COMMUNICATIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 6, Pages 844–895 (July 7, 2026)
https://doi.org/10.1090/cams/72

COMPARISON THEOREMS FOR THE MINIMUM EIGENVALUE OF A RANDOM POSITIVE-SEMIDEFINITE MATRIX

JOEL A. TROPP

ABSTRACT. This paper establishes a new comparison principle for the minimum eigenvalue of a sum of independent random positive-semidefinite matrices. The principle states that the minimum eigenvalue of the matrix sum is controlled by the minimum eigenvalue of a Gaussian random matrix whose statistics are determined by the summands. This methodology is powerful because of the vast arsenal of tools for treating Gaussian random matrices. As applications, the paper presents short, conceptual proofs of some old and new results in high-dimensional statistics. It also settles a longstanding open question in computational linear algebra about the injectivity properties of very sparse random matrices.

1. MOTIVATION

Random positive-semidefinite (*psd*) matrices appear throughout high-dimensional statistics and high-dimensional probability. In particular, random psd matrices model the sample covariance of a random vector, and they capture properties of random linear embeddings. For a psd matrix, the minimum eigenvalue provides a quantitative measure of invertibility, so it is often the crucial statistic of these random matrix models. This paper introduces a new technique for studying the minimum eigenvalue of a random psd matrix by establishing a comparison with the minimum eigenvalue of a Gaussian random matrix. It also showcases several applications of this methodology.

1.1. Background and context.

A *random matrix* is a matrix whose entries are random variables, not necessarily independent from each other. The distinctive concern of *random matrix theory* (*RMT*) is the action of the random matrix as a linear map. Some basic questions involve the localization of eigenvalues, the distribution of eigenvalues, and the structure of eigenvectors. One may also study embedding properties—how the random matrix distorts the geometry of a set, such as a linear subspace.

RMT emerged from applications in statistics (Wishart, 1927) and in nuclear physics (Wigner, 1955). The early research centered on random matrix models with special distributions (e.g., Gaussian) or with highly symmetric distributions (e.g., independent entries). Researchers have attained exquisite insight on these models because there are many ways to exploit the mathematical structure. See the textbook [Tao12] for an introduction to RMT fundamentals and recent trends.

Received by the editors February 2, 2025, and, in revised form, January 25, 2026, and March 17, 2026.
2020 *Mathematics Subject Classification.* Primary 15B52, 60B20.
Key words and phrases. Comparison theorem, high-dimensional probability, high-dimensional statistics, random matrix.
This research was supported in part by NSF FRG Award 1952777, ONR Award N-00014-24-1-2223, and the Caltech Carver Mead New Adventures Fund.

As applications of RMT have multiplied, attention has shifted from particular examples to more versatile templates. A key contemporary model is the sum of independent random self-adjoint matrices:

$$\mathbf{Y} = \sum_{i=1}^{n} \mathbf{W}_i \quad \text{where each } \mathbf{W}_i = \mathbf{W}_i^*, \text{ and } (\mathbf{W}_1, \dots, \mathbf{W}_n) \text{ is independent.}$$

This model encompasses Wishart and Wigner matrices, as well as a menagerie of other random matrices that arise in practice. Using *matrix concentration inequalities*, we can always bound the extreme eigenvalues of an independent sum in terms of accessible statistics of the summands. These results have made a meteoric impact because they reduce (formerly) challenging RMT problems to simple linear algebra calculations. The monograph [Tro15] surveys matrix concentration and its applications.

Nevertheless, the basic matrix concentration inequalities often exhibit a suboptimal dependence on the matrix dimension, so researchers have continued to search for general-purpose tools that can provide more precise information [Tro18, BBvH23, Ban+24, BvH24]. This body of work revisits the classic idea that we can analyze a probability model by relating it to a reference model with additional structure. In particular, it is fruitful to compare a random matrix to a Gaussian random matrix that we already understand.

This paper develops a new implementation of the comparison strategy: the minimum eigenvalue of a sum of independent random psd matrices is controlled by the minimum eigenvalue of a Gaussian random matrix with the same expectation and with slightly higher variance. This comparison has implications for discrete geometry, high-dimensional statistics, and computational linear algebra.

1.2. **Intuition: Positive random walks.** Our approach is inspired by some surprising properties of a positive random walk. Consider a random walk on the real line that can only move in the positive direction. What is the probability that the random walk remains close to its origin? This event occurs only when all of the increments are small, which is very unlikely. We can capture this insight with a standard probability inequality that is the starting point for our investigation.

To model the positive random walk, we introduce a sum of nonnegative real random variables that are independent and identically distributed (*iid*):

$$Y = \sum_{i=1}^{n} W_i \quad \text{where the } W_i \text{ are iid copies of } W \geq 0.$$

When the increment W has two moments, we can compare the moment generating function (*mgf*) for its lower tail with the mgf of a real-valued Gaussian random variable. For all $\theta \geq 0$,

$$\begin{aligned} \mathbb{E}[\mathrm{e}^{-\theta W}] &\leq \mathbb{E}[1 - \theta W + \tfrac{1}{2}\theta^2 W^2] = 1 - \theta\,\mathbb{E}[W] + \tfrac{1}{2}\theta^2\,\mathbb{E}[W^2] \\ &\leq \exp\bigl(-\theta\,\mathbb{E}[W] + \tfrac{1}{2}\theta^2\,\mathbb{E}[W^2]\bigr) \\ &= \mathbb{E}[\mathrm{e}^{-\theta X}] \quad \text{where } X \sim \textsc{normal}_{\mathbb{R}}(\mathbb{E}[W], \mathbb{E}[W^2]). \end{aligned} \tag{1.1}$$

Using independence, this calculation extends to the mgf of the lower tail of the sum Y:

$$\mathbb{E}[\mathrm{e}^{-\theta Y}] \leq \mathbb{E}[\mathrm{e}^{-\theta Z}] \quad \text{where} \quad Z \sim \textsc{normal}_{\mathbb{R}}\bigl(n \cdot \mathbb{E}[W],\ n \cdot \mathbb{E}[W^2]\bigr). \tag{1.2}$$

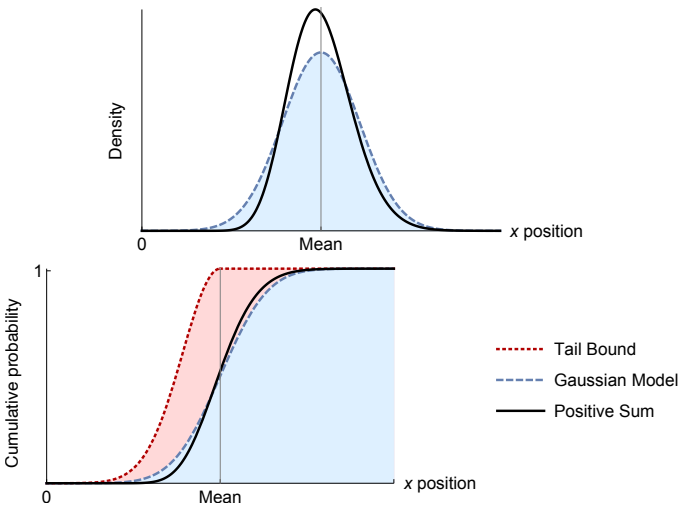

FIGURE 1.1. (Positive sum: Comparison). These plots illustrate the distribution Y of an iid sum of nonnegative real random variables (black line), along with the matching Gaussian distribution Z (dashed blue), described by (1.2), and the tail bound (dotted red), described by (1.3). The top panel shows the densities; the bottom panel shows the cumulative distributions. Each summand W is a chi-squared variable with 1 degree of freedom, and Y comprises $n = 32$ summands.

The mgf bound for the sum leads to a classic inequality for its lower tail:

$$\mathbb{P}\{Y \le \mathbb{E}[Z] - t\} \le \mathrm{e}^{-t^2/(2\operatorname{Var}[Z])} \quad \text{for } t \in [0, 1]. \tag{1.3}$$

In other words, the lower tail of the positive sum Y is related to the lower tail of a matching Gaussian random variable Z whose statistics depend on the summand W. See Figure 1.1 for an illustration.

1.3. **Random psd matrices.** This paper demonstrates that the same phenomena persist in the matrix setting. Consider a sum of iid random psd matrices, either real or complex:

$$\mathbf{Y} = \sum_{i=1}^{n} \mathbf{W}_i \quad \text{where the } \mathbf{W}_i \text{ are iid copies of a random psd matrix } \mathbf{W}. \tag{1.4}$$

The minimum eigenvalue $\lambda_{\min}(\mathbf{Y})$ can be expressed as the minimum of a family of iid positive sums:

$$\lambda_{\min}(\mathbf{Y}) = \min_{\|\mathbf{u}\|=1} \sum_{i=1}^{n} \mathbf{u}^* \mathbf{W}_i \mathbf{u}. \tag{1.5}$$

For each direction $\mathbf{u}$, the sum in (1.5) is very unlikely to be zero because of (1.3). On the other hand, the random summands $\mathbf{W}_i$ must cover every direction $\mathbf{u}$ before the minimum eigenvalue $\lambda_{\min}(\mathbf{Y})$ is strictly positive. It is not clear which of these two opposing principles prevails.

To resolve this dilemma, we adapt the strategy behind the scalar inequality (1.3) to the matrix setting. Construct a self-adjoint Gaussian random matrix $\mathbf{Z}$ that inherits its statistics from the summands:

$$\begin{aligned} \mathbb{E}[\mathbf{Z}] &= n \cdot \mathbb{E}[\mathbf{W}]; \\ \operatorname{Var}[\operatorname{Tr}[\mathbf{MZ}]] &= n \cdot \mathbb{E}[(\operatorname{Tr}[\mathbf{MW}])^2] \quad \text{for all self-adjoint } \mathbf{M}. \end{aligned} \tag{1.6}$$

Inspired by (1.2), we will establish a comparison between the trace mgfs of the two random matrices:

$$\mathbb{E}[\operatorname{Tr} \mathrm{e}^{-\theta \mathbf{Y}}] \le 2\, \mathbb{E}[\operatorname{Tr} \mathrm{e}^{-\theta \mathbf{Z}}] \quad \text{for all } \theta \ge 0. \tag{1.7}$$

Figure 1.2 illustrates the comparison. While the scalar case (1.2) is easy, the matrix inequality (1.7) relies on a deep fact from matrix analysis called Stahl's theorem, formerly the BMV conjecture [Sta13]. This is a novel approach for studying random matrices.

Using standard methods from high-dimensional probability [Tro15, vH16], the trace inequality (1.7) leads to a probabilistic comparison for the minimum eigenvalues:

$$\mathbb{P}\{\lambda_{\min}(\mathbf{Y}) \ge \mathbb{E}[\lambda_{\min}(\mathbf{Z})] - t\} \le 2d \cdot \mathrm{e}^{-t^2/(2\sigma_*^2(\mathbf{Z}))}, \tag{1.8}$$

where d is the matrix dimension. The weak variance $\sigma_*^2(\mathbf{Z}) := \max_{\|\mathbf{u}\|=1} \operatorname{Var}[\mathbf{u}^*\mathbf{Z}\mathbf{u}]$ controls the variance of $\lambda_{\min}(\mathbf{Z})$. The result (1.8) is powerful enough to yield sharp bounds for the minimum eigenvalue of some models. See Theorem 2.4 for the full statement of the comparison.

To analyze the minimum eigenvalue $\lambda_{\min}(\mathbf{Z})$ of the Gaussian matrix in (1.8), we have a bristling armamentarium of techniques at our disposal. This methodology leads to short, conceptual proofs of several important results from high-dimensional statistics (Section 5). It also addresses a vexing open question from computational linear algebra about very sparse random matrices (Section 6).

1.4. **Roadmap.** Section 2 states two comparison theorems and discusses related work. Section 3 provides background on Gaussian random matrices that aids in the analysis of the comparison model. Sections 4 to 6 apply the main results to several examples. Section 7 details a proof of (1.3) that generalizes to matrices. Sections 8 and 9 establish the main results, including (1.7) and (1.8).

1.5. **Notation.** We use standard conventions from linear algebra and probability. Nonlinear functions bind before the trace and the expectation. We may omit the brackets enclosing an argument.

Our results hold in both the real ($\mathbb{F} = \mathbb{R}$) and complex ($\mathbb{F} = \mathbb{C}$) setting; we often suppress the field. For a natural number $d \in \mathbb{N}$, the set $\mathbb{M}_d(\mathbb{F})$ is the linear space of $d \times d$ square matrices with entries in the field $\mathbb{F}$. The set $\mathbb{H}_d(\mathbb{F})$ denotes the real-linear space of *self-adjoint* (i.e., conjugate symmetric) matrices in $\mathbb{M}_d(\mathbb{F})$. As usual, $\operatorname{Tr}[\cdot]$ returns the (unnormalized) trace of a square matrix.

The symbol $^\intercal$ denotes the transpose of a vector or matrix, while * denotes the conjugate transpose. The ℓ_2 inner product is conjugate linear in the first coordinate: $\langle \mathbf{u}, \mathbf{v} \rangle := \mathbf{u}^*\mathbf{v}$ for $\mathbf{u}, \mathbf{v} \in \mathbb{F}^d$. Likewise, we employ the trace inner product $\langle \mathbf{M}, \mathbf{A} \rangle := \operatorname{Tr}[\mathbf{M}^*\mathbf{A}]$ for matrices $\mathbf{M}, \mathbf{A}$ with the same dimensions.

The symbol $\|\cdot\|$ denotes the ℓ_2 norm on vectors or the associated ℓ_2 operator norm on matrices. The Frobenius norm $\|\cdot\|_{\mathrm{F}}$ also appears regularly.

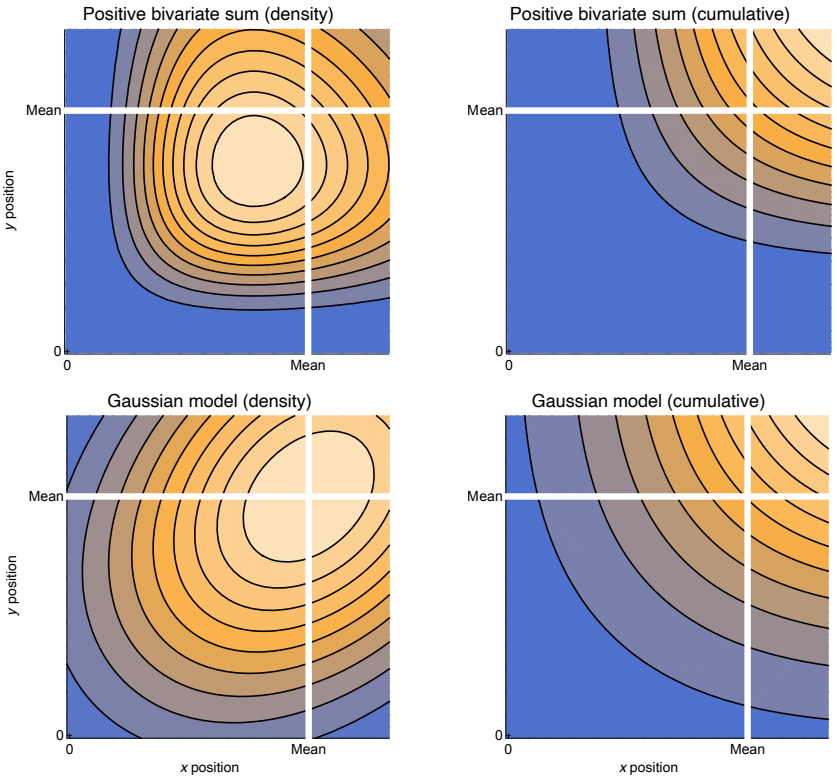

FIGURE 1.2. (Bivariate positive sum: Comparison). This figure illustrates the Gaussian comparison (1.6) for a 2×2 diagonal random matrix $\mathbf{W}$ with independent entries. Each diagonal entry W_{ii} is a chi-squared variable with 1 degree of freedom, and there are $n = 32$ summands. **[Top]** The bivariate distribution of a pair of independent positive sums. **[Bottom]** The distribution of the matching Gaussian model. **[Left]** Probability densities. **[Right]** Bivariate cumulative distribution functions. Brighter colors correspond to higher probabilities. The Gaussian comparison is valid southwest of the expectation (white lines). More precisely, the comparison concerns the minimum of the two coordinates, whose distribution looks similar to the univariate case (Figure 1.1).

As usual, $\mathbb{P}(\cdot)$ returns the probability of an event, and $\mathbb{E}[\cdot]$ computes the expectation of a random variable taking values in a finite-dimensional linear space. The operator $\mathrm{Var}[\cdot]$ reports the variance of a real random variable. The symbol $\sim$ means "has the distribution". The symbol $\asymp$ means equivalent within constant factors. In informal discussions, we may employ the orders $\lesssim$ and $\gtrsim$, which suppress universal constants. The symbol $\ll$ means "much less than".

2. Main results and related work

This section states our two main comparison theorems. The first concerns the sum of psd matrices with random weights. The second theorem concerns the sum of iid random psd matrices that was described in Section 1.3. Afterward, we present a simple first example, and we discuss related work.

2.1. Gaussian comparison: Randomly weighted sum of psd matrices.

The first result treats a random matrix model where we randomly weight the terms in a sum of fixed psd matrices.

Theorem 2.1 (Comparison: Randomly weighted psd matrices). *Fix a system of psd matrices $(\mathbf{A}_1, \dots, \mathbf{A}_n)$, real or complex, with common dimension d. Consider an independent family $(W_1, \dots, W_n)$ of nonnegative real random variables with two finite moments: $W_i \geq 0$ and $\mathbb{E}[W_i^2] < +\infty$. Define the random matrices*

$$\begin{aligned} \mathbf{Y} &= \sum_{i=1}^{n} W_i \mathbf{A}_i; \\ \mathbf{Z} &= \sum_{i=1}^{n} X_i \mathbf{A}_i \qquad \text{where } X_i \sim \textsc{normal}_{\mathbb{R}}(\mathbb{E}[W_i],\ \mathbb{E}[W_i^2]) \text{ are independent.} \end{aligned} \tag{2.1}$$

Then there is a stochastic comparison between the minimum eigenvalues of $\mathbf{Y}$ and $\mathbf{Z}$. In expectation,

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Y}) \geq \mathbb{E}\,\lambda_{\min}(\mathbf{Z}) - \sqrt{2\sigma_*^2(\mathbf{Z}) \log d}.$$

In addition, for all $t \geq 0$, the lower tail satisfies

$$\mathbb{P}\{\lambda_{\min}(\mathbf{Y}) \leq \mathbb{E}\,\lambda_{\min}(\mathbf{Z}) - t\} \leq d \cdot \mathrm{e}^{-t^2/(2\sigma_*^2(\mathbf{Z}))}.$$

For this model, the weak variance takes the form

$$\sigma_*^2(\mathbf{Z}) = \max_{\|\mathbf{u}\|=1} \sum_{i=1}^{n} \mathbb{E}[W_i^2] \cdot (\mathbf{u}^* \mathbf{A}_i \mathbf{u})^2.$$

The proof of Theorem 2.1 appears in Section 8. The short argument employs tools from Stein's method [Ros11, CGS11]. The critical ingredient is Stahl's theorem (Fact 8.2), a deep result on the structure of the trace exponential function [Sta13]. The weak variance $\sigma_*^2(\mathbf{Z})$ arises from Gaussian concentration [BLM13, Thm. 5.5] for the minimum eigenvalue; see Fact 3.4.

A valuable feature of Theorem 2.1 is that the psd coefficient matrices $\mathbf{A}_i$ are arbitrary and the random weights W_i can have different distributions. Of particular interest is the case where $W_i \sim \textsc{bernoulli}(p_i)$, which models a random sample from a family of fixed psd matrices.

Section 4 applies Theorem 2.1 to the geometric problem of sampling from a complex projective design. This result complements Rudelson's theorem [Rud99] on sampling from an isotropic system.

Remark 2.2 (Second moment versus variance). Compare the construction of the Gaussian comparison model (2.1) with the scalar case (1.2). It may be surprising that the variance of the Gaussian X_i depends on the raw second moment of W_i, but this discrepancy already arises in the scalar calculation (1.1) because of the Taylor expansion

at zero. In both the scalar and matrix setting, the probability bounds are most accurate when the random variables W_i place a lot of mass near zero. As a consequence, Theorem 2.1 is an effective tool for studying sparsely sampled sums of psd matrices.

2.2. **Second moments and Gaussians.** To state our next result compactly, we introduce notation that describes the second moments of a random matrix model.

Definition 2.3 (Random matrix: Second moments, self-adjoint case). Let $\mathbf{X} \in \mathbb{H}_d$ be a random self-adjoint matrix. Define the *second moment function* and the *variance function* of the random matrix:

$$\begin{aligned} \mathsf{Mom}[\mathbf{X}](\mathbf{M}) &:= \mathbb{E}\,|\langle \mathbf{M},\, \mathbf{X}\rangle|^2 = \mathbb{E}[(\mathrm{Tr}[\mathbf{MX}])^2], \\ \mathsf{Var}[\mathbf{X}](\mathbf{M}) &:= \mathrm{Var}[\langle \mathbf{M},\, \mathbf{X}\rangle] = \mathrm{Var}[\mathrm{Tr}[\mathbf{MX}]], \end{aligned} \qquad \text{for self-adjoint } \mathbf{M} \in \mathbb{H}_d.$$

These functions pack up the second-order statistics of the one-dimensional linear marginals of the random matrix.

Recall that a (real or complex) random matrix is *Gaussian* if and only if the real and imaginary parts of the entries compose a jointly Gaussian family of real random variables; e.g., see [BvH24, Sec. 2.1.2]. A self-adjoint Gaussian matrix $\mathbf{Z} \in \mathbb{H}_d$ is uniquely determined by its expectation $\mathbb{E}[\mathbf{Z}]$ and variance function $\mathsf{Var}[\mathbf{Z}]$. Given a self-adjoint matrix $\boldsymbol{\Delta} \in \mathbb{H}_d$ and a positive quadratic form $\mathsf{V} : \mathbb{H}_d \to \mathbb{R}_+$, we write $\textsc{normal}(\boldsymbol{\Delta}, \mathsf{V})$ for the unique Gaussian distribution on self-adjoint matrices with expectation $\boldsymbol{\Delta}$ and variance function V. For more background on Gaussian random matrices, see Section 3.

2.3. **Gaussian comparison: Sum of iid psd random matrices.** Our second result provides a Gaussian comparison for a sum of iid random psd matrices, the model presented in the introduction.

Theorem 2.4 (Comparison: Sum of iid psd matrices). *Let $\mathbf{W}$ be a random psd matrix, real or complex, with dimension d and with two finite moments: $\mathbb{E}\,\|\mathbf{W}\|^2 < +\infty$. Consider any self-adjoint Gaussian matrix $\mathbf{X}$ with dimension d whose first- and second-order statistics satisfy*

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{W}] \quad \text{and} \quad \mathsf{Var}[\mathbf{X}] \geq \mathsf{Mom}[\mathbf{W}]. \tag{2.2}$$

For a natural number $n \in \mathbb{N}$, define the random matrices

$$\begin{aligned} \mathbf{Y} &= \sum_{i=1}^{n} \mathbf{W}_i \quad \text{where } \mathbf{W}_i \sim \mathbf{W} \text{ iid;} \\ \mathbf{Z} &= \sum_{i=1}^{n} \mathbf{X}_i \quad \text{where } \mathbf{X}_i \sim \mathbf{X} \text{ iid.} \end{aligned} \tag{2.3}$$

Then there is a stochastic comparison between the minimum eigenvalues of $\mathbf{Y}$ and $\mathbf{Z}$. In expectation,

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Y}) \geq \mathbb{E}\,\lambda_{\min}(\mathbf{Z}) - \sqrt{2\sigma_*^2(\mathbf{Z}) \log(2d)}. \tag{2.4}$$

In addition, for all $t \geq 0$, the lower tail satisfies

$$\mathbb{P}\{\lambda_{\min}(\mathbf{Y}) \leq \mathbb{E}\,\lambda_{\min}(\mathbf{Z}) - t\} \leq 2d \cdot \mathrm{e}^{-t^2/(2\sigma_*^2(\mathbf{Z}))}. \tag{2.5}$$

The weak variance is defined as

$$\sigma_*^2(\mathbf{Z}) := \max_{\|\mathbf{u}\|=1} \mathrm{Var}[\mathbf{u}^*\mathbf{Z}\mathbf{u}]. \tag{2.6}$$

Theorem 2.4 is a (nontrivial) corollary of Theorem 2.1. For the proof, the strategy is to draw a large sample from the distribution of the summand $\mathbf{W}$ and to approximate the iid sum $\mathbf{Y}$ by extracting a sparse subsample. See Section 9 for the grisly details.

Let us expand on the difference between the two comparison theorems. While Theorem 2.1 forces the summands to take the simple form $W_i\mathbf{A}_i$, each summand has its own distribution. In contrast, Theorem 2.4 comprehends a general distribution $\mathbf{W}$, but it requires the summands to be identical copies of $\mathbf{W}$. We permit inequality in the comparison (2.2) to facilitate the application of the result. The Gaussian comparison model (2.3) for the sum $\mathbf{Y}$ can also be written in the form

$$\mathbf{Z} = n \cdot (\mathbb{E}\,\mathbf{X}) + \sqrt{n} \cdot (\mathbf{X} - \mathbb{E}\,\mathbf{X}) \sim \text{NORMAL}(n \cdot \mathbb{E}[\mathbf{X}],\ n \cdot \mathsf{Var}[\mathbf{X}]). \tag{2.7}$$

The statement (2.7) follows from the stability properties of the Gaussian. Compare with (1.2).

Theorem 2.4 supports many applications. As a first example, Section 2.4 provides a bound for the minimum eigenvalue of a Wishart matrix. Section 5 develops some old and new results on the minimum eigenvalue of a sample covariance matrix. Section 6 studies random linear embeddings, and it resolves a recalcitrant problem on the injectivity properties of very sparse random matrices.

Conjecture 2.5 (Non-iid sums). *It is natural to conjecture a comparison between the random matrices*

$$\mathbf{Y} = \sum_{i=1}^{n} \mathbf{W}_i \quad \textit{where } \mathbf{W}_i \textit{ are psd and independent;}$$

$$\mathbf{Z} = \sum_{i=1}^{n} \mathbf{X}_i \quad \textit{where } \mathbf{X}_i \sim \textit{NORMAL}(\mathbb{E}[\mathbf{W}_i],\ \mathsf{Mom}[\mathbf{W}_i]) \textit{ are independent.}$$

In this setting, we can establish the weaker expectation bound

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Y}) \geq \mathbb{E}\,\lambda_{\min}(\mathbf{Z}) - \sum_{i=1}^{n} \sqrt{2\sigma_*^2(\mathbf{X}_i)\log(2d)}.$$

We believe that stronger statements, parallel with Theorem 2.4*, should remain valid. See Section* 9 *for a summary of the argument and the obstacles to improving the result.*

2.4. **Example: Wishart matrix.** As a first example, we treat the minimum eigenvalue of a standard real Wishart matrix [Mui82, Sec. 3.2]. Introduce the random, rank-one psd matrix

$$\mathbf{W} = \mathbf{g}\mathbf{g}^{\mathsf{T}} \in \mathbb{H}_d(\mathbb{R}) \quad \text{where } \mathbf{g} \sim \text{NORMAL}_{\mathbb{R}}(\mathbf{0}, \mathbf{I}_d).$$

Draw independent copies $\mathbf{W}_1, \dots, \mathbf{W}_n$ of the random matrix $\mathbf{W}$, and form the sum:

$$\mathbf{Y} = \sum_{i=1}^{n} \mathbf{W}_i \sim \text{WISHART}_{\mathbb{R}}(\mathbf{I}_d, n). \tag{2.8}$$

After a calculation of the first and second moments of $\mathbf{W}$, Theorem 2.4 furnishes a comparison between the Wishart matrix $\mathbf{Y}$ and the Gaussian matrix

$$\mathbf{Z} = n \cdot \mathbf{I}_d + \gamma\sqrt{n} \cdot \mathbf{I}_d + \sqrt{n} \cdot \mathbf{G}_{\text{goe}} \in \mathbb{H}_d(\mathbb{R}), \tag{2.9}$$

where $\gamma \sim \text{NORMAL}_{\mathbb{R}}(0, 1)$ and $\mathbf{G}_{\text{goe}}$ is drawn independently from the (unnormalized) Gaussian orthogonal ensemble (GOE); see Section 3.7.3 for details. Exploiting standard

facts about the GOE matrix [AS17, Exer. 6.48], we find that the minimum eigenvalue and the weak variance satisfy

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Z}) \geq n - 2\sqrt{dn} \qquad \text{and} \qquad \sigma_*^2(\mathbf{Z}) \leq 3n.$$

Therefore, Theorem 2.4 yields the explicit, nonasymptotic bound

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Y}) \geq n - 2\sqrt{dn} - \sqrt{6n\log(2d)}. \tag{2.10}$$

The bound is nontrivial when $n > \left(2\sqrt{d} + \sqrt{6\log(2d)}\right)^2$. In particular, it suffices that $n \gtrsim d$.

How tight is the inequality (2.10)? Rescale by the number n of samples, and introduce the aspect ratio $\varrho := d/n \in (0, 1]$. When $d, n \to \infty$ with the ratio ϱ fixed, we determine that

$$\mathbb{E}\,\lambda_{\min}(n^{-1}\mathbf{Y}) \geq 1 - 2\sqrt{\varrho} - \sqrt{6\log(2d)/n} \to 1 - 2\sqrt{\varrho}.$$

In this regime, the sharp asymptotic [BY93, Thms. 1,2] is

$$\lambda_{\min}(n^{-1}\mathbf{Y}) \to 1 - 2\sqrt{\varrho} + \varrho \quad \text{almost surely.}$$

When the aspect ratio ϱ is small (that is, $n \gg d$), the bound (2.10) is correct to first order, including the numerical constant. On the other hand, the bound is only active inside the regime where $n > 4d$, so it does not speak to the more challenging case where $n \approx d$.

2.5. **Nonexample: Wishart matrix.** Instead, suppose that we express the Wishart matrix $\mathbf{Y} \sim \textsc{wishart}_{\mathbb{R}}(\mathbf{I}_d, n)$ as the sum of k iid copies of $\mathbf{W} \sim \textsc{wishart}_{\mathbb{R}}(\mathbf{I}_d, n/k)$, where k divides n.

After a calculation, we obtain the comparison model:

$$\mathbf{Z} = n \cdot \mathbf{I}_d + \gamma n k^{-1/2} \cdot \mathbf{I}_d + \sqrt{n} \cdot \mathbf{G}_{\text{goe}} \in \mathbb{H}_d(\mathbb{R}).$$

The minimum eigenvalue satisfies the same bound as before, but the weak variance is much larger:

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Z}) \geq n - 2\sqrt{dn} \quad \text{and} \quad \sigma_*^2(\mathbf{Z}) \leq 2n + n^2/k.$$

Theorem 2.4 results in the comparison

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Y}) \geq n - 2\sqrt{dn} - \sqrt{2(2n + n^2/k)\log(2d)}.$$

This inequality is vacuous unless the number of summands $k \gtrsim \log d$, and it does not make accurate predictions unless $k \gg \varrho^{-1}\log d$.

From this exercise, we discover that Theorem 2.4 furnishes different conclusions, depending on how we decompose a random matrix as a sum of iid psd terms. To extract the most leverage from the theorem, we want to break the random matrix into the smallest pieces we can. This phenomenon is typical of concentration inequalities for independent sums, which often deteriorate when individual summands are large relative to the sum. The example also supports a general prescription: Theorem 2.4 usually requires $\gtrsim \log d$ summands to produce nontrivial bounds.

2.6. **Equivariance.** An attractive feature of Theorems 2.1 and 2.4 is that the results are equivariant under linear transformations (Proposition 3.1). In particular, if $\mathbf{Z}$ is a Gaussian comparison model for the random psd sum $\mathbf{Y}$, then $\mathbf{K}^*\mathbf{Z}\mathbf{K}$ is a Gaussian comparison model for $\mathbf{K}^*\mathbf{Y}\mathbf{K}$. In this statement, $\mathbf{K}$ is any conformable matrix, not necessarily square.

This observation facilitates the computation of comparison models. For instance, it is often convenient to transform the random matrix $\mathbf{Y}$ so that its expectation $\mathbb{E}\,\mathbf{Y} = \mathbf{I}$. The equivariance property also plays a central role in the analysis of randomized subspace injections (Section 6).

2.7. **Gaussian comparison versus matrix concentration.** When is the Gaussian comparison method effective? Consider a self-adjoint Gaussian matrix $\mathbf{Z} \in \mathbb{H}_d$. Its minimum eigenvalue satisfies

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Z}) \geq \lambda_{\min}(\mathbb{E}\,\mathbf{Z}) - \mathbb{E}\,\|\mathbf{Z} - \mathbb{E}\,\mathbf{Z}\|. \tag{2.11}$$

A sufficient condition for the comparison (2.4) between $\lambda_{\min}(\mathbf{Y})$ and $\lambda_{\min}(\mathbf{Z})$ to be informative is that the right-hand side of (2.11) is positive. To check the latter condition, we need to understand the scale for the expected norm $\mathbb{E}\,\|\mathbf{Z} - \mathbb{E}\,\mathbf{Z}\|$.

To that end, define the *matrix variance* statistic [Tro15, Eqn. (2.2.4)]:

$$\sigma^2(\mathbf{Z}) := \|\mathbb{E}\,(\mathbf{Z} - \mathbb{E}\,\mathbf{Z})^2\|.$$

The matrix variance controls the expected norm of a self-adjoint, centered Gaussian matrix:

$$\sqrt{(2/\pi)\,\sigma^2(\mathbf{Z})} \leq \mathbb{E}\,\|\mathbf{Z} - \mathbb{E}\,\mathbf{Z}\| \leq \sqrt{2\sigma^2(\mathbf{Z})\log(2d)}. \tag{2.12}$$

This statement (2.12) is a variant of the matrix Khinchin inequality (Fact 3.3). Both bounds in (2.12) are saturated. We must undertake a more sensitive analysis to determine whether or not the expected norm includes the dimensional factor, $\log(2d)$. The matrix variance compares with the weak variance:

$$\sigma_*^2(\mathbf{Z}) \leq \sigma^2(\mathbf{Z}) \leq d \cdot \sigma_*^2(\mathbf{Z}). \tag{2.13}$$

Both bounds in (2.13) are attainable. In contrast to $\mathbb{E}\,\lambda_{\min}(\mathbf{Z})$, the statistics $\sigma^2(\mathbf{Z})$ and $\sigma_*^2(\mathbf{Z})$ are easy to compute, as they only depend on the second moments of the random matrix.

We can obtain a coarser version of Theorem 2.4 by incorporating the estimates (2.11), (2.12), and (2.13). For instance, the expectation bound (2.4) implies that

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Y}) \geq \lambda_{\min}(\mathbb{E}\,\mathbf{Z}) - 2\sqrt{2\sigma^2(\mathbf{Z})\log(2d)}. \tag{2.14}$$

In fact, an estimate similar with (2.14) follows from simpler arguments based on the scalar mgf inequality (1.2) and matrix concentration tools [Tro15]; see Section A for details.

The benefits of the Gaussian comparison method now come into sharper focus. Theorems 2.1 and 2.4 are most effective in case

$$\lambda_{\min}(\mathbb{E}\,\mathbf{Z}) \gg \mathbb{E}\,\|\mathbf{Z} - \mathbb{E}\,\mathbf{Z}\| \approx \sigma(\mathbf{Z}) \gg \sigma_*(\mathbf{Z}).$$

But we need a finer scalpel than the matrix Khinchin inequality (2.12) to assess whether the expected norm includes the dimensional factor, $\log(2d)$.

Example 2.6 (Wishart: Matrix concentration)**.** Suppose we apply the matrix concentration bound (2.14) to the comparison model (2.9) for the Wishart matrix $\mathbf{Y} \sim \textsc{wishart}_{\mathbb{R}}(\mathbf{I}_d, n)$. The matrix variance statistic $\sigma^2(\mathbf{Z}) = n(d+2)$, and we arrive at the bound

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Y}) \geq n - 2\sqrt{2n(d+2)\log(2d)}.$$

This bound, while nontrivial, does not capture the correct dimensional dependence that is visible in (2.10). We have squandered the valuable distributional information provided by Theorem 2.4.

Remark 2.7 (Intrinsic freeness)*.* Recent research [BBvH23, Ban+24], inspired by [Tro18], has demonstrated that we can sometimes compare the eigenvalue distribution of a Gaussian matrix with a free probability model, using simple summary statistics. These intrinsic freeness results can be valuable for handling the Gaussian comparison model, but they are unhelpful for the applications in this paper.

2.8. **Related work.** This paper is the first to recognize the potential applications of Stahl's theorem (Fact 8.2) in random matrix theory. A companion paper [Tro26a] exploits other facets of Stahl's theorem to derive comparison theorems for the extreme eigenvalues of a sum of independent random self-adjoint matrices. We believe that there are many further opportunities at this nexus.

We are not aware of previous results that are similar in detail with the Gaussian comparisons from Theorems 2.1 and 2.4. Nevertheless, the literature is scattered with breadcrumbs that form a broken trail toward these new results.

2.8.1. *Universality laws for random matrices.* Compared with our work, the results that are closest in spirit appear in a recent paper of Brailovskaya & van Handel [BvH24]. They derive quantitative universality theorems for sums of independent random matrices, in the spirit of the Berry–Esseen theorem. See [Tro26b] for an alternative approach to their results. As we will explain, their results are incomparable with the ones in this paper.

Brailovskaya & van Handel consider an independent sum of random self-adjoint matrices $\mathbf{W}_i$, not necessarily psd or identically distributed. They compare the sum with a Gaussian model that shares the same first- and second-order statistics, as in the multivariate central limit theorem:

$$\mathbf{Y} = \sum_{i=1}^{n} \mathbf{W}_i \quad \text{and} \quad \mathbf{Z}' \sim \textsc{normal}(\mathbb{E}[\mathbf{Y}],\, \mathsf{Var}[\mathbf{Y}]).$$

In contrast, our approach requires the summands $\mathbf{W}_i$ to be iid random psd matrices, and it compares the sum with a different Gaussian model that has slightly higher variance.

Under appropriate conditions on the summands, Brailovskaya & van Handel argue that the eigenvalue distribution of $\mathbf{Y}$ and the eigenvalue distribution of the Gaussian model $\mathbf{Z}'$ are similar. Their results address both the spectral density and the spectral support. As one may imagine, it appears to require stricter assumptions on the statistics of the random matrices to ensure this strong affinity.

For instance, to control the minimum eigenvalue of the random sum, Brailovskaya & van Handel assume that the uniform bound statistic

$$R \coloneqq \mathbb{E} \max_i \|\mathbf{W}_i - \mathbb{E}\,\mathbf{W}_i\|^2 \ll \sigma(\mathbf{Y})(\log d)^{-3}.$$

This condition ensures that each one of the summands makes a limited contribution to the sum. The restriction is particularly important when the random matrices have few moments or the summands have inhomogeneous distributions. Under this surmise, they prove [BvH24, Thm. 2.8] that the expected distance between the minimum eigenvalues satisfies

$$\mathbb{E}\,|\lambda_{\min}(\mathbf{Y}) - \lambda_{\min}(\mathbf{Z}')| \lesssim \sigma(\mathbf{Z}')^{5/6} R^{1/6} \log d + \sigma_*(\mathbf{Z}')(\log d)^{1/2}.$$

As in the present paper, the second term on the right arises from Gaussian concentration. To understand the first term, recall from (2.12) that the statistic $\sigma(\mathbf{Z}')$ controls the scale of the minimum eigenvalue: $\mathbb{E}\,\lambda_{\min}(\mathbf{Z}' - \mathbb{E}\,\mathbf{Z}')$. In some examples, the factor $R^{1/6} \log d$ submerges the first term below $\sigma(\mathbf{Z}')$.

As with our Gaussian comparison theorems, the proof strategy in the paper of Brailovskaya & van Handel is based on techniques from Stein's method. While there are small points of similarity (e.g., the use of interpolation), our technical apparatus follows an independent design.

The results of Brailovskaya & van Handel are powerful and wide ranging. For some of the applications we consider in this paper, however, their approach leads to parasitic error terms that make it impossible to reach the optimal bounds. These factors are particularly significant in applications to computational mathematics (Section 6), where constants and logarithms matter. The results in this paper and in the companion paper [Tro26a] sometimes allow for superior bounds.

2.8.2. *A sum of independent random psd matrices.* There is also a body of work that provides specialized bounds for the minimum eigenvalue of an independent sum of random psd matrices. Several of these papers are inspired by the same observation (1.3) that an independent sum of nonnegative real random variables has a Gaussian lower tail, and they pursue this insight in creative and multifarious ways. We focus on the earliest and most distinctive contributions.

This literature treats several distinct random matrix models. To facilitate comparisons, we summarize the implications for a special case involving sample covariance matrices. Consider a real random vector $\mathbf{w} \in \mathbb{R}^d$ that has four finite moments. For normalization, assume that the vector is *centered* and *isotropic*: $\mathbb{E}[\mathbf{w}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{w}\mathbf{w}^\mathsf{T}] = \mathbf{I}_d$. Form the sample covariance matrix based on n samples:

$$\mathbf{Y} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_i \mathbf{w}_i^\mathsf{T} \quad \text{where } \mathbf{w}_i \sim \mathbf{w} \text{ iid.}$$

For a parameter $\varepsilon \in (0, 1)$, how many samples $n = n(\varepsilon)$ are sufficient to ensure that $\lambda_{\min}(\mathbf{Y}) \geq 1 - \varepsilon$ with high probability? The weak assumptions on moments make this question very challenging.

Srivastava & Vershynin [SV13] formulated this problem and made the first contribution. Consider a random vector $\mathbf{w} \in \mathbb{R}^d$ that satisfies the uniform fourth moment condition

$$\max\nolimits_{\|\mathbf{u}\|=1} \mathbb{E}\,|\langle \mathbf{u},\, \mathbf{w}\rangle|^4 \leq L. \tag{2.15}$$

They established a sample complexity bound:

$$n \geq 400 L \varepsilon^{-3} d \quad \text{implies} \quad \mathbb{E}\,\lambda_{\min}(\mathbf{Y}) \geq 1 - \varepsilon.$$

Their paper is important because the sample complexity n has the correct (linear) dependence on the dimension d, although it exhibits a suboptimal dependence on ε. Their proof employs a Stieltjes transform argument inspired by work in spectral graph theory [BSS14].

Koltchinskii & Mendelson [KM15] developed a family of related results under a weak form of the moment condition (2.15). For some $\eta > 2$, assume that

$$\mathbb{P}\{|\langle \mathbf{u}, \mathbf{w}\rangle| > t\} \le Lt^{-(2+\eta)} \quad \text{when } \|\mathbf{u}\| = 1 \text{ and } t \ge 1.$$

In this regime, their work yields the correct dependence on the parameter ε. Indeed,

$$n \ge C\varepsilon^{-2}d \quad \text{implies} \quad \mathbb{P}\{\lambda_{\min}(\mathbf{Y}) < 1 - \varepsilon\} \le C\log(\mathrm{e}n/d) \cdot \mathrm{e}^{-d/C}.$$

The constant $C = C(\eta, L)$. Their proof involves truncation, small ball probabilities, and a Vapnik–Chervonenkis dimension argument.

Oliveira [Oli16] proposed a third approach. Under the assumption (2.15), he established that

$$n \ge 81L\varepsilon^{-2}(d + 2\log(2/\delta)) \quad \text{implies that} \quad \mathbb{P}\{\lambda_{\min}(\mathbf{Y}) < 1 - \varepsilon\} \le \delta. \tag{2.16}$$

The bound (2.16) has the correct dependence on all of the parameters. Oliveira's argument relies on the PAC-Bayesian method [LST02, McA03, AC11], a technique that smooths the distribution and employs the variational formulation of the entropy to obtain bounds. In Section 5, we will exploit the Gaussian comparison method to give a short proof of Oliveira's bound (2.16).

The papers discussed in this section depend heavily on moment assumptions, such as (2.15), which cannot capture the detailed distribution of a random vector. To highlight the benefits of the Gaussian comparison method, we will obtain new bounds for the sample covariance of a very sparse random vector (Theorem 5.4) that are not accessible via the methods outlined in this section.

3. Gaussian random matrices

To take advantage of the Gaussian comparison method, we must be able to construct the comparison model and determine its spectral properties. This section outlines some facts about Gaussian random matrices that will play a role in the applications and the proofs of the main theorems. For generality, we work in the complex setting, which includes the real setting as a special case.

3.1. The Cartesian decomposition.

To simplify some formulas, let us introduce the Cartesian decomposition of a square matrix. The *real part* and *imaginary part* of a square matrix $\mathbf{M} \in \mathbb{M}_d(\mathbb{C})$ with complex entries are the self-adjoint matrices

$$\operatorname{Re}\mathbf{M} := \frac{1}{2}(\mathbf{M} + \mathbf{M}^*) \in \mathbb{H}_d(\mathbb{C}) \quad \text{and} \quad \operatorname{Im}\mathbf{M} := \frac{1}{2\mathrm{i}}(\mathbf{M} - \mathbf{M}^*) \in \mathbb{H}_d(\mathbb{C}).$$

The *Cartesian decomposition* states that $\mathbf{M} = (\operatorname{Re}\mathbf{M}) + \mathrm{i}(\operatorname{Im}\mathbf{M})$, and the two terms are orthogonal with respect to the trace inner product. In other terms, the Cartesian decomposition presents the matrix as the sum of its Hermitian and skew-Hermitian parts.

3.2. **First and second moments of random matrices.** Let $\mathbf{X} \in \mathbb{H}_d(\mathbb{C})$ be a random self-adjoint matrix with complex entries. The second moment function (Definition 2.3) contains information about the second moment of each entry of the random matrix:

$$\mathsf{Mom}[\mathbf{X}](\operatorname{Re}\mathbf{E}_{ij}) = \mathbb{E}\,|\operatorname{Re} X_{ij}|^2 \quad \text{and} \quad \mathsf{Mom}[\mathbf{X}](\operatorname{Im}\mathbf{E}_{ij}) = \mathbb{E}\,|\operatorname{Im} X_{ij}|^2.$$

Here, $\mathbf{E}_{ij} \in \mathbb{M}_d(\mathbb{C})$ is the square matrix with a one in the (i, j) position and zeros elsewhere. We can extract mixed second moments via polarization. For example,

$$\mathsf{Mom}[\mathbf{X}]\big((\operatorname{Re}\mathbf{E}_{ij}) + (\operatorname{Re}\mathbf{E}_{k\ell})\big) - \mathsf{Mom}[\mathbf{X}]\big((\operatorname{Re}\mathbf{E}_{ij}) - (\operatorname{Re}\mathbf{E}_{k\ell})\big) = 4\,\mathbb{E}\big[(\operatorname{Re} X_{ij})(\operatorname{Re} X_{k\ell})\big].$$

In the last two displays, the range of the indices $i, j, k, \ell = 1, \dots, d$.

The variance function (Definition 2.3) collects the second moments of the centered random matrix:

$$\mathsf{Var}[\mathbf{X}] = \mathsf{Mom}[\mathbf{X} - \mathbb{E}\,\mathbf{X}] = \mathsf{Mom}[\mathbf{X}] - \mathsf{Mom}[\mathbb{E}\,\mathbf{X}].$$

For random self-adjoint matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{H}_d(\mathbb{C})$, the variance function is additive in the sense that

$$\mathsf{Var}[\mathbf{X} + \mathbf{Y}] = \mathsf{Var}[\mathbf{X}] + \mathsf{Var}[\mathbf{Y}] \quad \text{when } \mathbf{X}, \mathbf{Y} \text{ are independent.} \tag{3.1}$$

For contrast, the second moment function Mom does not satisfy the additivity rule (3.1).

The first and second moments of a random matrix are equivariant under linear transformations. This fact allows us to transfer the comparison theorems between models. We remark that there are many types of linear transformations that preserve the cone of psd matrices [Bha07, Sec. 2.1].

Proposition 3.1 (Moments: Equivariance). *Consider random self-adjoint matrices* $\mathbf{W}, \mathbf{X} \in \mathbb{H}_d(\mathbb{C})$ *with*

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{W}] \quad \textit{and} \quad \mathsf{Var}[\mathbf{X}] \succeq \mathsf{Mom}[\mathbf{W}].$$

Then, for each real-linear map $\mathcal{L} : \mathbb{H}_d(\mathbb{C}) \to \mathbb{H}_{d'}(\mathbb{C})$ *on self-adjoint matrices,*

$$\mathbb{E}[\mathcal{L}(\mathbf{X})] = \mathbb{E}[\mathcal{L}(\mathbf{W})] \quad \textit{and} \quad \mathsf{Var}[\mathcal{L}(\mathbf{X})] \succeq \mathsf{Mom}[\mathcal{L}(\mathbf{W})].$$

Proof. By linearity,

$$\mathbb{E}[\mathcal{L}(\mathbf{X})] = \mathcal{L}(\mathbb{E}\,\mathbf{X}) = \mathcal{L}(\mathbb{E}\,\mathbf{W}) = \mathbb{E}[\mathcal{L}(\mathbf{W})].$$

By duality, for each $\mathbf{M} \in \mathbb{H}_d(\mathbb{C})$,

$$\begin{aligned}
\mathsf{Var}[\mathcal{L}(\mathbf{X})](\mathbf{M}) &= \mathrm{Var}[\langle \mathbf{M},\ \mathcal{L}(\mathbf{X})\rangle] = \mathrm{Var}[\langle \mathcal{L}^*(\mathbf{M}),\ \mathbf{X}\rangle] \\
&= \mathsf{Var}[\mathbf{X}](\mathcal{L}^*(\mathbf{M}) \succeq \mathsf{Mom}[\mathbf{W}](\mathcal{L}^*(\mathbf{M})) \\
&= \mathbb{E}\,|\langle \mathcal{L}^*(\mathbf{M}),\ \mathbf{W}\rangle|^2 = \mathbb{E}\,|\langle \mathbf{M},\ \mathcal{L}(\mathbf{W})\rangle|^2 = \mathsf{Mom}[\mathcal{L}(\mathbf{W})](\mathbf{M}).
\end{aligned}$$

As usual, $\mathcal{L}^*$ is the adjoint of the linear map with respect to the trace inner product. □

3.3. **Matrix Gaussian series.** Section 2.2 provides an abstract definition of a Gaussian random matrix. This section introduces a concrete model that is often useful for calculations. A (self-adjoint) *matrix Gaussian series* is a random matrix model of the form

$$\mathbf{Z} = \mathbf{\Delta} + \sum_{i=1}^{n} \gamma_i \mathbf{H}_i \quad \text{where } \gamma_i \sim \textsc{normal}_{\mathbb{R}}(0, 1) \text{ iid.} \tag{3.2}$$

The coefficients $(\mathbf{\Delta}, \mathbf{H}_1, \dots, \mathbf{H}_n)$ are deterministic matrices in $\mathbb{H}_d(\mathbb{C})$. Let us emphasize that the Gaussian random variables γ_i are real-valued. Every self-adjoint Gaussian random matrix can be written in the form (3.2) in many different ways.

We can easily compute the mean and variance function of the random matrix (3.2):

$$\mathbb{E}[\mathbf{Z}] = \mathbf{\Delta} \quad \text{and} \quad \mathsf{Var}[\mathbf{Z}](\mathbf{M}) = \sum_{i=1}^{n} |\langle \mathbf{M},\ \mathbf{H}_i \rangle|^2 \quad \text{for } \mathbf{M} \in \mathbb{H}_d(\mathbb{C}). \tag{3.3}$$

The expression for the variance function follows from the additivity rule (3.1).

3.4. **Gaussian monotonicity.** Gaussian random matrices enjoy a strong monotonicity property. As the variance function increases, expectations of convex functions of the random matrix also increase.

Proposition 3.2 (Gaussians: Monotonicity)**.** *Consider two self-adjoint Gaussian matrices in $\mathbb{H}_d(\mathbb{C})$ with distributions $\mathbf{Z} \sim \textsc{normal}(\mathbf{\Delta}, \mathsf{V})$ and $\mathbf{Z}' \sim \textsc{normal}(\mathbf{\Delta}, \mathsf{V}')$ that share the same expectation. Suppose the variance functions V and V' satisfy the pointwise inequality*

$$\mathsf{V}(\mathbf{M}) \le \mathsf{V}'(\mathbf{M}) \quad \textit{for all } \mathbf{M} \in \mathbb{H}_d(\mathbb{C}). \tag{3.4}$$

For each convex function $f : \mathbb{H}_d(\mathbb{C}) \to \mathbb{R}$ where the expectations are finite,

$$\mathbb{E}\, f(\mathbf{Z}) \le \mathbb{E}\, f(\mathbf{Z}'). \tag{3.5}$$

Proof. Since $\mathsf{V}' - \mathsf{V}$ is a positive quadratic form on $\mathbb{H}_d(\mathbb{C})$, we can construct a Gaussian matrix $\mathbf{X} \sim \textsc{normal}(\mathbf{0}, \mathsf{V}' - \mathsf{V})$ that is independent from $\mathbf{Z}$. By linearity of expectation, $\mathbb{E}[\mathbf{Z} + \mathbf{X}] = \mathbf{\Delta}$. By additivity of the variance function (3.1),

$$\mathsf{Var}[\mathbf{Z} + \mathbf{X}] = \mathsf{Var}[\mathbf{Z}] + \mathsf{Var}[\mathbf{X}] = \mathsf{V} + (\mathsf{V}' - \mathsf{V}) = \mathsf{V}'.$$

Therefore, the sum $\mathbf{Z} + \mathbf{X} \sim \textsc{normal}(\mathbf{\Delta}, \mathsf{V}')$ follows the same distribution as $\mathbf{Z}'$. To obtain the convexity inequality (3.5), note that

$$\mathbb{E}\, f(\mathbf{Z}') = \mathbb{E}\, f(\mathbf{Z} + \mathbf{X}) \ge \mathbb{E}\, f(\mathbf{Z}).$$

We have invoked Jensen's inequality, conditional on $\mathbf{Z}$. □

3.5. **Matrix variance.** We can capture information about spectral features of a Gaussian matrix using scalar summary statistics. The *matrix variance* of a self-adjoint Gaussian matrix $\mathbf{Z} \in \mathbb{H}_d(\mathbb{C})$ is

$$\sigma^2(\mathbf{Z}) := \left\| \mathbb{E}\,(\mathbf{Z} - \mathbb{E}\,\mathbf{Z})^2 \right\| = \left\| \sum_{i=1}^{n} \mathbf{H}_i^2 \right\|. \tag{3.6}$$

The second identity in (3.6) is valid for an arbitrary representation (3.2) of $\mathbf{Z}$ as a Gaussian series. When $\mathbf{Z}, \mathbf{Z}'$ are independent, we have the subadditivity rule $\sigma^2(\mathbf{Z} + \mathbf{Z}') \le \sigma^2(\mathbf{Z}) + \sigma^2(\mathbf{Z}')$.

The matrix Khinchin inequality [Tro15, Thm. 4.6.1] states that the matrix variance controls the expectation of the extreme eigenvalues of the centered Gaussian matrix:

Fact 3.3 (Matrix Khinchin)**.** Consider a self-adjoint Gaussian matrix $\mathbf{Z}$ with dimension d. Then

$$-\mathbb{E}\,\lambda_{\min}(\mathbf{Z} - \mathbb{E}\,\mathbf{Z}) = \mathbb{E}\,\lambda_{\max}(\mathbf{Z} - \mathbb{E}\,\mathbf{Z}) \le \sqrt{2\sigma^2(\mathbf{Z}) \log d}. \tag{3.7}$$

In general, the logarithmic factor in (3.7) is required. The lower bound stated in (2.12) follows from a sharp moment comparison [LO99, Cor. 3] and Jensen's inequality.

3.6. **Weak variance.** For a self-adjoint Gaussian matrix $\mathbf{Z} \in \mathbb{H}_d(\mathbb{C})$, the *weak variance* statistic is defined as

$$\sigma_*^2(\mathbf{Z}) := \max_{\|\mathbf{u}\|=1} \mathsf{Var}[\mathbf{u}^*\mathbf{Z}\mathbf{u}] = \max_{\|\mathbf{u}\|=1} \sum_{i=1}^n (\mathbf{u}^*\mathbf{H}_i\mathbf{u})^2. \tag{3.8}$$

The second identity holds for an arbitrary representation (3.2) of $\mathbf{Z}$ as a Gaussian series. As stated in (2.13), the weak variance is controlled by the matrix variance, and both bounds are attainable.

The weak variance can easily be expressed in terms of the variance function:

$$\sigma_*^2(\mathbf{Z}) = \max_{\|\mathbf{u}\|=1} \mathsf{Var}[\mathbf{Z}](\mathbf{u}\mathbf{u}^*) = \max_{\|\mathbf{M}\|_*=1} \mathsf{Var}[\mathbf{Z}](\mathbf{M}).$$

We have written $\|\cdot\|_*$ for the Schatten 1-norm, also known as the nuclear norm. As a consequence, for self-adjoint Gaussian matrices $\mathbf{Z}, \mathbf{Z}' \in \mathbb{H}_d(\mathbb{C})$, the weak variance is monotone with respect to the variance function:

$$\mathsf{Var}[\mathbf{Z}] \le \mathsf{Var}[\mathbf{Z}'] \quad \text{implies} \quad \sigma_*^2(\mathbf{Z}) \le \sigma_*^2(\mathbf{Z}'). \tag{3.9}$$

When $\mathbf{Z}, \mathbf{Z}'$ are independent, we also have the subadditivity rule $\sigma_*^2(\mathbf{Z} + \mathbf{Z}') \le \sigma_*^2(\mathbf{Z}) + \sigma_*^2(\mathbf{Z}')$.

The weak variance arises when studying concentration properties of Gaussian random matrices.

Fact 3.4 (Gaussian concentration). Let $h : \mathbb{H}_d(\mathbb{C}) \to \mathbb{R}$ be a function that is 1-Lipschitz with respect to the ℓ_2 operator norm:

$$|h(\mathbf{A}) - h(\mathbf{B})| \le \|\mathbf{A} - \mathbf{B}\| \quad \text{for all } \mathbf{A}, \mathbf{B} \in \mathbb{H}_d(\mathbb{C}).$$

For a self-adjoint Gaussian matrix $\mathbf{Z} \in \mathbb{H}_d(\mathbb{C})$, the mgf of $h(\mathbf{Z})$ satisfies

$$\mathbb{E}\, \mathrm{e}^{\theta(h(\mathbf{Z}) - \mathbb{E}\, h(\mathbf{Z}))} \le \mathrm{e}^{\theta^2 \sigma_*^2(\mathbf{Z})/2} \quad \text{for all } \theta \in \mathbb{R}. \tag{3.10}$$

We can instantiate (3.10) with $h = \lambda_{\max}$ or $h = \lambda_{\min}$ because of Weyl's inequality [Bha97, Cor. III.2.6].

Sketch of proof. To confirm (3.10), define the composite function

$$f : (\gamma_1, \dots, \gamma_n) \mapsto h\left(\mathbf{\Delta} + \sum_{i=1}^n \gamma_i \mathbf{H}_i\right) =: h(\mathbf{Z}).$$

We quickly bound the Lipschitz constant of f with respect to the ℓ_2 norm:

$$\begin{aligned} |f(a_1, \dots, a_n) - f(b_1, \dots, b_n)| &\le \Big\| \sum_{i=1}^n (a_i - b_i)\mathbf{H}_i \Big\| \\ &= \max_{\|\mathbf{u}\|=1} \sum_{i=1}^n (a_i - b_i)(\mathbf{u}^*\mathbf{H}_i\mathbf{u}) \le \sigma_*(\mathbf{Z}) \cdot \|\mathbf{a} - \mathbf{b}\|. \end{aligned}$$

The last inequality is Cauchy–Schwarz. Gaussian Lipschitz concentration [BLM13, Thm. 5.5] yields the mgf bound (3.10). □

3.7. **Basic examples.** Let us collect some simple Gaussian matrices that we may combine to build comparison models. Indeed, we can represent a Gaussian matrix as a sum of independent Gaussian matrices by breaking the variance function into pieces and identifying a Gaussian model for each piece. This strategy is justified by the additivity (3.1) of the variance. In this section, the random variables $\gamma, \gamma_i, \gamma_{jk}, \gamma'_{jk}$ are iid $\textsc{normal}_{\mathbb{R}}(0,1)$. The distribution $\textsc{normal}_{\mathbb{C}}(0,1)$ generates a complex Gaussian random variable whose real and imaginary parts are iid $\textsc{normal}_{\mathbb{R}}(0,1/2)$ random variables.

3.7.1. *Scalar Gaussian matrix.* As a warmup, consider the scalar matrix $\mathbf{S} \coloneqq \gamma\mathbf{I} \in \mathbb{H}_d(\mathbb{C})$. It is easy to see that $\mathbb{E}\,\lambda_{\min}(\mathbf{S}) = \mathbb{E}\,\lambda_{\max}(\mathbf{S}) = 0$. The variance function of the scalar matrix acts as

$$\mathsf{Var}[\mathbf{S}](\mathbf{M}) = (\operatorname{Tr}\mathbf{M})^2 \quad \text{for } \mathbf{M} \in \mathbb{H}_d(\mathbb{C}).$$

Note that the matrix variance and weak variance coincide: $\sigma^2(\mathbf{S}) = \sigma_*^2(\mathbf{S}) = 1$.

3.7.2. *Diagonal Gaussian matrix.* Next, consider the diagonal matrix

$$\mathbf{D} \coloneqq \sum_{i=1}^{d} \gamma_i \mathbf{E}_{ii} \in \mathbb{H}_d(\mathbb{C}).$$

The extreme eigenvalues satisfy

$$-\mathbb{E}\,\lambda_{\min}(\mathbf{D}) = \mathbb{E}\,\lambda_{\max}(\mathbf{D}) = \mathbb{E}\max\{\gamma_i : i = 1, \dots, d\} \le \sqrt{2\log d}.$$

This expectation bound is asymptotically sharp. The variance function of the diagonal matrix is

$$\mathsf{Var}[\mathbf{D}](\mathbf{M}) = \sum_{i=1}^{d} |m_{ii}|^2 \quad \text{for } \mathbf{M} \in \mathbb{H}_d(\mathbb{C}).$$

The matrix variance and weak variance again coincide: $\sigma^2(\mathbf{D}) = \sigma_*^2(\mathbf{D}) = 1$.

3.7.3. *Gaussian orthogonal ensemble.* Now, we turn to a more sophisticated example. A random matrix from the (unnormalized) *Gaussian orthogonal ensemble* (*GOE*) takes the form

$$\mathbf{G}_{\text{goe}} \coloneqq \sqrt{2} \sum_{j,k=1}^{d} \gamma_{jk} \cdot (\operatorname{Re}\mathbf{E}_{jk}) \in \mathbb{H}_d(\mathbb{R}).$$

Note that this matrix is real and symmetric. The diagonal entries are iid $\textsc{normal}_{\mathbb{R}}(0,2)$ random variables, while the entries in the upper triangle are iid $\textsc{normal}_{\mathbb{R}}(0,1)$ random variables. The key property of the GOE distribution is its orthogonal invariance:

$$\mathbf{Q}^{\mathsf{T}}\mathbf{G}_{\text{goe}}\mathbf{Q} \sim \mathbf{G}_{\text{goe}} \quad \text{for each orthogonal } \mathbf{Q} \in \mathbb{M}_d(\mathbb{R}). \tag{3.11}$$

The extreme eigenvalues of the GOE matrix satisfy an elegant bound [AS17, Exer. 6.48]:

$$-\mathbb{E}\,\lambda_{\min}(\mathbf{G}_{\text{goe}}) = \mathbb{E}\,\lambda_{\max}(\mathbf{G}_{\text{goe}}) \le 2\sqrt{d}. \tag{3.12}$$

To recognize the GOE matrix, observe that its variance function is

$$\mathsf{Var}[\mathbf{G}_{\text{goe}}](\mathbf{M}) = 2\sum_{j,k=1}^{d} |m_{jk}|^2 = 2\|\mathbf{M}\|_{\text{F}}^2 \quad \text{for } \mathbf{M} \in \mathbb{H}_d(\mathbb{R}).$$

The matrix variance and weak variance differ almost as much as possible. Indeed, $\sigma^2(\mathbf{G}_{\mathrm{goe}}) = d+1$ and $\sigma_*^2(\mathbf{G}_{\mathrm{goe}}) = 2$. GOE matrices arise in several of our comparisons. (For example, see Section 2.4.)

3.7.4. *Gaussian unitary ensemble.* The *Gaussian unitary ensemble* (*GUE*) is the complex cousin of the GOE. A random matrix from this family takes the form

$$\mathbf{G}_{\mathrm{gue}} \coloneqq \sum_{j,k=1}^{d} \left[\gamma_{jk} \cdot (\operatorname{Re} \mathbf{E}_{jk}) + \gamma'_{jk} \cdot (\operatorname{Im} \mathbf{E}_{jk})\right] \in \mathbb{H}_d(\mathbb{C}).$$

The diagonal entries are iid real $\textsc{normal}_{\mathbb{R}}(0,1)$ random variables, while the entries in the upper triangle are iid complex $\textsc{normal}_{\mathbb{C}}(0,1)$ random variables. The GUE distribution is unitarily invariant:

$$\mathbf{U}^* \mathbf{G}_{\mathrm{gue}} \mathbf{U} \sim \mathbf{G}_{\mathrm{gue}} \quad \text{for each unitary } \mathbf{U} \in \mathbb{M}_d(\mathbb{C}).$$

The extreme eigenvalues of the GUE matrix satisfy the following bound [AS17, Exer. 6.38].

$$-\mathbb{E}\,\lambda_{\min}(\mathbf{G}_{\mathrm{gue}}) = \mathbb{E}\,\lambda_{\max}(\mathbf{G}_{\mathrm{gue}}) \le 2\sqrt{d}.$$

The variance function of the GUE matrix acts as

$$\mathsf{Var}[\mathbf{G}_{\mathrm{gue}}](\mathbf{M}) = \sum_{j,k=1}^{d} |m_{jk}|^2 = \|\mathbf{M}\|_{\mathrm{F}}^2 \quad \text{for } \mathbf{M} \in \mathbb{H}_d(\mathbb{C}).$$

The matrix variance and weak variance differ as much as possible: $\sigma^2(\mathbf{G}_{\mathrm{gue}}) = d$ and $\sigma_*^2(\mathbf{G}_{\mathrm{gue}}) = 1$. Among all random matrices, the GUE matrix is perhaps the most fundamental.

4. Application: Sampling from a design

Our first application is geometric. If we randomly subsample a set of vectors that spans a (complex) linear space, does the reduced set still span the space? In considering this question, we must enforce some regularity properties to avoid situations where most of the vectors fall in a proper subspace. For illustration, we impose a rather strong condition, which ensures that the vectors are distributed evenly across the unit sphere.

4.1. **Projective designs.** Consider a finite system $\mathcal{U} \coloneqq (\mathbf{u}_1, \dots, \mathbf{u}_n)$ of unit-norm vectors in $\mathbb{C}^d$. The system is called a *complex projective t-design* when it yields a quadrature rule for homogeneous degree-$2t$ polynomials on the complex unit sphere $\mathbb{S}^{d-1}(\mathbb{C})$. More precisely, for each vector $\mathbf{a} \in \mathbb{C}^d$,

$$\frac{1}{n} \sum_{i=1}^{n} |\langle \mathbf{a},\, \mathbf{u}_i \rangle|^{2t} = \mathbb{E}\,|\langle \mathbf{a},\, \mathbf{v} \rangle|^{2t} \quad \text{where } \mathbf{v} \sim \textsc{uniform}(\mathbb{S}^{d-1}).$$

In particular, a *projective 1-design* is the same as an *isotropic system* (also known as a *unit-norm tight frame*), so it spans the whole space:

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^* = \frac{1}{d} \cdot \mathbf{I}_d. \tag{4.1}$$

We are interested in a *projective 2-design*, which is characterized by the condition

$$\frac{1}{n}\sum_{i=1}^{n} |\langle \mathbf{M},\ \mathbf{u}_i\mathbf{u}_i^*\rangle|^2 = \frac{1}{d(d+1)}\big[\|\mathbf{M}\|_{\mathrm{F}}^2 + (\operatorname{Tr}\mathbf{M})^2\big] \quad \text{for all } \mathbf{M} \in \mathbb{H}_d(\mathbb{C}). \tag{4.2}$$

A projective 2-design is always a projective 1-design. Examples of projective 2-designs include systems of equiangular vectors, mutually unbiased bases, and other highly symmetric configurations. See [Wal18, Chap. 8] or [GKKT20, App. B.1] for more examples and discussion.

4.2. **Sampling from a projective design.** A classic question in nonasymptotic random matrix theory asks when a random subset of an isotropic system remains a spanning set. The fundamental result for this problem is due to Rudelson [Rud99]; see Section 4.2.1 for a proof sketch.

Fact 4.1 (Sampling: Projective 1-design). Consider a complex projective 1-design $\mathcal{U} \coloneqq (\mathbf{u}_1, \dots, \mathbf{u}_n)$ consisting of n unit-norm vectors in $\mathbb{C}^d$. For a parameter $1 \le s \le n$, construct a random subsystem $\mathcal{U}'$ with an average of s vectors by uniform sampling:

$$\mathcal{U}' = (\mathbf{u}_i : \xi_i = 1) \quad \text{where } \xi_i \sim \textsc{bernoulli}(s/n) \text{ iid.} \tag{4.3}$$

When $s \ge d\log(d/\delta)$, the random subset $\mathcal{U}'$ spans $\mathbb{C}^d$ with probability at least $1-\delta$. The logarithmic factor in the sampling complexity is necessary.

As a complement to Fact 4.1, we study the problem of sampling from a complex projective 2-design. We will establish the optimal result: a random subset remains a spanning set when the average number of vectors exceeds the ambient dimension by a constant factor. The proof appears in Section 4.2.2.

Theorem 4.2 (Sampling: Projective 2-design). *With the notation of Fact* 4.1*, assume that the system $\mathcal{U}$ is a complex projective 2-design. When the average number s of sampled vectors satisfies*

$$s \ge 4\big[\sqrt{d} + \sqrt{\log(d/\delta)}\big]^2,$$

the random subset $\mathcal{U}'$ spans $\mathbb{C}^d$ with probability at least $1-\delta$.

To analyze the subsampled system $\mathcal{U}'$, defined in (4.3), we introduce the random psd matrix

$$\mathbf{Y} \coloneqq \sum_{\mathbf{u}\in\mathcal{U}'} \mathbf{u}\mathbf{u}^* = \sum_{i=1}^{n} \xi_i \mathbf{u}_i\mathbf{u}_i^* \quad \text{where } \xi_i \sim \textsc{bernoulli}(s/n) \text{ iid.} \tag{4.4}$$

Observe that the system $\mathcal{U}'$ spans $\mathbb{C}^d$ if and only if $\lambda_{\min}(\mathbf{Y}) > 0$. We can obtain the minimum eigenvalue bound from a quick application of the Gaussian comparison method (Theorem 2.1).

In contrast, the existing techniques for minimum eigenvalues [SV13, KM15, Oli16] all fail to provide the correct bound for $\lambda_{\min}(\mathbf{Y})$ because the summands in (4.4) are both spiky and non-identically distributed. Spectral universality tools, such as [BvH24, Cor. 2.7], also produce the wrong answer when applied to the random matrix $\mathbf{Y}$.

4.2.1. *Proof of Fact* 4.1. Assume that $\mathcal{U}$ is a complex projective 1-design. For the proof, we express the average number of sampled vectors $s = \beta d$ in terms of a parameter in the range $1 \le \beta \le n/d$.

Consider the random matrix $\mathbf{Y}$, defined in (4.4). By the isotropy property (4.1),

$$\mathbb{E}\,\mathbf{Y} = \frac{\beta d}{n} \sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^* = \beta \cdot \mathbf{I}_d.$$

The $\mathbf{u}_i$ are unit vectors, so the spectral norm of each summand in (4.4) is bounded by one. An application of the matrix Chernoff inequality [Tro15, Thm. 5.1.1] yields

$$\mathbb{P}\{\lambda_{\min}(\mathbf{Y}) = 0\} \le d \cdot \mathrm{e}^{-\beta}.$$

If the oversampling parameter $\beta \ge \log(d/\delta)$, then the probability that $\mathbf{Y}$ is singular is at most δ.

Last, to confirm that the sampling complexity bound is unimprovable without further assumptions, consider the 1-design obtained by repeating the standard basis for $\mathbb{R}^d$:

$$\mathcal{U} = (\underbrace{\mathbf{e}_1, \dots, \mathbf{e}_1}_{R}, \underbrace{\mathbf{e}_2, \dots, \mathbf{e}_2}_{R}, \dots, \underbrace{\mathbf{e}_d, \dots, \mathbf{e}_d}_{R}).$$

The parameter R is a large natural number. To reliably select each distinct basis vector at least once, it takes $s \ge d \log d$ samples because of the coupon collector phenomenon [MU17, Sec. 2.4.1].

4.2.2. *Proof of Theorem* 4.2. Assume that $\mathcal{U}$ is a complex projective 2-design, and parameterize the average number of samples as $s = \beta d$.

Consider the random matrix $\mathbf{Y}$ defined in (4.4). Theorem 2.1 immediately yields a comparison with the Gaussian matrix

$$\begin{aligned} \mathbf{Z} &= \sum_{i=1}^{n} X_i \mathbf{u}_i \mathbf{u}_i^* && \text{where } X_i \sim \textsc{normal}_{\mathbb{R}}(\beta d/n,\ \beta d/n) \text{ iid;} \\ &\sim \beta \cdot \mathbf{I}_d + \sqrt{\frac{\beta d}{n}} \sum_{i=1}^{n} \gamma_i \mathbf{u}_i \mathbf{u}_i^* && \text{where } \gamma_i \sim \textsc{normal}_{\mathbb{R}}(0,1) \text{ iid.} \end{aligned}$$

Indeed, $\mathbb{E}\,\mathbf{Z} = \beta \cdot \mathbf{I}_d$ because a 2-design satisfies the isotropy property (4.1). The 2-design property (4.2) also allows us to compute the variance function. For self-adjoint $\mathbf{M} \in \mathbb{H}_d(\mathbb{C})$,

$$\mathsf{Var}[\mathbf{Z}](\mathbf{M}) = \frac{\beta d}{n} \sum_{i=1}^{n} |\langle \mathbf{M},\ \mathbf{u}_i \mathbf{u}_i^* \rangle|^2 = \frac{\beta}{d+1} \left[\|\mathbf{M}\|_{\mathrm{F}}^2 + (\operatorname{Tr} \mathbf{M})^2 \right].$$

The variance function determines the distribution of the random part of $\mathbf{Z}$. In view of the discussion in Section 3.7, we identify a GUE matrix and a scalar Gaussian matrix:

$$\mathbf{Z} = \beta \cdot \mathbf{I}_d + \sqrt{\frac{\beta}{d+1}} (\mathbf{G}_{\mathrm{gue}} + \gamma \mathbf{I}_d) \text{ where } \mathbf{G}_{\mathrm{gue}} \text{ and } \gamma \sim \textsc{normal}_{\mathbb{R}}(0,1) \text{ are independent.}$$

Using the inequality (3.12) for the GUE eigenvalues, we get

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Z}) = \beta + \sqrt{\frac{\beta}{d+1}} \cdot \mathbb{E}\left[\lambda_{\min}(\mathbf{G}_{\mathrm{gue}}) + \gamma\right] \ge \beta - 2\sqrt{\frac{\beta d}{d+1}} > \beta - 2\sqrt{\beta}.$$

Meanwhile, the weak variance satisfies

$$\sigma_*^2(\mathbf{Z}) \le \frac{\beta}{d+1}[\sigma_*^2(\mathbf{G}_{\text{gue}}) + \sigma_*^2(\gamma\mathbf{I})] < \frac{2\beta}{d}.$$

Instantiating Theorem 2.1, we arrive at the probability inequality

$$\mathbb{P}\left\{\lambda_{\min}(\mathbf{Y}) \le \beta - 2\sqrt{\beta} - t\right\} \le d \cdot \mathrm{e}^{-t^2 d/(4\beta)}.$$

Choose $t = \beta - 2\sqrt{\beta}$, set the failure probability equal to δ, and solve for the oversampling parameter β to complete the argument. □

5. Application: Sample covariance matrices

Our next application arises from high-dimensional statistics. Suppose that we want to detect which linear marginals of a random vector have strictly positive variance [SV13]. One procedure is to draw iid copies of the random vector and to form the sample covariance matrix. The sample covariance matrix provides estimates for the variance of each marginal. How many samples are sufficient to ensure that none of these estimates is too small?

Using the Gaussian comparison theorem (Theorem 2.4), we can reproduce and extend several major results on sample covariance matrices from the recent literature. When the second moments and fourth moments of the random vector are comparable, then the sampling complexity is proportional to the dimension of the random vector [Oli16]. We can also obtain useful information about the sample covariance matrix of a very sparse random vector [DZ24].

5.1. **The sample covariance matrix.** Consider a real random vector $\mathbf{w} \in \mathbb{R}^d$ with dimension d that has four finite moments: $\mathbb{E}\,\|\mathbf{w}\|^4 < +\infty$. Assume that the random vector is centered, and introduce the (population) covariance matrix:

$$\mathbb{E}[\mathbf{w}] = \mathbf{0} \quad \text{and} \quad \mathbf{K} \coloneqq \mathbb{E}[\mathbf{w}\mathbf{w}^\intercal].$$

For simplicity, we also assume that the covariance matrix $\mathbf{K}$ has full rank d. We can compute the variance of a linear marginal of the random vector in terms of the covariance matrix:

$$\operatorname{Var}[\langle \mathbf{a},\ \mathbf{w}\rangle] = \mathbf{a}^\intercal\mathbf{K}\mathbf{a} \quad \text{for each } \mathbf{a} \in \mathbb{R}^d. \tag{5.1}$$

Our goal is to obtain lower bounds for the variance in every direction, which we call the *variance detection* problem.

Suppose that we sample n iid copies of the random vector $\mathbf{w}$. The *sample covariance matrix* of this data is the random psd matrix

$$\widehat{\mathbf{K}}_n \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbf{w}_i\mathbf{w}_i^\intercal \quad \text{where } \mathbf{w}_i \sim \mathbf{w} \text{ iid}. \tag{5.2}$$

The sample covariance is an unbiased estimator for the population covariance: $\mathbb{E}[\widehat{\mathbf{K}}_n] = \mathbf{K}$ for each $n \in \mathbb{N}$. Moreover, the sample covariance matrix provides estimates for the variance of each linear marginal (5.1) of the distribution. For a parameter $\varepsilon \in (0, 1)$, we aspire that, with high probability,

$$\mathbf{a}^\intercal\widehat{\mathbf{K}}_n\mathbf{a} \ge (1-\varepsilon)\cdot\mathbf{a}^\intercal\mathbf{K}\mathbf{a} \quad \text{for all } \mathbf{a} \in \mathbb{R}^d. \tag{5.3}$$

How many samples $n = n(\mathbf{w}, \varepsilon)$ are sufficient to ensure that the property (5.3) is likely to prevail?

The answer to this question depends on the distribution of the random vector $\mathbf{w}$. Since $\mathbf{K}$ has rank d, it is necessary that $n \geq d$ because the rank of the sample covariance matrix $\widehat{\mathbf{K}}_n$ does not exceed the number n of samples. For a worst-case vector that satisfies $\|\mathbf{w}\| \lesssim \sqrt{d}$, it is necessary to draw $n \asymp d \log d$ samples [Rud99].

A major technical challenge is to find large classes of random vectors where the sample complexity of (5.3) is proportional to the dimension: $n \asymp d$. This problem has already inspired a vast literature that draws on a diverse set of technical ideas. The Gaussian comparison inequality (Theorem 2.4) offers a new methodology for addressing this problem.

Remark 5.1 (Sample covariance: Complex setting). Our approach adapts to the complex setting with minimal changes. We work in the real setting, as it is more typical in the statistics literature.

5.2. **Sample covariance: Four moment theorem.** Our first result reconstructs a prominent theorem from high-dimensional statistics, in a form due to Oliveira [Oli16]. When the second and fourth moments of a random vector are comparable, then the sample complexity of the variance detection problem is proportional to the dimension. The proof appears in Section 5.2.2.

Theorem 5.2 (Sample covariance matrices: Four moment theorem). *Suppose that* $\mathbf{w} \in \mathbb{R}^d$ *is a centered random vector with covariance matrix* $\mathbf{K}$ *and with four finite moments. For a constant* $\beta \geq 1$*, assume that*

$$\mathbb{E}\,|\langle \mathbf{a},\, \mathbf{w}\rangle|^4 \leq \beta^2 \cdot \big(\mathbb{E}\,|\langle \mathbf{a},\, \mathbf{w}\rangle|^2\big)^2 \quad \text{for each } \mathbf{a} \in \mathbb{R}^d. \tag{5.4}$$

For parameters $\varepsilon, \delta \in (0, 1)$*, suppose that the number* n *of samples satisfies*

$$n \geq \frac{24\beta^2 \cdot (d \vee \log(2d/\delta))}{\varepsilon^2}. \tag{5.5}$$

Then, with probability at least $1 - \delta$*, the sample covariance matrix* $\widehat{\mathbf{K}}_n$ *defined in* (5.2) *satisfies*

$$\mathbf{a}^\intercal \widehat{\mathbf{K}}_n \mathbf{a} \geq (1 - \varepsilon) \cdot \mathbf{a}^\intercal \mathbf{K} \mathbf{a} \quad \text{for all } \mathbf{a} \in \mathbb{R}^d. \tag{5.6}$$

The sample complexity (5.5) contains a "startup cost" of $n \gtrsim \beta^2 \log(d/\delta)$ samples to ensure that the failure probability is controlled. Usually, the sample complexity is governed by the other term: $n \gtrsim \beta^2 d$. The statistic β relates the second and fourth moments of marginals of the random vector.

The sample complexity $n \asymp d$ whenever the comparison factor β is independent of the dimension d. Many types of random vectors meet this criterion, including random vectors with independent bounded entries, log-concave random vectors, and tensor products of independent random vectors from these classes. We refer to the literature [SV13, KM15, Oli16, Zhi24] for more examples and discussion.

5.2.1. *Prior work.* Srivastava & Vershynin [SV13, Thm. 1.5] established the first result in the spirit of Theorem 5.2, with suboptimal dependence on the parameter ε. Koltchinskii & Mendelson [KM15, Thm. 1.3] obtained improvements to the dependence on ε

with slightly stricter moment assumptions. Oliveira [Oli16, Thm. 1.1] obtained a version of the result stated here, which gives the correct dependence on all the parameters. See Section 2.8.2 for more discussion.

Several papers [SV13, KM15, Tik18] present nonasymptotic bounds on the minimum eigenvalue of the sample covariance assuming control of the $2 + \eta$ moments, where $\eta > 0$. These conditions are weaker than (5.4). Nevertheless, we will see that the Gaussian comparison method can tease out structure from the random vector that is invisible to a uniform bound on moments (Section 5.3).

5.2.2. *Proof of Theorem* 5.2. The key idea is to exploit the superior concentration of the Gaussian comparison model. Indeed, we can invoke the most elementary net argument to bound the norm of the Gaussian matrix. The result we need is encapsulated in Lemma 5.3, which is established in Section 5.2.3.

Lemma 5.3 (Gaussian matrix: Bounded variance). *Let $\mathbf{Z} \in \mathbb{H}_d(\mathbb{R})$ be symmetric and Gaussian with*

$$\operatorname{Var}[\mathbf{a}^\intercal \mathbf{Z} \mathbf{a}] \le \beta^2 \cdot \|\mathbf{a}\|^2 \quad \text{for each } \mathbf{a} \in \mathbb{R}^d. \tag{5.7}$$

Then

$$\mathbb{E}\,\|\mathbf{Z} - \mathbb{E}\,\mathbf{Z}\| \le \sqrt{12\beta^2 d} \qquad \text{and} \qquad \sigma_*^2(\mathbf{Z}) \le \beta^2.$$

Let us continue with the proof of Theorem 5.2. We may assume that the covariance matrix $\mathbf{K}$ has full rank by restricting our attention to its range. Observe that the moment condition (5.4) and the conclusion (5.6) are both invariant under invertible linear transformations of the random vector. Therefore, we can make the change of variables $\mathbf{w} \mapsto \mathbf{K}^{-1/2}\mathbf{w}$ to ensure that the random vector $\mathbf{w}$ is isotropic: $\mathbb{E}[\mathbf{w}\mathbf{w}^\intercal] = \mathbf{I}_d$.

Introduce the Gaussian matrix $\mathbf{X} \in \mathbb{H}_d(\mathbb{R})$ with statistics

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{w}\mathbf{w}^\intercal] = \mathbf{I}_d \qquad \text{and} \qquad \operatorname{Var}[\mathbf{X}] = \operatorname{Mom}[\mathbf{w}\mathbf{w}^\intercal].$$

Note that the variances of the quadratic forms satisfy

$$\operatorname{Var}[\mathbf{a}^\intercal \mathbf{X} \mathbf{a}] = \operatorname{Var}[\mathbf{X}](\mathbf{a}\mathbf{a}^\intercal) = \operatorname{Mom}[\mathbf{w}\mathbf{w}^\intercal](\mathbf{a}\mathbf{a}^\intercal) = \mathbb{E}\,|\langle \mathbf{a},\ \mathbf{w}\rangle|^4 \le \beta^2 \cdot \|\mathbf{a}\|^2.$$

The last inequality is the hypothesis (5.4) of the theorem, plus isotropy.

According to Theorem 2.4, the Gaussian comparison model $\mathbf{Z} \in \mathbb{H}_d(\mathbb{R})$ for the sample covariance matrix $\widehat{\mathbf{K}}_n$ is

$$\mathbf{Z} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i \sim (\mathbb{E}\,\mathbf{X}) + \frac{1}{\sqrt{n}}(\mathbf{X} - \mathbb{E}\,\mathbf{X}) \quad \text{where } \mathbf{X}_i \sim \mathbf{X} \text{ iid.}$$

Using Lemma 5.3, we get

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Z}) \ge \lambda_{\min}(\mathbb{E}\,\mathbf{X}) - n^{-1/2}\,\mathbb{E}\,\|\mathbf{X} - \mathbb{E}\,\mathbf{X}\| \ge 1 - \sqrt{12\beta^2 d/n}.$$

Meanwhile, the weak variance $\sigma_*^2(\mathbf{Z}) \le \beta^2/n$. Activating Theorem 2.4, we discover that

$$\mathbb{P}\left\{\lambda_{\min}(\widehat{\mathbf{K}}_n) \ge 1 - \sqrt{12\beta^2 d/n} - \varepsilon/\sqrt{12}\right\} \le 2d \cdot \mathrm{e}^{-n\varepsilon^2/(24\beta^2)}.$$

Introduce the stated bound (5.5) for the sample complexity n and simplify the expression. The Rayleigh variational formula for the minimum eigenvalue yields the conclusion (5.6). □

5.2.3. *Proof of Lemma* 5.3. Without loss of generality, we may assume that $\mathbf{Z}$ is centered. The statement about the weak variance is immediate. Indeed, by its definition (3.8) and the hypothesis (5.7) of the lemma,

$$\sigma_*^2(\mathbf{Z}) = \max_{\|\mathbf{u}\|=1} \operatorname{Var}[\mathbf{u}^\intercal \mathbf{Z}\mathbf{u}] \le \beta^2.$$

For the norm bound, recall that the unit sphere in $\mathbb{R}^d$ admits an ε-net $\mathcal{N}$ with at most $(1 + 2/\varepsilon)^d$ points [Ver18, Cor. 4.2.13]. By the standard discretization argument [Ver18, Sec. 4.4],

$$\mathbb{E}\,\|\mathbf{Z}\| \le \mathbb{E} \max_{\mathbf{u}\in\mathcal{N}} |\mathbf{u}^\intercal \mathbf{Z}\mathbf{u}| + 2\varepsilon \cdot \mathbb{E}\,\|\mathbf{Z}\|.$$

Since $\operatorname{Var}[\mathbf{u}^\intercal \mathbf{Z}\mathbf{u}] \le \beta^2$ uniformly, the maximal inequality [vH16, Lem. 5.1] for Gaussian random variables asserts that

$$\mathbb{E} \max_{\mathbf{u}\in\mathcal{N}} |\mathbf{u}^\intercal \mathbf{Z}\mathbf{u}| \le \sqrt{2\beta^2 \log(2|\mathcal{N}|)} \le \sqrt{2\beta^2(\log 2 + d\log(1 + 2/\varepsilon))}.$$

The factor two inside the logarithm takes care of the absolute value in the maximum. Combine the last two displays, and solve for the expected norm to obtain

$$\mathbb{E}\,\|\mathbf{Z}\| \le \frac{1}{1 - 2\varepsilon}\sqrt{2\beta^2(\log 2 + d\log(1 + 2/\varepsilon))}.$$

Select $\varepsilon = 1/12$, and note that the numerical constant is bounded by $\sqrt{12}$ for all $d \ge 1$. □

5.3. **Example: Sparse covariance matrices.** For a more challenging example that goes beyond the scope of Theorem 5.2, we consider variance detection for a sparse random vector. For simplicity, we treat the case of iid entries, but we only require four finite moments.

Fix the dimension d, and let $\zeta \in (0, d]$ be a sparsity parameter. Introduce a real random variable ψ that is standardized and has bounded fourth moment:

$$\mathbb{E}[\psi] = 0 \quad \text{and} \quad \operatorname{Var}[\psi] = 1 \quad \text{and} \quad \mathbb{E}\,|\psi|^4 = B.$$

Construct a sparse random vector $\mathbf{w} \in \mathbb{R}^d$ with iid entries:

$$\mathbf{w} = \sqrt{\frac{d}{\zeta}} \begin{bmatrix} \xi_1\psi_1 \\ \vdots \\ \xi_d\psi_d \end{bmatrix} \quad \text{where} \quad \begin{aligned} \xi_i &\sim \text{BERNOULLI}(\zeta/d) \text{ iid;} \\ \psi_i &\sim \psi \text{ iid.} \end{aligned}$$

It is straightforward to check that $\mathbf{w}$ is centered and isotropic, and $\mathbf{w}$ has ζ nonzero entries on average. How does the sample complexity of the variance detection problem depend on the sparsity?

Theorem 5.4 (Sample covariance: Sparse vectors). *Instate the prevailing notation, and fix parameters $\varepsilon, \delta \in (0, 1)$. Suppose that the number n of samples satisfies*

$$n \ge \frac{25d \cdot (1 \vee (2B\zeta^{-1}\log(2d/\delta)))}{\varepsilon^2}. \tag{5.8}$$

With probability at least $1 - \delta$, the sample covariance matrix of the sparse random vector $\mathbf{w}$ has minimum eigenvalue $\lambda_{\min}(\widehat{\mathbf{K}}_n) \ge 1 - \varepsilon$.

Theorem 5.4 handles two different behavioral regimes:

- When $\zeta \ge 2B\log d$, the sample complexity $n \asymp d$.
- When $\zeta \le 2B\log d$, the sample complexity $n \asymp B\zeta^{-1} d\log d$.

It is quite difficult to confirm the optimal linear sample complexity in the first regime, especially when the distribution ψ has few moments. The scaling $n \gtrsim \zeta^{-1} d \log d$ in the second regime is also sharp. Indeed, with fewer samples, it is likely that one of the diagonal entries of $\widehat{\mathbf{K}}_n$ equals zero. In the ultra-sparse case ($\zeta \leq 1$), similar results follow from classic matrix concentration inequalities, such as Fact A.1.

We were unable to locate any prior work that matches Theorem 5.4, although there are several closely related results from the last few years. In particular, assume that the distribution ψ is bounded. In this case, the paper [CHZ22, Thm. 3.5] achieves the same complexity estimate (5.8) for the expectation of the minimum eigenvalue of the sample covariance matrix. In the regime where the sparsity $\zeta \gtrsim \log d$ and d/n is constant, Dumitriu & Zhu [DZ24, Thm. 2.9] achieve stronger estimates for the expectation of the minimum eigenvalue that are sufficient to attain the Bai–Yin limit, which is out of reach for our method. For a thorough literature review, refer to the paper [DZ24].

5.3.1. *Proof of Theorem* 5.4. By elementary considerations, we can calculate the second moment function of the random matrix $\mathbf{w}\mathbf{w}^\intercal$. Indeed, for symmetric $\mathbf{M} \in \mathbb{H}_d(\mathbb{R})$,

$$\begin{aligned}\mathsf{Mom}[\mathbf{w}\mathbf{w}^\intercal](\mathbf{M}) = \mathbb{E}\left[(\mathbf{w}^\intercal \mathbf{M} \mathbf{w})^2\right] &= \sum_{i \neq j} (2m_{ij}^2 + m_{ii} m_{jj}) + \frac{Bd}{\zeta} \sum_{i=1}^d m_{ii}^2 \\ &\leq 2\|\mathbf{M}\|_{\mathrm{F}}^2 + (\operatorname{Tr} \mathbf{M})^2 + \frac{Bd}{\zeta} \sum_{i=1}^d m_{ii}^2.\end{aligned}$$

A natural comparison model for $\mathbf{w}\mathbf{w}^\intercal$ is the Gaussian random matrix $\mathbf{X} \in \mathbb{H}_d(\mathbb{R})$ with statistics

$$\begin{aligned}\mathbb{E}[\mathbf{X}] &= \mathbb{E}[\mathbf{w}\mathbf{w}^\intercal] = \mathbf{I}_d; \\ \mathsf{Var}[\mathbf{X}](\mathbf{M}) &= 2\|\mathbf{M}\|_{\mathrm{F}}^2 + (\operatorname{Tr} \mathbf{M})^2 + \frac{Bd}{\zeta} \sum_{i=1}^d m_{ii}^2 \geq \mathsf{Mom}[\mathbf{w}\mathbf{w}^\intercal](\mathbf{M}).\end{aligned}$$

Referring back to Section 3.7, we quickly determine the Gaussian matrix with these statistics:

$$\mathbf{X} = \mathbf{I}_d + \mathbf{G}_{\mathrm{goe}} + \gamma \mathbf{I}_d + \sqrt{Bd/\zeta}\, \mathbf{D}.$$

The matrix $\mathbf{D}$ is a diagonal matrix of real standard Gaussian variables, independent from the GOE matrix $\mathbf{G}_{\mathrm{goe}}$ and the scalar Gaussian $\gamma \sim \textsc{normal}_{\mathbb{R}}(0, 1)$.

Therefore, the Gaussian comparison $\mathbf{Z} \in \mathbb{H}_d(\mathbb{R})$ for the sample covariance matrix $\widehat{\mathbf{K}}_n$ is

$$\mathbf{Z} \sim (\mathbb{E}\,\mathbf{X}) + n^{-1/2}(\mathbf{X} - \mathbb{E}\,\mathbf{X}) = \mathbf{I}_d + n^{-1/2}\mathbf{G}_{\mathrm{goe}} + n^{-1/2}\gamma \mathbf{I}_d + \sqrt{Bd/(\zeta n)}\, \mathbf{D}.$$

We may now calculate the spectral statistics:

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Z}) \geq 1 - 2\sqrt{\frac{d}{n}} - \sqrt{\frac{2Bd \log d}{\zeta n}} \qquad \text{and} \qquad \sigma_*^2(\mathbf{Z}) \leq \frac{3 + Bd/\zeta}{n}.$$

Invoking Theorem 2.4,

$$\mathbb{P}\left\{\lambda_{\min}(\widehat{\mathbf{K}}_n) \leq 1 - 2\sqrt{\frac{d}{n}} - \sqrt{\frac{2Bd \log d}{\zeta n}} - \frac{2\varepsilon}{5}\right\} \leq 2d \cdot \exp\left(\frac{-2n\varepsilon^2 \zeta}{25d(B + 3\zeta/d)}\right).$$

Introduce the sample complexity parameter from (5.8), and simplify to reach the stated result. □

6. Application: Randomized subspace injections

In numerical linear algebra, we can devise randomized algorithms for large-scale matrix computations; for example, see [MT20]. The design and analysis of these algorithms often relies on methods from random matrix theory. In fact, the tools in the present paper were developed to treat challenging mathematical problems from this field.

A *randomized subspace injection* is a random linear map whose action preserves the dimension of a fixed subspace [Sar06, OT18]. These injections serve as an important building block for randomized linear algebra algorithms [Woo14, MT20, Mur+23, CEMT25]. In this section, we employ the Gaussian comparison theorem (Theorem 2.4) to develop a new analysis of subspace injections based on very sparse random matrices [CW13, NN13]. The result largely settles an open question of Nelson & Nguyen [NN13, NN14] that has generated a substantial literature; Section 6.3.1 summarizes the prior work.

6.1. Subspace injections.

For a given subspace, a subspace injection is a linear map that does not annihilate any vector in that subspace; see [OT18, Sec. 7.1] or [CEMT25, Def. 1.2].

Definition 6.1 (Subspace injection). Fix a matrix $\mathbf{Q} \in \mathbb{R}^{n\times d}$ with d orthonormal columns. For a fixed *contraction factor* $\alpha > 0$, suppose that $\mathbf{\Phi} \in \mathbb{R}^{k\times n}$ is a matrix that satisfies

$$\|\mathbf{\Phi Q u}\|^2 \geq \alpha \cdot \|\mathbf{u}\|^2 \quad \text{for all } \mathbf{u} \in \mathbb{R}^d. \tag{6.1}$$

When (6.1) holds, we say that $\mathbf{\Phi}$ is an *α-subspace injection* for the d-dimensional range of $\mathbf{Q}$ with *embedding dimension* k. If (6.1) only holds at $\alpha = 0$, then $\mathbf{\Phi}$ is not a subspace injection for range$(\mathbf{Q})$.

The subspace injection property (6.1) has an equivalent spectral formulation:

$$\lambda_{\min}((\mathbf{\Phi Q})^{\mathsf{T}}(\mathbf{\Phi Q})) \geq \alpha > 0. \tag{6.2}$$

Of course, a necessary condition for the subspace injection property (6.1) or (6.2) is that the embedding dimension exceeds the subspace dimension: $k \geq d$. We want to design subspace injections where the embedding dimension $k \asymp d$.

Remark 6.2 (Subspace embedding [Sar06]). For an orthonormal matrix $\mathbf{Q} \in \mathbb{R}^{n\times d}$, suppose that $\mathbf{\Phi} \in \mathbb{R}^{k\times n}$ satisfies the two-sided bound

$$\alpha \cdot \|\mathbf{u}\|^2 \leq \|\mathbf{\Phi Q u}\|^2 \leq \beta \cdot \|\mathbf{u}\|^2 \quad \text{for all } \mathbf{u} \in \mathbb{R}^d. \tag{6.3}$$

Then we say that $\mathbf{\Phi}$ is an *(α, β)-subspace embedding* for the range of $\mathbf{Q}$. While it is critical that the contraction factor $\alpha > 0$, weak control on the ratio β/α suffices for most applications.

6.2. **Randomized subspace injections.** In many applications to computational linear algebra, we must construct a subspace injection without detailed knowledge of the fixed subspace range($\mathbf{Q}$). We can achieve this goal by drawing the matrix $\mathbf{\Phi}$ at random.

Consider an isotropic random vector $\boldsymbol{\phi} \in \mathbb{R}^n$; that is, $\mathbb{E}[\boldsymbol{\phi}\boldsymbol{\phi}^\intercal] = \mathbf{I}_n$. The distribution of the random vector $\boldsymbol{\phi}$ is an algorithm design choice. For an embedding dimension k, construct the random matrix

$$\mathbf{\Phi} = \frac{1}{\sqrt{k}} \begin{bmatrix} \boldsymbol{\phi}_1^\intercal \\ \vdots \\ \boldsymbol{\phi}_k^\intercal \end{bmatrix} \in \mathbb{R}^{k \times n} \quad \text{where } \boldsymbol{\phi}_i \sim \boldsymbol{\phi} \text{ iid.} \tag{6.4}$$

Since the rows are isotropic, the random matrix is an isometry on average:

$$\mathbb{E}\,\|\mathbf{\Phi}\mathbf{u}\|^2 = \|\mathbf{u}\|^2 \quad \text{for each } \mathbf{u} \in \mathbb{R}^n. \tag{6.5}$$

The normalization (6.5) is chosen so that $\mathbf{\Phi}$ typically has a contraction factor $\alpha \leq 1$.

For a fixed subspace $\mathbf{Q} \in \mathbb{R}^{d \times n}$ and embedding dimension k, we want to demonstrate that the random matrix $\mathbf{\Phi}$ is a subspace injection for the range of $\mathbf{Q}$ with high probability. Quantitatively, we need to understand how the contraction factor α depends on the embedding dimension k. In view of (6.2), it suffices to establish a lower bound on the minimum eigenvalue of the random psd matrix

$$\mathbf{Y} \coloneqq (\mathbf{\Phi}\mathbf{Q})^\intercal(\mathbf{\Phi}\mathbf{Q}) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{Q}^\intercal(\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\intercal)\mathbf{Q}. \tag{6.6}$$

By the construction (6.4), the matrix $\mathbf{Y}$ is a sum of iid random psd matrices, so we can activate our Gaussian comparison tools (Theorem 2.4).

6.3. **Sparse dimension reduction maps.** In computational applications, it is desirable to employ very sparse random matrices as subspace injections. While these maps are widely used [MT20, Mur+23], no existing analysis justifies the typical parameter choices. We outline prior work in Section 6.3.1.

Choose an embedding dimension $k \geq d$, and fix the sparsity parameter $\zeta \in (0, k]$. Construct the random vector

$$\boldsymbol{\phi}^\intercal = \sqrt{\frac{k}{\zeta}} \cdot \begin{bmatrix} \xi_1\psi_1 & \dots & \xi_n\psi_n \end{bmatrix} \in \mathbb{R}^n \quad \text{where} \quad \begin{matrix} \xi_i \sim \textsc{bernoulli}(\zeta/k) \text{ iid;} \\ \psi_i \sim \textsc{uniform}\{\pm 1\} \text{ iid.} \end{matrix} \tag{6.7}$$

It is easy to verify that $\boldsymbol{\phi}$ is centered and isotropic. Form the random matrix $\mathbf{\Phi}$, as in (6.4). Observe that $\mathbf{\Phi}$ has an average of ζ nonzero entries per column, for a total of ζn nonzero entries on average. We inquire when this sparse random matrix $\mathbf{\Phi}$ serves as a subspace injection for a fixed subspace. To achieve an embedding dimension $k \asymp d$, what is the minimal sparsity ζ?

The result depends on a geometric property of the subspace. Define the *coherence* $\mu(\mathbf{Q})$ of an orthonormal matrix $\mathbf{Q} \in \mathbb{R}^{n \times d}$ via

$$\mu(\mathbf{Q}) \coloneqq \max\{\|\mathbf{e}_i^\intercal\mathbf{Q}\|^2 : i = 1, \dots, n\}. \tag{6.8}$$

The coherence describes the alignment of the range of $\mathbf{Q}$ with the standard coordinate basis $(\mathbf{e}_1, \dots, \mathbf{e}_n)$. It satisfies the inequalities $d/n \leq \mu(\mathbf{Q}) \leq 1$.

Theorem 6.3 (Subspace injection: Sparse random matrix). *Fix an orthonormal matrix* $\mathbf{Q} \in \mathbb{R}^{n\times d}$ *with coherence* $\mu(\mathbf{Q})$*, defined in* (6.8)*. For parameters* $\varepsilon, \delta \in (0,1)$*, choose*

$$k \geq \frac{16(d \vee 6\log(2d/\delta))}{\varepsilon^2} \qquad \text{and} \qquad \zeta \geq \frac{32\mu(\mathbf{Q})\log(2d/\delta)}{\varepsilon^2}. \tag{6.9}$$

Construct the sparse random matrix $\mathbf{\Phi} \in \mathbb{R}^{k\times n}$ *using* (6.4) *and* (6.7)*. Then* $\mathbf{\Phi}$ *is a* $(1-\varepsilon)$*-subspace injection for* $\operatorname{range}(\mathbf{Q})$ *with probability at least* $1-\delta$*.*

The proof of Theorem 6.3 appears in Section 6.3.2. It follows from a quick application of Gaussian comparison, in the same manner as the result for sparse sample covariance matrices (Theorem 5.4).

Theorem 6.3 has several attractive features. First, we have achieved the optimal embedding dimension $k \asymp d$. Second, the sparsity level $\zeta \lesssim \log d$ for every subspace, nearly matching known lower bounds for sparse subspace embeddings; see (6.10) below. Third, the result shows that we can even reduce the sparsity for subspaces with small coherence. The ultimate limit, for a minimally coherent subspace, is $\zeta \gtrsim (d\log d)/n$, which allows for a random subspace injection with just $\mathcal{O}(d\log d)$ nonzero entries. Further improvements to the sparsity are only possible for special subspaces (e.g., when the rows of $\mathbf{Q}$ compose a complex projective 2-design).

For a sparse random matrix with iid entries, the ε^{-2} dependence in the embedding dimension k and the sparsity ζ seems to be necessary [CDD25]. For most applications, the parameter ε is a constant (say, $\varepsilon = 1/2$), so the poor scaling in ε is not a significant concern. In contrast, the subspace dimension d can be quite large, so it is essential to obtain the minimal dependence on d.

Remark 6.4 (Subspace embedding: Sparse random matrix). To verify the subspace embedding property (Remark 6.2) of the iid sparse random matrix $\mathbf{\Phi}$ in Theorem 6.3, we can easily obtain adequate upper bounds for the dilation factor β. For example, with $k \asymp d$ and $\zeta \asymp \mu(\mathbf{Q})\log d$,

$$\mathbb{E}\,\beta = \mathbb{E}\,\|\mathbf{\Phi Q}\|^2 \lesssim \log d.$$

This statement follows quickly when we apply the matrix Bernstein inequality [Tro15, Thm. 6.1.1] to the decomposition

$$\mathbf{\Phi Q} = \sqrt{\frac{1}{\zeta}} \sum_{i=1}^{k}\sum_{j=1}^{n} \xi_{ij}\psi_{ij}\mathbf{E}_{ij}\mathbf{Q} \quad \text{where} \quad \begin{matrix} \xi_{ij} \sim \textsc{bernoulli}(\zeta/k) \text{ iid;} \\ \psi_{ij} \sim \textsc{uniform}\{\pm 1\} \text{ iid.} \end{matrix}$$

For these parameter choices, the remaining question is whether we can reach the bound $\mathbb{E}\,\beta \leq \text{Const}$.

6.3.1. *Prior work and discussion.* Sarlós [Sar06] introduced the definition of a subspace embedding (Remark 6.2). Clarkson & Woodruff [CW13] proposed the first construction of a sparse subspace embedding. Soon after, Nelson & Nguyen [NN13] identified more effective constructions, including the iid entry model in Theorem 6.3, and a related model, called a *fixed-sparsity subspace embedding*, that has exactly ζ nonzero entries per column. There are several variants of these sparse subspace embeddings, which offer slightly different advantages and disadvantages. For brevity, we summarize results without listing the details of the embedding constructions.

What are the opportunities for and limitations on sparse subspace embeddings? For any fixed-sparsity subspace embedding with conditioning ratio $\beta/\alpha =: 1+\varepsilon$, Nelson &

Nguyen [NN14, Thm. 13] established lower bounds for the parameters:

$$k \gtrsim d/\varepsilon^2 \quad \text{and} \quad \zeta \gtrsim (\log d)/(\varepsilon \log\log d). \tag{6.10}$$

Bourgain et al. [BDN15, Thm. 5] obtained upper bounds for the parameters of sparse subspace embeddings that identified the role of the coherence statistic $\mu(\mathbf{Q})$. Using matrix concentration tools, Cohen [Coh16, Thm. 4.2] obtained upper bounds for fixed-sparsity subspace embeddings:

$$k \lesssim (d \log d)/\varepsilon^2 \quad \text{and} \quad \zeta \lesssim (\log d)/\varepsilon.$$

For a long time, Cohen's analysis was the best available, and it has remained a vexing problem to obtain an upper bound that matches the minimal dependence (6.10).

In applications, constants and logarithms are important. Based on computer experiments, Tropp et al. [TYUC19] recommended the following parameter settings for fixed-sparsity subspace embeddings:

$$k = 2d \quad \text{and} \quad \zeta = 8.$$

These parameters work well, but they lack theoretical justification. Regardless, fixed-sparsity subspace embeddings are widely used in computational linear algebra [MT20, Mur+23].

The last few years have witnessed a burst of new theoretical activity. Cartis et al. [CFS21] established upper bounds that improve over the Bourgain et al. result for subspaces with sufficiently small coherence. Chenakkod et al. [CDDR24, Thm. 1.2] removed the $\log d$ factor from the embedding dimension k at the cost of higher sparsity ζ by invoking results of Brailovskaya & van Handel [BvH24]. At the end of 2024, Chenakkod et al. [CDD25, Thm. 3.4] made further improvements to their arguments, reaching the following guarantee for fixed-sparsity subspace embeddings:

$$k \asymp d/\varepsilon^2 \quad \text{and} \quad \zeta \lesssim \log^2(d/\varepsilon)/\varepsilon + \log^3(d/\varepsilon).$$

The latter result is the state of the art. While the embedding dimension k has the optimal form, the excess logarithmic factors in the sparsity ζ remain a serious limitation. After this paper was written, Chenakkod et al. [CDD26] obtained further improvements, but these results do not operate with proportional embedding dimension $k \asymp d$.

As outlined, the prior work has focused on sparse subspace embeddings (Remark 6.2), rather than sparse subspace injections (Definition 6.1). Nevertheless, the injection property is by far the more important feature, both in theory and in practice; see [OT18, Sec. 7.1] or [MT20, Sec. 8] or [CEMT25]. Our result (Theorem 6.3) is the first to prove that sparse random matrices serve as subspace injections with (essentially) the minimal dependence (6.10) on the subspace dimension d in both the embedding dimension k and the sparsity ζ. The plain role of the subspace coherence $\mu(\mathbf{Q})$ is an added bonus.

Remark 6.5 (Fixed-sparsity subspace injection). We have treated the simplest model for a sparse subspace injection, where the random matrix $\mathbf{\Phi}$ has iid entries. The Gaussian comparison theorem also allows us to study the injection properties of a certain class of fixed-sparsity random matrices [CDD25, Def. 3.2]. To obtain a $(1-\varepsilon)$-subspace injection, it is sufficient that

$$k \asymp d/\varepsilon^2 \qquad \text{and} \quad \zeta \asymp (\log d)/\varepsilon^2.$$

As compared with Theorem 6.3, the subspace coherence $\mu(\mathbf{Q})$ does not appear in this result. See our follow-up paper [CEMT25, Thm. 1.8] for a full proof. As compared with Chennakod et al. [CDD25], we have reduced the dependence on $\log d$ significantly, at the cost of a worse dependence on ε.

6.3.2. *Proof of Theorem* 6.3. To apply the Gaussian comparison theorem, we need to construct an appropriate comparison model. We can accomplish this goal in a few short steps, building on the calculations for the sparse covariance matrix (Theorem 5.4).

As before, the second moment function of the random matrix $\mathbf{W} = \boldsymbol{\phi}\boldsymbol{\phi}^{\mathsf{T}} \in \mathbb{H}_n(\mathbb{R})$ takes the form

$$\mathsf{Mom}[\mathbf{W}](\mathbf{M}) = \sum_{i\neq j}(2m_{ij}^2 + m_{ii}m_{jj}) + \frac{k}{\zeta}\sum_{i=1}^n m_{ii}^2 \leq 2\|\mathbf{M}\|_{\mathrm{F}}^2 + (\operatorname{Tr}\mathbf{M})^2 + \frac{k}{\zeta}\sum_{i=1}^n m_{ii}^2.$$

For the random matrix $\mathbf{W}$, we construct a Gaussian comparison model $\mathbf{X} \in \mathbb{H}_d(\mathbb{R})$ with the statistics

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{W}] = \mathbf{I}_n;$$

$$\mathsf{Var}[\mathbf{X}](\mathbf{M}) = 2\|\mathbf{M}\|_{\mathrm{F}}^2 + (\operatorname{Tr}\mathbf{M})^2 + \frac{k}{\zeta}\sum_{i=1}^n m_{ii}^2 \geq \mathsf{Mom}[\mathbf{W}](\mathbf{M}).$$

Therefore, $\mathbf{X} \sim \mathbf{I}_n + \mathbf{G}_{\mathrm{goe}}^{(n)} + \gamma\mathbf{I}_n + \sqrt{k/\zeta}\,\mathbf{D}_n$. As usual, the GOE matrix $\mathbf{G}_{\mathrm{goe}}^{(n)} \in \mathbb{H}_n(\mathbb{R})$, the standard normal variable $\gamma \sim \textsc{normal}_{\mathbb{R}}(0,1)$, and the standard normal diagonal matrix $\mathbf{D}_n \in \mathbb{H}_n(\mathbb{R})$ are independent.

Since first and second moments are equivariant (Proposition 3.1), the Gaussian random matrix $\mathbf{Q}^{\mathsf{T}}\mathbf{X}\mathbf{Q}$ serves as a comparison model for $\mathbf{Q}^{\mathsf{T}}\mathbf{W}\mathbf{Q}$. That is,

$$\mathbb{E}[\mathbf{Q}^{\mathsf{T}}\mathbf{X}\mathbf{Q}] = \mathbb{E}[\mathbf{Q}^{\mathsf{T}}\mathbf{W}\mathbf{Q}] \qquad \text{and} \qquad \mathsf{Var}[\mathbf{Q}^{\mathsf{T}}\mathbf{X}\mathbf{Q}] \geq \mathsf{Mom}[\mathbf{Q}^{\mathsf{T}}\mathbf{W}\mathbf{Q}].$$

The matrix $\mathbf{Q} \in \mathbb{R}^{n\times d}$ has orthonormal columns, so

$$\mathbf{Q}^{\mathsf{T}}\mathbf{X}\mathbf{Q} \sim \mathbf{I}_d + \mathbf{G}_{\mathrm{goe}}^{(d)} + \gamma\mathbf{I}_d + \sqrt{k/\zeta}\,\mathbf{Q}^{\mathsf{T}}\mathbf{D}_n\mathbf{Q}.$$

We have exploited the fact (3.11) that the GOE distribution is invariant under rotations to confirm that the matrix $\mathbf{G}_{\mathrm{goe}}^{(d)} \in \mathbb{H}_d(\mathbb{R})$ also follows the GOE distribution.

Next, we construct a comparison model $\mathbf{Z}$ for the random matrix $\mathbf{Y}$, appearing in (6.6), that captures the subspace injection property. Since $\mathbf{Y} = k^{-1}\sum_{i=1}^k \mathbf{Q}^{\mathsf{T}}\mathbf{W}_i\mathbf{Q}$ for iid $\mathbf{W}_i \sim \mathbf{W}$, we quickly determine that

$$\begin{aligned}\mathbf{Z} &\sim \mathbf{Q}^{\mathsf{T}}\big[(\mathbb{E}\,\mathbf{X}) + k^{-1/2}(\mathbf{X} - \mathbb{E}\,\mathbf{X})\big]\mathbf{Q} \\ &\sim \mathbf{I}_d + k^{-1/2}\mathbf{G}_{\mathrm{goe}}^{(d)} + \gamma k^{-1/2}\mathbf{I}_d + \zeta^{-1/2}\mathbf{Q}^{\mathsf{T}}\mathbf{D}_n\mathbf{Q}.\end{aligned}$$

To finish the proof, we need to compute spectral statistics.

To that end, express the compression of the diagonal matrix as a Gaussian series:

$$\mathbf{Q}^{\mathsf{T}}\mathbf{D}_n\mathbf{Q} = \sum_{i=1}^n \gamma_i\mathbf{q}_i\mathbf{q}_i^{\mathsf{T}} \in \mathbb{H}_d(\mathbb{R}) \quad \text{where } \gamma_i \sim \textsc{normal}_{\mathbb{R}}(0,1) \text{ iid.}$$

The vector $\mathbf{q}_i^{\mathsf{T}} \in \mathbb{R}^d$ is the ith row of $\mathbf{Q}$. The matrix variance satisfies

$$\sigma^2(\mathbf{Q}^{\mathsf{T}}\mathbf{D}_n\mathbf{Q}) = \Big\|\sum_{i=1}^n(\mathbf{q}_i\mathbf{q}_i^{\mathsf{T}})^2\Big\| = \Big\|\sum_{i=1}^n \|\mathbf{q}_i\|^2\mathbf{q}_i\mathbf{q}_i^{\mathsf{T}}\Big\| \leq \mu(\mathbf{Q})\cdot\|\mathbf{Q}^{\mathsf{T}}\mathbf{Q}\| = \mu(\mathbf{Q}),$$

where $\mu(\mathbf{Q})$ is the coherence (6.8) of the subspace. The matrix Khinchin inequality (Fact 3.3) provides

$$\mathbb{E}\,\lambda_{\max}(\mathbf{Q}^{\mathsf{T}}\mathbf{D}_n\mathbf{Q}) \le \sqrt{2\mu(\mathbf{Q})\log d}.$$

Owing to the comparison (2.13), the weak variance $\sigma_*^2(\mathbf{Q}^{\mathsf{T}}\mathbf{D}_n\mathbf{Q}) \le \sigma^2(\mathbf{Q}^{\mathsf{T}}\mathbf{D}_n\mathbf{Q}) \le \mu(\mathbf{Q})$.

Last, bound the expected minimum eigenvalue and weak variance of the comparison model:

$$\begin{aligned}\mathbb{E}\,\lambda_{\min}(\mathbf{Z}) &\ge 1 - 2\sqrt{d/k} - \sqrt{2\zeta^{-1}\mu(\mathbf{Q})\log d};\\ \sigma_*^2(\mathbf{Z}) &\le 3/k + \mu(\mathbf{Q})/\zeta.\end{aligned}$$

Activating Theorem 2.4,

$$\mathbb{P}\left\{\lambda_{\min}(\mathbf{Y}) \ge 1 - 2\sqrt{d/k} - \sqrt{2\zeta^{-1}\mu(\mathbf{Q})\log d} - \varepsilon/4\right\} \le 2d\cdot\exp\left(\frac{-\varepsilon^2/32}{3/k + \mu(\mathbf{Q})/\zeta}\right).$$

Introduce the parameter choices (6.9) and simplify to reach the stated result. □

7. Gaussian comparison: Positive scalar sums

The technical development commences in this section. As a warmup, we present another proof of the lower tail bound for an independent sum of positive scalar random variables.

Theorem 7.1 (Positive sum: Gaussian lower tail). *Consider an independent family $(W_1, \dots, W_n)$ of nonnegative, square-integrable, real random variables: $W_i \ge 0$ and $\mathbb{E}\,W_i^2 < +\infty$. Introduce the sum of the random variables, along with the sum of second moments:*

$$X := \sum_{i=1}^n W_i \quad \text{and} \quad L_2 := \sum_{i=1}^n \mathbb{E}\,W_i^2.$$

Then the lower tail of the sum X satisfies

$$\mathbb{P}\{X \le \mathbb{E}\,X - t\} \le \mathrm{e}^{-t^2/(2L_2)} \quad \text{for all } t \ge 0.$$

Proof. Introduce the mgf of the lower tail of the centered sum:

$$\mathrm{mgf}_X(\theta) := \mathbb{E}\,\mathrm{e}^{-\theta(X - \mathbb{E}\,X)} \quad \text{for } \theta \ge 0. \tag{7.1}$$

Proposition 7.4 states that $\log \mathrm{mgf}_X(\theta) \le \theta^2 L_2/2$ for all $\theta \ge 0$. The Laplace transform method [BLM13, Sec. 2.2] yields the tail bound

$$\mathbb{P}\{X - \mathbb{E}\,X \le -t\} \le \inf_{\theta>0} \mathrm{e}^{-\theta t + \log \mathrm{mgf}_X(\theta)} \le \inf_{\theta>0} \mathrm{e}^{-\theta t + \theta^2 L_2/2} = \mathrm{e}^{-t^2/(2L_2)}.$$

The infimum is attained when $\theta = t/L_2$. □

The proof of Theorem 7.1 depends on a bound for mgf_X, defined in (7.1). Although this bound is classic [BLM13, Exer. 2.9] and rather elementary (Section 1.2), we will develop an alternative treatment that has more potential for generalization. The results in this section will reappear in the proof of the matrix comparison inequalities.

7.1. **Completely monotone functions.** Our approach takes advantage of a special feature of the decaying exponential that is encapsulated in Definition 7.2.

Definition 7.2 (Completely monotone function)**.** A continuous function $f : \mathsf{I} \to \mathbb{R}$ on the interval $\mathsf{I} \subseteq \mathbb{R}$ is *completely monotone to order K* when its first K derivatives exist and alternate sign:

$$(-1)^k f^{(k)}(w) \geq 0 \quad \text{for all } w \in \operatorname{int}(\mathsf{I}) \text{ and each } k = 0, 1, 2, \dots, K. \tag{7.2}$$

The function f is *completely monotone* when (7.2) holds for each $K \in \mathbb{N}$.

The fundamental example of a completely monotone function is a decaying exponential. For fixed $\theta \geq 0$ and variable $w \in \mathbb{R}$, the function

$$w \mapsto \mathrm{e}^{-\theta w} \quad \text{is completely monotone for } w \in \mathbb{R}.$$

In fact, a continuous function $f : \mathbb{R}_+ \to \mathbb{R}$ on the nonnegative real line is completely monotone if and only if it is the Laplace transform of a finite, positive Borel measure μ:

$$f(w) = \int_{[0,\infty)} \mathrm{e}^{-\theta w} \, \mu(\mathrm{d}\theta). \tag{7.3}$$

The representation (7.3) is called Bernstein's theorem [Wid41, Thm. IV.12a].

Completely monotone functions support a beautiful theory, elaborated in the classic book of Widder [Wid41, Chap. IV]. The monograph of Schilling et al. [SSV10] is a more recent reference. This background is not required for our purposes.

7.2. **Tools from Stein's method.** The main steps in the analysis are adapted from the literature on Charles Stein's method, a collection of tools for establishing distributional approximations and concentration inequalities. This section outlines some ideas from Stein's method. See the survey of Ross [Ros11] or the book of Chen et al. [CGS11] for more information.

To describe the variability in a distribution, we can employ exchangeable pairs of random variables. A pair (W, Y) of real random variables is *exchangeable* when $(W, Y) \sim (Y, W)$. Equivalently,

$$\mathbb{E}\, F(W, Y) = \mathbb{E}\, F(Y, W) \tag{7.4}$$

for every bivariate function $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ where the expectation exists. A pair of iid random variables is the most basic example of an exchangeable pair.

The next ingredient is an elegant covariance identity. Let $g, h : \mathsf{I} \to \mathbb{R}$ be functions on an interval $\mathsf{I} \subseteq \mathbb{R}$. For an iid pair (W, Y) of random variables taking values in I,

$$\operatorname{Cov}(g(W), h(W)) = \frac{1}{2}\, \mathbb{E}\big[(g(W) - g(Y))(h(W) - h(Y))\big]. \tag{7.5}$$

This formula can be verified by direct calculation. It is valid whenever the expectations are finite.

To bound differences of function values, as in (7.5), we employ a formula of Hermite. For a continuously differentiable function $h : \mathsf{I} \to \mathbb{R}$,

$$\frac{h(w) - h(y)}{w - y} = \int_0^1 h'(\tau w + (1 - \tau) y) \, \mathrm{d}\tau \quad \text{for all } w, y \in \mathsf{I}. \tag{7.6}$$

Identity (7.6) follows from the fundamental theorem of calculus. Moreover, if the derivative h' is convex on I, then

$$\frac{h(w) - h(y)}{w - y} \le \frac{h'(w) + h'(y)}{2} \quad \text{for all } w, y \in \mathsf{I}. \tag{7.7}$$

When $w = y$ in (7.6) or (7.7), we interpret the left-hand side as $h'(w)$.

The most important element in our proof of Theorem 7.1 is an association inequality for functions with opposite sense [BLM13, Thm. 2.14]. Assume that $g : \mathsf{I} \to \mathbb{R}$ is increasing, while $h : \mathsf{I} \to \mathbb{R}$ is decreasing. For any random variable W taking values in I,

$$\mathbb{E}[g(W)h(W)] \le \mathbb{E}[g(W)] \cdot \mathbb{E}[h(W)]. \tag{7.8}$$

Equivalently, the covariance of $g(W)$ and $h(W)$ is negative. To establish the result (7.8), note that

$$(g(s) - g(t))(h(s) - h(t)) \le 0 \quad \text{for all } s, t \in \mathsf{I}.$$

Combine this formula with the covariance identity (7.5).

7.3. **Covariance bounds for completely monotone functions.** This section shows how the tools from Stein's method lead to clean bounds for covariances involving a completely monotone function. In particular, this argument applies to the mgf of the lower tail.

Lemma 7.3 (Covariance bound: Completely monotone function). *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a function on the nonnegative real line that is completely monotone to order four. For each nonnegative real random variable W,*

$$\operatorname{Cov}(W, f'(W)) = \mathbb{E}\big[(W - \mathbb{E}\,W)f'(W)\big] \le \mathbb{E}[W^2] \cdot \mathbb{E}[f''(W)].$$

The bound is valid when the expectations are finite.

Proof. According to Definition 7.2, the derivatives of a completely monotone function f alternate sign. We only rely on the conditions that f'' is positive, decreasing, and convex, so complete monotonicity to order four is sufficient.

Let Y be an independent copy of W. By the exchangeable pairs formula (7.5),

$$\begin{aligned}\operatorname{Cov}(W, f'(W)) &= \frac{1}{2}\,\mathbb{E}\big[(W - Y)(f'(W) - f'(Y))\big] \\ &\le \frac{1}{4}\,\mathbb{E}\big[(W - Y)^2(f''(W) + f''(Y))\big] = \frac{1}{2}\,\mathbb{E}\big[(W - Y)^2 f''(W)\big].\end{aligned}$$

The inequality is the bound (7.7) for the divided difference of the function f', whose derivative f'' is convex. In the last step, we have used the exchangeability (7.4) of (W, Y) to simplify the expression.

To continue, let us separate the two factors in the expectation. Note that $(w - y)^2 \le w^2 + y^2$ for $w, y \ge 0$. Since f'' is positive,

$$\operatorname{Cov}(W, f'(W)) \le \frac{1}{2}\,\mathbb{E}\big[(W^2 + Y^2)f''(W)\big] = \frac{1}{2}\,\mathbb{E}[W^2 f''(W)] + \frac{1}{2}\,\mathbb{E}[W^2] \cdot \mathbb{E}[f''(W)].$$

The second step relies on the independence and identical distribution of (W, Y). On the nonnegative real line $\mathbb{R}_+$, the function $w \mapsto w^2$ is increasing, while f'' is decreasing.

Thus, the association inequality (7.8) furnishes a bound for the first expectation on the right-hand side:

$$\mathbb{E}[W^2 f''(W)] \le \mathbb{E}[W^2] \cdot \mathbb{E}[f''(W)].$$

Combine the last two displays to arrive at the desired result. □

Lemma 7.3 depends crucially on the assumption that f''' is negative, which allows us to invoke the association inequality (7.8) to decouple W^2 from $f''(W)$. The entropy method employs association inequalities in a similar fashion; for instance, see [BLM13, Thm. 6.27]. The overall structure of the proof is similar with classic arguments from Stein's method, exemplified in [Cha07, Thm. 1.5(ii)].

7.4. **Positive sum: mgf bound.** With Lemma 7.3 at hand, we quickly obtain a bound for the mgf of a sum of independent, nonnegative random variables.

Proposition 7.4 (Positive sum: mgf bound). *Instate the hypotheses of Theorem* 7.1. *The mgf* (7.1) *satisfies the bound*

$$\log \operatorname{mgf}_X(\theta) \le \theta^2 L_2/2 \quad \text{for all } \theta \ge 0.$$

We can rewrite the outcome of Proposition 7.4 in a more suggestive fashion:

$$\mathbb{E}\, \mathrm{e}^{-\theta X} \le \mathbb{E}\, \mathrm{e}^{-\theta Z} \quad \text{where } Z \sim \textsc{normal}(\mathbb{E}\, X, L_2). \tag{7.9}$$

In other words, we have compared the lower tail of the sum X with the lower tail of an appropriate normal random variable Z whose statistics derive from the summands in X. In Section 8, we will see that the inequality (7.9) generalizes to matrices.

Proof. Recall the definition (7.1) of $\operatorname{mgf}_X(\theta)$. Since $\log \operatorname{mgf}_X(0) = 0$, the result holds for $\theta = 0$, and we may assume that $\theta > 0$. The derivative of the mgf takes the form

$$\begin{aligned}\operatorname{mgf}_X'(\theta) &= -\mathbb{E}\left[(X - \mathbb{E}\, X)\mathrm{e}^{-\theta(X - \mathbb{E}\, X)}\right] \\ &= -\sum_{i=1}^n \mathbb{E}\left[(W_i - \mathbb{E}\, W_i)\mathrm{e}^{-\theta(X - \mathbb{E}\, X)}\right] =: -\sum_{i=1}^n \mathbb{E}\left[(W_i - \mathbb{E}\, W_i)\mathrm{e}^{\theta B_i - \theta W_i}\right].\end{aligned}$$

The second relation expands the sum $X - \mathbb{E}\, X = \sum_{i=1}^n (W_i - \mathbb{E}\, W_i)$. Note that the random variable $B_i := W_i - (X - \mathbb{E}\, X) = \mathbb{E}\, W_i - \sum_{j \ne i}(W_j - \mathbb{E}\, W_j)$ is statistically independent from W_i.

Conditional on B_i, apply Lemma 7.3 to the completely monotone function $f_i(w) := \mathrm{e}^{\theta B_i - \theta w}$ defined for $w \in \mathbb{R}_+$. This step results in the bound

$$\begin{aligned}\operatorname{mgf}_X'(\theta) &= \theta^{-1} \sum_{i=1}^n \operatorname{Cov}(W_i, f_i'(W_i)) \\ &\le \theta \sum_{i=1}^n \mathbb{E}[W_i^2] \cdot \mathbb{E}[\mathrm{e}^{\theta B_i - \theta W_i}] \\ &= \theta \sum_{i=1}^n \mathbb{E}[W_i^2] \cdot \mathbb{E}[\mathrm{e}^{-\theta(X - \mathbb{E}\, X)}] = \theta L_2 \cdot \operatorname{mgf}_X(\theta).\end{aligned}$$

Solve the differential inequality to complete the proof. □

8. Gaussian comparison: Randomly weighted sums of psd matrices

In this section, we turn to the proof of Section 2.1, which provides a bound for the minimum eigenvalue of a randomly weighted sum of fixed psd matrices. For technical reasons, we develop the result using slightly different notation and assumptions.

Fix a system $(\mathbf{A}_1, \dots, \mathbf{A}_n)$ of psd matrices, with common dimension d. These matrices need not be distinct from each other. Consider an independent family $(W_1, \dots, W_n)$ of square-integrable, nonnegative real random variables: $W_i \geq 0$ and $\mathbb{E}\, W_i^2 < +\infty$. Form the random psd matrix

$$\mathbf{X} \coloneqq \sum_{i=1}^{n} W_i \mathbf{A}_i. \tag{8.1}$$

We will compare the random psd matrix $\mathbf{X}$ with an appropriate Gaussian model. Define the deterministic self-adjoint matrices

$$\mathbf{H}_i \coloneqq (\mathbb{E}\, W_i^2)^{1/2} \mathbf{A}_i \quad \text{for } i = 1, \dots, n. \tag{8.2}$$

Construct the centered Gaussian matrix

$$\mathbf{Z} \coloneqq \sum_{i=1}^{n} \gamma_i \mathbf{H}_i \quad \text{where} \quad \gamma_i \sim \text{NORMAL}_{\mathbb{R}}(0,1) \text{ iid.} \tag{8.3}$$

Equivalently, $\mathbf{Z} \sim \text{NORMAL}(\mathbf{0}, \mathsf{V})$ with variance function

$$\mathsf{V}(\mathbf{M}) \coloneqq \sum_{i=1}^{n} |\langle \mathbf{M},\ \mathbf{H}_i \rangle|^2 = \sum_{i=1}^{n} (\mathbb{E}\, W_i^2) \cdot |\langle \mathbf{M},\ \mathbf{A}_i \rangle|^2 \quad \text{for } \mathbf{M} \in \mathbb{H}_d. \tag{8.4}$$

With these definitions, we can state a comparison inequality.

Theorem 8.1 (Gaussian comparison: Weighted psd sum). *Fix an arbitrary self-adjoint matrix $\mathbf{\Delta} \in \mathbb{H}_d$. Introduce the random d-dimensional matrices $\mathbf{X}$ and $\mathbf{Z}$ from* (8.1) *and* (8.3)*. Then*

$$\mathbb{E}\, \lambda_{\min}(\mathbf{X} - \mathbb{E}\, \mathbf{X} + \mathbf{\Delta}) \geq \mathbb{E}\, \lambda_{\min}(\mathbf{Z} + \mathbf{\Delta}) - \sqrt{2\sigma_*^2(\mathbf{Z}) \log d}. \tag{8.5}$$

In addition, for all $t \geq 0$,

$$\mathbb{P}\{\lambda_{\min}(\mathbf{X} - \mathbb{E}\, \mathbf{X} + \mathbf{\Delta}) \leq \mathbb{E}\, \lambda_{\min}(\mathbf{Z} + \mathbf{\Delta}) - t\} \leq d \cdot \mathrm{e}^{-t^2/(2\sigma_*^2(\mathbf{Z}))}. \tag{8.6}$$

The weak variance $\sigma_*^2(\mathbf{Z})$ is defined in (3.8)*.*

Theorem 8.1 generalizes Theorem 2.1 by allowing a shift $\mathbf{\Delta}$ of the expectation. This improvement comes for free and allows for a more transparent and natural argument.

Proof of Theorem 8.1. Proposition 8.5 provides a comparison for the trace mgfs:

$$\operatorname{mgf}_{\mathbf{X}}(\theta) \coloneqq \mathbb{E}\, \operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{X} - \mathbb{E}\, \mathbf{X} + \mathbf{\Delta})} \leq \mathbb{E}\, \operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{Z} + \mathbf{\Delta})} \quad \text{for } \theta \geq 0.$$

Bound the trace exponential on the right-hand side in terms of the minimum eigenvalue:

$$\operatorname{mgf}_{\mathbf{X}}(\theta) \leq d \cdot \mathbb{E}\, \lambda_{\max}\big(\mathrm{e}^{-\theta(\mathbf{Z} + \mathbf{\Delta})}\big) = d \cdot \mathbb{E}\, \mathrm{e}^{-\theta \lambda_{\min}(\mathbf{Z} + \mathbf{\Delta})}.$$

Indeed, the trace of a psd matrix does not exceed the dimension times the maximum eigenvalue. The second relation is the spectral mapping theorem, combined with the fact that the decreasing exponential function reverses the order of the eigenvalues.

To continue, add and subtract $\mathbb{E}\,\lambda_{\min}(\mathbf{Z}+\mathbf{\Delta})$ in the exponent:

$$\operatorname{mgf}_{\mathbf{X}}(\theta) \le d \cdot \mathrm{e}^{-\theta\,\mathbb{E}\,\lambda_{\min}(\mathbf{Z}+\mathbf{\Delta})} \cdot \mathbb{E}\,\mathrm{e}^{-\theta(\lambda_{\min}(\mathbf{Z}+\mathbf{\Delta})-\mathbb{E}\,\lambda_{\min}(\mathbf{Z}+\mathbf{\Delta}))}.$$

The Gaussian concentration inequality (3.10) controls the fluctuations of the minimum eigenvalue $\lambda_{\min}(\mathbf{Z}+\mathbf{\Delta})$ around its mean:

$$\operatorname{mgf}_{\mathbf{X}}(\theta) \le d \cdot \mathrm{e}^{-\theta\,\mathbb{E}\,\lambda_{\min}(\mathbf{Z}+\mathbf{\Delta})} \cdot \mathrm{e}^{\theta^2\sigma_*^2(\mathbf{Z})/2}. \tag{8.7}$$

The inequality (8.7) quickly leads to both the stated results.

To obtain the probability bound (8.6), we employ the matrix Laplace transform method [Tro15, Prop. 3.2.1]. For fixed $t \ge 0$ and arbitrary $\theta > 0$,

$$\begin{aligned}\mathbb{P}\{\lambda_{\min}(\mathbf{X}-\mathbb{E}\,\mathbf{X}+\mathbf{\Delta}) \le \mathbb{E}\,\lambda_{\min}(\mathbf{Z}+\mathbf{\Delta}) - t\} &\le \mathrm{e}^{\theta(\mathbb{E}\,\lambda_{\min}(\mathbf{Z}+\mathbf{\Delta})-t)} \cdot \operatorname{mgf}_{\mathbf{X}}(\theta)\\ &\le d \cdot \mathrm{e}^{-\theta t} \cdot \mathrm{e}^{\theta^2\sigma_*^2(\mathbf{Z})/2}.\end{aligned}$$

The second inequality is (8.7). Select $\theta = t/\sigma_*^2(\mathbf{Z})$ to reach the probability inequality.

The result (8.5) for the expectation follows from a similar argument. For arbitrary $\theta > 0$, the matrix Laplace transform method [Tro15, Prop. 3.2.2] yields

$$\begin{aligned}\mathbb{E}\,\lambda_{\min}(\mathbf{X}-\mathbb{E}\,\mathbf{X}-\mathbf{\Delta}) &\ge -\frac{1}{\theta}\log\operatorname{mgf}_{\mathbf{X}}(\theta)\\ &\ge -\frac{1}{\theta}\left[\log d - \theta\,\mathbb{E}\,\lambda_{\min}(\mathbf{Z}+\mathbf{\Delta}) + \theta^2\sigma_*^2(\mathbf{Z})/2\right].\end{aligned}$$

The second inequality is (8.7). Choose $\theta = \sqrt{(2\log d)/\sigma_*^2(\mathbf{Z})}$ to reach the expectation bound. □

Proof of Theorem 2.1 *from Theorem* 8.1. Choose the shift $\mathbf{\Delta} = \mathbb{E}\,\mathbf{X}$, and define the Gaussian matrix $\mathbf{Z}' := \mathbf{Z} + \mathbb{E}\,\mathbf{X} \sim \text{NORMAL}(\mathbb{E}\,\mathbf{X}, \mathsf{V})$, where V is defined in (8.4). Change variables to state Theorem 2.1 directly in terms of the distribution of $\mathbf{Z}'$. □

8.1. **Stahl's theorem: Complete monotonicity of the trace exponential.** To prove Theorem 8.1, we extend the considerations behind the scalar comparison (Theorem 7.1) to the matrix setting. This strategy depends on a profound fact from matrix analysis, called Stahl's theorem [Sta13].

Fact 8.2 (Stahl's theorem; formerly the BMV conjecture). Fix self-adjoint matrices $\mathbf{A}, \mathbf{B} \in \mathbb{H}_d$, and assume that $\mathbf{A}$ is psd. Then the trace exponential function

$$f(w) := \operatorname{Tr}\exp(\mathbf{B} - w\mathbf{A}) \quad \text{is completely monotone for } w \ge 0.$$

In particular, Lemma 7.3 applies to the function f.

While Stahl's theorem is recent, it has a long history. In 1975, Bessis–Moussa–Villani [BMV75] conjectured Fact 8.2 as part of their method for bounding partition functions of quantum mechanical systems. They established the result in two special cases: (1) when $\mathbf{A}, \mathbf{B}$ commute; or (2) when $\mathbf{A}, \mathbf{B}$ are 2×2 matrices. Over the next decades, the BMV conjecture received significant attention in mathematical physics, but it was not resolved.

In 2004, Lieb & Seiringer [LS04] proved that Fact 8.2 admits an equivalent formulation: For all psd $\mathbf{A}, \mathbf{B} \in \mathbb{H}_d$, the coefficients of the polynomial $w \mapsto \operatorname{Tr}(w\mathbf{A} + \mathbf{B})^p$ are nonnegative for all $p \in \mathbb{N}$. This link with real algebraic geometry ignited a new stage of research, based on sum-of-squares hierarchies and semidefinite programming, that generated new evidence supporting the BMV conjecture.

Finally, in 2012, Herbert Stahl [Sta13] established Fact 8.2 using classic methods from complex analysis. Bernstein's theorem (7.3) states that a completely monotone function is the Laplace transform of a finite, positive Borel measure. Roughly speaking, Stahl inverted the Laplace transform to obtain the representing measure. To prove that the representing measure is positive, he exploited the theory of Riemann surfaces. See [Erë15, Cli16] for other accounts of Stahl's proof.

Very recently, Otte Heinävaara constructed a rather different argument [Hei25, Cor. 1.2] that leads to a remarkable generalization of Fact 8.2:

Fact 8.3 (Heinävaara's theorem). Fix self-adjoint matrices $\mathbf{A}, \mathbf{B} \in \mathbb{H}_d$, and assume that $\mathbf{A}$ is psd. Consider a function $h : \mathbb{R} \to \mathbb{R}$ that is completely monotone to order K. Then the trace function

$$f(w) \coloneqq \operatorname{Tr} h(w\mathbf{A} - \mathbf{B}) \quad \text{is completely monotone to order } K \text{ for } w \in \mathbb{R}.$$

Stahl's theorem (Fact 8.2) follows from the choice $h(w) = \mathrm{e}^{-w}$.

Heinävaara's proof of Fact 8.3 appeals to a novel object, called a *tracial joint spectral measure*, that captures the behavior of trace functions of the form $(w, y) \mapsto \operatorname{Tr} h(w\mathbf{A} + y\mathbf{B})$ for self-adjoint $\mathbf{A}, \mathbf{B}$. His techniques yield many deep new statements about trace functions.

8.2. **Additional tools.** The argument involves some standard tools from probability and matrix analysis. First, we record the Gaussian integration by parts (IBP) rule [NP12, Lem. 1.1.1].

Fact 8.4 (Gaussian IBP). Consider iid real standard normal variables $(\gamma_1, \dots, \gamma_n)$. For each differentiable function $h : \mathbb{R}^n \to \mathbb{R}$,

$$\mathbb{E}\big[\gamma_i \cdot h(\gamma_1, \dots, \gamma_n)\big] = \mathbb{E}\big[(\partial_i h)(\gamma_1, \dots, \gamma_n)\big].$$

In this formula, $\partial_i h$ denotes the partial derivative of h with respect to its ith argument. The identity is valid whenever the right-hand side is finite.

Second, we recall Duhamel's formula from matrix calculus [Bha97, Example X.4.2(v)]. For self-adjoint matrices $\mathbf{A}, \mathbf{H} \in \mathbb{H}_d$, the derivative of the exponential at $\mathbf{A}$ in the direction $\mathbf{H}$ satisfies

$$(\mathrm{D}\,\mathrm{e}^{\mathbf{A}})[\mathbf{H}] \coloneqq \lim_{s \to 0} s^{-1}\big[\mathrm{e}^{\mathbf{A}+s\mathbf{H}} - \mathrm{e}^{\mathbf{A}}\big] = \int_0^1 \mathrm{e}^{\tau\mathbf{A}}\mathbf{H}\mathrm{e}^{(1-\tau)\mathbf{A}}\,\mathrm{d}\tau. \tag{8.8}$$

As a consequence, the derivative of the trace exponential simplifies to

$$\big(\mathrm{D} \operatorname{Tr} \mathrm{e}^{\mathbf{A}}\big)[\mathbf{H}] = \operatorname{Tr}\big[\mathbf{H}\mathrm{e}^{\mathbf{A}}\big]. \tag{8.9}$$

The statement (8.9) follows from (8.8) when we take the trace and cycle to combine the exponentials.

8.3. **Comparison for the trace mgf.** With Stahl's theorem (Fact 8.2) at hand, we can establish a bound for the trace mgf associated with the minimum eigenvalue of a random psd matrix.

Proposition 8.5 (Weighted psd sum: Comparison of trace mgfs). *Instate the hypotheses of Theorem* 8.1. *For* $\theta \geq 0$,

$$\operatorname{mgf}_{\mathbf{X}}(\theta) \coloneqq \mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{X} - \mathbb{E}\mathbf{X} + \boldsymbol{\Delta})} \leq \mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{Z} + \boldsymbol{\Delta})} =: \operatorname{mgf}_{\mathbf{Z}}(\theta).$$

In the scalar setting (Proposition 7.4), the mgf of a Gaussian random variable emerges from a direct argument. In the matrix setting, it is more expedient to compare the trace mgf of the psd model with the trace mgf of the Gaussian model. This argument relies on interpolation, much like classic Gaussian comparison inequalities [Kah86] or more recent work on universality for random matrices [BvH24].

Proof. To implement the comparison, we interpolate between the two random matrix models. Define a path in the space of random matrices:

$$\mathbf{Y}_s := \sqrt{s}\,(\mathbf{X} - \mathbb{E}\,\mathbf{X}) + \sqrt{1-s}\,\mathbf{Z} + \mathbf{\Delta} \quad \text{for } s \in [0,1]. \tag{8.10}$$

Introduce the trace mgf of the interpolants:

$$u(s) := \mathbb{E}\,\operatorname{Tr} \mathrm{e}^{-\theta \mathbf{Y}_s} \quad \text{for } s \in [0,1].$$

Note that $u(0) = \operatorname{mgf}_{\mathbf{Z}}(\theta)$ and $u(1) = \operatorname{mgf}_{\mathbf{X}}(\theta)$. We claim that the derivative $u'(s) \le 0$ for $s \in (0,1)$. Once this point is settled, Proposition 7.4 follows inexorably.

By a scaling argument, we can assume that $\theta = 1$. Now, for fixed $s \in (0,1)$, the derivative $u'(s)$ along the interpolation path takes the form

$$u'(s) = \frac{-1}{2\sqrt{s}}\,\mathbb{E}\,\operatorname{Tr}\big[(\mathbf{X} - \mathbb{E}\,\mathbf{X})\mathrm{e}^{-\mathbf{Y}_s}\big] - \frac{-1}{2\sqrt{1-s}}\,\mathbb{E}\,\operatorname{Tr}\big[\mathbf{Z}\mathrm{e}^{-\mathbf{Y}_s}\big] =: ① - ②. \tag{8.11}$$

We start with the second term ②, involving the Gaussian random matrix, as it offers a template for how to bound the first term ①.

To handle the term ② from (8.11), we invoke Gaussian integration by parts. In preparation, expand the random matrix $\mathbf{Z} = \sum_{i=1}^n \gamma_i \mathbf{H}_i$ as a sum:

$$\begin{aligned} ② &= \frac{-1}{2\sqrt{1-s}}\,\mathbb{E}\,\operatorname{Tr}\big[\mathbf{Z}\mathrm{e}^{-\mathbf{Y}_s}\big] = \frac{-1}{2\sqrt{1-s}} \sum_{i=1}^n \mathbb{E}\big[\gamma_i \operatorname{Tr}\big[\mathbf{H}_i \mathrm{e}^{-\mathbf{Y}_s}\big]\big] \\ &= \frac{1}{2}\sum_{i=1}^n \int_0^1 \mathbb{E}\,\operatorname{Tr}\big[\mathbf{H}_i \mathrm{e}^{-\tau\mathbf{Y}_s}\mathbf{H}_i \mathrm{e}^{-(1-\tau)\mathbf{Y}_s}\big]\,\mathrm{d}\tau \\ &= \frac{1}{2}\sum_{i=1}^n \mathbb{E}[W_i^2] \cdot \int_0^1 \mathbb{E}\,\operatorname{Tr}\big[\mathbf{A}_i \mathrm{e}^{-\tau\mathbf{Y}_s}\mathbf{A}_i \mathrm{e}^{-(1-\tau)\mathbf{Y}_s}\big]\,\mathrm{d}\tau. \end{aligned}$$

The second line is the Gaussian IBP rule (Fact 8.4). This calculation relies on Duhamel's formula (8.8) for the derivative of the matrix exponential. The partial derivative $\partial_{\gamma_i}\mathbf{Y}_s = \sqrt{1-s}\,\mathbf{H}_i$, owing to the expressions (8.10) for $\mathbf{Y}_s$ and (8.3) for $\mathbf{Z}$. For the last step, recall the definition (8.2) of the matrix coefficients $\mathbf{H}_i$.

To obtain a bound for the term ① in (8.11), we plan to exploit Lemma 7.3. Expand the centered random matrix $\mathbf{X} - \mathbb{E}\,\mathbf{X} = \sum_{i=1}^n (W_i - \mathbb{E}\,W_i)\mathbf{A}_i$ as a sum:

$$\begin{aligned} ① &= \frac{-1}{2\sqrt{s}}\,\mathbb{E}\,\operatorname{Tr}\big[(\mathbf{X} - \mathbb{E}\,\mathbf{X})\mathrm{e}^{-\mathbf{Y}_s}\big] = \frac{-1}{2\sqrt{s}} \sum_{i=1}^n \mathbb{E}\,\operatorname{Tr}\big[(W_i - \mathbb{E}\,W_i)\mathbf{A}_i \mathrm{e}^{-\mathbf{Y}_s}\big] \\ &=: \frac{1}{2s}\sum_{i=1}^n \mathbb{E}\,\operatorname{Tr}\big[-\sqrt{s}(W_i - \mathbb{E}\,W_i)\mathbf{A}_i \mathrm{e}^{\mathbf{B}_i - \sqrt{s}W_i\mathbf{A}_i}\big]. \end{aligned}$$

For each index i, we have defined the random matrix $\mathbf{B}_i := \sqrt{s}W_i\mathbf{A}_i - \mathbf{Y}_s$. Observe that $\mathbf{B}_i$ is statistically independent from the random variable W_i.

Conditional on $\mathbf{B}_i$, introduce the (deterministic) functions

$$f_i(w) \coloneqq \operatorname{Tr} \mathrm{e}^{\mathbf{B}_i - \sqrt{s}w\mathbf{A}_i} \quad \text{for } w \geq 0 \text{ and } i = 1, \dots, n.$$

Stahl's theorem (Fact 8.2) asserts that each f_i is completely monotone for fixed self-adjoint $\mathbf{A}_i, \mathbf{B}_i$ with $\mathbf{A}_i$ psd. In light of (8.8) and (8.9), the first two derivatives of f_i satisfy

$$f_i'(w) = -\sqrt{s} \cdot \operatorname{Tr}\left[\mathbf{A}_i \mathrm{e}^{\mathbf{B}_i - \sqrt{s}w\mathbf{A}_i}\right];$$
$$f_i''(w) = s \cdot \int_0^1 \operatorname{Tr}\left[\mathbf{A}_i \mathrm{e}^{\tau(\mathbf{B}_i - \sqrt{s}w\mathbf{A}_i)} \mathbf{A}_i \mathrm{e}^{(1-\tau)(\mathbf{B}_i - \sqrt{s}w\mathbf{A}_i)}\right] \mathrm{d}\tau.$$

We may now express ① in terms of the functions f_i. This revision allows us to invoke Lemma 7.3, conditional on $\mathbf{B}_i$. We arrive at the inequality

$$① = \frac{1}{2s} \sum_{i=1}^n \mathbb{E}\left[(W_i - \mathbb{E}\, W_i) f_i'(W_i)\right] \leq \frac{1}{2s} \sum_{i=1}^n \mathbb{E}[W_i^2] \cdot \mathbb{E}[f_i''(W_i)]$$
$$= \frac{1}{2} \sum_{i=1}^n \mathbb{E}[W_i^2] \cdot \int_0^1 \mathbb{E} \operatorname{Tr}\left[\mathbf{A}_i \mathrm{e}^{-\tau \mathbf{Y}_s} \mathbf{A}_i \mathrm{e}^{-(1-\tau)\mathbf{Y}_s}\right] \mathrm{d}\tau = ②.$$

The last line depends on the relations $\mathbf{B}_i - \sqrt{s}W_i\mathbf{A}_i = -\mathbf{Y}_s$. As promised, $u'(s) = ① - ② \leq 0$. □

8.4. **Extension: Polynomial moments.** The proof of Proposition 8.5 can be adapted to obtain a comparison theorem for one-sided polynomial moments.

Proposition 8.6 (Randomly weighted sum: Polynomial moment bound)**.** *Instate the hypotheses of Theorem* 8.1. *For each* $p \geq 4$,

$$\mathbb{E} \operatorname{Tr} (\mathbf{X} - \mathbb{E}\,\mathbf{X} + \mathbf{\Delta})_-^p \leq \mathbb{E} \operatorname{Tr} (\mathbf{Z} + \mathbf{\Delta})_-^p.$$

As usual, the negative part $(a)_- \coloneqq \max\{-a, 0\}$ *for* $a \in \mathbb{R}$ *binds before the power.*

Sketch of proof. The structure of the argument is identical with the proof of Proposition 8.5. For justification, when $p \geq 4$, note that $h : w \mapsto (w)_-^p$ is completely monotone to order four on the real line. Heineväara's theorem (Fact 8.3) ensures that the associated trace function $f(w) \coloneqq \operatorname{Tr} (w\mathbf{A} - \mathbf{B})_-^p$ is also completely monotone to order four. Therefore, we can activate Lemma 7.3.

While it is not really necessary to compute the derivatives of the trace function f explicitly, one may employ the Daleckiĭ–Kreĭn formula [Bha97, Thm. V.3.3] to obtain the detailed expressions. □

9. Gaussian comparison: Sum of iid random psd matrices

This section develops a comparison theorem for the minimum eigenvalue of a sum of iid random psd matrices. This formulation includes covariance matrices and related models. Surprisingly, this result is a consequence of the comparison (Theorem 8.1) for sums of randomly weighted psd matrices.

Let $\mathbf{W}$ be a random psd matrix with dimension d. Assume that $\mathbf{W}$ has two finite moments: $\mathbb{E}\,\|\mathbf{W}\|^2 < +\infty$. For a fixed natural number $k \in \mathbb{N}$, consider the random psd

matrix $\mathbf{Y}$ obtained by adding k independent copies of $\mathbf{W}$. That is,

$$\mathbf{Y} \coloneqq \sum_{j=1}^{k} \mathbf{W}_j \quad \text{where } \mathbf{W}_j \sim \mathbf{W} \text{ iid.} \tag{9.1}$$

We compare the random psd matrix $\mathbf{Y}$ with an appropriate Gaussian model. Consider a centered Gaussian matrix that follows the distribution

$$\mathbf{Z} \sim \text{NORMAL}(\mathbf{0}, \mathsf{V}) \quad \text{where} \quad \mathsf{V} \coloneqq k \cdot \mathsf{Mom}[\mathbf{W}]. \tag{9.2}$$

With these definitions, we can state another comparison theorem.

Theorem 9.1 (Gaussian comparison: Sum of iid psd matrices). *Fix an arbitrary self-adjoint matrix $\mathbf{\Delta} \in \mathbb{H}_d$. Introduce the random d-dimensional matrices $\mathbf{Y}$ and $\mathbf{Z}$ from* (9.1) *and* (9.2)*. Then*

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Y} - \mathbb{E}\,\mathbf{Y} + \mathbf{\Delta}) \geq \mathbb{E}\,\lambda_{\min}(\mathbf{Z} + \mathbf{\Delta}) - \sqrt{2\sigma_*^2(\mathbf{Z}) \log(2d)}.$$

In addition, for all $t \geq 0$,

$$\mathbb{P}\{\lambda_{\min}(\mathbf{Y} - \mathbb{E}\,\mathbf{Y} + \mathbf{\Delta}) \leq \mathbb{E}\,\lambda_{\min}(\mathbf{Z} + \mathbf{\Delta}) - t\} \leq 2d \cdot \mathrm{e}^{-t^2/(2\sigma_*^2(\mathbf{Z}))}.$$

The weak variance $\sigma_*^2(\mathbf{Z})$ is defined in (3.8)*.*

Proof of Theorem 9.1. The content of the argument is a comparison inequality for the trace mgfs:

$$\mathrm{mgf}_{\mathbf{Y}}(\theta) \coloneqq \mathbb{E}\,\mathrm{Tr}\,\mathrm{e}^{-\theta(\mathbf{Y} - \mathbb{E}\,\mathbf{Y} + \mathbf{\Delta})} \leq 2\,\mathbb{E}\,\mathrm{Tr}\,\mathrm{e}^{-\theta(\mathbf{Z} + \mathbf{\Delta})} \quad \text{for } \theta \geq 0. \tag{9.3}$$

See (9.9). The rest of the argument follows the same route as the proof of Theorem 8.1. □

Proof of Theorem 2.4 *from Theorem* 9.1. To derive the result in Section 2.3, we choose the shift $\mathbf{\Delta} = \mathbb{E}\,\mathbf{Y}$, and we employ monotonicity properties of the Gaussian distribution to relax the assumption on the variance function.

Here are the details. Construct the Gaussian distribution $\mathbf{Z}' \sim \text{NORMAL}(\mathbf{\Delta}, \mathsf{V}')$ where $\mathbf{\Delta} = \mathbb{E}\,\mathbf{Y}$ and the variance function $\mathsf{V}' \geq \mathsf{V} = k \cdot \mathsf{Mom}[\mathbf{W}]$. Invoke Theorem 9.1 with the centered Gaussian matrix $\mathbf{Z} \sim \text{NORMAL}(\mathbf{0}, \mathsf{V})$ to obtain the expectation bound

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Y}) \geq \mathbb{E}\,\lambda_{\min}(\mathbf{Z} + \mathbf{\Delta}) - \sqrt{2\sigma_*^2(\mathbf{Z}) \log(2d)}.$$

Gaussian monotonicity (Proposition 3.2) for the concave function $\mathbf{A} \mapsto \lambda_{\min}(\mathbf{A} + \mathbf{\Delta})$ and monotonicity of the weak variance (3.9) imply that

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Z} + \mathbf{\Delta}) \geq \mathbb{E}\,\lambda_{\min}(\mathbf{Z}') \quad \text{and} \quad \sigma_*^2(\mathbf{Z}) \leq \sigma_*^2(\mathbf{Z}').$$

Combine the last two displays and adjust the notation to reach the expectation bound (2.4) in Theorem 2.4. The proof of the tail bound (2.5) is similar. □

Remark 9.2 (Non-iid case). To obtain the expectation bound for non-iid sums stated in Conjecture 2.5, we simply iterate the expectation bound from Theorem 9.1 to substitute a Gaussian matrix for one summand at a time. At each step, we use the shift matrix $\mathbf{\Delta}$ to capture the contributions of the remaining random matrices.

9.1. **Proof of the trace mgf bound: Overview.** The trace mgf bound (9.3) for an iid sum is a nontrivial corollary of the trace mgf bound for a randomly weighted sum (Proposition 8.5). Our strategy depends on an empirical approximation of the distribution of the iid sum $\mathbf{Y}$. We draw and fix a large sample from the distribution of the summand $\mathbf{W}$, and we form a proxy $\widehat{\mathbf{Y}}$ for the random matrix $\mathbf{Y}$ by adding exactly k of the sample points. The empirical approximation $\widehat{\mathbf{Y}}$ is a randomly weighted sum of fixed psd matrices with dependent coefficients. Via the classic Poissonization technique, we can decouple the coefficients to arrive at a randomly weighted sum with independent coefficients. Proposition 8.5 allows us to compare the trace mgf of the independent model with a Gaussian distribution. Related arguments have appeared in the probability literature; for example, see [Tal21, Sec. 11.11].

Remark 9.3 (Non-iid case). We cannot extend this proof strategy to address the non-iid case because each application of Poissonization introduces a factor of two into the trace mgf bound; cf. (9.3). After replacing each of n non-identical summands by its Gaussian proxy, we have increased the trace mgf by an intolerable factor of 2^n.

9.2. **Step 1: Empirical approximation.** To lighten notation, assume that the shift $\mathbf{\Delta} \in \mathbb{H}_d$ equals the zero matrix. The proof for a general shift is no different.

For a large parameter $n \in \mathbb{N}$, draw a sample $\mathcal{A}_n \coloneqq (\mathbf{A}_1, \dots, \mathbf{A}_n)$ where the matrices $\mathbf{A}_i$ are iid copies of $\mathbf{W}$. The sampled matrices may not be distinct. Until the last steps of the proof, we regard the sample $\mathcal{A}_n$ as fixed.

Consider a random matrix $\widehat{\mathbf{W}}$ that takes a uniformly random value from the fixed sample $\mathcal{A}_n$:

$$\widehat{\mathbf{W}} = \mathbf{A}_I \quad \text{where} \quad I \sim \textsc{uniform}\{1, \dots, n\}.$$

We construct a proxy $\widehat{\mathbf{Y}}_n$ for the iid sum $\mathbf{Y}$ by forming an iid sum of copies of $\widehat{\mathbf{W}}$.

$$\widehat{\mathbf{Y}}_n \coloneqq \sum_{j=1}^{k} \widehat{\mathbf{W}}_j \quad \text{where } \widehat{\mathbf{W}}_j \sim \widehat{\mathbf{W}}.$$

As a heuristic, when the number n of sample points is large, the distribution of the proxy $\widehat{\mathbf{Y}}_n$ is close to the distribution of the original sum $\mathbf{Y}$. Lemma 9.8 justifies this claim.

9.3. **Step 2: Multinomial model.** To analyze the proxy $\widehat{\mathbf{Y}}_n$, we work with an alternative representation. The k independent summands in the proxy take the form

$$\widehat{\mathbf{W}}_j = \mathbf{A}_{I_j} \quad \text{where } I_j \sim \textsc{uniform}\{1, \dots, n\} \text{ iid for } j = 1, \dots, k.$$

For each index $i = 1, \dots, n$, define a scalar random variable δ_i that counts the number of the I_j that select the index i. That is,

$$\delta_i \coloneqq \#\big\{j \in \{1, \dots, k\} : I_j = i\big\}.$$

As a consequence, $\boldsymbol{\delta} \coloneqq (\delta_1, \dots, \delta_n) \sim \textsc{multinomial}(k, n)$ follows the multinomial distribution of k balls placed independently and uniformly at random in n bins. Recall that

$$\delta_i \sim \textsc{binomial}(1/n, k) \quad \text{and} \quad \sum_{i=1}^{n} \delta_i = k.$$

Because of the coupling between the distributions,

$$\widehat{\mathbf{Y}}_n = \sum_{j=1}^{k} \widehat{\mathbf{W}}_j = \sum_{i=1}^{n} \delta_i \mathbf{A}_i. \tag{9.4}$$

Let us emphasize that the coefficient vector $\boldsymbol{\delta}$ is independent from the sample $\mathcal{A}_n$.

9.4. **Step 3: Poissonization.** The coefficients δ_i in the representation (9.4) are not independent, but we can pass to a model that has independent coefficients:

$$\mathbf{X}_n \coloneqq \sum_{i=1}^{n} Q_i \mathbf{A}_i \quad \text{where } Q_i \sim \textsc{poisson}(k/n) \text{ iid.} \tag{9.5}$$

The Poisson variables Q_i are independent from each other and from the multinomial variables δ_i. Since $\mathbb{E}\, Q_i = \mathbb{E}\, \delta_i$, the conditional expectations of the random matrices $\mathbf{X}_n$ and $\widehat{\mathbf{Y}}_n$ coincide.

$$\mathbb{E}\left[\mathbf{X}_n \,|\, \mathcal{A}_n\right] = \mathbb{E}\left[\widehat{\mathbf{Y}}_n \,|\, \mathcal{A}_n\right] = \mathbb{E}\left[k\widehat{\mathbf{W}} \,|\, \mathcal{A}_n\right]. \tag{9.6}$$

Standard arguments quickly lead to a comparison between the two random models $\widehat{\mathbf{Y}}_n$ and $\mathbf{X}_n$.

Lemma 9.4 (Poissonization). *Consider the random matrices $\widehat{\mathbf{Y}}_n$ and $\mathbf{X}_n$ defined in* (9.4) *and* (9.5). *For all $\theta \geq 0$,*

$$\mathbb{E}\left[\operatorname{Tr} \mathrm{e}^{-\theta(\widehat{\mathbf{Y}}_n - \mathbb{E}[\widehat{\mathbf{Y}}_n \,|\, \mathcal{A}_n])} \,|\, \mathcal{A}_n\right] \leq 2\, \mathbb{E}\left[\operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{X}_n - \mathbb{E}[\mathbf{X}_n \,|\, \mathcal{A}_n])} \,|\, \mathcal{A}_n\right].$$

Proof. With the sample $\mathcal{A}_n$ fixed, consider the deterministic, positive function

$$f(s_1, \dots, s_n) \coloneqq \operatorname{Tr} \exp\left(-\theta\left(\sum_{i=1}^{n} s_i \mathbf{A}_i - \mathbb{E}\left[k\widehat{\mathbf{W}} \,|\, \mathcal{A}_n\right]\right)\right) \quad \text{for } (s_1, \dots, s_n) \in \mathbb{R}_+^n.$$

According to (9.6), the conditional expectation in the definition of f coincides with the conditional expectation of both $\widehat{\mathbf{Y}}_n$ and $\mathbf{X}_n$. Thus, we obtain the trace exponential with $\widehat{\mathbf{Y}}_n$ by evaluating $f(\delta_1, \dots, \delta_n)$, and we obtain the trace exponential with $\mathbf{X}_n$ by evaluating $f(Q_1, \dots, Q_n)$.

The key insight is that the function f is monotone decreasing with respect to each argument s_i because each matrix $\mathbf{A}_i$ is psd. This is a well-established and easy fact [Car10, Thm. 2.10], and it is also contained in Stahl's theorem (Fact 8.2).

For completeness, we include the familiar argument [MU17, Sec. 5.4] that leads to the conclusion of the lemma. Recall that $(\delta_1, \dots, \delta_n) \sim \textsc{multinomial}(k, n)$ and that $Q_i \sim \textsc{poisson}(k/n)$ iid for $i = 1, \dots, n$. Define the sum $Q \coloneqq \sum_{i=1}^{n} Q_i$ of the Poisson variables, and note that $Q \sim \textsc{poisson}(k)$. Conditional on the sample $\mathcal{A}_n$, calculate that

$$\begin{aligned}
\mathbb{E}[f(Q_1, \dots, Q_n) \,|\, \mathcal{A}_n] &\geq \sum_{\ell=0}^{k} \mathbb{E}[f(Q_1, \dots, Q_n) \,|\, \mathcal{A}_n, Q = \ell] \cdot \mathbb{P}\{Q = \ell\} \\
&\geq \mathbb{E}[f(Q_1, \dots, Q_n) \,|\, \mathcal{A}_n, Q = k] \cdot \sum_{\ell=0}^{k} \mathbb{P}\{Q = \ell\} \\
&\geq \frac{1}{2}\, \mathbb{E}[f(\delta_1, \dots, \delta_n) \,|\, \mathcal{A}_n].
\end{aligned}$$

The first inequality follows from the law of total expectation and the positivity of f. The second inequality depends on the monotonicity of f. The third inequality holds

because k is a median of the POISSON(k) distribution [Cho94, Thm. 2]. Last, conditioning the Poisson variables on the event $\{Q = k\}$ produces the multinomial distribution [MU17, Thm. 5.6]. □

Remark 9.5 (Poissonization). The analog of Lemma 9.4 holds if we replace the trace exponential with any other trace function that is both positive and monotone. In particular, it applies to the one-sided power function $\mathbf{M} \mapsto \operatorname{Tr}(\mathbf{M})^{\underline{p}}$ for $p > 0$.

9.5. **Step 4: Comparison with the Gaussian model.** We may now invoke the existing trace mgf bound (Proposition 8.5) to compare the independent model $\mathbf{X}_n$ with a suitable Gaussian matrix.

Conditional on the sample $\mathcal{A}_n$, define the variance function

$$\mathsf{V}_n(\mathbf{M}) := \sum_{i=1}^{n} (\mathbb{E}\, Q_i^2) \cdot |\langle \mathbf{M},\ \mathbf{A}_i \rangle|^2 = \left(1 + \frac{k}{n}\right) \frac{k}{n} \sum_{i=1}^{n} |\langle \mathbf{M},\ \mathbf{A}_i \rangle|^2 \quad \text{for } \mathbf{M} \in \mathbb{H}_d.$$

Indeed, since $Q_i \sim$ POISSON(k/n), its second moment is $(1 + k/n)(k/n)$. Construct the centered (conditionally) Gaussian random matrix

$$\mathbf{Z}_n \sim \text{NORMAL}(\mathbf{0}, \mathsf{V}_n). \tag{9.7}$$

By the law of large numbers, when the number n of samples is large, $\mathsf{V}_n \approx \mathsf{V}$, where V is defined in (9.2). As a consequence, the distribution of $\mathbf{Z}_n$ is close to the distribution of the original Gaussian model $\mathbf{Z}$ with covariance V. Lemma 9.9 fully justifies this claim.

To compare the trace mgfs of the proxy $\widehat{\mathbf{Y}}_n$ and the Gaussian matrix $\mathbf{Z}_n$, first apply the Poissonization result (Lemma 9.4). Then invoke the trace mgf bound (Proposition 8.5), conditional on $\mathcal{A}_n$. We arrive at the comparison

$$\begin{aligned} \mathbb{E}\left[\operatorname{Tr} \mathrm{e}^{-\theta(\widehat{\mathbf{Y}}_n - \mathbb{E}[\widehat{\mathbf{Y}}_n \,|\, \mathcal{A}_n])} \,\middle|\, \mathcal{A}_n \right] &\le 2\, \mathbb{E}\left[\operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{X}_n - \mathbb{E}[\mathbf{X}_n \,|\, \mathcal{A}_n])} \,\middle|\, \mathcal{A}_n \right] \\ &\le 2\, \mathbb{E}\left[\operatorname{Tr} \mathrm{e}^{-\theta \mathbf{Z}_n} \,\middle|\, \mathcal{A}_n \right]. \end{aligned} \tag{9.8}$$

It remains to relate the random matrices $\widehat{\mathbf{Y}}_n$ and $\mathbf{Z}_n$ to the target models $\mathbf{Y}$ and $\mathbf{Z}$.

Remark 9.6 (Sparse sampling). We are employing Proposition 8.5 in a situation where each random weight Q_i places most of its mass at zero. This is precisely the setting where we anticipate that the result is most accurate.

9.6. **Step 5: Limits.** At this stage, we can unfreeze the random sample $\mathcal{A}_n$ and take limits. This process will produce the bound

$$\mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{Y} - \mathbb{E}\,\mathbf{Y})} \le 2\, \mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta \mathbf{Z}}. \tag{9.9}$$

The random matrices $\mathbf{Y}$ and $\mathbf{Z}$ are defined in (9.1) and (9.2). The remaining steps leading to this result are technical.

First, Lemma 9.7 shows that it is enough to prove (9.9) under the additional assumption that the random summand $\mathbf{W}$ is bounded. In this case, we may calculate that

$$\begin{aligned} \mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{Y} - \mathbb{E}\,\mathbf{Y})} &= \lim\nolimits_{n\to\infty} \mathbb{E}\left[\mathbb{E}\left[\operatorname{Tr} \mathrm{e}^{-\theta(\widehat{\mathbf{Y}}_n - \mathbb{E}[\widehat{\mathbf{Y}}_n \,|\, \mathcal{A}_n])} \,\middle|\, \mathcal{A}_n \right] \right] \\ &\le 2 \lim\nolimits_{n\to\infty} \mathbb{E}\left[\mathbb{E}\left[\operatorname{Tr} \mathrm{e}^{-\theta \mathbf{Z}_n} \,\middle|\, \mathcal{A}_n \right] \right] = 2\, \mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta \mathbf{Z}}. \end{aligned}$$

The first limit follows from Lemma 9.8. The second relation is the mgf bound (9.8). The second limit follows from Lemma 9.9. The details of these computations occupy the upcoming subsections.

9.7. **Technical step 6: Truncation.** To continue, we restrict our attention to the setting where the random summand $\mathbf{W}$ is bounded in norm.

Lemma 9.7 (Truncation)**.** *Suppose that* (9.9) *holds when the ℓ_2 operator norm $\|\mathbf{W}\|$ is uniformly bounded. Then* (9.9) *remains valid when $\mathbb{E}\,\|\mathbf{W}\|^2 < +\infty$.*

Proof. Suppose that $\mathbb{E}\,\|\mathbf{W}\|^2 < +\infty$. Let $R > 0$ be a truncation parameter. We will consider the random matrix models arising from the truncated summand: $\mathbf{W}\mathbb{1}\{\|\mathbf{W}\| \le R\}$.

For iid summands $\mathbf{W}_j \sim \mathbf{W}$, define the coupled random matrix models

$$\mathbf{Y} := \sum_{j=1}^{k} \mathbf{W}_j \quad \text{and} \quad \mathbf{Y}_{\wedge R} := \sum_{j=1}^{k} \mathbf{W}_j \mathbb{1}\{\|\mathbf{W}_j\| \le R\}.$$

Since the trace exponential is monotone with respect to the psd order [Car10, Thm. 2.10],

$$\operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{Y}_{\wedge R} - \mathbb{E}\,\mathbf{Y}_{\wedge R})} \le \operatorname{Tr} \mathrm{e}^{\theta\,\mathbb{E}\,\mathbf{Y}} < +\infty \quad \text{pointwise and for all } R > 0.$$

We have used the semidefinite relations $\mathbf{0} \preccurlyeq \mathbf{Y}_{\wedge R}$ and $\mathbb{E}\,\mathbf{Y}_{\wedge R} \preccurlyeq \mathbb{E}\,\mathbf{Y}$, where $\preccurlyeq$ is the *semidefinite partial order*, also known as the *Loewner partial order*. Bounded convergence yields

$$\lim_{R\to\infty} \mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{Y}_{\wedge R} - \mathbb{E}\,\mathbf{Y}_{\wedge R})} = \mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{Y} - \mathbb{E}\,\mathbf{Y})}.$$

Indeed, $\mathbf{Y}_{\wedge R} \to \mathbf{Y}$ pointwise, and so $\mathbb{E}\,\mathbf{Y}_{\wedge R} \to \mathbb{E}\,\mathbf{Y}$. The limit of the expectation is confirmed by applying monotone convergence to the quadratic forms induced by the random psd matrices.

As for the Gaussian model, define the variance functions for an argument $\mathbf{M} \in \mathbb{H}_d$:

$$\mathsf{V}(\mathbf{M}) := k \cdot \mathbb{E}\left[|\langle \mathbf{M},\ \mathbf{W}\rangle|^2\right] \quad \text{and} \quad \mathsf{V}_{\wedge R}(\mathbf{M}) := k \cdot \mathbb{E}\left[|\langle \mathbf{M},\ \mathbf{W}\rangle|^2 \cdot \mathbb{1}\{\|\mathbf{W}\| \le R\}\right].$$

It is easy to see that $\mathsf{V}_{\wedge R}(\mathbf{M}) \le \mathsf{V}(\mathbf{M})$. Therefore, the Gaussian random matrices $\mathbf{Z}_{\wedge R} \sim \textsc{normal}(\mathbf{0}, \mathsf{V}_{\wedge R})$ and $\mathbf{Z} \sim \textsc{normal}(\mathbf{0}, \mathsf{V})$ satisfy

$$\mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta \mathbf{Z}_{\wedge R}} \le \mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta \mathbf{Z}} \quad \text{for all } R > 0.$$

This statement follows from monotonicity (Proposition 3.2) for the expectation of a convex function of a Gaussian. Convexity of the trace exponential is an easy classical fact [Car10, Thm. 2.10], which is also contained in Stahl's theorem (Fact 8.2).

We conclude that

$$\begin{aligned} \mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{Y} - \mathbb{E}\,\mathbf{Y})} &= \limsup\nolimits_{R\to\infty} \mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{Y}_{\wedge R} - \mathbb{E}\,\mathbf{Y}_{\wedge R})} \\ &\le 2 \limsup\nolimits_{R\to\infty} \mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta \mathbf{Z}_{\wedge R}} \le 2\,\mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta \mathbf{Z}}. \end{aligned}$$

The inequality in the last display depends on the hypothesis of the lemma, namely that the relation (9.9) holds when $\|\mathbf{W}\|$ is uniformly bounded. □

9.8. **Technical step 7: Convergence of the empirical model.** Next, we must verify that the empirical approximations $\widehat{\mathbf{Y}}_n$ converge weakly to the original random matrix model $\mathbf{Y}$.

Lemma 9.8 (Empirical approximation: Weak convergence). *Assume that the random summand has two finite moments:* $\mathbb{E}\,\|\mathbf{W}\|^2 < +\infty$. *Define random matrices* $\mathbf{Y}$ *and* $\widehat{\mathbf{Y}}_n$ *as in* (9.1) *and* (9.4).

(1) *For each bounded, Lipschitz function* $h : \mathbb{H}_d \to \mathbb{R}$,

$$\mathbb{E}\, h\big(\widehat{\mathbf{Y}}_n - \mathbb{E}\big[\widehat{\mathbf{Y}}_n \,|\, \mathcal{A}_n\big]\big) \to \mathbb{E}\, h(\mathbf{Y} - \mathbb{E}\,\mathbf{Y}) \quad \text{as } n \to \infty. \tag{9.10}$$

(2) *If* $\|\mathbf{W}\|$ *is uniformly bounded, the limit* (9.10) *also holds for* $h(\mathbf{M}) := \operatorname{Tr} \mathrm{e}^{-\theta \mathbf{M}}$ *with* $\theta \in \mathbb{R}$.

Lipschitz functions are defined with respect to the Frobenius norm on self-adjoint matrices. The expectations average over everything, including the random sample $\mathcal{A}_n$.

Proof. To begin, let us explore some properties of the empirical approximation $\widehat{\mathbf{Y}}_n$. The representation (9.4) shows that

$$\widehat{\mathbf{Y}}_n = \sum_{i=1}^{n} \delta_i^{(n)} \mathbf{A}_i \quad \text{where } \boldsymbol{\delta}^{(n)} \sim \textsc{multinomial}(k, n).$$

The multinomial coefficients $\boldsymbol{\delta}^{(n)} := \big(\delta_1^{(n)}, \dots, \delta_n^{(n)}\big)$ are independent from the sample $\mathcal{A}_n$. Define the event D_n where the multinomial selects k distinct summands:

$$\mathsf{D}_n := \big\{ \#\operatorname{supp}\big(\boldsymbol{\delta}^{(n)}\big) = k \big\}.$$

For large sample size n, it is likely that D_n occurs. By the birthday paradox argument [MU17, Sec. 5.1],

$$\mathbb{P}(\mathsf{D}_n) = \prod_{j=1}^{k} \left(1 - \frac{j-1}{n}\right) \geq 1 - \sum_{j=1}^{k} \frac{j-1}{n} > 1 - \frac{k^2}{n}. \tag{9.11}$$

Conditional on the event D_n occurring, the distribution of the empirical approximation $\widehat{\mathbf{Y}}_n$ is the same as the distribution $\mathbf{Y}$. More precisely, for each Borel set $\mathsf{B} \subseteq \mathbb{H}_d$,

$$\mathbb{P}\big\{\widehat{\mathbf{Y}}_n \in \mathsf{B} \,|\, \mathsf{D}_n\big\} = \mathbb{P}\left\{ \sum_{i \in \operatorname{supp}(\boldsymbol{\delta}^{(n)})} \mathbf{A}_i \in \mathsf{B} \,|\, \mathsf{D}_n \right\} = \mathbb{P}\left\{ \sum_{j=1}^{k} \mathbf{W}_j \in \mathsf{B} \right\} = \mathbb{P}\{\mathbf{Y} \in \mathsf{B}\}.$$

Indeed, each sample $\mathbf{A}_i$ is an independent draw from the distribution $\mathbf{W}$, as are the random matrices $\mathbf{W}_1, \dots, \mathbf{W}_k$. This argument formalizes the intuition that the empirical approximation is a good proxy for the original random matrix.

Next, we turn to the conditional expectation of the empirical approximation. Since each coefficient $\delta_i \sim \textsc{binomial}(1/n, k)$,

$$\mathbb{E}\big[\widehat{\mathbf{Y}}_n \,|\, \mathcal{A}_n\big] = \frac{k}{n} \sum_{i=1}^{n} \mathbf{A}_i.$$

Each sample $\mathbf{A}_i$ is an independent copy of $\mathbf{W}$, so its expectation satisfies $\mathbb{E}\,\mathbf{A}_i = \mathbb{E}\,\mathbf{W} = k^{-1}\,\mathbb{E}\,\mathbf{Y}$. By orthogonality,

$$\mathbb{E}\,\big\|\mathbb{E}\big[\widehat{\mathbf{Y}}_n \,|\, \mathcal{A}_n\big] - \mathbb{E}\,\mathbf{Y}\big\|_{\mathrm{F}}^2 = \mathbb{E}\,\left\|\frac{k}{n}\sum_{i=1}^{n}(\mathbf{A}_i - \mathbb{E}\,\mathbf{A}_i)\right\|_{\mathrm{F}}^2 = \frac{k^2}{n}\,\mathbb{E}\,\|\mathbf{W} - \mathbb{E}\,\mathbf{W}\|_{\mathrm{F}}^2 \to 0. \tag{9.12}$$

The limit occurs as $n \to \infty$, while k is a fixed number. We have used the fact that $\mathbf{W}$ has two moments.

Now, suppose that h has bounded Lipschitz norm L. In other words, both the uniform norm of h and the Lipschitz constant of h are at most L. Define the sequence of moments

$$E_n := \mathbb{E}\, h\big(\widehat{\mathbf{Y}}_n - \mathbb{E}\big[\widehat{\mathbf{Y}}_n \,|\, \mathcal{A}_n\big]\big).$$

Add and subtract the matrix $\mathbb{E}\,\mathbf{Y}$, and invoke the Lipschitz property:

$$E_n = \mathbb{E}\, h\big(\widehat{\mathbf{Y}}_n - \mathbb{E}\,\mathbf{Y}\big) \pm L \cdot \mathbb{E}\,\big\|\mathbb{E}\big[\widehat{\mathbf{Y}}_n \,|\, \mathcal{A}_n\big] - \mathbb{E}\,\mathbf{Y}\big\|_{\mathrm{F}}.$$

The notation $x = a \pm b$ is shorthand for the pair of inequalities $a - b \le x \le a + b$. From (9.12), we see that the second term on the right-hand side of the last display tends to zero. As for the first term, we compute the expectation by conditioning on the event D_n:

$$\begin{aligned}\mathbb{E}\, h\big(\widehat{\mathbf{Y}}_n - \mathbb{E}\,\mathbf{Y}\big) &= \mathbb{E}\big[h\big(\widehat{\mathbf{Y}}_n - \mathbb{E}\,\mathbf{Y}\big) \,|\, \mathsf{D}_n\big] \cdot \mathbb{P}(\mathsf{D}_n) + \mathbb{E}\big[h\big(\widehat{\mathbf{Y}}_n - \mathbb{E}\,\mathbf{Y}\big) \,|\, \mathsf{D}_n^{\mathsf{c}}\big] \cdot \mathbb{P}(\mathsf{D}_n^{\mathsf{c}}) \\ &= \mathbb{E}\big[h(\mathbf{Y} - \mathbb{E}\,\mathbf{Y})\big] \pm \frac{2k^2 L}{n} \to \mathbb{E}\, h(\mathbf{Y} - \mathbb{E}\,\mathbf{Y}).\end{aligned}$$

Indeed, the conditional distribution $\widehat{\mathbf{Y}}_n \,|\, \mathsf{D}_n \sim \mathbf{Y}$. The inequalities depend on two applications of the probability estimate (9.11) and the fact that the magnitude of h is uniformly bounded by L. Altogether,

$$E_n = \mathbb{E}\, h\big(\widehat{\mathbf{Y}}_n - \mathbb{E}\big[\widehat{\mathbf{Y}}_n \,|\, \mathcal{A}_n\big]\big) \to \mathbb{E}\, h(\mathbf{Y} - \mathbb{E}\,\mathbf{Y}) \quad \text{as } n \to \infty.$$

We have established the weak convergence claim.

Last, assume that the random summand satisfies $\|\mathbf{W}\| \le R$ for some $R > 0$. In this case, the random matrices $\widehat{\mathbf{Y}}_n$ and $\mathbf{Y}$ are all bounded in norm by kR. The trace exponential function $h(\mathbf{M}) = \operatorname{Tr} \mathrm{e}^{-\theta \mathbf{M}}$ is bounded and Lipschitz on the common support of these random matrices. Therefore, the weak convergence result (9.10) ensures that

$$\mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta(\widehat{\mathbf{Y}}_n - \mathbb{E}[\widehat{\mathbf{Y}}_n \,|\, \mathcal{A}_n])} \to \mathbb{E} \operatorname{Tr} \mathrm{e}^{-\theta(\mathbf{Y} - \mathbb{E}\,\mathbf{Y})} \quad \text{as } n \to \infty.$$

This is the second conclusion. □

9.9. **Technical step 8: Convergence of the Gaussian model.** Finally, we must argue that the sequence $\mathbf{Z}_n$ of Gaussian models converges weakly to the target Gaussian distribution $\mathbf{Z}$. The proof relies on characteristic functions.

Lemma 9.9 (Gaussian model: Weak convergence). *Assume that the random summand has two finite moments:* $\mathbb{E}\,\|\mathbf{W}\|^2 < +\infty$. *Define Gaussian matrices* $\mathbf{Z}$ *and* $\mathbf{Z}_n$ *as in* (9.2) *and* (9.7).

(1) *For each bounded Lipschitz function* $h : \mathbb{H}_d \to \mathbb{R}$,

$$\mathbb{E}\, h(\mathbf{Z}_n) \to \mathbb{E}\, h(\mathbf{Z}) \quad \text{as } n \to \infty. \tag{9.13}$$

(2) *If* $\|\mathbf{W}\|$ *is uniformly bounded, the limit* (9.13) *also holds for* $h(\mathbf{M}) := \operatorname{Tr} \mathrm{e}^{-\theta \mathbf{M}}$ *with* $\theta \in \mathbb{R}$.

The expectation averages over everything, including the random sample $\mathcal{A}_n$.

Proof. Consider a self-adjoint Gaussian matrix $\mathbf{G} \sim \textsc{normal}(\mathbf{0}, \mathsf{C})$. For a self-adjoint argument $\mathbf{M} \in \mathbb{H}_d$, the characteristic function $\chi_{\mathbf{G}}$ of the Gaussian $\mathbf{G}$ takes the form

$$\chi_{\mathbf{G}}(\mathbf{M}) := \mathbb{E}\, \mathrm{e}^{\mathrm{i}\langle \mathbf{M},\, \mathbf{G}\rangle} = \mathrm{e}^{-\mathsf{C}(\mathbf{M})/2}.$$

This just reinterprets the usual formula for the Gaussian characteristic function [Dud02, Sec. 9.5].

Now, the centered Gaussian matrices $\mathbf{Z}_n$ and $\mathbf{Z}$ have variance functions

$$\mathsf{V}_n(\mathbf{M}) \coloneqq \left(1 + \frac{k}{n}\right) \frac{k}{n} \sum_{i=1}^{n} |\langle \mathbf{M},\ \mathbf{A}_i \rangle|^2 \quad \text{and} \quad \mathsf{V}(\mathbf{M}) \coloneqq k \cdot \mathbb{E}\, |\langle \mathbf{M},\ \mathbf{W} \rangle|^2.$$

Since V_n is a random function that depends on the sample $\mathcal{A}_n$, the random matrix $\mathbf{Z}_n$ is actually a Gaussian mixture. Each random sample $\mathbf{A}_i$ is an independent copy of $\mathbf{W}$. Therefore,

$$\mathsf{V}_n(\mathbf{M}) = \left(1 + \frac{k}{n}\right) \frac{k}{n} \sum_{i=1}^{n} |\langle \mathbf{M},\ \mathbf{A}_i \rangle|^2 \to k \cdot \mathbb{E}\, |\langle \mathbf{M},\ \mathbf{W} \rangle|^2 \quad \text{almost surely.} \tag{9.14}$$

The formula (9.14) is just the strong law of large numbers [Dud02, Thm. 8.3.5]. The statement is justified by the fact that $\mathbb{E}\, \|\mathbf{W}\|^2$ is finite.

Recall that a sequence of random matrices converges weakly if and only if the characteristic functions converge pointwise to a limit that is continuous at the origin [Dud02, Thm. 9.8.2]. Therefore, to prove the weak convergence statement (9.13), it suffices to verify that

$$\chi_{\mathbf{Z}_n}(\mathbf{M}) \to \chi_{\mathbf{Z}}(\mathbf{M}) \quad \text{for each } \mathbf{M} \in \mathbb{H}_d.$$

Equivalently, we can obtain weak convergence from the limit

$$\mathbb{E}\, \mathrm{e}^{-\mathsf{V}_n(\mathbf{M})/2} \to \mathrm{e}^{-\mathsf{V}(\mathbf{M})/2} \quad \text{for each } \mathbf{M} \in \mathbb{H}_d.$$

But this statement follows instantly from bounded convergence and the almost sure limit (9.14). Indeed, variance functions are positive, so the exponentials are uniformly bounded by one.

Last, assume that the random summand satisfies $\|\mathbf{W}\| \le R$ for some $R > 0$. We must upgrade the weak convergence (9.13) to convergence for the trace mgf function $h(\mathbf{M}) \coloneqq \operatorname{Tr} \mathrm{e}^{-\theta \mathbf{M}}$ with $\theta \in \mathbb{R}$. This step requires asymptotic uniform integrability [vdV98, Thm. 2.20]. We claim that

$$\lim_{B \to \infty} \limsup_{n \to \infty} \mathbb{E}\left[\operatorname{Tr} \mathrm{e}^{-\theta \mathbf{Z}_n} \mathbb{1}\{\|\mathbf{Z}_n\| \ge B\} \right] \to 0. \tag{9.15}$$

Granted (9.15), we obtain the limit $\mathbb{E}\, h(\mathbf{Z}_n) \to \mathbb{E}\, h(\mathbf{Z})$, and the proof is complete.

Let us establish the claim. Since $\|\mathbf{W}\| \le R$, the variance functions V_n are uniformly bounded by a fixed variance function. More precisely, for $n \ge k$, since $\|\mathbf{A}_i\| \le R$ as well,

$$\mathsf{V}_n(\mathbf{M}) = \left(1 + \frac{k}{n}\right) \frac{k}{n} \sum_{i=1}^{n} |\langle \mathbf{M},\ \mathbf{A}_i \rangle|^2 \le 2kR^2 d \|\mathbf{M}\|_{\mathrm{F}}^2 =: \mathsf{C}(\mathbf{M}).$$

Define the Gaussian random matrix $\mathbf{G} \sim \textsc{normal}(\mathbf{0}, \mathsf{C})$. By monotonicity (Proposition 3.2), applied conditionally on $\mathcal{A}_n$, the comparison $\mathbb{E}\, f(\mathbf{Z}_n) \le \mathbb{E}\, f(\mathbf{G})$ holds for each convex function $f : \mathbb{H}_d \to \mathbb{R}$.

To confirm asymptotic uniform integrability (9.15), make a simple bound on the trace exponential. Then apply the inequalities of Cauchy–Schwarz and Markov to the expectation in (9.15) to obtain

$$\begin{aligned}
\mathbb{E}\left[\operatorname{Tr} \mathrm{e}^{-\theta \mathbf{Z}_n} \mathbb{1}\{\|\mathbf{Z}_n\| \ge B\} \right] &\le d \cdot \mathbb{E}\left[\mathrm{e}^{|\theta|\, \|\mathbf{Z}_n\|} \mathbb{1}\{\|\mathbf{Z}_n\| \ge B\} \right] \\
&\le d \cdot \left(\mathbb{E}\, \mathrm{e}^{2|\theta|\, \|\mathbf{Z}_n\|} \right)^{1/2} (B^{-1}\, \mathbb{E}\, \|\mathbf{Z}_n\|)^{1/2} \\
&\le d \cdot \left(\mathbb{E}\, \mathrm{e}^{2|\theta|\, \|\mathbf{G}\|} \right)^{1/2} (B^{-1}\, \mathbb{E}\, \|\mathbf{G}\|)^{1/2}.
\end{aligned}$$

The last inequality follows from the comparison of $\mathbf{Z}_n$ with $\mathbf{G}$. Since $\mathbf{G}$ is a particular Gaussian matrix with dimension d, the two expectations with respect to $\mathbf{G}$ are finite, and the asymptotic uniform integrability condition (9.15) is valid. □

9.10. **Extension: Polynomial moments.** The proof of the trace mgf comparison (9.9) can be adapted to obtain a comparison theorem for polynomial moments.

Proposition 9.10 (iid psd sum: Polynomial moment bound). *Instate the hypotheses of Theorem* 9.1. *For each* $p \geq 4$,

$$\mathbb{E}\,\mathrm{Tr}\,(\mathbf{Y} - \mathbb{E}\,\mathbf{Y} + \mathbf{\Delta})_+^p \leq 2\,\mathbb{E}\,\mathrm{Tr}\,(\mathbf{Z} + \mathbf{\Delta})_+^p.$$

Sketch of proof. Note that the trace function $f(w) \coloneqq \mathrm{Tr}\,(w\mathbf{A} - \mathbf{B})_+^p$ is completely monotone of order four when $p \geq 4$. Therefore, we can activate the Poissonization result (Remark 9.5) and the polynomial moment bound for weighted sums (Proposition 8.6). These are the main changes, as compared with the proof of (9.9). There are also some minor technical differences in the proofs of Lemmas 9.7 to 9.9. □

Appendix A. Matrix concentration for psd sums

For completeness, we include a short proof of a matrix concentration inequality for the lower tail of a psd sum. This result was communicated to the author by Andreas Maurer in 2011, but there does not seem to be a citation available.

Fact A.1 (Matrix concentration: Exponential Paley–Zygmund). Consider an independent family $(\mathbf{W}_1, \dots, \mathbf{W}_n)$ of random psd matrices with common dimension d and with two finite moments. Define the matrix sum and the sum of second moments:

$$\mathbf{Y} \coloneqq \sum_{i=1}^n \mathbf{W}_i \quad \text{and} \quad L_2 \coloneqq \Big\| \sum_{i=1}^n \mathbb{E}[\mathbf{W}_i^2] \Big\|.$$

Then

$$\mathbb{E}\,\lambda_{\min}(\mathbf{Y}) \geq \lambda_{\min}(\mathbb{E}\,\mathbf{Y}) - \sqrt{2L_2 \log d}. \tag{A.1}$$

In addition, for $t \geq 0$,

$$\mathbb{P}\{\lambda_{\min}(\mathbf{Y}) \leq \lambda_{\min}(\mathbb{E}\,\mathbf{Y}) - t\} \leq d \cdot \mathrm{e}^{-t^2/(2L_2)}. \tag{A.2}$$

To identify the connection between Fact A.1 and the Paley–Zygmund inequality [BLM13, Exer. 2.4], select $t = \varepsilon \cdot \lambda_{\min}(\mathbb{E}\,\mathbf{Y})$ in the tail bound (A.2). We obtain a ratio of the squared norm of the first moment to the norm of the sum of second moments. The proof depends on a trace mgf bound.

Lemma A.2 (Exponential Paley–Zygmund: log-mgf bound). *Let* $\mathbf{W}$ *be a random psd matrix with two finite moments. For* $\theta \geq 0$, *we have the semidefinite relation*

$$\log \mathbb{E}\,\mathrm{e}^{-\theta \mathbf{W}} \preccurlyeq -\theta\,\mathbb{E}[\mathbf{W}] + \tfrac{1}{2}\theta^2\,\mathbb{E}[\mathbf{W}^2].$$

Proof. Recall the numerical inequality $\mathrm{e}^{-a} \leq 1 - a + a^2/2$, valid for $a \geq 0$. By the transfer rule [Tro15, Prop. 2.1.4], the inequality extends to matrices:

$$\mathbb{E}\,\mathrm{e}^{-\theta \mathbf{W}} \preccurlyeq \mathbb{E}[\mathbf{I} - \theta\mathbf{W} + \tfrac{1}{2}\theta^2\mathbf{W}^2] = \mathbf{I} - \theta\,\mathbb{E}[\mathbf{W}] + \tfrac{1}{2}\theta^2\,\mathbb{E}[\mathbf{W}^2] \quad \text{for } \theta \geq 0.$$

Since the logarithm is matrix monotone [Tro15, Prop. 8.4.4], we can extract the logarithm. Apply the numerical inequality $\log(1 + a) \leq a$ for $a > -1$ using the transfer rule. □

Proof of Fact A.1. The result follows quickly from standard matrix concentration arguments. For example, to derive the probability inequality, we apply the matrix Laplace transform method [Tro15, Prop. 3.2.1]. For each $\theta > 0$,

$$\begin{aligned}\mathbb{P}\{\lambda_{\min}(\mathbf{Y} - \mathbb{E}\,\mathbf{Y}) \le -t\} &\le \mathrm{e}^{-\theta t}\,\mathbb{E}\,\mathrm{Tr}\exp(-\theta\mathbf{Y} + \theta\,\mathbb{E}[\mathbf{Y}]) \\ &\le \mathrm{e}^{-\theta t}\,\mathrm{Tr}\exp\left(\sum_{i=1}^{n}\big(\log\mathbb{E}\,\mathrm{e}^{-\theta\mathbf{W}_i} + \theta\,\mathbb{E}[\mathbf{W}_i]\big)\right) \\ &\le \mathrm{e}^{-\theta t}\,\mathrm{Tr}\exp\left(\frac{1}{2}\theta^2\sum_{i=1}^{n}\mathbb{E}[\mathbf{W}_i^2]\right) \le d\cdot\mathrm{e}^{-\theta t+\theta^2 L_2/2}.\end{aligned}$$

The second inequality requires the subadditivity of matrix log-mgfs [Tro15, Lem. 3.1] and the monotonicity of the trace exponential. The third inequality is Lemma A.2. The last inequality depends on the spectral mapping theorem and the definition of L_2. Select $\theta = t/(2L_2)$ to complete the bound. The stated result follows from an application of Weyl's inequality [Bha97, Cor. III.2.2]. □

Acknowledgement

The author thanks Ethan Epperly, Jorge Garza-Vargas, Otte Heinävaara, De Huang, Raphael Meyer, Ramon van Handel, Roman Vershynin, and Rob Webber for valuable discussions and feedback. The anonymous reviewer also offered thoughtful comments that improved the paper.

References

[AS17] Guillaume Aubrun and Stanisław J. Szarek, *Alice and Bob meet Banach*, Mathematical Surveys and Monographs, vol. 223, American Mathematical Society, Providence, RI, 2017. The interface of asymptotic geometric analysis and quantum information theory, DOI 10.1090/surv/223. MR3699754

[AC11] Jean-Yves Audibert and Olivier Catoni, *Robust linear least squares regression*, Ann. Statist. **39** (2011), no. 5, 2766–2794, DOI 10.1214/11-AOS918. MR2906886

[BY93] Z. D. Bai and Y. Q. Yin, *Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix*, Ann. Probab. **21** (1993), no. 3, 1275–1294. MR1235416

[BBvH23] Afonso S. Bandeira, March T. Boedihardjo, and Ramon van Handel, *Matrix concentration inequalities and free probability*, Invent. Math. **234** (2023), no. 1, 419–487, DOI 10.1007/s00222-023-01204-6. MR4635836

[Ban+24] Afonso S. Bandeira et al., *Matrix concentration inequalities and free probability II. Two-sided bounds and applications*, 2024. `arXiv:2406.11453 [math.PR]`.

[BSS14] Joshua Batson, Daniel A. Spielman, and Nikhil Srivastava, *Twice-Ramanujan sparsifiers*, SIAM Rev. **56** (2014), no. 2, 315–334, DOI 10.1137/130949117. MR3201185

[BMV75] D. Bessis, P. Moussa, and M. Villani, *Monotonic converging variational approximations to the functional integrals in quantum statistical mechanics*, J. Mathematical Phys. **16** (1975), no. 11, 2318–2325, DOI 10.1063/1.522463. MR416396

[Bha97] Rajendra Bhatia, *Matrix analysis*, Graduate Texts in Mathematics, vol. 169, Springer-Verlag, New York, 1997, DOI 10.1007/978-1-4612-0653-8. MR1477662

[Bha07] Rajendra Bhatia, *Positive definite matrices*, Princeton Series in Applied Mathematics, Princeton University Press, Princeton, NJ, 2007. MR2284176

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities*, Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence; With a foreword by Michel Ledoux, DOI 10.1093/acprof:oso/9780199535255.001.0001. MR3185193

[BDN15] Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson, *Toward a unified theory of sparse dimensionality reduction in Euclidean space*, Geom. Funct. Anal. **25** (2015), no. 4, 1009–1088, DOI 10.1007/s00039-015-0332-9. MR3385629

[BvH24] Tatiana Brailovskaya and Ramon van Handel, *Universality and sharp matrix concentration inequalities*, Geom. Funct. Anal. **34** (2024), no. 6, 1734–1838, DOI 10.1007/s00039-024-00692-9. MR4823211

[CHZ22] T. Tony Cai, Rungang Han, and Anru R. Zhang, *On the non-asymptotic concentration of heteroskedastic Wishart-type matrix*, Electron. J. Probab. **27** (2022), Paper No. 29, 40, DOI 10.1214/22-ejp758. MR4385832

[CEMT25] C. Camaño, E. N. Epperly, R. A. Meyer, and J. A. Tropp, *Faster linear algebra algorithms with structured random matrices*, 2025. `arXiv:2508.21189 [cs.DS]`.

[Car10] Eric Carlen, *Trace inequalities and quantum entropy: an introductory course*, Entropy and the quantum, Contemp. Math., vol. 529, Amer. Math. Soc., Providence, RI, 2010, pp. 73–140, DOI 10.1090/conm/529/10428. MR2681769

[CFS21] Coralia Cartis, Jan Fiala, and Zhen Shao, *Hashing embeddings of optimal dimension, with applications to linear least squares*, 2021. `arXiv:2105.11815 [math.NA]`.

[Cha07] Sourav Chatterjee, *Stein's method for concentration inequalities*, Probab. Theory Related Fields **138** (2007), no. 1-2, 305–321, DOI 10.1007/s00440-006-0029-y. MR2288072

[CGS11] Louis H. Y. Chen, Larry Goldstein, and Qi-Man Shao, *Normal approximation by Stein's method*, Probability and its Applications (New York), Springer, Heidelberg, 2011, DOI 10.1007/978-3-642-15007-4. MR2732624

[CDD25] Shabarish Chenakkod, Michał Dereziński, and Xiaoyu Dong, *Optimal oblivious subspace embeddings with near-optimal sparsity*, 52nd International Colloquium on Automata, Languages, and Programming, LIPIcs. Leibniz Int. Proc. Inform., vol. 334, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2025, pp. Art. No. 55, 20, DOI 10.4230/lipics.icalp.2025.55. MR4934207

[CDD26] Shabarish Chenakkod, Michał Dereziński, and Xiaoyu Dong, *Optimal subspace embeddings: resolving Nelson-Nguyen conjecture up to sub-polylogarithmic factors*, Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), SIAM, Philadelphia, PA, 2026, pp. 4501–4559, DOI 10.1137/1.9781611978971.163. MR5027130

[CDDR24] Shabarish Chenakkod, Michał Dereziński, Xiaoyu Dong, and Mark Rudelson, *Optimal embedding dimension for sparse subspace embeddings*, STOC'24—Proceedings of the 56th Annual ACM Symposium on Theory of Computing, ACM, New York, 2024, pp. 1106–1117, DOI 10.1145/3618260.3649762. MR4764888

[Cho94] K. P. Choi, *On the medians of gamma distributions and an equation of Ramanujan*, Proc. Amer. Math. Soc. **121** (1994), no. 1, 245–251, DOI 10.2307/2160389. MR1195477

[CW13] Kenneth L. Clarkson and David P. Woodruff, *Low rank approximation and regression in input sparsity time*, STOC'13—Proceedings of the 2013 ACM Symposium on Theory of Computing, ACM, New York, 2013, pp. 81–90, DOI 10.1145/2488608.2488620. MR3210769

[Cli16] Fabien Clivaz, *Stahl's theorem (aka BMV conjecture): insights and intuition on its proof*, Spectral theory and mathematical physics, Oper. Theory Adv. Appl., vol. 254, Birkhäuser/Springer, [Cham], 2016, pp. 107–117, DOI 10.1007/978-3-319-29992-1_6. MR3526447

[Coh16] Michael B. Cohen, *Nearly tight oblivious subspace embeddings by trace inequalities*, Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York, 2016, pp. 278–287, DOI 10.1137/1.9781611974331.ch21. MR3478397

[Dud02] R. M. Dudley, *Real analysis and probability*, Cambridge Studies in Advanced Mathematics, vol. 74, Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original, DOI 10.1017/CBO9780511755347. MR1932358

[DZ24] Ioana Dumitriu and Yizhe Zhu, *Extreme singular values of inhomogeneous sparse random rectangular matrices*, Bernoulli **30** (2024), no. 4, 2904–2931, DOI 10.3150/23-bej1699. MR4779853

[Erë15] A. È. Erëmenko, *Herbert Stahl's proof of the BMV conjecture*, Mat. Sb., 206.1, 2015, pp. 97–102.

[GKKT20] M. Guţă, J. Kahn, R. Kueng, and J. A. Tropp, *Fast state tomography with optimal error bounds*, J. Phys. A **53** (2020), no. 20, 204001, 28, DOI 10.1088/1751-8121/ab8111. MR4093475

[Hei25] Otte Heinävaara, *Tracial joint spectral measures*, Invent. Math. **239** (2025), no. 2, 505–526, DOI 10.1007/s00222-024-01309-6. MR4850603

[Kah86] Jean-Pierre Kahane, *Une inégalité du type de Slepian et Gordon sur les processus gaussiens* (French, with English summary), Israel J. Math. **55** (1986), no. 1, 109–110, DOI 10.1007/BF02772698. MR858463

[KM15] Vladimir Koltchinskii and Shahar Mendelson, *Bounding the smallest singular value of a random matrix without concentration*, Int. Math. Res. Not. IMRN **23** (2015), 12991–13008, DOI 10.1093/imrn/rnv096. MR3431642

[LST02] John Langford and John Shawe-Taylor, *PAC-Bayes & Margins*, Adv. Neural Information Processing Systems, volume 15, MIT Press, 2002.

[LO99] Rafał Latała and Krzysztof Oleszkiewicz, *Gaussian measures of dilatations of convex symmetric sets*, Ann. Probab. **27** (1999), no. 4, 1922–1938, DOI 10.1214/aop/1022677554. MR1742894

[LS04] Elliott H. Lieb and Robert Seiringer, *Equivalent forms of the Bessis-Moussa-Villani conjecture*, J. Statist. Phys. **115** (2004), no. 1-2, 185–190, DOI 10.1023/B:JOSS.0000019811.15510.27. MR2070093

[MT20] Per-Gunnar Martinsson and Joel A. Tropp, *Randomized numerical linear algebra: foundations and algorithms*, Acta Numer. **29** (2020), 403–572, DOI 10.1017/s0962492920000021. MR4189294

[McA03] David McAllester, *Simplified PAC Bayesian Margin bounds*, Learning Theory and Kernel Machines, Springer, Berlin, 2003, pp. 203–215.

[MU17] Michael Mitzenmacher and Eli Upfal, *Probability and computing*, 2nd ed., Cambridge University Press, Cambridge, 2017. Randomization and probabilistic techniques in algorithms and data analysis. MR3674428

[Mui82] Robb J. Muirhead, *Aspects of multivariate statistical theory*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York, 1982. MR652932

[Mur+23] Riley Murray et al. *Randomized numerical linear algebra: A perspective on the field with an eye to software*, 2023. `arXiv:2302.11474 [math.NA]`.

[NN13] Jelani Nelson and Huy L. Nguyễn, *OSNAP: faster numerical linear algebra algorithms via sparser subspace embeddings*, 2013 IEEE 54th Annual Symposium on Foundations of Computer Science—FOCS 2013, IEEE Computer Soc., Los Alamitos, CA, 2013, pp. 117–126, DOI 10.1109/FOCS.2013.21. MR3246213

[NN14] Jelani Nelson and Huy L. Nguyễn, *Lower bounds for oblivious subspace embeddings*, Automata, languages, and programming. Part I, Lecture Notes in Comput. Sci., vol. 8572, Springer, Heidelberg, 2014, pp. 883–894, DOI 10.1007/978-3-662-43948-7_73. MR3238678

[NP12] Ivan Nourdin and Giovanni Peccati, *Normal approximations with Malliavin calculus*, Cambridge Tracts in Mathematics, vol. 192, Cambridge University Press, Cambridge, 2012. From Stein's method to universality, DOI 10.1017/CBO9781139084659. MR2962301

[Oli16] Roberto Imbuzeiro Oliveira, *The lower tail of random quadratic forms with applications to ordinary least squares*, Probab. Theory Related Fields **166** (2016), no. 3-4, 1175–1194, DOI 10.1007/s00440-016-0738-9. MR3568047

[OT18] Samet Oymak and Joel A. Tropp, *Universality laws for randomized dimension reduction, with applications*, Inf. Inference **7** (2018), no. 3, 337–446, DOI 10.1093/imaiai/iax011. MR3858331

[Ros11] Nathan Ross, *Fundamentals of Stein's method*, Probab. Surv. **8** (2011), 210–293, DOI 10.1214/11-PS182. MR2861132

[Rud99] M. Rudelson, *Random vectors in the isotropic position*, J. Funct. Anal. **164** (1999), no. 1, 60–72, DOI 10.1006/jfan.1998.3384. MR1694526

[Sar06] T. Sarlós. *Improved approximation algorithms for large matrices via random projections*, FOCS'06: Proc. 2006 IEEE 47th Ann. Symp. Foundations of Computer Science, 2006, pp. 143–152.

[SSV10] René L. Schilling, Renming Song, and Zoran Vondraček, *Bernstein functions*, De Gruyter Studies in Mathematics, vol. 37, Walter de Gruyter & Co., Berlin, 2010. Theory and applications. MR2598208

[SV13] Nikhil Srivastava and Roman Vershynin, *Covariance estimation for distributions with* $2 + \varepsilon$ *moments*, Ann. Probab. **41** (2013), no. 5, 3081–3111, DOI 10.1214/12-AOP760. MR3127875

[Sta13] Herbert R. Stahl, *Proof of the BMV conjecture*, Acta Math. **211** (2013), no. 2, 255–290, DOI 10.1007/s11511-013-0104-z. MR3143891

[Tal21] Michel Talagrand, *Upper and lower bounds for stochastic processes—decomposition theorems*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics], vol. 60, Springer, Cham, 2021. Second edition [of 3184689], DOI 10.1007/978-3-030-82595-9. MR4381414

[Tao12] Terence Tao, *Topics in random matrix theory*, Graduate Studies in Mathematics, vol. 132, American Mathematical Society, Providence, RI, 2012, DOI 10.1090/gsm/132. MR2906465

[Tik18] Konstantin Tikhomirov, *Sample covariance matrices of heavy-tailed distributions*, Int. Math. Res. Not. IMRN **20** (2018), 6254–6289, DOI 10.1093/imrn/rnx067. MR3872323

[Tro26a] J. A. Tropp, *Comparison theorems for the extreme eigenvalues of a random symmetric matrix*, 2026, `arXiv:2603.043365 [ma.PR]`.

[Tro26b] J. A. Tropp, *Universality laws for random matrices via exchangeable counterparts*, 2026, arXiv:2603.05803 [ma.PR].

[Tro15] Joel A. Tropp, *An introduction to matrix concentration inequalities*, Found. Trends Machine Learning, 8(1-2):1–230, 2015.

[Tro18] Joel A. Tropp, *Second-order matrix concentration inequalities*, Appl. Comput. Harmon. Anal. **44** (2018), no. 3, 700–736, DOI 10.1016/j.acha.2016.07.005. MR3768857

[TYUC19] Joel A. Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher, *Streaming low-rank matrix approximation with an application to scientific simulation*, SIAM J. Sci. Comput. **41** (2019), no. 4, A2430–A2463, DOI 10.1137/18M1201068. MR3986561

[vdV98] A. W. van der Vaart, *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 3, Cambridge University Press, Cambridge, 1998, DOI 10.1017/CBO9780511802256. MR1652247

[vH16] Ramon van Handel, *Probability in high dimension*, APC 550 Lecture Notes, Princeton Univ., 2016.

[Ver18] Roman Vershynin, *High-dimensional probability*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 47, Cambridge University Press, Cambridge, 2018. An introduction with applications in data science; With a foreword by Sara van de Geer, DOI 10.1017/9781108231596. MR3837109

[Wal18] Shayne F. D. Waldron, *An introduction to finite tight frames*, Applied and Numerical Harmonic Analysis, Birkhäuser/Springer, New York, 2018, DOI 10.1007/978-0-8176-4815-2. MR3752185

[Wid41] David Vernon Widder, *The Laplace Transform*, Princeton Mathematical Series, vol. 6, Princeton University Press, Princeton, NJ, 1941. MR5923

[Woo14] David P. Woodruff, *Sketching as a tool for numerical linear algebra*, Found. Trends Theor. Comput. Sci. **10** (2014), no. 1-2, iv+157, DOI 10.1561/0400000060. MR3285427

[Zhi24] Nikita Zhivotovskiy, *Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle*, Electron. J. Probab. **29** (2024), Paper No. 13, 28, DOI 10.1214/23-ejp1021. MR4693860

Department of Computing and Mathematical Science, Caltech, Pasadena, California 91125-5000

URL: https://tropp.caltech.edu

Email address: jtropp@caltech.edu