The Annals of Probability
2016, Vol. 44, No. 5, 3431–3473
DOI: 10.1214/15-AOP1054

EFRON–STEIN INEQUALITIES FOR RANDOM MATRICES

BY DANIEL PAULIN*,[1], LESTER MACKEY† AND JOEL A. TROPP‡

National University of Singapore, Stanford University† and California Institute of Technology‡*

This paper establishes new concentration inequalities for random matrices constructed from independent random variables. These results are analogous with the generalized Efron–Stein inequalities developed by Boucheron et al. The proofs rely on the method of exchangeable pairs.

1. Introduction. Matrix concentration inequalities provide probabilistic bounds for the spectral-norm deviation of a random matrix from its mean value. The monograph [Tropp (2015)] contains an overview of this theory and an extensive bibliography. This machinery has revolutionized the analysis of nonclassical random matrices that arise in statistics [Koltchinskii (2011)], machine learning [Morvant, Koço and Ralaivola (2012)], signal processing [Netrapalli, Jain and Sanghavi (2013)], numerical analysis [Avron and Toledo (2014)], theoretical computer science [Wigderson and Xiao (2008)] and combinatorics [Oliveira (2009)].

In the scalar setting, the core concentration results concern sums of independent random variables. Likewise, in the matrix setting, the central results concern independent sums. For example, the matrix Bernstein inequality [Tropp (2012), Theorem 1.4] describes the behavior of independent, centered random matrices that are subject to a uniform bound. There are also a few results that apply to more general classes of random matrices, for example, the matrix bounded difference inequality [Tropp (2012), Corollary 7.5] and the dependent matrix inequalities of Mackey et al. (2014). Nevertheless, it is common to encounter random matrices that we cannot treat using these techniques.

In the scalar setting, there are concentration inequalities that can provide information about the fluctuations of more complicated random variables. In particular, Efron–Stein inequalities [Boucheron, Lugosi and Massart (2003), Boucheron et al. (2005)] describe the concentration of functions of independent random variables in terms of random estimates for the local Lipschitz behavior of those functions. These results have found extensive applications [Boucheron, Lugosi and Massart (2013)].

Received August 2014; revised August 2015.

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

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

Key words and phrases. Concentration inequalities, Stein's method, random matrix, noncommutative, exchangeable pairs, coupling, bounded differences, Efron–Stein inequality, trace inequality.

The goal of this paper is to establish new Efron–Stein inequalities that describe the concentration properties of a matrix-valued function of independent random variables. The main results appear below as Theorems 4.2 and 4.3.

To highlight the value of this work, we establish an improved version of the matrix bounded difference inequality (Corollary 6.1). We also develop a more substantial application to compound sample covariance matrices (Theorem 7.1).

We anticipate that our results have many additional consequences. For instance, we envision new proofs of consistency for correlation matrix estimation Cai and Jiang (2011), Shao and Zhou (2014) and inverse covariance matrix estimation Ravikumar et al. (2011) under sparsity constraints.

1.1. *Technical approach.* In the scalar setting, the generalized Efron–Stein inequalities were originally established using entropy methods Boucheron, Lugosi and Massart (2003), Boucheron et al. (2005). Unfortunately, in the matrix setting, entropy methods do not seem to have the same strength [Chen and Tropp (2014)].

Instead, our argument is based on ideas from the method of exchangeable pairs [Stein (1972, 1986)]. In the scalar setting, this approach for proving concentration inequalities was initiated in the paper Chatterjee (2007) and the thesis Chatterjee (2005). The extension to random matrices appears in the recent paper Mackey et al. (2014).

The method of exchangeable pairs has two chief advantages over alternative approaches to matrix concentration. First, it offers a straightforward way to prove polynomial moment inequalities, which are not easy to obtain using earlier techniques. Second, exchangeable pair arguments also apply to random matrices constructed from weakly dependent random variables.

Mackey et al. (2014) focuses on sums of weakly dependent random matrices because the techniques are less effective for general matrix-valued functionals. In this work, we address this shortcoming by developing a matrix version of the kernel coupling construction from Chatterjee (2005), Section 4.1. This argument requires some challenging new matrix inequalities that may have independent interest. We also describe some new techniques for controlling the evolution of the kernel coupling.

We believe that our proof of the Efron–Stein inequality via the method of exchangeable pairs is novel, even in the scalar setting. As a consequence, our paper contributes to the growing literature that uses Stein's ideas to develop concentration inequalities.

2. Notation and preliminaries from matrix analysis. This section summarizes our notation, as well as some background results from matrix analysis. The reader may prefer to skip this material at first; we have included detailed cross-references throughout the paper.

2.1. *Elementary matrices.* First, we introduce the identity matrix $\mathbf{I}$ and the zero matrix $\mathbf{0}$. The standard basis matrix $\mathbf{E}_{ij}$ has a one in the (i, j) position and zeros elsewhere. The dimensions of these matrices are determined by context.

2.2. *Sets of matrices and the semidefinite order.* We write $\mathbb{M}^d$ for the algebra of $d \times d$ complex matrices. The *trace* and *normalized trace* are given by

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

The symbol $\|\cdot\|$ always refers to the usual operator norm on $\mathbb{M}^d$ induced by the ℓ_2^d vector norm. We also equip $\mathbb{M}^d$ with the trace inner product $\langle \mathbf{B}, \mathbf{C} \rangle := \operatorname{tr}[\mathbf{B}^* \mathbf{C}]$ to form a Hilbert space.

Let $\mathbb{H}^d$ denote the real-linear subspace of $\mathbb{M}^d$ consisting of $d \times d$ Hermitian matrices. The cone of positive-semidefinite matrices will be abbreviated as $\mathbb{H}_+^d$. Given an interval I of the real line, we also define $\mathbb{H}^d(I)$ to be the convex set of Hermitian matrices whose eigenvalues are all contained in I.

We use curly inequalities, such as $\preccurlyeq$, for the positive-semidefinite order on the Hilbert space $\mathbb{H}^d$. That is, for $\mathbf{A}, \mathbf{B} \in \mathbb{H}^d$, we write $\mathbf{A} \preccurlyeq \mathbf{B}$ if and only if $\mathbf{B} - \mathbf{A}$ is positive semidefinite.

2.3. *Matrix functions.* Let $f : I \to \mathbb{R}$ be a function on an interval I of the real line. We can lift f to form a *standard matrix function* $f : \mathbb{H}^d(I) \to \mathbb{H}^d$. More precisely, for each matrix $\mathbf{A} \in \mathbb{H}^d(I)$, we define the standard matrix function via the rule

$$f(\mathbf{A}) := \sum_{k=1}^{d} f(\lambda_k) \mathbf{u}_k \mathbf{u}_k^* \qquad \text{where } \mathbf{A} = \sum_{k=1}^{d} \lambda_k \mathbf{u}_k \mathbf{u}_k^*$$

is an eigenvalue decomposition of the Hermitian matrix $\mathbf{A}$. When we apply a familiar scalar function to an Hermitian matrix, we are always referring to the associated standard matrix function. To denote general matrix-valued functions, we use bold uppercase letters, such as $\mathbf{F}$, $\mathbf{H}$, $\mathbf{\Psi}$.

2.4. *Monotonicity and convexity of trace functions.* The trace of a standard matrix function inherits certain properties from the scalar function. Let I be an interval, and assume that $\mathbf{A}, \mathbf{B} \in \mathbb{H}(I)$. When the function $f : I \to \mathbb{R}$ is increasing,[2]

$$\mathbf{A} \preccurlyeq \mathbf{B} \quad \text{implies} \quad \operatorname{tr} f(\mathbf{A}) \le \operatorname{tr} f(\mathbf{B}). \tag{2.1}$$

When the function $f : I \to \mathbb{R}$ is convex,

$$\operatorname{tr} f\big(\tau \mathbf{A} + (1 - \tau)\mathbf{B}\big) \le \tau \operatorname{tr} f(\mathbf{A}) + (1 - \tau) \operatorname{tr} f(\mathbf{B}) \qquad \text{for } \tau \in [0, 1]. \tag{2.2}$$

See Petz (1994), Propositions 1 and 2, for proofs.

[2]We place the convention that "increasing" means "nondecreasing."

2.5. *The real part of a matrix and the matrix square.* For each matrix $\mathbf{M} \in \mathbb{M}^d$, we introduce the real and imaginary parts,

$$\begin{aligned} \operatorname{Re}(\mathbf{M}) &:= \frac{1}{2}(\mathbf{M} + \mathbf{M}^*) \in \mathbb{H}^d \quad \text{and} \\ \operatorname{Im}(\mathbf{M}) &:= \frac{1}{2\mathrm{i}}(\mathbf{M} - \mathbf{M}^*) \in \mathbb{H}^d. \end{aligned} \tag{2.3}$$

Note the semidefinite bound

$$\operatorname{Re}(\mathbf{M})^2 \preccurlyeq \tfrac{1}{2}(\mathbf{M}\mathbf{M}^* + \mathbf{M}^*\mathbf{M}) \qquad \text{for each } \mathbf{M} \in \mathbb{M}^d. \tag{2.4}$$

Indeed, $\operatorname{Re}(\mathbf{M})^2 + \operatorname{Im}(\mathbf{M})^2 = \frac{1}{2}(\mathbf{M}\mathbf{M}^* + \mathbf{M}^*\mathbf{M})$ and $\operatorname{Im}(\mathbf{M})^2 \succcurlyeq \mathbf{0}$.

The real part of a product of Hermitian matrices satisfies

$$\operatorname{Re}(\mathbf{AB}) = \frac{\mathbf{AB} + \mathbf{BA}}{2} \preccurlyeq \frac{\mathbf{A}^2 + \mathbf{B}^2}{2} \qquad \text{for all } \mathbf{A}, \mathbf{B} \in \mathbb{H}^d. \tag{2.5}$$

This result follows when we expand $(\mathbf{A} - \mathbf{B})^2 \succcurlyeq \mathbf{0}$. As a consequence,

$$\left(\frac{\mathbf{A} + \mathbf{B}}{2}\right)^2 \preccurlyeq \frac{\mathbf{A}^2 + \mathbf{B}^2}{2} \qquad \text{for all } \mathbf{A}, \mathbf{B} \in \mathbb{H}^d. \tag{2.6}$$

In other words, the matrix square is operator convex.

2.6. *Some matrix norms.* Finally, we will make use of two additional families of matrix norms. For $p \in [1, \infty]$, the Schatten p-norm is given by

$$\|\mathbf{B}\|_{S_p} := \left(\operatorname{tr}|\mathbf{B}|^p\right)^{1/p} \qquad \text{for each } \mathbf{B} \in \mathbb{M}^d, \tag{2.7}$$

where $|\mathbf{B}| := (\mathbf{B}^*\mathbf{B})^{1/2}$. For $p \geq 1$, we introduce the matrix norm induced by the ℓ_p^d vector norm:

$$\|\mathbf{B}\|_{p \to p} := \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{Bx}\|_p}{\|\mathbf{x}\|_p} \qquad \text{for each } \mathbf{B} \in \mathbb{M}^d. \tag{2.8}$$

In particular, the matrix norm induced by the ℓ_1^d vector norm returns the maximum ℓ_1^d norm of a column; the norm induced by ℓ_∞^d returns the maximum ℓ_1^d norm of a row.

3. Matrix moments and concentration. Our goal is to develop expectation and tail bounds for the spectral norm of a random matrix. As in the scalar setting, these results follow from bounds for polynomial and exponential moments. This section describes the mechanism by which we convert bounds for matrix moments into concentration inequalities.

3.1. *The matrix Chebyshev inequality.* We can obtain concentration inequalities for a random matrix in terms of the Schatten p-norm. This fact extends Chebyshev's inequality.

PROPOSITION 3.1 (Matrix Chebyshev inequality). *Let* $\mathbf{X} \in \mathbb{H}^d$ *be a random matrix. For all* $t > 0$,

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

Furthermore,

$$\mathbb{E}\,\|\mathbf{X}\| \leq \inf_{p \geq 1} \big(\mathbb{E}\,\|\mathbf{X}\|_{S_p}^p\big)^{1/p}.$$

This statement repeats Mackey et al. (2014), Proposition 6.2. See also Ahlswede and Winter (2002), Appendix, for earlier work.

3.2. *The matrix Laplace transform method.* We can also obtain exponential concentration inequalities from a matrix version of the moment generating function.

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

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

We believe this definition is due to Ahlswede and Winter (2002).

The following proposition is an extension of Bernstein's method. It converts bounds for the trace m.g.f. of a random matrix into bounds on its maximum eigenvalue.

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

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

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

Proposition 3.3 restates Mackey et al. (2014), Proposition 3.3, which collects results from Ahlswede and Winter (2002), Chen, Gittens and Tropp (2012), Oliveira (2010), Tropp (2012).

We will use a special case of Proposition 3.3. This result delineates the consequences of a specific bound for the trace m.g.f.

PROPOSITION 3.4. *Let* $\mathbf{X} \in \mathbb{H}^d$ *be a random matrix with normalized trace m.g.f.* $m(\theta) := \mathbb{E}\,\bar{\operatorname{tr}}\, \mathrm{e}^{\theta \mathbf{X}}$. *Assume that there are nonnegative constants* c, v *for which*

$$\log m(\theta) \leq \frac{v\theta^2}{2(1 - c\theta)} \qquad \text{when } 0 \leq \theta < 1/c.$$

Then, for all $t \geq 0$,

$$\mathbb{P}\{\lambda_{\max}(\mathbf{X}) \geq t\} \leq d \exp\left(\frac{-t^2}{2v + 2ct}\right). \tag{3.5}$$

Furthermore,

$$\mathbb{E}\,\lambda_{\max}(\mathbf{X}) \leq \sqrt{2v \log d} + c \log d.$$

See Mackey et al. (2014), Section 4.2.4, for the proof of Proposition 3.4.

4. Matrix Efron–Stein inequalities. The main outcome of this paper is a family of Efron–Stein inequalities for random matrices. These estimates provide powerful tools for controlling the trace moments of a random matrix in terms of the trace moments of a randomized "variance proxy." Combining these inequalities with the results from Section 3, we can obtain concentration inequalities for the spectral norm.

4.1. *Setup for Efron–Stein inequalities.* Efron–Stein inequalities apply to random matrices constructed from a family of independent random variables. Introduce the random vector

$$Z := (Z_1, \dots, Z_n) \in \mathcal{Z},$$

where $Z_1, \dots, Z_n$ are mutually independent random variables. We assume that $\mathcal{Z}$ is a Polish space to avoid problems with conditioning [Dudley (2002), Theorem 12.2.2]. Let $\mathbf{H} : \mathcal{Z} \to \mathbb{H}^d$ be a measurable function that takes values in the space of Hermitian matrices. We will assume that $\mathbb{E}\,\|\mathbf{H}(Z)\| < \infty$, and construct the centered random matrix

$$\mathbf{X} := \mathbf{X}(Z) := \mathbf{H}(Z) - \mathbb{E}\,\mathbf{H}(Z).$$

Our goal is to study the behavior of $\mathbf{X}$, which describes the fluctuations of the random matrix $\mathbf{H}(Z)$ about its mean value.

A function of independent random variables will concentrate about its mean if it depends smoothly on all of its inputs. We can quantify smoothness by assessing the influence of each coordinate on the matrix-valued function. For each coordinate j, construct the random vector

$$Z^{(j)} := (Z_1, \dots, Z_{j-1}, \widetilde{Z}_j, Z_{j+1}, \dots, Z_n) \in \mathcal{Z},$$

where $\widetilde{Z}_j$ is an independent copy of Z_j. It is clear that Z and $Z^{(j)}$ have the same distribution, and they differ only in coordinate j. Form the random matrices

$$\mathbf{X}^{(j)} := \mathbf{X}(Z^{(j)}) = \mathbf{H}(Z^{(j)}) - \mathbb{E}\,\mathbf{H}(Z) \qquad \text{for } j = 1, \dots, n. \tag{4.1}$$

Note that each $\mathbf{X}^{(j)}$ follows the same distribution as $\mathbf{X}$.

Efron–Stein inequalities control the fluctuations of the centered random matrix $\mathbf{X}$ in terms of the discrepancies between $\mathbf{X}$ and the $\mathbf{X}^{(j)}$. To present these results, let us define the *variance proxy*

$$\mathbf{V} := \frac{1}{2}\sum_{j=1}^{n} \mathbb{E}[(\mathbf{X} - \mathbf{X}^{(j)})^2 | Z]. \tag{4.2}$$

Efron–Stein inequalities bound the trace moments of the random matrix $\mathbf{X}$ in terms of the moments of the variance proxy $\mathbf{V}$. This is similar to the estimate provided by a Poincaré inequality [Boucheron, Lugosi and Massart (2013), Section 3.5].

Passing from the random matrix $\mathbf{X}$ to the variance proxy $\mathbf{V}$ has a number of advantages. There are many situations where the variance proxy admits an accurate deterministic bound, so we can reduce problems involving random matrices to simpler matrix arithmetic. Moreover, the variance proxy is a sum of positive semidefinite terms, which are easier to control than arbitrary random matrices. The examples in Sections 5, 6 and 7 support these claims.

REMARK 4.1. In the scalar setting, Efron–Stein inequalities [Boucheron, Lugosi and Massart (2003), Boucheron et al. (2005)] can alternatively be expressed in terms of the positive part of the fluctuations:

$$V_+ := \frac{1}{2}\sum_{j=1}^{n} \mathbb{E}[(X - X^{(j)})_+^2 | Z],$$

where $(a)_+ := \max\{0, a\}$. Our approach can reproduce these positive-part bounds in the scalar setting but does not deliver positive-part expressions in the general matrix setting. See Section 13.3 for more discussion.

4.2. *Polynomial Efron–Stein inequalities for random matrices.* The first main result of the paper is a polynomial Efron–Stein inequality for a random matrix constructed from independent random variables.

THEOREM 4.2 (Matrix polynomial Efron–Stein). *Instate the notation of Section 4.1. For each natural number $p \geq 1$,*

$$\big(\mathbb{E}\,\|\mathbf{X}\|_{S_{2p}}^{2p}\big)^{1/(2p)} \leq \sqrt{2(2p-1)}\big(\mathbb{E}\,\|\mathbf{V}\|_{S_p}^{p}\big)^{1/(2p)}.$$

The proof appears in Section 11.

We can regard Theorem 4.2 as a matrix extension of the scalar concentration inequality [Boucheron et al. (2005), Theorem 1], which was obtained using the entropy method. In contrast, our results depend on a different style of argument, based on the theory of exchangeable pairs [Chatterjee (2005), Stein (1986)]. Our approach is novel, even in the scalar setting. Unfortunately, it leads to slightly worse constants.

Theorem 4.2 allows us to control the trace moments of a random Hermitian matrix in terms of the trace moments of the variance proxy. We can obtain probability inequalities for the spectral norm by combining this result with the matrix Chebyshev inequality, Proposition 3.1.

4.3. *Exponential Efron–Stein inequalities for random matrices.* The second main result of the paper is an exponential Efron–Stein inequality for a random matrix built from independent random variables.

THEOREM 4.3 (Matrix exponential Efron–Stein). *Instate the notation of Section 4.1, and assume that $\|\mathbf{X}\|$ is bounded. When $|\theta| < \sqrt{\psi/2}$,*

$$\begin{aligned}\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta\mathbf{X}} &\leq \frac{1}{2}\log\left(\frac{1}{1-2\theta^2/\psi}\right)\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi\mathbf{V}} \\ &\leq \frac{\theta^2/\psi}{1-2\theta^2/\psi}\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi\mathbf{V}}.\end{aligned} \tag{4.3}$$

In general, without assuming that $\|\mathbf{X}\|$ is bounded, we have

$$\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta\mathbf{X}} \leq \log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\mathrm{e}\theta^2\mathbf{V}}. \tag{4.4}$$

In particular, when $|\theta| < \sqrt{\psi/\mathrm{e}}$,

$$\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta\mathbf{X}} \leq \big(\mathrm{e}\theta^2/\psi\big)\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi\mathbf{V}}. \tag{4.5}$$

The proof appears in Section 12.1.

Theorem 4.3 is a matrix extension of the exponential Efron–Stein inequalities for scalar random variables established in Boucheron, Lugosi and Massart (2003), Theorem 1, by means of the entropy method. As in the polynomial case, we use a new argument based on exchangeable pairs.

Theorem 4.3 allows us to control trace exponential moments of a random Hermitian matrix in terms of the trace exponential moments of the variance proxy.

We arrive at probability inequalities for the spectral norm by combining this result with the matrix Laplace transform method, Proposition 3.3. Although bounds on polynomial trace moments are stronger than bounds on exponential trace moments [Mackey et al. (2014), Section 6], the exponential inequalities are often more useful in practice.

4.4. *Rectangular matrices*. Suppose now that $\mathbf{H} : \mathcal{Z} \to \mathbb{C}^{d_1+d_2}$ is a measurable function taking *rectangular* matrix values. We can also develop Efron–Stein inequalities for the random rectangular matrix $\mathbf{X} := \mathbf{H}(Z) - \mathbb{E}\,\mathbf{H}(Z)$ as a formal consequence of the results for Hermitian random matrices.

The approach is based on a device from operator theory called the *Hermitian dilation*, which is defined as

$$\mathscr{H}(\mathbf{B}) := \begin{bmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{0} \end{bmatrix} \in \mathbb{H}^{d_1+d_2} \qquad \text{for } \mathbf{B} \in \mathbb{C}^{d_1+d_2}.$$

To obtain Efron–Stein inequalities for random rectangular matrices, we simply apply Theorem 4.2 and Theorem 4.3 to the dilation $\mathscr{H}(\mathbf{X})$. We omit the details. For more information about these arguments, see Tropp (2012), Section 2.6, Mackey et al. (2014), Section 8, or Tropp (2015), Section 2.1.13.

5. Example: Self-bounded random matrices. As a first example, we consider the case where the variance proxy is dominated by an affine function of the centered random matrix.

COROLLARY 5.1 (Self-bounded random matrices). *Instate the notation of Section 4.1. Assume that $\|\mathbf{X}\|$ is bounded, and suppose that there are nonnegative constants c, v for which*

$$\mathbf{V} \preccurlyeq v\mathbf{I} + c\mathbf{X} \qquad \textit{almost surely.} \tag{5.1}$$

Then, for all $t \geq 0$,

$$\begin{aligned} &\mathbb{P}\{\lambda_{\max}(\mathbf{X}) \geq t\} \leq d \exp\left(\frac{-t^2}{4v + 4ct}\right) \quad \textit{and} \\ &\mathbb{P}\{\lambda_{\min}(\mathbf{X}) \leq -t\} \leq d \exp\left(\frac{-t^2}{4v + 4ct}\right). \end{aligned} \tag{5.2}$$

Furthermore,

$$\begin{aligned} &\mathbb{E}\,\lambda_{\max}(\mathbf{X}) \leq \sqrt{4v \log d} + 2c \log d \quad \textit{and} \\ &\mathbb{E}\,\lambda_{\min}(\mathbf{X}) \geq -\sqrt{4v \log d} - 2c \log d. \end{aligned} \tag{5.3}$$

Without assuming that $\|\mathbf{X}\|$ is bounded, (5.2) and (5.3) hold with c and v replaced by $\mathrm{e}c/2$ and $\mathrm{e}v$, respectively.

PROOF. The result is a consequence of the exponential Efron–Stein inequality for random matrices, Theorem 4.3. When $0 \le \theta < \sqrt{\psi/2}$, we may calculate that

$$
\begin{aligned}
\log m_{\mathbf{X}}(\theta) = \log \mathbb{E}\,\bar{\operatorname{tr}}\, \mathrm{e}^{\theta \mathbf{X}} &\le \frac{1}{2}\log\left(\frac{1}{1-2\theta^2/\psi}\right)\log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\psi \mathbf{V}} \\
&\le \frac{1}{2}\log\left(\frac{1}{1-2\theta^2/\psi}\right)\log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\psi(v\mathbf{I}+c\mathbf{X})} \\
&= \frac{1}{2}\log\left(\frac{1}{1-2\theta^2/\psi}\right)(\psi v + \log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\psi c\mathbf{X}}) \\
&\le \frac{(\theta^2/\psi)(1-\theta^2/\psi)}{1-2\theta^2/\psi}(\psi v + \log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\psi c\mathbf{X}}).
\end{aligned}
\tag{5.4}
$$

In the first inequality, we can introduce the bound (5.1) for $\mathbf{V}$ because the trace exponential is monotone (2.1). The final inequality follows from the numerical fact

$$\log\bigl(1/(1-2x)\bigr) \le 2x(1-x)/(1-2x) \qquad \text{for } x \in [0, 1/2).$$

Select $\psi = \theta/c$ to obtain a copy of $m_{\mathbf{X}}(\theta)$ on the right-hand side of (5.4). Solve for $m_{\mathbf{X}}(\theta)$, and appeal to the numerical fact

$$(1-x)(1-2x) \le 1 - 3x(1-x) \qquad \text{for } x \in [0, 1/2)$$

to reach

$$\log m_{\mathbf{X}}(\theta) \le \frac{v\theta^2(1-c\theta)}{1-3c\theta(1-c\theta)} \le \frac{v\theta^2}{1-2c\theta} \qquad \text{when } 0 \le \theta < 1/(2c). \tag{5.5}$$

Invoke Proposition 3.4 to complete the proof.

For the lower tails, notice that we can employ the same argument for $-1/(2c) < \theta \le 0$ up to (5.4). Then we can choose $\psi := |\theta|/c$ and substitute the inequality (5.5) for $|\theta|$ to obtain that

$$\log m_{\mathbf{X}}(\theta) \le \frac{v\theta^2}{1-2c|\theta|} \qquad \text{when } -1/(2c) < \theta \le 0, \tag{5.6}$$

which implies the same bounds as for the upper tail. The advertised inequalities without the boundedness assumption follows from an analogous argument based on (4.5). □

Hypothesis (5.1) is analogous with the assumptions in the result Mackey et al. (2014), Theorem 4.1. In Section 6, we explain how this estimate supports a matrix version of the bounded difference inequality, but the result also extends well beyond this example.

6. Example: Matrix bounded differences. The matrix bounded difference inequality [Mackey et al. (2014), Tropp (2012)] controls the fluctuations of a matrix-valued function of independent random variables. This result has been used to analyze algorithms for multiclass classification [Machart and Ralaivola (2012), Morvant, Koço and Ralaivola (2012)], crowdsourcing [Dalvi et al. (2013)], and nondifferentiable optimization [Zhou and Hu (2014)].

Let us explain how to derive a refined version of the matrix bounded differences inequality from Theorem 4.3.

COROLLARY 6.1 (Matrix bounded differences). *Instate the notation of Section 4.1. Assume there is a deterministic matrix* $\mathbf{A} \in \mathbb{H}^d$ *for which*

$$\sum_{j=1}^{n} \left(\mathbf{H}(z_1, \dots, z_n) - \mathbf{H}(z_1, \dots, z_j', \dots, z_n)\right)^2 \preccurlyeq \mathbf{A}^2$$

for any $z_1, \dots, z_n, z_1', \dots, z_n'$. *Define*

$$\sigma^2 := \|\mathbf{A}\|^2. \tag{6.1}$$

Then, for all $t \geq 0$,

$$\mathbb{P}\{\lambda_{\max}(\mathbf{H}(Z) - \mathbb{E}\,\mathbf{H}(Z)) \geq t\} \leq d \cdot \mathrm{e}^{-t^2/(2\sigma^2)} \quad \textit{and}$$
$$\mathbb{P}\{\lambda_{\min}(\mathbf{H}(Z) - \mathbb{E}\,\mathbf{H}(Z)) \leq -t\} \leq d \cdot \mathrm{e}^{-t^2/(2\sigma^2)}.$$

Furthermore,

$$\mathbb{E}\,\lambda_{\max}(\mathbf{H}(Z) - \mathbb{E}\,\mathbf{H}(Z)) \leq \sigma\sqrt{2\log d} \quad \textit{and}$$
$$\mathbb{E}\,\lambda_{\min}(\mathbf{H}(Z) - \mathbb{E}\,\mathbf{H}(Z)) \geq -\sigma\sqrt{2\log d}.$$

PROOF. Observe that the variance proxy satisfies

$$\begin{aligned}\mathbf{V} &= \frac{1}{2}\sum_{j=1}^{n} \mathbb{E}[(\mathbf{X} - \mathbf{X}^{(j)})^2 | Z] \\ &= \frac{1}{2}\sum_{j=1}^{n} \mathbb{E}[(\mathbf{H}(Z) - \mathbf{H}(Z^{(j)}))^2 | Z] \preccurlyeq \frac{1}{2}\mathbf{A}^2.\end{aligned}$$

It follows from definition (6.1) that $\mathbf{V} \preccurlyeq \frac{1}{2}\sigma^2 \mathbf{I}$. Invoke Corollary 5.1 to complete the argument. Finally, we obtain the same results for the lower tail by applying the argument to $-\mathbf{H}(Z)$. □

REMARK 6.2 (Related work). Previous results in the literature made assumptions of the type

$$\left(\mathbf{H}(z_1, \dots, z_n) - \mathbf{H}(z_1, \dots, z_j', \dots, z_n)\right)^2 \preccurlyeq \mathbf{A}_j^2 \qquad \text{for each index } j,$$

for some deterministic matrices $\mathbf{A}_1, \dots, \mathbf{A}_n \in \mathbb{H}^d$, and let $\sigma^2 := \|\sum_{j=1}^n \mathbf{A}_j^2\|$. Corollary 6.1 is more general, it improves the constants in Tropp (2012), Corollary 7.5, and it removes an extraneous assumption from Mackey et al. (2014), Corollary 11.1. It is possible to further improve the constants in the exponent by a factor of 2 to obtain a bound of the form $d \cdot \mathrm{e}^{-t^2/\sigma^2}$; see the original argument in Paulin, Mackey and Tropp (2013).

7. Application: Compound sample covariance matrices. In this section, we consider the *compound sample covariance matrix*:

$$\widehat{\mathbf{\Lambda}}_n := \frac{1}{n}\mathbf{Z}\mathbf{B}\mathbf{Z}^*. \tag{7.1}$$

The central matrix $\mathbf{B} \in \mathbb{H}^n$ is fixed, and the columns of $\mathbf{Z} \in \mathbb{C}^{p\times n}$ are random vectors drawn independently from a common distribution on $\mathbb{C}^p$.

When the matrix $\mathbf{B} = n^{-1}\mathbf{I}$, the compound sample covariance matrix $\widehat{\mathbf{\Lambda}}_n$ reduces to the classical empirical covariance $n^{-1}\mathbf{Z}\mathbf{Z}^*$. The latter matrix can be written as a sum of independent rank-one matrices, and its concentration properties are well established [Adamczak et al. (2011)]. For general $\mathbf{B}$, however, the random matrix $\widehat{\mathbf{\Lambda}}_n$ cannot be expressed as an independent sum, so the behavior becomes significantly harder to characterize. See, for example, the analysis of Soloveychik (2014).

The most common example of a compound sample covariance matrix is the compound Wishart matrix [Speicher (1998)], where the columns of $\mathbf{Z}$ are drawn from a multivariate normal distribution. These matrices have been used to estimate the sample covariance under correlated sampling [Burda et al. (2011)]. They also arise in risk estimation for portfolio management [Collins, McDonald and Saad (2013)].

We will use Theorem 4.3 to develop an exponential concentration inequality for one class of compound sample covariance matrices.

THEOREM 7.1 (Concentration of compound sample covariance). *Suppose that the entries of $\mathbf{Z} \in \mathbb{C}^{p\times n}$ are independent random variables with mean zero, variance σ^2, and magnitude bounded by L. Let $\mathbf{B} \in \mathbb{H}^n$ be fixed. For any $t \geq 0$ we have*

$$\begin{aligned}\mathbb{P}\{\|\mathbf{Z}\mathbf{B}\mathbf{Z}^* - \mathbb{E}[\mathbf{Z}\mathbf{B}\mathbf{Z}^*]\| \geq t\} \\ \leq 2p \exp\left(\frac{-t^2}{44(p\sigma^2 + L^2)\|\mathbf{B}\|_{\mathrm{F}}^2 + 32\sqrt{3}Lp\|\mathbf{B}\|t}\right).\end{aligned}$$

Furthermore,

$$\mathbb{E}\|\mathbf{Z}\mathbf{B}\mathbf{Z}^* - \mathbb{E}[\mathbf{Z}\mathbf{B}\mathbf{Z}^*]\| \leq 2\sqrt{44(p\sigma^2 + L^2)\log p\|\mathbf{B}\|_{\mathrm{F}}^2} + 32\sqrt{3}Lp\log p\|\mathbf{B}\|.$$

Let us emphasize that Theorem 7.1 is valid even when $\mathbf{B}$ is not positive semidefinite, in contrast to some previous work on this problem. Our proof gives finer results when $\mathbf{B}$ is positive semidefinite. We have also made a number of loose estimates in order to obtain a clear statement of the bound.

7.1. *Setup*. Let $\mathbf{Z}$ be a $p \times n$ random matrix whose entries are independent, identically distributed, zero-mean random variables with variance σ^2 and bounded in magnitude by $L = 1$. The general case follows by a homogeneity argument. Define the centered random matrix

$$\mathbf{X}(\mathbf{Z}) = \mathbf{ZBZ}^* - \mathbb{E}[\mathbf{ZBZ}^*],$$

where $\mathbf{B} \in \mathbb{H}^d$. By direct calculation, the expectation takes the form

$$\mathbb{E}[\mathbf{ZBZ}^*] = \sigma^2(\operatorname{tr}\mathbf{B})\mathbf{I}. \tag{7.2}$$

As in Section 4.1, we introduce independent copies $\widetilde{Z}_{ij}$ of the entries Z_{ij} of $\mathbf{Z}$ and define the random matrices

$$\mathbf{Z}^{(ij)} = \mathbf{Z} + (\widetilde{Z}_{ij} - Z_{ij})\mathbf{E}_{ij} \qquad \text{for } i = 1, \dots, p \text{ and } j = 1, \dots, n.$$

Introduce the variance proxy

$$\mathbf{V} := \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{n}\mathbb{E}\big[\big(\mathbf{X}(\mathbf{Z}) - \mathbf{X}(\mathbf{Z}^{(ij)})\big)^2 | \mathbf{Z}\big]. \tag{7.3}$$

Theorem 4.3 allows us to bound $\lambda_{\max}(\mathbf{X}(\mathbf{Z}))$ in terms of the trace m.g.f. of $\mathbf{V}$. Our task is to develop bounds on the trace m.g.f. of $\mathbf{V}$ in terms of the problem data.

7.2. *A bound for the variance proxy*. We begin with a general bound for $\mathbf{V}$. First, we use the definitions to simplify the expression (7.3), and then we invoke the operator convexity (2.6) of the square function:

$$\begin{aligned}\mathbf{V} &= \frac{1}{2}\sum_{ij}\mathbb{E}\big[\big(2(\widetilde{Z}_{ij} - Z_{ij})\operatorname{Re}\big(\mathbf{E}_{ij}\mathbf{BZ}^*\big) + |\widetilde{Z}_{ij} - Z_{ij}|^2\mathbf{E}_{ij}\mathbf{BE}_{ij}^*\big)^2 | \mathbf{Z}\big] \\ &\preccurlyeq \frac{1}{2}\sum_{ij}\mathbb{E}\big[8|\widetilde{Z}_{ij} - Z_{ij}|^2\operatorname{Re}\big(\mathbf{E}_{ij}\mathbf{BZ}^*\big)^2 + 2|\widetilde{Z}_{ij} - Z_{ij}|^4|b_{jj}|^2\mathbf{E}_{ii} | \mathbf{Z}\big],\end{aligned}$$

where $\{b_{ij}\}_{1 \le i,j \le n}$ denote the elements of $\mathbf{B}$. Since Z_{ij} and $\widetilde{Z}_{ij}$ are centered variables that are bounded in magnitude by one,

$$\mathbb{E}\big[|\widetilde{Z}_{ij} - Z_{ij}|^2 | \mathbf{Z}\big] \le 2 \quad \text{and} \quad \mathbb{E}\big[|\widetilde{Z}_{ij} - Z_{ij}|^4 | \mathbf{Z}\big] \le 8.$$

Using the bound (2.4) for the square of the real part, we obtain

$$\begin{aligned}\mathbf{V} &\preccurlyeq \sum_{ij}\big[4\big(\mathbf{BZ}^*\mathbf{ZB}\big)_{jj}\mathbf{E}_{ii} + 4\mathbf{ZBE}_{jj}\mathbf{BZ}^* + 8|b_{jj}|^2\mathbf{E}_{ii}\big] \\ &= 4p\,\bar{\operatorname{tr}}[\mathbf{ZB}^2\mathbf{Z}^*]\mathbf{I} + 4p\mathbf{ZB}^2\mathbf{Z}^* + 8\Big(\sum_j |b_{jj}|^2\Big)\mathbf{I}.\end{aligned} \tag{7.4}$$

In the first term on the right-hand side of (7.4), we have used cyclicity of the standard trace, and then we have rescaled to obtain the normalized trace.

7.3. *A bound for the trace m.g.f. of the random matrix.* Next, we apply the matrix exponential Efron–Stein inequality, Theorem 4.3, to bound the logarithm of the trace m.g.f. of the random matrix,

$$\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta\mathbf{X}} \le \frac{\theta^2/\psi}{1-2\theta^2/\psi}\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi\mathbf{V}}. \tag{7.5}$$

Let us focus on the trace m.g.f. of $\mathbf{V}$. The trace exponential is monotone (2.1), so we can introduce the bound (7.4) for $\mathbf{V}$ and simplify the expression

$$\begin{aligned}\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\psi\mathbf{V}} &\le \log \mathbb{E}\big[\mathrm{e}^{4\psi p\bar{\mathrm{tr}}[\mathbf{Z}\mathbf{B}^2\mathbf{Z}^*]}\bar{\mathrm{tr}}\,\mathrm{e}^{4\psi p\mathbf{Z}\mathbf{B}^2\mathbf{Z}^*}\big] + 8\psi\Big(\sum_j |b_{jj}|^2\Big)\\ &\le \frac{1}{2}\log \mathbb{E}\,\mathrm{e}^{8\psi p\bar{\mathrm{tr}}[\mathbf{Z}\mathbf{B}^2\mathbf{Z}^*]} + \frac{1}{2}\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{8\psi p\mathbf{Z}\mathbf{B}^2\mathbf{Z}^*} + 8\psi\Big(\sum_j |b_{jj}|^2\Big)\\ &\le \log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{8\psi p\mathbf{Z}\mathbf{B}^2\mathbf{Z}^*} + 8\psi\Big(\sum_j |b_{jj}|^2\Big).\end{aligned}$$

To reach the second line, we use the Cauchy–Schwarz inequality for expectation, and we use Jensen's inequality to pull the normalized trace through the square. To arrive at the last expression, we apply Jensen's inequality to draw out the normalized trace from the exponential. Substitute the last display into (7.5) and write out the definition of $\mathbf{X}$ to conclude that

$$\begin{aligned}&\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta(\mathbf{Z}\mathbf{B}\mathbf{Z}^* - \mathbb{E}[\mathbf{Z}\mathbf{B}\mathbf{Z}^*])}\\ &\qquad\le \frac{\theta^2/\psi}{1-2\theta^2/\psi}\Big[\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{8\psi p\mathbf{Z}\mathbf{B}^2\mathbf{Z}^*} + 8\psi\Big(\sum_j |b_{jj}|^2\Big)\Big].\end{aligned} \tag{7.6}$$

This m.g.f. bound (7.6) is the central point in the argument. The rest of the proof consists of elementary (but messy) manipulations.

7.4. *The positive-semidefinite case.* First, we develop an m.g.f. bound for a compound sample covariance matrix based on a positive-semidefinite matrix $\mathbf{A} \succcurlyeq \mathbf{0}$. Invoke the bound (7.6) with the choice $\mathbf{B} = \mathbf{A}$, and introduce the estimate $\mathbf{A}^2 \preccurlyeq \|\mathbf{A}\|\mathbf{A}$ to reach

$$\begin{aligned}&\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta(\mathbf{Z}\mathbf{A}\mathbf{Z}^* - \mathbb{E}[\mathbf{Z}\mathbf{A}\mathbf{Z}^*])}\\ &\qquad\le \frac{\theta^2/\psi}{1-2\theta^2/\psi}\Big[\log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{8\psi p\|\mathbf{A}\|\mathbf{Z}\mathbf{A}\mathbf{Z}^*} + 8\psi\Big(\max_j a_{jj}\Big)(\mathrm{tr}\,\mathbf{A})\Big].\end{aligned}$$

Select $\psi = \theta/(8p\|\mathbf{A}\|)$, which yields

$$\log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta(\mathbf{ZAZ}^* - \mathbb{E}[\mathbf{ZAZ}^*])} \le \frac{1}{1-16p\|\mathbf{A}\|\theta}\Big[8p\|\mathbf{A}\|\theta \log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta\mathbf{ZAZ}^*} + 8\theta^2\Big(\max_j a_{jj}\Big)(\operatorname{tr}\mathbf{A})\Big].$$

Referring to the calculation (7.2), we see that

$$\log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta\mathbf{ZAZ}^*} = \log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta(\mathbf{ZAZ}^* - \mathbb{E}[\mathbf{ZAZ}^*])} + \sigma^2(\operatorname{tr}\mathbf{A})\theta.$$

Combine the last two displays, and rearrange to arrive at

$$\log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta(\mathbf{ZAZ}^* - \mathbb{E}[\mathbf{ZAZ}^*])} \le \frac{8\theta^2 \operatorname{tr}\mathbf{A}}{1-24p\|\mathbf{A}\|\theta}\Big(p\sigma^2\|\mathbf{A}\| + \max_j a_{jj}\Big). \tag{7.7}$$

At this point, we can derive probabilistic bounds for $\lambda_{\max}(\mathbf{ZAZ}^* - \mathbb{E}[\mathbf{ZAZ}^*])$ by applying Corollary 5.1.

7.5. *The general case.* To analyze the case where $\mathbf{B} \in \mathbb{H}^n$ is arbitrary, we begin once again with (7.6). To control the m.g.f. on the right-hand side, we need to center the random matrix $\mathbf{ZB}^2\mathbf{Z}^*$. Applying the calculation (7.2) with $\mathbf{B} \mapsto \mathbf{B}^2$, we obtain

$$\log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{8\psi p\mathbf{ZB}^2\mathbf{Z}^*} = \log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{8\psi p(\mathbf{ZB}^2\mathbf{Z}^* - \mathbb{E}[\mathbf{ZB}^2\mathbf{Z}^*])} + 8p\sigma^2\|\mathbf{B}\|_{\mathrm{F}}^2\psi.$$

We have used the fact that $\operatorname{tr}\mathbf{B}^2 = \|\mathbf{B}\|_{\mathrm{F}}^2$. Since $\mathbf{B}^2$ is positive semidefinite, we may introduce the bound (7.7) with $\mathbf{A} = \mathbf{B}^2$ and $\theta = 8\psi p$. This step yields

$$\log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{8\psi p\mathbf{ZB}^2\mathbf{Z}^*} \le \frac{512p^2\|\mathbf{B}\|_{\mathrm{F}}^2\|\mathbf{B}\|^2(p\sigma^2+1)\psi^2}{1-192p^2\|\mathbf{B}\|^2\psi} + 8p\sigma^2\|\mathbf{B}\|_{\mathrm{F}}^2\psi.$$

This argument relies on the estimate $\max_j(\mathbf{B}^2)_{jj} \le \|\mathbf{B}\|^2$.

Introduce the latter display into (7.6). Select $\psi = (384p^2\|\mathbf{B}\|^2)^{-1}$, and invoke the inequality $\sum_j |b_{jj}|^2 \le \|\mathbf{B}\|_{\mathrm{F}}^2$. A numerical simplification delivers

$$\begin{aligned}
\log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta(\mathbf{ZBZ}^* - \mathbb{E}[\mathbf{ZBZ}^*])} &\le \frac{11\|\mathbf{B}\|_{\mathrm{F}}^2(p\sigma^2+1)\theta^2}{1-768p^2\|\mathbf{B}\|^2\theta^2} \\
&= \frac{11\|\mathbf{B}\|_{\mathrm{F}}^2(p\sigma^2+1)\theta^2}{(1-\sqrt{768}p\|\mathbf{B}\|\theta)(1+\sqrt{768}p\|\mathbf{B}\|\theta)} \\
&\le \frac{11\|\mathbf{B}\|_{\mathrm{F}}^2(p\sigma^2+1)\theta^2}{1-\sqrt{768}p\|\mathbf{B}\|\theta}.
\end{aligned} \tag{7.8}$$

Tail and expectation bounds for the maximal eigenvalue follow from Proposition 3.4 with $v = 22\|\mathbf{B}\|_{\mathrm{F}}^2(p\sigma^2+1)$ and $c = 16\sqrt{3}p\|\mathbf{B}\|$.

The bounds for the minimum eigenvalue follow from the same argument. In this case, we must consider negative values of the parameter θ, but we can transfer the sign to the matrix $\mathbf{B}$ and proceed as before, since (7.8) remains unchanged by the transformation $\mathbf{B} \to -\mathbf{B}$. Together, the bounds on the maximum and minimum eigenvalue lead to estimates for the spectral norm.

8. Random matrices, exchangeable pairs and kernels. Now, we embark on our quest to prove the matrix Efron–Stein inequalities of Section 4. This section outlines some basic concepts from the theory of exchangeable pairs; cf. Chatterjee (2007, 2005), Stein (1972, 1986). Afterward, we explain how these ideas lead to concentration inequalities.

8.1. *Exchangeable pairs.* In our analysis, the primal concept is an exchangeable pair of random variables.

DEFINITION 8.1 (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* when it has the same distribution as the pair (Z', Z).

In particular, Z and Z' have the same distribution, and $\mathbb{E} f(Z, Z') = \mathbb{E} f(Z', Z)$ for every integrable function f.

8.2. *Kernel Stein pairs.* We are interested in a special class of exchangeable pairs of random matrices. There must be an antisymmetric bivariate kernel that "reproduces" the matrices in the pair. This approach is motivated by Chatterjee (2007).

DEFINITION 8.2 (Kernel Stein pair). Let (Z, Z') be an exchangeable pair of random variables taking values in a Polish space $\mathcal{Z}$, and let $\mathbf{\Psi} : \mathcal{Z} \to \mathbb{H}^d$ be a measurable function. Define the random Hermitian matrices

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

We say that $(\mathbf{X}, \mathbf{X}')$ is a *kernel Stein pair* if there exists a bivariate function $\mathbf{K} : \mathcal{Z}^2 \to \mathbb{H}^d$ for which

$$\mathbf{K}(z, z') = -\mathbf{K}(z', z) \qquad \text{for all } z, z' \in \mathcal{Z} \tag{8.1}$$

and

$$\mathbb{E}[\mathbf{K}(Z, Z')|Z] = \mathbf{X} \qquad \text{almost surely.} \tag{8.2}$$

When discussing a kernel Stein pair $(\mathbf{X}, \mathbf{X}')$, we assume that $\mathbb{E}\,\|\mathbf{X}\|^2 < \infty$. We sometimes write $\mathbf{K}$-*Stein pair* to emphasize the specific kernel $\mathbf{K}$.

The kernel is always centered in the sense that

$$\mathbb{E}[\mathbf{K}(Z, Z')] = \mathbf{0}. \tag{8.3}$$

Indeed, $\mathbb{E}[\mathbf{K}(Z, Z')] = -\mathbb{E}[\mathbf{K}(Z', Z)] = -\mathbb{E}[\mathbf{K}(Z, Z')]$, where the first identity follows from antisymmetry and the second follows from exchangeability.

REMARK 8.3 (Matrix Stein pairs). The analysis in [Mackey et al. (2014)] is based on a subclass of kernel Stein pairs called *matrix Stein pairs*. A matrix Stein pair $(\mathbf{X}, \mathbf{X}')$ derived from an auxiliary exchangeable pair (Z, Z') satisfies the stronger condition

$$\mathbb{E}[\mathbf{X} - \mathbf{X}' | Z] = \alpha \mathbf{X} \qquad \text{for some } \alpha > 0. \tag{8.4}$$

That is, a matrix Stein pair is a kernel Stein pair with $\mathbf{K}(Z, Z') = \alpha^{-1}(\mathbf{X} - \mathbf{X}')$. Although Mackey et al. (2014) describe several classes of matrix Stein pairs, most exchangeable pairs of random matrices do not satisfy (8.4). Kernel Stein pairs are more common, so they are commensurately more useful.

8.3. *The method of exchangeable pairs*. Kernel Stein pairs are valuable because they offer a powerful tool for evaluating moments of a random matrix. We express this claim in a fundamental technical lemma, which generalizes both Chatterjee (2007), equation (6) and Mackey et al. (2014), Lemma 2.3.

LEMMA 8.4 (Method of exchangeable pairs). *Suppose that* $(\mathbf{X}, \mathbf{X}') \in \mathbb{H}^d \times \mathbb{H}^d$ *is a* $\mathbf{K}$*-Stein pair constructed from an auxiliary exchangeable pair* $(Z, Z') \in \mathcal{Z}^2$. *Let* $\mathbf{F} : \mathbb{H}^d \to \mathbb{H}^d$ *be a measurable function that satisfies the regularity condition*

$$\mathbb{E}\|\mathbf{K}(Z, Z')\mathbf{F}(\mathbf{X})\| < \infty. \tag{8.5}$$

Then

$$\mathbb{E}[\mathbf{X}\mathbf{F}(\mathbf{X})] = \tfrac{1}{2}\,\mathbb{E}[\mathbf{K}(Z, Z')(\mathbf{F}(\mathbf{X}) - \mathbf{F}(\mathbf{X}'))]. \tag{8.6}$$

PROOF. Definition 8.2, of a kernel Stein pair, implies that

$$\mathbb{E}[\mathbf{X}\mathbf{F}(\mathbf{X})] = \mathbb{E}[\mathbb{E}[\mathbf{K}(Z, Z')|Z]\mathbf{F}(\mathbf{X})] = \mathbb{E}[\mathbf{K}(Z, Z')\mathbf{F}(\mathbf{X})],$$

where we justify the pull-through property of conditional expectation using the regularity condition (8.5). The antisymmetry (8.1) of the kernel $\mathbf{K}$ delivers the relation

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

Average the two preceding displays to reach the identity (8.6). □

Lemma 8.4 has several immediate consequences for the structure of a $\mathbf{K}$-Stein pair $(\mathbf{X}, \mathbf{X}')$ constructed from an auxiliary exchangeable pair (Z, Z'). First, the matrix $\mathbf{X}$ must be centered:

$$\mathbb{E}\,\mathbf{X} = \mathbf{0}. \tag{8.7}$$

This result follows from the choice $\mathbf{F}(\mathbf{X}) = \mathbf{I}$.

Second, we can develop a bound for the variance of the random matrix $\mathbf{X}$. Since $\mathbf{X}$ is centered,

$$\mathrm{Var}[\mathbf{X}] = \mathbb{E}[\mathbf{X}^2] = \tfrac{1}{2}\,\mathbb{E}[\mathrm{Re}(\mathbf{K}(Z, Z')(\mathbf{X} - \mathbf{X}'))].$$

This claim follows when we apply Lemma 8.4 with $\mathbf{F}(\mathbf{X}) = \mathbf{X}$ and extract the real part (2.3) of the result. Invoke the matrix inequality (2.5) to obtain

$$\mathrm{Var}[\mathbf{X}] \preccurlyeq \tfrac{1}{4}\,\mathbb{E}[\mathbf{K}(Z, Z')^2 + (\mathbf{X} - \mathbf{X}')^2]. \tag{8.8}$$

In other words, we can obtain bounds for the variance in terms of the variance of the kernel $\mathbf{K}$ and the variance of $\mathbf{X} - \mathbf{X}'$.

8.4. *Conditional variances.* To each kernel Stein pair $(\mathbf{X}, \mathbf{X}')$, we may associate two random matrices called the *conditional variance* and *kernel conditional variance* of $\mathbf{X}$. We will see that $\mathbf{X}$ is concentrated around the zero matrix whenever the conditional variance and the kernel conditional variance are both small.

DEFINITION 8.5 (Conditional variances). Suppose that $(\mathbf{X}, \mathbf{X}')$ is a $\mathbf{K}$-Stein pair, constructed from an auxiliary exchangeable pair (Z, Z'). The *conditional variance* is the random matrix

$$\mathbf{V}_{\mathbf{X}} := \tfrac{1}{2}\,\mathbb{E}[(\mathbf{X} - \mathbf{X}')^2 | Z], \tag{8.9}$$

and the *kernel conditional variance* is the random matrix

$$\mathbf{V}^{\mathbf{K}} := \tfrac{1}{2}\,\mathbb{E}[\mathbf{K}(Z, Z')^2 | Z]. \tag{8.10}$$

Because of the bound (8.8), the conditional variances satisfy

$$\mathrm{Var}[\mathbf{X}] \preccurlyeq \tfrac{1}{2}\,\mathbb{E}[\mathbf{V}_{\mathbf{X}} + \mathbf{V}^{\mathbf{K}}],$$

so it is natural to seek concentration results stated in terms of these quantities.

9. Polynomial moments of a random matrix.

We begin by developing a polynomial moment bound for a kernel Stein pair. This result shows that we can control the expectation of the Schatten p-norm in terms of the conditional variance and the kernel conditional variance.

THEOREM 9.1 (Polynomial moments for a kernel Stein pair). *Let $(\mathbf{X}, \mathbf{X}')$ be a $\mathbf{K}$-Stein pair based on an auxiliary exchangeable pair (Z, Z'). For a natural number $p \geq 1$, assume the regularity conditions*

$$\mathbb{E}\,\|\mathbf{X}\|_{S_{2p}}^{2p} < \infty \quad \text{and} \quad \mathbb{E}\|\mathbf{K}(Z, Z')\|^{2p} < \infty.$$

Then, for each $s > 0$,

$$(\mathbb{E}\,\|\mathbf{X}\|_{S_{2p}}^{2p})^{1/(2p)} \leq \sqrt{2p-1}\,\big(\mathbb{E}\|\tfrac{1}{2}(s\mathbf{V}_{\mathbf{X}} + s^{-1}\mathbf{V}^{\mathbf{K}})\|_{S_p}^{p}\big)^{1/(2p)}.$$

The symbol $\|\cdot\|_{S_p}$ refers to the Schatten p-norm (2.7), and the conditional variances $\mathbf{V}_{\mathbf{X}}$ and $\mathbf{V}^{\mathbf{K}}$ are defined in (8.9) and (8.10).

We establish this result, which holds equally for infinite dimensional operators $\mathbf{X}$, in the remainder of this section. The pattern of argument is similar to the proofs of Chatterjee (2005), Theorem 3.14, and Mackey et al. (2014), Theorem 7.1, but we require a nontrivial new matrix inequality.

9.1. *The polynomial mean value trace inequality.* The main new ingredient in the proof of Theorem 9.1 is the following matrix trace inequality.

LEMMA 9.2 (Polynomial mean value trace inequality). *For all matrices* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C} \in \mathbb{H}^d$, *all integers* $q \geq 1$, *and all* $s > 0$, *it holds that*

$$\left|\operatorname{tr}\left[\mathbf{C}(\mathbf{A}^q - \mathbf{B}^q)\right]\right| \leq \frac{q}{4} \operatorname{tr}\left[\left(s(\mathbf{A}-\mathbf{B})^2 + s^{-1}\mathbf{C}^2\right)\left(|\mathbf{A}|^{q-1} + |\mathbf{B}|^{q-1}\right)\right].$$

Lemma 9.2 improves on the estimate [Mackey et al. (2014), Lemma 3.4], which drives concentration inequalities for matrix Stein pairs. Since the result does not have any probabilistic content, we defer the proof until Appendix B.

9.2. *Proof of Theorem* 9.1. The argument follows the same lines as the proof of Mackey et al. (2014), Theorem 7.1, so we pass lightly over certain details. Let us examine the quantity of interest:

$$E := \mathbb{E}\, \|\mathbf{X}\|_{S_{2p}}^{2p} = \mathbb{E} \operatorname{tr} |\mathbf{X}|^{2p} = \mathbb{E} \operatorname{tr}\left[\mathbf{X} \cdot \mathbf{X}^{2p-1}\right],$$

where $\cdot$ denotes the usual matrix product. To apply the method of exchangeable pairs, Lemma 8.4, we first check the regularity condition (8.5):

$$\begin{aligned}\mathbb{E}\left\|\mathbf{K}(Z, Z') \cdot \mathbf{X}^{2p-1}\right\| &\leq \mathbb{E}\left(\left\|\mathbf{K}(Z, Z')\right\| \|\mathbf{X}\|^{2p-1}\right) \\ &\leq \left(\mathbb{E}\left\|\mathbf{K}(Z, Z')\right\|^{2p}\right)^{1/(2p)} \left(\mathbb{E}\, \|\mathbf{X}\|^{2p}\right)^{(2p-1)/(2p)} < \infty,\end{aligned}$$

where we have applied Hölder's inequality for expectation and the fact that the spectral norm is dominated by the Schatten $2p$-norm. Thus, we may invoke Lemma 8.4 with $\mathbf{F}(\mathbf{X}) = \mathbf{X}^{2p-1}$ to reach

$$E = \tfrac{1}{2}\, \mathbb{E} \operatorname{tr}\left[\mathbf{K}(Z, Z') \cdot \left(\mathbf{X}^{2p-1} - (\mathbf{X}')^{2p-1}\right)\right].$$

Next, fix a parameter $s > 0$. Apply the polynomial mean value trace inequality, Lemma 9.2, with $q = 2p - 1$ to obtain the estimate

$$\begin{aligned}E &\leq \frac{2p-1}{8}\, \mathbb{E} \operatorname{tr}\left[\left(s(\mathbf{X}-\mathbf{X}')^2 + s^{-1}\mathbf{K}(Z, Z')^2\right) \cdot \left(\mathbf{X}^{2p-2} + (\mathbf{X}')^{2p-2}\right)\right] \\ &= \frac{2p-1}{4}\, \mathbb{E} \operatorname{tr}\left[\left(s(\mathbf{X}-\mathbf{X}')^2 + s^{-1}\mathbf{K}(Z, Z')^2\right) \cdot \mathbf{X}^{2p-2}\right] \\ &= (2p-1)\, \mathbb{E} \operatorname{tr}\left[\frac{1}{2}\left(s\mathbf{V}_{\mathbf{X}} + s^{-1}\mathbf{V}^{\mathbf{K}}\right) \cdot \mathbf{X}^{2p-2}\right].\end{aligned}$$

The second line follows from the fact that $(\mathbf{X}, \mathbf{X}')$ is an exchangeable pair, and the third line depends on the definitions (8.9) and (8.10) of the conditional variances. We have used the regularity condition $\mathbb{E}\,\|\mathbf{X}\|_{S_{2p}}^{2p} < \infty$ to justify the pull-through property of conditional expectation.

Now, we apply Hölder's inequality for the trace followed by Hölder's inequality for the expectation. These steps yield

$$\begin{aligned} E &\le (2p-1)\big(\mathbb{E}\big\|\tfrac{1}{2}\big(s\mathbf{V}_{\mathbf{X}} + s^{-1}\mathbf{V}^{\mathbf{K}}\big)\big\|_{S_p}^{p}\big)^{1/p}\big(\mathbb{E}\,\|\mathbf{X}\|_{S_{2p}}^{2p}\big)^{(p-1)/p} \\ &= (2p-1)\big(\mathbb{E}\big\|\tfrac{1}{2}\big(s\mathbf{V}_{\mathbf{X}} + s^{-1}\mathbf{V}^{\mathbf{K}}\big)\big\|_{S_p}^{p}\big)^{1/p} E^{(p-1)/p}. \end{aligned}$$

Solve this algebraic identity for E to determine that

$$E^{1/(2p)} \le \sqrt{2p-1}\big(\mathbb{E}\big\|\tfrac{1}{2}\big(s\mathbf{V}_{\mathbf{X}} + s^{-1}\mathbf{V}^{\mathbf{K}}\big)\big\|_{S_p}\big)^{1/(2p)}.$$

This completes the proof of Theorem 9.1.

10. Constructing a kernel via Markov chain coupling. Theorem 9.1 is one of the main steps toward the polynomial Efron–Stein inequality for random matrices. To reach the latter result, we need to develop an explicit construction for the kernel Stein pair along with concrete bounds for the conditional variance. We present this material in the current section, and we establish the Efron–Stein bound in Section 11. The analysis leading to exponential concentration inequalities is somewhat more involved. We postpone these results until Section 12.

10.1. *Overview.* For a random matrix $\mathbf{X}$ that is presented as part of a kernel Stein pair, Theorem 9.1 provides strong bounds on the polynomial moments in terms of the conditional variances. To make this result effective, we need to address several more questions.

First, given an exchangeable pair of random matrices, we can ask whether it is possible to equip the pair with a kernel that satisfies (8.2). In fact, there is a general construction that works whenever the exchangeable pair is suitably ergodic. This method depends on an idea [Chatterjee (2005), Section 4.1] that ultimately relies on an observation of Stein; cf. Stein (1986). We describe this approach in Sections 10.2 and 10.3.

Second, we can ask whether there is a mechanism for bounding the conditional variances in terms of simpler quantities. We have developed some new tools for performing these estimates. These methods appear in Sections 10.4 and 10.5.

10.2. *Kernel couplings.* Stein noticed that each exchangeable pair (Z, Z') of $\mathcal{Z}$-valued random variables yields a reversible Markov chain with a symmetric transition kernel P given by

$$Pf(z) := \mathbb{E}[f(Z')|Z = z]$$

for each function $f : \mathcal{Z} \to \mathbb{R}$ that satisfies $\mathbb{E}|f(Z)| < \infty$. In other words, for any initial value $Z_{(0)} \in \mathcal{Z}$, we can construct a Markov chain

$$Z_{(0)} \to Z_{(1)} \to Z_{(2)} \to Z_{(3)} \to \cdots,$$

where $\mathbb{E}[f(Z_{(i+1)})|Z_{(i)}] = Pf(Z_{(i)})$ for each integrable function f. This requirement suffices to determine the distribution of each $Z_{(i+1)}$.

When the chain $(Z_{(i)})_{i\geq 0}$ is ergodic enough, we can explicitly construct a kernel that satisfies (8.2) for any exchangeable pair of random matrices constructed from the auxiliary exchangeable pair (Z, Z'). To explain this idea, we adapt a definition from Chatterjee (2005), Section 4.1.

DEFINITION 10.1 (Kernel coupling). Let $(Z, Z') \in \mathcal{Z}^2$ be an exchangeable pair. Let $(Z_{(i)})_{i\geq 0}$ and $(Z'_{(i)})_{i\geq 0}$ be two Markov chains with arbitrary initial values, each evolving according to the transition kernel P induced by (Z, Z'). We call $(Z_{(i)}, Z'_{(i)})_{i\geq 0}$ a *kernel coupling* for (Z, Z') if

$$Z_{(i)} \perp\!\!\!\perp Z'_{(0)}|Z_{(0)} \quad \text{and} \quad Z'_{(i)} \perp\!\!\!\perp Z_{(0)}|Z'_{(0)} \qquad \text{for all } i. \tag{10.1}$$

The notation $U \perp\!\!\!\perp V|W$ means U and V are independent conditional on W.

For an example of kernel coupling, consider the simple random walk on the hypercube $\{\pm 1\}^n$ where two vertices are neighbors when they differ in exactly one coordinate. We can start two random walks at two different locations on the cube. At each step, we select a uniformly random coordinate from $\{1, \dots, n\}$ and a uniformly random value from $\{\pm 1\}$. We update *both* of the walks by replacing the *same* chosen coordinate with the *same* chosen value. The two walks arrive at the same vertex (i.e., they *couple*) as soon as we have updated each coordinate at least once.

10.3. *Kernel Stein pairs from the Poisson equation.* Chatterjee (2005), Section 4.1, observed that it is often possible to construct a kernel coupling by solving the Poisson equation for the Markov chain with transition kernel P.

PROPOSITION 10.2. *Let $(Z_{(i)}, Z'_{(i)})_{i\geq 0}$ be a kernel coupling for an exchangeable pair $(Z, Z') \in \mathcal{Z}^2$. Let $\mathbf{\Psi} : \mathcal{Z} \to \mathbb{H}^d$ be a bounded, measurable function with $\mathbb{E}\,\mathbf{\Psi}(Z) = \mathbf{0}$. Suppose there is a positive constant L for which*

$$\sum_{i=0}^{\infty} \|\mathbb{E}[\mathbf{\Psi}(Z_{(i)}) - \mathbf{\Psi}(Z'_{(i)})|Z_{(0)} = z, Z'_{(0)} = z']\| \leq L \qquad \text{for all } z, z' \in \mathcal{Z}. \tag{10.2}$$

Then $(\mathbf{\Psi}(Z), \mathbf{\Psi}(Z'))$ is a $\mathbf{K}$-Stein pair with kernel

$$\mathbf{K}(z, z') := \sum_{i=0}^{\infty} \mathbb{E}[\mathbf{\Psi}(Z_{(i)}) - \mathbf{\Psi}(Z'_{(i)})|Z_{(0)} = z, Z'_{(0)} = z']. \tag{10.3}$$

The proof of this result is identical with that of Chatterjee (2005), Lemmas 4.1 and 4.2, which establishes Proposition 10.2 in the scalar setting. We omit the details.

REMARK 10.3 (Regularity). Proposition 10.2 holds for functions $\mathbf{\Psi}$ that satisfy conditions weaker than boundedness. We focus on the simplest case to reduce the technical burden.

10.4. *Bounding the conditional variances I.* The construction described in Proposition 10.2 is valuable because it leads to an explicit description of the kernel. In many examples, this formula allows us to develop a succinct bound on the conditional variances. We encapsulate the required calculations in a technical lemma.

LEMMA 10.4. *Instate the notation and hypotheses of Proposition* 10.2, *and define the kernel Stein pair* $(\mathbf{X}, \mathbf{X}') = (\mathbf{\Psi}(Z), \mathbf{\Psi}(Z'))$. *For each* $i = 0, 1, 2, \ldots,$ *assume that*

$$\mathbb{E}\big[\big(\mathbb{E}[\mathbf{\Psi}(Z_{(i)}) - \mathbf{\Psi}(Z'_{(i)})|Z_{(0)} = Z, Z'_{(0)} = Z']\big)^2|Z\big] \preccurlyeq \beta_i^2 \mathbf{\Gamma}_i, \tag{10.4}$$

where β_i *is a nonnegative number and* $\mathbf{\Gamma}_i \in \mathbb{H}^d$ *is a random matrix. Then the conditional variance* (8.9) *and kernel conditional variance* (8.10) *satisfy*

$$\mathbf{V_X} \preccurlyeq \frac{1}{2}\beta_0^2 \mathbf{\Gamma}_0 \quad \textit{and} \quad \mathbf{V^K} \preccurlyeq \frac{1}{2}\left(\sum_{j=0}^{\infty} \beta_j\right) \sum_{i=0}^{\infty} \beta_i \mathbf{\Gamma}_i.$$

PROOF. By a continuity argument, we may assume that $\beta_i > 0$ for each index i. Write

$$\mathbf{Y}_i := \mathbb{E}[\mathbf{\Psi}(Z_{(i)}) - \mathbf{\Psi}(Z'_{(i)})|Z_{(0)} = Z, Z'_{(0)} = Z'].$$

The definition (8.9) of the conditional variance $\mathbf{V_X}$ immediately implies

$$\mathbf{V_X} = \tfrac{1}{2}\,\mathbb{E}\big[(\mathbf{X} - \mathbf{X}')^2|Z\big] = \tfrac{1}{2}\,\mathbb{E}[\mathbf{Y}_0^2|Z] \preccurlyeq \tfrac{1}{2}\beta_0^2 \mathbf{\Gamma}_0.$$

The semidefinite relation follows from the hypothesis (10.4).

According to the definition (8.10) of the kernel conditional variance $\mathbf{V^K}$ and the kernel construction (10.3), we have

$$\begin{aligned}\mathbf{V^K} &= \frac{1}{2}\,\mathbb{E}[\mathbf{K}(Z, Z')|Z] = \frac{1}{2}\,\mathbb{E}\left[\left(\sum_{i=0}^{\infty} \mathbf{Y}_i\right)^2 \Big| Z\right] \\ &= \frac{1}{2}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \mathbb{E}[\mathrm{Re}(\mathbf{Y}_i \mathbf{Y}_j)|Z].\end{aligned}$$

The semidefinite bound (2.5) for the real part of a product implies that

$$\begin{aligned}\mathbf{V}^{\mathbf{K}} &\preccurlyeq \frac{1}{2}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{1}{2}\left(\frac{\beta_j}{\beta_i}\mathbb{E}[\mathbf{Y}_i^2|Z] + \frac{\beta_i}{\beta_j}\mathbb{E}[\mathbf{Y}_j^2|Z]\right)\\ &\preccurlyeq \frac{1}{2}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{1}{2}\left(\frac{\beta_j}{\beta_i}\beta_i^2\mathbf{\Gamma}_i + \frac{\beta_i}{\beta_j}\beta_j^2\mathbf{\Gamma}_j\right)\\ &= \frac{1}{2}\left(\sum_{j=0}^{\infty}\beta_j\right)\sum_{i=0}^{\infty}\beta_i\mathbf{\Gamma}_i.\end{aligned}$$

The second relation depends on the hypothesis (10.4). □

10.5. *Bounding the conditional variances II.* The random matrices $\mathbf{\Gamma}_i$ that arise in Lemma 10.4 often share a common form. We can use this property to obtain a succinct bound for the conditional variance expression that appears in Theorem 9.1. This reduction allows us to establish Efron–Stein inequalities.

LEMMA 10.5. *Instate the notation and hypotheses of Lemma 10.4. Suppose*

$$\mathbf{\Gamma}_i = \mathbb{E}[\mathbf{W}_{(i)}|Z] \qquad \text{where } \mathbf{W}_{(i)} \stackrel{d}{=} \mathbf{\Gamma}_0 \qquad \text{for each } i \geq 1. \tag{10.5}$$

Then, for each increasing and convex function $f : \mathbb{R}_+ \to \mathbb{R}$,

$$\mathbb{E}\operatorname{tr} f\left(\beta_0^{-2}\mathbf{V}_{\mathbf{X}} + \left(\sum_{i=0}^{\infty}\beta_i\right)^{-2}\mathbf{V}^{\mathbf{K}}\right) \leq \mathbb{E}\operatorname{tr} f(\mathbf{\Gamma}_0).$$

PROOF. Abbreviate $B = \sum_{i=0}^{\infty}\beta_i$. Lemma 10.4 provides that

$$\mathbf{V}_{\mathbf{X}} \preccurlyeq \frac{1}{2}\beta_0^2\mathbf{\Gamma}_0 \quad \text{and} \quad \mathbf{V}^{\mathbf{K}} \preccurlyeq \frac{1}{2}B\sum_{i=0}^{\infty}\beta_i\mathbf{\Gamma}_i.$$

Since f is increasing and convex on $\mathbb{R}_+$, the function $\operatorname{tr} f : \mathbb{H}_+^d \to \mathbb{R}$ is increasing (2.1) and convex (2.2). Therefore,

$$\begin{aligned}\mathbb{E}\operatorname{tr} f\big(\beta_0^{-2}\mathbf{V}_{\mathbf{X}} + B^{-2}\mathbf{V}^{\mathbf{K}}\big) &\leq \mathbb{E}\operatorname{tr} f\left(\frac{1}{2}\mathbf{\Gamma}_0 + \frac{1}{2B}\sum_{i=0}^{\infty}\beta_i\mathbf{\Gamma}_i\right)\\ &\leq \frac{1}{2}\mathbb{E}\operatorname{tr} f(\mathbf{\Gamma}_0) + \frac{1}{2B}\sum_{i=0}^{\infty}\beta_i\,\mathbb{E}\operatorname{tr} f(\mathbf{\Gamma}_i).\end{aligned}$$

In view of (10.5), Jensen's inequality and the tower property together yield

$$\mathbb{E}\operatorname{tr} f(\mathbf{\Gamma}_i) = \mathbb{E}\operatorname{tr} f\big(\mathbb{E}[\mathbf{W}_{(i)}|Z]\big) \leq \mathbb{E}\operatorname{tr} f(\mathbf{W}_{(i)}) = \mathbb{E}\operatorname{tr} f(\mathbf{\Gamma}_0).$$

Combine the latter two displays to complete the argument. □

11. The polynomial Efron–Stein inequality for a random matrix. We are now prepared to establish the polynomial Efron–Stein inequality, Theorem 4.2. We retain the notation and hypotheses from Section 4.1, and we encourage the reader to review this material before continuing. The proof is divided into two parts. First, we assume that the random matrix is bounded so that the kernel coupling tools apply. Then we use a truncation argument to remove the boundedness assumption.

11.1. *A kernel coupling for a vector of independent variables.* We begin with the construction of an exchangeable pair. Recall that $Z := (Z_1, \dots, Z_n) \in \mathcal{Z}$ is a vector of mutually independent random variables. For each coordinate j,

$$Z^{(j)} := (Z_1, \dots, \widetilde{Z}_j, \dots, Z_n) \in \mathcal{Z},$$

where $\widetilde{Z}_j$ is an independent copy of Z_j. Form the random vector

$$Z' := Z^{(J)} \qquad \text{where } J \sim \text{UNIFORM}\{1, \dots, n\}. \tag{11.1}$$

We may assume that J is drawn independently from Z. It follows that (Z, Z') is exchangeable.

Next, we build an explicit kernel coupling $(Z_{(i)}, Z'_{(i)})_{i\geq 0}$ for the exchangeable pair (Z, Z'). The Markov chains may take arbitrary initial values $Z_{(0)}$ and $Z'_{(0)}$. For each time $i \geq 1$, we let both chains evolve via the same random choice:

1. Independent of prior choices, draw a coordinate $J_i \sim \text{UNIFORM}\{1, \dots, n\}$.
2. Draw an independent copy $\widetilde{Z}_{(i)}$ of Z.
3. Form $Z_{(i)}$ by replicating $Z_{(i-1)}$ and then replacing the J_ith coordinate with the J_ith coordinate of $\widetilde{Z}_{(i)}$.
4. Form $Z'_{(i)}$ by replicating $Z'_{(i-1)}$ and then replacing the J_ith coordinate with the J_ith coordinate of $\widetilde{Z}_{(i)}$.

By construction, $(Z_{(i)}, Z'_{(i)})_{i\geq 0}$ satisfies the kernel coupling property (10.1). This coupling is drawn from Chatterjee (2005), Section 4.1. Note that this is just a glorification of the hypercube example in Section 10.2.

11.2. *A kernel Stein pair.* Let $\mathbf{H} : \mathcal{Z} \to \mathbb{H}^d$ be a bounded, measurable function. Construct the random matrices

$$\mathbf{X} := \mathbf{H}(Z) - \mathbb{E}\,\mathbf{H}(Z) \quad \text{and} \quad \mathbf{X}' := \mathbf{H}(Z') - \mathbb{E}\,\mathbf{H}(Z). \tag{11.2}$$

To verify that $(\mathbf{X}, \mathbf{X}')$ is a kernel Stein pair, we use Lemma 10.2 to construct a kernel. For all $z, z' \in \mathcal{Z}$,

$$\mathbf{K}(z, z') := \sum_{i=0}^{\infty} \mathbb{E}\big[\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z'_{(i)}) | Z_{(0)} = z, Z'_{(0)} = z'\big]. \tag{11.3}$$

To verify the regularity condition for the lemma, notice that the two chains couple as soon as we have refreshed all n coordinates. According to the analysis of the coupon collector problem [Levin, Peres and Wilmer (2009), Section 2.2], the expected coupling time is bounded by $n(1 + \log n)$. Since $\|\mathbf{H}(Z)\|$ is bounded, the hypothesis (10.2) is in force.

11.3. *The evolution of the kernel coupling*. Draw a realization (Z, Z') of the exchangeable pair, and write J for the coordinate where Z and Z' differ. Let $(Z_{(i)}, Z'_{(i)})_{i\geq 0}$ be the kernel coupling described in the last section, starting at $Z_{(0)} = Z$ and $Z'_{(0)} = Z'$. Therefore, the initial value of the kernel coupling is a pair of vectors that differ in precisely one coordinate. Because of the coupling construction,

$$\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z'_{(i)}) = \big(\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z'_{(i)})\big) \cdot \mathbb{1}[J \notin \{J_1, \dots, J_i\}].$$

The operator Schwarz inequality [Bhatia (2007), equation (3.19)] implies that

$$\begin{aligned}
&\big(\mathbb{E}[\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z'_{(i)})|Z, Z']\big)^2 \\
&\quad = \big(\mathbb{E}\big[\big(\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z'_{(i)})\big) \cdot \mathbb{1}[J \notin \{J_1, \dots, J_i\}]|Z, Z'\big]\big)^2 \\
&\quad \preccurlyeq \mathbb{E}\big[\big(\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z'_{(i)})\big)^2|Z, Z'\big] \cdot \mathbb{E}\big[\mathbb{1}[J \notin \{J_1, \dots, J_i\}]|Z, Z'\big] \\
&\quad = (1 - 1/n)^i \cdot \mathbb{E}\big[\big(\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z'_{(i)})\big)^2|Z, Z'\big].
\end{aligned} \tag{11.4}$$

Take the conditional expectation with respect to Z, and invoke the tower property to reach

$$\begin{aligned}
&\mathbb{E}\big[\big(\mathbb{E}[\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z'_{(i)})|Z, Z']\big)^2|Z\big] \\
&\qquad \preccurlyeq (1 - 1/n)^i \cdot \mathbb{E}\big[\big(\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z'_{(i)})\big)^2|Z\big].
\end{aligned} \tag{11.5}$$

11.4. *Conditional variance bounds*. To obtain a bound for the expression (11.5) that satisfies the prerequisites of Lemma 10.5, we will replace $Z'_{(i)}$ with a variable $Z^*_{(i)}$ that satisfies

$$\big(Z_{(i)}, Z^*_{(i)}\big) \stackrel{d}{=} \big(Z, Z'\big) \quad \text{and} \quad Z^*_{(i)} \perp\!\!\!\perp Z|Z_{(i)}.$$

For $i \geq 0$, define $Z^*_{(i)}$ as being equal to $Z_{(i)}$ everywhere except in coordinate J, where it equals Z'_J. Since $(J, Z'_J) \perp\!\!\!\perp Z|Z_{(i)}$, we have our desired conditional independence. Moreover, this definition ensures that $Z^*_{(i)} = Z'_{(i)}$ whenever $J \notin \{J_1, \dots, J_i\}$. Therefore,

$$\mathbb{E}\big[\big(\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z'_{(i)})\big)^2|Z\big] \preccurlyeq \mathbb{E}\big[\big(\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z^*_{(i)})\big)^2|Z\big].$$

Consequently, the hypothesis (10.4) of Lemma 10.4 is valid with

$$\mathbf{\Gamma}_i := \mathbb{E}\big[\big(\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z^*_{(i)})\big)^2|Z\big] \tag{11.6}$$

and $\beta_i := (1 - 1/n)^{i/2}$.

Now, let us have a closer look at the form of $\mathbf{\Gamma}_i$. The tower property and conditional independence of $(Z^*_{(i)}, Z)$ imply that

$$\begin{aligned}
\mathbf{\Gamma}_i &= \mathbb{E}\big[\mathbb{E}\big[\big(\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z^*_{(i)})\big)^2|Z_{(i)}, Z\big]|Z\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\big(\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z^*_{(i)})\big)^2|Z_{(i)}\big]|Z\big].
\end{aligned}$$

Since

$$\mathbf{\Gamma}_0 = \mathbb{E}\big[\big(\mathbf{H}(Z) - \mathbf{H}(Z')\big)^2 | Z\big], \tag{11.7}$$

we can express the latter observation as

$$\mathbf{\Gamma}_i = \mathbb{E}[\mathbf{W}_{(i)} | Z] \qquad \text{where } \mathbf{W}_{(i)} \stackrel{d}{=} \mathbf{\Gamma}_0$$

by setting

$$\mathbf{W}_{(i)} := \mathbb{E}\big[\big(\mathbf{H}(Z_{(i)}) - \mathbf{H}(Z^*_{(i)})\big)^2 | Z_{(i)}\big].$$

This is the second hypothesis required by Lemma 10.5.

11.5. *The polynomial Efron–Stein inequality*: *Bounded case*. We are prepared to prove the polynomial Efron–Stein inequality, Theorem 4.2, for a *bounded* random matrix $\mathbf{X}$ of the form (11.2).

Let p be a natural number. Since $(\mathbf{X}, \mathbf{X}')$ is a kernel Stein pair, Theorem 9.1 provides that for any $s > 0$,

$$\big(\mathbb{E}\,\|\mathbf{X}\|_{S_{2p}}^{2p}\big)^{1/(2p)} \le \sqrt{2p-1}\big(\mathbb{E}\big\|\tfrac{1}{2}\big(s\mathbf{V}_{\mathbf{X}} + s^{-1}\mathbf{V}^{\mathbf{K}}\big)\big\|_{S_p}^{p}\big)^{1/(2p)}. \tag{11.8}$$

The regularity condition holds because both the random matrix $\mathbf{X}$ and the kernel $\mathbf{K}$ are bounded.

Rewrite the Schatten p-norm in terms of the trace:

$$\mathbb{E}\left\|\frac{1}{2}\big(s\mathbf{V}_{\mathbf{X}} + s^{-1}\mathbf{V}^{\mathbf{K}}\big)\right\|_{S_p}^{p} = \mathbb{E}\operatorname{tr}\left[\frac{s}{2}\big(\mathbf{V}_{\mathbf{X}} + s^{-2}\mathbf{V}^{\mathbf{K}}\big)\right]^{p}. \tag{11.9}$$

This expression has the form required by Lemma 10.5. Indeed, the function $t \mapsto (st/2)^p$ is increasing and convex on $\mathbb{R}_+$. Furthermore, we may choose $\beta_0 = 1$ and

$$s := \sum_{i=0}^{\infty} \beta_i = \left(1 - \left(1 - \frac{1}{n}\right)^{-1/2}\right)^{-1} < 2n.$$

Lemma 10.5 now delivers the bound

$$\mathbb{E}\operatorname{tr}\left[\frac{s}{2}\big(\mathbf{V}_{\mathbf{X}} + s^{-2}\mathbf{V}^{\mathbf{K}}\big)\right]^{p} \le \mathbb{E}\operatorname{tr}\left[\frac{1}{2}s\mathbf{\Gamma}_0\right]^{p} \le \mathbb{E}\left\|2 \cdot \frac{n}{2}\mathbf{\Gamma}_0\right\|_{S_p}^{p}. \tag{11.10}$$

Next, we observe that the random matrix $\frac{1}{2}n\mathbf{\Gamma}_0$ coincides with the variance proxy $\mathbf{V}$ defined in (4.2). Indeed,

$$\begin{aligned} \frac{1}{2}n\mathbf{\Gamma}_0 &= \frac{1}{2}n\,\mathbb{E}\big[\big(\mathbf{H}(Z) - \mathbf{H}(Z')\big)^2 | Z\big] \\ &= \frac{1}{2}\sum_{j=1}^{n} \mathbb{E}\big[\big(\mathbf{H}(Z) - \mathbf{H}(Z^{(j)})\big)^2 | Z\big] \qquad (11.11) \\ &= \frac{1}{2}\sum_{j=1}^{n} \mathbb{E}\big[\big(\mathbf{X} - \mathbf{X}^{(j)}\big)^2 | Z\big] = \mathbf{V}. \end{aligned}$$

The first identity is (11.7). The second follows from the definition (11.1) of Z'. The last line harks back to the definition (4.1) of $\mathbf{X}^{(j)}$ and the variance proxy (4.2).

Sequence the displays (11.8)–(11.11) to reach

$$\left(\mathbb{E}\,\|\mathbf{X}\|_{S_{2p}}^{2p}\right)^{1/(2p)} \le \sqrt{2(2p-1)}\left(\mathbb{E}\,\|\mathbf{V}\|_{S_p}^{p}\right)^{1/(2p)} \qquad \text{when } \mathbf{X} \text{ is bounded.} \tag{11.12}$$

This complete the proof of Theorem 4.2 under the assumption that $\mathbf{X}$ is bounded.

11.6. *The polynomial Efron–Stein inequality*: *General case*. Finally, we establish Theorem 4.2 by removing the stipulation that $\mathbf{X}$ is bounded from (11.12). Fix any truncation level $R > 0$, and introduce the truncation function $\varphi_R : \mathbb{C} \to \mathbb{C}$ as

$$\varphi_R(z) := z \cdot \mathbb{1}\big[|z| \le R\big] + R \cdot \frac{z}{|z|} \cdot \mathbb{1}\big[|z| > R\big]. \tag{11.13}$$

By convention, $\varphi_R(0) = 0$. It is straightforward to show that this function satisfies the contraction property

$$\big|\varphi_R(z) - \varphi_R(w)\big| \le |z - w| \qquad \text{for any } z, w \in \mathbb{C} \text{ and } R \in \mathbb{R}_+. \tag{11.14}$$

We further define the truncated random matrix $\mathbf{\Psi}_R(Z)$ by applying the function φ_R element-wise on each element of the matrix $\mathbf{\Psi}(Z)$, and we set

$$\mathbf{X}_R := \mathbf{\Psi}_R(Z) - \mathbb{E}\big(\mathbf{\Psi}_R(Z)\big).$$

Finally, we let

$$\mathbf{V}_R := \frac{1}{2}\sum_{j=1}^{n} \mathbb{E}\big[\big(\mathbf{X}_R - \mathbf{X}_R^{(j)}\big)^2 | Z\big].$$

Since $\mathbf{X}_R$ is bounded, the results of Section 11.5 imply that

$$\left(\mathbb{E}\,\|\mathbf{X}_R\|_{S_{2p}}^{2p}\right)^{1/(2p)} \le \sqrt{2(2p-1)}\left(\mathbb{E}\,\|\mathbf{V}_R\|_{S_p}^{p}\right)^{1/(2p)}. \tag{11.15}$$

Moreover, the contraction property (11.14) implies that each element of $\mathbf{V}_R$ is no larger than the corresponding element of $\mathbf{V}$. Hence, $\operatorname{tr}(\mathbf{V}_R) \le \operatorname{tr}(\mathbf{V})$. By standard Schatten p-norm inequalities, we now have

$$\|\mathbf{V}_R\|_{S_p}^{p} \le \|\mathbf{V}_R\|_{S_1}^{p} = \big[\operatorname{tr}(\mathbf{V}_R)\big]^p \le \big[\operatorname{tr}(\mathbf{V})\big]^p = \|\mathbf{V}\|_{S_1}^{p} \le d^{p-1} \cdot \|\mathbf{V}\|_{S_p}^{p}.$$

Let us assume that $\mathbb{E}\,\|\mathbf{V}\|_{S_p}^{p} < \infty$, for otherwise there is nothing to prove. Since $\|\mathbf{V}_R\|_{S_p}^{p} \overset{\text{a.s.}}{\to} \|\mathbf{V}\|_{S_p}^{p}$ as $R \to \infty$, the dominated convergence theorem implies that

$$\lim_{R\to\infty} \mathbb{E}\,\|\mathbf{V}_R\|_{S_p}^{p} = \mathbb{E}\,\|\mathbf{V}\|_{S_p}^{p}.$$

Finally, since $\|\mathbf{X}_R\|_{S_{2p}}^{2p} \xrightarrow{\text{a.s.}} \|\mathbf{X}\|_{S_{2p}}^{2p}$, Fatou's lemma and (11.15) imply

$$\begin{aligned}\mathbb{E}\,\|\mathbf{X}\|_{S_{2p}}^{2p} &= \mathbb{E}\Big(\liminf_{R\to\infty}\|\mathbf{X}_R\|_{S_{2p}}^{2p}\Big) \le \liminf_{R\to\infty}\mathbb{E}\,\|\mathbf{X}_R\|_{S_{2p}}^{2p} \\ &\le \big(2(2p-1)\big)^p \liminf_{R\to\infty}\mathbb{E}\,\|\mathbf{V}_R\|_{S_p}^{p} = \big(2(2p-1)\big)^p\,\mathbb{E}\,\|\mathbf{V}\|_{S_p}^{p}.\end{aligned}$$

The claim of the theorem follows by taking the $1/(2p)$th power of both sides.

12. Exponential concentration inequalities. In this section, we develop an exponential moment bound for a kernel Stein pair. This result shows that we can control the trace m.g.f. in terms of the conditional variance and the kernel conditional variance.

THEOREM 12.1 (Exponential moments for a kernel Stein pair). *Suppose that* $(\mathbf{X},\mathbf{X}')$ *is a* $\mathbf{K}$*-Stein pair, and assume that* $\|\mathbf{X}\|$ *is bounded. For* $\psi>0$, *define*

$$r(\psi) := \frac{1}{\psi}\inf_{s>0}\log\mathbb{E}\,\bar{\operatorname{tr}}\exp\Big(\frac{\psi}{2}\big(s\mathbf{V}_{\mathbf{X}}+s^{-1}\mathbf{V}^{\mathbf{K}}\big)\Big). \tag{12.1}$$

When $|\theta|<\sqrt{\psi}$,

$$\log\mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta\mathbf{X}} \le \frac{\psi r(\psi)}{2}\log\Big(\frac{1}{1-\theta^2/\psi}\Big) \le \frac{r(\psi)\theta^2}{2(1-\theta^2/\psi)}.$$

The conditional variances $\mathbf{V}_{\mathbf{X}}$ *and* $\mathbf{V}^{\mathbf{K}}$ *are defined in* (8.9) *and* (8.10).

The rest of this section is devoted to establishing this result. The pattern of argument is similar with the proofs of Chatterjee (2005), Theorem 3.13, and Mackey et al. (2014), Theorem 5.1, but we require another nontrivial new matrix inequality.

Theorem 12.1 has a variety of consequences. In Section 12.1, we use it to derive the exponential Efron–Stein inequality, Theorem 4.3. Additional applications of the result appear in Section 13. The result is stated under the boundedness assumption. In the general unbounded case, Theorem 4.3 is established in a different way, by deduction from the polynomial Efron–Stein bounds.

12.1. *Proof of exponential Efron–Stein inequality*: *Bounded case*. Theorem 12.1 is the last major step toward the matrix exponential Efron–Stein inequality, Theorem 4.3. The proof is similar to the argument in Section 11.5 leading up to the polynomial Efron–Stein inequality so we proceed quickly.

Recall the setup from Section 4.1. In this section, we assume that $\mathbf{X}$ is bounded. We rely on the kernel Stein pair $(\mathbf{X},\mathbf{X}')$ that we constructed in Section 11.1, as well as the analysis from Section 11.3. From Theorem 12.1 we obtain that for any

$s > 0$,

$$\log \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta \mathbf{X}} \le \frac{1}{2}\log\left(\frac{1}{1-\theta^2/\psi}\right)\log \mathbb{E}\,\bar{\operatorname{tr}}\,\exp\left(\frac{s\psi}{2}(\mathbf{V_X} + s^{-2}\mathbf{V^K})\right)$$
$$\le \frac{\theta^2/\psi}{2(1-\theta^2/\psi)}\log \mathbb{E}\,\bar{\operatorname{tr}}\exp\left(\frac{s\psi}{2}(\mathbf{V_X} + s^{-2}\mathbf{V^K})\right).$$

Since $t \mapsto \mathrm{e}^{s\psi t/2}$ is increasing and convex on $\mathbb{R}_+$, by choosing s as in (11.5), Lemma 10.5 implies that

$$\mathbb{E}\,\bar{\operatorname{tr}}\exp\left(\frac{s\psi}{2}(\mathbf{V_X} + s^{-2}\mathbf{V^K})\right) \le \mathbb{E}\,\bar{\operatorname{tr}}\exp\left(\frac{s\psi}{2}\mathbf{\Gamma}_0(Z)\right)$$
$$\le \mathbb{E}\,\bar{\operatorname{tr}}\exp\left(2\frac{n\psi}{2}\mathbf{\Gamma}_0(Z)\right) = \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{2\psi\mathbf{V}}.$$

The identity $\frac{n}{2}\mathbf{\Gamma}_0(Z) = \mathbf{V}$ was established in (11.11). Combine the two displays, and make the change of variables $\psi \mapsto \psi/2$ to complete the proof of Theorem 4.3 in the bounded case.

12.2. *The exponential Efron–Stein inequality*: *General case*. Here, we will establish the exponential Efron–Stein inequality without the boundedness assumption. The truncation argument we have used for the polynomial case does not seem to be applicable here, so we use an alternative approach. Assume, without loss of generality, that $\|\mathbf{V}\|$ has finite moments of all order (otherwise there is nothing to prove). From the polynomial Efron–Stein inequality (Theorem 4.2), we know that for any $p \in \mathbb{N}$,

$$\mathbb{E}\,\bar{\operatorname{tr}}(\mathbf{X}^{2p}) \le (2(2p-1))^p\,\mathbb{E}\,\bar{\operatorname{tr}}(\mathbf{V}^p). \tag{12.2}$$

For any $x, \theta \in \mathbb{R}$, we have the numerical inequality

$$\mathrm{e}^{\theta x} < \mathrm{e}^{\theta x} + \mathrm{e}^{-\theta x} = 2\sum_{p=0}^{\infty}\frac{\theta^{2p}}{(2p)!}x^{2p}.$$

Therefore,

$$\bar{\operatorname{tr}}\,\mathrm{e}^{\theta\mathbf{X}} \le 2\sum_{p=0}^{\infty}\frac{\theta^{2p}}{(2p)!}\bar{\operatorname{tr}}(\mathbf{X}^{2p}).$$

Take the expectation on both sides, using monotone convergence, to see that

$$\begin{aligned}\mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta\mathbf{X}} &\le 2\sum_{p=0}^{\infty}\frac{\theta^{2p}}{(2p)!}\mathbb{E}\,\bar{\operatorname{tr}}(\mathbf{X}^{2p})\\ &\le 2\sum_{p=0}^{\infty}\frac{(2(2p-1))^p\theta^{2p}}{(2p)!}\mathbb{E}\,\bar{\operatorname{tr}}(\mathbf{V}^p)\\ &\le \sum_{p=0}^{\infty}\frac{\mathrm{e}^p\theta^{2p}}{p!}\mathbb{E}\,\bar{\operatorname{tr}}(\mathbf{V}^p) = \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\mathrm{e}\theta^2\mathbf{V}}.\end{aligned} \tag{12.3}$$

To reach the second line, we invoke (12.2). The last inequality depends on the numerical bound

$$\frac{(2p)!}{p!} \geq \frac{(2(2p-1))^p}{2\mathrm{e}^p} \qquad \text{for each integer } p \geq 0.$$

This inequality can be established using Robbins's error estimate Robbins (1955) for Stirling's formula. We obtain (4.4) by taking the logarithm of (12.3). Finally, (4.5) follows from the fact that $t \mapsto t^{-1} \log \mathbb{E}\,\bar{\mathrm{tr}}\,(\mathrm{e}^{t\mathbf{V}})$ is monotone increasing for $t > 0$ by virtue of Jensen's inequality.

12.3. *The exponential mean value trace inequality.* To establish Theorem 12.1, we require another trace inequality.

LEMMA 12.2 (Exponential mean value trace inequality). *For all matrices* $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{H}^d$ *and all* $s > 0$ *it holds that*

$$\left|\bar{\mathrm{tr}}\left[\mathbf{C}\left(\mathrm{e}^{\mathbf{A}} - \mathrm{e}^{\mathbf{B}}\right)\right]\right| \leq \tfrac{1}{4}\bar{\mathrm{tr}}\left[\left(s(\mathbf{A}-\mathbf{B})^2 + s^{-1}\mathbf{C}^2\right)\left(\mathrm{e}^{\mathbf{A}} + \mathrm{e}^{\mathbf{B}}\right)\right].$$

We defer the proof to Appendix C.

12.4. *Some properties of the trace m.g.f.* For the proof, we need to develop some basic facts about the trace moment generating function.

LEMMA 12.3 (Properties of the trace m.g.f.). *Assume that* $\mathbf{X} \in \mathbb{H}^d$ *is a centered random matrix that is bounded in norm. Define the normalized trace m.g.f.* $m(\theta) = \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta\mathbf{X}}$ *for* $\theta \in \mathbb{R}$. *Then*

$$\log m(\theta) \geq 0 \quad \textit{and} \quad \log m(0) = 0. \tag{12.4}$$

The derivative of the trace m.g.f. satisfies

$$m'(\theta) = \mathbb{E}\,\bar{\mathrm{tr}}\left[\mathbf{X}\mathrm{e}^{\theta\mathbf{X}}\right] \quad \textit{and} \quad m'(0) = 0. \tag{12.5}$$

The trace m.g.f. is a convex function; *in particular*

$$m'(\theta) \leq 0 \qquad \textit{for } \theta \leq 0 \quad \textit{and} \quad m'(\theta) \geq 0 \qquad \textit{for } \theta \geq 0. \tag{12.6}$$

PROOF. The result $m(0) = 0$ follows immediately from the definition of the trace m.g.f. Since $\mathbb{E}\,\mathbf{X} = \mathbf{0}$,

$$\log m(\theta) = \log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta\mathbf{X}} \geq \log \bar{\mathrm{tr}}\,\mathrm{e}^{\theta\,\mathbb{E}\,\mathbf{X}} \geq 0.$$

The first inequality is Jensen's, which depends on the fact (2.1) that the trace exponential is a convex function.

Next, consider the derivative of the trace m.g.f. For each $\theta \in \mathbb{R}$,

$$m'(\theta) = \mathbb{E}\,\bar{\mathrm{tr}}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{e}^{\theta\mathbf{X}}\right] = \mathbb{E}\,\bar{\mathrm{tr}}\left[\mathbf{X}\mathrm{e}^{\theta\mathbf{X}}\right], \tag{12.7}$$

where the dominated convergence theorem and the boundedness of $\mathbf{X}$ justify the exchange of expectation and derivative. The claim $m'(0) = 0$ follows from (12.7) and the fact that $\mathbb{E}\,\mathbf{X} = \mathbf{0}$.

Similarly, the second derivative of the trace m.g.f. satisfies

$$m''(\theta) = \mathbb{E}\,\bar{\operatorname{tr}}[\mathbf{X}^2 \mathrm{e}^{\theta \mathbf{X}}] \geq 0 \qquad \text{for each } \theta \in \mathbb{R}.$$

The inequality holds because $\mathbf{X}^2$ and $\mathrm{e}^{\theta\mathbf{X}}$ are both positive semidefinite, so the trace of their product must be nonnegative. We discover that the trace m.g.f. is convex, which means that the derivative m' is an increasing function. □

12.5. *Bounding the derivative of the trace m.g.f.* The first step in the proof of Theorem 12.1 is to bound the trace m.g.f. of the random matrix $\mathbf{X}$ in terms of the two conditional variance measures.

LEMMA 12.4 (The derivative of the trace m.g.f.). *Instate the notation and hypotheses of Theorem 12.1. Define the normalized trace m.g.f.* $m(\theta) := \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta\mathbf{X}}$. *Then*

$$|m'(\theta)| \leq \frac{1}{2}|\theta| \cdot \inf_{s>0} \mathbb{E}\,\bar{\operatorname{tr}}[(s\mathbf{V}_{\mathbf{X}} + s^{-1}\mathbf{V}^{\mathbf{K}})\mathrm{e}^{\theta\mathbf{X}}] \qquad \textit{for all } \theta \in \mathbb{R}. \tag{12.8}$$

PROOF. Assume that the kernel Stein pair $(\mathbf{X}, \mathbf{X}')$ is constructed from an auxiliary exchangeable pair (Z, Z'). By (12.5), the result holds trivially for $\theta = 0$, so we may assume that $\theta \neq 0$. The form of the derivative (12.5) is suitable for an application of the method of exchangeable pairs, Lemma 8.4. Since $\mathbf{X}$ is bounded, the regularity condition (8.5) is satisfied, and we obtain

$$m'(\theta) = \tfrac{1}{2}\,\mathbb{E}\,\bar{\operatorname{tr}}[\mathbf{K}(Z, Z')(\mathrm{e}^{\theta\mathbf{X}} - \mathrm{e}^{\theta\mathbf{X}'})]. \tag{12.9}$$

The exponential mean value trace inequality, Lemma 12.2, implies that

$$\begin{aligned}
|m'(\theta)| &\leq \frac{1}{8} \cdot \inf_{s>0} \mathbb{E}\,\bar{\operatorname{tr}}[(s(\theta\mathbf{X} - \theta\mathbf{X}')^2 + s^{-1}\mathbf{K}(Z, Z')^2) \cdot (\mathrm{e}^{\theta\mathbf{X}} + \mathrm{e}^{\theta\mathbf{X}'})] \\
&= \frac{1}{4} \cdot \inf_{s>0} \mathbb{E}\,\bar{\operatorname{tr}}[(s(\theta\mathbf{X} - \theta\mathbf{X}')^2 + s^{-1}\mathbf{K}(Z, Z')^2) \cdot \mathrm{e}^{\theta\mathbf{X}}] \\
&= \frac{1}{4}|\theta| \cdot \inf_{t>0} \mathbb{E}\,\bar{\operatorname{tr}}[(t(\mathbf{X} - \mathbf{X}')^2 + t^{-1}\mathbf{K}(Z, Z')^2) \cdot \mathrm{e}^{\theta\mathbf{X}}] \\
&= \frac{1}{2}|\theta| \cdot \inf_{t>0} \mathbb{E}\,\bar{\operatorname{tr}}\left[\frac{t}{2}\,\mathbb{E}[(\mathbf{X} - \mathbf{X}')^2 | Z] \cdot \mathrm{e}^{\theta\mathbf{X}} + \frac{1}{2t}\,\mathbb{E}[\mathbf{K}(Z, Z')^2 | Z] \cdot \mathrm{e}^{\theta\mathbf{X}}\right].
\end{aligned}$$

The first equality follows from the exchangeability of $(\mathbf{X}, \mathbf{X}')$; the second follows from the change of variables $s = |\theta|^{-1}t$; and the final one depends on the pull-through property of conditional expectation. We reach the result (12.8) by introducing the definitions (8.9) and (8.10) of the conditional variance and the kernel conditional variance. □

12.6. *Decoupling via an entropy inequality.* The next step in the proof uses an entropy inequality to separate the conditional variances in (12.8) from the matrix exponential.

FACT 12.5 (Young's inequality for matrix entropy). Let $\mathbf{U}$ be a random matrix in $\mathbb{H}^d$ that is bounded in norm, and suppose that $\mathbf{W}$ is a random matrix in $\mathbb{H}_+^d$ that is subject to the normalization $\mathbb{E}\,\bar{\mathrm{tr}}\,\mathbf{W} = 1$. Then

$$\mathbb{E}\,\bar{\mathrm{tr}}(\mathbf{U}\mathbf{W}) \le \log \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\mathbf{U}} + \mathbb{E}\,\bar{\mathrm{tr}}[\mathbf{W}\log \mathbf{W}].$$

This fact appears as Mackey et al. (2014), Proposition A.3; see also Carlen (2010), Theorem 2.13.

12.7. *A differential inequality.* To continue the argument, we fix a parameter $\psi > 0$. Rewrite (12.8) as

$$\left|m'(\theta)\right| \le \frac{|\theta| m(\theta)}{\psi} \inf_{s>0} \mathbb{E}\,\bar{\mathrm{tr}}\left[\left(\frac{\psi}{2}\left(s\mathbf{V}_{\mathbf{X}} + s^{-1}\mathbf{V}^{\mathbf{K}}\right)\right) \cdot \frac{\mathrm{e}^{\theta \mathbf{X}}}{m(\theta)}\right].$$

Invoke Fact 12.5 to obtain

$$\left|m'(\theta)\right| \le \frac{|\theta| m(\theta)}{\psi}\left(\inf_{s>0} \log \mathbb{E}\,\bar{\mathrm{tr}}\exp\left(\frac{\psi}{2}\left(s\mathbf{V}_{\mathbf{X}} + s^{-1}\mathbf{V}^{\mathbf{K}}\right)\right) + \mathbb{E}\,\bar{\mathrm{tr}}\left[\frac{\mathrm{e}^{\theta\mathbf{X}}}{m(\theta)} \log \frac{\mathrm{e}^{\theta\mathbf{X}}}{m(\theta)}\right]\right).$$

In view of (12.4),

$$\log \frac{\mathrm{e}^{\theta\mathbf{X}}}{m(\theta)} = \theta\mathbf{X} - \log m(\theta) \cdot \mathbf{I} \preccurlyeq \theta \mathbf{X}.$$

Identify the function $r(\psi)$ defined in (12.1) and the derivative of the trace m.g.f. to reach

$$\left|m'(\theta)\right| \le |\theta| m(\theta) r(\psi) + \frac{\theta|\theta|}{\psi} \cdot m'(\theta). \tag{12.10}$$

This inequality is valid for all $\psi > 0$, and all $\theta \in \mathbb{R}$.

12.8. *Solving the differential inequality.* We begin with the case where $\theta \ge 0$. The result (12.6) shows that $m'(\varphi) \ge 0$ for $\varphi \in [0, \theta]$. Therefore, the differential inequality (12.10) reads

$$m'(\varphi) \le \varphi m(\varphi) r(\psi) + \left(\varphi^2/\psi\right) m'(\varphi) \qquad \text{for } \varphi \in [0, \theta].$$

Rearrange this expression to isolate the log-derivative $m'(\varphi)/m(\varphi)$:

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

Recall the fact (12.4) that $\log m(0) = 0$, and integrate to obtain

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

when $0 \le \theta < \sqrt{\psi}$. Making an additional approximation, we find that

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

for the same parameter range.

Finally, we treat the case where $\theta \le 0$. The result (12.6) shows that $m'(\varphi) \le 0$ for $\varphi \in [\theta, 0]$, so the differential inequality (12.10) becomes

$$m'(\varphi) \ge \varphi m(\varphi) r(\psi) + (\varphi^2/\psi) m'(\varphi) \qquad \text{for } \varphi \in [\theta, 0].$$

The rest of the argument parallels the situation where θ is positive.

13. Complements.

The tools in this paper are applicable in a wide variety of settings. To indicate what might be possible, we briefly present two additional concentration results for random matrices arising as functions of dependent random variables. We also indicate some prospects for future research.

13.1. *Matrix bounded differences without independence.*

A key strength of the method of exchangeable pairs is the fact that it also applies to random matrices that are built from weakly dependent random variables. This section describes an extension of Corollary 6.1 that holds even when the input variables exhibit some interactions.

To quantify the amount of dependency among the variables, we use a Dobrushin interdependence matrix [Dobrushin (1970)]. This concept involves a certain amount of auxiliary notation. Given a vector $\mathbf{x} = (x_1, \dots, x_n)$, we write

$$\mathbf{x}_{-i} = (x_1, \dots x_{i-1}, x_{i+1}, \dots, x_n)$$

for the vector with its ith component deleted. Let $Z = (Z_1, \dots, Z_n)$ be a vector of random variables taking values in a Polish space $\mathcal{Z}$ with sigma algebra $\mathcal{F}$. The symbol $\mu_i(\cdot|Z_{-i})$ refers to the distribution of Z_i conditional on the random vector Z_{-i}. We also require the total variation distance d_{TV} between probability measures μ and ν on $(\mathcal{Z}, \mathcal{F})$:

$$d_{\mathrm{TV}}(\nu, \mu) := \sup_{A \in \mathcal{F}} \bigl|\nu(A) - \mu(A)\bigr|. \tag{13.1}$$

With this foundation in place, we can state the definition.

DEFINITION 13.1 (Dobrushin interdependence matrix). Let $Z = (Z_1, \dots, Z_n)$ be a random vector taking values in a Polish space $\mathcal{Z}$. Let $\mathbf{D} \in \mathbb{R}^{n\times n}$ be a matrix with a zero diagonal that satisfies the condition

$$d_{\mathrm{TV}}\big(\mu_i(\cdot|\mathbf{x}_{-i}), \mu_i(\cdot|\mathbf{y}_{-i})\big) \le \sum_{j=1}^{n} D_{ij}\mathbb{1}[x_j \ne y_j] \tag{13.2}$$

for each index i and for all vectors $\mathbf{x}, \mathbf{y} \in \mathcal{Z}$. Then $\mathbf{D}$ is called a *Dobrushin interdependence matrix* for the random vector Z.

The kernel coupling method extends readily to the setting of weak dependence. We obtain a new matrix bounded differences inequality, which is a significant extension of Corollary 6.1. This statement can be viewed as a matrix version of Chatterjee's result [Chatterjee (2005), Theorem 4.3].

COROLLARY 13.2 (Dobrushin matrix bounded differences). *Suppose that $Z := (Z_1, \dots, Z_n)$ in a Polish space $\mathcal{Z}$ is a vector of dependent random variables with a Dobrushin interdependence matrix $\mathbf{D}$ with the property that*

$$\max\big\{\|\mathbf{D}\|_{1\to1}, \|\mathbf{D}\|_{\infty\to\infty}\big\} < 1. \tag{13.3}$$

Let $\mathbf{H} : \mathcal{Z} \to \mathbb{H}^d$ be a measurable function, and let $(\mathbf{A}_1, \dots, \mathbf{A}_n)$ be a deterministic sequence of Hermitian matrices that satisfy

$$\big(\mathbf{H}(z_1, \dots, z_n) - \mathbf{H}(z_1, \dots, z_j', \dots, z_n)\big)^2 \preccurlyeq \mathbf{A}_j^2,$$

where z_k, z_k' range over the possible values of Z_k for each k. Compute the boundedness and dependence parameters

$$\sigma^2 := \left\|\sum_{j=1}^{n} \mathbf{A}_j^2\right\| \quad \textit{and} \quad b := \left[1 - \frac{1}{2}\big(\|\mathbf{D}\|_{1\to1} + \|\mathbf{D}\|_{\infty\to\infty}\big)\right]^{-1}.$$

Then, for all $t \ge 0$,

$$\mathbb{P}\big\{\lambda_{\max}\big(\mathbf{H}(Z) - \mathbb{E}\,\mathbf{H}(Z)\big) \ge t\big\} \le d \cdot \mathrm{e}^{-t^2/(b\sigma^2)}.$$

Furthermore,

$$\mathbb{E}\,\lambda_{\max}\big(\mathbf{H}(Z) - \mathbb{E}\,\mathbf{H}(Z)\big) \le \sigma\sqrt{b \log d}.$$

Observe that the bounds here are a factor of b worse than the independent case outlined in Corollary 6.1. The proof is similar to the proof in the scalar case in Chatterjee (2005). We refer the reader to our earlier report [Paulin, Mackey and Tropp (2013)] for details.

13.2. *Matrix-valued functions of Haar random elements.* This section describes a concentration result for a matrix-valued function of a random element drawn uniformly from a compact group. This corollary can be viewed as a matrix extension of Chatterjee (2005), Theorem 4.6.

COROLLARY 13.3 (Concentration for Hermitian functions of Haar measures). *Let $Z \sim \mu$ be Haar distributed on a compact topological group G, and let $\mathbf{\Psi} : G \to \mathbb{H}^d$ be a measurable function satisfying $\mathbb{E}\,\mathbf{\Psi}(Z) = \mathbf{0}$. Let $Y, Y_1, Y_2, \dots$ be i.i.d. random variables in G satisfying*

$$Y \overset{d}{=} Y^{-1} \quad \text{and} \quad zYz^{-1} \overset{d}{=} Y \qquad \text{for all } z \in G. \tag{13.4}$$

Assume

$$\|\mathbf{\Psi}(z)\| \le R \qquad \text{for all } z \in G$$

and

$$S^2 = \sup_{g \in G} \big\|\mathbb{E}\big[\big(\mathbf{\Psi}(g) - \mathbf{\Psi}(Yg)\big)^2\big]\big\| < \infty.$$

Compute the boundedness parameter

$$\sigma^2 := \frac{S^2}{2} \sum_{i=0}^{\infty} \min\{1, 4RS^{-1} d_{\mathrm{TV}}(\mu_i, \mu)\},$$

where μ_i is the distribution of the product $Y_i \cdots Y_1$. Then, for all $t \ge 0$,

$$\mathbb{P}\{\lambda_{\max}\big(\mathbf{\Psi}(Z)\big) \ge t\} \le d \cdot \mathrm{e}^{-t^2/(2\sigma^2)}.$$

Furthermore,

$$\mathbb{E}\,\lambda_{\max}\big(\mathbf{\Psi}(Z)\big) \le \sigma\sqrt{2\log d}.$$

Corollary 13.3 relates the concentration of Hermitian functions to the convergence of random walks on a group. In particular, Corollary 13.3 can be used to study matrices constructed from random permutations or random unitary matrices. The proof is similar to the proof of the scalar result; see our earlier report [Paulin, Mackey and Tropp (2013)] for details.

13.3. *Conjectures and consequences.* We conjecture that the following trace inequalities hold.

CONJECTURE 13.4 (Signed mean value trace inequalities). For all matrices $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{H}^d$, all positive integers q, and any $s > 0$ it holds that

$$\begin{aligned} \operatorname{tr}[\mathbf{C}(\mathrm{e}^{\mathbf{A}} - \mathrm{e}^{\mathbf{B}})] \le \tfrac{1}{2} \operatorname{tr}\big[&\big(s(\mathbf{A}-\mathbf{B})_+^2 + s^{-1}\mathbf{C}_+^2\big)\mathrm{e}^{\mathbf{A}} \\ &+ \big(s(\mathbf{A}-\mathbf{B})_-^2 + s^{-1}\mathbf{C}_-^2\big)\mathrm{e}^{\mathbf{B}}\big] \end{aligned}$$

and

$$\begin{aligned}&\operatorname{tr}\bigl[\mathbf{C}(\mathbf{A}^q-\mathbf{B}^q)\bigr]\\&\quad\le\frac{q}{2}\operatorname{tr}\bigl[\bigl(s(\mathbf{A}-\mathbf{B})_+^2+s^{-1}\mathbf{C}_+^2\bigr)|\mathbf{A}|^{q-1}\\&\qquad+\bigl(s(\mathbf{A}-\mathbf{B})_-^2+s^{-1}\mathbf{C}_-^2\bigr)|\mathbf{B}|^{q-1}\bigr].\end{aligned}$$

This statement involves the standard matrix functions that lift the scalar functions $(a)_+ := \max\{a, 0\}$ and $(a)_- := \max\{-a, 0\}$. Extensive simulations with random matrices suggest that Conjecture 13.4 holds, but we did not find a proof.

These inequalities would imply one-sided matrix versions of the exponential Efron–Stein and moment bounds, similar to those formulated for the scalar setting in Boucheron, Lugosi and Massart (2003) and Boucheron et al. (2005). In the scalar case, Conjecture 13.4 is valid, so it is possible to obtain the results of Boucheron, Lugosi and Massart (2003) and Boucheron et al. (2005) by the exchangeable pair method.

APPENDIX A: OPERATOR INEQUALITIES

Our main results rely on some basic inequalities from operator theory. We are not aware of good references for this material, so we have included short proofs.

A.1. Young's inequality for commuting operators. In the scalar setting, Young's inequality provides an additive bound for the product of two numbers. More precisely, for indices $p, q \in (1, \infty)$ that satisfy the conjugacy relation $p^{-1} + q^{-1} = 1$, we have

$$ab \le \frac{1}{p}|a|^p + \frac{1}{q}|b|^q \qquad \text{for all } a, b \in \mathbb{R}. \tag{A.1}$$

The same result has a natural extension for commuting operators.

LEMMA A.1 (Young's inequality for commuting operators). *Suppose that $\mathcal{A}$ and $\mathcal{B}$ are self-adjoint linear maps on the Hilbert space $\mathbb{M}^d$ that commute with each other. Let $p, q \in (1, \infty)$ satisfy the conjugacy relation $p^{-1} + q^{-1} = 1$. Then*

$$\mathcal{A}\mathcal{B} \preccurlyeq \frac{1}{p}|\mathcal{A}|^p + \frac{1}{q}|\mathcal{B}|^q.$$

PROOF. Since $\mathcal{A}$ and $\mathcal{B}$ commute, there exists a unitary operator $\mathcal{U}$ and diagonal operators $\mathcal{D}$ and $\mathcal{M}$ for which $\mathcal{A} = \mathcal{U}\mathcal{D}\mathcal{U}^*$ and $\mathcal{B} = \mathcal{U}\mathcal{M}\mathcal{U}^*$. Young's inequality (A.1) for scalars immediately implies that

$$\mathcal{D}\mathcal{M} \preccurlyeq \frac{1}{p}|\mathcal{D}|^p + \frac{1}{q}|\mathcal{M}|^q.$$

Conjugating both sides of this inequality by $\mathcal{U}$, we obtain

$$\mathcal{AB} = \mathcal{U}(\mathcal{DM})\mathcal{U}^* \preccurlyeq \frac{1}{p}\mathcal{U}|\mathcal{D}|^p\mathcal{U}^* + \frac{1}{q}\mathcal{U}|\mathcal{M}|^q\mathcal{U}^* = \frac{1}{p}|\mathcal{A}|^p + \frac{1}{q}|\mathcal{B}|^q.$$

The last identity follows from the definition of a standard function of an operator. □

A.2. An operator version of Cauchy–Schwarz. We also need a simple version of the Cauchy–Schwarz inequality for operators. The proof follows a classical argument, but it also involves an operator decomposition.

LEMMA A.2 (Operator Cauchy–Schwarz). *Let $\mathcal{A}$ be a self-adjoint linear operator on the Hilbert space $\mathbb{M}^d$, and let $\mathbf{M}$ and $\mathbf{N}$ be matrices in $\mathbb{M}^d$. Then*

$$|\langle \mathbf{M}, \mathcal{A}(\mathbf{N})\rangle| \le [\langle \mathbf{M}, |\mathcal{A}|(\mathbf{M})\rangle \cdot \langle \mathbf{N}, |\mathcal{A}|(\mathbf{N})\rangle]^{1/2}.$$

The inner product symbol refers to the trace, or Frobenius, inner product.

PROOF. Consider the Jordan decomposition $\mathcal{A} = \mathcal{A}_+ - \mathcal{A}_-$, where $\mathcal{A}_+$ and $\mathcal{A}_-$ are both positive semidefinite. For all $s > 0$,

$$\begin{aligned} 0 &\le \langle (s\mathbf{M} - s^{-1}\mathbf{N}), \mathcal{A}_+(s\mathbf{M} - s^{-1}\mathbf{N})\rangle \\ &= s^2\langle \mathbf{M}, \mathcal{A}_+(\mathbf{M})\rangle + s^{-2}\langle \mathbf{N}, \mathcal{A}_+(\mathbf{N})\rangle - 2\langle \mathbf{M}, \mathcal{A}_+(\mathbf{N})\rangle. \end{aligned}$$

Likewise,

$$\begin{aligned} 0 &\le \langle (s\mathbf{M} + s^{-1}\mathbf{N}), \mathcal{A}_-(s\mathbf{M} + s^{-1}\mathbf{N})\rangle \\ &= s^2\langle \mathbf{M}, \mathcal{A}_-(\mathbf{M})\rangle + s^{-2}\langle \mathbf{N}, \mathcal{A}_-(\mathbf{N})\rangle + 2\langle \mathbf{M}, \mathcal{A}_-(\mathbf{N})\rangle. \end{aligned}$$

Add the latter two inequalities and rearrange the terms to obtain

$$2\langle \mathbf{M}, \mathcal{A}(\mathbf{N})\rangle \le s^2\langle \mathbf{M}, |\mathcal{A}|(\mathbf{M})\rangle + s^{-2}\langle \mathbf{N}, |\mathcal{A}|(\mathbf{N})\rangle,$$

where we have used the relation $|\mathcal{A}| = \mathcal{A}_+ + \mathcal{A}_-$. Take the infimum of the right-hand side over $s > 0$ to reach

$$\langle \mathbf{M}, \mathcal{A}(\mathbf{N})\rangle \le [\langle \mathbf{M}, |\mathcal{A}|(\mathbf{M})\rangle \cdot \langle \mathbf{M}, |\mathcal{A}|(\mathbf{N})\rangle]^{1/2}. \tag{A.2}$$

Repeat the same argument, interchanging the roles of the matrices $s\mathbf{M} - s^{-1}\mathbf{N}$ and $s\mathbf{M} + s^{-1}\mathbf{N}$. We conclude that (A.2) also holds with an absolute value on the left-hand side. This observation completes the proof. □

APPENDIX B: THE POLYNOMIAL MEAN VALUE TRACE INEQUALITY

The critical new ingredient in Theorem 9.1 is the polynomial mean value trace inequality, Lemma 9.2. Let us proceed with a proof of this result.

PROOF OF LEMMA 9.2. First, we need to develop another representation for the trace quantity that we are analyzing. Assume that $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{H}^d$. A direct calculation shows that

$$\mathbf{A}^q - \mathbf{B}^q = \sum_{k=0}^{q-1} \mathbf{A}^k (\mathbf{A} - \mathbf{B}) \mathbf{B}^{q-1-k}.$$

As a consequence,

$$\operatorname{tr}\big[\mathbf{C}\big(\mathbf{A}^q - \mathbf{B}^q\big)\big] = \sum_{k=0}^{q-1} \big\langle \mathbf{C}, \mathbf{A}^k (\mathbf{A} - \mathbf{B}) \mathbf{B}^{q-1-k} \big\rangle. \tag{B.1}$$

To bound the right-hand side of (B.1), we require an appropriate mean inequality.

To that end, we define some self-adjoint operators on $\mathbb{M}^d$:

$$\mathcal{A}_k(\mathbf{M}) := \mathbf{A}^k \mathbf{M} \quad \text{and} \quad \mathcal{B}_k(\mathbf{M}) := \mathbf{M}\mathbf{B}^k \qquad \text{for each } k = 0, 1, \dots, q-1.$$

The absolute values of these operators satisfy

$$|\mathcal{A}_k|(\mathbf{M}) = |\mathbf{A}|^k \mathbf{M} \quad \text{and} \quad |\mathcal{B}_k|(\mathbf{M}) = \mathbf{M}|\mathbf{B}|^k \qquad \text{for each } k = 0, 1, \dots, q-1.$$

Note that $|\mathcal{A}_k|$ and $|\mathcal{B}_{q-k-1}|$ commute with each other for each k. Therefore, Young's inequality for commuting operators, Lemma A.1, yields the bound

$$\begin{aligned} |\mathcal{A}_k \mathcal{B}_{q-k-1}| &= |\mathcal{A}_k||\mathcal{B}_{q-k-1}| \\ &\preccurlyeq \frac{k}{q-1} |\mathcal{A}_k|^{(q-1)/k} + \frac{q-k-1}{q-1} |\mathcal{B}_{q-k-1}|^{(q-1)/(q-k-1)} \\ &= \frac{k}{q-1} |\mathcal{A}_1|^{q-1} + \frac{q-k-1}{q-1} |\mathcal{B}_1|^{q-1}. \end{aligned} \tag{B.2}$$

Summing over k, we discover that

$$\sum_{k=0}^{q-1} |\mathcal{A}_k \mathcal{B}_{q-k-1}| \preccurlyeq \frac{q}{2} |\mathcal{A}_1|^{q-1} + \frac{q}{2} |\mathcal{B}_1|^{q-1}. \tag{B.3}$$

This is the mean inequality that we require.

To apply this result, we need to rewrite (B.1) using the operators $\mathcal{A}_k$ and $\mathcal{A}_{q-k-1}$. It holds that

$$\begin{aligned} &\operatorname{tr}\big[\mathbf{C}\big(\mathbf{A}^q - \mathbf{B}^q\big)\big] \\ &\quad = \sum_{k=0}^{q-1} \big\langle \mathbf{C}, (\mathcal{A}_k \mathcal{B}_{q-k-1})(\mathbf{A} - \mathbf{B}) \big\rangle \\ &\quad \le \left[\sum_{k=0}^{q-1} \big\langle \mathbf{C}, |\mathcal{A}_k \mathcal{B}_{q-k-1}|(\mathbf{C}) \big\rangle \cdot \sum_{k=0}^{q-1} \big\langle \mathbf{A} - \mathbf{B}, |\mathcal{A}_k \mathcal{B}_{q-k-1}|(\mathbf{A} - \mathbf{B}) \big\rangle \right]^{1/2}. \end{aligned} \tag{B.4}$$

The second relation follows from the operator Cauchy–Schwarz inequality, Lemma A.2, and the usual Cauchy–Schwarz inequality for the sum.

It remains to bound to two sums on the right-hand side of (B.4). The mean inequality (B.2) ensures that

$$\begin{aligned}&\sum_{k=0}^{q-1}\langle \mathbf{C}, |\mathcal{A}_k\mathcal{B}_{q-k-1}|(\mathbf{C})\rangle \\ &\quad\le \frac{q}{2}\langle \mathbf{C}, (|\mathcal{A}_1|^{q-1}+|\mathcal{B}_1|^{q-1})(\mathbf{C})\rangle \\ &\quad= \frac{q}{2}\langle \mathbf{C}, |\mathbf{A}|^{q-1}\mathbf{C}+\mathbf{C}|\mathbf{B}|^{q-1}\rangle = \frac{q}{2}\operatorname{tr}\bigl[\mathbf{C}^2(|\mathbf{A}|^{q-1}+|\mathbf{B}|^{q-1})\bigr].\end{aligned} \tag{B.5}$$

Likewise,

$$\sum_{k=0}^{q-1}\langle \mathbf{A}-\mathbf{B}, |\mathcal{A}_k\mathcal{B}_{q-k-1}|(\mathbf{A}-\mathbf{B})\rangle \le \frac{q}{2}\operatorname{tr}\bigl[(\mathbf{A}-\mathbf{B})^2(|\mathbf{A}|^{q-1}+|\mathbf{B}|^{q-1})\bigr]. \tag{B.6}$$

Introduce the two inequalities (B.5) and (B.6) into (B.4) to reach

$$\begin{aligned}&\operatorname{tr}\bigl[\mathbf{C}(\mathbf{A}^q-\mathbf{B}^q)\bigr] \\ &\quad\le \frac{q}{2}\bigl(\operatorname{tr}\bigl[\mathbf{C}^2(|\mathbf{A}|^{q-1}+|\mathbf{B}|^{q-1})\bigr]\cdot\operatorname{tr}\bigl[(\mathbf{A}-\mathbf{B})^2(|\mathbf{A}|^{q-1}+|\mathbf{B}|^{q-1})\bigr]\bigr)^{1/2}.\end{aligned}$$

The result follows when we apply the numerical inequality between the geometric mean and the arithmetic mean. □

APPENDIX C: THE EXPONENTIAL MEAN VALUE TRACE INEQUALITY

Finally, we establish the trace inequality stated in Lemma 12.2. See the manuscript Paulin (2012) for an alternative proof.

PROOF OF LEMMA 12.2. To begin, we develop an alternative expression for the trace quantity that we need to bound. Observe that

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathrm{e}^{\tau\mathbf{A}}\mathrm{e}^{(1-\tau)\mathbf{B}} = \mathrm{e}^{\tau\mathbf{A}}(\mathbf{A}-\mathbf{B})\mathrm{e}^{(1-\tau)\mathbf{B}}.$$

The fundamental theorem of calculus delivers the identity

$$\mathrm{e}^{\mathbf{A}}-\mathrm{e}^{\mathbf{B}} = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\tau}\mathrm{e}^{\tau\mathbf{A}}\mathrm{e}^{(1-\tau)\mathbf{B}}\,\mathrm{d}\tau = \int_0^1 \mathrm{e}^{\tau\mathbf{A}}(\mathbf{A}-\mathbf{B})\mathrm{e}^{(1-\tau)\mathbf{B}}\,\mathrm{d}\tau.$$

Therefore, using the definition of the trace inner product, we reach

$$\operatorname{tr}\bigl[\mathbf{C}(\mathrm{e}^{\mathbf{A}}-\mathrm{e}^{\mathbf{B}})\bigr] = \int_0^1 \langle \mathbf{C}, \mathrm{e}^{\tau\mathbf{A}}(\mathbf{A}-\mathbf{B})\mathrm{e}^{(1-\tau)\mathbf{B}}\rangle\,\mathrm{d}\tau. \tag{C.1}$$

We can bound the right-hand side by developing an appropriate matrix version of the inequality between the logarithmic mean and the arithmetic mean.

Let us define two families of positive-definite operators on the Hilbert space $\mathbb{M}^d$:

$$\mathcal{A}_\tau(\mathbf{M}) = \mathrm{e}^{\tau\mathbf{A}}\mathbf{M} \quad \text{and} \quad \mathcal{B}_{1-\tau}(\mathbf{M}) = \mathbf{M}\mathrm{e}^{(1-\tau)\mathbf{B}} \qquad \text{for each } \tau \in [0, 1].$$

In other words, $\mathcal{A}_\tau$ is a left-multiplication operator, and $\mathcal{B}_{1-\tau}$ is a right-multiplication operator. It follows immediately that $\mathcal{A}_\tau$ and $\mathcal{B}_{1-\tau}$ commute for each $\tau \in [0, 1]$. Young's inequality for commuting operators, Lemma A.1, implies that

$$\mathcal{A}_\tau\mathcal{B}_{1-\tau} \preccurlyeq \tau \cdot |\mathcal{A}_\tau|^{1/\tau} + (1-\tau) \cdot |\mathcal{B}_{1-\tau}|^{1/(1-\tau)} = \tau \cdot |\mathcal{A}_1| + (1-\tau) \cdot |\mathcal{B}_1|.$$

Integrating over τ, we discover that

$$\text{(C.2)} \qquad \int_0^1 \mathcal{A}_\tau\mathcal{B}_{1-\tau}\, d\tau \preccurlyeq \frac{1}{2}\big(|\mathcal{A}_1| + |\mathcal{B}_1|\big) = \frac{1}{2}(\mathcal{A}_1 + \mathcal{B}_1).$$

This is our matrix extension of the logarithmic–arithmetic mean inequality.

To relate this result to the problem at hand, we rewrite the expression (C.1) using the operators $\mathcal{A}_\tau$ and $\mathcal{B}_{1-\tau}$. Indeed,

$$\begin{aligned} &\operatorname{tr}\big[\mathbf{C}(\mathrm{e}^{\mathbf{A}} - \mathrm{e}^{\mathbf{B}})\big] \\ \text{(C.3)} \qquad &= \int_0^1 \big\langle \mathbf{C}, (\mathcal{A}_\tau\mathcal{B}_{1-\tau})(\mathbf{A}-\mathbf{B})\big\rangle\, \mathrm{d}\tau \\ &\le \bigg[\int_0^1 \big\langle \mathbf{C}, (\mathcal{A}_\tau\mathcal{B}_{1-\tau})(\mathbf{C})\big\rangle\, \mathrm{d}\tau \cdot \int_0^1 \big\langle \mathbf{A}-\mathbf{B}, (\mathcal{A}_\tau\mathcal{B}_{1-\tau})(\mathbf{A}-\mathbf{B})\big\rangle\, \mathrm{d}\tau\bigg]^{1/2}. \end{aligned}$$

The second identity follows from the definition of the trace inner product. The last relation follows from the operator Cauchy–Schwarz inequality, Lemma A.2, and the usual Cauchy–Schwarz inequality for the integral.

It remains to bound the two integrals in (C.3). These estimates are an immediate consequence of (C.2). First,

$$\begin{aligned} &\int_0^1 \big\langle \mathbf{C}, (\mathcal{A}_\tau\mathcal{B}_{1-\tau})(\mathbf{C})\big\rangle\, \mathrm{d}\tau \\ \text{(C.4)} \qquad &\le \frac{1}{2}\big\langle \mathbf{C}, (\mathcal{A}_1 + \mathcal{B}_1)(\mathbf{C})\big\rangle \\ &= \frac{1}{2}\big\langle \mathbf{C}, \mathrm{e}^{\mathbf{A}}\mathbf{C} + \mathbf{C}\mathrm{e}^{\mathbf{B}}\big\rangle = \frac{1}{2}\operatorname{tr}\big[\mathbf{C}^2(\mathrm{e}^{\mathbf{A}} + \mathrm{e}^{\mathbf{B}})\big]. \end{aligned}$$

The last two relations follow from the definitions of the operators $\mathcal{A}_1$ and $\mathcal{B}_1$, the definition of the trace inner product, and the cyclicity of the trace. Likewise,

$$\text{(C.5)} \qquad \int_0^1 \big\langle \mathbf{A}-\mathbf{B}, (\mathcal{A}_\tau\mathcal{B}_{1-\tau})(\mathbf{A}-\mathbf{B})\big\rangle\, \mathrm{d}\tau = \frac{1}{2}\operatorname{tr}\big[(\mathbf{A}-\mathbf{B})^2(\mathrm{e}^{\mathbf{A}} + \mathrm{e}^{\mathbf{B}})\big].$$

Substitute (C.4) and (C.5) into inequality (C.3) to reach

$$\begin{aligned}&\operatorname{tr}\big[\mathbf{C}(\mathrm{e}^{\mathbf{A}}-\mathrm{e}^{\mathbf{B}})\big]\\&\quad\le\tfrac{1}{2}\big(\operatorname{tr}\big[\mathbf{C}^2(\mathrm{e}^{\mathbf{A}}+\mathrm{e}^{\mathbf{B}})\big]\cdot\operatorname{tr}\big[(\mathbf{A}-\mathbf{B})^2(\mathrm{e}^{\mathbf{A}}+\mathrm{e}^{\mathbf{B}})\big]\big)^{1/2}.\end{aligned}$$

We obtain the result stated in Lemma 12.2 by applying the numerical inequality between the geometric mean and the arithmetic mean. □

REFERENCES

ADAMCZAK, R., LITVAK, A. E., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2011). Sharp bounds on the rate of convergence of the empirical covariance matrix. *C. R. Math. Acad. Sci. Paris* **349** 195–200. MR2769907

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

AVRON, H. and TOLEDO, S. (2014). Effective stiffness: Generalizing effective resistance sampling to finite element matrices. Available at arXiv:1110.4437.

BHATIA, R. (2007). *Positive Definite Matrices*. Princeton Univ. Press, Princeton, NJ. MR2284176

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2003). Concentration inequalities using the entropy method. *Ann. Probab.* **31** 1583–1614. MR1989444

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193

BOUCHERON, S., BOUSQUET, O., LUGOSI, G. and MASSART, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.* **33** 514–560. MR2123200

BURDA, Z., JAROSZ, A., NOWAK, M. A., JURKIEWICZ, J., PAPP, G. and ZAHED, I. (2011). Applying free random variables to random matrix analysis of financial data. Part I: The Gaussian case. *Quant. Finance* **11** 1103–1124. MR2813839

CAI, T. T. and JIANG, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann. Statist.* **39** 1496–1525. MR2850210

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

CHATTERJEE, S. (2005). Concentration inequalities with exchangeable pairs. PhD thesis, Stanford Univ., Palo Alto, February 2008. Available at arXiv:math/0507526.

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

CHEN, R. Y., GITTENS, A. and TROPP, J. A. (2012). The masked sample covariance estimator: An analysis via the matrix Laplace transform method. *Information and Inference* **1** 2–20. DOI:10.1093/imaiai/ias001.

CHEN, R. Y. and TROPP, J. A. (2014). Subadditivity of matrix ϕ-entropy and concentration of random matrices. *Electron. J. Probab.* **19** no. 27, 30. MR3174839

COLLINS, B., MCDONALD, D. and SAAD, N. (2013). Compound wishart matrices and noisy covariance matrices: Risk underestimation. Available at arXiv:1306.5510.

DALVI, N., DASGUPTA, A., KUMAR, R. and RASTOGI, V. (2013). Aggregating crowdsourced binary ratings. In *Proceedings of the* 22*nd International Conference on World Wide Web*, *WWW'*13 285–294. Republic and Canton of Geneva, Switzerland.

DOBRUSHIN, R. L. (1970). Prescribing a system of random variables by conditional distributions. *Theory Probab. Appl.* **15** 458–486.

DUDLEY, R. M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge Univ. Press, Cambridge. MR1932358

KOLTCHINSKII, V. (2011). Von Neumann entropy penalization and low-rank matrix estimation. *Ann. Statist*. **39** 2936–2973. MR3012397

LEVIN, D. A., PERES, Y. and WILMER, E. L. (2009). *Markov Chains and Mixing Times*. Amer. Math. Soc., Providence, RI. MR2466937

MACHART, P. and RALAIVOLA, L. (2012). Confusion matrix stability bounds for multiclass classification. Available at arXiv:1202.6221.

MACKEY, L., JORDAN, M. I., CHEN, R. Y., FARRELL, B. and TROPP, J. A. (2014). Matrix concentration inequalities via the method of exchangeable pairs. *Ann. Probab.* **42** 906–945. MR3189061

MORVANT, E., KOÇO, S. and RALAIVOLA, L. (2012). Pac-Bayesian generalization bound on confusion matrix for multi-class classification. In *Proceedings of the* 29*th International Conference on Machine Learning* (*ICML*-12) 815–822. International Machine Learning Society.

NETRAPALLI, P., JAIN, P. and SANGHAVI, S. (2013). Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems* 2796–2804. Neural Information Processing Systems Foundation.

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

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

PAULIN, D. (2012). A Note on matrix concentration inequalities via the method of exchangeable pairs. Available at arXiv:1212.2012.

PAULIN, D., MACKEY, L. and TROPP, J. A. (2013). Deriving matrix concentration inequalities from kernel couplings. Available at arXiv:1305.0612.

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

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron*. *J*. *Stat*. **5** 935–980. MR2836766

ROBBINS, H. (1955). A remark on Stirling's formula. *Amer. Math*. *Monthly* **62** 26–29. MR0069328

SHAO, Q.-M. and ZHOU, W.-X. (2014). Necessary and sufficient conditions for the asymptotic distributions of coherence of ultra-high dimensional random matrices. *Ann*. *Probab.* **42** 623–648. MR3178469

SOLOVEYCHIK, I. (2014). Error bound for compound Wishart matrices. Available at arXiv:1402.5581.

SPEICHER, R. (1998). Combinatorial theory of the free product with amalgamation and operator-valued free probability theory. *Mem*. *Amer. Math*. *Soc*. **132** x+88. MR1407898

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

STEIN, C. (1986). *Approximate Computation of Expectations*. IMS, Hayward, CA. MR0882007

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

TROPP, J. A. (2015). An introduction to matrix concentration inequalities. *Found*. *Trends Mach*. *Learning* **8**. Available at arXiv:1501.01571.

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

ZHOU, E. and HU, J. (2014). Gradient-based adaptive stochastic search for non-differentiable optimization. *IEEE Trans. Automat. Control* **59** 1818–1832. MR3232075

D. PAULIN
DEPARTMENT OF STATISTICS
AND APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
6 SCIENCE DRIVE 2, BLOCK S16, 06-127
SINGAPORE 117546
USA
E-MAIL: paulindani@gmail.com

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

J. A. TROPP
DEPARTMENT OF COMPUTING
AND MATHEMATICAL SCIENCES
ANNENBERG CENTER, ROOM 307
1200 E. CALIFORNIA BLVD.
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91125
USA
E-MAIL: jtropp@cms.caltech.edu