THE EXPECTED NORM OF A SUM OF INDEPENDENT RANDOM MATRICES: AN ELEMENTARY APPROACH

JOEL A. TROPP

ABSTRACT. In contemporary applied and computational mathematics, a frequent challenge is to bound the expectation of the spectral norm of a sum of independent random matrices. This quantity is controlled by the norm of the expected square of the random matrix and the expectation of the maximum squared norm achieved by one of the summands; there is also a weak dependence on the dimension of the random matrix. The purpose of this paper is to give a complete, elementary proof of this important, but underappreciated, inequality.

1. MOTIVATION

Over the last decade, random matrices have become ubiquitous in applied and computational mathematics. As this trend accelerates, more and more researchers must confront random matrices as part of their work. Classical random matrix theory can be difficult to use, and it is often silent about the questions that come up in modern applications. As a consequence, it has become imperative to develop and disseminate new tools that are easy to use and that apply to a wide range of random matrices.

1.1. Matrix Concentration Inequalities.

Matrix concentration inequalities are among the most popular of these new methods. For a random matrix $\boldsymbol{Z}$ with appropriate structure, these results use simple parameters associated with the random matrix to provide bounds of the form

$$\mathbb{E}\|\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z}\| \quad \le \quad \dots \quad \text{and} \quad \mathbb{P}\{\|\boldsymbol{Z}-\mathbb{E}\boldsymbol{Z}\|\ge t\} \quad \le \quad \dots$$

where $\|\cdot\|$ denotes the spectral norm, also known as the ℓ_2 operator norm. These tools have already found a place in a huge number of mathematical research fields, including

- numerical linear algebra [Tro11]
- numerical analysis [MB14]
- uncertainty quantification [CG14]
- statistics [Kol11]
- econometrics [CC13]
- approximation theory [CDL13]
- sampling theory [BG13]
- machine learning [DKC13, LPSS$^+$14]
- learning theory [FSV12, MKR12]
- mathematical signal processing [CBSW14]
- optimization [CSW12]
- computer graphics and vision [CGH14]
- quantum information theory [Hol12]
- theory of algorithms [HO14, CKM$^+$14] and
- combinatorics [Oli10].

These references are chosen more or less at random from a long menu of possibilities. See the monograph [Tro15a] for an overview of the main results on matrix concentration, many detailed applications, and additional background references.

Date: 15 June 2015.

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

Key words and phrases. Probability inequality; random matrix; sum of independent random variables.

1.2. **The Expected Norm.** The purpose of this paper is to provide a complete proof of the following important, but underappreciated, theorem. This result is adapted from [CGT12, Thm. A.1].

Theorem I (The Expected Norm of an Independent Sum of Random Matrices)**.** *Consider an independent family $\{\boldsymbol{S}_1, \dots, \boldsymbol{S}_n\}$ of random $d_1 \times d_2$ complex-valued matrices with $\mathbb{E}\,\boldsymbol{S}_i = \mathbf{0}$ for each index i, and define*

$$\boldsymbol{Z} := \textstyle\sum_{i=1}^{n} \boldsymbol{S}_i. \tag{1.1}$$

Introduce the matrix variance parameter

$$\begin{aligned} v(\boldsymbol{Z}) &:= \max\left\{ \left\| \mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^* \right] \right\|, \left\| \mathbb{E}\left[\boldsymbol{Z}^*\boldsymbol{Z} \right] \right\| \right\} \\ &= \max\left\{ \left\| \textstyle\sum_i \mathbb{E}\left[\boldsymbol{S}_i\boldsymbol{S}_i^* \right] \right\|, \left\| \textstyle\sum_i \mathbb{E}\left[\boldsymbol{S}_i^*\boldsymbol{S}_i \right] \right\| \right\} \end{aligned} \tag{1.2}$$

and the large deviation parameter

$$L := \left(\mathbb{E} \max_i \|\boldsymbol{S}_i\|^2 \right)^{1/2}. \tag{1.3}$$

Define the dimensional constant

$$C(\boldsymbol{d}) := C(d_1, d_2) := 4 \cdot \left(1 + 2\lceil \log(d_1 + d_2) \rceil \right). \tag{1.4}$$

Then we have the matching estimates

$$\sqrt{c \cdot v(\boldsymbol{Z})} \; + \; c \cdot L \quad \leq \quad \left(\mathbb{E}\,\|\boldsymbol{Z}\|^2 \right)^{1/2} \quad \leq \quad \sqrt{C(\boldsymbol{d}) \cdot v(\boldsymbol{Z})} \; + \; C(\boldsymbol{d}) \cdot L. \tag{1.5}$$

In the lower inequality, we can take $c := 1/4$. The symbol $\|\cdot\|$ denotes the ℓ_2 operator norm, also known as the spectral norm, and * refers to the conjugate transpose operation. The map $\lceil\cdot\rceil$ returns the smallest integer that exceeds its argument.

The proof of this result occupies the bulk of this paper. Most of the page count is attributed to a detailed presentation of the required background material from linear algebra and probability. We have based the argument on the most elementary considerations possible, and we have tried to make the work self-contained. Once the reader has digested these ideas, the related—but more sophisticated —approach in the paper [MJC$^+$14] should be accessible.

1.3. **Discussion.** Before we continue, some remarks about Theorem I are in order. First, although it may seem restrictive to focus on independent sums, as in (1.1), this model captures an enormous number of useful examples. See the monograph [Tro15a] for justification.

We have chosen the term *variance parameter* because the quantity (1.2) is a direct generalization of the variance of a scalar random variable. The passage from the first formula to the second formula in (1.2) is an immediate consequence of the assumption that the summands $\boldsymbol{S}_i$ are independent and have zero mean (see Section 5). We use the term *large-deviation parameter* because the quantity (1.3) reflects the part of the expected norm of the random matrix that is attributable to one of the summands taking an unusually large value. In practice, both parameters are easy to compute using matrix arithmetic and some basic probabilistic considerations.

In applications, it is common that we need high-probability bounds on the norm of a random matrix. Typically, the bigger challenge is to estimate the expectation of the norm, which is what Theorem I achieves. Once we have a bound for the expectation, we can use scalar concentration inequalities, such as [BLM13, Thm. 6.10], to obtain high-probability bounds on the deviation between the norm and its mean value.

We have stated Theorem I as a bound on the second moment of $\|\boldsymbol{Z}\|$ because this is the most natural form of the result. Equivalent bounds hold for the first moment:

$$\sqrt{c' \cdot v(\boldsymbol{Z})} \; + \; c' \cdot L \quad \leq \quad \mathbb{E}\,\|\boldsymbol{Z}\| \quad \leq \quad \sqrt{C(\boldsymbol{d}) \cdot v(\boldsymbol{Z})} \; + \; C(\boldsymbol{d}) \cdot L.$$

We can take $c' = 1/8$. The upper bound follows easily from (1.5) and Jensen's inequality. The lower bound requires the Khintchine–Kahane inequality [LO94].

Observe that the lower and upper estimates in (1.5) differ only by the factor $C(\boldsymbol{d})$. As a consequence, the lower bound has no explicit dimensional dependence, while the upper bound has only a weak dependence on the dimension. Under the assumptions of the theorem, it is not possible to make substantial improvements to either the lower bound or the upper bound. Section 7 provides examples that support this claim.

In the theory of matrix concentration, one of the major challenges is to understand what properties of the random matrix $\boldsymbol{Z}$ allow us to remove the dimensional factor $C(\boldsymbol{d})$ from the estimate (1.5). This question is largely open, but the recent papers [Oli13, BH14, Tro15b] make some progress.

1.4. The Uncentered Case.

Although Theorem I concerns a centered random matrix, it can also be used to study a general random matrix. The following result is an immediate corollary of Theorem I.

Theorem II. *Consider an independent family $\{\boldsymbol{S}_1, \dots, \boldsymbol{S}_n\}$ of random $d_1 \times d_2$ complex-valued matrices, not necessarily centered. Define*

$$\boldsymbol{R} := \textstyle\sum_{i=1}^n \boldsymbol{S}_i$$

Introduce the variance parameter

$$\begin{aligned} v(\boldsymbol{R}) &:= \max\left\{ \left\| \mathbb{E}\left[(\boldsymbol{R} - \mathbb{E}\,\boldsymbol{R})(\boldsymbol{R} - \mathbb{E}\,\boldsymbol{R})^* \right] \right\|, \ \left\| \mathbb{E}\left[(\boldsymbol{R} - \mathbb{E}\,\boldsymbol{R})^* (\boldsymbol{R} - \mathbb{E}\,\boldsymbol{R}) \right] \right\| \right\} \\ &= \max\left\{ \left\| \textstyle\sum_{i=1}^n \mathbb{E}\left[(\boldsymbol{S}_i - \mathbb{E}\,\boldsymbol{S}_i)(\boldsymbol{S}_i - \mathbb{E}\,\boldsymbol{S}_i)^* \right] \right\|, \ \left\| \mathbb{E}\left[(\boldsymbol{S}_i - \mathbb{E}\,\boldsymbol{S}_i)^* (\boldsymbol{S}_i - \mathbb{E}\,\boldsymbol{S}_i) \right] \right\| \right\} \end{aligned}$$

and the large-deviation parameter

$$L^2 := \mathbb{E} \max_i \left\| \boldsymbol{S}_i - \mathbb{E}\,\boldsymbol{S}_i \right\|^2.$$

Then we have the matching estimates

$$\sqrt{c \cdot v(\boldsymbol{R})} \ + \ c \cdot L \quad \le \quad \left(\mathbb{E} \left\| \boldsymbol{R} - \mathbb{E}\,\boldsymbol{R} \right\|^2 \right)^{1/2} \quad \le \quad \sqrt{C(\boldsymbol{d}) \cdot v(\boldsymbol{R})} \ + \ C(\boldsymbol{d}) \cdot L.$$

We can take $c = 1/4$, and the dimensional constant $C(\boldsymbol{d})$ is defined in (1.4).

Theorem II can also be used to study $\|\boldsymbol{R}\|$ by combining it with the estimates

$$\left\| \mathbb{E}\,\boldsymbol{R} \right\| \ - \ \left(\mathbb{E} \left\| \boldsymbol{R} - \mathbb{E}\,\boldsymbol{R} \right\|^2 \right)^{1/2} \quad \le \quad \left(\mathbb{E} \left\| \boldsymbol{R} \right\|^2 \right)^{1/2} \quad \le \quad \left\| \mathbb{E}\,\boldsymbol{R} \right\| \ + \ \left(\mathbb{E} \left\| \boldsymbol{R} - \mathbb{E}\,\boldsymbol{R} \right\|^2 \right)^{1/2}.$$

These bounds follow from the triangle inequality for the spectral norm.

It is productive to interpret Theorem II as a perturbation result because it describes how far the random matrix $\boldsymbol{R}$ deviates from its mean $\mathbb{E}\,\boldsymbol{R}$. We can derive many useful consequences from a bound of the form

$$\left(\mathbb{E} \left\| \boldsymbol{R} - \mathbb{E}\,\boldsymbol{R} \right\|^2 \right)^{1/2} \quad \le \quad \dots$$

This estimate shows that, on average, all of the singular values of $\boldsymbol{R}$ are close to the corresponding singular values of $\mathbb{E}\,\boldsymbol{R}$. It also implies that, on average, the singular vectors of $\boldsymbol{R}$ are close to the corresponding singular vectors of $\mathbb{E}\,\boldsymbol{R}$, provided that the associated singular values are isolated. Furthermore, we discover that, on average, each linear functional $\operatorname{tr}[\boldsymbol{C}\boldsymbol{R}]$ is uniformly close to $\mathbb{E} \operatorname{tr}[\boldsymbol{C}\boldsymbol{R}]$ for each fixed matrix $\boldsymbol{C} \in \mathbb{M}^{d_2 \times d_1}$ with bounded Schatten 1-norm $\|\boldsymbol{C}\|_{S_1} \le 1$.

1.5. History.

Theorem I is not new. An early version of the upper bound appeared in Rudelson's work [Rud99, Thm. 1]; see also [RV07, Thm. 3.1] and [Tro08, Sec. 9]. The first explicit statement of the upper bound appeared in [CGT12, Thm. A.1]. All of these results depend on the noncommutative Khintchine inequality [LP86, Pis98, Buc01]. In our approach, the main innovation is a particularly easy proof of a Khintchine-type inequality for matrices, patterned after [MJC$^+$14, Cor 7.3] and [Tro15b, Thm. 8.1].

The ideas behind the proof of the lower bound in Theorem I are older. This estimate depends on generic considerations about the behavior of a sum of independent random variables in a Banach space. These techniques are explained in detail in [LT11, Ch. 6]. Our presentation expands on a proof sketch that appears in the monograph [Tro15a, Secs. 5.1.2 and 6.1.2].

1.6. **Target Audience.** This paper is intended for students and researchers who want to develop a detailed understanding of the foundations of matrix concentration. The preparation required is modest.

- **Basic Convexity.** Some simple ideas from convexity play a role, notably the concept of a convex function and Jensen's inequality.
- **Intermediate Linear Algebra.** The requirements from linear algebra are more substantial. The reader should be familiar with the spectral theorem for Hermitian (or symmetric) matrices, Rayleigh's variational principle, the trace of a matrix, and the spectral norm. The paper includes reminders about this material. The paper elaborates on some less familiar ideas, including inequalities for the trace and the spectral norm.
- **Intermediate Probability.** The paper demands some comfort with probability. The most important concepts are expectation and the elementary theory of conditional expectation. We develop the other key ideas, including the notion of symmetrization.

Although many readers will find the background material unnecessary, it is hard to locate these ideas in one place and we prefer to make the paper self-contained. In any case, we provide detailed cross-references so that the reader may dive into the proofs of the main results without wading through the shallower part of the paper.

1.7. **Roadmap.** Section 2 and Section 3 contain the background material from linear algebra and probability. To prove the upper bound in Theorem I, the key step is to establish the result for the special case of a sum of fixed matrices, each modulated by a random sign. This result appears in Section 4. In Section 5, we exploit this result to obtain the upper bound in (1.5). In Section 6, we present the easier proof of the lower bound in (1.5). Finally, Section 7 shows that it is not possible to improve (1.5) substantially.

2. Linear Algebra Background

Our aim is to make this paper as accessible as possible. To that end, this section presents some background material from linear algebra. Good references include [Hal74, Bha97, HJ13]. We also assume some familiarity with basic ideas from the theory of convexity, which may be found in the books [Lue69, Roc97, Bar02, BV04].

2.1. **Convexity.** Let V be a finite-dimensional linear space. A subset $E \subset V$ is *convex* when

$$\boldsymbol{x}, \boldsymbol{y} \in E \quad \text{implies} \quad \tau \cdot \boldsymbol{x} + (1-\tau) \cdot \boldsymbol{y} \in E \quad \text{for each } \tau \in [0,1].$$

Let E be a convex subset of a linear space V. A function $f : E \to \mathbb{R}$ is *convex* if

$$f\big(\tau \boldsymbol{x} + (1-\tau)\boldsymbol{y}\big) \le \tau \cdot f(\boldsymbol{x}) + (1-\tau) \cdot f(\boldsymbol{y}) \quad \text{for all } \tau \in [0,1] \text{ and all } \boldsymbol{x}, \boldsymbol{y} \in V. \tag{2.1}$$

We say that f is *concave* when $-f$ is convex.

2.2. **Vector Basics.** Let $\mathbb{C}^d$ be the complex linear space of d-dimensional complex vectors, equipped with the usual componentwise addition and scalar multiplication. The ℓ_2 norm $\|\cdot\|$ is defined on $\mathbb{C}^d$ via the expression

$$\|\boldsymbol{x}\|^2 := \boldsymbol{x}^* \boldsymbol{x} \quad \text{for each } \boldsymbol{x} \in \mathbb{C}^d.$$

The symbol * denotes the conjugate transpose of a vector. Recall that the ℓ_2 norm is a convex function.

A family $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_d\} \subset \mathbb{C}^d$ is called an *orthonormal basis* if it satisfies the relations

$$\boldsymbol{u}_i^* \boldsymbol{u}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

The orthonormal basis also has the property

$$\textstyle\sum_{i=1}^d \boldsymbol{u}_i \boldsymbol{u}_i^* = \mathbf{I}_d$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix.

2.3. **Matrix Basics.** A *matrix* is a rectangular array of complex numbers. Addition and multiplication by a complex scalar are defined componentwise, and we can multiply two matrices with compatible dimensions. We write $\mathbb{M}^{d_1 \times d_2}$ for the complex linear space of $d_1 \times d_2$ matrices. The symbol * also refers to the conjugate transpose operation on matrices.

A square matrix $\boldsymbol{H}$ is *Hermitian* when $\boldsymbol{H} = \boldsymbol{H}^*$. Hermitian matrices are sometimes called *conjugate symmetric*. We introduce the set of $d \times d$ Hermitian matrices:

$$\mathbb{H}_d := \left\{ \boldsymbol{H} \in \mathbb{M}^{d \times d} : \boldsymbol{H} = \boldsymbol{H}^* \right\}.$$

Note that the set $\mathbb{H}_d$ is a linear space over the real field.

An Hermitian matrix $\boldsymbol{A} \in \mathbb{H}_d$ is *positive semidefinite* when

$$\boldsymbol{u}^* \boldsymbol{A} \boldsymbol{u} \geq 0 \quad \text{for each } \boldsymbol{u} \in \mathbb{C}^d.$$

It is convenient to use the notation $\boldsymbol{A} \preccurlyeq \boldsymbol{H}$ to mean that $\boldsymbol{H} - \boldsymbol{A}$ is positive semidefinite. In particular, the relation $\mathbf{0} \preccurlyeq \boldsymbol{H}$ is equivalent to $\boldsymbol{H}$ being positive semidefinite. Observe that

$$\mathbf{0} \preccurlyeq \boldsymbol{A} \quad \text{and} \quad \mathbf{0} \preccurlyeq \boldsymbol{H} \quad \text{implies} \quad \mathbf{0} \preccurlyeq \alpha \cdot (\boldsymbol{A} + \boldsymbol{H}) \quad \text{for each } \alpha \geq 0.$$

In other words, addition and nonnegative scaling preserve the positive-semidefinite property.

For every matrix $\boldsymbol{B}$, both of its squares $\boldsymbol{B}\boldsymbol{B}^*$ and $\boldsymbol{B}^*\boldsymbol{B}$ are Hermitian and positive semidefinite.

2.4. **Basic Spectral Theory.** Each Hermitian matrix $\boldsymbol{H} \in \mathbb{H}_d$ can be expressed in the form

$$\boldsymbol{H} = \textstyle\sum_{i=1}^d \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^* \tag{2.2}$$

where the λ_i are uniquely determined real numbers, called *eigenvalues*, and $\{\boldsymbol{u}_i\}$ is an orthonormal basis for $\mathbb{C}^d$. The representation (2.2) is called an *eigenvalue decomposition*.

An Hermitian matrix $\boldsymbol{H}$ is positive semidefinite if and only if its eigenvalues λ_i are all nonnegative. Indeed, using the eigenvalue decomposition (2.2), we see that

$$\boldsymbol{u}^* \boldsymbol{H} \boldsymbol{u} = \textstyle\sum_{i=1}^n \lambda_i \cdot \boldsymbol{u}^* \boldsymbol{u}_i \boldsymbol{u}_i^* \boldsymbol{u} = \sum_{i=1}^n \lambda_i \cdot |\boldsymbol{u}^* \boldsymbol{u}_i|^2.$$

To verify the forward direction, select $\boldsymbol{u} = \boldsymbol{u}_j$ for each index j. The reverse direction should be obvious.

We define a *monomial* function of an Hermitian matrix $\boldsymbol{H} \in \mathbb{H}_d$ by repeated multiplication:

$$\boldsymbol{H}^0 = \mathbf{I}_d, \quad \boldsymbol{H}^1 = \boldsymbol{H}, \quad \boldsymbol{H}^2 = \boldsymbol{H} \cdot \boldsymbol{H}, \quad \boldsymbol{H}^3 = \boldsymbol{H} \cdot \boldsymbol{H}^2, \quad \text{etc.}$$

For each nonnegative integer r, it is not hard to check that

$$\boldsymbol{H} = \textstyle\sum_{i=1}^d \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^* \quad \text{implies} \quad \boldsymbol{H}^r = \sum_{i=1}^d \lambda_i^r \boldsymbol{u}_i \boldsymbol{u}_i^*. \tag{2.3}$$

In particular, $\boldsymbol{H}^{2p}$ is positive semidefinite for each nonnegative integer p.

2.5. **Rayleigh's Variational Principle.** The *Rayleigh principle* is an attractive expression for the maximum eigenvalue $\lambda_{\max}(\boldsymbol{H})$ of an Hermitian matrix $\boldsymbol{H} \in \mathbb{H}_d$. This result states that

$$\lambda_{\max}(\boldsymbol{H}) = \max_{\|\boldsymbol{u}\|=1} \boldsymbol{u}^* \boldsymbol{H} \boldsymbol{u}. \tag{2.4}$$

The maximum takes place over all unit-norm vectors $\boldsymbol{u} \in \mathbb{C}^d$. The identity (2.4) follows from the Lagrange multiplier theorem and the existence of the eigenvalue decomposition (2.2). Similarly, the minimum eigenvalue $\lambda_{\min}(\boldsymbol{H})$ satisfies

$$\lambda_{\min}(\boldsymbol{H}) = \min_{\|\boldsymbol{u}\|=1} \boldsymbol{u}^* \boldsymbol{H} \boldsymbol{u}. \tag{2.5}$$

We can obtain (2.5) by applying (2.4) to $-\boldsymbol{H}$.

Rayleigh's principle implies that order relations for positive-semidefinite matrices lead to order relations for their eigenvalues.

Fact 2.1 (Monotonicity)**.** *Let $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_d$ be Hermitian matrices. Then*

$$\boldsymbol{A} \preccurlyeq \boldsymbol{H} \quad \textit{implies} \quad \lambda_{\max}(\boldsymbol{A}) \leq \lambda_{\max}(\boldsymbol{H}).$$

Proof. The condition $\boldsymbol{A} \preccurlyeq \boldsymbol{H}$ implies that the eigenvalues of $\boldsymbol{H} - \boldsymbol{A}$ are nonnegative. Therefore, Rayleigh's principle (2.5) yields

$$0 \le \lambda_{\min}(\boldsymbol{H} - \boldsymbol{A}) = \min_{\|\boldsymbol{u}\|=1} \left(\boldsymbol{u}^* \boldsymbol{H}\boldsymbol{u} - \boldsymbol{u}^* \boldsymbol{A}\boldsymbol{u}\right) \le \boldsymbol{v}^* \boldsymbol{H}\boldsymbol{v} - \boldsymbol{v}^* \boldsymbol{A}\boldsymbol{v}$$

for any unit-norm vector $\boldsymbol{v}$. Select a unit-norm vector $\boldsymbol{v}$ for which $\lambda_{\max}(\boldsymbol{A}) = \boldsymbol{v}^* \boldsymbol{A}\boldsymbol{v}$, and then rearrange:

$$\lambda_{\max}(\boldsymbol{A}) = \boldsymbol{v}^* \boldsymbol{A}\boldsymbol{v} \le \boldsymbol{v}^* \boldsymbol{H}\boldsymbol{v} \le \lambda_{\max}(\boldsymbol{H}).$$

The last relation is Rayleigh's principle (2.4). □

2.6. **The Trace.** The *trace* of a square matrix $\boldsymbol{B} \in \mathbb{M}^{d\times d}$ is defined as

$$\operatorname{tr} \boldsymbol{B} := \textstyle\sum_{i=1}^{d} b_{ii}. \tag{2.6}$$

It is clear that the trace is a linear functional on $\mathbb{M}^{d\times d}$. By direct calculation, one may verify that

$$\operatorname{tr}[\boldsymbol{B}\boldsymbol{C}] = \operatorname{tr}[\boldsymbol{C}\boldsymbol{B}] \quad \text{for all } \boldsymbol{B} \in \mathbb{M}^{d\times r} \text{ and } \boldsymbol{C} \in \mathbb{M}^{r\times d}.$$

This property is called the *cyclicity* of the trace.

The trace of an Hermitian matrix $\boldsymbol{H} \in \mathbb{H}_d$ can also be expressed in terms of its eigenvalues:

$$\operatorname{tr} \boldsymbol{H} = \textstyle\sum_{i=1}^{d} \lambda_i. \tag{2.7}$$

This formula follows when we introduce the eigenvalue decomposition (2.2) into (2.6). Then we invoke the linearity and the cyclicity properties of the trace, as well as the properties of an orthonormal basis. We also instate the convention that monomials bind before the trace: $\operatorname{tr} \boldsymbol{H}^r := \operatorname{tr}[\boldsymbol{H}^r]$ for each nonnegative integer r.

2.7. **The Spectral Norm.** The *spectral norm* of a matrix $\boldsymbol{B} \in \mathbb{M}^{d_1\times d_2}$ is defined as

$$\|\boldsymbol{B}\| := \max_{\|\boldsymbol{u}\|=1} \|\boldsymbol{B}\boldsymbol{u}\|.$$

The maximum takes place over unit-norm vectors $\boldsymbol{u} \in \mathbb{C}^{d_2}$. We have the important identity

$$\|\boldsymbol{B}\|^2 = \|\boldsymbol{B}^*\boldsymbol{B}\| = \|\boldsymbol{B}\boldsymbol{B}^*\| \quad \text{for every matrix } \boldsymbol{B}. \tag{2.8}$$

Furthermore, the spectral norm is a convex function, and it satisfies the triangle inequality.

For an Hermitian matrix, the spectral norm can be written in terms of the eigenvalues:

$$\|\boldsymbol{H}\| = \max\left\{\lambda_{\max}(\boldsymbol{H}),\ -\lambda_{\min}(\boldsymbol{H})\right\} \quad \text{for each Hermitian matrix } \boldsymbol{H}. \tag{2.9}$$

As a consequence,

$$\|\boldsymbol{A}\| = \lambda_{\max}(\boldsymbol{A}) \quad \text{for each positive-semidefinite matrix } \boldsymbol{A}. \tag{2.10}$$

This discussion implies that

$$\|\boldsymbol{H}\|^{2p} = \|\boldsymbol{H}^{2p}\| \quad \text{for each Hermitian } \boldsymbol{H} \text{ and each nonnegative integer } p. \tag{2.11}$$

Use the relations (2.3) and (2.9) to verify this fact.

2.8. **Some Spectral Norm Inequalities.** We need some basic inequalities for the spectral norm. First, note that

$$\|\boldsymbol{A}\| \le \operatorname{tr} \boldsymbol{A} \quad \text{when } \boldsymbol{A} \text{ is positive semidefinite.} \tag{2.12}$$

This point follows from (2.10) and (2.7) because the eigenvalues of a positive-semidefinite matrix are nonnegative.

The next result uses the spectral norm to bound the trace of a product.

Fact 2.2 (Bound for the Trace of a Product)**.** *Consider Hermitian matrices $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_d$, and assume that $\boldsymbol{A}$ is positive semidefinite. Then*

$$\operatorname{tr}[\boldsymbol{H}\boldsymbol{A}] \le \|\boldsymbol{H}\| \cdot \operatorname{tr} \boldsymbol{A}.$$

Proof. Introducing the eigenvalue decomposition $\boldsymbol{A} = \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^*$, we see that

$$\operatorname{tr}[\boldsymbol{HA}] = \sum_i \lambda_i \operatorname{tr}[\boldsymbol{H}\boldsymbol{u}_i\boldsymbol{u}_i^*] = \sum_i \lambda_i \boldsymbol{u}_i^* \boldsymbol{H}\boldsymbol{u}_i \leq \lambda_{\max}(\boldsymbol{H}) \cdot \sum_i \lambda_i \leq \|\boldsymbol{H}\| \cdot \operatorname{tr} \boldsymbol{A}.$$

The first two relations follow from linearity and cyclicity of the trace. The first inequality depends on Rayleigh's principle (2.4) and the nonnegativity of the eigenvalues λ_i. The last bound follows from (2.9). □

We also need a bound for the norm of a sum of squared positive-semidefinite matrices.

Fact 2.3 (Bound for a Sum of Squares)**.** *Consider positive-semidefinite matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_n \in \mathbb{H}_d$. Then*

$$\left\| \sum_{i=1}^n \boldsymbol{A}_i^2 \right\| \leq \max_i \|\boldsymbol{A}_i\| \cdot \left\| \sum_{i=1}^n \boldsymbol{A}_i \right\|.$$

Proof. Let $\boldsymbol{A}$ be positive semidefinite. We claim that

$$\boldsymbol{A}^2 \preccurlyeq M \cdot \boldsymbol{A} \quad \text{whenever } \lambda_{\max}(\boldsymbol{A}) \leq M. \tag{2.13}$$

Indeed, introducing the eigenvalue decomposition $\boldsymbol{A} = \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^*$, we find that

$$M \cdot \boldsymbol{A} - \boldsymbol{A}^2 = M \cdot \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^* - \sum_i \lambda_i^2 \boldsymbol{u}_i \boldsymbol{u}_i^* = \sum_i (M - \lambda_i) \lambda_i \cdot \boldsymbol{u}_i \boldsymbol{u}_i^*.$$

The first relation uses (2.3). Since $0 \leq \lambda_i \leq M$, the scalar coefficients in the sum are nonnegative. Therefore, the matrix $M \cdot \boldsymbol{A} - \boldsymbol{A}^2$ is positive semidefinite, which is what we needed to show.

Select $M := \max_i \lambda_{\max}(\boldsymbol{A}_i)$. The inequality (2.13) ensures that

$$\boldsymbol{A}_i^2 \preccurlyeq M \cdot \boldsymbol{A}_i \quad \text{for each index } i.$$

Summing these relations, we see that

$$\sum_{i=1}^n \boldsymbol{A}_i^2 \preccurlyeq M \cdot \sum_{i=1}^n \boldsymbol{A}_i.$$

The monotonicity principle, Fact 2.1, yields the inequality

$$\lambda_{\max}\left(\sum_{i=1}^n \boldsymbol{A}_i^2\right) \leq \lambda_{\max}\left(M \cdot \sum_{i=1}^n \boldsymbol{A}_i\right) = M \cdot \lambda_{\max}\left(\sum_{i=1}^n \boldsymbol{A}_i\right).$$

We have used the fact that the maximum eigenvalue of an Hermitian matrix is positive homogeneous. Finally, recall that, per (2.10), the spectral norm of a positive-semidefinite matrix equals its maximum eigenvalue. □

2.9. **GM–AM Inequality for the Trace.** We require another substantial matrix inequality, which is one (of several) matrix analogs of the inequality between the geometric mean and the arithmetic mean.

Fact 2.4 (GM–AM Trace Inequality)**.** *Consider Hermitian matrices $\boldsymbol{H}, \boldsymbol{W}, \boldsymbol{Y} \in \mathbb{H}_d$. For each nonnegative integer r and each integer q in the range $0 \leq q \leq 2r$,*

$$\operatorname{tr}\left[\boldsymbol{H}\boldsymbol{W}^q \boldsymbol{H}\boldsymbol{Y}^{2r-q}\right] + \operatorname{tr}\left[\boldsymbol{H}\boldsymbol{W}^{2r-q}\boldsymbol{H}\boldsymbol{Y}^q\right] \leq \operatorname{tr}\left[\boldsymbol{H}^2 \cdot \left(\boldsymbol{W}^{2r} + \boldsymbol{Y}^{2r}\right)\right]. \tag{2.14}$$

In particular,

$$\sum_{q=0}^{2r} \operatorname{tr}\left[\boldsymbol{H}\boldsymbol{W}^q \boldsymbol{H}\boldsymbol{Y}^{2r-q}\right] \leq \frac{2r+1}{2} \operatorname{tr}\left[\boldsymbol{H}^2 \cdot \left(\boldsymbol{W}^{2r} + \boldsymbol{Y}^{2r}\right)\right].$$

Proof. The result (2.14) is a matrix version of the following numerical inequality. For $\lambda, \mu \geq 0$,

$$\lambda^\theta \mu^{1-\theta} + \lambda^{1-\theta} \mu^\theta \leq \lambda + \mu \quad \text{for each } \theta \in [0,1]. \tag{2.15}$$

To verify this bound, we may assume that $\lambda, \mu > 0$ because it is trivial to check when either λ or μ equals zero. Notice that the left-hand side of the bound is a convex function of θ on the interval $[0,1]$. This point follows easily from the representation

$$f(\theta) := \lambda^\theta \mu^{1-\theta} + \lambda^{1-\theta}\mu^\theta = \mathrm{e}^{\theta \log\lambda + (1-\theta)\log\mu} + \mathrm{e}^{(1-\theta)\log\lambda + \theta\log\mu}.$$

The value of the convex function f on the interval $[0,1]$ is controlled by the maximum value it achieves at one of the endpoints:

$$f(\theta) \leq \max\{f(0), f(1)\} = \lambda + \mu.$$

This inequality coincides with (2.15).

To prove (2.14) from (2.15), we use eigenvalue decompositions. The case $r = 0$ is immediate, so we may assume that $r \geq 1$. Let q be an integer in the range $0 \leq q \leq 2r$. Introduce eigenvalue decompositions:

$$\boldsymbol{W} = \sum_{i=1}^{d} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^* \quad \text{and} \quad \boldsymbol{Y} = \sum_{j=1}^{d} \mu_j \boldsymbol{v}_j \boldsymbol{v}_j^*.$$

Calculate that

$$\begin{aligned} \operatorname{tr}\left[\boldsymbol{H}\boldsymbol{W}^q \boldsymbol{H}\boldsymbol{Y}^{2r-q}\right] &= \operatorname{tr}\left[\boldsymbol{H}\big(\textstyle\sum_{i=1}^{d} \lambda_i^q \boldsymbol{u}_i \boldsymbol{u}_i^*\big)\boldsymbol{H}\big(\textstyle\sum_{j=1}^{d} \mu_j^{2r-q} \boldsymbol{v}_j \boldsymbol{v}_j^*\big)\right] \\ &= \textstyle\sum_{i,j=1}^{d} \lambda_i^q \mu_j^{2r-q} \cdot \operatorname{tr}\left[\boldsymbol{H}\boldsymbol{u}_i \boldsymbol{u}_i^* \boldsymbol{H}\boldsymbol{v}_j \boldsymbol{v}_j^*\right] \\ &\leq \textstyle\sum_{i,j=1}^{d} |\lambda_i|^q |\mu_j|^{2r-q} \cdot \left|\boldsymbol{u}_i^* \boldsymbol{H}\boldsymbol{v}_j\right|^2. \end{aligned} \tag{2.16}$$

The first identity relies on the formula (2.3) for the eigenvalue decomposition of a monomial. The second step depends on the linearity of the trace. In the last line, we rewrite the trace using cyclicity, and the inequality emerges when we apply absolute values. The representation $|\boldsymbol{u}_i^* \boldsymbol{H}\boldsymbol{v}_j|^2$ emphasizes that this quantity is nonnegative, which we use to justify several inequalities.

Invoking the inequality (2.16) twice, we arrive at the bound

$$\begin{aligned} \operatorname{tr}\left[\boldsymbol{H}\boldsymbol{W}^q \boldsymbol{H}\boldsymbol{Y}^{2r-q}\right] + \operatorname{tr}\left[\boldsymbol{H}\boldsymbol{W}^{2r-q} \boldsymbol{H}\boldsymbol{Y}^{q}\right] &\leq \textstyle\sum_{i,j=1}^{d} \left(|\lambda_i|^q |\mu_j|^{2r-q} + |\lambda_i|^{2r-q} |\mu_j|^q\right) \cdot \left|\boldsymbol{u}_i^* \boldsymbol{H}\boldsymbol{v}_j\right|^2 \\ &\leq \textstyle\sum_{i,j=1}^{d} \left(\lambda_i^{2r} + \mu_j^{2r}\right) \cdot \left|\boldsymbol{u}_i^* \boldsymbol{H}\boldsymbol{v}_j\right|^2. \end{aligned} \tag{2.17}$$

The second inequality is (2.15), with $\theta = q/(2r)$ and $\lambda = \lambda_i^{2r}$ and $\mu = \mu_j^{2r}$.

It remains to rewrite the right-hand side of (2.17) a more recognizable form. To that end, observe that

$$\begin{aligned} &\operatorname{tr}\left[\boldsymbol{H}\boldsymbol{W}^q \boldsymbol{H}\boldsymbol{Y}^{2r-q}\right] + \operatorname{tr}\left[\boldsymbol{H}\boldsymbol{W}^{2r-q} \boldsymbol{H}\boldsymbol{Y}^{q}\right] \\ &\qquad \leq \textstyle\sum_{i,j=1}^{d} \left(\lambda_i^{2r} + \mu_j^{2r}\right) \cdot \operatorname{tr}\left[\boldsymbol{H}\boldsymbol{u}_i \boldsymbol{u}_i^* \boldsymbol{H}\boldsymbol{v}_j \boldsymbol{v}_j^*\right] \\ &\qquad = \operatorname{tr}\left[\boldsymbol{H}\big(\textstyle\sum_{i=1}^{d} \lambda_i^{2r} \boldsymbol{u}_i \boldsymbol{u}_i^*\big)\boldsymbol{H}\big(\textstyle\sum_{j=1}^{d} \boldsymbol{v}_j \boldsymbol{v}_j^*\big)\right] + \operatorname{tr}\left[\boldsymbol{H}\big(\textstyle\sum_{i=1}^{d} \boldsymbol{u}_i \boldsymbol{u}_i^*\big)\boldsymbol{H}\big(\textstyle\sum_{j=1}^{d} \mu_j^{2r} \boldsymbol{v}_j \boldsymbol{v}_j^*\big)\right] \\ &\qquad = \operatorname{tr}\left[\boldsymbol{H}^2 \cdot \boldsymbol{W}^{2r}\right] + \operatorname{tr}\left[\boldsymbol{H}^2 \cdot \boldsymbol{Y}^{2r}\right]. \end{aligned}$$

In the first step, we return the squared magnitude to its representation as a trace. Then we use linearity to draw the sums back inside the trace. Next, invoke (2.3) to identify the powers $\boldsymbol{W}^{2r}$ and $\boldsymbol{Y}^0 = \mathbf{I}_d$ and $\boldsymbol{W}^0 = \mathbf{I}_d$ and $\boldsymbol{Y}^{2r}$. Last, use the cyclicity of the trace to combine the factors of $\boldsymbol{H}$. The result (2.14) follows from the linearity of the trace. □

Remark 2.5 (The Power of Abstraction)**.** There is a cleaner, but more abstract, proof of the inequality (2.14). Consider the left- and right-multiplication operators

$$\mathsf{W} : \boldsymbol{H} \mapsto \boldsymbol{W}\boldsymbol{H} \quad \text{and} \quad \mathsf{Y} : \boldsymbol{H} \mapsto \boldsymbol{H}\boldsymbol{Y}.$$

Observe that powers of W and Y correspond to left- and right-multiplication by powers of $\boldsymbol{W}$ and $\boldsymbol{Y}$. Now, the operators W and Y commute, so there is a basis (orthonormal with respect to the trace inner product) for $\mathbb{H}_d$ in which they are simultaneously diagonalizable. Representing the operators in this basis, we can use (2.15) to check that

$$\mathsf{W}^q \mathsf{Y}^{2r-q} + \mathsf{W}^{2r-q} \mathsf{Y}^q \preccurlyeq \mathsf{W}^{2r} + \mathsf{Y}^{2r}.$$

Now, calculate that

$$\begin{aligned} \operatorname{tr}\left[\boldsymbol{H}\boldsymbol{W}^q \boldsymbol{H}\boldsymbol{Y}^{2r-q}\right] + \operatorname{tr}\left[\boldsymbol{H}\boldsymbol{W}^{2r-q} \boldsymbol{H}\boldsymbol{Y}^{q}\right] &= \operatorname{tr}\left[\boldsymbol{H} \cdot \left(\mathsf{W}^q \mathsf{Y}^{2r-q} + \mathsf{W}^{2r-q} \mathsf{Y}^q\right)(\boldsymbol{H})\right] \\ &\leq \operatorname{tr}\left[\boldsymbol{H} \cdot \left(\mathsf{W}^{2r} + \mathsf{Y}^{2r}\right)(\boldsymbol{H})\right] \\ &= \operatorname{tr}\left[\boldsymbol{H}^2 \cdot \left(\boldsymbol{W}^{2r} + \boldsymbol{Y}^{2r}\right)\right]. \end{aligned}$$

We omit the details.

2.10. **The Hermitian Dilation.** Last, we introduce the *Hermitian dilation* $\mathscr{H}(\boldsymbol{B})$ of a rectangular matrix $\boldsymbol{B} \in \mathbb{M}^{d_1 \times d_2}$. This is the Hermitian matrix

$$\mathscr{H}(\boldsymbol{B}) := \begin{bmatrix} \mathbf{0} & \boldsymbol{B} \\ \boldsymbol{B}^* & \mathbf{0} \end{bmatrix} \in \mathbb{H}_{d_1+d_2}. \tag{2.18}$$

Note that the map $\mathscr{H}$ is real linear. By direct calculation,

$$\mathscr{H}(\boldsymbol{B})^2 = \begin{bmatrix} \boldsymbol{B}\boldsymbol{B}^* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B}^*\boldsymbol{B} \end{bmatrix}. \tag{2.19}$$

We also have the spectral-norm identity

$$\|\mathscr{H}(\boldsymbol{B})\| = \|\boldsymbol{B}\|. \tag{2.20}$$

To verify (2.20), calculate that

$$\|\mathscr{H}(\boldsymbol{B})\|^2 = \left\|\mathscr{H}(\boldsymbol{B})^2\right\| = \left\| \begin{bmatrix} \boldsymbol{B}\boldsymbol{B}^* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B}^*\boldsymbol{B} \end{bmatrix} \right\| = \max\left\{ \|\boldsymbol{B}\boldsymbol{B}^*\|, \|\boldsymbol{B}^*\boldsymbol{B}\| \right\} = \|\boldsymbol{B}\|^2.$$

The first identity is (2.11); the second is (2.19). The norm of a block-diagonal Hermitian matrix is the maximum spectral norm of a block, which follows from the Rayleigh principle (2.4) with a bit of work. Finally, invoke the property (2.8).

3. Probability Background

This section contains some background material from the field of probability. Good references include the books [LT11, GS01, Tao12].

3.1. **Expectation.** The symbol $\mathbb{E}$ denotes the expectation operator. We will not define expectation formally or spend any energy on technical details. No issues arise if we assume, for example, that all random variables are bounded.

We use brackets to enclose the argument of the expectation when it is important for clarity, and we instate the convention that nonlinear functions bind before expectation. For instance, $\mathbb{E}\, X^p := \mathbb{E}[X^p]$ and $\mathbb{E}\max_i X_i := \mathbb{E}[\max_i X_i]$.

Sometimes, we add a subscript to indicate a partial expectation. For example, if J is a random variable, $\mathbb{E}_J$ refers to the average over J, with all other random variables fixed. We only use this notation when J is independent from the other random variables, so there are no complications. In particular, we can compute iterated expectations: $\mathbb{E}[\mathbb{E}_J[\cdot]] = \mathbb{E}[\cdot]$ whenever all the expectations are finite.

3.2. **Random Matrices.** A *random matrix* is a matrix whose entries are complex random variables, not necessarily independent. We compute the expectation of a random matrix $\boldsymbol{Z}$ componentwise:

$$(\mathbb{E}[\boldsymbol{Z}])_{ij} = \mathbb{E}[z_{ij}] \quad \text{for each pair } (i,j) \text{ of indices.}$$

As in the scalar case, if $\boldsymbol{W}$ and $\boldsymbol{Z}$ are independent,

$$\mathbb{E}[\boldsymbol{W}\boldsymbol{Z}] = (\mathbb{E}\,\boldsymbol{W})(\mathbb{E}\,\boldsymbol{Z}).$$

Since the expectation is linear, it also commutes with all of the simple linear operations we perform on matrices.

It suffices to take a naïve view of independence, expectation, and so forth. For the technically inclined, let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. A $d_1 \times d_2$ random matrix $\boldsymbol{Z}$ is simply a measurable function

$$\boldsymbol{Z} : \Omega \to \mathbb{M}^{d_1 \times d_2}.$$

A family $\{\boldsymbol{Z}_i : i = 1, \dots, n\}$ of random matrices is independent when

$$\mathbb{P}\{\boldsymbol{Z}_i \in E_i \text{ for } i = 1, \dots, n\} = \textstyle\prod_{i=1}^n \mathbb{P}\{\boldsymbol{Z}_i \in E_i\}.$$

for any collection of Borel[1] subsets $E_i \subset \mathbb{M}^{d_1 \times d_2}$.

[1] Open sets in $\mathbb{M}^{d_1 \times d_2}$ are defined with respect to the metric topology induced by the spectral norm.

3.3. **Inequalities for Expectation.** We need several basic inequalities for expectation. We set these out for future reference. Let X, Y be (arbitrary) real random variables. The Cauchy–Schwarz inequality states that

$$|\mathbb{E}[XY]| \le \left(\mathbb{E}\, X^2\right)^{1/2} \cdot \left(\mathbb{E}\, Y^2\right)^{1/2}. \tag{3.1}$$

For $r \ge 1$, the triangle inequality states that

$$\left(\mathbb{E}\,|X+Y|^r\right)^{1/r} \le \left(\mathbb{E}\,|X|^r\right)^{1/r} + \left(\mathbb{E}\,|Y|^r\right)^{1/r}. \tag{3.2}$$

Each of these inequalities is vacuous precisely when its right-hand side is infinite.

Jensen's inequality describes how expectation interacts with a convex or concave function; cf. (2.1). Let X be a random variable taking values in a finite-dimensional linear space V, and let $f : V \to \mathbb{R}$ be a function. Then

$$\begin{aligned} f(\mathbb{E}\, X) &\le \mathbb{E}\, f(X) \quad \text{when } f \text{ is convex, and} \\ \mathbb{E}\, f(X) &\le f(\mathbb{E}\, X) \quad \text{when } f \text{ is concave.} \end{aligned} \tag{3.3}$$

The inequalities (3.3) also hold when we replace $\mathbb{E}$ with a partial expectation. Let us emphasize that these bounds do require that all of the expectations exist.

3.4. **Symmetrization.** Symmetrization is an important technique for studying the expectation of a function of independent random variables. The idea is to inject auxiliary randomness into the function. Then we condition on the original random variables and average with respect to the extra randomness. When the auxiliary random variables are more pliable, this approach can lead to significant simplifications.

A *Rademacher* random variable ε takes the two values ± 1 with equal probability. The following result shows how we can use Rademacher random variables to study a sum of independent random matrices.

Fact 3.1 (Symmetrization)**.** *Let $\boldsymbol{S}_1, \dots, \boldsymbol{S}_n \in \mathbb{M}^{d_1 \times d_2}$ be independent random matrices. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables that are also independent from the random matrices. For each $r \ge 1$,*

$$\frac{1}{2} \cdot \left(\mathbb{E}\left\|\sum\nolimits_{i=1}^n \varepsilon_i \boldsymbol{S}_i\right\|^r\right)^{1/r} \le \left(\mathbb{E}\left\|\sum\nolimits_{i=1}^n (\boldsymbol{S}_i - \mathbb{E}\,\boldsymbol{S}_i)\right\|^r\right)^{1/r} \le 2 \cdot \left(\mathbb{E}\left\|\sum\nolimits_{i=1}^n \varepsilon_i \boldsymbol{S}_i\right\|^r\right)^{1/r}.$$

This result holds whenever $\mathbb{E}\,\|\boldsymbol{S}_i\|^r < \infty$ for each index i.

Proof. For notational simplicity, assume that $r = 1$. We discuss the general case at the end of the argument.

Let $\{\boldsymbol{S}_i' : i = 1, \dots, n\}$ be an independent copy of the sequence $\{\boldsymbol{S}_i : i = 1, \dots, n\}$, and let $\mathbb{E}'$ denote partial expectation with respect to the independent copy. Then

$$\begin{aligned} \mathbb{E}\left\|\sum\nolimits_{i=1}^n (\boldsymbol{S}_i - \mathbb{E}\,\boldsymbol{S}_i)\right\| &= \mathbb{E}\left\|\sum\nolimits_{i=1}^n \left[(\boldsymbol{S}_i - \mathbb{E}\,\boldsymbol{S}_i) - \mathbb{E}'(\boldsymbol{S}_i' - \mathbb{E}\,\boldsymbol{S}_i)\right]\right\| \\ &\le \mathbb{E}\left[\mathbb{E}'\left\|\sum\nolimits_{i=1}^n \left[(\boldsymbol{S}_i - \mathbb{E}\,\boldsymbol{S}_i) - (\boldsymbol{S}_i' - \mathbb{E}\,\boldsymbol{S}_i)\right]\right\|\right] \\ &= \mathbb{E}\left\|\sum\nolimits_{i=1}^n (\boldsymbol{S}_i - \boldsymbol{S}_i')\right\|. \end{aligned}$$

The first identity holds because $\mathbb{E}'\,\boldsymbol{S}_i' = \mathbb{E}\,\boldsymbol{S}_i$ by identical distribution. Since the spectral norm is convex, we can apply Jensen's inequality (3.3) conditionally to draw out the partial expectation $\mathbb{E}'$. Last, we combine the iterated expectation into a single expectation.

Observe that $\boldsymbol{S}_i - \boldsymbol{S}_i'$ has the same distribution as its negation $\boldsymbol{S}_i' - \boldsymbol{S}_i$. It follows that the independent sequence $\{\varepsilon_i(\boldsymbol{S}_i - \boldsymbol{S}_i') : i = 1, \dots, n\}$ has the same distribution as $\{\boldsymbol{S}_i - \boldsymbol{S}_i' : i = 1, \dots, n\}$. Therefore, the expectation of any nonnegative function takes the same value for both sequences. In particular,

$$\begin{aligned} \mathbb{E}\left\|\sum\nolimits_{i=1}^n (\boldsymbol{S}_i - \boldsymbol{S}_i')\right\| &= \mathbb{E}\left\|\sum\nolimits_{i=1}^n \varepsilon_i (\boldsymbol{S}_i - \boldsymbol{S}_i')\right\| \\ &\le \mathbb{E}\left\|\sum\nolimits_{i=1}^n \varepsilon_i \boldsymbol{S}_i\right\| + \mathbb{E}\left\|\sum\nolimits_{i=1}^n (-\varepsilon_i)\boldsymbol{S}_i'\right\| \\ &= 2\,\mathbb{E}\left\|\sum\nolimits_{i=1}^n \varepsilon_i \boldsymbol{S}_i\right\|. \end{aligned}$$

The second step is the triangle inequality, and the last line follows from the identical distribution of $\{\varepsilon_i \boldsymbol{S}_i\}$ and $\{-\varepsilon_i \boldsymbol{S}_i'\}$. Combine the last two displays to obtain the upper bound.

To obtain results for $r > 1$, we pursue the same approach. We require the additional observation that $\|\cdot\|^r$ is a convex function, and we also need to invoke the triangle inequality (3.2). Finally, we remark that the lower bound follows from a similar procedure, so we omit the demonstration. □

4. The Expected Norm of a Matrix Rademacher Series

To prove Theorem I, our overall strategy is to use symmetrization. This approach allows us to reduce the study of an independent sum of random matrices to the study of a sum of fixed matrices modulated by independent Rademacher random variables. This type of random matrix is called a *matrix Rademacher series*. In this section, we establish a bound on the spectral norm of a matrix Rademacher series. This is the key technical step in the proof of Theorem I.

Theorem 4.1 (Matrix Rademacher Series)**.** *Let $\boldsymbol{H}_1, \dots, \boldsymbol{H}_n$ be fixed Hermitian matrices with dimension d. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables. Then*

$$\left(\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i \boldsymbol{H}_i\right\|^2\right)^{1/2} \leq \sqrt{1+2\lceil\log d\rceil} \cdot \left\|\sum_{i=1}^n \boldsymbol{H}_i^2\right\|^{1/2}. \tag{4.1}$$

The proof of Theorem 4.1 occupies the bulk of this section, beginning with Section 4.2. The argument is really just a fancy version of the familiar calculation of the moments of a centered standard normal random variable; see Section 4.8 for details.

4.1. **Discussion.** Before we establish Theorem 4.1, let us make a few comments. First, it is helpful to interpret the result in the same language we have used to state Theorem I. Introduce the matrix Rademacher series

$$\boldsymbol{X} := \sum_{i=1}^n \varepsilon_i \boldsymbol{H}_i.$$

Compute the matrix variance, defined in (1.2):

$$v(\boldsymbol{X}) := \left\|\mathbb{E}\boldsymbol{X}^2\right\| = \left\|\sum_{i,j=1}^n \mathbb{E}[\varepsilon_i\varepsilon_j] \cdot \boldsymbol{H}_i\boldsymbol{H}_j\right\| = \left\|\sum_{i=1}^n \boldsymbol{H}_i^2\right\|.$$

We may rewrite Theorem 4.1 as the statement that

$$\left(\mathbb{E}\,\|\boldsymbol{X}\|^2\right)^{1/2} \leq \sqrt{(1+2\lceil\log d\rceil) \cdot v(\boldsymbol{X})}.$$

In other words, Theorem 4.1 is a sharper version of Theorem I for the special case of a matrix Rademacher series.

Next, we have focused on bounding the second moment of $\|\boldsymbol{X}\|$ because this is the most natural form of the result. Note that we also control the first moment because of Jensen's inequality (3.3):

$$\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i \boldsymbol{H}_i\right\| \leq \left(\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i \boldsymbol{H}_i\right\|^2\right)^{1/2} \leq \sqrt{1+2\lceil\log d\rceil} \cdot \left\|\sum_{i=1}^n \boldsymbol{H}_i^2\right\|^{1/2}. \tag{4.2}$$

A simple variant on the proof of Theorem 4.1 provides bounds for higher moments.

Third, the dimensional factor on the right-hand side of (4.1) is asymptotically sharp. Indeed, let us write $K(d)$ for the minimum possible constant in the inequality

$$\left(\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i \boldsymbol{H}_i\right\|^2\right)^{1/2} \leq K(d) \cdot \left\|\sum_{i=1}^n \boldsymbol{H}_i^2\right\|^{1/2} \quad \text{for } \boldsymbol{H}_i \in \mathbb{H}_d \text{ and } n \in \mathbb{N}.$$

The example in Section 7.1 shows that

$$K(d) \geq \sqrt{2\log d}.$$

In other words, (4.1) cannot be improved without making further assumptions.

Theorem 4.1 is a variant on the noncommutative Khintchine inequality, first established by Lust-Piquard [LP86] and later improved by Pisier [Pis98] and by Buchholz [Buc01]. The noncommutative Khintchine inequality gives bounds for the Schatten norm of a matrix Rademacher series, rather than for the spectral norm. Rudelson [Rud99] pointed out that the noncommutative Khintchine inequality also implies bounds for the spectral norm of a matrix Rademacher series. In our presentation, we choose to control the spectral norm directly.

4.2. **The Spectral Norm and the Trace Moments.** To begin the proof of Theorem 4.1, we introduce the random Hermitian matrix

$$\boldsymbol{X} := \sum_{i=1}^n \varepsilon_i \boldsymbol{H}_i \tag{4.3}$$

Our goal is to bound the expected spectral norm of $\boldsymbol{X}$. We may proceed by estimating the expected trace of a power of the random matrix, which is known as a *trace moment*. Fix a nonnegative integer p. Observe that

$$\left(\mathbb{E}\,\|\boldsymbol{X}\|^2\right)^{1/2} \le \left(\mathbb{E}\,\|\boldsymbol{X}\|^{2p}\right)^{1/(2p)} = \left(\mathbb{E}\,\|\boldsymbol{X}^{2p}\|\right)^{1/(2p)} \le \left(\mathbb{E}\operatorname{tr}\boldsymbol{X}^{2p}\right)^{1/(2p)}. \tag{4.4}$$

The first identity is Jensen's inequality (3.3), applied to the concave function $t \mapsto t^{1/p}$. The second relation is (2.11). The final inequality is the bound (2.12) on the norm of the positive-semidefinite matrix $\boldsymbol{X}^{2p}$ by its trace.

Remark 4.2 (Higher Moments)**.** It should be clear that we can also bound expected powers of the spectral norm using the same technique. For simplicity, we omit this development.

4.3. **Summation by Parts.** To study the trace moments of the random matrix $\boldsymbol{X}$, we rely on a discrete analog of integration by parts. This approach is clearer if we introduce some more notation. For each index i, define the random matrices

$$\boldsymbol{X}_{+i} := \boldsymbol{H}_i + \sum_{j\neq i} \varepsilon_j \boldsymbol{H}_j \quad \text{and} \quad \boldsymbol{X}_{-i} := -\boldsymbol{H}_i + \sum_{j\neq i} \varepsilon_j \boldsymbol{H}_j$$

In other words, the distribution of $\boldsymbol{X}_{\varepsilon_i i}$ is the conditional distribution of the random matrix $\boldsymbol{X}$ given the value ε_i of the ith Rademacher variable. This interpretation depends on the assumption that the Rademacher variables are independent.

Beginning with the trace moment, observe that

$$\begin{aligned}
\mathbb{E}\operatorname{tr}\boldsymbol{X}^{2p} &= \mathbb{E}\operatorname{tr}\left[\boldsymbol{X}\cdot\boldsymbol{X}^{2p-1}\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\mathbb{E}_{\varepsilon_i}\operatorname{tr}\left[\varepsilon_i \boldsymbol{H}_i \cdot \boldsymbol{X}^{2p-1}\right]\right] \\
&= \frac{1}{2}\sum_{i=1}^n \mathbb{E}\operatorname{tr}\left[\boldsymbol{H}_i\cdot\left(\boldsymbol{X}_{+i}^{2p-1} - \boldsymbol{X}_{-i}^{2p-1}\right)\right]
\end{aligned} \tag{4.5}$$

In the second step, we simply write out the definition (4.3) of the random matrix $\boldsymbol{X}$ and use the linearity of the trace to draw out the sum. Then we write the expectation as an iterated expectation. To reach the next line, write out the partial expectation using the notation $\boldsymbol{X}_{\pm i}$ and the linearity of the trace.

4.4. **A Difference of Powers.** Next, let us apply an algebraic identity to reduce the difference of powers in (4.5). For matrices $\boldsymbol{W}, \boldsymbol{Y} \in \mathbb{H}_d$, it holds that

$$\boldsymbol{W}^{2p-1} - \boldsymbol{Y}^{2p-1} = \sum_{q=0}^{2p-2} \boldsymbol{W}^q(\boldsymbol{W}-\boldsymbol{Y})\boldsymbol{W}^{2p-2-q}. \tag{4.6}$$

To check this expression, just expand the matrix products and notice that the sum telescopes.

Introducing the relation (4.6) with $\boldsymbol{W} = \boldsymbol{X}_{+i}$ and $\boldsymbol{Y} = \boldsymbol{X}_{-i}$ into the formula (4.5), we find that

$$\begin{aligned}
\mathbb{E}\operatorname{tr}\boldsymbol{X}^{2p} &= \frac{1}{2}\sum_{i=1}^n \mathbb{E}\operatorname{tr}\left[\boldsymbol{H}_i\cdot\sum_{q=0}^{2p-2}\boldsymbol{X}_{+i}^q(\boldsymbol{X}_{+i}-\boldsymbol{X}_{-i})\boldsymbol{X}_{-i}^{2p-2-q}\right] \\
&= \sum_{i=1}^n\sum_{q=0}^{2p-2}\mathbb{E}\operatorname{tr}\left[\boldsymbol{H}_i\boldsymbol{X}_{+i}^q\boldsymbol{H}_i\boldsymbol{X}_{-i}^{2p-2-q}\right].
\end{aligned} \tag{4.7}$$

Linearity of the trace allows us to draw out the sum over q, and we have used the observation that $\boldsymbol{X}_{+i} - \boldsymbol{X}_{-i} = 2\boldsymbol{H}_i$.

4.5. **A Bound for the Trace Moments.** We are now in a position to obtain a bound for the trace moments of $\boldsymbol{X}$. Beginning with (4.7), we compute that

$$\begin{aligned}\mathbb{E}\operatorname{tr}\boldsymbol{X}^{2p} &= \textstyle\sum_{i=1}^{n}\sum_{q=0}^{2p-2}\mathbb{E}\operatorname{tr}\left[\boldsymbol{H}_i\boldsymbol{X}_{+i}^{q}\boldsymbol{H}_i\boldsymbol{X}_{-i}^{2p-2-q}\right]\\ &\le \textstyle\sum_{i=1}^{n}\dfrac{2p-1}{2}\mathbb{E}\operatorname{tr}\left[\boldsymbol{H}_i^2\cdot\left(\boldsymbol{X}_{+i}^{2p-2}+\boldsymbol{X}_{-i}^{2p-2}\right)\right]\\ &= (2p-1)\cdot\textstyle\sum_{i=1}^{n}\mathbb{E}\operatorname{tr}\left[\boldsymbol{H}_i^2\cdot\left(\mathbb{E}_{\varepsilon_i}\boldsymbol{X}^{2p-2}\right)\right]\\ &= (2p-1)\cdot\mathbb{E}\operatorname{tr}\left[\left(\textstyle\sum_{i=1}^{n}\boldsymbol{H}_i^2\right)\cdot\boldsymbol{X}^{2p-2}\right]\\ &\le (2p-1)\cdot\left\|\textstyle\sum_{i=1}^{n}\boldsymbol{H}_i^2\right\|\cdot\mathbb{E}\operatorname{tr}\boldsymbol{X}^{2p-2}.\end{aligned}\tag{4.8}$$

The bound in the second line is the trace GM–AM inequality, Fact 2.4, with $r = p-1$ and $\boldsymbol{W} = \boldsymbol{X}_{+i}$ and $\boldsymbol{Y} = \boldsymbol{X}_{-i}$. To reach the third line, observe that the parenthesis in the second line is twice the partial expectation of $\boldsymbol{X}^{2p-2}$ with respect to ε_i. Afterward, we use linearity of the expectation and the trace to draw in the sum over i, and then we combine the expectations. Last, invoke the trace inequality from Fact 2.2.

4.6. **Iteration and the Spectral Norm Bound.** The expression (4.8) shows that the trace moment is controlled by a trace moment with a smaller power:

$$\mathbb{E}\operatorname{tr}\boldsymbol{X}^{2p} \le (2p-1)\cdot\left\|\textstyle\sum_{i=1}^{n}\boldsymbol{H}_i^2\right\|\cdot\mathbb{E}\operatorname{tr}\boldsymbol{X}^{2p-2}.$$

Iterating this bound p times, we arrive at the result

$$\begin{aligned}\mathbb{E}\operatorname{tr}\boldsymbol{X}^{2p} &\le (2p-1)!!\cdot\left\|\textstyle\sum_{i=1}^{n}\boldsymbol{H}_i^2\right\|^p\cdot\operatorname{tr}\boldsymbol{X}^0\\ &= d\cdot(2p-1)!!\cdot\left\|\textstyle\sum_{i=1}^{n}\boldsymbol{H}_i^2\right\|^p.\end{aligned}\tag{4.9}$$

The double factorial is defined as $(2p-1)!! := (2p-1)(2p-3)(2p-5)\cdots(5)(3)(1)$.

The expression (4.4) shows that we can control the expected spectral norm of $\boldsymbol{X}$ by means of a trace moment. Therefore, for any nonnegative integer p, it holds that

$$\mathbb{E}\|\boldsymbol{X}\| \le \left(\mathbb{E}\operatorname{tr}\boldsymbol{X}^{2p}\right)^{1/(2p)} \le \left(d\cdot(2p-1)!!\right)^{1/(2p)}\cdot\left\|\textstyle\sum_{i=1}^{n}\boldsymbol{H}_i^2\right\|^{1/2}.\tag{4.10}$$

The second inequality is simply our bound (4.9). All that remains is to choose the value of p to minimize the factor on the right-hand side.

4.7. **Calculating the Constant.** Finally, let us develop an accurate bound for the leading factor on the right-hand side of (4.10). We claim that

$$(2p-1)!! \le \left(\frac{2p+1}{\mathrm{e}}\right)^p.\tag{4.11}$$

Given this estimate, select $p = \lceil\log d\rceil$ to reach

$$\left(d\cdot(2p-1)!!\right)^{1/(2p)} \le d^{1/(2p)}\sqrt{\frac{2p+1}{\mathrm{e}}} \le \sqrt{2p+1} = \sqrt{1+2\lceil\log d\rceil}.\tag{4.12}$$

Introduce the inequality (4.12) into (4.10) to complete the proof of Theorem 4.1.

To check that (4.11) is valid, we use some tools from integral calculus:

$$\begin{aligned}\log\left((2p-1)!!\right) &= \textstyle\sum_{i=1}^{p-1}\log(2i+1)\\ &= \left[\frac{1}{2}\log(2\cdot 0+1)+\textstyle\sum_{i=1}^{p-1}\log(2i+1)+\frac{1}{2}\log(2p+1)\right]-\frac{1}{2}\log(2p+1)\\ &\le \int_0^p\log(2x+1)\,\mathrm{d}x-\frac{1}{2}\log(2p+1)\\ &= p\log(2p+1)-p.\end{aligned}$$

The bracket in the second line is the trapezoid rule approximation of the integral in the third line. Since the integrand is concave, the trapezoid rule underestimates the integral. Exponentiating this formula, we arrive at (4.11).

4.8. **Context.** The proof of Theorem 4.1 is really just a discrete, matrix version of the familiar calculation of the $(2p)$th moment of a centered normal random variable. Let us elaborate. Recall the Gaussian integration by parts formula:

$$\mathbb{E}[\gamma \cdot f(\gamma)] = \sigma^2 \cdot \mathbb{E}[f'(\gamma)] \tag{4.13}$$

where $\gamma \sim \textsc{normal}(0, \sigma^2)$ and $f : \mathbb{R} \to \mathbb{R}$ is any function for which the integrals are finite. This result follows when we write the expectations as integrals with respect to the normal density tand invoke the usual integration by parts rule. Now, suppose that we wish to compute the $(2p)$th moment of γ. We have

$$\mathbb{E}\,\gamma^{2p} = \mathbb{E}\left[\gamma \cdot \gamma^{2p-1}\right] = (2p-1) \cdot \sigma^2 \cdot \mathbb{E}\,\gamma^{2p-2}. \tag{4.14}$$

The second identity is just (4.13) with the choice $f(t) = t^{2p-1}$. Iterating (4.14), we discover that

$$\mathbb{E}\,\gamma^{2p} = (2p-1)!! \cdot \sigma^{2p}.$$

In Theorem 4.1, the matrix variance parameter $v(\boldsymbol{X})$ plays the role of the scalar variance σ^2.

In fact, the link with Gaussian integration by parts is even stronger. Consider a matrix Gaussian series

$$\boldsymbol{Y} := \sum_{i=1}^n \gamma_i \boldsymbol{H}_i$$

where $\{\gamma_i\}$ is an independent family of standard normal variables. If we replace the discrete integration by parts in the proof of Theorem 4.1 with Gaussian integration by parts, the argument leads to the bound

$$\left(\mathbb{E}\left\|\sum_{i=1}^n \gamma_i \boldsymbol{H}_i\right\|^2\right)^{1/2} \leq \sqrt{1 + 2\lceil \log d \rceil} \cdot \left\|\sum_{i=1}^n \boldsymbol{H}_i^2\right\|^{1/2}.$$

This approach requires matrix calculus, but it is slightly simpler than the argument for matrix Rademacher series in other respects. See [Tro15b, Thm. 8.1] for a proof of the noncommutative Khintchine inequality for Gaussian series along these lines.

5. Upper Bounds for the Expected Norm

We are now prepared to establish the upper bound for an arbitrary sum of independent random matrices. The argument is based on the specialized result, Theorem 4.1, for matrix Rademacher series. It proceeds by steps through more and more general classes of random matrices: first positive semidefinite, then Hermitian, and finally rectangular. Here is what we will show.

Theorem 5.1 (Expected Norm: Upper Bounds)**.** *Define the dimensional constant* $C(d) := 4(1 + 2\lceil \log d \rceil)$. *The expected spectral norm of a sum of independent random matrices satisfies the following upper bounds.*

(1) ***The Positive-Semidefinite Case.*** *Consider an independent family* $\{\boldsymbol{T}_1, \dots, \boldsymbol{T}_n\}$ *of random* $d \times d$ *positive-semidefinite matrices, and define the sum*

$$\boldsymbol{W} := \sum_{i=1}^n \boldsymbol{T}_i.$$

Then

$$\mathbb{E}\,\|\boldsymbol{W}\| \leq \left[\|\mathbb{E}\,\boldsymbol{W}\|^{1/2} + \sqrt{C(d)} \cdot \left(\mathbb{E} \max_i \|\boldsymbol{T}_i\|\right)^{1/2}\right]^2. \tag{5.1}$$

(2) ***The Centered Hermitian Case.*** *Consider an independent family* $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\}$ *of random* $d \times d$ *Hermitian matrices with* $\mathbb{E}\,\boldsymbol{Y}_i = \boldsymbol{0}$ *for each index* i*, and define the sum*

$$\boldsymbol{X} := \sum_{i=1}^n \boldsymbol{Y}_i.$$

Then

$$\left(\mathbb{E}\,\|\boldsymbol{X}\|^2\right)^{1/2} \leq \sqrt{C(d)} \cdot \left\|\mathbb{E}\,\boldsymbol{X}^2\right\|^{1/2} + C(d) \cdot \left(\mathbb{E} \max_i \|\boldsymbol{Y}_i\|^2\right)^{1/2}. \tag{5.2}$$

(3) ***The Centered Rectangular Case.*** *Consider an independent family* $\{\boldsymbol{S}_1, \dots, \boldsymbol{S}_n\}$ *of random* $d_1 \times d_2$ *matrices with* $\mathbb{E}\,\boldsymbol{S}_i = \boldsymbol{0}$ *for each index* i*, and define the sum*

$$\boldsymbol{Z} := \textstyle\sum_{i=1}^n \boldsymbol{S}_i.$$

Then

$$\mathbb{E}\,\|\boldsymbol{Z}\| \leq \sqrt{C(d)} \cdot \max\left\{ \left\| \mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^*\right] \right\|^{1/2},\ \left\| \mathbb{E}\left[\boldsymbol{Z}^*\boldsymbol{Z}\right] \right\|^{1/2} \right\} + C(d) \cdot \left(\mathbb{E}\max_i \|\boldsymbol{S}_i\|^2 \right)^{1/2} \tag{5.3}$$

where $d := d_1 + d_2$.

The proof of Theorem 5.1 takes up the rest of this section. The presentation includes notes about the provenance of various parts of the argument.

The upper bound in Theorem I follows instantly from Case (3) of Theorem 5.1. We just introduce the notation $v(\boldsymbol{Z})$ for the variance parameter, and we calculate that

$$\mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^*\right] = \textstyle\sum_{i,j=1}^n \mathbb{E}\left[\boldsymbol{S}_i\boldsymbol{S}_j^*\right] = \sum_{i=1}^n \mathbb{E}\left[\boldsymbol{S}_i\boldsymbol{S}_i^*\right].$$

The first expression follows immediately from the definition of $\boldsymbol{Z}$ and the linearity of the expectation; the second identity holds because the random matrices $\boldsymbol{S}_i$ are independent and have mean zero. The formula for $\mathbb{E}\left[\boldsymbol{Z}^*\boldsymbol{Z}\right]$ is valid for precisely the same reasons.

5.1. **Proof of the Positive-Semidefinite Case.**

Recall that $\boldsymbol{W}$ is a random $d \times d$ positive-semidefinite matrix of the form

$$\boldsymbol{W} := \textstyle\sum_{i=1}^n \boldsymbol{T}_i \quad \text{where the } \boldsymbol{T}_i \text{ are positive semidefinite.}$$

Let us introduce notation for the quantity of interest:

$$E := \mathbb{E}\,\|\boldsymbol{W}\| = \mathbb{E}\,\left\|\textstyle\sum_{i=1}^n \boldsymbol{T}_i\right\|$$

By the triangle inequality for the spectral norm,

$$E \leq \left\|\textstyle\sum_{i=1}^n \mathbb{E}\,\boldsymbol{T}_i\right\| + \mathbb{E}\left\|\textstyle\sum_{i=1}^n (\boldsymbol{T}_i - \mathbb{E}\,\boldsymbol{T}_i)\right\| \leq \left\|\textstyle\sum_{i=1}^n \mathbb{E}\,\boldsymbol{T}_i\right\| + 2\,\mathbb{E}\left\|\textstyle\sum_{i=1}^n \varepsilon_i \boldsymbol{T}_i\right\|.$$

The second inequality follows from symmetrization, Fact 3.1. In this expression, $\{\varepsilon_i\}$ is an independent family of Rademacher random variables, independent from $\{\boldsymbol{T}_i\}$. Conditioning on the choice of the random matrices $\boldsymbol{T}_i$, we apply Theorem 4.1 via the bound (4.2):

$$\mathbb{E}\left\|\textstyle\sum_{i=1}^n \varepsilon_i \boldsymbol{T}_i\right\| = \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\varepsilon}}\left\|\textstyle\sum_{i=1}^n \varepsilon_i \boldsymbol{T}_i\right\|\right] \leq \sqrt{1 + 2\lceil\log d\rceil} \cdot \mathbb{E}\left[\left\|\textstyle\sum_{i=1}^n \boldsymbol{T}_i^2\right\|^{1/2}\right].$$

The operator $\mathbb{E}_{\boldsymbol{\varepsilon}}$ averages over the choice of the Rademacher random variables, with the matrices $\boldsymbol{T}_i$ fixed. Now, since the matrices $\boldsymbol{T}_i$ are positive-semidefinite,

$$\begin{aligned} \mathbb{E}\left[\left\|\textstyle\sum_{i=1}^n \boldsymbol{T}_i^2\right\|^{1/2}\right] &\leq \mathbb{E}\left[\left(\max_i \|\boldsymbol{T}_i\|\right)^{1/2} \cdot \left\|\textstyle\sum_{i=1}^n \boldsymbol{T}_i\right\|^{1/2}\right] \\ &\leq \left(\mathbb{E}\max_i \|\boldsymbol{T}_i\|\right)^{1/2} \cdot \left(\mathbb{E}\left\|\textstyle\sum_{i=1}^n \boldsymbol{T}_i\right\|\right)^{1/2} \\ &= \left(\mathbb{E}\max_i \|\boldsymbol{T}_i\|\right)^{1/2} \cdot E^{1/2}. \end{aligned}$$

The first inequality is Fact 2.3, and the second is the Cauchy–Schwarz inequality (3.1) for expectation. In the last step, we identified a copy of the quantity E.

Combine the last three displays to see that

$$E \leq \left\|\textstyle\sum_{i=1}^n \mathbb{E}\,\boldsymbol{T}_i\right\| + \sqrt{4(1 + 2\lceil\log d\rceil)} \cdot \left(\mathbb{E}\max_i \|\boldsymbol{T}_i\|\right)^{1/2} \cdot E^{1/2}. \tag{5.4}$$

For any $\alpha, \beta \geq 0$, the quadratic inequality $t^2 \leq \alpha + \beta t$ implies that

$$t \leq \frac{1}{2}\left[\beta + \sqrt{\beta^2 + 4\alpha}\right] \leq \frac{1}{2}\left[\beta + \beta + 2\sqrt{\alpha}\right] = \sqrt{\alpha} + \beta$$

because the square root is subadditive. Applying this fact to the quadratic relation (5.4) for $E^{1/2}$, we obtain

$$E^{1/2} \leq \left\|\textstyle\sum_{i=1}^n \mathbb{E}\,\boldsymbol{T}_i\right\|^{1/2} + \sqrt{4(1 + 2\lceil\log d\rceil)} \cdot \left(\mathbb{E}\max_i \|\boldsymbol{T}_i\|\right)^{1/2}.$$

Square both sides to reach the conclusion (5.1).

This argument is adapted from Rudelson's paper [Rud99], which develops a version of this result for the case where the matrices $\boldsymbol{T}_i$ have rank one; see also [RV07]. The paper [Tro08] contains the first estimates for the constants. Magen & Zouzias [MZ11] observed that similar considerations apply when the matrices $\boldsymbol{T}_i$ have higher rank. The complete result (5.1) first appeared in [CGT12, App.]. The constants in this paper are marginally better. Related bounds for Schatten norms appear in [MJC+14, Sec. 7] and in [JZ13].

The results described in the last paragraph are all matrix versions of the classical inequalities due to Rosenthal [Ros70b, Lem. 1]. These bounds can be interpreted as polynomial moment versions of the Chernoff inequality.

5.2. **Proof of the Hermitian Case.** The result (5.2) for Hermitian matrices is a corollary of Theorem 4.1 and the positive-semidefinite result (5.1). Recall that $\boldsymbol{X}$ is a $d \times d$ random Hermitian matrix of the form

$$\boldsymbol{X} := \sum_{i=1}^n \boldsymbol{Y}_i \quad \text{where } \mathbb{E}\,\boldsymbol{Y}_i = \boldsymbol{0}.$$

We may calculate that

$$\begin{aligned}\left(\mathbb{E}\,\|\boldsymbol{X}\|^2\right)^{1/2} &= \left(\mathbb{E}\,\left\|\sum_{i=1}^n \boldsymbol{Y}_i\right\|^2\right)^{1/2}\\ &\le 2\left(\mathbb{E}\left[\mathbb{E}_{\boldsymbol{\varepsilon}}\left\|\sum_{i=1}^n \varepsilon_i \boldsymbol{Y}_i\right\|^2\right]\right)^{1/2}\\ &\le \sqrt{4(1+2\lceil\log d\rceil)}\cdot\left(\mathbb{E}\left\|\sum_{i=1}^n \boldsymbol{Y}_i^2\right\|\right)^{1/2}.\end{aligned}$$

The first inequality follows from the symmetrization procedure, Fact 3.1. The second inequality applies Theorem 4.1, conditional on the choice of $\boldsymbol{Y}_i$. The remaining expectation contains a sum of independent positive-semidefinite matrices. Therefore, we may invoke (5.1) with $\boldsymbol{T}_i = \boldsymbol{Y}_i^2$. We obtain

$$\mathbb{E}\left\|\sum_{i=1}^n \boldsymbol{Y}_i^2\right\| \le \left[\left\|\sum_{i=1}^n \mathbb{E}\,\boldsymbol{Y}_i^2\right\|^{1/2} + \sqrt{4(1+2\lceil\log d\rceil)}\cdot\left(\mathbb{E}\max_i\|\boldsymbol{Y}_i^2\|\right)^{1/2}\right]^2.$$

Combine the last two displays to reach

$$\left(\mathbb{E}\,\|\boldsymbol{X}\|^2\right)^{1/2} \le \sqrt{4(1+2\lceil\log d\rceil)}\cdot\left[\left\|\sum_{i=1}^n \mathbb{E}\,\boldsymbol{Y}_i^2\right\|^{1/2} + \sqrt{4(1+2\lceil\log d\rceil)}\cdot\left(\mathbb{E}\max_i\|\boldsymbol{Y}_i\|^2\right)^{1/2}\right].$$

Rewrite this expression to reach (5.2).

A version of the result (5.2) first appeared in [CGT12, App.]; the constants here are marginally better. Related results for the Schatten norm appear in the papers [JX03, JX08, MJC+14, JZ13]. These bounds are matrix extensions of the scalar inequalities due to Rosenthal [Ros70b, Thm. 3] and to Rosén [Ros70a, Thm. 1]; see also Nagaev–Pinelis [NP77, Thm. 2]. They can be interpreted as the polynomial moment inequalities that sharpen the Bernstein inequality.

5.3. **Proof of the Rectangular Case.** Finally, we establish the rectangular result (5.3). Recall that $\boldsymbol{Z}$ is a $d_1 \times d_2$ random rectangular matrix of the form

$$\boldsymbol{Z} := \sum_{i=1}^n \boldsymbol{S}_i \quad \text{where } \mathbb{E}\,\boldsymbol{S}_i = \boldsymbol{0}.$$

Set $d := d_1 + d_2$, and form a random $d \times d$ Hermitian matrix $\boldsymbol{X}$ by dilating $\boldsymbol{Z}$:

$$\boldsymbol{X} := \mathscr{H}(\boldsymbol{Z}) = \sum_{i=1}^n \mathscr{H}(\boldsymbol{S}_i).$$

The Hermitian dilation $\mathscr{H}$ is defined in (2.18); the second relation holds because the dilation is a real-linear map.

Evidently, the random matrix $\boldsymbol{X}$ is a sum of independent, centered, random Hermitian matrices $\mathscr{H}(\boldsymbol{S}_i)$. Therefore, we may apply (5.2) to $\boldsymbol{X}$ to see that

$$\left(\mathbb{E}\,\|\mathscr{H}(\boldsymbol{Z})\|^2\right)^{1/2} \le \sqrt{4(1+2\lceil\log d\rceil)}\cdot\left\|\mathbb{E}\left[\mathscr{H}(\boldsymbol{Z})^2\right]\right\|^{1/2} + 4(1+2\lceil\log d\rceil)\cdot\left(\mathbb{E}\max_i\|\mathscr{H}(\boldsymbol{S}_i)\|^2\right)^{1/2}. \tag{5.5}$$

Since the dilation preserves norms (2.20), the left-hand side of (5.5) is exactly what we want:

$$\left(\mathbb{E}\,\|\mathscr{H}(\boldsymbol{Z})\|^2\right)^{1/2} = \left(\mathbb{E}\,\|\boldsymbol{Z}\|^2\right)^{1/2}.$$

To simplify the first term on the right-hand side of (5.5), invoke the formula (2.19) for the square of the dilation:

$$\left\|\mathbb{E}\left[\mathscr{H}(\boldsymbol{Z})^2\right]\right\| = \left\|\begin{bmatrix}\mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^*\right] & \mathbf{0} \\ \mathbf{0} & \mathbb{E}\left[\boldsymbol{Z}^*\boldsymbol{Z}\right]\end{bmatrix}\right\| = \max\left\{\left\|\mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^*\right]\right\|,\ \left\|\mathbb{E}\left[\boldsymbol{Z}^*\boldsymbol{Z}\right]\right\|\right\}.$$

The second identity relies on the fact that the norm of a block-diagonal matrix is the maximum norm of a diagonal block. To simplify the second term on the right-hand side of (5.5), we use (2.20) again:

$$\|\mathscr{H}(\boldsymbol{S}_i)\| = \|\boldsymbol{S}_i\|.$$

Introduce the last three displays into (5.5) to arrive at the result (5.3).

The result (5.3) first appeared in the monograph [Tro15a, Eqn. (6.16)] with (possibly) incorrect constants. The current paper contains the first complete presentation of the bound.

6. Lower Bounds for the Expected Norm

Finally, let us demonstrate that each of the upper bounds in Theorem 5.1 is sharp up to the dimensional constant $C(d)$. The following result gives matching lower bounds in each of the three cases.

Theorem 6.1 (Expected Norm: Lower Bounds)**.** *The expected spectral norm of a sum of independent random matrices satisfies the following lower bounds.*

(1) ***The Positive-Semidefinite Case.*** *Consider an independent family $\{\boldsymbol{T}_1,\dots,\boldsymbol{T}_n\}$ of random $d\times d$ positive-semidefinite matrices, and define the sum*

$$\boldsymbol{W} := \textstyle\sum_{i=1}^n \boldsymbol{T}_i.$$

Then

$$\mathbb{E}\|\boldsymbol{W}\| \ge \frac{1}{4}\left[\|\mathbb{E}\boldsymbol{W}\|^{1/2} + \left(\mathbb{E}\max_i\|\boldsymbol{T}_i\|\right)^{1/2}\right]^2. \tag{6.1}$$

(2) ***The Centered Hermitian Case.*** *Consider an independent family $\{\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n\}$ of random $d\times d$ Hermitian matrices with $\mathbb{E}\boldsymbol{Y}_i = \mathbf{0}$ for each index i, and define the sum*

$$\boldsymbol{X} := \textstyle\sum_{i=1}^n \boldsymbol{Y}_i.$$

Then

$$\left(\mathbb{E}\|\boldsymbol{X}\|^2\right)^{1/2} \ge \frac{1}{2}\left\|\mathbb{E}\boldsymbol{X}^2\right\|^{1/2} + \frac{1}{4}\left(\mathbb{E}\max_i\|\boldsymbol{Y}_i\|^2\right)^{1/2}. \tag{6.2}$$

(3) ***The Centered Rectangular Case.*** *Consider an independent family $\{\boldsymbol{S}_1,\dots,\boldsymbol{S}_n\}$ of random $d_1\times d_2$ matrices with $\mathbb{E}\boldsymbol{S}_i = \mathbf{0}$ for each index i, and define the sum*

$$\boldsymbol{Z} := \textstyle\sum_{i=1}^n \boldsymbol{S}_i.$$

Then

$$\mathbb{E}\|\boldsymbol{Z}\| \ge \frac{1}{2}\max\left\{\left\|\mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^*\right]\right\|^{1/2},\ \left\|\mathbb{E}\left[\boldsymbol{Z}^*\boldsymbol{Z}\right]\right\|^{1/2}\right\} + \frac{1}{4}\left(\mathbb{E}\max_i\|\boldsymbol{S}_i\|^2\right)^{1/2}. \tag{6.3}$$

The rest of the section describes the proof of Theorem 6.1.

The lower bound in Theorem I is an immediate consequence of Case (3) of Theorem 6.1. We simply introduce the notation $v(\boldsymbol{Z})$ for the variance parameter.

6.1. The Positive-Semidefinite Case.

The lower bound (6.1) in the positive-semidefinite case is relatively easy. Recall that

$$\boldsymbol{W} := \textstyle\sum_{i=1}^n \boldsymbol{T}_i \quad \text{where the } \boldsymbol{T}_i \text{ are positive semidefinite.}$$

First, by Jensen's inequality (3.3) and the convexity of the spectral norm,

$$\mathbb{E}\|\boldsymbol{W}\| \ge \|\mathbb{E}\boldsymbol{W}\|. \tag{6.4}$$

Second, let I be the minimum value of the index i where $\max_i\|\boldsymbol{T}_i\|$ is achieved; note that I is a random variable. Since the summands $\boldsymbol{T}_i$ are positive semidefinite, it is easy to see that

$$\boldsymbol{T}_I \preccurlyeq \textstyle\sum_{i=1}^n \boldsymbol{T}_i.$$

Therefore, by the norm identity (2.10) for a positive-semidefinite matrix and the monotonicity of the maximum eigenvalue, Fact 2.1, we have

$$\max_i \|\boldsymbol{T}_i\| = \|\boldsymbol{T}_I\| = \lambda_{\max}(\boldsymbol{T}_I) \le \lambda_{\max}\left(\sum_{i=1}^n \boldsymbol{T}_i\right) = \left\|\sum_{i=1}^n \boldsymbol{T}_i\right\| = \|\boldsymbol{W}\|.$$

Take the expectation to arrive at

$$\mathbb{E}\max_i \|\boldsymbol{T}_i\| \le \mathbb{E}\|\boldsymbol{W}\|. \tag{6.5}$$

Average the two bounds (6.4) and (6.5) to obtain

$$\mathbb{E}\|\boldsymbol{W}\| \ge \frac{1}{2}\left[\, \|\mathbb{E}\boldsymbol{W}\| + \mathbb{E}\max_i \|\boldsymbol{T}_i\| \,\right].$$

To reach (6.1), apply the numerical fact that $2(a+b) \ge \left(\sqrt{a}+\sqrt{b}\right)^2$, valid for all $a, b \ge 0$.

6.2. **Hermitian Case.** The Hermitian case (6.2) is similar in spirit, but the details are a little more involved. Recall that

$$\boldsymbol{X} := \sum_{i=1}^n \boldsymbol{Y}_i \quad \text{where } \mathbb{E}\,\boldsymbol{Y}_i = \boldsymbol{0}.$$

First, using the identity (2.11), we have

$$\left(\mathbb{E}\|\boldsymbol{X}\|^2\right)^{1/2} = \left(\mathbb{E}\|\boldsymbol{X}^2\|\right)^{1/2} \ge \left\|\mathbb{E}\boldsymbol{X}^2\right\|^{1/2}. \tag{6.6}$$

The second relation is Jensen's inequality (3.3).

To obtain the other part of our lower bound, we use the lower bound from the symmetrization result, Fact 3.1:

$$\mathbb{E}\|\boldsymbol{X}\|^2 = \mathbb{E}\left\|\sum_{i=1}^n \boldsymbol{Y}_i\right\|^2 \ge \frac{1}{4}\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i \boldsymbol{Y}_i\right\|^2$$

where $\{\varepsilon_i\}$ is an independent family of Rademacher random variables, independent from $\{\boldsymbol{Y}_i\}$. Now, we condition on the choice of $\{\boldsymbol{Y}_i\}$, and we compute the partial expectation with respect to the ε_i. Let I be the minimum value of the index i where $\max_i \|\boldsymbol{Y}_i\|^2$ is achieved. By Jensen's inequality (3.3), applied conditionally,

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left\|\sum_{i=1}^n \varepsilon_i \boldsymbol{Y}_i\right\|^2 \ge \mathbb{E}_{\varepsilon_I}\left\|\mathbb{E}\left[\sum_{i=1}^n \varepsilon_i \boldsymbol{Y}_i \,|\, \varepsilon_I\right]\right\|^2 = \mathbb{E}_{\varepsilon_I}\|\varepsilon_I \boldsymbol{Y}_I\|^2 = \max_i \|\boldsymbol{Y}_i\|^2.$$

Combining the last two displays and taking a square root, we discover that

$$\left(\mathbb{E}\|\boldsymbol{X}\|^2\right)^{1/2} \ge \frac{1}{2}\left(\mathbb{E}\max_i \|\boldsymbol{Y}_i\|^2\right)^{1/2}. \tag{6.7}$$

Average the two bounds (6.6) and (6.7) to conclude that (6.2) is valid.

6.3. **The Rectangular Case.** The rectangular case (6.3) follows instantly from the Hermitian case when we apply (6.2) to the Hermitian dilation. Recall that

$$\boldsymbol{Z} := \sum_{i=1}^n \boldsymbol{S}_i \quad \text{where } \mathbb{E}\,\boldsymbol{S}_i = \boldsymbol{0}.$$

Define a random matrix $\boldsymbol{X}$ by applying the Hermitian dilation (2.18) to $\boldsymbol{Z}$:

$$\boldsymbol{X} := \mathscr{H}(\boldsymbol{Z}) = \sum_{i=1}^n \mathscr{H}(\boldsymbol{S}_i).$$

Since $\boldsymbol{X}$ is a sum of independent, centered, random Hermitian matrices, the bound (6.2) yields

$$\left(\mathbb{E}\|\mathscr{H}(\boldsymbol{Z})\|^2\right)^{1/2} \ge \frac{1}{2}\left\|\mathbb{E}\left[\mathscr{H}(\boldsymbol{Z})^2\right]\right\| + \frac{1}{4}\left(\mathbb{E}\max_i \|\mathscr{H}(\boldsymbol{S}_i)\|^2\right)^{1/2}.$$

Repeating the calculations in Section 5.3, we arrive at the advertised result (6.3).

7. Optimality of Theorem I

The lower bounds and upper bounds in Theorem I match, except for the dimensional factor $C(\boldsymbol{d})$. In this section, we show by example that neither the lower bounds nor the upper bounds can be sharpened substantially. More precisely, the logarithms cannot appear in the lower bound, and they must appear in the upper bound. As a consequence, unless we make further assumptions, Theorem I cannot be improved except by constant factors and, in one place, by an iterated logarithm.

7.1. **Upper Bound: Variance Term.** First, let us show that the variance term in the upper bound in (1.5) must contain a logarithm. This example is drawn from [Tro15a, Sec. 6.1.2].

For a large parameter n, consider the $d \times d$ random matrix

$$\boldsymbol{Z} := \sum_{i=1}^{d} \sum_{j=1}^{n} \frac{1}{\sqrt{n}} \varepsilon_{ij} \mathbf{E}_{ii}$$

As before, $\{\varepsilon_{ij}\}$ is an independent family of Rademacher random variables, and $\mathbf{E}_{ii}$ is a $d \times d$ matrix with a one in the (i,i) position and zeroes elsewhere. The variance parameter satisfies

$$v(\boldsymbol{Z}) = \left\| \sum_{i=1}^{d} \sum_{j=1}^{n} \frac{1}{n} \mathbf{E}_{ii} \right\| = \|\mathbf{I}_d\| = 1.$$

The large deviation parameter satisfies

$$L^2 = \mathbb{E} \max_{i,j} \left\| \frac{1}{\sqrt{n}} \varepsilon_{ij} \mathbf{E}_{ii} \right\|^2 = \frac{1}{n}.$$

Therefore, the variance term drives the upper bound (1.5). For this example, it is easy to estimate the norm directly. Indeed,

$$\mathbb{E} \|\boldsymbol{Z}\|^2 \approx \mathbb{E} \left\| \sum_{i=1}^{d} \gamma_i \mathbf{E}_{ii} \right\|^2 = \mathbb{E} \max_{i=1,\dots,d} |\gamma_i|^2 \approx 2 \log d.$$

Here, $\{\gamma_i\}$ is an independent family of standard normal variables, and the first approximation follows from the central limit theorem. The norm of a diagonal matrix is the maximum absolute value of one of the diagonal entries. Last, we use the well-known fact that the expected maximum among d squared standard normal variables is asymptotic to $2 \log d$. In summary,

$$\left(\mathbb{E} \|\boldsymbol{Z}\|^2 \right)^{1/2} \approx \sqrt{2 \log d \cdot v(\boldsymbol{X})}.$$

We conclude that the variance term in the upper bound must carry a logarithm. Furthermore, it follows that Theorem 4.1 is numerically sharp.

7.2. **Upper Bound: Large-Deviation Term.** Next, we verify that the large-deviation term in the upper bound in (1.5) must also contain a logarithm, although the bound is slightly suboptimal. This example is drawn from [Tro15a, Sec. 6.1.2].

For a large parameter n, consider the $d \times d$ random matrix

$$\boldsymbol{Z} := \sum_{i=1}^{d} \sum_{j=1}^{n} \left(\delta_{ij} - n^{-1} \right) \cdot \mathbf{E}_{ii}$$

where $\{\delta_{ij}\}$ is an independent family of BERNOULLI(n^{-1}) random variables. That is, δ_{ij} takes only the values zero and one, and its expectation is n^{-1}. The variance parameter for the random matrix is

$$v(\boldsymbol{Z}) = \left\| \sum_{i=1}^{d} \sum_{j=1}^{n} \mathbb{E} \left(\delta_{ij} - n^{-1} \right)^2 \cdot \mathbf{E}_{ii} \right\| = \left\| \sum_{i=1}^{d} \sum_{j=1}^{n} n^{-1} \left(1 - n^{-1} \right) \cdot \mathbf{E}_{ii} \right\| \approx 1.$$

The large deviation parameter is

$$L^2 = \mathbb{E} \max_{i,j} \left\| \left(\delta_{ij} - n^{-1} \right) \cdot \mathbf{E}_{ii} \right\|^2 \approx 1.$$

Therefore, the large-deviation term drives the upper bound in (1.5):

$$\left(\mathbb{E} \|\boldsymbol{Z}\|^2 \right)^{1/2} \leq \sqrt{4(1 + 2\lceil \log d \rceil)} + 4(1 + 2\lceil \log d \rceil).$$

On the other hand, by direct calculation

$$\left(\mathbb{E} \|\boldsymbol{Z}\|^2 \right)^{1/2} \approx \left(\mathbb{E} \left\| \sum_{i=1}^{d} (Q_i - 1) \cdot \mathbf{E}_{ii} \right\|^2 \right)^{1/2} = \left(\mathbb{E} \max_{i=1,\dots,d} |Q_i - 1|^2 \right)^{1/2} \approx \text{const} \cdot \frac{\log d}{\log \log d}.$$

Here, $\{Q_i\}$ is an independent family of POISSON(1) random variables, and the first approximation follows from the Poisson limit of a binomial. The second approximation depends on a (messy) calculation for the expected squared maximum of a family of independent Poisson variables. We see that the large deviation term in the upper bound (1.5) cannot be improved, except by an iterated logarithm factor.

7.3. **Lower Bound: Variance Term.** Next, we argue that there are examples where the variance term in the lower bound from (1.5) cannot have a logarithmic factor.

Consider a $d \times d$ random matrix of the form

$$\boldsymbol{Z} := \textstyle\sum_{i,j=1}^{d} \varepsilon_{ij} \mathbf{E}_{ij}.$$

Here, $\{\varepsilon_{ij}\}$ is an independent family of Rademacher random variables. The variance parameter satisfies

$$v(\boldsymbol{Z}) = \max\left\{ \left\| \textstyle\sum_{i,j=1}^{d} \left(\mathbb{E}\,\varepsilon_{ij}^2\right) \cdot \mathbf{E}_{ij}\mathbf{E}_{ij}^* \right\|, \left\| \textstyle\sum_{i,j=1}^{d} \left(\mathbb{E}\,\varepsilon_{ij}^2\right) \cdot \mathbf{E}_{ij}^*\mathbf{E}_{ij} \right\| \right\} = \max\{\|d \cdot \mathbf{I}_d\|, \|d \cdot \mathbf{I}_d\|\} = d.$$

The large-deviation parameter is

$$L^2 = \mathbb{E}\max_{i,j} \left\| \varepsilon_{ij} \mathbf{E}_{ij} \right\|^2 = 1.$$

Therefore, the variance term controls the lower bound in (1.5):

$$\left(\mathbb{E}\,\|\boldsymbol{Z}\|^2\right)^{1/2} \geq \sqrt{cd} + c.$$

Meanwhile, it can be shown that the norm of the random matrix $\boldsymbol{Z}$ satisfies

$$\left(\mathbb{E}\,\|\boldsymbol{Z}\|^2\right)^{1/2} \approx \sqrt{2d}.$$

See the paper [BH14] for an elegant proof of this nontrivial result. We see that the variance term in the lower bound in (1.5) cannot have a logarithmic factor.

7.4. **Lower Bound: Large-Deviation Term.** Finally, we produce an example where the large-deviation term in the lower bound from (1.5) cannot have a logarithmic factor.

Consider a $d \times d$ random matrix of the form

$$\boldsymbol{Z} := \textstyle\sum_{i=1}^{d} P_i \mathbf{E}_{ii}.$$

Here, $\{P_i\}$ is an independent family of symmetric random variables whose tails satisfy

$$\mathbb{P}\,\{|P_i| \geq t\} = \begin{cases} t^{-4}, & t \geq 1 \\ 1, & t \leq 1. \end{cases}$$

The key properties of these variables are that

$$\mathbb{E}\,P_i^2 = 2 \quad \text{and} \quad \mathbb{E}\max_{i=1,\dots,d} P_i^2 \approx \text{const} \cdot d^2.$$

The second expression just describes the asymptotic order of the expected maximum. We quickly compute that the variance term satisfies

$$v(\boldsymbol{Z}) = \left\| \textstyle\sum_{i=1}^{d} \left(\mathbb{E}\,P_i^2\right) \mathbf{E}_{ii} \right\| = 2.$$

Meanwhile, the large-deviation factor satisfies

$$L^2 = \mathbb{E}\max_{i=1,\dots,d} \|P_i \mathbf{E}_{ii}\|^2 = \mathbb{E}\max_{i=1,\dots,d} |P_i|^2 \approx \text{const} \cdot d^2.$$

Therefore, the large-deviation term drives the lower bound (1.5):

$$\left(\mathbb{E}\,\|\boldsymbol{Z}\|^2\right)^{1/2} \gtrapprox \text{const} \cdot d.$$

On the other hand, by direct calculation,

$$\left(\mathbb{E}\,\|\boldsymbol{Z}\|^2\right)^{1/2} = \left(\mathbb{E}\left\| \textstyle\sum_{i=1}^{d} P_i \mathbf{E}_{ii} \right\|^2\right)^{1/2} = \left(\mathbb{E}\max_{i=1,\dots,d} |P_i|^2\right)^{1/2} \approx \text{const} \cdot d.$$

We conclude that the large-deviation term in the lower bound (1.5) cannot carry a logarithmic factor.

Acknowledgments

The author wishes to thank Ryan Lee for a careful reading of the manuscript. The author gratefully acknowledges support from ONR award N00014-11-1002 and the Gordon & Betty Moore Foundation.

References

[Bar02] A. Barvinok. *A course in convexity*, volume 54 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.

[BG13] R. F. Bass and K. Gröchenig. Relevant sampling of band-limited functions. *Illinois J. Math.*, 57(1):43–58, 2013.

[BH14] A. Bandeira and R. V. Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. Available at `http://arXiv.org/abs/1408.6185`, Aug. 2014.

[Bha97] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

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

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[CBSW14] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Coherent matrix completion. In *Proc. 31st Intl. Conf. Machine Learning*, Beijing, 2014.

[CC13] X. Chen and T. M. Christensen. Optimal uniform convergence rates for sieve nonparametric instrumental variables regression. Available at `http://arXiv.org/abs/1311.0412`, Nov. 2013.

[CDL13] A. Cohen, M. A. Davenport, and D. Leviatan. On the stability and accuracy of least squares approximations. *Found. Comput. Math.*, 13(5):819–834, 2013.

[CG14] P. Constantine and D. Gleich. Computing active subspaces. Available at `http://arXiv.org/abs/1408.0545`, Aug. 2014.

[CGH14] Y. Chen, L. Guibas, and Q. Huang. Near-optimal joint object matching via convex relaxation. In *Proc. 31st Intl. Conf. Machine Learning*, Beijing, 2014.

[CGT12] R. Y. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Inf. Inference*, 1(1):2–20, 2012.

[CKM$^+$14] M. B. Cohen, R. Kyng, G. L. Miller, J. W. Pachocki, R. Peng, A. B. Rao, and S. C. Xu. Solving SDD linear systems in nearly $m\log^{1/2}(n)$ time. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 343–352, New York, NY, USA, 2014. ACM.

[CSW12] S.-S. Cheung, A. M.-C. So, and K. Wang. Linear matrix inequalities with stochastically dependent perturbations and applications to chance-constrained semidefinite optimization. *SIAM J. Optim.*, 22(4):1394–1430, 2012.

[DKC13] J. Djolonga, A. Krause, and V. Cevher. High-dimensional Gaussian process bandits. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1025–1033. Curran Associates, Inc., 2013.

[FSV12] M. Fornasier, K. Schnass, and J. Vybiral. Learning functions of few arbitrary linear parameters in high dimensions. *Found. Comput. Math.*, 12(2):229–262, 2012.

[GS01] G. R. Grimmett and D. R. Stirzaker. *Probability and random processes*. Oxford University Press, New York, third edition, 2001.

[Hal74] P. R. Halmos. *Finite-dimensional vector spaces*. Springer-Verlag, New York-Heidelberg, second edition, 1974. Undergraduate Texts in Mathematics.

[HJ13] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[HO14] N. J. A. Harvey and N. Olver. Pipage rounding, pessimistic estimators and matrix concentration. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 926–945. SIAM, 2014.

[Hol12] A. S. Holevo. *Quantum systems, channels, information*, volume 16 of *De Gruyter Studies in Mathematical Physics*. De Gruyter, Berlin, 2012. A mathematical introduction.

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

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

[JZ13] M. Junge and Q. Zeng. Noncommutative Bennett and Rosenthal inequalities. *Ann. Probab.*, 41(6):4287–4316, 2013.

[Kol11] V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

[LO94] R. Latała and K. Oleszkiewicz. On the best constant in the Khinchin-Kahane inequality. *Studia Math.*, 109(1):101–104, 1994.

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

[LPSS$^+$14] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schölkopf. Randomized nonlinear component analysis. In *Proc. 31st Intl. Conf. Machine Learning*, Beijing, July 2014.

[LT11] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.

[Lue69] D. G. Luenberger. *Optimization by vector space methods*. John Wiley & Sons, Inc., New York-London-Sydney, 1969.

[MB14] W. B. March and G. Biros. Far-field compression for fast kernel summation methods in high dimensions. Available at http://arXiv.org/abs/1409.2802, Sep. 2014.

[MJC$^+$14] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *Ann. Probab.*, 42(3):906–945, 2014.

[MKR12] E. Morvant, S. Koço, and L. Ralaivola. PAC-Bayesian generalization bound on confusion matrix for multi-class classification. In *Proc. 29th Intl. Conf. Machine Learning*, Edinburgh, 2012.

[MZ11] A. Magen and A. Zouzias. Low rank matrix-valued Chernoff bounds and approximate matrix multiplication. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1422–1436. SIAM, Philadelphia, PA, 2011.

[NP77] S. V. Nagaev and I. F. Pinelis. Some inequalities for the distributions of sums of independent random variables. *Teor. Verojatnost. i Primenen.*, 22(2):254–263, 1977.

[Oli10] R. I. Oliveira. The spectrum of random k-lifts of large graphs (with possibly large k). *J. Comb.*, 1(3-4):285–306, 2010.

[Oli13] R. I. Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. Available at http://arXiv.org/abs/1312.2903, Dec. 2013.

[Pis98] G. Pisier. Non-commutative vector valued L_p-spaces and completely p-summing maps. *Astérisque*, (247):vi+131, 1998.

[Roc97] R. T. Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.

[Ros70a] B. Rosén. On bounds on the central moments of even order of a sum of independent random variables. *Ann. Math. Statist.*, 41:1074–1077, 1970.

[Ros70b] H. P. Rosenthal. On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 8:273–303, 1970.

[Rud99] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, 1999.

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

[Tao12] T. Tao. *Topics in random matrix theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

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

[Tro11] J. A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Adv. Adapt. Data Anal.*, 3(1-2):115–126, 2011.

[Tro15a] J. A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learning*, 8(1–2), May 2015.

[Tro15b] J. A. Tropp. Second-order matrix concentration inequalities. Available at http://arXiv.org/abs/1504.05919, Apr. 2015.