DERIVING MATRIX CONCENTRATION INEQUALITIES FROM KERNEL COUPLINGS

By

Daniel Paulin
Lester Mackey
Joel A. Tropp

Technical Report No. 2014-10
August 2014

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

DERIVING MATRIX CONCENTRATION INEQUALITIES FROM KERNEL COUPLINGS

By

Daniel Paulin
National University of Singapore

Lester Mackey
Stanford University

Joel A. Tropp
California Institute of Technology

Technical Report No. 2014-10
August 2014

This research was supported in part by grants from the Office of Naval Research and the Air Force Office of Scientific Research.

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

Deriving Matrix Concentration Inequalities from Kernel Couplings

Daniel Paulin[1]
Lester Mackey[2] **and Joel A. Tropp**[3]

[1] *Department of Mathematics, National University of Singapore, e-mail:* paulindani@gmail.com

[2] *Department of Statistics, Stanford University, e-mail:* lmackey@stanford.edu

[3] *Department of Computing and Mathematical Sciences, California Institute of Technology, e-mail:* jtropp@cms.caltech.edu

Abstract: This paper derives exponential tail bounds and polynomial moment inequalities for the spectral norm deviation of a random matrix from its mean value. The argument depends on a matrix extension of Stein's method of exchangeable pairs for concentration of measure, as introduced by Chatterjee. Recent work of Mackey et al. uses these techniques to analyze random matrices with additive structure, while the enhancements in this paper cover a wider class of matrix-valued random elements. In particular, these ideas lead to a bounded differences inequality that applies to random matrices constructed from weakly dependent random variables. The proofs require novel trace inequalities that may be of independent interest.

AMS 2000 subject classifications: Primary 60B20, 60E15; secondary 60G09, 60F10.
Keywords and phrases: Concentration inequalities, Stein's method, random matrix, non-commutative, exchangeable pairs, coupling, bounded differences, Dobrushin dependence, Ising model, Haar measure, trace inequality.

This paper is based on two independent manuscripts from late 2012 that both used kernel couplings to establish matrix concentration inequalities. One manuscript is by Paulin; the other is by Mackey and Tropp. The authors have combined this research into a unified presentation, with equal contributions from both groups.

1. Introduction

Matrix concentration inequalities provide probabilistic bounds on the spectral-norm deviation of a random matrix from its mean value. Over the last decade, a growing field of research has established that many scalar concentration results have direct analogs for matrices. For example, see [1, 16, 23]. This machinery has simplified the study of random matrices that arise in applications from statistics [8], machine learning [15], signal processing [2], numerical analysis [22], theoretical computer science [24], and combinatorics [17].

Most of the recent research on matrix concentration depends on a matrix extension of the Laplace transform method from elementary probability. In the matrix setting, it is a serious technical challenge to obtain bounds on the matrix analog of the moment generating function. The earlier works [1, 16] use the Golden–Thompson inequality to accomplish this task. A more powerful argument [23] invokes Lieb's Theorem [10, Thm. 6] to complete the estimates.

Very recently, Mackey et al. [13] have shown that it is also possible to use Stein's method of exchangeable pairs to control the matrix moment generating function. This argument depends

on a matrix version of Chatterjee's technique [5, 4] for establishing concentration inequalities using exchangeable pairs. This approach has two chief advantages. First, it offers a straightforward way to prove polynomial moment inequalities for matrices, which are not easy to obtain using earlier techniques. Second, exchangeable pair arguments also apply to random matrices constructed from weakly dependent random variables.

The work [13] focuses on sums of weakly dependent random matrices because its techniques are less effective for other examples. The goal of the current research is to adapt ideas from Chatterjee's thesis [4] to establish concentration inequalities for more general types of random matrices. In particular, we have obtained new versions of the matrix bounded difference inequality (see [23, Cor. 7.5] or [13, Cor. 11.1]) that hold for a random matrix that is expressed as a measurable function of weakly dependent random variables. These results appear as Corollary 4.1 and Corollary 5.2.

1.1. A First Look at Exchangeable Pairs

The method of exchangeable pairs depends on the idea that an exchangeable counterpart of a random variable encodes information about the symmetries in the distribution. Here is a simple but fundamental example of an exchangeable pair of random matrices:

$$\boldsymbol{X} = \sum\nolimits_{j=1}^{n} \boldsymbol{Y}_j \quad \text{and} \quad \boldsymbol{X}' = \boldsymbol{X} + (\widetilde{\boldsymbol{Y}}_J - \boldsymbol{Y}_J) \tag{1.1}$$

where $\{\boldsymbol{Y}_j\}$ is an independent family of random Hermitian matrices, J is a random index chosen uniformly from $\{1, \dots, n\}$, and $\widetilde{\boldsymbol{Y}}_J$ is an independent copy of $\boldsymbol{Y}_J$. Notice that

$$\frac{n}{2}\,\mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{X}')^2\right] = \mathbb{E}\left[\boldsymbol{X}^2\right] - [\mathbb{E}\,\boldsymbol{X}]^2 = \operatorname{Var}(\boldsymbol{X}).$$

As a consequence, we can interpret the random matrix $\frac{n}{2}(\boldsymbol{X} - \boldsymbol{X}')^2$ as a stochastic estimate for the variance of the independent sum $\boldsymbol{X}$. When this random matrix is uniformly small in norm, we can prove that the sum $\boldsymbol{X}$ concentrates around its mean value. We refer to Theorem 3.1 or the result [13, Thm. 4.1] for a rigorous statement.

1.2. Roadmap

Section 1.3 continues with some notation and preliminary remarks. In Section 2, we describe the concept of a kernel Stein pair of random matrices, which stands at the center of our analysis. In Section 3, we state abstract concentration inequalities for kernel Stein pairs. Afterward, Sections 4 and 5 derive bounded difference inequalities for random matrices constructed from independent and weakly dependent random variables. As an application, we consider the problem of estimating the correlations in a two-dimensional Ising model in Section 6. We close with some complementary material in Section 7. The proofs of the main results appear in three Appendices.

1.3. Notation and Preliminaries

First, we introduce the identity matrix $\mathbf{I}$ and the zero matrix $\mathbf{0}$. Their dimensions are determined by context.

We write $\mathbb{M}^d$ for the algebra of $d \times d$ complex matrices. 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 $\langle \boldsymbol{B}, \boldsymbol{C} \rangle := \operatorname{tr}[\boldsymbol{B}^*\boldsymbol{C}]$ to form a Hilbert space.

Let $\mathbb{H}^d$ denote the subspace of $\mathbb{M}^d$ consisting of $d \times d$ Hermitian matrices. Given an interval I of the real line, we define $\mathbb{H}^d(I)$ to be the family of Hermitian matrices with eigenvalues contained in I. We use curly inequalities, such as $\preccurlyeq$, for the positive semidefinite order on the Hilbert space ℓ_2^d and the Hilbert space $\mathbb{H}^d$.

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 $\boldsymbol{A} \in \mathbb{H}^d(I)$, we define the standard matrix function via the rule

$$f(\boldsymbol{A}) := \sum\nolimits_{k=1}^{d} f(\lambda_k)\, \boldsymbol{u}_k \boldsymbol{u}_k^* \quad \text{where} \quad \boldsymbol{A} = \sum\nolimits_{k=1}^{d} \lambda_k\, \boldsymbol{u}_k \boldsymbol{u}_k^*$$

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

For $\boldsymbol{M} \in \mathbb{M}^d$, we write $\operatorname{Re}(\boldsymbol{M}) := \frac{1}{2}(\boldsymbol{M} + \boldsymbol{M}^*)$ for the Hermitian part of $\boldsymbol{M}$. The following semidefinite relation holds.

$$\operatorname{Re}(\boldsymbol{A}\boldsymbol{B}) = \frac{\boldsymbol{A}\boldsymbol{B} + \boldsymbol{B}\boldsymbol{A}}{2} \preccurlyeq \frac{\boldsymbol{A}^2 + \boldsymbol{B}^2}{2} \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}^d. \tag{1.2}$$

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

$$\left(\frac{\boldsymbol{A} + \boldsymbol{B}}{2}\right)^2 \preccurlyeq \frac{\boldsymbol{A}^2 + \boldsymbol{B}^2}{2} \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}^d. \tag{1.3}$$

In other words, the matrix square is operator convex.

Finally, we need two additional families of matrix norms. For $p \in [1, \infty]$, the Schatten p-norm is given by

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

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

$$\|\boldsymbol{B}\|_{p \to p} := \sup_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{B}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p} \quad \text{for each } \boldsymbol{B} \in \mathbb{M}^d \tag{1.4}$$

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.

2. Exchangeable Pairs of Random Matrices

The basic principle behind this paper is that we can exploit the symmetries of the distribution of a random matrix to obtain matrix concentration inequalities. One way to encode symmetries is to identify an exchangeable counterpart of the random matrix. This section outlines the main concepts from the method of exchangeable pairs, including an example of fundamental importance. Once we have an exchangeable pair, we can apply ideas of Chatterjee [4] to obtain concentration inequalities, which is the subject of Section 3.

2.1. Kernel Stein Pairs

In this work, the primal concept is an exchangeable pair of random variables.

Definition 2.1 (Exchangeable Pair)**.** Let Z and Z' be a pair of random variables taking values in a Polish space $\mathcal{Z}$. We say that a (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 function f where the expectations are finite.

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.

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

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

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

$$\boldsymbol{K}(Z, Z') = -\boldsymbol{K}(Z', Z) \quad \text{and} \quad \mathbb{E}[\boldsymbol{K}(Z, Z') \,|\, Z] = \boldsymbol{X} \quad \text{almost surely.} \tag{2.1}$$

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

It turns out that most exchangeable pairs of random matrices admit a kernel $\boldsymbol{K}$ that satisfies (2.1). We describe the construction in Section 2.2.

Kernel Stein Pairs versus Matrix Stein Pairs. The analysis in the article [13] is based on an important subclass of kernel Stein pairs termed *matrix Stein pairs*. A matrix Stein pair $(\boldsymbol{X}, \boldsymbol{X}')$ derived from an auxiliary exchangeable pair (Z, Z') satisfies the stronger condition

$$\mathbb{E}[\boldsymbol{X} - \boldsymbol{X}' \,|\, Z] = \alpha \boldsymbol{X} \quad \text{for some } \alpha > 0. \tag{2.2}$$

That is, a matrix Stein pair is a kernel Stein pair with $\boldsymbol{K}(Z, Z') = \alpha^{-1}(\boldsymbol{X} - \boldsymbol{X}')$. Although the paper [13] describes several fundamental classes of matrix Stein pairs, most exchangeable pairs of random matrices do not satisfy the condition (2.2). Kernel Stein pairs are much more common, so they are commensurately more useful.

2.2. Kernel Couplings

Given an exchangeable pair of random matrices, we can ask whether it is possible to equip the pair with a kernel that satisfies (2.1). In fact, there is a very general construction that works whenever the exchangeable pair is suitably ergodic. This method depends on an idea of Chatterjee [4, Sec. 4.1] that ultimately relies on an observation of Stein [21].

Stein noticed that any exchangeable pair (Z, Z') of $\mathcal{Z}$-valued random variables defines a reversible Markov chain with a symmetric transition kernel P given by

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

for each integrable function $f : \mathcal{Z} \to \mathbb{R}$. 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 (2.1) for any exchangeable pair of random matrices constructed from the auxiliary exchangeable pair (Z, Z'). To explain this idea, we begin with a definition.

Definition 2.3 (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,

$$Z_{(i)} \perp\!\!\!\perp Z'_{(0)} \,|\, Z_{(0)} \quad \text{and} \quad Z'_{(i)} \perp\!\!\!\perp Z_{(0)} \,|\, Z'_{(0)} \quad \text{for all } i. \tag{2.3}$$

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

The key lemma, essentially due to Chatterjee [4, Sec. 4.1], allows us to construct a kernel Stein pair by way of a kernel coupling.

Lemma 2.4. *Let $(Z_{(i)}, Z'_{(i)})_{i\geq 0}$ be a kernel coupling for an exchangeable pair $(Z, Z') \in \mathcal{Z}^2$. Let $\boldsymbol{\Psi} : \mathcal{Z} \to \mathbb{H}^d$ be a measurable function with $\mathbb{E}\,\boldsymbol{\Psi}(Z) = \mathbf{0}$. Suppose that there is a positive constant L for which*

$$\sum\nolimits_{i=0}^{\infty} \left\| \mathbb{E}[\boldsymbol{\Psi}(Z_{(i)}) - \boldsymbol{\Psi}(Z'_{(i)}) \,|\, Z_{(0)} = z, Z'_{(0)} = z'] \right\| \leq L \quad \text{for all } z, z' \in \mathcal{Z}. \tag{2.4}$$

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

$$\boldsymbol{K}(Z, Z') := \sum\nolimits_{i=0}^{\infty} \mathbb{E}[\boldsymbol{\Psi}(Z_{(i)}) - \boldsymbol{\Psi}(Z'_{(i)}) \,|\, Z_{(0)} = Z, Z'_{(0)} = Z']. \tag{2.5}$$

The proof of this result is identical with that of [4, Lem. 4.2], which establishes the same formula (2.5) in the scalar setting. Lemma 2.4 indicates that the kernel $\boldsymbol{K}$ associated with an exchangeable pair (Z, Z') and a map $\boldsymbol{\Psi}$ tends to be small when the two Markov chains in the kernel coupling have a small coupling time.

2.3. Conditional Variance

To each kernel Stein pair $(\boldsymbol{X}, \boldsymbol{X}')$, we may associate two random matrices called the *conditional variance* and *kernel conditional variance* of $\boldsymbol{X}$. Ultimately, we show that $\boldsymbol{X}$ is concentrated around the zero matrix whenever the conditional variance and the kernel conditional variance are both small.

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

$$\boldsymbol{V}_{\boldsymbol{X}} := \boldsymbol{V}_{\boldsymbol{X}}(Z) := \frac{1}{2}\,\mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{X}')^2 \,|\, Z\right], \tag{2.6}$$

and the *kernel conditional variance* is the random matrix

$$\boldsymbol{V}^{\boldsymbol{K}} := \boldsymbol{V}^{\boldsymbol{K}}(Z) := \frac{1}{2}\,\mathbb{E}\left[\boldsymbol{K}(Z, Z')^2 \,|\, Z\right]. \tag{2.7}$$

The following lemma provides a convenient way to control the conditional variance and the kernel conditional variance when the kernel is obtained from a kernel coupling as in Lemma 2.4.

Lemma 2.6. *Let $(Z_{(i)}, Z'_{(i)})_{i\geq 0}$ be a kernel coupling for an exchangeable pair $(Z, Z') \in \mathcal{Z}^2$, and let $\boldsymbol{\Psi} : \mathcal{Z} \to \mathbb{H}^d$ be a measurable map. Suppose that $(\boldsymbol{X}, \boldsymbol{X}') = (\boldsymbol{\Psi}(Z), \boldsymbol{\Psi}(Z'))$ is a kernel Stein pair where the kernel $\boldsymbol{K}$ is constructed via (2.5). For each $i = 0, 1, 2, \dots$, assume that*

$$\mathbb{E}\left[\, \mathbb{E}[\boldsymbol{\Psi}(Z_{(i)}) - \boldsymbol{\Psi}(Z'_{(i)}) \,|\, Z, Z']^2 \,|\, Z\right] \preccurlyeq s_i^2\, \boldsymbol{\Gamma}(Z) \quad \text{almost surely,} \tag{2.8}$$

where $\boldsymbol{\Gamma} : \mathcal{Z} \to \mathbb{H}^d$ is a measurable map and $(s_i)_{i\geq 0}$ is a deterministic sequence of nonnegative numbers. Then the conditional variance (2.6) satisfies

$$\boldsymbol{V}_{\boldsymbol{X}} \preccurlyeq \frac{1}{2} s_0^2\, \boldsymbol{\Gamma}(Z) \quad \text{almost surely,}$$

and the kernel conditional variance (2.7) satisfies

$$\boldsymbol{V}^{\boldsymbol{K}} \preccurlyeq \frac{1}{2} \left(\sum\nolimits_{i=0}^{\infty} s_i \right)^2 \boldsymbol{\Gamma}(Z) \quad \text{almost surely.}$$

Proof. Using a continuity argument, we may assume that each $s_i > 0$ for each integer $i \geq 0$. For each i, define $\boldsymbol{Y}_i := \mathbb{E}[\boldsymbol{\Psi}(Z_{(i)}) - \boldsymbol{\Psi}(Z'_{(i)}) \,|\, Z, Z']$. By the kernel coupling construction (2.5), we have

$$\begin{aligned}
\boldsymbol{V}^{\boldsymbol{K}} &= \frac{1}{2} \sum\nolimits_{i=0}^{\infty} \sum\nolimits_{j=0}^{\infty} \mathbb{E}[\boldsymbol{Y}_i \boldsymbol{Y}_j \,|\, Z] = \frac{1}{2} \sum\nolimits_{i=0}^{\infty} \sum\nolimits_{j=0}^{\infty} \mathbb{E}[\mathrm{Re}(\boldsymbol{Y}_i \boldsymbol{Y}_j) \,|\, Z] \\
&\preccurlyeq \frac{1}{2} \sum\nolimits_{i=0}^{\infty} \sum\nolimits_{j=0}^{\infty} \frac{1}{2} \left(\frac{s_j}{s_i} \mathbb{E}[\boldsymbol{Y}_i^2 \,|\, Z] + \frac{s_i}{s_j} \mathbb{E}[\boldsymbol{Y}_j^2 \,|\, Z] \right) \\
&\preccurlyeq \frac{1}{2} \sum\nolimits_{i=0}^{\infty} \sum\nolimits_{j=0}^{\infty} \frac{1}{2} \left(\frac{s_j}{s_i} s_i^2\, \boldsymbol{\Gamma}(Z) + \frac{s_i}{s_j} s_j^2\, \boldsymbol{\Gamma}(Z) \right) \\
&= \frac{1}{2} \left(\sum\nolimits_{i=0}^{\infty} s_i \sum\nolimits_{j=0}^{\infty} s_j \right) \boldsymbol{\Gamma}(Z) = \frac{1}{2} \left(\sum\nolimits_{i=0}^{\infty} s_i \right)^2 \boldsymbol{\Gamma}(Z),
\end{aligned}$$

where the first semidefinite inequality follows from (1.2) and the second inequality depends on the hypothesis (2.8). Similarly,

$$\boldsymbol{V}_{\boldsymbol{X}} = \frac{1}{2} \mathbb{E}[\boldsymbol{Y}_0^2 \,|\, Z] \preccurlyeq \frac{1}{2} s_0^2\, \boldsymbol{\Gamma}(Z).$$

This observation completes the proof. □

2.4. Example: Matrix Functions of Independent Variables

To illustrate the definitions in this section, we describe a simple but important example of a kernel Stein pair. Suppose that $Z := (Z_1, \dots, Z_n)$ is a vector of independent random variables taking values in a Polish space $\mathcal{Z}$. Let $\boldsymbol{H} : \mathcal{Z} \to \mathbb{H}^d$ be a measurable function, and let $(\boldsymbol{A}_j)_{j\geq 1}$ be a sequence of deterministic Hermitian matrices satisfying

$$(\boldsymbol{H}(z_1, \dots, z_n) - \boldsymbol{H}(z_1, \dots, z'_j, \dots, z_n))^2 \preccurlyeq \boldsymbol{A}_j^2 \tag{2.9}$$

where z_j, z_j' range over the possible values of Z_j for each j. We aim to analyze the random matrix

$$\boldsymbol{X} := \boldsymbol{H}(Z) - \mathbb{E}\,\boldsymbol{H}(Z). \tag{2.10}$$

We encounter matrices of this form in a variety of applications. For instance, concentration inequalities for the norm of $\boldsymbol{X}$ have immediate implications for the generalization properties of algorithms for multiclass classification [11, 15].

In this section, we explain how to construct a kernel exchangeable pair for studying the random matrix (2.10), and we compute the conditional variance and kernel conditional variance. Later, in Section 4, we use these calculations to establish a matrix bounded difference inequality that improves on [23, Cor 7.5].

To begin, we form an exchangeable counterpart for Z:

$$Z' := (Z_1, \dots, Z_{J-1}, \tilde{Z}_J, Z_{J+1}, \dots, Z_n)$$

where $\tilde{Z} := (\tilde{Z}_1, \dots, \tilde{Z}_n)$ is an independent copy of Z. We draw the coordinate J uniformly at random from $\{1, \dots, n\}$, independent from everything else. Then the random matrix

$$\boldsymbol{X}' := \boldsymbol{H}(Z') - \mathbb{E}\,\boldsymbol{H}(Z)$$

is an exchangeable counterpart for the matrix $\boldsymbol{X}$.

To verify that $(\boldsymbol{X}, \boldsymbol{X}')$ is a kernel Stein pair for a suitable kernel $\boldsymbol{K}$, we establish an explicit kernel coupling $(Z_{(i)}, Z'_{(i)})_{i\geq 0}$. For each $i \geq 1$, define $\tilde{Z}_{(i)}$ to be an independent copy of Z. We generate the pair $(Z_{(i)}, Z'_{(i)})$ from the previous pair $(Z_{(i-1)}, Z'_{(i-1)})$ by selecting an independent random index J_i uniformly from $\{1, \dots, n\}$ and replacing the J_i-th coordinates of both $Z_{(i-1)}$ and $Z'_{(i-1)}$ with the J_i-th coordinate of $\tilde{Z}_{(i)}$. By construction, the two marginal chains $(Z_{(i)})_{i\geq 0}$ and $(Z'_{(i)})_{i\geq 0}$ evolve according to the transition kernel induced by (Z, Z'), and they satisfy the kernel coupling property (2.3). The analysis of the coupon collector's problem [9, Sec. 2.2] shows that the expected coupling time for this pair of Markov chains is bounded by $n(1+\log n)$. Therefore, Lemma 2.4 implies that $(\boldsymbol{X}, \boldsymbol{X}')$ is a kernel Stein pair with

$$\boldsymbol{K}(Z, Z') := \sum\nolimits_{i=0}^{\infty} \mathbb{E}[\boldsymbol{H}(Z_{(i)}) - \boldsymbol{H}(Z'_{(i)}) \,|\, Z_{(0)} = Z, Z'_{(0)} = Z'].$$

Since the two Markov chains couple rapidly, we expect that the kernel is small.

To bound the size of the kernel, we use Lemma 2.6. For each integer $i \geq 0$, define the event $\mathcal{E}_i := \{J \notin \{J_1, \dots, J_i\}\}$. Off of the event $\mathcal{E}_i$, we have $\boldsymbol{H}(Z_{(i)}) = \boldsymbol{H}(Z'_{(i)})$; on the event $\mathcal{E}_i$, the random vectors $Z_{(i)}$ and $Z'_{(i)}$ can differ only in the J-th coordinate. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\,\mathbb{E}[\boldsymbol{H}(Z_{(i)}) - \boldsymbol{H}(Z'_{(i)}) \,|\, Z, Z']^2 \,|\, Z\right] \\
&\qquad = \mathbb{E}\left[(\mathbb{P}\,\{\mathcal{E}_i\} \cdot \mathbb{E}[\boldsymbol{H}(Z_{(i)}) - \boldsymbol{H}(Z'_{(i)}) \,|\, Z, Z', \mathcal{E}_i])^2 \,|\, Z\right] \\
&\qquad \preccurlyeq (1 - 1/n)^{2i} \cdot \mathbb{E}[(\boldsymbol{H}(Z_{(i)}) - \boldsymbol{H}(Z'_{(i)}))^2 \,|\, Z, \mathcal{E}_i] \\
&\qquad \preccurlyeq (1 - 1/n)^{2i} \cdot \mathbb{E}[\boldsymbol{A}_J^2].
\end{aligned}$$

The first semidefinite inequality follows from the convexity (1.3) of the matrix square, and the second depends on our bounded differences assumption (2.9). Apply Lemma 2.6 with $s_i = (1 - 1/n)^i$ and $\boldsymbol{\Gamma}(Z) = \mathbb{E}[\boldsymbol{A}_J^2]$ to conclude that

$$\boldsymbol{V}^{\boldsymbol{K}} \preccurlyeq \frac{1}{2}\,\mathbb{E}[\boldsymbol{A}_J^2] \left(\sum\nolimits_{i=0}^{\infty} (1 - 1/n)^i\right)^2 = \frac{n^2}{2}\,\mathbb{E}[\boldsymbol{A}_J^2] = \frac{n}{2} \sum\nolimits_{j=1}^{n} \boldsymbol{A}_j^2 \tag{2.11}$$

and that

$$\boldsymbol{V}_{\boldsymbol{X}} \preccurlyeq \frac{1}{2}\,\mathbb{E}[\boldsymbol{A}_J^2] = \frac{1}{2n}\sum\nolimits_{j=1}^{n} \boldsymbol{A}_j^2. \tag{2.12}$$

We discover that the conditional variance and the kernel conditional variance are under control when $\boldsymbol{H}$ has bounded coordinate differences. Section 4 discusses how these estimates imply that the matrix $\boldsymbol{X}$ concentrates well.

3. Concentration Inequalities for Random Matrices

This section contains our main results on concentration for random matrices. Given a kernel Stein pair, we explain how the conditional variance and kernel conditional variance allow us to obtain exponential tail bounds and polynomial moment inequalities.

At a high level, our work suggests the following plan of action. You begin with a random matrix, $\boldsymbol{X}$. You use the symmetries of the random matrix to construct an exchangeable counterpart, $\boldsymbol{X}'$, that is close but not identical to $\boldsymbol{X}$. You construct a kernel coupling from this exchangeable pair, and you compute the conditional variances, $\boldsymbol{V}_{\boldsymbol{X}}$ and $\boldsymbol{V}_{\boldsymbol{K}}$. Then you apply the concentration results from this section to control the deviation of $\boldsymbol{X}$ from its mean. In the sections to come, we provide specific examples and applications of this template.

3.1. Exponential Tail Bounds

Our first result establishes exponential concentration for the maximum and minimum eigenvalues of a random matrix.

Theorem 3.1 (Concentration for Bounded Random Matrices)**.** *Consider a $\boldsymbol{K}$-Stein pair $(\boldsymbol{X}, \boldsymbol{X}') \in \mathbb{H}^d \times \mathbb{H}^d$. Suppose there exist nonnegative constants c, v, s for which the conditional variance (2.6) and the kernel conditional variance (2.7) of the pair satisfy*

$$\boldsymbol{V}_{\boldsymbol{X}} \preccurlyeq s^{-1} \cdot (c\boldsymbol{X} + v\,\mathbf{I}) \quad \text{and} \quad \boldsymbol{V}^{\boldsymbol{K}} \preccurlyeq s \cdot (c\boldsymbol{X} + v\,\mathbf{I}) \quad \text{almost surely.} \tag{3.1}$$

Then, for all $t \geq 0$,

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

Furthermore,

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

Theorem 3.1 extends the concentration result of [13, Thm. 4.1], which only applies to matrix Stein pairs. The argument leading up to Theorem 3.1 is very similar with the proof of the earlier result. The main innovation is a new type of mean value inequality for matrices that improves on [13, Lem. 3.4].

Lemma 3.2 (Exponential Mean Value Trace Inequality)**.** *For all matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{H}^d$ and all $s > 0$ it holds that*

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

See Appendix B for the proofs of Theorem 3.1 and Lemma 3.2.

3.2. Polynomial Moment Inequalities

The second main result shows that we can bound the polynomial moments of a random matrix in terms of the conditional variance and the kernel conditional variance.

Theorem 3.3 (Matrix BDG Inequality)**.** *Suppose that $(\boldsymbol{X}, \boldsymbol{X}')$ is a $\boldsymbol{K}$-Stein pair based on an auxiliary exchangeable pair (Z, Z'). Let $p \geq 1$ be a natural number, and assume that $\mathbb{E}\,\|\boldsymbol{X}\|_{S_{2p}}^{2p} < \infty$ and $\mathbb{E}\,\|\boldsymbol{K}(Z, Z')\|^{2p} < \infty$. Then, for any $s > 0$,*

$$\left(\mathbb{E}\,\|\boldsymbol{X}\|_{S_{2p}}^{2p}\right)^{1/2p} \leq \sqrt{2p-1}\left(\mathbb{E}\left\|\frac{1}{2}(s\,\boldsymbol{V}_{\boldsymbol{X}} + s^{-1}\,\boldsymbol{V}^{\boldsymbol{K}})\right\|_{S_p}^{p}\right)^{1/2p}.$$

We have written $\|\cdot\|_{S_p}$ for the Schatten p-norm.

Theorem 3.3 generalizes the matrix Burkholder–Davis–Gundy inequality [13, Thm. 7.1], which only applies to matrix Stein pairs. This result depends on another novel mean value inequality for matrices.

Lemma 3.4 (Polynomial Mean Value Trace Inequality)**.** *For all matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{H}^d$, all integers $q \geq 1$, and all $s > 0$, it holds that*

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

The proofs of Theorem 3.3 and Lemma 3.4 can be found in Appendix C. We remark that both results extend directly to infinite-dimensional Schatten-class operators.

4. Example: Matrix Bounded Differences Inequality

As a first example, we show how to use Theorem 3.1 to derive a matrix version of McDiarmid's bounded differences inequality [14].

Corollary 4.1 (Matrix Bounded Differences)**.** *Suppose that $Z := (Z_1, \ldots, Z_n) \in \mathcal{Z}$ is a vector of independent random variables that takes values in a Polish space $\mathcal{Z}$. Let $\boldsymbol{H} : \mathcal{Z} \to \mathbb{H}^d$ be a measurable function, and let $(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n)$ be a deterministic sequence of Hermitian matrices that satisfy*

$$(\boldsymbol{H}(z_1, \ldots, z_n) - \boldsymbol{H}(z_1, \ldots, z_j', \ldots, z_n))^2 \preccurlyeq \boldsymbol{A}_j^2$$

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

$$\sigma^2 := \left\|\sum\nolimits_{j=1}^{n} \boldsymbol{A}_j^2\right\|.$$

Then, for all $t \geq 0$,

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

Furthermore,

$$\mathbb{E}\,\lambda_{\max}\left(\boldsymbol{H}(Z) - \mathbb{E}\,\boldsymbol{H}(Z)\right) \leq \sigma\sqrt{\log d}.$$

Proof. Introduce the random matrix $\boldsymbol{X} = \boldsymbol{H}(Z) - \mathbb{E}\,\boldsymbol{H}(Z)$. We can use the kernel Stein pair constructed in Section 2.4 to study the behavior of $\boldsymbol{X}$. According to (2.12), the conditional variance satisfies

$$\boldsymbol{V}_{\boldsymbol{X}} \preccurlyeq \frac{1}{2n}\sum\nolimits_{j=1}^{n} \boldsymbol{A}_j^2 \preccurlyeq \left(\frac{\sigma^2}{2}\mathbf{I}\right)/n.$$

According to (2.11), the kernel conditional variance satisfies

$$\boldsymbol{V}^{\boldsymbol{K}} \preccurlyeq \frac{n}{2}\sum\nolimits_{j=1}^{n} \boldsymbol{A}_j^2 \preccurlyeq n\left(\frac{\sigma^2}{2}\mathbf{I}\right),$$

Invoke Theorem 3.1 with $c = 0$, $v = \sigma^2/2$, and $s = n$ to complete the bound. □

Corollary 4.1 improves on the matrix bounded differences inequality [23, Cor. 7.5], which features an additional factor of 1/8 in the exponent of the tail bound. It also strengthens the bounded differences inequality [13, Cor. 11.1] for matrix Stein pairs, which requires an extra assumption that the function $\boldsymbol{H}$ is "self-reproducing."

Remark 4.2 (Extensions). The conclusions of Corollary 4.1 hold with $\sigma^2 := \left\|\boldsymbol{A}^2\right\|$ under either one of the weaker hypotheses

$$\sum\nolimits_j (\boldsymbol{H}(z_1, \ldots, z_n) - \boldsymbol{H}(z_1, \ldots, z_j', \ldots, z_n))^2 \preccurlyeq \boldsymbol{A}^2$$

or

$$\sum\nolimits_j \mathbb{E}\left[(\boldsymbol{H}(z_1, \ldots, z_n) - \boldsymbol{H}(z_1, \ldots, Z_j \ldots, z_n))^2\right] \preccurlyeq \boldsymbol{A}^2$$

where $\boldsymbol{A} \in \mathbb{H}^d$ is deterministic and z_k, z_k' range over all possible values of Z_k for each index k. This claim follows from a simple adaptation of the argument in Section 2.4.

We can also obtain moment inequalities for the random matrix $\boldsymbol{H}(\boldsymbol{Z})$ by invoking Theorem 3.3. We have omitted a detailed statement because exponential tail bounds are more popular in applications.

5. Example: 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 4.1 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 [7]. This concept involves a certain amount of auxiliary notation. Given a vector $\boldsymbol{x} = (x_1, \ldots, x_n)$, we write $\boldsymbol{x}_{-i} = (x_1, \ldots x_{i-1}, x_{i+1}, \ldots, x_n)$ for the vector with its ith component deleted. Let $Z = (Z_1, \ldots, Z_n)$ 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 $(\mathcal{Z}, \mathcal{F})$:

$$d_{\mathrm{TV}}(\nu, \mu) := \sup_{A \in \mathcal{F}} |\nu(A) - \mu(A)| . \tag{5.1}$$

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

Definition 5.1 (Dobrushin Interdependence Matrix)**.** Let $Z = (Z_1, \dots, Z_n)$ be a random vector taking values in a Polish space $\mathcal{Z}$. Let $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ be a matrix with a zero diagonal that satisfies the condition

$$d_{\mathrm{TV}}\big(\mu_i(\cdot \,|\, \boldsymbol{x}_{-i}), \mu_i(\cdot \,|\, \boldsymbol{y}_{-i})\big) \le \sum\nolimits_{j=1}^{n} D_{ij} \mathbb{1}[x_j \neq y_j] \tag{5.2}$$

for each index i and for all vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{Z}$. Then $\boldsymbol{D}$ 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 4.1. This statement can be viewed as a matrix version of Chatterjee's result [4, Thm. 4.3].

Corollary 5.2 (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 $\boldsymbol{D}$ with the property that*

$$\max\left\{ \|\boldsymbol{D}\|_{1\to 1}, \ \|\boldsymbol{D}\|_{\infty\to\infty} \right\} < 1. \tag{5.3}$$

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

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

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

$$\sigma^2 := \Big\| \sum\nolimits_{j=1}^{n} \boldsymbol{A}_j^2 \Big\| \quad \text{and} \quad b := \left[1 - \frac{1}{2}\big(\|\boldsymbol{D}\|_{1\to 1} + \|\boldsymbol{D}\|_{\infty\to\infty} \big) \right]^{-1}.$$

Then, for all $t \ge 0$,

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

Furthermore,

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

The proof of Corollary 5.2 appears below in Section 5.1. In Section 6, we describe an application of the result to physical spin systems in Section 6. Observe that the bounds here are a factor of b worse than the independent case outlined in Corollary 4.1.

5.1. Proof of Concentration under Dobrushin Assumptions

The proof of Corollary 5.2 is longer than the argument behind Corollary 4.1, but it follows the same pattern.

Exchangeable Counterparts. Let $\boldsymbol{X} = \boldsymbol{H}(Z) - \mathbb{E}\,\boldsymbol{H}(Z)$. To begin, we form exchangeable counterparts for the random input Z and the random matrix $\boldsymbol{X}$.

$$Z' := (Z_1, \dots, Z_{J-1}, \tilde{Z}_J, Z_{J+1}, \dots, Z_n) \quad \text{and} \quad \boldsymbol{X}' := \boldsymbol{H}(Z') - \mathbb{E}\,\boldsymbol{H}(Z)$$

where J is an independent index drawn uniformly from $\{1, \dots, n\}$ and $\tilde{Z}_i$ and Z_i are conditionally i.i.d. given Z_{-i} for each index i.

A Kernel Coupling. Next, we construct a kernel coupling $(Z_{(i)}, Z'_{(i)})_{i\geq 0}$ by adapting the proof of [4, Thm. 4.3]. For each $i \geq 1$, we generate $(Z_{(i)}, Z'_{(i)})$ from $(Z_{(i-1)}, Z'_{(i-1)})$ by selecting an independent random index J_i uniformly from $\{1, \dots, n\}$ and replacing the J_i-th coordinates of $Z_{(i-1)}$ and $Z'_{(i-1)}$ with $\tilde{Z}_{(i-1),J_i}$ and $\tilde{Z}'_{(i-1),J_i}$ respectively. The replacement variables are sampled so that

$$Z_{(i-1),j} \perp\!\!\!\perp \tilde{Z}_{(i-1),j} \,|\, Z_{(i-1),-j} \quad \text{and} \quad Z'_{(i-1),j} \perp\!\!\!\perp \tilde{Z}'_{(i-1),j} \,|\, Z'_{(i-1),-j}.$$

We require that $\tilde{Z}_{(i-1),j}$ and $\tilde{Z}'_{(i-1),j}$ are maximally coupled, i.e.,

$$\mathbb{P}\left\{\tilde{Z}_{(i-1),j} \neq \tilde{Z}'_{(i-1),j} \,|\, Z_{(i-1)}, Z'_{(i-1)}\right\} = d_{\mathrm{TV}}\big(\mu_j(\cdot \,|\, Z_{(i-1),-j}), \mu_j(\cdot \,|\, Z'_{(i-1),-j})\big).$$

By construction, the two marginal chains $(Z_{(i)})_{i\geq 0}$ and $(Z'_{(i)})_{i\geq 0}$ have the same the kernel as (Z, Z'), and they satisfy the kernel coupling property (2.3). Furthermore, the coupling boundedness criterion (2.4) is met, just as in the scalar setting [4, p. 78]. Lemma 2.4 now implies that $(\boldsymbol{X}, \boldsymbol{X}')$ is a kernel Stein pair with kernel

$$\boldsymbol{K}(\boldsymbol{z}, \boldsymbol{z}') := \sum\nolimits_{i=0}^{\infty} \mathbb{E}\left[\boldsymbol{H}(Z_{(i)}) - \boldsymbol{H}(Z'_{(i)}) \,|\, Z_{(0)} = \boldsymbol{z}, Z'_{(0)} = \boldsymbol{z}'\right].$$

The Conditional Variances. With the kernel coupling established, we may proceed to analyze the conditional variances $\boldsymbol{V}_{\boldsymbol{X}}$ and $\boldsymbol{V}^{\boldsymbol{K}}$. First, we collect the information necessary to apply Lemma 2.6. Fix an index $i \geq 0$, and write $\boldsymbol{H}(Z_{(i)}) - \boldsymbol{H}(Z'_{(i)})$ as a telescoping sum:

$$\begin{aligned}\boldsymbol{H}(Z_{(i)}) - \boldsymbol{H}(Z'_{(i)}) = \sum_{j=1}^{n} \Big[&\boldsymbol{H}\left(Z_{(i),1}, \dots, Z_{(i),j}, Z'_{(i),j+1}, \dots, Z'_{(i),n}\right) \\ &- \boldsymbol{H}\left(Z_{(i),1}, \dots, Z_{(i),j-1}, Z'_{(i),j}, \dots, Z'_{(i),n}\right)\Big] =: \sum_{j=1}^{n} \boldsymbol{W}_{(i),j}.\end{aligned}$$

Introduce the event $\mathcal{E}_{(i),j} := \{Z_{(i),j} \neq Z'_{(i),j}\}$. Abbreviate $p_{(i),j} = \mathbb{P}\left\{\mathcal{E}_{(i),j} \,|\, Z, Z'\right\}$ and $\tilde{\boldsymbol{W}}_{(i),j} = \mathbb{E}[\boldsymbol{W}_{(i),j} \,|\, Z, Z', \mathcal{E}_{(i),j}]$. Off of the event $\mathcal{E}_{(i),j}$, it holds that $\boldsymbol{W}_{(i),j} = \boldsymbol{0}$. Therefore,

$$\mathbb{E}[\boldsymbol{W}_{(i),j} \,|\, Z, Z'] = \tilde{\boldsymbol{W}}_{(i),j}\; p_{(i),j}.$$

In [4, pp. 77–78], Chatterjee established that, for each i and j,

$$p_{(i),j} \leq \mathbf{e}_j^* \boldsymbol{B}^i \mathbf{e}_J \quad \text{for} \quad \boldsymbol{B} := \left(1 - \frac{1}{n}\right)\mathbf{I} + \frac{1}{n}\boldsymbol{D}. \tag{5.4}$$

We use $\mathbf{e}_k$ to denote the kth standard basis vector, and $\boldsymbol{B}^i$ refers to the ith power of the square, nonnegative matrix $\boldsymbol{B}$.

To continue, make the calculation

$$\begin{aligned}\Big(\sum\nolimits_{j=1}^{n} \mathbb{E}[\boldsymbol{W}_{(i),j} \,|\, Z, Z']\Big)^2 &= \sum\nolimits_{j=1}^{n}\sum\nolimits_{k=1}^{n} \tilde{\boldsymbol{W}}_{(i),j}\tilde{\boldsymbol{W}}_{(i),k}\; p_{(i),j}p_{(i),k} \\ &\preccurlyeq \sum\nolimits_{1\le j,k\le n} \frac{1}{2}(\tilde{\boldsymbol{W}}_{(i),j}^2 + \tilde{\boldsymbol{W}}_{(i),k}^2)\; p_{(i),j}p_{(i),k} \\ &\preccurlyeq \sum\nolimits_{1\le j,k\le n} \boldsymbol{A}_k^2 \cdot \mathbf{e}_j^*\boldsymbol{B}^i\mathbf{e}_J \cdot \mathbf{e}_k^*\boldsymbol{B}^i\mathbf{e}_J \\ &= \big\|\boldsymbol{B}^i\mathbf{e}_J\big\|_1 \cdot \sum\nolimits_{k=1}^{n} \boldsymbol{A}_k^2 \cdot \mathbf{e}_k^*\boldsymbol{B}^i\mathbf{e}_J \\ &\preccurlyeq \|\boldsymbol{B}\|_{1\to 1}^i \cdot \sum\nolimits_{k=1}^{n} \boldsymbol{A}_k^2 \cdot \mathbf{e}_k^*\boldsymbol{B}^i\mathbf{e}_J.\end{aligned}$$

The first semidefinite inequality follows from (1.2). The second relation depends on (5.4). We reach the next identity by summing over j, noting that $\mathbf{e}_j^*\boldsymbol{B}^i\mathbf{e}_J$ is nonnegative. The last inequality follows from the definition (1.4) of $\|\cdot\|_{1\to 1}$ and the fact that this norm is submultiplicative. Next, take the expectation of the latter display with respect to J. We obtain

$$\begin{aligned}\mathbb{E}\left[\Big(\sum\nolimits_{j=1}^{n} \mathbb{E}[\boldsymbol{W}_{(i),j} \,|\, Z, Z']\Big)^2 \,\middle|\, Z\right] &\preccurlyeq \|\boldsymbol{B}\|_{1\to 1}^i \cdot \frac{1}{n}\sum\nolimits_{1\le j,k\le n} \boldsymbol{A}_k^2 \cdot \mathbf{e}_k^*\boldsymbol{B}^i\mathbf{e}_j \\ &= \|\boldsymbol{B}\|_{1\to 1}^i \cdot \frac{1}{n}\sum\nolimits_{k=1}^{n} \boldsymbol{A}_k^2 \cdot \big\|\mathbf{e}_k^*\boldsymbol{B}^i\big\|_1 \\ &\preccurlyeq \|\boldsymbol{B}\|_{1\to 1}^i \cdot \|\boldsymbol{B}\|_{\infty\to\infty}^i \cdot \frac{1}{n}\sum\nolimits_{k=1}^{n} \boldsymbol{A}_k^2.\end{aligned}$$

The justifications are similar with those for the preceding calculation.

As a consequence of this bound, we are in a position to apply Lemma 2.6. Set $\boldsymbol{\Gamma}(Z) = n^{-1}\sum_{k=1}^{n} \boldsymbol{A}_k^2$ and $s_i = \|\boldsymbol{B}\|_{1\to 1}^{i/2}\,\|\boldsymbol{B}\|_{\infty\to\infty}^{i/2}$ for each $i \ge 0$. The lemma delivers

$$\begin{aligned}\boldsymbol{V}_{\boldsymbol{X}} &\preccurlyeq \frac{1}{2}\boldsymbol{\Gamma}(Z) \preccurlyeq \frac{\sigma^2}{2n}\cdot\mathbf{I}, \quad \text{and} \\ \boldsymbol{V}^{\boldsymbol{K}} &\preccurlyeq \frac{1}{2}\Big(\sum\nolimits_{i=0}^{\infty} s_i\Big)^2 \boldsymbol{\Gamma}(Z) \preccurlyeq \left(1 - \sqrt{\|\boldsymbol{B}\|_{\infty\to\infty}\,\|\boldsymbol{B}\|_{1\to 1}}\right)^{-2} \frac{\sigma^2}{2n}\cdot\mathbf{I}.\end{aligned}$$

where σ^2 is defined in the statement of Corollary 5.2. It remains to simplify the formula for the kernel conditional variance.

The definition of $\boldsymbol{B}$ ensures that

$$\|\boldsymbol{B}\|_{1\to 1} = 1 - \frac{1}{n}\left(1 - \|\boldsymbol{D}\|_{1\to 1}\right) \quad \text{and} \quad \|\boldsymbol{B}\|_{\infty\to\infty} = 1 - \frac{1}{n}\left(1 - \|\boldsymbol{D}\|_{\infty\to\infty}\right).$$

As a consequence of the geometric–arithmetic mean inequality,

$$1 - \sqrt{\|\boldsymbol{B}\|_{1\to 1}\,\|\boldsymbol{B}\|_{\infty\to\infty}} \ge \frac{1}{n}\left[1 - \frac{1}{2}\big(\,\|\boldsymbol{D}\|_{1\to 1} + \|\boldsymbol{D}\|_{\infty\to\infty}\,\big)\right].$$

We conclude that

$$\boldsymbol{V}^{\boldsymbol{K}} \preccurlyeq \left[1 - \frac{1}{2}\big(\,\|\boldsymbol{D}\|_{1\to 1} + \|\boldsymbol{D}\|_{\infty\to\infty}\,\big)\right]^{-2} \cdot \frac{n\sigma^2}{2}\cdot\mathbf{I} = \frac{nb^2\sigma^2}{2}\cdot\mathbf{I},$$

where b is defined in the statement of Corollary 5.2.

Finally, we invoke Theorem 3.1 with $c = 0$ and $v = b\sigma^2/2$ and $s = nb$ to obtain the advertised conclusions.

6. Application: Correlation in the 2D Ising Model

In this section, we apply the dependent matrix bounded differences inequality of Corollary 5.2 to study correlations in a simple spin system. Consider the 2D Ising model without an external field on an $n \times n$ square lattice with a periodic boundary. Let $\boldsymbol{\sigma} := (\sigma_{ij} : 1 \le i, j \le n)$ be an array of random spins taking values in $\{+1, -1\}$. To simplify the discussion, we treat array indices periodically, so we interpret the index i to mean $((i-1) \bmod n) + 1$. We also write $(i, j) \sim (k, l)$ to indicate that the vertices are neighbors in the periodic square lattice; that is, $k = i \pm 1$ and $l = j$ or else $k = i$ and $l = j \pm 1$. With this notation, the Hamiltonian may be expressed as

$$H(\boldsymbol{\sigma}) = \sum_{(i,j)\sim(k,l)} \sigma_{ij}\, \sigma_{kl},$$

where the sum occurs over distinct pairs of neighboring vertices. We assign a probability distribution to the array $\boldsymbol{\sigma}$ of spins:

$$\mathbb{P}\{\boldsymbol{\sigma}\} = \frac{1}{A} \exp\big(\beta\, H(\boldsymbol{\sigma})\big), \tag{6.1}$$

where $A = \sum_{\boldsymbol{\sigma}'} \exp\big(\beta\, H(\boldsymbol{\sigma}')\big)$ denotes the normalizing constant (also known as the partition function). This model has been studied extensively, and it is known to exhibit a phase transition at $\beta_c = \frac{1}{2}\log(1+\sqrt{2})$. For example, see [19].

Fix a positive number $d \le n$. For indices $1 \le i, j \le d$, we define the spin–spin correlation function as

$$c_{ij} = \mathbb{E}[\sigma_{11}\, \sigma_{ij}].$$

We write $\boldsymbol{C}$ for the $d \times d$ matrix whose entries are c_{ij}. The paper [25] of Wu offers an explicit expression for the correlations in the limit as the size n of the lattice tends to infinity. In particular, in the high-temperature regime $\beta < \beta_c$, the correlations decay exponentially. On the other hand, this is a limiting result and there is no analytic formula for finite lattices.

One may wish to estimate the spin–spin correlation matrix from a sampled value $\boldsymbol{\sigma}$ of the spins. We propose the estimator

$$\widehat{C}_{ij} := \frac{1}{n^2} \sum_{1 \le k,l \le n} \sigma_{kl} \cdot \sigma_{k+i-1, l+j-1} \quad \text{for } 1 \le i, j \le d. \tag{6.2}$$

The mean of $\widehat{\boldsymbol{C}}$ is the spin–spin correlation matrix $\boldsymbol{C}$, so it is natural to wonder about the deviations of the estimator from its mean value. We can use the concentration results from the previous section to quantify these fluctuations.

6.1. Concentration for General Matrices

Since the estimator $\widehat{\boldsymbol{C}}$ need not be Hermitian, we need a way to extend our techniques to general matrices. We employ a well-known device from operator theory, called the *Hermitian dilation* [23, Sec. 2.6].

Definition 6.1 (Hermitian dilation)**.** Consider a matrix $\boldsymbol{B} \in \mathbb{C}^{d_1 \times d_2}$, and set $d = d_1 + d_2$. The *Hermitian dilation* of $\boldsymbol{B}$ is the matrix

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

The dilation preserves spectral properties in the sense that $\lambda_{\max}(\mathcal{D}(\boldsymbol{B})) = \|\mathcal{D}(\boldsymbol{B})\| = \|\boldsymbol{B}\|$. Therefore,

$$\mathbb{P}\left\{\left\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\right\| \geq t\right\} = \mathbb{P}\left\{\lambda_{\max}\big(\mathcal{D}(\widehat{\boldsymbol{C}}) - \mathcal{D}(\boldsymbol{C})\big) \geq t\right\}. \tag{6.3}$$

Using this observation, we can obtain a tail bound for the spectral-norm error in the estimator $\widehat{\boldsymbol{C}}$ by studying its dilation.

6.2. Bounding the Dobrushin Coefficients

To apply Corollary 5.2, we need to bound the Dobrushin coefficients of the array $\boldsymbol{\sigma}$ of spins. Let $\boldsymbol{\sigma}' \in \{\pm 1\}^{n\times n}$ be a second independent draw from the Ising model. Extending our notation from before, we write $\mu_{ij}(\cdot \,|\, \boldsymbol{\sigma}_{-(i,j)})$ for the conditional distribution of σ_{ij} given the remaining variables. In our setting,

$$\begin{aligned} &d_{\mathrm{TV}}\big(\mu_{ij}(\cdot \,|\, \boldsymbol{\sigma}_{-(i,j)}),\ \mu_{ij}(\cdot \,|\, \boldsymbol{\sigma}'_{-(i,j)})\big) \\ &\qquad = \left|\mathbb{P}\left\{\sigma_{ij} = 1 \,\middle|\, \sum\nolimits_{(k,l):(i,j)\sim(k,l)} \sigma_{kl}\right\} - \mathbb{P}\left\{\sigma'_{ij} = 1 \,\middle|\, \sum\nolimits_{(k,l):(i,j)\sim(k,l)} \sigma'_{kl}\right\}\right|. \end{aligned} \tag{6.4}$$

It follows from (6.1) that

$$\mathbb{P}\left\{\sigma_{ij} = 1 \,\middle|\, \sum\nolimits_{(k,l):(i,j)\sim(k,l)} \sigma_{kl} = s\right\} = \frac{\exp(s\beta)}{\exp(s\beta) + \exp(-s\beta)} = \frac{1}{1 + \exp(-2s\beta)}$$

for each possible value $s \in \{-4, -2, 0, 2, 4\}$. Therefore, the expression (6.4) admits the upper bound

$$\frac{1}{1 + \exp(-4\beta)} - \frac{1}{2}$$

when $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ differ in a single coordinate. We may select the Dobrushin interdependence matrix

$$D_{(i,j),(k,l)} = \begin{cases} (1 + \exp(-4\beta))^{-1} - \frac{1}{2}, & \text{when} (i,j) \sim (k,l) \\ 0, & \text{otherwise.} \end{cases}$$

This matrix satisfies the Dobrushin condition (5.2). By direct computation,

$$\max\left\{\, \|\boldsymbol{D}\|_{1\to 1}\,,\ \|\boldsymbol{D}\|_{\infty\to\infty} \,\right\} \leq \frac{4}{1 - \exp(-4\beta)} - 2. \tag{6.5}$$

because every vertex has four neighbors. The right-hand side of (6.5) is smaller than one precisely when $\beta < \beta_D = \frac{1}{4}\log(3)$. Since $\beta_D < \beta_c$, the hypotheses of Corollary 5.2 are satisfied for only part of the high-temperature regime.

6.3. Tail Bound for the Estimator

We intend to apply Corollary 5.2 to the Hermitian matrix $\mathcal{D}(\widehat{\boldsymbol{C}})$. For each index $1 \leq i, j \leq n$, write $\widehat{\boldsymbol{C}}^{ij}$ for the value of $\widehat{\boldsymbol{C}}$ when the sign of σ_{ij} is flipped. From (6.2), we have the inequalities

$$\left|\widehat{C}_{kl} - \widehat{C}^{ij}_{kl}\right| \leq \frac{4}{n^2} \quad \text{for } 1 \leq k, l \leq d.$$

As a consequence, we reach the semidefinite relation

$$\left(\mathcal{D}(\widehat{\boldsymbol{C}}) - \mathcal{D}(\widehat{\boldsymbol{C}}_{ij})\right)^2 \preccurlyeq \frac{16d^2}{n^4}\mathbf{I}.$$

Summing over all vertices in the lattice, we obtain an inequality for the boundedness parameter

$$\sigma^2 = \left\| \sum\nolimits_{1\le i,j\le n} \frac{16d^2}{n^4} \cdot \mathbf{I} \right\| = \frac{16d^2}{n^2}.$$

For $\beta < \beta_D$, Corollary 5.2 implies that

$$\mathbb{P}\left\{ \left\| \widehat{\boldsymbol{C}} - \boldsymbol{C} \right\| \ge t \right\} \le 2d \cdot \exp\left(\frac{-t^2}{3 - 4(1+\exp(-4\beta))^{-1}} \cdot \frac{n^2}{16d^2} \right).$$

Therefore, the typical deviation $\mathbb{E}\left\| \widehat{\boldsymbol{C}} - \boldsymbol{C} \right\|$ has order $(d\sqrt{\log d})/n$. Therefore, in the regime where $\beta < \beta_D$, one sample suffices to obtain an accurate estimate of the spin–spin correlation matrix $\boldsymbol{C}$, provided that $n \gg d$.

7. Complements

The tools of Section 3 are applicable in a wide variety of settings. To indicate what might be possible, we briefly present another packaged concentration result. We also indicate some prospects for future research.

7.1. 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 [4, Thm. 4.6]. We provide the proof in Appendix D.

Corollary 7.1 (Concentration for Hermitian Functions of Haar Measures)**.** *Let $Z \sim \mu$ be Haar distributed on a compact topological group G, and let $\boldsymbol{\Psi} : G \to \mathbb{H}^d$ be a measurable function satisfying $\mathbb{E}\,\boldsymbol{\Psi}(Z) = \mathbf{0}$. Let $Y, Y_1, Y_2, \dots$ be i.i.d. random variables in G satisfying*

$$Y =_d Y^{-1} \quad \text{and} \quad zYz^{-1} =_d Y \quad \text{for all} \quad z \in G. \tag{7.1}$$

Compute the boundedness parameter

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

where μ_i is the distribution of the product $Y_i \cdots Y_1$,

$$\|\boldsymbol{\Psi}(z)\| \le R \quad \text{for all} \quad z \in G, \quad \text{and} \quad S^2 = \sup_{g\in G} \left\| \mathbb{E}\left[(\boldsymbol{\Psi}(g) - \boldsymbol{\Psi}(Yg))^2 \right] \right\|.$$

Then, for all $t \ge 0$,

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

Furthermore,

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

Corollary 7.1 relates the concentration of Hermitian functions to the convergence of random walks on a group. Since representation theory leads to a matrix model of compact groups, it is often natural to build random matrices from a group representation. In particular, Corollary 7.1 can be used to study matrices constructed from random permutations or random unitary matrices. We omit the details.

7.2. Conjectures and Consequences

Lugosi et al. [3] study a class of self-bounding (scalar) functions, which arise in applications in statistics and learning theory. They use log-Sobolev inequalities to obtain information about the concentration properties of these functions. It is also possible to perform the analysis using the method of exchangeable pairs.

Let us introduce the matrix analog of a self-bounding function.

Definition 7.2 (Self-bounding Matrix Function)**.** A function $\boldsymbol{H} : \mathcal{Z} \to \mathbb{H}^d$ is called (a, b) *matrix self-bounding* if, for any $Z, Z' \in \mathcal{Z}$,

1. $\boldsymbol{H}(Z) - \boldsymbol{H}(z_1, \dots, z_i', \dots, z_n) \preccurlyeq \mathbf{I}$, and
2. $\sum_{i=1}^n (\boldsymbol{H}(Z) - \boldsymbol{H}(z_1, \dots, z_i', \dots, z_n))_+ \preccurlyeq a\boldsymbol{H}(Z) + b\mathbf{I}$.

$\boldsymbol{H}$ is *weakly* (a, b) *matrix self-bounding* if, for any $Z, Z' \in \mathcal{Z}$,

$$\sum\nolimits_{i=1}^n (\boldsymbol{H}(Z) - \boldsymbol{H}(z_1, \dots, z_i', \dots, z_n))_+^2 \preccurlyeq a\boldsymbol{H}(Z) + b\mathbf{I}.$$

Mackey [12, Thm. 25] proposed a slightly different definition that includes an additional self-reciprocity condition. His analysis requires this extra hypothesis because it is based on matrix Stein pairs.

The approach in this paper is not quite strong enough to develop concentration inequalities for self-bounding matrix functions. Our techniques would work if the following mean value trace inequality were valid.

Conjecture 7.3 (Signed Mean Value Trace Inequalities)**.** *For all matrices* $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{H}^d$*, all positive integers* q*, and any* $s > 0$ *it holds that*

$$\operatorname{tr}\left[\boldsymbol{C}(\mathrm{e}^{\boldsymbol{A}} - \mathrm{e}^{\boldsymbol{B}})\right] \le \frac{1}{2}\operatorname{tr}\left[(s(\boldsymbol{A}-\boldsymbol{B})_+^2 + s^{-1}\boldsymbol{C}_+^2)\mathrm{e}^{\boldsymbol{A}} + (s(\boldsymbol{A}-\boldsymbol{B})_-^2 + s^{-1}\boldsymbol{C}_-^2)\mathrm{e}^{\boldsymbol{B}})\right].$$

and

$$\operatorname{tr}\left[\boldsymbol{C}(\boldsymbol{A}^q - \boldsymbol{B}^q)\right] \le \frac{q}{2}\left[(s(\boldsymbol{A}-\boldsymbol{B})_+^2 + s^{-1}\boldsymbol{C}_+^2)\,|\boldsymbol{A}|^{q-1} + (s(\boldsymbol{A}-\boldsymbol{B})_-^2 + s^{-1}\boldsymbol{C}_-^2)\,|\boldsymbol{B}|^{q-1})\right].$$

This statement involves the standard matrix functions that lift the scalar functions $x_+ := \max\{x, 0\}$ and $x_- := \max\{-a, 0\}$. Extensive simulations with random matrices suggest that Conjecture 7.3 holds, but we did not find a proof.

Acknowledgements

Paulin thanks his thesis advisors, Louis Chen and Adrian Röllin, for their helpful comments on this manuscript. Tropp was supported by ONR awards N00014-08-1-0883 and N00014-11-1002, AFOSR award FA9550-09-1-0643, and a Sloan Research Fellowship.

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

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

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

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

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 $\boldsymbol{M}$ and $\boldsymbol{N}$ be matrices in $\mathbb{M}^d$. Then*

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

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 $s > 0$,

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

Likewise,

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

Add the latter two inequalities and rearrange the terms to obtain

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

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

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

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

Appendix B: Proof of the Exponential Tail Bound

This appendix contains a proof of the exponential tail bound Theorem 3.1. The argument parallels the approach developed in [13], but we require more powerful estimates along the way. In view of the similarities, we emphasize the places where the proofs differ, and we suppress details that are identical with the earlier work.

B.1. The Matrix Laplace Transform Method

A central tool in our investigation is a matrix variant of the classical moment generating function. Ahlswede & Winter [1, App.] introduced this definition in their investigation of matrix concentration.

Definition B.1 (Trace Mgf)**.** Let $\boldsymbol{X}$ be a random Hermitian matrix. The *(normalized) trace moment generating function* of $\boldsymbol{X}$ is defined as

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

The following proposition from [13, Prop. 3.3] collects results from [1, 18, 23, 6].

Proposition B.2 (Matrix Laplace Transform Method)**.** *Let $\boldsymbol{X} \in \mathbb{H}^d$ be a random matrix with normalized trace mgf $m(\theta) := \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta \boldsymbol{X}}$. For each $t \in \mathbb{R}$,*

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{X}) \ge t\} \le d \cdot \inf_{\theta > 0} \exp\{-\theta t + \log m(\theta)\}. \tag{B.1}$$

$$\mathbb{P}\{\lambda_{\min}(\boldsymbol{X}) \le t\} \le d \cdot \inf_{\theta < 0} \exp\{-\theta t + \log m(\theta)\}. \tag{B.2}$$

Furthermore,

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{X}) \le \inf_{\theta > 0} \frac{1}{\theta} [\log d + \log m(\theta)]. \tag{B.3}$$

$$\mathbb{E}\,\lambda_{\min}(\boldsymbol{X}) \ge \sup_{\theta < 0} \frac{1}{\theta} [\log d + \log m(\theta)]. \tag{B.4}$$

In summary, we can bound the extreme eigenvalues of a random matrix by controlling the trace mgf.

B.2. The Method of Exchangeable Pairs

The main technical challenge in developing concentration inequalities is to obtain bounds for the trace mgf. In this work, we follow the approach from the paper [13], which extends Chatterjee's concentration argument [5] to the matrix setting. The key idea is to use an exchangeable pair to bound the derivative of the trace mgf, which in turns allows us to control the growth of the trace mgf.

We begin with a technical lemma, which generalizes [13, Lem. 2.3] and [4, Lem. 3.1]. This result permits us to rewrite certain matrix expectations using kernel Stein pairs.

Lemma B.3 (Method of Exchangeable Pairs)**.** *Suppose that* $(\boldsymbol{X}, \boldsymbol{X}') \in \mathbb{H}^d \times \mathbb{H}^d$ *is a* $\boldsymbol{K}$*-Stein pair constructed from an auxiliary exchangeable pair* (Z, Z')*. Let* $\boldsymbol{F} : \mathbb{H}^d \to \mathbb{H}^d$ *be a measurable function that satisfies the regularity condition*

$$\mathbb{E}\,\|\boldsymbol{K}(Z, Z') \cdot \boldsymbol{F}(\boldsymbol{X})\| < \infty. \tag{B.5}$$

Then

$$\mathbb{E}\,[\boldsymbol{X} \cdot \boldsymbol{F}(\boldsymbol{X})] = \frac{1}{2}\,\mathbb{E}\,[\boldsymbol{K}(Z, Z')(\boldsymbol{F}(\boldsymbol{X}) - \boldsymbol{F}(\boldsymbol{X}'))]\,. \tag{B.6}$$

Proof. Definition 2.2, of a kernel Stein pair, implies that

$$\mathbb{E}[\boldsymbol{X} \cdot \boldsymbol{F}(\boldsymbol{X})] = \mathbb{E}\left[\mathbb{E}[\boldsymbol{K}(Z, Z') \,|\, Z] \cdot \boldsymbol{F}(\boldsymbol{X})\right] = \mathbb{E}[\boldsymbol{K}(Z, Z')\,\boldsymbol{F}(\boldsymbol{X})],$$

where we justify the pull-through property of conditional expectation using the regularity condition (B.5). Since the kernel $\boldsymbol{K}$ satisfies the antisymmetry property (2.1), we also have the relation

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

Average the two preceding displays to reach the identity (B.6). □

Under suitable regularity conditions, the derivative of the trace mgf of a random matrix $\boldsymbol{X}$ has precisely the form needed to invoke to the method of exchangeable pairs:

$$m'(\theta) = \mathbb{E}\,\bar{\mathrm{tr}}\left[\boldsymbol{X}\mathrm{e}^{\theta\boldsymbol{X}}\right].$$

Hence, we may apply Lemma B.3 with $\boldsymbol{F}(\boldsymbol{X}) = \mathrm{e}^{\theta\boldsymbol{X}}$ to obtain the expression

$$m'(\theta) = \frac{1}{2}\,\mathbb{E}\,\bar{\mathrm{tr}}\left[\boldsymbol{K}(Z, Z')\left(\mathrm{e}^{\theta\boldsymbol{X}} - \mathrm{e}^{\theta\boldsymbol{X}'}\right)\right]. \tag{B.7}$$

The primary novelty in this work is a method for bounding the right-hand side of (B.7).

B.3. The Exponential Mean Value Trace Inequality

To control the expression (B.7) for the derivative of the trace mgf, we will invoke Lemma 3.2, the exponential mean value trace inequality. We establish this key lemma in this section. See the manuscript [20] for an alternative proof.

Proof of Lemma 3.2. To begin, we develop an alternative expression for the trace quantity that we need to bound. Observe that

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

The Fundamental Theorem of Calculus delivers the identity

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

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

$$\operatorname{tr}\left[\boldsymbol{C}\big(\mathrm{e}^{\boldsymbol{A}} - \mathrm{e}^{\boldsymbol{B}}\big)\right] = \int_0^1 \left\langle \boldsymbol{C},\ \mathrm{e}^{\tau \boldsymbol{A}} (\boldsymbol{A} - \boldsymbol{B}) \mathrm{e}^{(1-\tau)\boldsymbol{B}} \right\rangle \mathrm{d}\tau. \tag{B.8}$$

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 $\mathbb{M}^d$:

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

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

$$\int_0^1 \mathcal{A}_\tau \mathcal{B}_{1-\tau} d\tau \preccurlyeq \frac{1}{2}(|\mathcal{A}_1| + |\mathcal{B}_1|) = \frac{1}{2}(\mathcal{A}_1 + \mathcal{B}_1). \tag{B.9}$$

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

To relate this result to the problem at hand, we rewrite the expression (B.8) using the operators $\mathcal{A}_\tau$ and $\mathcal{B}_{1-\tau}$. Indeed,

$$\begin{aligned} \operatorname{tr}\left[\boldsymbol{C}\big(\mathrm{e}^{\boldsymbol{A}} - \mathrm{e}^{\boldsymbol{B}}\big)\right] &= \int_0^1 \langle \boldsymbol{C},\ (\mathcal{A}_\tau \mathcal{B}_{1-\tau})(\boldsymbol{A} - \boldsymbol{B}) \rangle \,\mathrm{d}\tau \\ &\le \left[\int_0^1 \langle \boldsymbol{C},\ (\mathcal{A}_\tau \mathcal{B}_{1-\tau})(\boldsymbol{C}) \rangle \,\mathrm{d}\tau \cdot \int_0^1 \langle \boldsymbol{A} - \boldsymbol{B},\ (\mathcal{A}_\tau \mathcal{B}_{1-\tau})(\boldsymbol{A} - \boldsymbol{B}) \rangle \,\mathrm{d}\tau\right]^{1/2}. \end{aligned} \tag{B.10}$$

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 (B.10). These estimates are an immediate consequence of (B.9). First,

$$\begin{aligned} \int_0^1 \langle \boldsymbol{C},\ (\mathcal{A}_\tau \mathcal{B}_{1-\tau})(\boldsymbol{C}) \rangle \,\mathrm{d}\tau &\le \frac{1}{2} \langle \boldsymbol{C},\ (\mathcal{A}_1 + \mathcal{B}_1)(\boldsymbol{C}) \rangle \\ &= \frac{1}{2} \left\langle \boldsymbol{C},\ \mathrm{e}^{\boldsymbol{A}} \boldsymbol{C} + \boldsymbol{C} \mathrm{e}^{\boldsymbol{B}} \right\rangle = \frac{1}{2} \operatorname{tr}\left[\boldsymbol{C}^2 \big(\mathrm{e}^{\boldsymbol{A}} + \mathrm{e}^{\boldsymbol{B}}\big)\right]. \end{aligned} \tag{B.11}$$

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,

$$\int_0^1 \langle \boldsymbol{A} - \boldsymbol{B},\ (\mathcal{A}_\tau \mathcal{B}_{1-\tau})(\boldsymbol{A} - \boldsymbol{B}) \rangle \,\mathrm{d}\tau = \frac{1}{2} \operatorname{tr}\left[(\boldsymbol{A} - \boldsymbol{B})^2 \left(\mathrm{e}^{\boldsymbol{A}} + \mathrm{e}^{\boldsymbol{B}}\right)\right]. \tag{B.12}$$

Substitute (B.11) and (B.12) into the inequality (B.10) to reach

$$\operatorname{tr}\left[\boldsymbol{C}\left(\mathrm{e}^{\boldsymbol{A}} - \mathrm{e}^{\boldsymbol{B}}\right)\right] \le \frac{1}{2}\left(\operatorname{tr}\left[\boldsymbol{C}^2\left(\mathrm{e}^{\boldsymbol{A}} + \mathrm{e}^{\boldsymbol{B}}\right)\right] \cdot \operatorname{tr}\left[(\boldsymbol{A} - \boldsymbol{B})^2\left(\mathrm{e}^{\boldsymbol{A}} + \mathrm{e}^{\boldsymbol{B}}\right)\right]\right)^{1/2}.$$

We obtain the result stated in Lemma 3.2 by applying the numerical inequality between the geometric mean and the arithmetic mean. □

B.4. Bounding the Derivative of the Trace Mgf

We are now prepared to obtain a bound for the derivative of the trace mgf in terms of the conditional variance and the kernel conditional variance.

Lemma B.4 (The Derivative of the Trace Mgf)**.** *Suppose that* $(\boldsymbol{X}, \boldsymbol{X}')$ *is a* $\boldsymbol{K}$*-Stein pair, and assume that* $\boldsymbol{X}$ *is almost surely bounded in norm. Define the normalized trace mgf* $m(\theta) := \mathbb{E}\,\bar{\operatorname{tr}}\,\mathrm{e}^{\theta \boldsymbol{X}}$*. Then*

$$|m'(\theta)| \le \frac{1}{2}\,|\theta| \cdot \inf_{s>0}\ \mathbb{E}\,\bar{\operatorname{tr}}\left[\left(s\boldsymbol{V}_{\boldsymbol{X}} + s^{-1}\boldsymbol{V}^{\boldsymbol{K}}\right)\mathrm{e}^{\theta \boldsymbol{X}}\right] \quad \text{for all } \theta \in \mathbb{R}. \tag{B.13}$$

The conditional variances $\boldsymbol{V}_{\boldsymbol{X}}$ *and* $\boldsymbol{V}^{\boldsymbol{K}}$ *are defined in* (2.6) *and* (2.7)*.*

Proof. Consider the derivative of the trace mgf

$$m'(\theta) = \mathbb{E}\,\bar{\operatorname{tr}}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\,\mathrm{e}^{\theta \boldsymbol{X}}\right] = \mathbb{E}\,\bar{\operatorname{tr}}\left[\boldsymbol{X}\mathrm{e}^{\theta \boldsymbol{X}}\right], \tag{B.14}$$

where the dominated convergence theorem and the boundedness of $\boldsymbol{X}$ justify the exchange of expectation and derivative. When $\theta = 0$, we have $m'(\theta) = 0$, as advertised. When $\theta \neq 0$, the form of this derivative is ripe for an application of the method of exchangeable pairs, Lemma B.3. Since $\boldsymbol{X}$ is bounded, the regularity condition (B.5) is satisfied, and we obtain

$$m'(\theta) = \frac{1}{2}\,\mathbb{E}\,\bar{\operatorname{tr}}\left[\boldsymbol{K}(Z, Z')\left(\mathrm{e}^{\theta \boldsymbol{X}} - \mathrm{e}^{\theta \boldsymbol{X}'}\right)\right]. \tag{B.15}$$

The exponential mean value trace inequality, Lemma 3.2, implies that

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

The first equality follows from the exchangeability of $(\boldsymbol{X}, \boldsymbol{X}')$; 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 (B.13) by introducing the definitions (2.6) and (2.7) of the conditional variance and the kernel conditional variance. □

B.5. Bounding the Trace Mgf

Lemma B.4 gives us a powerful tool for bounding the trace mgf of a random matrix $\boldsymbol{X}$ that is presented as part of a kernel Stein pair. The following lemma shows how to derive a trace mgf bound from bounds on the kernel conditional variance.

Lemma B.5 (Trace Mgf Estimates for Bounded Random Matrices)**.** *Let $(\boldsymbol{X}, \boldsymbol{X}')$ be a $\boldsymbol{K}$-Stein pair, and suppose there exist nonnegative constants c, v, s for which*

$$\boldsymbol{V}_{\boldsymbol{X}} \preccurlyeq s^{-1}(c\boldsymbol{X} + v\,\mathbf{I}) \quad \text{and} \quad \boldsymbol{V}^{\boldsymbol{K}} \preccurlyeq s\,(c\boldsymbol{X} + v\,\mathbf{I}) \quad \text{almost surely.} \tag{B.16}$$

Then the normalized trace mgf $m(\theta) := \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta \boldsymbol{X}}$ satisfies the bounds

$$\log m(\theta) \le \frac{v\theta^2}{2} \qquad \text{when } \theta \le 0. \tag{B.17}$$

$$\log m(\theta) \le \frac{v}{c^2}\left[\log\left(\frac{1}{1-c\theta}\right) - c\theta\right] \tag{B.18}$$

$$\le \frac{v\theta^2}{2(1-c\theta)} \qquad \text{when } 0 \le \theta < 1/c. \tag{B.19}$$

The two conditional variances are defined in (2.6) and (2.7).

Proof. As demonstrated in [13, Lem. 4.3], the assumption (B.16) implies that $\boldsymbol{X}$ is almost surely bounded in norm. Hence, we may apply Lemma B.4 along with our conditional variance bounds (B.16) to obtain

$$\begin{aligned}
|m'(\theta)| &\le \frac{1}{2}\,|\theta| \cdot \inf_{t>0}\ \mathbb{E}\,\bar{\mathrm{tr}}\left[(t\,\boldsymbol{V}_{\boldsymbol{X}} + t^{-1}\boldsymbol{V}^{\boldsymbol{K}})\,\mathrm{e}^{\theta \boldsymbol{X}}\right] \\
&\le \frac{1}{2}\,|\theta| \cdot \mathbb{E}\,\bar{\mathrm{tr}}\left[(s\,\boldsymbol{V}_{\boldsymbol{X}} + s^{-1}\boldsymbol{V}^{\boldsymbol{K}})\,\mathrm{e}^{\theta \boldsymbol{X}}\right] \\
&\le |\theta| \cdot \mathbb{E}\,\bar{\mathrm{tr}}\left[(c\boldsymbol{X} + v\,\mathbf{I})\,\mathrm{e}^{\theta \boldsymbol{X}}\right] \\
&= c\,|\theta| \cdot \mathbb{E}\,\bar{\mathrm{tr}}\left[\boldsymbol{X}\mathrm{e}^{\theta \boldsymbol{X}}\right] + v\,|\theta| \cdot \mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta \boldsymbol{X}} \\
&= c\,|\theta| \cdot m'(\theta) + v\,|\theta| \cdot m(\theta),
\end{aligned}$$

where the third inequality follows from the positivity of $\mathrm{e}^{\theta \boldsymbol{X}}$. The remainder of the argument now proceeds as in [13, Lem. 4.3]. □

B.6. Proof of Theorem 3.1

The remainder of the proof of Theorem 3.1 is identical to that of [13, Thm. 4.1], once we substitute the trace mgf estimates from Lemma B.5 in place of the result [13, Lem. 4.3]. We omit the details.

Appendix C: Proof of the Polynomial Moment Inequality

Next, we develop a proof of the matrix Burkholder–Davis–Gundy inequality, Theorem 3.3. The proof parallels the argument in [13], but we need some new matrix inequalities to make the extension to kernel Stein pairs.

C.1. The Polynomial Mean Value Trace Inequality

The critical new ingredient in Theorem 3.3 is the polynomial mean value trace inequality, Lemma 3.4. Let us proceed with a proof of this result.

Proof of Lemma 3.4. First, we need to develop another representation for the trace quantity that we are analyzing. Assume that $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{H}^d$. A direct calculation shows that

$$\boldsymbol{A}^q - \boldsymbol{B}^q = \sum\nolimits_{k=0}^{q-1} \boldsymbol{A}^k(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{B}^{q-1-k}.$$

As a consequence,

$$\operatorname{tr}\left[\boldsymbol{C}(\boldsymbol{A}^q - \boldsymbol{B}^q)\right] = \sum\nolimits_{k=0}^{q-1} \left\langle \boldsymbol{C},\ \boldsymbol{A}^k(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{B}^{q-1-k} \right\rangle. \tag{C.1}$$

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

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

$$\mathcal{A}_k(\boldsymbol{M}) := \boldsymbol{A}^k\boldsymbol{M} \quad \text{and} \quad \mathcal{B}_k(\boldsymbol{M}) := \boldsymbol{M}\boldsymbol{B}^k \quad \text{for each } k = 0, 1, 2, \dots, q-1.$$

The absolute values of these operators satisfy

$$|\mathcal{A}_k|\,(\boldsymbol{M}) = |\boldsymbol{A}|^k\,\boldsymbol{M} \quad \text{and} \quad |\mathcal{B}_k|\,(\boldsymbol{M}) = \boldsymbol{M}\,|\boldsymbol{B}|^k \quad \text{for each } k = 0, 1, 2, \dots, q-1.$$

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

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

Summing over k, we discover that

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

This is the mean inequality that we require.

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

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

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 (C.4). The mean inequality (C.2) ensures that

$$\sum_{k=0}^{q-1} \langle \boldsymbol{C},\ |\mathcal{A}_k\mathcal{B}_{q-k-1}|\,(\boldsymbol{C})\rangle \le \frac{q}{2}\,\big\langle \boldsymbol{C},\ \big(\,|\mathcal{A}_1|^{q-1} + |\mathcal{B}_1|^{q-1}\,\big)(\boldsymbol{C})\big\rangle$$
$$= \frac{q}{2}\,\big\langle \boldsymbol{C},\ |\boldsymbol{A}|^{q-1}\,\boldsymbol{C} + \boldsymbol{C}\,|\boldsymbol{B}|^{q-1}\big\rangle = \frac{q}{2}\operatorname{tr}\big[\boldsymbol{C}^2\big(\,|\boldsymbol{A}|^{q-1} + |\boldsymbol{B}|^{q-1}\,\big)\big]. \quad \text{(C.5)}$$

Likewise,

$$\sum_{k=0}^{q-1} \langle \boldsymbol{A}-\boldsymbol{B},\ |\mathcal{A}_k\mathcal{B}_{q-k-1}|\,(\boldsymbol{A}-\boldsymbol{B})\rangle \le \frac{q}{2}\operatorname{tr}\big[(\boldsymbol{A}-\boldsymbol{B})^2\big(\,|\boldsymbol{A}|^{q-1} + |\boldsymbol{B}|^{q-1}\,\big)\big]. \quad \text{(C.6)}$$

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

$$\operatorname{tr}\big[\boldsymbol{C}(\boldsymbol{A}^q - \boldsymbol{B}^q)\big] \le \frac{q}{2}\left(\operatorname{tr}\big[\boldsymbol{C}^2\big(\,|\boldsymbol{A}|^{q-1} + |\boldsymbol{B}|^{q-1}\,\big)\big]\cdot \operatorname{tr}\big[(\boldsymbol{A}-\boldsymbol{B})^2\big(\,|\boldsymbol{A}|^{q-1} + |\boldsymbol{B}|^{q-1}\,\big)\big]\right)^{1/2}.$$

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

C.2. Proof of Theorem 3.3

Abbreviate

$$E := \mathbb{E}\,\|\boldsymbol{X}\|_{S_{2p}}^{2p} = \mathbb{E}\operatorname{tr}|\boldsymbol{X}|^{2p} = \mathbb{E}\operatorname{tr}\big[\boldsymbol{X}\cdot\boldsymbol{X}^{2p-1}\big].$$

To apply the method of exchangeable pairs, Lemma B.3, we check the regularity condition (B.5):

$$\mathbb{E}\,\big\|\boldsymbol{K}(Z,Z')\cdot\boldsymbol{X}^{2p-1}\big\| \le \mathbb{E}\,\big(\,\|\boldsymbol{K}(Z,Z')\|\,\|\boldsymbol{X}\|^{2p-1}\,\big)$$
$$\le \big(\,\mathbb{E}\,\|\boldsymbol{K}(Z,Z')\|^{2p}\,\big)^{1/2p}\big(\,\mathbb{E}\,\|\boldsymbol{X}\|^{2p}\,\big)^{(2p-1)/2p} < \infty,$$

where we have applied Hölder's inequality for expectation and the fact that the spectral norm is dominated by the Schatten $(2p)$-norm. Invoke Lemma B.3 with $\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{X}^{2p-1}$ to reach

$$E = \frac{1}{2}\,\mathbb{E}\operatorname{tr}\big[\boldsymbol{K}(Z,Z')\cdot\big(\boldsymbol{X}^{2p-1} - \boldsymbol{X}'^{2p-1}\big)\big].$$

Next, fix a parameter $s > 0$. Apply the polynomial mean value trace inequality, Lemma 3.4, with $q = 2p-1$ to obtain the estimate

$$E \le \frac{2p-1}{8}\,\mathbb{E}\operatorname{tr}[(s\,(\boldsymbol{X}-\boldsymbol{X}')^2 + s^{-1}\boldsymbol{K}(Z,Z')^2)\cdot(\boldsymbol{X}^{2p-2} + \boldsymbol{X}'^{2p-2})]$$
$$= \frac{2p-1}{4}\,\mathbb{E}\operatorname{tr}[(s\,(\boldsymbol{X}-\boldsymbol{X}')^2 + s^{-1}\boldsymbol{K}(Z,Z')^2)\cdot\boldsymbol{X}^{2p-2}]$$
$$= (2p-1)\,\mathbb{E}\operatorname{tr}\left[\frac{1}{2}(s\,\boldsymbol{V}_{\boldsymbol{X}} + s^{-1}\boldsymbol{V}^{\boldsymbol{K}})\cdot\boldsymbol{X}^{2p-2}\right],$$

where we have used the exchangeability of $(\boldsymbol{X},\boldsymbol{X}')$ and the definitions (2.6) and (2.7) of the conditional variances. In the last step, we justify the pull-through property with the regularity condition $\mathbb{E}\,\|\boldsymbol{X}\|_{S_{2p}}^{2p} < \infty$. The remainder of the argument is identical with the proof of [13, Thm. 8.1].

Appendix D: Haar Measures and Controlled Total Variation

In this section, we prove Corollary 7.1 by studying the behavior of Hermitian functions of group-valued random elements. Under the notation of Corollary 7.1, we define $\boldsymbol{X} := \boldsymbol{\Psi}(Z)$. [4, Thm. 4.6] showed that scalar functions of the Haar measure are well concentrated whenever particular random walks on G converge rapidly to the Haar distribution. In the sections to follow, we develop a Hermitian analogue of this relationship using the tools of Sections 2 and 3. As in [4], we will adopt the total variation distance between measures (5.1) as our convergence metric.

D.1. A Kernel Coupling

We begin by establishing a kernel coupling suitable for analyzing $\boldsymbol{X}$. Since $Y \in G$ is independent of Z and satisfies (7.1), $Z' = YZ$ is exchangeable counterpart for Z, and hence $(\boldsymbol{X}, \boldsymbol{X}') = (\boldsymbol{\Psi}(Z), \boldsymbol{\Psi}(Z'))$ is an exchangeable pair.

Moreover, the sequence of pairs

$$(Z_{(i)}, Z'_{(i)}) := (Y_i \cdots Y_1 Z, Y_i \cdots Y_1 Z') \quad \text{for each} \quad i \geq 0 \tag{D.1}$$

defines a kernel coupling for $(\boldsymbol{X}, \boldsymbol{X}')$. Thus, $(\boldsymbol{X}, \boldsymbol{X}')$ is a kernel Stein pair with $\boldsymbol{K}$ defined as in Lemma 2.4 whenever the precondition (2.4) is met.

D.2. The Conditional Variances

The sequence of multipliers $(Y_i \cdots Y_1)_{i=1}^{\infty}$ in our kernel coupling (D.1) can be viewed as a random walk on the group G, and, for many choices of Y, this sequence will converge to a Haar distributed random variable. Intuitively, a faster rate of convergence implies a faster coupling time for the Markov chains $(Z_{(i)})_{i\geq 0}$ and $(Z'_{(i)})_{i\geq 0}$ and hence a smaller $\boldsymbol{K}$-conditional variance (2.7). Our next lemma makes this intuition more precise by bounding the $\boldsymbol{K}$-conditional variance in terms of the total variation distance between $Y_i \cdots Y_1$ and Z.

Lemma D.1. *Let $Z \sim \mu$ be Haar distributed on a group G. Let $(\boldsymbol{X}, \boldsymbol{X}') := (\boldsymbol{\Psi}(Z), \boldsymbol{\Psi}(Z'))$ with $\boldsymbol{K}$ constructed as in Section D.1. Suppose that μ_i is the distribution of $Y_i \cdots Y_1$ and that*

$$S^2 := \sup_{g \in G} \left\| \mathbb{E}[(\boldsymbol{\Psi}(g) - \boldsymbol{\Psi}(Yg))^2] \right\|.$$

Then $(\boldsymbol{X}, \boldsymbol{X}')$ is a $\boldsymbol{K}$-Stein pair whenever $\sum_{i=0}^{\infty} d_{TV}(\mu_i, \mu) < \infty$. Moreover, the conditional variance (2.6) satisfies

$$\lambda_{\max}(\boldsymbol{V}_{\boldsymbol{X}}) \leq \frac{S^2}{2} \quad \text{almost surely,}$$

and the $\boldsymbol{K}$-conditional variance (2.7) satisfies

$$\lambda_{\max}(\boldsymbol{V}^{\boldsymbol{K}}) \leq \frac{S^2}{2} \left(\sum_{i=0}^{\infty} \min\left\{1,\ 4RS^{-1}\, d_{TV}(\mu_i, \mu)\right\} \right)^2 \quad \text{almost surely.}$$

Proof. Fix any $i \geq 0$. We aim to bound

$$\begin{aligned}\mathbb{E}[\boldsymbol{\Psi}(Z_{(i)}) - \boldsymbol{\Psi}(Z'_{(i)}) \,|\, Z = z, Z' = z']^2 &= (\mathbb{E}[\boldsymbol{\Psi}(Y_i \cdots Y_1 z)] - \mathbb{E}[\boldsymbol{\Psi}(Y_i \cdots Y_1 z')])^2 \\ &\preccurlyeq 2\,\mathbb{E}[\boldsymbol{\Psi}(Y_i \cdots Y_1 z)]^2 + 2\,\mathbb{E}[\boldsymbol{\Psi}(Y_i \cdots Y_1 z')]^2,\end{aligned}$$

where the inequality follows from the convexity of the matrix square. For any $z \in G$,

$$\begin{aligned}\mathbb{E}[\boldsymbol{\Psi}(Y_i \cdots Y_1 z)] &= \mathbb{E}[\boldsymbol{\Psi}(Y_i \cdots Y_1 z)] - \mathbb{E}[\boldsymbol{\Psi}(Z)] \\ &= \mathbb{E}[\boldsymbol{\Psi}(Y_i \cdots Y_1 z)] - \mathbb{E}[\boldsymbol{\Psi}(Zz)],\end{aligned}$$

since Z is Haar distributed, and hence $Zz =_d Z$. Furthermore, for any positive measure ν that dominates μ and μ_i,

$$\begin{aligned}\|\mathbb{E}[\boldsymbol{\Psi}(Y_i \cdots Y_1 z)]\| &= \left\| \int \boldsymbol{\Psi}(yz) \left(\frac{d\mu_i}{d\nu}(y) - \frac{d\mu}{d\nu}(y) \right) d\nu(y) \right\| \\ &\leq R \int \left| \frac{d\mu_i}{d\nu}(y) - \frac{d\mu}{d\nu}(y) \right| d\nu(y) \\ &\leq 2R\; d_{\mathrm{TV}}(\mu_i, \mu),\end{aligned}$$

by our bound on $\boldsymbol{\Psi}$ and the definition of total variation. Therefore,

$$\mathbb{E}[\mathbb{E}[\boldsymbol{\Psi}(Z_{(i)}) - \boldsymbol{\Psi}(Z'_{(i)}) \,|\, Z, Z']^2 \,|\, Z] \preccurlyeq 16R^2 d_{\mathrm{TV}}^2(\mu_i, \mu)\; \mathbf{I}.$$

We note moreover that

$$\left\| \sum\nolimits_{i=0}^{\infty} |\, \mathbb{E}[\boldsymbol{\Psi}(Z_{(i)}) - \boldsymbol{\Psi}(Z'_{(i)}) \,|\, Z_{(0)} = z, Z'_{(0)} = z']| \right\| \leq \sum\nolimits_{i=0}^{\infty} 4R\; d_{\mathrm{TV}}(\mu_i, \mu)$$

for all z and z'. Hence, by Lemma 2.4, $(\boldsymbol{X}, \boldsymbol{X}')$ is a valid $\boldsymbol{K}$-Stein pair whenever the total variation distances are summable.

Next, let $W_i := Y_i \cdots Y_1$, and notice that

$$\begin{aligned}\left\|\mathbb{E}[\mathbb{E}[\boldsymbol{\Psi}(Z_{(i)}) - \boldsymbol{\Psi}(Z'_{(i)}) \,|\, Z, Z']^2 \,|\, Z = z]\right\| &\leq \left\|\mathbb{E}[(\boldsymbol{\Psi}(Z_{(i)}) - \boldsymbol{\Psi}(Z'_{(i)}))^2 \,|\, Z = z]\right\| \\ &= \left\|\mathbb{E}[(\boldsymbol{\Psi}(W_i z) - \boldsymbol{\Psi}(W_i Y z))^2]\right\| \\ &\leq \sup_{g \in G} \left\|\mathbb{E}[(\boldsymbol{\Psi}(gz) - \boldsymbol{\Psi}(gYz))^2]\right\| \\ &= \sup_{g \in G} \left\|\mathbb{E}[(\boldsymbol{\Psi}(gz) - \boldsymbol{\Psi}(Ygz))^2]\right\| \leq S^2,\end{aligned}$$

where the first inequality is a consequence of the convexity of the matrix square, and the final equality follows from the property $g^{-1} Y g =_d Y$ for all $g \in G$. Hence, we may apply Lemma 2.6 with $s_0 = S$, with $s_i = \min\left\{S,\; 4R\, d_{\mathrm{TV}}(\mu_i, \mu)\right\}$ for $i > 0$, and with $\boldsymbol{\Gamma}(Z) = \mathbf{I}$ to obtain the result. □

D.3. Exponential Concentration

We are finally equipped to prove Corollary 7.1. Under the kernel coupling construction of Section D.1, the conditional variance bounds of Lemma D.1 imply that Theorem 3.1 holds with $c = 0$, with $v = \frac{1}{2} S^2 \sum_{i=0}^{\infty} \min\left\{1,\; 4RS^{-1}\, d_{\mathrm{TV}}(\mu_i, \mu)\right\}$, and with $s = \sum_{i=0}^{\infty} \min\left\{1,\; 4RS^{-1}\, d_{\mathrm{TV}}(\mu_i, \mu)\right\}$. This establishes the result.

References

[1] R. Ahlswede and A. Winter, *Strong converse for identification via quantum channels*, IEEE Trans. Inform. Theory **48** (2002), no. 3, 569–579.

[2] A. Ahmed, B. Recht, and J Romberg, *Blind deconvolution using convex programming*, arXiv e-prints (2012), Available at http://arxiv.org/abs/1211.5608.

[3] S. Boucheron, G. Lugosi, and P. Massart, *On concentration of self-bounding functions*, Electron. J. Probab. **14** (2009), no. 64, 1884–1899.

[4] S. Chatterjee, *Concentration inequalities with exchangeable pairs*, Ph.D. thesis, Stanford University, Palo Alto, Feb. 2005, Available at http://arxiv.org/abs/math/0507526.

[5] ———, *Stein's method for concentration inequalities*, Probab. Theory Related Fields **138** (2007), 305–321.

[6] R. Y. Chen, A. Gittens, and J. A. Tropp, *The masked sample covariance estimator: An analysis via the matrix Laplace transform method*, Information and Inference (2012), To appear.

[7] R. L. Dobrushin, *Prescribing a system of random variables by conditional distributions*, Theory of Probability & Its Applications **15** (1970), no. 3, 458–486.

[8] V. Koltchinskii, *Von neumann entropy penalization and low-rank matrix estimation*, The Annals of Statistics **39** (2012), no. 6, 2936–2973.

[9] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*, American Mathematical Society, Providence, RI, 2009, With a chapter by James G. Propp and David B. Wilson.

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

[11] P. Machart and L. Ralaivola, *Confusion Matrix Stability Bounds for Multiclass Classification*, arXiv e-prints (2012), Available at http://arxiv.org/abs/1202.6221.

[12] L. Mackey, *Matrix factorization and matrix concentration*, Ph.D. thesis, EECS Department, University of California, Berkeley, May 2012, Available at http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-99.html.

[13] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp, *Matrix concentration inequalities via the method of exchangeable pairs*, Available at http://arxiv.org/abs/1201.6002, 2012.

[14] C. McDiarmid, *On the method of bounded differences*, Surveys in Combinatorics 1989 (London), London Mathematical Society Lecture Notes, 1989, pp. 148–188.

[15] E. Morvant, S. Koço, and L. Ralaivola, *PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification*, arXiv e-prints (2012), Available at http://arxiv.org/abs/1202.6228.

[16] R. I. Oliveira, *Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges*, Available at http://arxiv.org/abs/0911.0600, Nov. 2009.

[17] ———, *Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges*, Available at http://arxiv.org/abs/0911.0600, Nov. 2009.

[18] ———, *Sums of random Hermitian matrices and an inequality by Rudelson*, Electron. Commun. Probab. **15** (2010), 203–212.

[19] L. Onsager, *Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition*, Physical Review **65** (1944), 117–149.

[20] D. Paulin, *A Note on Matrix Concentration Inequalities via the Method of Exchangeable*

Pairs, arXiv e-prints (2012), Available at http://arxiv.org/abs/1212.2012.

[21] C. Stein, *Approximate computation of expectations*, Lecture Notes-Monograph Series **7** (1986), pp. i–iii+1–7+9–51+53–57+59–93+95–103+105–123+125–135+137–143+145–159+161–164.

[22] J. A. Tropp, *Improved analysis of the subsampled randomized Hadamard transform*, Adv. Adapt. Data Anal. **3** (2011), no. 1-2, 115–126.

[23] J. A. Tropp, *User-friendly tail bounds for sums of random matrices*, Found. Comput. Math. (2011).

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

[25] T. T. Wu, B. M. McCoy, C. A. Tracy, and E. Barouch, *Spin-spin correlation functions for the two-dimensional ising model: Exact theory in the scaling region*, Physical Review B **13** (1976), no. 1, 316.