[\log d + \log m(\theta)]. \tag{3.4}$$

Estimates (3.3) and (3.4) for the expectations are usually sharp up to the logarithm of the dimension. In many situations, tail bounds (3.1) and (3.2) are reasonable for moderate $t$, but they tend to overestimate the probability of a large deviation. Note that, in general, we cannot dispense with the dimensional factor $d$. See [49], Section 4, for a detailed discussion of these issues. Additional inequalities for the interior eigenvalues can be established using the minimax Laplace transform method [15].

PROOF OF PROPOSITION 3.3. To establish (3.1), fix $\theta > 0$. Owing to Markov's inequality,

$$\begin{aligned}\mathbb{P}\{\lambda_{\max}(\mathbf{X}) \geq t\} &= \mathbb{P}\{\mathrm{e}^{\lambda_{\max}(\theta \mathbf{X})} \geq \mathrm{e}^{\theta t}\} \leq \mathrm{e}^{-\theta t} \cdot \mathbb{E}\,\mathrm{e}^{\lambda_{\max}(\theta \mathbf{X})} \\ &= \mathrm{e}^{-\theta t} \cdot \mathbb{E}\,\lambda_{\max}(\mathrm{e}^{\theta \mathbf{X}}) \leq \mathrm{e}^{-\theta t} \cdot \mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta \mathbf{X}}.\end{aligned}$$

The third relation depends on the spectral mapping theorem and the monotonicity of the exponential. The last inequality holds because the trace of a positive-definite matrix exceeds its maximum eigenvalue. Identify the normalized trace m.g.f., and take the infimum over $\theta$ to complete the argument.

The proof of (3.2) parallels the proof of (3.1). For $\theta < 0$,

$$\mathbb{P}\{\lambda_{\min}(\mathbf{X}) \leq t\} = \mathbb{P}\{\theta \lambda_{\min}(\mathbf{X}) \geq \theta t\} = \mathbb{P}\{\lambda_{\max}(\theta \mathbf{X}) \geq \theta t\}.$$

We used the property that $-\lambda_{\min}(\mathbf{A}) = \lambda_{\max}(-\mathbf{A})$ for each Hermitian matrix $\mathbf{A}$. The rest of the argument is the same as in the preceding paragraph.

For the expectation bound (3.3), fix $\theta > 0$. Jensen's inequality yields

$$\mathbb{E}\,\lambda_{\max}(\mathbf{X}) = \theta^{-1}\,\mathbb{E}\,\lambda_{\max}(\theta\mathbf{X}) \le \theta^{-1}\log\mathbb{E}\,\mathrm{e}^{\lambda_{\max}(\theta\mathbf{X})} \le \theta^{-1}\log\mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta\mathbf{X}}.$$

The justification is the same as above. Identify the normalized trace m.g.f., and take the infimum over $\theta > 0$. Similar considerations yield (3.4). □

3.3. *Studying the trace m.g.f. with exchangeable pairs.* The technical difficulty in the matrix Laplace transform method arises because we need to estimate the trace m.g.f. Previous authors have applied deep results from matrix analysis to accomplish this bound: the Golden–Thompson inequality is central to [1, 35, 36], while Lieb's result [26], Theorem 6, animates [20, 48, 49].

In this paper, we develop a fundamentally different technique for studying the trace m.g.f. The main idea is to control the *growth* of the trace m.g.f. by bounding its *derivative*. To see why we have adopted this strategy, consider a random Hermitian matrix $\mathbf{X}$, and observe that the derivative of its trace m.g.f. can be written as

$$m'(\theta) = \mathbb{E}\,\bar{\mathrm{tr}}\big[\mathbf{X}\mathrm{e}^{\theta\mathbf{X}}\big]$$

under appropriate regularity conditions. This expression has just the form that we need to invoke the method of exchangeable pairs, Lemma 2.4, with $\mathbf{F}(\mathbf{X}) = \mathrm{e}^{\theta\mathbf{X}}$. We obtain

$$m'(\theta) = \frac{1}{2\alpha}\,\mathbb{E}\,\bar{\mathrm{tr}}\big[(\mathbf{X}-\mathbf{X}')(\mathrm{e}^{\theta\mathbf{X}} - \mathrm{e}^{\theta\mathbf{X}'})\big]. \tag{3.5}$$

This formula strongly suggests that we should apply a mean value theorem to control the derivative; we establish the result that we need in Section 3.4 below. Ultimately, this argument leads to a differential inequality for $m'(\theta)$, which we can integrate to obtain an estimate for $m(\theta)$.

The technique of bounding the derivative of an m.g.f. lies at the heart of the log-Sobolev method for studying concentration phenomena [25], Chapter 5. Recently, Chatterjee [6, 7] demonstrated that the method of exchangeable pairs provides another way to control the derivative of an m.g.f. Our arguments closely follow the pattern set by Chatterjee; the novelty inheres in the extension of these ideas to the matrix setting and the striking applications that this extension permits.

3.4. *The mean value trace inequality.* To bound expression (3.5) for the derivative of the trace m.g.f., we need a matrix generalization of the mean value theorem for a function with a convex derivative. We state the result in full generality because it plays a role later.

LEMMA 3.4 (Mean value trace inequality). *Let $I$ be an interval of the real line. Suppose that $g : I \to \mathbb{R}$ is a weakly increasing function and that $h : I \to \mathbb{R}$ is*

*a function whose derivative $h'$ is convex. For all matrices $\mathbf{A}, \mathbf{B} \in \mathbb{H}^d(I)$, it holds that*

$$\begin{aligned}&\bar{\mathrm{tr}}\big[\big(g(\mathbf{A}) - g(\mathbf{B})\big) \cdot \big(h(\mathbf{A}) - h(\mathbf{B})\big)\big] \\ &\qquad \le \tfrac{1}{2}\,\bar{\mathrm{tr}}\big[\big(g(\mathbf{A}) - g(\mathbf{B})\big) \cdot (\mathbf{A} - \mathbf{B}) \cdot \big(h'(\mathbf{A}) + h'(\mathbf{B})\big)\big].\end{aligned}$$

*When $h'$ is concave, the inequality is reversed. The same results hold for the standard trace.*

To prove Lemma 3.4, we require a trace inequality [38], Proposition 3, that follows from the definition of a matrix function and the spectral theorem for Hermitian matrices.

PROPOSITION 3.5 (Generalized Klein inequality). *Let $u_1, \dots, u_n$ and $v_1, \dots, v_n$ be real-valued functions on an interval $I$ of the real line. Suppose*

$$\sum_k u_k(a) v_k(b) \ge 0 \qquad \text{for all } a, b \in I. \tag{3.6}$$

*Then*

$$\bar{\mathrm{tr}}\Big[\sum_k u_k(\mathbf{A}) v_k(\mathbf{B})\Big] \ge 0 \qquad \text{for all } \mathbf{A}, \mathbf{B} \in \mathbb{H}^d(I).$$

With the generalized Klein inequality, we can establish Lemma 3.4 by developing the appropriate scalar inequality.

PROOF OF LEMMA 3.4. Fix $a, b \in I$. Since $g$ is weakly increasing, $(g(a) - g(b)) \cdot (a - b) \ge 0$. The fundamental theorem of calculus and the convexity of $h'$ yield the estimate

$$\begin{aligned}&\big(g(a) - g(b)\big) \cdot \big(h(a) - h(b)\big) \\ &\qquad = \big(g(a) - g(b)\big) \cdot (a - b) \int_0^1 h'\big(\tau a + (1 - \tau) b\big)\,\mathrm{d}\tau \\ &\qquad \le \big(g(a) - g(b)\big) \cdot (a - b) \int_0^1 \big[\tau \cdot h'(a) + (1 - \tau) \cdot h'(b)\big]\,\mathrm{d}\tau \\ &\qquad = \frac{1}{2}\big[\big(g(a) - g(b)\big) \cdot (a - b) \cdot \big(h'(a) + h'(b)\big)\big].\end{aligned} \tag{3.7}$$

The inequality is reversed when $h'$ is concave.

Bound (3.7) can be written in the form (3.6) by expanding the products and collecting terms depending on $a$ into functions $u_k(a)$ and terms depending on $b$ into functions $v_k(b)$. Proposition 3.5 then delivers a trace inequality, which can be massaged into the desired form using the cyclicity of the trace and the fact that standard functions of the same matrix commute. We omit the algebraic details. □

REMARK 3.6. We must warn the reader that the proof of Lemma 3.4 succeeds because the trace contains a product of *three* terms involving *two* matrices. The obstacle to proving more general results is that we cannot reorganize expressions like tr(**ABAB**) and tr(**ABC**) at will.

3.5. *Bounding the derivative of the trace m.g.f.* The central result in this section applies the method of exchangeable pairs and the mean value trace inequality to bound the derivative of the trace m.g.f. in terms of the conditional variance. This is the most important step in our theory on the exponential concentration of random matrices.

LEMMA 3.7 (The derivative of the trace m.g.f.). *Suppose that* $(\mathbf{X}, \mathbf{X}') \in \mathbb{H}^d \times \mathbb{H}^d$ *is a matrix Stein pair, and assume that* $\mathbf{X}$ *is almost surely bounded in norm. Define the trace m.g.f.* $m(\theta) := \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta\mathbf{X}}$. *Then*

$$m'(\theta) \le \theta \cdot \mathbb{E}\,\bar{\mathrm{tr}}\big[\Delta_{\mathbf{X}}\mathrm{e}^{\theta\mathbf{X}}\big] \qquad \text{when } \theta \ge 0; \tag{3.8}$$

$$m'(\theta) \ge \theta \cdot \mathbb{E}\,\bar{\mathrm{tr}}\big[\Delta_{\mathbf{X}}\mathrm{e}^{\theta\mathbf{X}}\big] \qquad \text{when } \theta \le 0. \tag{3.9}$$

*The conditional variance* $\boldsymbol{\Delta}_{\mathbf{X}}$ *is defined in* (2.4).

PROOF. We begin with the expression for the derivative of the trace m.g.f.,

$$m'(\theta) = \mathbb{E}\,\bar{\mathrm{tr}}\bigg[\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{e}^{\theta\mathbf{X}}\bigg] = \mathbb{E}\,\bar{\mathrm{tr}}\big[\mathbf{X}\mathrm{e}^{\theta\mathbf{X}}\big]. \tag{3.10}$$

We can move the derivative inside the expectation because of the dominated convergence theorem and the boundedness of $\mathbf{X}$.

Apply the method of exchangeable pairs, Lemma 2.4, with the function $\mathbf{F}(\mathbf{X}) = \mathrm{e}^{\theta\mathbf{X}}$ to reach an alternative representation of the derivative (3.10),

$$m'(\theta) = \frac{1}{2\alpha}\,\mathbb{E}\,\bar{\mathrm{tr}}\big[(\mathbf{X} - \mathbf{X}')\big(\mathrm{e}^{\theta\mathbf{X}} - \mathrm{e}^{\theta\mathbf{X}'}\big)\big]. \tag{3.11}$$

We have used the boundedness of $\mathbf{X}$ to verify the regularity condition (2.2).

Expression (3.11) is perfectly suited for an application of the mean value trace inequality, Lemma 3.4. First, assume that $\theta \ge 0$, and consider the function $h : s \mapsto \mathrm{e}^{\theta s}$. The derivative $h' : s \mapsto \theta\mathrm{e}^{\theta s}$ is convex, so Lemma 3.4 implies that

$$\begin{aligned} m'(\theta) &\le \frac{\theta}{4\alpha}\,\mathbb{E}\,\bar{\mathrm{tr}}\big[(\mathbf{X} - \mathbf{X}')^2 \cdot \big(\mathrm{e}^{\theta\mathbf{X}} + \mathrm{e}^{\theta\mathbf{X}'}\big)\big] \\ &= \frac{\theta}{2\alpha}\,\mathbb{E}\,\bar{\mathrm{tr}}\big[(\mathbf{X} - \mathbf{X}')^2 \cdot \mathrm{e}^{\theta\mathbf{X}}\big] \\ &= \theta \cdot \mathbb{E}\,\bar{\mathrm{tr}}\bigg[\frac{1}{2\alpha}\,\mathbb{E}\big[(\mathbf{X} - \mathbf{X}')^2 \mid Z\big] \cdot \mathrm{e}^{\theta\mathbf{X}}\bigg]. \end{aligned}$$

The second line follows from the fact that $(\mathbf{X}, \mathbf{X}')$ is an exchangeable pair. In the last line, we have used the boundedness of $\mathbf{X}$ and $\mathbf{X}'$ to invoke the pull-through

property of conditional expectation. Identify the conditional variance $\boldsymbol{\Delta}_{\mathbf{X}}$, defined in (2.4), to complete the argument.

The result for $\theta \le 0$ follows from an analogous argument. In this case, we simply observe that the derivative of the function $h : s \mapsto \mathrm{e}^{\theta s}$ is now concave, so the mean value trace inequality, Lemma 3.4, produces a lower bound. The remaining steps are identical. □

REMARK 3.8 (Regularity conditions). To simplify the presentation, we have instated a boundedness assumption in Lemma 3.7. All the examples we discuss satisfy this requirement. When $\mathbf{X}$ is unbounded, Lemma 3.7 still holds provided that $\mathbf{X}$ meets an integrability condition.

**4. Exponential concentration for bounded random matrices.** We are now prepared to establish exponential concentration inequalities. Our first major result demonstrates that an almost-sure bound for the conditional variance yields exponential tail bounds for the extreme eigenvalues of a random Hermitian matrix. We can also obtain estimates for the expectation of the extreme eigenvalues.

THEOREM 4.1 (Concentration for bounded random matrices). *Consider a matrix Stein pair* $(\mathbf{X}, \mathbf{X}') \in \mathbb{H}^d \times \mathbb{H}^d$. *Suppose there exist nonnegative constants* $c, v$ *for which the conditional variance* (2.4) *of the pair satisfies*

$$\boldsymbol{\Delta}_{\mathbf{X}} \preccurlyeq c\mathbf{X} + v\mathbf{I} \qquad \text{almost surely}. \tag{4.1}$$

*Then, for all* $t \ge 0$,

$$\begin{aligned}
\mathbb{P}\{\lambda_{\min}(\mathbf{X}) \le -t\} &\le d \cdot \exp\left\{\frac{-t^2}{2v}\right\}, \\
\mathbb{P}\{\lambda_{\max}(\mathbf{X}) \ge t\} &\le d \cdot \exp\left\{-\frac{t}{c} + \frac{v}{c^2}\log\left(1 + \frac{ct}{v}\right)\right\} \\
&\le d \cdot \exp\left\{\frac{-t^2}{2v + 2ct}\right\}.
\end{aligned}$$

*Furthermore*,

$$\begin{aligned}
\mathbb{E}\,\lambda_{\min}(\mathbf{X}) &\ge -\sqrt{2v\log d}, \\
\mathbb{E}\,\lambda_{\max}(\mathbf{X}) &\le \sqrt{2v\log d} + c\log d.
\end{aligned}$$

This result may be viewed as a matrix analogue of Chatterjee's concentration inequality for scalar random variables [6], Theorem 1.5(ii). The proof of Theorem 4.1 appears below in Section 4.2. Before we present the argument, let us explain how the result provides a short proof of a Hoeffding-type inequality for matrices.

4.1. *Application*: *Matrix Hoeffding inequality.* Theorem 4.1 yields an extension of Hoeffding's inequality [19] that holds for an independent sum of bounded random matrices.

COROLLARY 4.2 (Matrix Hoeffding). *Consider a finite sequence* $(\mathbf{Y}_k)_{k\geq 1}$ *of independent random matrices in* $\mathbb{H}^d$ *and a finite sequence* $(\mathbf{A}_k)_{k\geq 1}$ *of deterministic matrices in* $\mathbb{H}^d$. *Assume that*

$$\mathbb{E}\,\mathbf{Y}_k = \mathbf{0} \quad \text{and} \quad \mathbf{Y}_k^2 \preccurlyeq \mathbf{A}_k^2 \qquad \text{almost surely for each index } k.$$

*Then, for all* $t \geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \mathbf{Y}_k\right) \geq t\right\} \leq d \cdot \mathrm{e}^{-t^2/2\sigma^2} \qquad \text{for } \sigma^2 := \frac{1}{2}\left\|\sum_k (\mathbf{A}_k^2 + \mathbb{E}\,\mathbf{Y}_k^2)\right\|.$$

*Furthermore*,

$$\mathbb{E}\,\lambda_{\max}\left(\sum_k \mathbf{Y}_k\right) \leq \sigma\sqrt{2\log d}.$$

PROOF. Let $\mathbf{X} = \sum_k \mathbf{Y}_k$. Since $\mathbf{X}$ is a sum of centered, independent random matrices, we can use the matrix Stein pair constructed in Section 2.4. According to (2.6), the conditional variance satisfies

$$\boldsymbol{\Delta}_{\mathbf{X}} = \frac{1}{2}\sum_k (\mathbf{Y}_k^2 + \mathbb{E}\,\mathbf{Y}_k^2) \preccurlyeq \sigma^2\mathbf{I}$$

because $\mathbf{Y}_k^2 \preccurlyeq \mathbf{A}_k^2$. Invoke Theorem 4.1 with $c = 0$ and $v = \sigma^2$ to complete the bound. □

In the scalar setting $d = 1$, Corollary 4.2 reproduces an inequality of Chatterjee [6], Section 1.5, which itself is an improvement over the classical scalar Hoeffding bound. In turn, Corollary 4.2 improves upon the matrix Hoeffding inequality of [49], Theorem 1.3, in two ways. First, we have improved the constant in the exponent to its optimal value $1/2$. Second, we have decreased the size of the variance measure because $\sigma^2 \leq \|\sum_k \mathbf{A}_k^2\|$. Finally, let us remark that a similar result holds under the weaker assumption that $\sum_k \mathbf{Y}_k^2 \preccurlyeq \mathbf{A}^2$ almost surely.

Corollary 4.2 admits a plethora of applications. For example, in theoretical computer science, Widgerson and Xiao employ a suboptimal matrix Hoeffding inequality [50], Theorem 2.6, to derive efficient, derandomized algorithms for homomorphism testing and semidefinite covering problems. Under the improvements of Corollary 4.2, their results improve accordingly.

4.2. *Proof of Theorem 4.1: Exponential concentration.* Suppose that $(\mathbf{X}, \mathbf{X}')$ is a matrix Stein pair constructed from an auxiliary exchangeable pair $(Z, Z')$. Our aim is to bound the normalized trace m.g.f.

$$m(\theta) := \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta \mathbf{X}} \qquad \text{for } \theta \in \mathbb{R}. \tag{4.2}$$

The basic strategy is to develop a differential inequality, which we integrate to control $m(\theta)$ itself. Once these estimates are in place, the matrix Laplace transform method, Proposition 3.3, furnishes probability inequalities for the extreme eigenvalues of $\mathbf{X}$.

The following result summarizes our bounds for the trace m.g.f. $m(\theta)$.

LEMMA 4.3 (Trace m.g.f. estimates for bounded random matrices). *Let $(\mathbf{X}, \mathbf{X}')$ be a matrix Stein pair, and suppose there exist nonnegative constants $c, v$ for which*

$$\boldsymbol{\Delta}_{\mathbf{X}} \preccurlyeq c\mathbf{X} + v\mathbf{I} \qquad \textit{almost surely}. \tag{4.3}$$

*Then the normalized trace m.g.f. $m(\theta) := \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta \mathbf{X}}$ satisfies the bounds*

$$\log m(\theta) \le \frac{v\theta^2}{2} \qquad \textit{when } \theta \le 0, \tag{4.4}$$

$$\log m(\theta) \le \frac{v}{c^2}\left[\log\left(\frac{1}{1-c\theta}\right) - c\theta\right] \tag{4.5}$$

$$\le \frac{v\theta^2}{2(1-c\theta)} \qquad \textit{when } 0 \le \theta < 1/c. \tag{4.6}$$

We establish Lemma 4.3 in Section 4.2.1 et seq. In Section 4.2.4, we finish the proof of Theorem 4.1 by combining these bounds with the matrix Laplace transform method.

4.2.1. *Boundedness of the random matrix.* First, we confirm that the random matrix $\mathbf{X}$ is almost surely bounded under hypothesis (4.3) on the conditional variance $\boldsymbol{\Delta}_{\mathbf{X}}$. Recall definition (2.4) of the conditional variance, and compute that

$$\boldsymbol{\Delta}_{\mathbf{X}} = \frac{1}{2\alpha}\,\mathbb{E}\big[(\mathbf{X}-\mathbf{X}')^2 \mid Z\big] \succcurlyeq \frac{1}{2\alpha}\big(\mathbb{E}[\mathbf{X}-\mathbf{X}' \mid Z]\big)^2 = \frac{1}{2\alpha}(\alpha\mathbf{X})^2 = \frac{\alpha}{2}\mathbf{X}^2.$$

The semidefinite bound is the operator Jensen inequality (1.2), applied conditionally. The third relation follows from definition (2.1) of a matrix Stein pair. Owing to assumption (4.3), we reach the quadratic inequality $\frac{1}{2}\alpha\mathbf{X}^2 \preccurlyeq c\mathbf{X} + v\mathbf{I}$. The scale factor $\alpha$ is positive, so we may conclude that the eigenvalues of $\mathbf{X}$ are almost surely restricted to a bounded interval.

4.2.2. *Differential inequalities for the trace m.g.f.* Since the matrix $\mathbf{X}$ is almost surely bounded, the derivative of the trace m.g.f. has the form

$$m'(\theta) = \mathbb{E}\,\bar{\mathrm{tr}}\big[\mathbf{X}\mathrm{e}^{\theta\mathbf{X}}\big] \qquad \text{for } \theta \in \mathbb{R}. \tag{4.7}$$

To control the derivative, we combine Lemma 3.7 with the assumed inequality (4.3) for the conditional variance. For $\theta \geq 0$, we obtain

$$\begin{aligned} m'(\theta) &\leq \theta \cdot \mathbb{E}\,\bar{\mathrm{tr}}\big[\boldsymbol{\Delta}_{\mathbf{X}}\mathrm{e}^{\theta\mathbf{X}}\big] \\ &\leq \theta \cdot \mathbb{E}\,\bar{\mathrm{tr}}\big[(c\mathbf{X} + v\mathbf{I})\mathrm{e}^{\theta\mathbf{X}}\big] \\ &= c\theta \cdot \mathbb{E}\,\bar{\mathrm{tr}}\big[\mathbf{X}\mathrm{e}^{\theta\mathbf{X}}\big] + v\theta \cdot \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta\mathbf{X}} \\ &= c\theta \cdot m'(\theta) + v\theta \cdot m(\theta). \end{aligned}$$

In the last line, we have identified the trace m.g.f. (4.2) and its derivative (4.7). The second relation holds because the matrix $\mathrm{e}^{\theta\mathbf{X}}$ is positive definite. Indeed, when $\mathbf{P}$ is psd, $\mathbf{A} \preccurlyeq \mathbf{B}$ implies that $\mathrm{tr}(\mathbf{AP}) \leq \mathrm{tr}(\mathbf{BP})$.

For $\theta \leq 0$, the same argument yields a lower bound

$$m'(\theta) \geq c\theta \cdot m'(\theta) + v\theta \cdot m(\theta).$$

Rearrange these inequalities to isolate the log-derivative $m'(\theta)/m(\theta)$ of the trace m.g.f. We reach

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log m(\theta) \leq \frac{v\theta}{1 - c\theta} \qquad \text{for } 0 \leq \theta < 1/c \quad \text{and} \tag{4.8}$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log m(\theta) \geq \frac{v\theta}{1 - c\theta} \qquad \text{for } \theta \leq 0. \tag{4.9}$$

4.2.3. *Solving the differential inequalities.* Observe that

$$\log m(0) = \log \bar{\mathrm{tr}}\,\mathrm{e}^{\mathbf{0}} = \log \bar{\mathrm{tr}}\,\mathbf{I} = \log 1 = 0. \tag{4.10}$$

Therefore, we may integrate the differential inequalities (4.8) and (4.9), starting at zero, to obtain bounds on $\log m(\theta)$ elsewhere.

First, assume that $0 \leq \theta < 1/c$. In view of (4.10), the fundamental theorem of calculus and the differential inequality (4.8) imply that

$$\log m(\theta) = \int_0^\theta \frac{\mathrm{d}}{\mathrm{d}s}\log m(s)\,\mathrm{d}s \leq \int_0^\theta \frac{vs}{1 - cs}\,\mathrm{d}s = -\frac{v}{c^2}\big(c\theta + \log(1 - c\theta)\big).$$

We can develop a weaker inequality by making a further approximation within the integral,

$$\log m(\theta) \leq \int_0^\theta \frac{vs}{1 - cs}\,\mathrm{d}s \leq \int_0^\theta \frac{vs}{1 - c\theta}\,\mathrm{d}s = \frac{v\theta^2}{2(1 - c\theta)}.$$

These inequalities are the trace m.g.f. estimates (4.5) and (4.6) appearing in Lemma 4.3.

Next, assume that $\theta \le 0$. In this case, the differential inequality (4.9) yields

$$-\log m(\theta) = \int_\theta^0 \frac{\mathrm{d}}{\mathrm{d}s} \log m(s)\,\mathrm{d}s \ge \int_\theta^0 \frac{vs}{1-cs}\,\mathrm{d}s \ge \int_\theta^0 vs\,\mathrm{d}s = -\frac{v\theta^2}{2}.$$

This calculation delivers the trace m.g.f. bound (4.4). The proof of Lemma 4.3 is complete.

4.2.4. *The matrix Laplace transform argument.* With Lemma 4.3 at hand, we quickly finish the proof of Theorem 4.1. First, let us establish probability inequalities for the maximum eigenvalue. The Laplace transform bound (3.1) and the trace m.g.f. estimate (4.5) together yield

$$\begin{aligned}\mathbb{P}\{\lambda_{\max}(\mathbf{X}) \ge t\} &\le \inf_{0<\theta<1/c} d \cdot \exp\left\{-\theta t - \frac{v}{c^2}(c\theta + \log(1-c\theta))\right\} \\ &\le d \cdot \exp\left\{-\frac{t}{c} + \frac{v}{c^2}\log\left(1 + \frac{ct}{v}\right)\right\}.\end{aligned}$$

The second relation follows when we choose $\theta = t/(v+ct)$. Similarly, the trace m.g.f. bound (4.6) delivers

$$\begin{aligned}\mathbb{P}\{\lambda_{\max}(\mathbf{X}) \ge t\} &\le \inf_{0<\theta<1/c} d \cdot \exp\left\{-\theta t + \frac{v\theta^2}{2(1-c\theta)}\right\} \\ &= d \cdot \exp\left\{-\frac{v}{2c^2}(1-\sqrt{1+2ct/v})^2\right\} \\ &\le d \cdot \exp\left\{-\frac{t^2}{2v+2ct}\right\},\end{aligned}$$

because the infimum occurs at $\theta = (1 - 1/\sqrt{1+2ct/v})/c$. The final inequality depends on the numerical fact

$$(1-\sqrt{1+2x})^2 \ge \frac{x^2}{1+x} \qquad \text{for all } x \ge 0.$$

To control the expectation of the maximum eigenvalue, we combine the Laplace transform bound (3.3) and the trace m.g.f. bound (4.6) to see that

$$\mathbb{E}\,\lambda_{\max}(\mathbf{X}) \le \inf_{0<\theta<1/c} \frac{1}{\theta}\left[\log d + \frac{v\theta^2}{2(1-c\theta)}\right] = \sqrt{2v\log d} + c\log d.$$

The second relation can be verified using a computer algebra system.

Next, we turn to results for the minimum eigenvalue. Combine the matrix Laplace transform bound (3.2) with the trace m.g.f. bound (4.4) to reach

$$\mathbb{P}\{\lambda_{\min}(\mathbf{X}) \le -t\} \le d \cdot \inf_{\theta<0} \exp\left\{\theta t + \frac{v\theta^2}{2}\right\} = d \cdot \mathrm{e}^{-t^2/2v}.$$

The infimum is attained at $\theta = -t/v$. To compute the expectation of the minimum eigenvalue, we apply the Laplace transform bound (3.4) and the trace m.g.f. bound (4.4), whence

$$\mathbb{E}\,\lambda_{\min}(\mathbf{X}) \geq \sup_{\theta<0} \frac{1}{\theta}\left[\log d + \frac{v\theta^2}{2}\right] = -\sqrt{2v\log d}.$$

The supremum is attained at $\theta = -\sqrt{2v^{-1}\log d}$.

**5. Refined exponential concentration for random matrices.** Although Theorem 4.1 is a strong result, the hypothesis $\mathbf{\Delta_X} \preccurlyeq c\mathbf{X} + v\mathbf{I}$ on the conditional variance is too stringent for many situations of interest. Our second major result shows that we can use the typical behavior of the conditional variance to obtain tail bounds for the maximum eigenvalue of a random Hermitian matrix.

THEOREM 5.1 (Refined concentration for random matrices). *Suppose that* $(\mathbf{X}, \mathbf{X}') \in \mathbb{H}^d \times \mathbb{H}^d$ *is a matrix Stein pair, and assume that* $\mathbf{X}$ *is almost surely bounded in norm. Define the function*

$$r(\psi) := \frac{1}{\psi}\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi\mathbf{\Delta_X}} \qquad \text{for each } \psi > 0, \tag{5.1}$$

*where* $\mathbf{\Delta_X}$ *is the conditional variance* (2.4). *Then, for all* $t \geq 0$ *and all* $\psi > 0$,

$$\mathbb{P}\{\lambda_{\max}(\mathbf{X}) \geq t\} \leq d \cdot \exp\left\{\frac{-t^2}{2r(\psi) + 2t/\sqrt{\psi}}\right\}. \tag{5.2}$$

*Furthermore, for all* $\psi > 0$,

$$\mathbb{E}\,\lambda_{\max}(\mathbf{X}) \leq \sqrt{2r(\psi)\log d} + \frac{\log d}{\sqrt{\psi}}. \tag{5.3}$$

This theorem is essentially a matrix version of a result from Chatterjee's thesis [7], Theorem 3.13. The proof of Theorem 5.1 is similar in spirit to the proof of Theorem 4.1, so we postpone the demonstration until Appendix A.

Let us offer some remarks to clarify the meaning of this result. Recall that $\mathbf{\Delta_X}$ is a stochastic approximation for the variance of the random matrix $\mathbf{X}$. We can interpret the function $r(\psi)$ as a measure of the typical magnitude of the conditional variance. Indeed, the matrix Laplace transform result, Proposition 3.3, ensures that

$$\mathbb{E}\,\lambda_{\max}(\mathbf{\Delta_X}) \leq \inf_{\psi>0}\left[r(\psi) + \frac{\log d}{\psi}\right].$$

The import of this inequality is that we can often identify a value of $\psi$ to make $r(\psi) \approx \mathbb{E}\,\lambda_{\max}(\mathbf{\Delta_X})$. Ideally, we also want to choose $r(\psi) \gg \psi^{-1/2}$ so that the term $r(\psi)$ drives the tail bound (5.2) when the parameter $t$ is small. In the next subsection, we show that these heuristics yield a matrix Bernstein inequality.

5.1. *Application*: *The matrix Bernstein inequality.* As an illustration of Theorem 5.1, we establish a tail bound for a sum of centered, independent random matrices that are subject to a uniform norm bound.

COROLLARY 5.2 (Matrix Bernstein). *Consider an independent sequence* $(\mathbf{Y}_k)_{k\geq 1}$ *of random matrices in* $\mathbb{H}^d$ *that satisfy*

$$\mathbb{E}\,\mathbf{Y}_k = \mathbf{0} \quad \text{and} \quad \|\mathbf{Y}_k\| \leq R \qquad \text{for each index } k.$$

*Then, for all* $t \geq 0$,

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

*Furthermore*,

$$\mathbb{E}\,\lambda_{\max}\left(\sum_k \mathbf{Y}_k\right) \leq \sigma\sqrt{3\log d} + R\log d.$$

Corollary 5.2 is directly comparable with other matrix Bernstein inequalities in the literature. The constants are slightly worse than [49], Theorem 1.4 and slightly better than [35], Theorem 1.2. The hypotheses in the current result are somewhat stricter than those in the prior works. Nevertheless, the proof provides a template for studying more complicated random matrices that involve dependent random variables.

PROOF OF COROLLARY 5.2. Consider the matrix Stein pair $(\mathbf{X}, \mathbf{X}')$ described in Section 2.4. Calculation (2.6) shows that the conditional variance of $\mathbf{X}$ satisfies

$$\boldsymbol{\Delta}_{\mathbf{X}} = \frac{1}{2}\sum_k (\mathbf{Y}_k^2 + \mathbb{E}\,\mathbf{Y}_k^2).$$

The function $r(\psi)$ measures the typical size of $\boldsymbol{\Delta}_{\mathbf{X}}$. To control $r(\psi)$, we center the conditional variance and reduce the expression as follows:

$$\begin{aligned} r(\psi) &:= \frac{1}{\psi}\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi\boldsymbol{\Delta}_{\mathbf{X}}} \leq \frac{1}{\psi}\log \mathbb{E}\,\bar{\mathrm{tr}}\exp\{\psi(\Delta_{\mathbf{X}} - \mathbb{E}\,\boldsymbol{\Delta}_{\mathbf{X}}) + \psi\|\mathbb{E}\,\boldsymbol{\Delta}_{\mathbf{X}}\|\cdot\mathbf{I}\} \\ &= \frac{1}{\psi}\log \mathbb{E}\,\bar{\mathrm{tr}}[\mathrm{e}^{\psi\sigma^2}\cdot\exp\{\psi(\boldsymbol{\Delta}_{\mathbf{X}} - \mathbb{E}\,\boldsymbol{\Delta}_{\mathbf{X}})\}] \qquad (5.4) \\ &= \sigma^2 + \frac{1}{\psi}\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi(\boldsymbol{\Delta}_{\mathbf{X}} - \mathbb{E}\,\boldsymbol{\Delta}_{\mathbf{X}})}. \end{aligned}$$

The inequality depends on the monotonicity of the trace exponential [38], Section 2. Afterward, we have applied the identity $\|\mathbb{E}\,\boldsymbol{\Delta}_{\mathbf{X}}\| = \|\mathbb{E}\,\mathbf{X}^2\| = \sigma^2$, which follows from (2.5) and the independence of the sequence $(\mathbf{Y}_k)_{k\geq 1}$.

Introduce the centered random matrix

$$\mathbf{W} := \boldsymbol{\Delta}_{\mathbf{X}} - \mathbb{E}\,\Delta_{\mathbf{X}} = \frac{1}{2}\sum_k (\mathbf{Y}_k^2 - \mathbb{E}\,\mathbf{Y}_k^2). \tag{5.5}$$

Observe that $\mathbf{W}$ consists of a sum of centered, independent random matrices, so we can study it using the matrix Stein pair discussed in Section 2.4. Adapt the conditional variance calculation (2.6) to obtain

$$\begin{aligned}\boldsymbol{\Delta}_{\mathbf{W}} &= \frac{1}{2}\cdot\frac{1}{4}\sum_k \big[(\mathbf{Y}_k^2 - \mathbb{E}\,\mathbf{Y}_k^2)^2 + \mathbb{E}(\mathbf{Y}_k^2 - \mathbb{E}\,\mathbf{Y}_k^2)^2\big] \\ &\preccurlyeq \frac{1}{8}\sum_k \big[2\mathbf{Y}_k^4 + 2(\mathbb{E}\,\mathbf{Y}_k^2)^2 + \mathbb{E}\,\mathbf{Y}_k^4 - (\mathbb{E}\,\mathbf{Y}_k^2)^2\big] \\ &\preccurlyeq \frac{1}{4}\sum_k (\mathbf{Y}_k^4 + \mathbb{E}\,\mathbf{Y}_k^4).\end{aligned}$$

To reach the second line, we apply the operator convexity (1.1) of the matrix square to the first parenthesis, and we compute the second expectation explicitly. The third line follows from the operator Jensen inequality (1.2). To continue, make the estimate $\mathbf{Y}_k^4 \preccurlyeq R^2\mathbf{Y}_k^2$ in both terms. Thus,

$$\boldsymbol{\Delta}_{\mathbf{W}} \preccurlyeq \frac{R^2}{4}\sum_{k=1}^n (\mathbf{Y}_k^2 + \mathbb{E}\,\mathbf{Y}_k^2) \preccurlyeq \frac{R^2}{2}\cdot\mathbf{W} + \frac{R^2\sigma^2}{2}\cdot\mathbf{I}.$$

The trace m.g.f. bound, Lemma 4.3, delivers

$$\log m_{\mathbf{W}}(\psi) = \log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi\mathbf{W}} \le \frac{R^2\sigma^2\psi^2}{4 - 2R^2\psi}. \tag{5.6}$$

To complete the proof, combine the bounds (5.4) and (5.6) to reach

$$r(\psi) \le \sigma^2 + \frac{R^2\sigma^2\psi}{4 - 2R^2\psi}.$$

In particular, it holds that $r(R^{-2}) \le 1.5\sigma^2$. The result now follows from Theorem 5.1. □

**6. Polynomial moments and the spectral norm of a random matrix.** We can also study the spectral norm of a random matrix by bounding its polynomial moments. To present these results, we must introduce the family of Schatten norms.

DEFINITION 6.1 (Schatten norm). For each $p \ge 1$, the Schatten $p$-norm is defined as

$$\|\mathbf{B}\|_p := \big(\mathrm{tr}\,|\mathbf{B}|^p\big)^{1/p} \qquad \text{for } \mathbf{B} \in \mathbb{M}^d.$$

In this setting, $|\mathbf{B}| := (\mathbf{B}^*\mathbf{B})^{1/2}$. Bhatia's book [2], Chapter IV, contains a detailed discussion of these norms and their properties.

The following proposition is a matrix analog of the Chebyshev bound from classical probability. As in the scalar case [27], Exercise 1, this bound is at least as tight as the analogous matrix Laplace transform bound (3.1).

PROPOSITION 6.2 (Matrix Chebyshev method). *Let* $\mathbf{X}$ *be a random matrix. For all* $t > 0$,

$$\mathbb{P}\{\|\mathbf{X}\| \geq t\} \leq \inf_{p \geq 1} t^{-p} \cdot \mathbb{E}\,\|\mathbf{X}\|_p^p. \tag{6.1}$$

*Furthermore*,

$$\mathbb{E}\,\|\mathbf{X}\| \leq \inf_{p \geq 1} \left(\mathbb{E}\,\|\mathbf{X}\|_p^p\right)^{1/p}. \tag{6.2}$$

PROOF. To prove (6.1), we use Markov's inequality. For $p \geq 1$,

$$\mathbb{P}\{\|\mathbf{X}\| \geq t\} \leq t^{-p} \cdot \mathbb{E}\,\|\mathbf{X}\|^p = t^{-p} \cdot \mathbb{E}\big\||\mathbf{X}|^p\big\| \leq t^{-p} \cdot \mathbb{E}\,\mathrm{tr}\,|\mathbf{X}|^p,$$

since the trace of a positive matrix dominates the maximum eigenvalue. To verify (6.2), select $p \geq 1$. Jensen's inequality implies that

$$\mathbb{E}\,\|\mathbf{X}\| \leq \left(\mathbb{E}\,\|\mathbf{X}\|^p\right)^{1/p} = \left(\mathbb{E}\big\||\mathbf{X}|^p\big\|\right)^{1/p} \leq \left(\mathbb{E}\,\mathrm{tr}\,|\mathbf{X}|^p\right)^{1/p}.$$

Identify the Schatten $p$-norm and take infima to complete the bounds. □

**7. Polynomial moment inequalities for random matrices.** Our last major result demonstrates that the polynomial moments of a random Hermitian matrix are controlled by the moments of the conditional variance. By combining this result with the matrix Chebyshev method, Proposition 6.2, we can obtain probability inequalities for the spectral norm of a random Hermitian matrix.

THEOREM 7.1 (Matrix BDG inequality). *Let* $p = 1$ *or* $p \geq 1.5$. *Suppose that* $(\mathbf{X}, \mathbf{X}')$ *is a matrix Stein pair where* $\mathbb{E}\,\|\mathbf{X}\|_{2p}^{2p} < \infty$. *Then*

$$\left(\mathbb{E}\,\|\mathbf{X}\|_{2p}^{2p}\right)^{1/(2p)} \leq \sqrt{2p-1} \cdot \left(\mathbb{E}\,\|\mathbf{\Delta}_{\mathbf{X}}\|_p^p\right)^{1/(2p)}.$$

*The conditional variance* $\mathbf{\Delta}_{\mathbf{X}}$ *is defined in* (2.4).

REMARK 7.2 (Missing values). Theorem 7.1 also holds when $1 < p < 1.5$. In this range, our bound for the constant is $\sqrt{4p-2}$. The proof requires a variant of the mean value trace inequality for a convex function $h$.

Theorem 7.1 extends a scalar result of Chatterjee [6], Theorem 1.5(iii), to the matrix setting. Chatterjee's bound can be viewed as an exchangeable pairs version of the Burkholder–Davis–Gundy (BDG) inequality from classical martingale theory [4]. Other matrix extensions of the BDG inequality appear in the work of Pisier–Xu [40] and the work of Junge–Xu [21, 22]. The proof of Theorem 7.1, which applies equally to infinite dimensional operators $\mathbf{X}$, appears below in Section 7.3.

7.1. *Application*: *Matrix Khintchine inequality.* First, we demonstrate that the matrix BDG inequality contains an improvement of the noncommutative Khintchine inequality [28, 29] in the matrix setting. This result has been a dominant tool in several application areas over the last few years, largely because of the articles [44, 45].

COROLLARY 7.3 (Matrix Khintchine). *Suppose that* $p = 1$ *or* $p \geq 1.5$. *Consider a finite sequence* $(\mathbf{Y}_k)_{k\geq 1}$ *of independent, random, Hermitian matrices and a deterministic sequence* $(\mathbf{A}_k)_{k\geq 1}$ *for which*

$$\mathbb{E}\,\mathbf{Y}_k = \mathbf{0} \quad \text{and} \quad \mathbf{Y}_k^2 \preccurlyeq \mathbf{A}_k^2 \qquad \text{almost surely for each index } k. \tag{7.1}$$

*Then*

$$\left(\mathbb{E}\left\|\sum_k \mathbf{Y}_k\right\|_{2p}^{2p}\right)^{1/(2p)} \leq \sqrt{p-0.5}\cdot\left\|\left(\sum_k (\mathbf{A}_k^2 + \mathbb{E}\,\mathbf{Y}_k^2)\right)^{1/2}\right\|_{2p}.$$

*In particular, when* $(\varepsilon_k)_{k\geq 1}$ *is an independent sequence of Rademacher random variables*,

$$\left(\mathbb{E}\left\|\sum_k \varepsilon_k \mathbf{A}_k\right\|_{2p}^{2p}\right)^{1/(2p)} \leq \sqrt{2p-1}\cdot\left\|\left(\sum_k \mathbf{A}_k^2\right)^{1/2}\right\|_{2p}. \tag{7.2}$$

PROOF. Consider the random matrix $\mathbf{X} = \sum_k \mathbf{Y}_k$. We use the matrix Stein pair constructed in Section 2.4. According to (2.6), the conditional variance $\boldsymbol{\Delta}_{\mathbf{X}}$ satisfies

$$\boldsymbol{\Delta}_{\mathbf{X}} = \frac{1}{2}\sum_k (\mathbf{Y}_k^2 + \mathbb{E}\,\mathbf{Y}_k^2) \preccurlyeq \frac{1}{2}\sum_k (\mathbf{A}_k^2 + \mathbb{E}\,\mathbf{Y}_k^2).$$

An application of Theorem 7.1 completes the argument. □

For each positive integer $p$, the optimal constant $\mathrm{C}_{2p}$ on the right-hand side of (7.2) satisfies

$$\mathrm{C}_{2p}^{2p} = (2p-1)!! = (2p)!/\big(2^p\, p!\big)$$

as shown by Buchholz [3], Theorem 5. Since $(2p-1)^p/(2p-1)!! < \mathrm{e}^{p-1/2}$ for each positive integer $p$, the constant in (7.2) lies within a factor $\sqrt{\mathrm{e}}$ of optimal.

Previous methods for establishing the matrix Khintchine inequality are rather involved, so it is remarkable that the simple argument based on exchangeable pairs leads to a result that is so accurate. The same argument even yields a result under the weaker assumption that $\sum_k \mathbf{Y}_k^2 \preccurlyeq \mathbf{A}^2$ almost surely.

7.2. *Application*: *Matrix Rosenthal inequality.* As a second example, we can develop a more sophisticated set of moment inequalities that are roughly the polynomial equivalent of the exponential moment bound underlying the matrix Bernstein inequality.

COROLLARY 7.4 (Matrix Rosenthal inequality). *Suppose that* $p = 1$ *or* $p \geq 1.5$. *Consider a finite sequence* $(\mathbf{P}_k)_{k\geq 1}$ *of independent, random psd matrices that satisfy* $\mathbb{E}\,\|\mathbf{P}_k\|_{2p}^{2p} < \infty$. *Then*

$$\begin{aligned}\left(\mathbb{E}\left\|\sum_k \mathbf{P}_k\right\|_{2p}^{2p}\right)^{1/(2p)} \\ \leq \left[\left\|\sum_k \mathbb{E}\,\mathbf{P}_k\right\|_{2p}^{1/2} + \sqrt{4p-2}\cdot\left(\sum_k \mathbb{E}\,\|\mathbf{P}_k\|_{2p}^{2p}\right)^{1/(4p)}\right]^2.\end{aligned} \tag{7.3}$$

*Now, consider a finite sequence* $(\mathbf{Y}_k)_{k\geq 1}$ *of centered, independent, random Hermitian matrices, and assume that* $\mathbb{E}\,\|\mathbf{Y}_k\|_{4p}^{4p} < \infty$. *Then*

$$\begin{aligned}\left(\mathbb{E}\left\|\sum_k \mathbf{Y}_k\right\|_{4p}^{4p}\right)^{1/(4p)} \\ \leq \sqrt{4p-1}\cdot\left\|\left(\sum_k \mathbb{E}\,\mathbf{Y}_k^2\right)^{1/2}\right\|_{4p} + (4p-1)\cdot\left(\sum_k \mathbb{E}\,\|\mathbf{Y}_k\|_{4p}^{4p}\right)^{1/(4p)}.\end{aligned} \tag{7.4}$$

Turn to Appendix B for the proof of Corollary 7.4. This result extends a moment inequality due to Nagaev and Pinelis [33], which refines the constants in Rosenthal's inequality [43], Lemma 1. See the historical discussion [39], Section 5, for details. An interesting application of Corollary 7.4 is to establish improved sample complexity bounds for masked sample covariance estimation [8] when the dimension of a covariance matrix exceeds the number of samples. As we were finishing this paper, we learned that Junge and Zheng have recently established a noncommutative moment inequality [23], Theorem 0.4, that is quite similar to Corollary 7.4.

7.3. *Proof of the matrix BDG inequality.* In many respects, the proof of the matrix BDG inequality is similar to the proof of the exponential concentration result, Theorem 4.1. Both are based on moment comparison arguments that ultimately depend on the method of exchangeable pairs and the mean value trace inequality.

Suppose that $(\mathbf{X}, \mathbf{X}')$ is a matrix Stein pair with scale factor $\alpha$. First, observe that the result for $p = 1$ already follows from (2.5). Therefore, we may assume that $p \geq 1.5$. Introduce notation for the quantity of interest,

$$E := \mathbb{E}\,\|\mathbf{X}\|_{2p}^{2p} = \mathbb{E}\,\mathrm{tr}\,|\mathbf{X}|^{2p}.$$

Rewrite the expression for $E$ by peeling off a copy of $|\mathbf{X}|$. This move yields

$$E = \mathbb{E}\,\mathrm{tr}\big[|\mathbf{X}| \cdot |\mathbf{X}|^{2p-1}\big] = \mathbb{E}\,\mathrm{tr}\big[\mathbf{X} \cdot \mathrm{sgn}\,(\mathbf{X}) \cdot |\mathbf{X}|^{2p-1}\big].$$

Apply the method of exchangeable pairs, Lemma 2.4, with $\mathbf{F}(\mathbf{X}) = \mathrm{sgn}\,(\mathbf{X}) \cdot |\mathbf{X}|^{2p-1}$ to reach

$$E = \frac{1}{2\alpha}\,\mathbb{E}\,\mathrm{tr}\big[(\mathbf{X} - \mathbf{X}') \cdot (\mathrm{sgn}\,(\mathbf{X}) \cdot |\mathbf{X}|^{2p-1} - \mathrm{sgn}(\mathbf{X}') \cdot |\mathbf{X}'|^{2p-1})\big].$$

To verify the regularity condition (2.2) in Lemma 2.4, compute that

$$\begin{aligned}
&\mathbb{E}\,\big\|(\mathbf{X} - \mathbf{X}') \cdot \mathrm{sgn}\,(\mathbf{X}) \cdot |\mathbf{X}|^{2p-1}\big\| \\
&\quad\leq \mathbb{E}\big(\|\mathbf{X}\|\,\|\mathbf{X}\|^{2p-1}\big) + \mathbb{E}\big(\|\mathbf{X}'\|\,\|\mathbf{X}\|^{2p-1}\big) \\
&\quad\leq 2\big(\mathbb{E}\,\|\mathbf{X}\|^{2p}\big)^{1/(2p)}\big(\mathbb{E}\,\|\mathbf{X}\|^{2p}\big)^{(2p-1)/2p} \\
&\quad= 2\,\mathbb{E}\,\|\mathbf{X}\|^{2p} < \infty.
\end{aligned}$$

We have used the fact that $\mathrm{sgn}\,(\mathbf{X})$ is a unitary matrix, the exchangeability of $(\mathbf{X}, \mathbf{X}')$, Hölder's inequality for expectation and the fact that the Schatten $2p$-norm dominates the spectral norm.

We intend to apply the mean value trace inequality to obtain an estimate for the quantity $E$. Consider the function $h : s \mapsto \mathrm{sgn}\,(s) \cdot |s|^{2p-1}$. Its derivative $h'(s) = (2p-1) \cdot |s|^{2p-2}$ is convex because $p \geq 1.5$. Lemma 3.4 delivers the bound

$$\begin{aligned}
E &\leq \frac{2p-1}{4\alpha}\,\mathbb{E}\,\mathrm{tr}\big[(\mathbf{X} - \mathbf{X}')^2 \cdot (|\mathbf{X}|^{2p-2} + |\mathbf{X}'|^{2p-2})\big] \\
&= \frac{2p-1}{2\alpha}\,\mathbb{E}\,\mathrm{tr}\big[(\mathbf{X} - \mathbf{X}')^2 \cdot |\mathbf{X}|^{2p-2}\big] \\
&= (2p-1) \cdot \mathbb{E}\,\mathrm{tr}\big[\boldsymbol{\Delta}_{\mathbf{X}} \cdot |\mathbf{X}|^{2p-2}\big].
\end{aligned}$$

The second line follows from the exchangeability of $\mathbf{X}$ and $\mathbf{X}'$. In the last line, we identify the conditional variance $\boldsymbol{\Delta}_{\mathbf{X}}$, defined in (2.4). As before, the moment bound $\mathbb{E}\,\|\mathbf{X}\|_{2p}^{2p} < \infty$ is strong enough to justify using the pull-through property in this step.

To continue, we must find a copy of $E$ within the latter expression. We can accomplish this goal using one of the basic results from the theory of Schatten norms [2], Corollary IV.2.6.

PROPOSITION 7.5 (Hölder inequality for trace). *Let $p$ and $q$ be Hölder conjugate indices, that is, positive numbers with the relationship $q = p/(p-1)$. Then*

$$\operatorname{tr}(\mathbf{BC}) \le \|\mathbf{B}\|_p \|\mathbf{C}\|_q \qquad \text{for all } \mathbf{B}, \mathbf{C} \in \mathbb{M}^d.$$

To complete the argument, apply the Hölder inequality for the trace followed by the Hölder inequality for the expectation. Thus

$$\begin{aligned} E &\le (2p-1) \cdot \mathbb{E}\big[\|\mathbf{\Delta_X}\|_p \cdot \big\||\mathbf{X}|^{2p-2}\big\|_{p/(p-1)}\big] \\ &= (2p-1) \cdot \mathbb{E}\big[\|\mathbf{\Delta_X}\|_p \cdot \|\mathbf{X}\|_{2p}^{2p-2}\big] \\ &\le (2p-1) \cdot \big(\mathbb{E}\,\|\mathbf{\Delta_X}\|_p^p\big)^{1/p} \cdot \big(\mathbb{E}\,\|\mathbf{X}\|_{2p}^{2p}\big)^{(p-1)/p} \\ &= (2p-1) \cdot \big(\mathbb{E}\,\|\mathbf{\Delta_X}\|_p^p\big)^{1/p} \cdot E^{(p-1)/p}. \end{aligned}$$

Solve this algebraic inequality for the positive number $E$ to conclude that

$$E \le (2p-1)^p \cdot \mathbb{E}\,\|\mathbf{\Delta_X}\|_p^p.$$

Extract the $(2p)$th root to establish the matrix BDG inequality.

**8. Extension to general complex matrices.** Although, at first sight, it may seem that our theory is limited to random Hermitian matrices, results for general random matrices follow as a formal corollary [42, 49]. The approach is based on a device from operator theory [37].

DEFINITION 8.1 (Hermitian dilation). Let $\mathbf{B}$ be a matrix in $\mathbb{C}^{d_1 \times d_2}$, and set $d = d_1 + d_2$. The *Hermitian dilation* of $\mathbf{B}$ is the matrix

$$\mathscr{D}(\mathbf{B}) := \begin{bmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{0} \end{bmatrix} \in \mathbb{H}^d.$$

The dilation has two valuable properties. First, it preserves spectral information,

$$\lambda_{\max}\big(\mathscr{D}(\mathbf{B})\big) = \big\|\mathscr{D}(\mathbf{B})\big\| = \|\mathbf{B}\|. \tag{8.1}$$

Second, the square of the dilation satisfies

$$\mathscr{D}(\mathbf{B})^2 = \begin{bmatrix} \mathbf{B}\mathbf{B}^* & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^*\mathbf{B} \end{bmatrix}. \tag{8.2}$$

We can study a random matrix—not necessarily Hermitian—by applying our matrix concentration inequalities to the Hermitian dilation of the random matrix. As an illustration, let us prove a Bernstein inequality for general random matrices.

COROLLARY 8.2 (Bernstein inequality for general matrices). *Consider a finite sequence $(\mathbf{Z}_k)_{k \ge 1}$ of independent random matrices in $\mathbb{C}^{d_1 \times d_2}$ that satisfy*

$$\mathbb{E}\,\mathbf{Z}_k = \mathbf{0} \quad \text{and} \quad \|\mathbf{Z}_k\| \le R \qquad \text{almost surely for each index } k.$$

*Define* $d := d_1 + d_2$, *and introduce the variance measure*

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

*Then, for all* $t \geq 0$,

$$\mathbb{P}\left\{ \left\| \sum_k \mathbf{Z}_k \right\| \geq t \right\} \leq d \cdot \exp\left\{ \frac{-t^2}{3\sigma^2 + 2Rt} \right\}. \tag{8.3}$$

*Furthermore*,

$$\mathbb{E}\left\| \sum_k \mathbf{Z}_k \right\| \leq \sigma \sqrt{3 \log d} + R \log d. \tag{8.4}$$

PROOF. Consider the random series $\sum_k \mathscr{D}(\mathbf{Z}_k)$. The summands are independent, random Hermitian matrices that satisfy

$$\mathbb{E}\, \mathscr{D}(\mathbf{Z}_k) = \mathbf{0} \quad \text{and} \quad \left\| \mathscr{D}(\mathbf{Z}_k) \right\| \leq R.$$

The second identity depends on the spectral property (8.1). Therefore, the matrix Bernstein inequality, Corollary 5.2, applies. To state the outcome, we first note that $\lambda_{\max}(\sum_k \mathscr{D}(\mathbf{Z}_k)) = \|\sum_k \mathbf{Z}_k\|$, again because of the spectral property (8.1). Next, use the formula (8.2) to compute that

$$\left\| \sum_k \mathbb{E}[\mathscr{D}(\mathbf{Z}_k)^2] \right\| = \left\| \begin{bmatrix} \sum_k \mathbb{E}(\mathbf{Z}_k \mathbf{Z}_k^*) & \mathbf{0} \\ \mathbf{0} & \sum_k \mathbb{E}(\mathbf{Z}_k^* \mathbf{Z}_k) \end{bmatrix} \right\| = \sigma^2.$$

This observation completes the proof. □

Corollary 8.2 has important implications for the problem of estimating a matrix from noisy measurements. Indeed, bound (8.4) leads to a sample complexity analysis for matrix completion [13]. Moreover, a variety of authors have used tail bounds of the form (8.3) to control the error of convex optimization methods for matrix estimation [16, 30, 34, 42].

**9. A sum of conditionally independent, zero-mean matrices.** A chief advantage of the method of exchangeable pairs is its ability to handle random matrices constructed from *dependent* random variables. In this section, we briefly describe a way to relax the independence requirement when studying a sum of random matrices. In Sections 10 and 11, we develop more elaborate examples.

9.1. *Formulation.* Let us consider a finite sequence $(\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ of random Hermitian matrices that are conditionally independent given an auxiliary random element $Z$. Suppose moreover that

$$\mathbb{E}[\mathbf{Y}_k \mid Z] = \mathbf{0} \qquad \text{almost surely for each index } k. \tag{9.1}$$

We are interested in the sum of these conditionally independent, zero-mean random matrices

$$\mathbf{X} := \mathbf{Y}_1 + \cdots + \mathbf{Y}_n. \tag{9.2}$$

This type of series includes many examples that arise in practice.

EXAMPLE 9.1 (Rademacher series with random matrix coefficients). Consider a finite sequence $(\mathbf{W}_k)_{k\geq 1}$ of random Hermitian matrices. Suppose the sequence $(\varepsilon_k)_{k\geq 1}$ consists of independent Rademacher random variables that are independent from the random matrices. Consider the random series

$$\sum_k \varepsilon_k \mathbf{W}_k.$$

The summands may be strongly dependent on each other, but the independence of the Rademacher variables ensures that the summands are conditionally independent and of zero mean (9.1) given $Z := (\mathbf{W}_k)_{k\geq 1}$.

9.2. *A matrix Stein pair.* Let us describe how to build a matrix Stein pair $(\mathbf{X}, \mathbf{X}')$ for the sum (9.2) of conditionally independent, zero-mean random matrices. The approach is similar to the case of an independent sum, which appears in Section 2.4. For each $k$, we draw a random matrix $\mathbf{Y}_k'$ so that $\mathbf{Y}_k'$ and $\mathbf{Y}_k$ are conditionally i.i.d. given $(\mathbf{Y}_j)_{j\neq k}$. Then, independently, we draw an index $K$ uniformly at random from $\{1, \dots, n\}$. As in Section 2.4, the random matrix

$$\mathbf{X}' := \mathbf{Y}_1 + \cdots + \mathbf{Y}_{K-1} + \mathbf{Y}_K' + \mathbf{Y}_{K+1} + \cdots + \mathbf{Y}_n$$

is an exchangeable counterpart to $\mathbf{X}$. The conditional independence and conditional zero-mean (9.1) assumptions imply that, almost surely,

$$\mathbb{E}\big[\mathbf{Y}_k' \mid (\mathbf{Y}_j)_{j\neq k}\big] = \mathbb{E}\big[\mathbf{Y}_k \mid (\mathbf{Y}_j)_{j\neq k}\big] = \mathbb{E}\big[\mathbb{E}[\mathbf{Y}_k \mid Z] \mid (\mathbf{Y}_j)_{j\neq k}\big] = \mathbf{0}.$$

Hence,

$$\begin{aligned}\mathbb{E}\big[\mathbf{X} - \mathbf{X}' \mid (\mathbf{Y}_j)_{j\geq 1}\big] &= \mathbb{E}\big[\mathbf{Y}_K - \mathbf{Y}_K' \mid (\mathbf{Y}_j)_{j\geq 1}\big] \\ &= \frac{1}{n}\sum_{k=1}^n \big(\mathbf{Y}_k - \mathbb{E}\big[\mathbf{Y}_k' \mid (\mathbf{Y}_j)_{j\neq k}\big]\big) = \frac{1}{n}\sum_{k=1}^n \mathbf{Y}_k = \frac{1}{n}\mathbf{X}.\end{aligned}$$

Therefore, $(\mathbf{X}, \mathbf{X}')$ is a matrix Stein pair with scale factor $\alpha = n^{-1}$.

We can determine the conditional variance after a short argument that parallels computation (2.6) in the independent setting,

$$\begin{aligned}\boldsymbol{\Delta}_{\mathbf{X}} &= \frac{n}{2} \cdot \mathbb{E}\big[(\mathbf{Y}_K - \mathbf{Y}'_K)^2 \mid (\mathbf{Y}_j)_{j\geq 1}\big] \\ &= \frac{1}{2}\sum_{k=1}^{n}(\mathbf{Y}_k^2 + \mathbb{E}[\mathbf{Y}_k^2 \mid (\mathbf{Y}_j)_{j\neq k}]).\end{aligned} \tag{9.3}$$

Expression (9.3) shows that, even in the presence of some dependence, we can control the size of the conditional expectation uniformly if we control the size of the individual summands.

Using the Stein pair $(\mathbf{X}, \mathbf{X}')$ and expression (9.3), we may develop a variety of concentration inequalities for conditionally independent, zero-mean sums that are analogous to our results for independent sums. We omit detailed examples.

**10. Combinatorial sums of matrices.** The method of exchangeable pairs can also be applied to many types of highly symmetric distributions. In this section, we study a class of *combinatorial matrix statistics*, which generalize the scalar statistics studied by Hoeffding [18].

10.1. *Formulation.* Consider a deterministic array $(\mathbf{A}_{jk})_{j,k=1}^n$ of Hermitian matrices, and let $\pi$ be a uniformly random permutation on $\{1, \dots, n\}$. Define the random matrix

$$\mathbf{Y} := \sum_{j=1}^{n} \mathbf{A}_{j\pi(j)} \qquad \text{whose mean } \mathbb{E}\,\mathbf{Y} = \frac{1}{n}\sum_{j,k=1}^{n} \mathbf{A}_{jk}. \tag{10.1}$$

The combinatorial sum $\mathbf{Y}$ is a natural candidate for an exchangeable pair analysis. Before we describe how to construct a matrix Stein pair, let us mention a few problems that lead to a random matrix of the form $\mathbf{Y}$.

EXAMPLE 10.1 (Sampling without replacement). Consider a finite collection $\mathscr{B} := \{\mathbf{B}_1, \dots, \mathbf{B}_n\}$ of deterministic Hermitian matrices. Suppose that we want to study a sum of $s$ matrices sampled randomly from $\mathscr{B}$ without replacement. We can express this type of series in the form

$$\mathbf{W} := \sum_{j=1}^{s} \mathbf{B}_{\pi(j)},$$

where $\pi$ is a random permutation on $\{1, \dots, n\}$. The matrix $\mathbf{W}$ is therefore an example of a combinatorial sum.

EXAMPLE 10.2 (A randomized "inner product"). Consider two fixed sequences of complex matrices

$$\mathbf{B}_1, \dots, \mathbf{B}_n \in \mathbb{C}^{d_1 \times s} \quad \text{and} \quad \mathbf{C}_1, \dots, \mathbf{C}_n \in \mathbb{C}^{s \times d_2}.$$

We may form a permuted matrix "inner product" by arranging one sequence in random order, multiplying the elements of the two sequences together, and summing the terms. That is, we are interested in the random matrix

$$\mathbf{Z} := \sum_{j=1}^{n} \mathbf{B}_j \mathbf{C}_{\pi(j)}.$$

This random matrix $\mathscr{D}(\mathbf{Z})$ is a combinatorial sum of Hermitian matrices.

10.2. *A matrix Stein pair.* To study the combinatorial sum (10.1) of matrices using the method of exchangeable pairs, we first introduce the zero-mean random matrix

$$\mathbf{X} := \mathbf{Y} - \mathbb{E}\,\mathbf{Y}.$$

To construct a matrix Stein pair $(\mathbf{X}, \mathbf{X}')$, we draw a pair $(J, K)$ of indices independently of $\pi$ and uniformly at random from $\{1, \dots, n\}^2$. Define a second random permutation $\pi' := \pi \circ (J, K)$ by composing $\pi$ with the transposition of the random indices $J$ and $K$. The pair $(\pi, \pi')$ is exchangeable, so

$$\mathbf{X}' := \sum_{j=1}^{n} \mathbf{A}_{j\pi'(j)} - \mathbb{E}\,\mathbf{Y}$$

is an exchangeable counterpart to $\mathbf{X}$.

To verify that $(\mathbf{X}, \mathbf{X}')$ is a matrix Stein pair, we calculate that

$$\begin{aligned}
\mathbb{E}\big[\mathbf{X} - \mathbf{X}' \mid \pi\big] &= \mathbb{E}[\mathbf{A}_{J\pi(J)} + \mathbf{A}_{K\pi(K)} - \mathbf{A}_{J\pi(K)} - \mathbf{A}_{K\pi(J)} \mid \pi] \\
&= \frac{1}{n^2} \sum_{j,k=1}^{n} [\mathbf{A}_{j\pi(j)} + \mathbf{A}_{k\pi(k)} - \mathbf{A}_{j\pi(k)} - \mathbf{A}_{k\pi(j)}] \\
&= \frac{2}{n}(\mathbf{Y} - \mathbb{E}\,\mathbf{Y}) = \frac{2}{n}\mathbf{X}.
\end{aligned}$$

The first identity holds because the sums $\mathbf{X}$ and $\mathbf{X}'$ differ for only four choices of indices. Thus $(\mathbf{X}, \mathbf{X}')$ is a Stein pair with scale factor $\alpha = 2/n$.

Turning to the conditional variance, we find that

$$\begin{aligned}
\boldsymbol{\Delta}_{\mathbf{X}}(\pi) &= \frac{n}{4}\,\mathbb{E}\big[(\mathbf{X} - \mathbf{X}')^2 \mid \pi\big] \\
&= \frac{1}{4n} \sum_{j,k=1}^{n} [\mathbf{A}_{j\pi(j)} + \mathbf{A}_{k\pi(k)} - \mathbf{A}_{j\pi(k)} - \mathbf{A}_{k\pi(j)}]^2.
\end{aligned} \tag{10.2}$$

The structure of the conditional variance differs from previous examples, but we recognize that $\boldsymbol{\Delta}_{\mathbf{X}}$ is controlled when the matrices $\mathbf{A}_{jk}$ are bounded.

10.3. *Exponential concentration for a combinatorial sum.* We can apply our matrix concentration results to study the behavior of a combinatorial sum of matrices. As an example, let us present a Bernstein-type inequality. The argument is similar to the proof of Corollary 5.2, so we leave the details to Appendix C.

COROLLARY 10.3 (Bernstein inequality for a combinatorial matrix sum). *Consider an array* $(\mathbf{A}_{jk})_{j,k=1}^{n}$ *of deterministic matrices in* $\mathbb{H}^d$ *that satisfy*

$$\sum_{j,k=1}^{n} \mathbf{A}_{jk} = \mathbf{0} \quad \text{and} \quad \|\mathbf{A}_{jk}\| \le R \qquad \text{for each pair } (j,k) \text{ of indices.}$$

*Define the random matrix* $\mathbf{X} := \sum_{j=1}^{n} \mathbf{A}_{j\pi(j)}$, *where* $\pi$ *is a uniformly random permutation on* $\{1,\dots,n\}$. *Then, for all* $t \ge 0$,

$$\mathbb{P}\{\lambda_{\max}(\mathbf{X}) \ge t\} \le d \cdot \exp\left\{\frac{-t^2}{12\sigma^2 + 4\sqrt{2}Rt}\right\} \qquad \text{for } \sigma^2 := \frac{1}{n}\left\|\sum_{j,k=1}^{n} \mathbf{A}_{jk}^2\right\|.$$

*Furthermore*,

$$\mathbb{E}\,\lambda_{\max}(\mathbf{X}) \le \sigma\sqrt{12\log d} + 2\sqrt{2}R\log d.$$

**11. Self-reproducing matrix functions.** The method of exchangeable pairs can also be used to analyze nonlinear matrix-valued functions of random variables. In this section, we explain how to analyze matrix functions that satisfy a *self-reproducing property*.

11.1. *Example*: *Matrix second-order Rademacher chaos.* We begin with an example that shows how the self-reproducing property might arise. Consider a quadratic form that takes on random matrix values

$$\mathbf{H}(\boldsymbol{\varepsilon}) := \sum_{k} \sum_{j<k} \varepsilon_j \varepsilon_k \mathbf{A}_{jk}. \tag{11.1}$$

In this expression, $\boldsymbol{\varepsilon}$ is a finite vector of independent Rademacher random variables. The array $(\mathbf{A}_{jk})_{j,k\ge 1}$ consists of deterministic Hermitian matrices, and we assume that $\mathbf{A}_{jk} = \mathbf{A}_{kj}$.

Observe that the summands in $\mathbf{H}(\boldsymbol{\varepsilon})$ are dependent, and they do not satisfy the conditional zero-mean property (9.1) in general. Nevertheless, $\mathbf{H}(\boldsymbol{\varepsilon})$ does satisfy a fruitful self-reproducing property

$$\begin{aligned}\sum_{k}\big(\mathbf{H}(\boldsymbol{\varepsilon}) - \mathbb{E}[\mathbf{H}(\boldsymbol{\varepsilon}) \mid (\varepsilon_j)_{j\ne k}]\big) &= \sum_{k}\sum_{j\ne k} \varepsilon_j\big(\varepsilon_k - \mathbb{E}[\varepsilon_k]\big)\mathbf{A}_{jk} \\ &= \sum_{k}\sum_{j\ne k} \varepsilon_j\varepsilon_k\mathbf{A}_{jk} = 2\mathbf{H}(\boldsymbol{\varepsilon}).\end{aligned}$$

We have applied the pull-through property of conditional expectation, the assumption that the Rademacher variables are independent and the fact that $\mathbf{A}_{jk} = \mathbf{A}_{kj}$. As we will see, this type of self-reproducing condition can be used to construct a matrix Stein pair.

A random matrix of the form (11.1) is called a *second-order Rademacher chaos*. This class of random matrices arises in a variety of situations, including randomized linear algebra [11], compressed sensing [41], Section 9, and chance-constrained optimization [9]. Indeed, concentration inequalities for the matrix-valued Rademacher chaos have many potential applications.

11.2. *Formulation and matrix Stein pair.* In this section, we describe a more general version of the self-reproducing property. Suppose that $\mathbf{z} := (Z_1, \dots, Z_n)$ is a random vector taking values in a Polish space $\mathcal{Z}$. First, we construct an exchangeable counterpart

$$\mathbf{z}' := (Z_1, \dots, Z_{K-1}, Z'_K, Z_{K+1}, \dots, Z_n), \tag{11.2}$$

where $Z_k$ and $Z'_k$ are conditionally i.i.d. given $(Z_j)_{j \neq k}$, and $K$ is an independent coordinate drawn uniformly at random from $\{1, \dots, n\}$.

Next, let $\mathbf{H} : \mathcal{Z} \to \mathbb{H}^d$ be a bounded measurable function. Assume that $\mathbf{H}(\mathbf{z})$ satisfies an abstract *self-reproducing property*: for a parameter $s > 0$,

$$\sum_{k=1}^{n} \big(\mathbf{H}(\mathbf{z}) - \mathbb{E}\big[\mathbf{H}(\mathbf{z}) \mid (Z_j)_{j \neq k}\big]\big) = s \cdot \big(\mathbf{H}(\mathbf{z}) - \mathbb{E}\,\mathbf{H}(\mathbf{z})\big) \qquad \text{almost surely.}$$

Under this assumption, we can easily check that the random matrices

$$\mathbf{X} := \mathbf{H}(\mathbf{z}) - \mathbb{E}\,\mathbf{H}(\mathbf{z}) \quad \text{and} \quad \mathbf{X}' := \mathbf{H}(\mathbf{z}') - \mathbb{E}\,\mathbf{H}(\mathbf{z})$$

form a matrix Stein pair. Indeed,

$$\mathbb{E}\big[\mathbf{X} - \mathbf{X}' \mid \mathbf{z}\big] = \mathbb{E}\big[\mathbf{H}(\mathbf{z}) - \mathbf{H}(\mathbf{z}') \mid \mathbf{z}\big] = \frac{s}{n}\big(\mathbf{H}(\mathbf{z}) - \mathbb{E}\,\mathbf{H}(\mathbf{z})\big) = \frac{s}{n}\mathbf{X}.$$

We see that $(\mathbf{X}, \mathbf{X}')$ is a matrix Stein pair with scaling factor $\alpha = s/n$.

Finally, we compute the conditional variance

$$\begin{aligned} \boldsymbol{\Delta}_{\mathbf{X}}(\mathbf{z}) &= \frac{n}{2s}\,\mathbb{E}\big[\big(\mathbf{H}(\mathbf{z}) - \mathbf{H}(\mathbf{z}')\big)^2 \mid \mathbf{z}\big] \\ &= \frac{1}{2s}\sum_{k=1}^{n} \mathbb{E}\big[\big(\mathbf{H}(\mathbf{z}) - \mathbf{H}(Z_1, \dots, Z'_k, \dots, Z_n)\big)^2 \mid \mathbf{z}\big]. \end{aligned} \tag{11.3}$$

We discover that the conditional variance is small when $\mathbf{H}$ has controlled coordinate differences. In this case, the method of exchangeable pairs provides good concentration inequalities for the random matrix $\mathbf{X}$.

11.3. *Matrix bounded differences inequality.* As an example, we can develop a bounded differences inequality for random matrices by appealing to Theorem 4.1.

COROLLARY 11.1 (Matrix bounded differences). *Let* $\mathbf{z} := (Z_1, \dots, Z_n)$ *be a random vector taking values in a Polish space* $\mathcal{Z}$, *and, for each index* $k$, *let* $Z_k'$ *and* $Z_k$ *be conditionally i.i.d. given* $(Z_j)_{j \neq k}$. *Suppose that* $\mathbf{H} : \mathcal{Z} \to \mathbb{H}^d$ *is a function that satisfies the self-reproducing property*

$$\sum_{k=1}^{n} \big(\mathbf{H}(\mathbf{z}) - \mathbb{E}\big[\mathbf{H}(\mathbf{z}) \mid (Z_j)_{j \neq k}\big]\big) = s \cdot \big(\mathbf{H}(\mathbf{z}) - \mathbb{E}\,\mathbf{H}(\mathbf{z})\big) \qquad \textit{almost surely}$$

*for a parameter* $s > 0$ *as well as the bounded differences condition*

$$\mathbb{E}\big[\big(\mathbf{H}(\mathbf{z}) - \mathbf{H}(Z_1, \dots, Z_k', \dots, Z_n)\big)^2 \mid \mathbf{z}\big] \preccurlyeq \mathbf{A}_k^2 \qquad \textit{for each index } k \tag{11.4}$$

*almost surely, where* $\mathbf{A}_k$ *is a deterministic matrix in* $\mathbb{H}^d$. *Then, for all* $t \geq 0$,

$$\mathbb{P}\big\{\lambda_{\max}\big(\mathbf{H}(\mathbf{z}) - \mathbb{E}\,\mathbf{H}(\mathbf{z})\big) \geq t\big\} \leq d \cdot \mathrm{e}^{-st^2/L} \qquad \textit{for } L := \Big\| \sum_{k=1}^{n} \mathbf{A}_k^2 \Big\|.$$

*Furthermore*,

$$\mathbb{E}\,\lambda_{\max}\big(\mathbf{H}(\mathbf{z}) - \mathbb{E}\,\mathbf{H}(\mathbf{z})\big) \leq \sqrt{\frac{L \log d}{s}}.$$

In the scalar setting, Corollary 11.1 reduces to a version of McDiarmid's bounded difference inequality [31]. The result also complements the matrix bounded difference inequality of [49], Corollary 7.5, which requires independent input variables but makes no self-reproducing assumption.

PROOF OF COROLLARY 11.1. Since $\mathbf{H}(\mathbf{z})$ is self-reproducing, we may construct a matrix Stein pair $(\mathbf{X}, \mathbf{X}')$ with scale factor $\alpha = s/n$ as in Section 11. According to (11.3), the conditional variance of the pair satisfies

$$\begin{aligned} \boldsymbol{\Delta}_{\mathbf{X}} &= \frac{1}{2s} \sum_{k=1}^{n} \mathbb{E}\big[\big(\mathbf{H}(\mathbf{z}) - \mathbf{H}(Z_1, \dots, Z_k', \dots, Z_n)\big)^2 \mid \mathbf{z}\big] \\ &\preccurlyeq \frac{1}{2s} \sum_{k=1}^{n} \mathbf{A}_k^2 \preccurlyeq \frac{L}{2s} \cdot \mathbf{I}. \end{aligned}$$

We have used the bounded differences condition (11.4) and the definition of the bound $L$. To complete the proof, we apply the concentration result, Theorem 4.1, with the parameters $c = 0$ and $v = L/2s$. $\square$

## APPENDIX A: PROOF OF THEOREM 5.1

The proof of the refined exponential concentration bound, Theorem 5.1, parallels the argument in Theorem 4.1, but it differs at an important point. In the earlier result, we used an almost sure bound on the conditional variance to control the derivative of the trace m.g.f. This time, we use entropy inequalities to introduce finer information about the behavior of the conditional variance. The proof is essentially a matrix version of Chatterjee's argument [7], Theorem 3.13.

Our main object is to bound the trace m.g.f. of $\mathbf{X}$ in terms of the trace m.g.f. of the conditional variance. The next result summarizes our bounds.

LEMMA A.1 (Refined trace m.g.f. estimates). *Let $(\mathbf{X}, \mathbf{X}')$ be a matrix Stein pair, and assume that $\mathbf{X}$ is almost surely bounded in norm. Then the normalized trace m.g.f. $m(\theta) := \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta \mathbf{X}}$ satisfies the bounds*

$$\begin{aligned} \log m(\theta) &\le \frac{1}{2}\log\left(\frac{1}{1-\theta^2/\psi}\right)\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi \boldsymbol{\Delta}_{\mathbf{X}}} \\ &\le \frac{\theta^2/\psi}{2(1-\theta^2/\psi)}\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi \boldsymbol{\Delta}_{\mathbf{X}}} \qquad \textit{for } \psi > 0 \textit{ and } 0 \le \theta < \sqrt{\psi}. \end{aligned} \tag{A.1}$$

We establish Lemma A.1 in Section A.1 et seq. Afterward, in Section A.5, we invoke the matrix Laplace transform bound to complete the proof of Theorem 5.1.

**A.1. The derivative of the trace m.g.f.** The first steps of the argument are the same as in the proof of Theorem 4.1. Since $\mathbf{X}$ is almost surely bounded, we need not worry about regularity conditions. The derivative of the trace m.g.f. satisfies

$$m'(\theta) = \mathbb{E}\,\mathrm{tr}\big[\mathbf{X}\mathrm{e}^{\theta \mathbf{X}}\big] \qquad \text{for } \theta \in \mathbb{R}. \tag{A.2}$$

Lemma 3.7 provides a bound for the derivative in terms of the conditional variance,

$$m'(\theta) \le \theta \cdot \mathbb{E}\,\bar{\mathrm{tr}}\big[\Delta_{\mathbf{X}}\mathrm{e}^{\theta \mathbf{X}}\big] \qquad \text{for } \theta \ge 0. \tag{A.3}$$

In the proof of Lemma 4.3, we applied an almost sure bound for the conditional variance to control the derivative of the m.g.f. This time, we incorporate information about the typical size of $\boldsymbol{\Delta}_{\mathbf{X}}$ by developing a bound in terms of the function $r(\psi)$.

**A.2. Entropy for random matrices and duality.** Let us introduce an entropy function for random matrices.

DEFINITION A.2 (Entropy for random matrices). Let $\mathbf{W}$ be a random matrix in $\mathbb{H}_+^d$ subject to the normalization $\mathbb{E}\,\bar{\mathrm{tr}}\,\mathbf{W} = 1$. The (*negative*) *matrix entropy* is defined as

$$\mathrm{ent}(\mathbf{W}) := \mathbb{E}\,\bar{\mathrm{tr}}(\mathbf{W}\log \mathbf{W}). \tag{A.4}$$

We enforce the convention that $0 \log 0 = 0$.

The matrix entropy is relevant to our discussion because its Fenchel–Legendre conjugate is the cumulant generating function. The Young inequality for matrix entropy offers one way to formulate this duality relationship.

PROPOSITION A.3 (Young inequality for matrix entropy). *Suppose that* $\mathbf{V}$ *is a random matrix in* $\mathbb{H}^d$ *that is almost surely bounded in norm, and suppose that* $\mathbf{W}$ *is a random matrix in* $\mathbb{H}^d_+$ *subject to the normalization* $\mathbb{E}\,\bar{\mathrm{tr}}\,\mathbf{W} = 1$. *Then*

$$\mathbb{E}\,\bar{\mathrm{tr}}(\mathbf{VW}) \le \log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\mathbf{V}} + \mathrm{ent}(\mathbf{W}).$$

Proposition A.3 follows from a variant of the argument in [5], Theorem 2.13.

**A.3. A refined differential inequality for the trace m.g.f.** We intend to apply the Young inequality for matrix entropy to decouple the product of random matrices in (A.3). First, we must rescale the exponential in (A.3), so its expected trace equals one,

$$\mathbf{W}(\theta) := \frac{1}{\mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta\mathbf{X}}} \cdot \mathrm{e}^{\theta\mathbf{X}} = \frac{1}{m(\theta)} \cdot \mathrm{e}^{\theta\mathbf{X}}. \tag{A.5}$$

For each $\psi > 0$, we can rewrite (A.3) as

$$m'(\theta) \le \frac{\theta m(\theta)}{\psi} \cdot \mathbb{E}\,\bar{\mathrm{tr}}\big[\psi\,\Delta_{\mathbf{X}} \cdot \mathbf{W}(\theta)\big].$$

The Young inequality for matrix entropy, Proposition A.3, implies that

$$m'(\theta) \le \frac{\theta m(\theta)}{\psi}\big[\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi\,\boldsymbol{\Delta}_{\mathbf{X}}} + \mathrm{ent}\big(\mathbf{W}(\theta)\big)\big]. \tag{A.6}$$

The first term in the bracket is precisely $\psi r(\psi)$. Let us examine the second term more closely.

To control the matrix entropy of $\mathbf{W}(\theta)$, we need to bound its logarithm. Referring back to definition (A.5), we see that

$$\log \mathbf{W}(\theta) = \theta\mathbf{X} - \big(\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta\mathbf{X}}\big) \cdot \mathbf{I} \preccurlyeq \theta\mathbf{X} - \big(\log \bar{\mathrm{tr}}\,\mathrm{e}^{\theta\,\mathbb{E}\,\mathbf{X}}\big) \cdot \mathbf{I} = \theta\mathbf{X}. \tag{A.7}$$

The second relation depends on Jensen's inequality and the fact that the trace exponential is convex [38], Section 2. The third relation relies on the property that $\mathbb{E}\,\mathbf{X} = \mathbf{0}$. Since the matrix $\mathbf{W}(\theta)$ is positive, we can substitute the semidefinite bound (A.7) into the definition (A.4) of the matrix entropy,

$$\begin{aligned}\mathrm{ent}\big(\mathbf{W}(\theta)\big) &= \mathbb{E}\,\bar{\mathrm{tr}}\big[\mathbf{W}(\theta) \cdot \log \mathbf{W}(\theta)\big] \\ &\le \theta \cdot \mathbb{E}\,\bar{\mathrm{tr}}\big[\mathbf{W}(\theta) \cdot \mathbf{X}\big] = \frac{\theta}{m(\theta)} \cdot \mathbb{E}\,\bar{\mathrm{tr}}\big[\mathbf{X}\mathrm{e}^{\theta\mathbf{X}}\big].\end{aligned}$$

We have reintroduced the definition (A.5) of $\mathbf{W}(\theta)$ in the last relation. Identify the derivative (A.2) of the trace m.g.f. to reach

$$\operatorname{ent}\bigl(\mathbf{W}(\theta)\bigr) \le \frac{\theta m'(\theta)}{m(\theta)}. \tag{A.8}$$

To establish a differential inequality, substitute the definition (5.1) of $r(\psi)$ and the bound (A.8) into the estimate (A.6) to discover that

$$m'(\theta) \le \frac{\theta m(\theta)}{\psi}\biggl[\psi r(\psi) + \frac{\theta m'(\theta)}{m(\theta)}\biggr] = r(\psi)\theta \cdot m(\theta) + \frac{\theta^2}{\psi} \cdot m'(\theta).$$

Rearrange this formula to isolate the log-derivative $m'(\theta)/m(\theta)$ of the trace m.g.f. We conclude that

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log m(\theta) \le \frac{r(\psi)\theta}{1 - \theta^2/\psi} \qquad \text{for } 0 \le \theta < \sqrt{\psi}. \tag{A.9}$$

**A.4. Solving the differential inequality.** To integrate (A.9), recall that $\log m(0) = 0$, and invoke the fundamental theorem of calculus to reach

$$\log m(\theta) = \int_0^\theta \frac{\mathrm{d}}{\mathrm{d}s} \log m(s)\,\mathrm{d}s \le \int_0^\theta \frac{r(\psi)s}{1 - s^2/\psi}\,\mathrm{d}s = \frac{\psi r(\psi)}{2} \log\biggl(\frac{1}{1 - \theta^2/\psi}\biggr).$$

We can develop a weaker inequality by making a further approximation within the integral

$$\log m(\theta) \le \int_0^\theta \frac{r(\psi)s}{1 - s^2/\psi}\,\mathrm{d}s \le \int_0^\theta \frac{r(\psi)s}{1 - \theta^2/\psi}\,\mathrm{d}s = \frac{r(\psi)\theta^2}{2(1 - \theta^2/\psi)}.$$

These calculations are valid when $0 \le \theta < \sqrt{\psi}$, so claim (A.1) follows.

**A.5. The matrix Laplace transform argument.** With the trace m.g.f. bound (A.1) at hand, we can complete the proof of Theorem 5.1. Proposition 3.3, the matrix Laplace transform method, yields the estimate

$$\begin{aligned}
\mathbb{P}\{\lambda_{\max}(\mathbf{X}) \ge t\} &\le d \cdot \inf_{0<\theta<\sqrt{\psi}} \exp\biggl\{-\theta t + \frac{r(\psi)\theta^2}{2(1 - \theta^2/\psi)}\biggr\} \\
&\le d \cdot \inf_{0<\theta<\sqrt{\psi}} \exp\biggl\{-\theta t + \frac{r(\psi)\theta^2}{2(1 - \theta/\sqrt{\psi})}\biggr\} \\
&= d \cdot \exp\biggl\{-\frac{r(\psi)\psi}{2}\bigl(1 - \sqrt{1 + 2t/\bigl(r(\psi)\sqrt{\psi}\bigr)}\bigr)^2\biggr\} \\
&\le d \cdot \exp\biggl\{-\frac{t^2}{2r(\psi) + 2t/\sqrt{\psi}}\biggr\},
\end{aligned}$$

since the infimum occurs at $\theta = \sqrt{\psi} - \sqrt{\psi}/\sqrt{1 + 2t/(r(\psi)\sqrt{\psi})}$. This delivers the tail bound (5.2).

To establish inequality (5.3) for the expectation of the maximum eigenvalue, we can apply Proposition 3.3 and the trace m.g.f. bound (A.1) a second time. Indeed,

$$\begin{aligned}\mathbb{E}\,\lambda_{\max}(\mathbf{X}) &\le \inf_{0<\theta<\sqrt{\psi}} \frac{1}{\theta}\left[\log d + \frac{r(\psi)\theta^2}{2(1-\theta^2/\psi)}\right] \\ &\le \inf_{0<\theta<\sqrt{\psi}} \frac{1}{\theta}\left[\log d + \frac{r(\psi)\theta^2}{2(1-\theta/\sqrt{\psi})}\right] = \sqrt{2r(\psi)\log d} + \frac{\log d}{\sqrt{\psi}}.\end{aligned}$$

This completes the proof of Theorem 5.1.

## APPENDIX B: PROOF OF THEOREM 7.4

The proof of the matrix Rosenthal inequality takes place in two steps. First, we verify that the bound (7.3) holds for psd random matrices. Then, we use this result to provide a short proof of the bound (7.4) for Hermitian random matrices. Before we start, let us remind the reader that the $L_p$ *norm* of a scalar random variable $Z$ is given by $(\mathbb{E}\,|Z|^p)^{1/p}$ for each $p \ge 1$.

**B.1. A sum of random psd matrices.** We begin with the moment bound (7.3) for an independent sum of random psd matrices. Introduce the quantity of interest

$$E^2 := \left(\mathbb{E}\left\|\sum_k \mathbf{P}_k\right\|_{2p}^{2p}\right)^{1/(2p)}.$$

We may invoke the triangle inequality for the $L_{2p}$ norm to obtain

$$\begin{aligned}E^2 &\le \left(\mathbb{E}\left\|\sum_k (\mathbf{P}_k - \mathbb{E}\,\mathbf{P}_k)\right\|_{2p}^{2p}\right)^{1/(2p)} + \left\|\sum_k \mathbb{E}\,\mathbf{P}_k\right\|_{2p} \\ &=: \left(\mathbb{E}\,\|\mathbf{X}\|_{2p}^{2p}\right)^{1/(2p)} + \mu.\end{aligned}$$

We can apply the matrix BDG inequality to control this expectation, which yields an algebraic inequality between $E^2$ and $E$. We solve this inequality to bound $E^2$.

The series $\mathbf{X}$ consists of centered, independent random matrices, so we can use the Stein pair described in Section 2.4. According to (2.6), the conditional variance $\boldsymbol{\Delta}_{\mathbf{X}}$ takes the form

$$\begin{aligned}\boldsymbol{\Delta}_{\mathbf{X}} &= \frac{1}{2}\sum_k \left[(\mathbf{P}_k - \mathbb{E}\,\mathbf{P}_k)^2 + \mathbb{E}(\mathbf{P}_k - \mathbb{E}\,\mathbf{P}_k)^2\right] \\ &\preccurlyeq \frac{1}{2}\sum_k \left[2\mathbf{P}_k^2 + 2(\mathbb{E}\,\mathbf{P}_k)^2 + \mathbb{E}\,\mathbf{P}_k^2 - (\mathbb{E}\,\mathbf{P}_k)^2\right] \\ &\preccurlyeq \sum_k (\mathbf{P}_k^2 + \mathbb{E}\,\mathbf{P}_k^2).\end{aligned}$$

The first inequality follows from the operator convexity (1.1) of the square function; the second expectation is computed exactly. The last bound uses the operator Jensen inequality (1.2). Now, the matrix BDG inequality yields

$$\begin{aligned} E^2 &\le \sqrt{2p-1}\cdot\left(\mathbb{E}\,\|\boldsymbol{\Delta}_{\mathbf{X}}\|_p^p\right)^{1/(2p)} + \mu \\ &\le \sqrt{2p-1}\cdot\left(\mathbb{E}\left\|\sum_k (\mathbf{P}_k^2 + \mathbb{E}\,\mathbf{P}_k^2)\right\|_p^p\right)^{1/(2p)} + \mu \\ &\le \sqrt{4p-2}\cdot\left(\mathbb{E}\left\|\sum_k \mathbf{P}_k^2\right\|_p^p\right)^{1/(2p)} + \mu. \end{aligned}$$

The third line follows from the triangle inequality for the $L_p$ norm and Jensen's inequality.

Next, we search for a copy of $E^2$ inside this expectation. To accomplish this goal, we want to draw a factor $\mathbf{P}_k$ off of each term in the sum. The following result of Pisier and Xu [40], Lemma 2.6, has the form we desire.

PROPOSITION B.1 (A matrix Schwarz-type inequality). *Consider a finite sequence $(\mathbf{A}_k)_{k\ge 1}$ of deterministic psd matrices. For each $p \ge 1$,*

$$\left\|\sum_k \mathbf{A}_k^2\right\|_p \le \left(\sum_k \|\mathbf{A}_k\|_{2p}^{2p}\right)^{1/(2p)} \left\|\sum_k \mathbf{A}_k\right\|_{2p}.$$

Apply the matrix Schwarz-type inequality, Proposition B.1, to reach

$$\begin{aligned} E^2 &\le \sqrt{4p-2}\cdot\left[\mathbb{E}\left(\sum_k \|\mathbf{P}_k\|_{2p}^{2p}\right)^{1/2}\left\|\sum_k \mathbf{P}_k\right\|_{2p}^p\right]^{1/(2p)} + \mu \\ &\le \sqrt{4p-2}\cdot\left(\sum_k \mathbb{E}\,\|\mathbf{P}_k\|_{2p}^{2p}\right)^{1/(4p)}\left(\mathbb{E}\left\|\sum_k \mathbf{P}_k\right\|_{2p}^{2p}\right)^{1/(4p)} + \mu. \end{aligned}$$

The second bound is the Cauchy–Schwarz inequality for expectation. The resulting estimate takes the form $E^2 \le cE + \mu$. Solutions of this quadratic inequality must satisfy $E \le c + \sqrt{\mu}$. We reach

$$E \le \sqrt{4p-2}\cdot\left(\sum_k \mathbb{E}\,\|\mathbf{P}_k\|_{2p}^{2p}\right)^{1/(4p)} + \left\|\sum_k \mathbb{E}\,\mathbf{P}_k\right\|_{2p}^{1/2}.$$

Square this expression to complete the proof of (7.3).

**B.2. A sum of centered, random Hermitian matrices.** We are now prepared to establish bound (7.4) for a sum of centered, independent, random Hermitian matrices. Define the random matrix $\mathbf{X} := \sum_k \mathbf{Y}_k$. We may use the matrix Stein pair

described in Section 2.4. According to (2.6), the conditional variance $\mathbf{\Delta_X}$ takes the form

$$\mathbf{\Delta_X} = \frac{1}{2}\sum_k (\mathbf{Y}_k^2 + \mathbb{E}\,\mathbf{Y}_k^2).$$

The matrix BDG inequality, Theorem 7.1, yields

$$\begin{aligned}
\big(\mathbb{E}\,\|\mathbf{X}\|_{4p}^{4p}\big)^{1/(4p)} &\le \sqrt{4p-1}\cdot\big(\mathbb{E}\,\|\mathbf{\Delta_X}\|_{2p}^{2p}\big)^{1/(4p)} \\
&= \sqrt{4p-1}\cdot\left(\mathbb{E}\left\|\frac{1}{2}\sum_k(\mathbf{Y}_k^2+\mathbb{E}\,\mathbf{Y}_k^2)\right\|_{2p}^{2p}\right)^{1/(4p)} \\
&\le \sqrt{4p-1}\cdot\left(\mathbb{E}\left\|\sum_k \mathbf{Y}_k^2\right\|_{2p}^{2p}\right)^{1/(4p)}.
\end{aligned}$$

The third line follows from the triangle inequality for the $L_{2p}$ norm and Jensen's inequality. To bound the remaining expectation, we simply note that the sum consists of independent, random psd matrices. We complete the proof by invoking the matrix Rosenthal inequality (7.3) and simplifying.

## APPENDIX C: PROOF OF THEOREM 10.3

Consider the matrix Stein pair $(\mathbf{X}, \mathbf{X}')$ constructed in Section 10.2. Expression (10.2) and the operator convexity (1.1) of the matrix square allow us to bound the conditional variance as follows.

$$\begin{aligned}
\mathbf{\Delta_X}(\pi) &= \frac{1}{4n}\sum_{j,k=1}^n [\mathbf{A}_{j\pi(j)} + \mathbf{A}_{k\pi(k)} - \mathbf{A}_{j\pi(k)} - \mathbf{A}_{k\pi(j)}]^2 \\
&\preccurlyeq \frac{1}{n}\sum_{j,k=1}^n [\mathbf{A}_{j\pi(j)}^2 + \mathbf{A}_{k\pi(k)}^2 + \mathbf{A}_{j\pi(k)}^2 + \mathbf{A}_{k\pi(j)}^2] \\
&= 2\sum_{j=1}^n \mathbf{A}_{j\pi(j)}^2 + \frac{2}{n}\sum_{j,k=1}^n \mathbf{A}_{jk}^2 = \mathbf{W} + 4\mathbf{\Sigma},
\end{aligned}$$

where

$$\mathbf{W} := 2\left(\sum_{j=1}^n \mathbf{A}_{j\pi(j)}^2\right) - 2\mathbf{\Sigma} \quad \text{and} \quad \Sigma := \frac{1}{n}\sum_{j,k=1}^n \mathbf{A}_{jk}^2.$$

Substitute the bound for $\mathbf{\Delta_X}(\pi)$ into the definition (5.1) of $r(\psi)$ to see that

$$\begin{aligned}
r(\psi) &:= \frac{1}{\psi}\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi\mathbf{\Delta_X}(\pi)} \\
&\le \frac{1}{\psi}\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi(\mathbf{W}+4\mathbf{\Sigma})} \le 4\sigma^2 + \frac{1}{\psi}\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi\mathbf{W}}.
\end{aligned} \tag{C.1}$$

The inequalities follow from the monotonicity of the trace exponential [38], Section 2 and the fact that $\sigma^2 = \|\mathbf{\Sigma}\|$. Therefore, it suffices to bound the trace m.g.f. of $\mathbf{W}$.

Our approach is to construct a matrix Stein pair for $\mathbf{W}$ and to argue that the associated conditional variance $\mathbf{\Delta}_{\mathbf{W}}(\pi)$ satisfies a semidefinite bound. We may then exploit the trace m.g.f. bounds from Lemma 4.3. Observe that $\mathbf{W}$ and $\mathbf{X}$ take the same form: both have mean zero and share the structure of a combinatorial sum. Therefore, we can study the behavior of $\mathbf{W}$ using the matrix Stein pair from Section 10.2. Adapting (10.2), we see that the conditional variance of $\mathbf{W}$ satisfies

$$\begin{aligned}\mathbf{\Delta}_{\mathbf{W}}(\pi) &= \frac{1}{n}\sum_{j,k=1}^{n}\left[\mathbf{A}_{j\pi(j)}^2 + \mathbf{A}_{k\pi(k)}^2 - \mathbf{A}_{j\pi(k)}^2 - \mathbf{A}_{k\pi(j)}^2\right]^2 \\ &\preccurlyeq \frac{4}{n}\sum_{j,k=1}^{n}\left[\mathbf{A}_{j\pi(j)}^4 + \mathbf{A}_{k\pi(k)}^4 + \mathbf{A}_{j\pi(k)}^4 + \mathbf{A}_{k\pi(j)}^4\right] \\ &\preccurlyeq \frac{4R^2}{n}\sum_{j,k=1}^{n}\left[\mathbf{A}_{j\pi(j)}^2 + \mathbf{A}_{k\pi(k)}^2 + \mathbf{A}_{j\pi(k)}^2 + \mathbf{A}_{k\pi(j)}^2\right].\end{aligned}$$

In the first line, the centering terms in $\mathbf{W}$ cancel each other out. Then we apply the operator convexity (1.1) of the matrix square and the bound $\mathbf{A}_{jk}^4 \preccurlyeq R^2\mathbf{A}_{jk}^2$. Finally, identify $\mathbf{W}$ and $\mathbf{\Sigma}$ to reach

$$\mathbf{\Delta}_{\mathbf{W}}(\pi) \preccurlyeq 4R^2(\mathbf{W} + 4\mathbf{\Sigma}) \preccurlyeq 4R^2 \cdot \mathbf{W} + 16R^2\sigma^2 \cdot \mathbf{I}. \tag{C.2}$$

Matrix inequality (C.2) gives us access to established trace m.g.f. bounds. Indeed,

$$\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi\mathbf{W}} \le \frac{8R^2\sigma^2\psi^2}{1 - 4R^2\psi}$$

as a consequence of Lemma 4.3 with parameters $c = 4R^2$ and $v = 16R^2\sigma^2$.

At last, we substitute the latter bound into (C.1) to discover that

$$r(\psi) \le 4\sigma^2 + \frac{8R^2\sigma^2\psi}{1 - 4R^2\psi}.$$

In particular, setting $\psi = (8R^2)^{-1}$, we find that $r(\psi) \le 6\sigma^2$. Apply Theorem 5.1 to wrap up.

**Acknowledgments** The authors thank Houman Owhadi for helpful conversations. This paper is based on two independent manuscripts from mid-2011 that both applied the method of exchangeable pairs to establish matrix concentration inequalities. One manuscript is by Mackey and Jordan; the other is by Chen, Farrell and Tropp. The authors have combined this research into a single unified presentation, with equal contributions from both groups.

## REFERENCES


[1] AHLSWEDE, R. and WINTER, A. (2002). Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory* **48** 569–579. MR1889969

[2] BHATIA, R. (1997). *Matrix Analysis. Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

[3] BUCHHOLZ, A. (2001). Operator Khintchine inequality in noncommutative probability. *Math. Ann.* **319** 1–16. MR1812816

[4] BURKHOLDER, D. L. (1973). Distribution function inequalities for martingales. *Ann. Probab.* **1** 19–42. MR0365692

[5] CARLEN, E. (2010). Trace inequalities and quantum entropy: An introductory course. In *Entropy and the Quantum. Contemp. Math.* **529** 73–140. Amer. Math. Soc., Providence, RI. MR2681769

[6] CHATTERJEE, S. (2007). Stein's method for concentration inequalities. *Probab. Theory Related Fields* **138** 305–321. MR2288072

[7] CHATTERJEE, S. (2008). Concentration inequalities with exchangeable pairs. Ph.D. thesis, Stanford Univ., Palo Alto.

[8] CHEN, R. Y., GITTENS, A. and TROPP, J. A. (2012). The masked sample covariance estimator: An analysis using matrix concentration inequalities. *Information and Inference* **1** 2–20.

[9] CHEUNG, S. S., SO, A. M.-C. and WANG, K. (2011). Chance-constrained linear matrix inequalities with dependent perturbations: A safe tractable approximation approach. Available at http://www.optimization-online.org/DB_FILE/2011/01/2898.pdf.

[10] CHIU, J. and DEMANET, L. (2011). Sublinear randomized algorithms for skeleton decomposition. Available at arXiv:1110.4193.

[11] CHIU, J. and DEMANET, L. (2012). Matrix probing and its conditioning. *SIAM J. Numer. Anal.* **50** 171–193. MR2888309

[12] COHEN, A., DAVENPORT, M. and LEVIATAN, D. (2011). On the stability and accuracy of least-squares approximation. Available at arXiv:1111.4422.

[13] FOYGEL, R. and SREBRO, N. (2011). Concentration-based guarantees for low-rank matrix reconstruction. *J. Mach. Learn. Res.* **19** 315–340.

[14] GITTENS, A. (2011). The spectral norm error of the naïve Nyström extension. Available at arXiv:1110.5305.

[15] GITTENS, A. and TROPP, J. A. (2011). Tail bounds for all eigenvalues of a sum of random matrices. Available at arXiv:1104.4513.

[16] GROSS, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory* **57** 1548–1566. MR2815834

[17] HIGHAM, N. J. (2008). *Functions of Matrices*: *Theory and Computation*. SIAM, Philadelphia, PA. MR2396439

[18] HOEFFDING, W. (1951). A combinatorial central limit theorem. *Ann. Math. Statistics* **22** 558–566. MR0044058

[19] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

[20] HSU, D., KAKADE, S. M. and ZHANG, T. (2012). Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electron. Commun. Probab.* **17** 13. MR2900355

[21] JUNGE, M. and XU, Q. (2003). Noncommutative Burkholder/Rosenthal inequalities. *Ann. Probab.* **31** 948–995. MR1964955

[22] JUNGE, M. and XU, Q. (2008). Noncommutative Burkholder/Rosenthal inequalities. II. Applications. *Israel J. Math.* **167** 227–282. MR2448025

[23] JUNGE, M. and ZHENG, Q. (2011). Noncommutative Bennett and Rosenthal inequalities. Available at arXiv:1111.1027.

[24] KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Math*. **2033**. Springer, Heidelberg. MR2829871

[25] LEDOUX, M. (2001). *The Concentration of Measure Phenomenon*. *Mathematical Surveys and Monographs* **89**. Amer. Math. Soc., Providence, RI. MR1849347

[26] LIEB, E. H. (1973). Convex trace functions and the Wigner–Yanase–Dyson conjecture. *Adv. Math*. **11** 267–288. MR0332080

[27] LUGOSI, G. (2009). Concentration-of-measure inequalities. Available at http://www.econ.upf.edu/~lugosi/anu.pdf.

[28] LUST-PIQUARD, F. (1986). Inégalités de Khintchine dans $C_p$ $(1 < p < \infty)$. *C. R. Acad. Sci. Paris Sér. I Math*. **303** 289–292. MR0859804

[29] LUST-PIQUARD, F. and PISIER, G. (1991). Noncommutative Khintchine and Paley inequalities. *Ark. Mat*. **29** 241–260. MR1150376

[30] MACKEY, L., TALWALKAR, A. and JORDAN, M. I. (2011). Divide-and-conquer matrix factorization. In *Advances in Neural Information Processing Systems* 24 (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira and K. Q. Weinberger, eds.) 1134–1142.

[31] MCDIARMID, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, 1989 (*Norwich*, 1989). *London Mathematical Society Lecture Note Series* **141** 148–188. Cambridge Univ. Press, Cambridge. MR1036755

[32] MINSKER, S. (2011). Some extensions of Bernstein's inequality for self-adjoint operators. Available at arXiv:1112.5448.

[33] NAGAEV, S. V. and PINELIS, I. F. (1977). Some inequalities for the distributions of sums of independent random variables. *Theory Probab. Appl*. **22** 248–256.

[34] NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res*. **13** 1665–1697. MR2930649

[35] OLIVEIRA, R. I. (2009). Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Available at arXiv:0911.0600.

[36] OLIVEIRA, R. I. (2010). Sums of random Hermitian matrices and an inequality by Rudelson. *Electron. Commun. Probab*. **15** 203–212. MR2653725

[37] PAULSEN, V. (2002). *Completely Bounded Maps and Operator Algebras*. *Cambridge Studies in Advanced Mathematics* **78**. Cambridge Univ. Press, Cambridge. MR1976867

[38] PETZ, D. (1994). A survey of certain trace inequalities. In *Functional Analysis and Operator Theory* (*Warsaw*, 1992). *Banach Center Publ*. **30** 287–298. Polish Acad. Sci., Warsaw. MR1285615

[39] PINELIS, I. (1994). Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab*. **22** 1679–1706. MR1331198

[40] PISIER, G. and XU, Q. (1997). Noncommutative martingale inequalities. *Comm. Math. Phys*. **189** 667–698. MR1482934

[41] RAUHUT, H. (2010). Compressive sensing and structured random matrices. In *Theoretical Foundations and Numerical Methods for Sparse Recovery*. *Radon Ser. Comput. Appl. Math*. **9** 1–92. de Gruyter, Berlin. MR2731597

[42] RECHT, B. (2011). A simpler approach to matrix completion. *J. Mach. Learn. Res*. **12** 3413–3430. MR2877360

[43] ROSENTHAL, H. P. (1970). On the subspaces of $L_p$ $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math*. **8** 273–303. MR0271721

[44] RUDELSON, M. (1999). Random vectors in the isotropic position. *J. Funct. Anal*. **164** 60–72. MR1694526

[45] RUDELSON, M. and VERSHYNIN, R. (2007). Sampling from large matrices: An approach through geometric functional analysis. *J. ACM* **54** Art. 21, 19 pp. (electronic). MR2351844

[46] SO, A. M.-C. (2011). Moment inequalities for sums of random matrices and their applications in optimization. *Math. Program.* **130** 125–151. MR2853163

[47] STEIN, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol. II*: *Probability Theory* 583–602. Univ. California Press, Berkeley, CA. MR0402873

[48] TROPP, J. A. (2011). Freedman's inequality for matrix martingales. *Electron. Commun. Probab.* **16** 262–270. MR2802042

[49] TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459

[50] WIGDERSON, A. and XIAO, D. (2008). Derandomizing the Ahlswede–Winter matrix-valued Chernoff bound using pessimistic estimators, and applications. *Theory Comput.* **4** 53–76. MR2403380

L. MACKEY
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL
390 SERRA MALL
STANFORD, CALIFORNIA 94305-4065
USA
E-MAIL: lmackey@stanford.edu

M. I. JORDAN
DEPARTMENTS OF EECS AND STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
427 EVANS HALL
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: jordan@stat.berkeley.edu

R. Y. CHEN
B. FARRELL
J. A. TROPP
DEPARTMENT OF COMPUTING
AND MATHEMATICAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
1200 E. CALIFORNIA BLVD.
PASADENA, CALIFORNIA 91125
USA
E-MAIL: ycchen@caltech.edu
farrell@cms.caltech.edu
jtropp@cms.caltech.edu