TAIL BOUNDS FOR ALL EIGENVALUES OF A SUM OF RANDOM MATRICES

ALEX GITTENS AND JOEL A. TROPP

Abstract. This work introduces the minimax Laplace transform method, a modification of the cumulant-based matrix Laplace transform method developed in [Tro11c] that yields *both* upper and lower bounds on *each* eigenvalue of a sum of random self-adjoint matrices. This machinery is used to derive eigenvalue analogs of the classical Chernoff, Bennett, and Bernstein bounds.

Two examples demonstrate the efficacy of the minimax Laplace transform. The first concerns the effects of column sparsification on the spectrum of a matrix with orthonormal rows. Here, the behavior of the singular values can be described in terms of coherence-like quantities. The second example addresses the question of relative accuracy in the estimation of eigenvalues of the covariance matrix of a random process. Standard results on the convergence of sample covariance matrices provide bounds on the number of samples needed to obtain relative accuracy in the spectral norm, but these results only guarantee relative accuracy in the estimate of the maximum eigenvalue. The minimax Laplace transform argument establishes that if the lowest eigenvalues decay sufficiently fast, $\Omega(\varepsilon^{-2}\kappa_\ell^2 \ell \log p)$ samples, where $\kappa_\ell = \lambda_1(\boldsymbol{C})/\lambda_\ell(\boldsymbol{C})$, are sufficient to ensure that the dominant ℓ eigenvalues of the covariance matrix of a $\mathcal{N}(\mathbf{0}, \boldsymbol{C})$ random vector are estimated to within a factor of $1 \pm \varepsilon$ with high probability.

1. Introduction

The field of nonasymptotic random matrix theory has traditionally focused on the problem of bounding the extreme eigenvalues of a random matrix. In some circumstances, however, we may also be interested in studying the behavior of the interior eigenvalues. In this case, classical tools do not readily apply. Indeed, the interior eigenvalues are determined by the min-max of a random process, which is very challenging to control.

This paper demonstrates that it is possible to combine the matrix Laplace transform method detailed in [Tro11c] with the Courant–Fischer characterization of eigenvalues to obtain nontrivial bounds on the interior eigenvalues of a sum of random self-adjoint matrices. This approach expands the scope of the matrix probability inequalities from [Tro11c] so that they provide interesting information about the bulk spectrum.

As one application of our approach, we investigate estimates for the covariance matrix of a centered stationary random process. We show that the eigenvalues of the sample covariance matrix provide relative-error approximations to the eigenvalues of the covariance matrix. We focus on Gaussian processes, but our arguments can be extended to other distributions. The following theorem distills the results in section 7.

Theorem 1.1. *Let $\boldsymbol{C} \in \mathbb{R}^{p\times p}$ be positive semidefinite. Fix an integer $\ell \leq p$ and assume the tail $\{\lambda_i(\boldsymbol{C})\}_{i>\ell}$ of the spectrum of $\boldsymbol{C}$ decays sufficiently fast that*

$$\sum\nolimits_{i>\ell} \lambda_i(\boldsymbol{C}) = \mathrm{O}(\lambda_1(\boldsymbol{C})).$$

Date: July 21, 2011.

Both authors can be reached at Annenberg Center, MC 305-16, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91125. Email: gittens@cms.caltech.edu and jtropp@cms.caltech.edu. Research supported by ONR awards N00014-08-1-0883 and N00014-11-1-0025, AFOSR award FA9550-09-1-0643, and a Sloan Fellowship.

Let $\{\boldsymbol{\eta}_j\}_{j=1}^n \subset \mathbb{R}^p$ be i.i.d. samples drawn from a $\mathcal{N}(\mathbf{0}, \boldsymbol{C})$ distribution. Define the sample covariance matrix

$$\widehat{\boldsymbol{C}}_n = \frac{1}{n} \sum\nolimits_{j=1}^n \boldsymbol{\eta}_j \boldsymbol{\eta}_j^*.$$

Let κ_ℓ be the condition number associated with a dominant ℓ-dimensional invariant subspace of $\boldsymbol{C}$,

$$\kappa_\ell = \frac{\lambda_1(\boldsymbol{C})}{\lambda_\ell(\boldsymbol{C})}.$$

If $n = \Omega(\varepsilon^{-2}\kappa_\ell^2 \ell \log p)$, then with high probability

$$|\lambda_k(\widehat{\boldsymbol{C}}_n) - \lambda_k(\boldsymbol{C})| \leq \varepsilon \lambda_k(\boldsymbol{C}) \quad \text{for } k = 1, \ldots, \ell.$$

Thus, assuming sufficiently fast decay of the residual eigenvalues, $n = \Omega(\varepsilon^{-2}\kappa_\ell^2 \ell \log p)$ samples ensure that the top ℓ eigenvalues of $\boldsymbol{C}$ are captured to relative precision. Spectral decay of this sort is encountered when, e.g., the residual eigenvalues of $\boldsymbol{C}$ decay like $k^{-(1+\delta)}$ for some $\delta > 0$ or when they arise from measurements corrupted by low-power white noise.

We contrast Theorem 1.1 with established spectral norm error bounds for covariance estimation, which do not exploit spectral decay and require that $n = \Omega(\varepsilon^{-2}\kappa_\ell^2 p)$ samples be taken to capture the top ℓ eigenvalues to relative precision (see section 7). The estimate in Theorem 1.1 can be sharpened using information about the spectrum of $\boldsymbol{C}$ and the desired failure probability or modified to account for different types of spectral decay. The same tools used in the proof of the theorem can be used to estimate $\lambda_k(\widehat{\boldsymbol{C}}_n - \boldsymbol{C})$.

1.1. **Related Work.** We believe that this paper contains the first general-purpose tools for studying the full spectrum of a finite-dimensional random matrix. The literature on random matrix theory (RMT) contains some complementary results, but they do not seem to apply with the same generality. Methods from RMT fall into two rough categories: asymptotic methods and nonasymptotic methods. We discuss the relevant results from each in turn.

The modern asymptotic theory began in the 1950s when physicists observed that, on certain scales, the behavior of a quantum system is described by the spectrum of a random matrix [Meh04]. They further observed the phenomenon of *universality*: as the dimension increases, the spectral statistics become independent of the distribution of the random matrix; instead, they are determined by the symmetries of the distribution [Dei07]. Since these initial observations, physicists, statisticians, engineers, and mathematicians have found manifold applications of the asymptotic theory in high-dimensional statistics [Joh01, Joh07, El 08], physics [GMGW98, Meh04], wireless communication [TV04, ST06], and pure mathematics [RS96, BK99], to mention only a few areas.

Asymptotic random matrix theory has developed primarily through the examination of specific classes of random matrices. We mention two well-studied classes. Sample covariance matrices take the form $n^{-1}\boldsymbol{B}_n\boldsymbol{B}_n^*$, where the columns of $\boldsymbol{B}_n$ comprise n independent observations. Wigner matrices are Hermitian matrices whose superdiagonal entries are independent, zero-mean, and have unit variance and whose diagonal entries are i.i.d., real, and have finite variance.

The fundamental object of study in asymptotic random matrix theory is the empirical spectral distribution function (ESD). Given a random Hermitian matrix $\boldsymbol{A}$ of order n, its ESD

$$F^{\boldsymbol{A}}(x) = \frac{1}{n}\#\{1 \leq i \leq n : \lambda_i(\boldsymbol{A}) \leq x\}$$

is a random distribution function which encodes the statistics of the spectrum of $\boldsymbol{A}$. Wigner's theorem [Wig55], the seminal result of the asymptotic theory, establishes that if $\{\boldsymbol{A}_n\}$ is a sequence of independent, symmetric $n \times n$ matrices with i.i.d. $\mathcal{N}(0,1)$ entries on and above the diagonal,

then the expected ESD of $n^{-1/2}\boldsymbol{A}_n$ converges weakly in probability, as n approaches infinity, to the semicircular law given by

$$F(x) = \frac{1}{2\pi} \int_{-\infty}^{x} \sqrt{4 - y^2}\, \mathbf{1}_{[-2,2]}(y)\, \mathrm{d}y.$$

Thus, at least in the limiting sense, the spectra of these random matrices are well characterized. Development of the classical asymptotic theory has been driven by the natural question raised by Wigner's result: to what extent is the semicircular law, and more generally, the existence of a limiting spectral distribution (LSD) universal?

The literature on the existence and universality of LSDs is massive; we mention only the highlights. It is now known that the semicircular law is universal for Wigner matrices. Suppose that $\{\boldsymbol{A}_n\}$ is a sequence of independent $n \times n$ Wigner matrices. Grenander established that if all the moments are finite, then the ESD of $n^{-1/2}\boldsymbol{A}_n$ converges weakly to the semicircular law in probability [Gre63]. Arnold showed that, assuming a finite fourth moment, the ESD almost surely converges weakly to the semicircular law [Arn71]. Around the same time, Marčenko and Pastur determined the form of the limiting spectral distribution of sample covariance matrices [MP67].

More recently, Tao and Vu confirmed the long-conjectured circular law hypothesis. Let $\{\boldsymbol{C}_n\}$ be a sequence of independent $n \times n$ matrices whose entries are i.i.d. and have unit variance. Then the ESD of $n^{-1/2}\boldsymbol{C}_n$ converges weakly to the uniform measure on the unit disk, both in probability and almost surely [TV10b].

Although the convergence rate of the ESD has considerable practical interest, it was not until 1993 that theoretical results became available when Bai showed that for Wigner matrices [Bai93a] and sample covariance matrices [Bai93b] the expected ESDs of $n^{-1/2}\boldsymbol{A}_n$ and $n^{-1}\boldsymbol{B}_n\boldsymbol{B}_n^*$, respectively, both converge pointwise at a rate of $\mathrm{O}(n^{-1/4})$. Later, Bai and coauthors established the pointwise convergence in probability of the ESD of the normalized Wigner matrix $n^{-1/2}\boldsymbol{A}_n$ [BMT97] and greatly improved the convergence rates [BMT99, BMT02, BMY03]. The strongest result to date is due to Bai et al., who have shown that, if the entries of the Wigner matrix possess finite sixth moments, then pointwise convergence in probability of the ESD of $n^{-1/2}\boldsymbol{A}_n$ occurs at the rate of $\mathrm{O}(n^{-1/2})$ [BHPZ11].

Classically, individual eigenvalues have been studied through the limiting behavior of the extremal eigenvalues and the asymptotic joint distribution of several eigenvalues. Much is known about the limiting distribution of the largest eigenvalues of Wigner and covariance matrices. Geman showed that if the columns of $\boldsymbol{B}_n$ are drawn from a sufficiently regular distribution, then the largest eigenvalue of the sample covariance matrix $n^{-1}\boldsymbol{B}_n\boldsymbol{B}_n^*$ converges almost surely to a limit [Gem80]. Bai, Yin, and coauthors showed that the existence of a fourth moment is both necessary and sufficient for the existence of such a limit [YBK88, BSY88]. They also identified necessary and sufficient conditions for the existence of limits for the smallest and largest eigenvalues of a normalized Wigner matrix $n^{-1/2}\boldsymbol{A}_n$ [BY88b]. El Karoui has recently described the limiting behavior of the leading eigenvalues of a large class of sample covariance matrices [El 07].

Less is known about the rate of convergence of the eigenvalues, but some results are available. Write the eigenvalues of a self-adjoint matrix $\boldsymbol{A}$ in nonincreasing order $\lambda_1 \geq \ldots \geq \lambda_n$. For $1 \leq j \leq n$, the classical location γ_j of the jth eigenvalue of the normalized Wigner matrix $n^{-1/2}\boldsymbol{A}_n$ is defined via the relation

$$\int_{-\infty}^{\gamma_j} \rho_{sc}(x)\, dx = \frac{j}{n},$$

where ρ_{sc} is the density associated with the semicircular law. Intuitively, the facts that $F^{\frac{1}{\sqrt{n}}\boldsymbol{A}_n} \to F^{sc}$ and $F^{\frac{1}{\sqrt{n}}\boldsymbol{A}_n}(\lambda_j) = j/n$ suggest that $\frac{1}{\sqrt{n}}\lambda_j \to \gamma_j$. Indeed, it follows from [BY88a, BY88b] that

$$\lambda_j = \sqrt{n}\gamma_j + \mathrm{o}(\sqrt{n})$$

asymptotically almost surely. Under the assumption that the entries exhibit uniform subgaussian decay, Erdös, Yau, and Yin have strengthened this result by showing that, up to log factors, the eigenvalues of $n^{-1/2}\boldsymbol{A}_n$ are within $\mathrm{O}(n^{-2/3})$ of their classical position with high probability [EYY10]. More generally, Tao and Vu have established the universality of a result due to Gustavsson [Gus05] in the complex Gaussian Wigner case: $(\log n)^{-1/2}(\sqrt{n}\lambda_j - n\gamma_j)$ is asymptotically normally distributed [TV11]. Further, they have shown that eigenvalues in the bulk of the spectrum ($j = \Omega(n)$) of a Wigner matrix satisfy

$$\mathbb{E}|\lambda_j - \sqrt{n}\gamma_j|^2 = \mathrm{O}(n^{-c}),$$

for some universal constant $c > 0$ [TV10a].

In contrast to the asymptotic theory, which remains to a large extent driven by the study of particular classes of random matrices, the nonasymptotic theory has developed as a collection of techniques for addressing the behavior of a broad range of random matrices. The nonasymptotic theory has its roots in geometric functional analysis in the 1970s, where random matrices were used to investigate the local properties of Banach spaces [LM93, SD01, Ver10]. Since then, the nonasymptotic theory has found applications in areas including theoretical computer science [Ach03, Vem04, SS08], machine learning [DM05], optimization [Nem07, So09], and numerical linear algebra [DM10, HMT11, Mah11].

As is the case in the asymptotic theory, the sharpest and most comprehensive results available in the nonasymptotic theory concern the behavior of Gaussian matrices. The amenability of the Gaussian distribution makes it possible to obtain results such as Szarek's nonasymptotic analog of the Wigner semicircle theorem for Gaussian matrices [Sza90] and Chen and Dongarra's bounds on the condition number of Gaussian matrices [CD05]. The properties of less well-behaved random matrices can sometimes be related back to those of Gaussian matrices using probabilistic tools, such as symmetrization; see, e.g., the derivation of Latała's bound on the norms of zero-mean random matrices [Lat05].

More generally, bounds on extremal eigenvalues can be obtained from knowledge of the moments of the entries. For example, the smallest singular value of a square matrix with i.i.d. zero-mean subgaussian entries with unit variance is $\mathrm{O}(n^{-1/2})$ with high probability [RV08]. Concentration of measure results, such as Talagrand's concentration inequality for product spaces [Tal95], have also contributed greatly to the nonasymptotic theory. We mention in particular the work of Achlioptas and McSherry on randomized sparsification of matrices [AM01, AM07], that of Meckes on the norms of random matrices [Mec04], and that of Alon, Krivelevich and Vu [AKV02] on the concentration of the largest eigenvalues of random symmetric matrices, all of which are applications of Talagrand's inequality. In cases where geometric information on the distribution of the random matrices is available, the tools of empirical process theory—such as the generic chaining, also due to Talagrand [Tal05]—can be used to convert this geometric information into information on the spectra. One natural example of such a case consists of matrices whose rows are independently drawn from a log-concave distribution [MP06, ALPTJ11].

The noncommutative Khintchine inequality (NCKI), which bounds the moments of the norm of a sum of fixed matrices modulated by random signs [LP86, LPP91], is a widely used tool in the nonasymptotic theory. Despite its power, the NCKI is unwieldy. To use it, one must reduce the problem to a suitable form by applying symmetrization and decoupling arguments and exploiting the equivalence between moments and tail bounds. It is often more convenient to apply the NCKI in the guise of a lemma, due to Rudelson [Rud99], that provides an analog of the law of large numbers for sums of rank-one matrices. This result has found many applications, including column-subset selection [RV07] and the fast approximate solution of least-squares problems [DMMS11]. The NCKI and its corollaries do not always yield sharp results because parasitic logarithmic factors arise in many settings.

The current paper is ultimately based on the influential work of Ahlswede and Winter [AW02]. This line of research leads to explicit tail bounds for the maximum eigenvalue of a sum of random matrices. These probability inequalities parallel the classical scalar tail bounds due to Bernstein and others. Matrix probability inequalities allow us to obtain valuable information about the maximum eigenvalue of a random matrix with very little effort. Furthermore, they apply to a wide variety of random matrices. We note, however, that matrix probability inequalities can lead to parasitic logarithmic factors similar to those that emerge from the NCKI.

Major contributions to the literature on matrix probability inequalities include the papers [CM08, Rec09, Gro11]. We emphasize two works of Oliveira [Oli09, Oli10] that go well beyond earlier research. The sharpest current results appear in the works of Tropp [Tro11c, Tro11b, Tro11a]. Recently, Hsu, Kakade, and Zhang [HKZ11] have modified Tropp's approach to establish matrix probability inequalities that depend on an intrinsic dimension parameter, rather than the ambient dimension.

1.2. **Outline.** In section 2, we introduce the notation used in this paper and state a convenient version of the Courant–Fischer theorem. In section 3, we use the Courant–Fischer theorem to extend the Laplace transform technique from [Tro11c] to apply to all the eigenvalues of self-adjoint matrices, thereby obtaining the minimax Laplace transform. We apply this technique in sections 4 and 5 to develop eigenvalue analogs of the classical Chernoff and Bernstein bounds. The final two sections illustrate, using two familiar problems, that the minimax Laplace technique gives us significantly more information on the spectra of random matrices than current approaches. In section 6, we use the Chernoff bounds to quantify the effects of column sparsification on all the singular values of matrices with orthogonal rows. In section 7, we consider the question of how fast, in relative error, the eigenvalues of empirical covariance matrices converge.

2. Background and Notation

We establish the notation used in the sequel and state a convenient version of the Courant–Fischer theorem.

Unless otherwise stated, we work over the complex field. The kth column of the matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{a}_k$, and the entries are denoted a_{jk} or $(\boldsymbol{A})_{jk}$. We define $\mathbb{M}_{\mathrm{sa}}^n$ to be the set of self-adjoint matrices with dimension n. The eigenvalues of a matrix $\boldsymbol{A}$ in $\mathbb{M}_{\mathrm{sa}}^n$ are arranged in weakly decreasing order: $\lambda_{\max}(\boldsymbol{A}) = \lambda_1(\boldsymbol{A}) \geq \lambda_2(\boldsymbol{A}) \geq \cdots \geq \lambda_n(\boldsymbol{A}) = \lambda_{\min}(\boldsymbol{A})$. Likewise, singular values of a rectangular matrix $\boldsymbol{B}$ with rank r are ordered $s_1(\boldsymbol{B}) \geq s_2(\boldsymbol{B}) \geq \cdots \geq s_r(\boldsymbol{B})$. The spectral norm of a matrix $\boldsymbol{B}$ is expressed as $\|\boldsymbol{B}\|$. We often compare self-adjoint matrices using the semidefinite ordering. In this ordering, $\boldsymbol{A}$ is greater than or equal to $\boldsymbol{B}$, written $\boldsymbol{A} \succeq \boldsymbol{B}$ or $\boldsymbol{B} \preceq \boldsymbol{A}$, when $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite.

The expectation of a random variable is denoted by $\mathbb{E}X$. We write $X \sim \mathrm{Bern}(p)$ to indicate that X has a Bernoulli distribution with mean p.

One of our central tools is the variational characterization of the eigenvalues of a self-adjoint matrix given by the Courant–Fischer theorem. For integers d and n satisfying $1 \leq d \leq n$, the complex Stiefel manifold

$$\mathbb{V}_d^n = \{\boldsymbol{V} \in \mathbb{C}^{n \times d} : \boldsymbol{V}^*\boldsymbol{V} = \mathbf{I}\}$$

is the collection of orthonormal bases for the d-dimensional subspaces of $\mathbb{C}^n$, or, equivalently, the collection of all isometric embeddings of $\mathbb{C}^d$ into $\mathbb{C}^n$. Let $\boldsymbol{A}$ be a self-adjoint matrix with dimension n, and let $\boldsymbol{V} \in \mathbb{V}_d^n$ be an orthonormal basis for a subspace of $\mathbb{C}^n$. Then the matrix $\boldsymbol{V}^*\boldsymbol{A}\boldsymbol{V}$ can be interpreted as the compression of $\boldsymbol{A}$ to the space spanned by $\boldsymbol{V}$.

Proposition 2.1 (Courant–Fischer)**.** *Let $\boldsymbol{A}$ be a self-adjoint matrix with dimension n. Then*

$$\lambda_k(\boldsymbol{A}) = \min_{\boldsymbol{V} \in \mathbb{V}_{n-k+1}^n} \lambda_{\max}\left(\boldsymbol{V}^* \boldsymbol{A} \boldsymbol{V}\right) \quad \text{and} \tag{2.1}$$

$$\lambda_k(\boldsymbol{A}) = \max_{\boldsymbol{V} \in \mathbb{V}_k^n} \lambda_{\min}\left(\boldsymbol{V}^* \boldsymbol{A} \boldsymbol{V}\right). \tag{2.2}$$

A matrix $\boldsymbol{V}_- \in \mathbb{V}_k^n$ achieves equality in (2.2) *if and only if its columns span a dominant k-dimensional invariant subspace of $\boldsymbol{A}$. Likewise, a matrix $\boldsymbol{V}_+ \in \mathbb{V}_{n-k+1}^n$ achieves equality in* (2.1) *if and only if its columns span a bottom $(n-k+1)$-dimensional invariant subspace of $\boldsymbol{A}$.*

The $\pm$ subscripts in Proposition 2.1 are chosen to reflect the fact that $\lambda_k(\boldsymbol{A})$ is the *minimum* eigenvalue of $\boldsymbol{V}_-^* \boldsymbol{A} \boldsymbol{V}_-$ and the *maximum* eigenvalue of $\boldsymbol{V}_+^* \boldsymbol{A} \boldsymbol{V}_+$. As a consequence of Proposition 2.1, when $\boldsymbol{A}$ is self-adjoint, $\lambda_k(-\boldsymbol{A}) = -\lambda_{n-k+1}(\boldsymbol{A})$. This fact allows us to use the same techniques we develop for bounding the eigenvalues from above to bound them from below.

3. Tail Bounds For Interior Eigenvalues

In this section we develop a generic bound on the tail probabilities of eigenvalues of sums of independent, random, self-adjoint matrices. We establish this bound by supplementing the matrix Laplace transform methodology of [Tro11c] with Proposition 2.1 and a new result, due to Lieb and Seiringer [LS05], on the concavity of a certain trace function on the cone of positive-definite matrices.

First we observe that the Courant–Fischer theorem allows us relate the behavior of the kth eigenvalue of a matrix to the behavior of the largest eigenvalue of an appropriate compression of the matrix.

Theorem 3.1. *Let $\boldsymbol{X}$ be a random self-adjoint matrix with dimension n, and let $k \leq n$ be an integer. Then, for all $t \in \mathbb{R}$,*

$$\mathbb{P}\left\{\lambda_k(\boldsymbol{X}) \geq t\right\} \leq \inf_{\theta > 0} \min_{\boldsymbol{V} \in \mathbb{V}_{n-k+1}^n} \left\{ \mathrm{e}^{-\theta t} \cdot \mathbb{E} \operatorname{tr} \mathrm{e}^{\theta \boldsymbol{V}^* \boldsymbol{X} \boldsymbol{V}} \right\}. \tag{3.1}$$

Proof. Let θ be a fixed positive number. Then

$$\begin{aligned} \mathbb{P}\left\{\lambda_k(\boldsymbol{X}) \geq t\right\} = \mathbb{P}\left\{\lambda_k(\theta \boldsymbol{X}) \geq \theta t\right\} &= \mathbb{P}\left\{ \mathrm{e}^{\lambda_k(\theta \boldsymbol{X})} \geq \mathrm{e}^{\theta t} \right\} \\ &\leq \mathrm{e}^{-\theta t} \cdot \mathbb{E} \mathrm{e}^{\lambda_k(\theta \boldsymbol{X})} = \mathrm{e}^{-\theta t} \cdot \mathbb{E} \exp\left\{ \min_{\boldsymbol{V} \in \mathbb{V}_{n-k+1}^n} \lambda_{\max}\left(\theta \boldsymbol{V}^* \boldsymbol{X} \boldsymbol{V}\right) \right\}. \end{aligned}$$

The first identity follows from the positive homogeneity of eigenvalue maps and the second from the monotonicity of the scalar exponential function. The final two relations are Markov's inequality and (2.1).

To continue, we need to bound the expectation. Interchange the order of the exponential and the minimum; then apply the spectral mapping theorem to see that

$$\begin{aligned} \mathbb{E} \exp\left\{ \min_{\boldsymbol{V} \in \mathbb{V}_{n-k+1}^n} \lambda_{\max}\left(\theta \boldsymbol{V}^* \boldsymbol{X} \boldsymbol{V}\right) \right\} &= \mathbb{E} \min_{\boldsymbol{V} \in \mathbb{V}_{n-k+1}^n} \lambda_{\max}\left(\exp(\theta \boldsymbol{V}^* \boldsymbol{X} \boldsymbol{V})\right) \\ &\leq \min_{\boldsymbol{V} \in \mathbb{V}_{n-k+1}^n} \mathbb{E} \lambda_{\max}\left(\exp(\theta \boldsymbol{V}^* \boldsymbol{X} \boldsymbol{V})\right) \\ &\leq \min_{\boldsymbol{V} \in \mathbb{V}_{n-k+1}^n} \mathbb{E} \operatorname{tr} \exp(\theta \boldsymbol{V}^* \boldsymbol{X} \boldsymbol{V}). \end{aligned}$$

The first inequality is Jensen's. The second inequality follows because the exponential of a self-adjoint matrix is positive definite, so its largest eigenvalue is smaller than its trace.

Combine these observations and take the infimum over all positive θ to complete the argument. □

We are interested in the case where the matrix $\boldsymbol{X}$ in Theorem 3.1 can be expressed as a sum of independent random matrices. In this case, we use the following result to develop the right-hand side of the Laplace transform bound (3.1).

Theorem 3.2. *Consider a finite sequence $\{\boldsymbol{X}_j\}$ of independent, random, self-adjoint matrices with dimension n and a sequence $\{\boldsymbol{A}_j\}$ of fixed self-adjoint matrices with dimension n that satisfy the relations*

$$\mathbb{E}\mathrm{e}^{\boldsymbol{X}_j} \preceq \mathrm{e}^{\boldsymbol{A}_j}. \tag{3.2}$$

Let $\boldsymbol{V} \in \mathbb{V}_k^n$ be an isometric embedding of $\mathbb{C}^k$ into $\mathbb{C}^n$ for some $k \le n$. Then

$$\mathbb{E}\operatorname{tr}\exp\left\{\sum\nolimits_j \boldsymbol{V}^*\boldsymbol{X}_j\boldsymbol{V}\right\} \le \operatorname{tr}\exp\left\{\sum\nolimits_j \boldsymbol{V}^*\boldsymbol{A}_j\boldsymbol{V}\right\}. \tag{3.3}$$

In particular,

$$\mathbb{E}\operatorname{tr}\exp\left\{\sum\nolimits_j \boldsymbol{X}_j\right\} \le \operatorname{tr}\exp\left\{\sum\nolimits_j \boldsymbol{A}_j\right\}. \tag{3.4}$$

Theorem 3.2 is an extension of Lemma 3.4 of [Tro11c], which establishes the special case (3.4). The proof depends upon a recent result due to Lieb and Seiringer [LS05, Thm. 3] that extends Lieb's earlier result [Lie73, Thm. 6].

Proposition 3.1 (Lieb–Seiringer 2005)**.** *Let $\boldsymbol{H}$ be a self-adjoint matrix with dimension k. Let $\boldsymbol{V} \in \mathbb{V}_k^n$ be an isometric embedding of $\mathbb{C}^k$ into $\mathbb{C}^n$ for some $k \le n$. Then the function*

$$\boldsymbol{A} \longmapsto \operatorname{tr}\exp\{\boldsymbol{H} + \boldsymbol{V}^*(\log \boldsymbol{A})\boldsymbol{V}\}$$

is concave on the cone of positive-definite matrices in $\mathbb{M}_{\mathrm{sa}}^n$.

Proof of Theorem 3.2. First, note that (3.2) and the operator monotonicity of the matrix logarithm yield the following inequality for each k:

$$\log \mathbb{E}\mathrm{e}^{\boldsymbol{X}_k} \preceq \boldsymbol{A}_k. \tag{3.5}$$

Let $\mathbb{E}_k$ denote expectation conditioned on the first k summands, $\boldsymbol{X}_1$ through $\boldsymbol{X}_k$. Then

$$\begin{aligned}
\mathbb{E}\operatorname{tr}\exp\left\{\sum\nolimits_{j\le\ell} \boldsymbol{V}^*\boldsymbol{X}_j\boldsymbol{V}\right\} &= \mathbb{E}\mathbb{E}_1\cdots\mathbb{E}_{\ell-1}\operatorname{tr}\exp\left\{\sum\nolimits_{j\le\ell-1} \boldsymbol{V}^*\boldsymbol{X}_j\boldsymbol{V} + \boldsymbol{V}^*\left(\log \mathrm{e}^{\boldsymbol{X}_\ell}\right)\boldsymbol{V}\right\} \\
&\le \mathbb{E}\mathbb{E}_1\cdots\mathbb{E}_{\ell-2}\operatorname{tr}\exp\left\{\sum\nolimits_{j\le\ell-1} \boldsymbol{V}^*\boldsymbol{X}_j\boldsymbol{V} + \boldsymbol{V}^*\left(\log \mathbb{E}\mathrm{e}^{\boldsymbol{X}_\ell}\right)\boldsymbol{V}\right\} \\
&\le \mathbb{E}\mathbb{E}_1\cdots\mathbb{E}_{\ell-2}\operatorname{tr}\exp\left\{\sum\nolimits_{j\le\ell-1} \boldsymbol{V}^*\boldsymbol{X}_j\boldsymbol{V} + \boldsymbol{V}^*\left(\log \mathrm{e}^{\boldsymbol{A}_\ell}\right)\boldsymbol{V}\right\} \\
&= \mathbb{E}\mathbb{E}_1\cdots\mathbb{E}_{\ell-2}\operatorname{tr}\exp\left\{\sum\nolimits_{j\le\ell-1} \boldsymbol{V}^*\boldsymbol{X}_j\boldsymbol{V} + \boldsymbol{V}^*\boldsymbol{A}_\ell\boldsymbol{V}\right\}.
\end{aligned}$$

The first inequality follows from Proposition 3.1 and Jensen's inequality, and the second depends on (3.5) and the monotonicity of the trace exponential. Iterate this argument to complete the proof. □

Our main result follows from combining Theorem 3.1 and Theorem 3.2.

Theorem 3.3 (Minimax Laplace Transform)**.** *Consider a finite sequence $\{\boldsymbol{X}_j\}$ of independent, random, self-adjoint matrices with dimension n, and let $k \le n$ be an integer.*

(i) Let $\{\boldsymbol{A}_j\}$ be a sequence of self-adjoint matrices that satisfy the semidefinite relations

$$\mathbb{E}\mathrm{e}^{\theta\boldsymbol{X}_j} \preceq \mathrm{e}^{g(\theta)\boldsymbol{A}_j}$$

where $g : (0,\infty) \to [0,\infty)$. Then, for all $t \in \mathbb{R}$,

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \boldsymbol{X}_j\right) \ge t\right\} \le \inf_{\theta>0} \min_{\boldsymbol{V}\in\mathbb{V}_{n-k+1}^n} \left[\mathrm{e}^{-\theta t}\cdot \operatorname{tr}\exp\left\{g(\theta)\sum\nolimits_j \boldsymbol{V}^*\boldsymbol{A}_j\boldsymbol{V}\right\}\right].$$

(ii) Let $\{\boldsymbol{A}_j : \mathbb{V}_{n-k+1}^n \to \mathbb{M}_{\mathrm{sa}}^n\}$ *be a sequence of functions that satisfy the semidefinite relations*

$$\mathbb{E}\mathrm{e}^{\theta \boldsymbol{V}^* \boldsymbol{X}_j \boldsymbol{V}} \preceq \mathrm{e}^{g(\theta)\boldsymbol{A}_j(\boldsymbol{V})}$$

for all $\boldsymbol{V} \in \mathbb{V}_{n-k+1}^n$, *where* $g : (0,\infty) \to [0,\infty)$. *Then, for all* $t \in \mathbb{R}$,

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \boldsymbol{X}_j\right) \geq t\right\} \leq \inf_{\theta>0} \min_{\boldsymbol{V}\in\mathbb{V}_{n-k+1}^n} \left[\mathrm{e}^{-\theta t} \cdot \operatorname{tr}\exp\left\{g(\theta)\sum\nolimits_j \boldsymbol{A}_j(\boldsymbol{V})\right\}\right].$$

The first bound in Theorem 3.3 requires less detailed information on how compression affects the summands but correspondingly does not give as sharp results as the second.

In the following two sections, we use the minimax Laplace transform method to derive Chernoff and Bernstein inequalities for the interior eigenvalues of a sum of independent random matrices. Tail bounds for the eigenvalues of matrix Rademacher and Gaussian series, eigenvalue Hoeffding, and matrix martingale eigenvalue tail bounds can all be derived in a similar manner; see [Tro11c] for relevant details.

4. Chernoff bounds

Classical Chernoff bounds establish that the tails of a sum of independent nonnegative random variables decay subexponentially. [Tro11c] develops Chernoff bounds for the maximum and minimum eigenvalues of a sum of independent positive-semidefinite matrices. We extend this analysis to study the interior eigenvalues.

Intuitively, the eigenvalue tail bounds should depend on how concentrated the summands are; e.g., the maximum eigenvalue of a sum of operators whose ranges are aligned is likely to vary more than that of a sum of operators whose ranges are orthogonal. To measure how much a finite sequence of random summands $\{\boldsymbol{X}_j\}$ concentrates in a given subspace, we define a function $\Psi : \bigcup_{1\leq k\leq n} \mathbb{V}_k^n \to \mathbb{R}$ that satisfies

$$\max\nolimits_j \lambda_{\max}\left(\boldsymbol{V}^*\boldsymbol{X}_j\boldsymbol{V}\right) \leq \Psi(\boldsymbol{V}) \qquad \text{almost surely for each } \boldsymbol{V} \in \bigcup\nolimits_{1\leq k\leq n} \mathbb{V}_k^n. \tag{4.1}$$

The sequence $\{\boldsymbol{X}_j\}$ associated with Ψ will always be clear from context. We have the following result.

Theorem 4.1 (Eigenvalue Chernoff Bounds)**.** *Consider a finite sequence* $\{\boldsymbol{X}_j\}$ *of independent, random, positive-semidefinite matrices with dimension* n. *Given an integer* $k \leq n$, *define*

$$\mu_k = \lambda_k\left(\sum\nolimits_j \mathbb{E}\boldsymbol{X}_j\right),$$

and let $\boldsymbol{V}_+ \in \mathbb{V}_{n-k+1}^n$ *and* $\boldsymbol{V}_- \in \mathbb{V}_k^n$ *be isometric embeddings that satisfy*

$$\mu_k = \lambda_{\max}\left(\sum\nolimits_j \boldsymbol{V}_+^*(\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_+\right) = \lambda_{\min}\left(\sum\nolimits_j \boldsymbol{V}_-^*(\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_-\right).$$

Then

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \boldsymbol{X}_j\right) \geq (1+\delta)\mu_k\right\} \leq (n-k+1)\cdot\left[\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_k/\Psi(\boldsymbol{V}_+)} \qquad \text{for } \delta > 0, \text{ and}$$

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \boldsymbol{X}_j\right) \leq (1-\delta)\mu_k\right\} \leq k\cdot\left[\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_k/\Psi(\boldsymbol{V}_-)} \qquad \text{for } \delta \in [0,1),$$

where Ψ *is a function that satisfies* (4.1).

Theorem 4.1 tells us how the tails of the kth eigenvalue are controlled by the variation of the random summands in the top and bottom invariant subspaces of $\sum_j \mathbb{E}\boldsymbol{X}_j$. Up to the dimensional factors k and $n-k+1$, the eigenvalues exhibit binomial-type tails. When $k = 1$ (respectively, $k = n$) Theorem 4.1 controls the probability that the largest eigenvalue of the sum is small (respectively, the probability that the smallest eigenvalue of the sum is large), thereby complementing the one-sided Chernoff bounds of [Tro11c].

Remark 4.1. If it is difficult to estimate $\Psi(\boldsymbol{V}_+)$ or $\Psi(\boldsymbol{V}_-)$, one can resort to the weaker estimates

$$\Psi(\boldsymbol{V}_+) \le \max_{\boldsymbol{V} \in \mathbb{V}_{n-k+1}^n} \max_j \|\boldsymbol{V}^*\boldsymbol{X}_j\boldsymbol{V}\| = \max_j \|\boldsymbol{X}_j\|$$
$$\Psi(\boldsymbol{V}_-) \le \max_{\boldsymbol{V} \in \mathbb{V}_k^n} \max_j \|\boldsymbol{V}^*\boldsymbol{X}_j\boldsymbol{V}\| = \max_j \|\boldsymbol{X}_j\|.$$

Theorem 4.1 follows from Theorem 3.3 using an appropriate bound on the matrix moment generating functions. The following lemma is due to Ahlswede and Winter [AW02]; see also [Tro11c, Lem. 5.8].

Lemma 4.2. *Suppose that $\boldsymbol{X}$ is a random positive-semidefinite matrix that satisfies $\lambda_{\max}(\boldsymbol{X}) \le 1$. Then*

$$\mathbb{E}\mathrm{e}^{\theta\boldsymbol{X}} \preceq \exp\Big((\mathrm{e}^\theta - 1)(\mathbb{E}\boldsymbol{X})\Big) \qquad \text{for } \theta \in \mathbb{R}.$$

Proof of Theorem 4.1, upper bound. We consider the case where $\Psi(\boldsymbol{V}_+) = 1$; the general case follows by homogeneity. Define

$$\boldsymbol{A}_j(\boldsymbol{V}_+) = \boldsymbol{V}_+^*(\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_+ \quad \text{and} \quad g(\theta) = \mathrm{e}^\theta - 1.$$

Theorem 3.3(ii) and Lemma 4.2 imply that

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \boldsymbol{X}_j\right) \ge (1+\delta)\mu_k\right\} \le \inf_{\theta>0} \mathrm{e}^{-\theta(1+\delta)\mu_k} \cdot \operatorname{tr}\exp\left\{g(\theta)\sum\nolimits_j \boldsymbol{V}_+^*(\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_+\right\}.$$

Bound the trace by the maximum eigenvalue, taking into account the reduced dimension of the summands:

$$\begin{aligned}\operatorname{tr}\exp\left\{g(\theta)\sum\nolimits_j \boldsymbol{V}_+^*(\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_+\right\} &\le (n-k+1)\cdot\lambda_{\max}\left(\exp\left\{g(\theta)\sum\nolimits_j \boldsymbol{V}_+^*(\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_+\right\}\right)\\ &= (n-k+1)\cdot\exp\left\{g(\theta)\cdot\lambda_{\max}\left(\sum\nolimits_j \boldsymbol{V}_+^*(\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_+\right)\right\}.\end{aligned}$$

The equality follows from the spectral mapping theorem. Identify the quantity μ_k; then combine the last two inequalities to obtain

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \boldsymbol{X}_j\right) \ge (1+\delta)\mu_k\right\} \le (n-k+1)\cdot\inf_{\theta>0}\mathrm{e}^{[g(\theta)-\theta(1+\delta)]\mu_k}.$$

The right-hand side is minimized when $\theta = \log(1+\delta)$, which gives the desired upper tail bound. □

Proof of Theorem 4.1, lower bound. As before, we consider the case where $\Psi(\boldsymbol{V}_-) = 1$. Clearly,

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \boldsymbol{X}_j\right) \le (1-\delta)\mu_k\right\} = \mathbb{P}\left\{\lambda_{n-k+1}\left(\sum\nolimits_j -\boldsymbol{X}_j\right) \ge -(1-\delta)\mu_k\right\}. \tag{4.2}$$

Apply Lemma 4.2 to see that, for $\theta > 0$,

$$\mathbb{E}\mathrm{e}^{\theta(-\boldsymbol{V}_-^*\boldsymbol{X}_j\boldsymbol{V}_-)} = \mathbb{E}\mathrm{e}^{(-\theta)\boldsymbol{V}_-^*\boldsymbol{X}_j\boldsymbol{V}_-} \preceq \exp\big(g(\theta)\cdot\boldsymbol{V}_-^*(-\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_-\big),$$

where $g(\theta) = 1 - \mathrm{e}^{-\theta}$. Theorem 3.3(ii) thus implies that the latter probability in (4.2) is bounded by

$$\inf_{\theta>0}\mathrm{e}^{\theta(1-\delta)\mu_k}\cdot\operatorname{tr}\exp\left\{g(\theta)\sum\nolimits_j \boldsymbol{V}_-^*(-\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_-\right\}.$$

Using reasoning analogous to that in the proof of the upper bound, we justify the first of the following inequalities:

$$\begin{aligned}\operatorname{tr}\exp\left\{g(\theta)\sum\nolimits_j \boldsymbol{V}_-^*(-\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_-\right\} &\le k\cdot\exp\left\{\lambda_{\max}\left(g(\theta)\sum\nolimits_j \boldsymbol{V}_-^*(-\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_-\right)\right\}\\ &= k\cdot\exp\left\{-g(\theta)\cdot\lambda_{\min}\left(\sum\nolimits_j \boldsymbol{V}_-^*(\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_-\right)\right\}\\ &= k\cdot\exp\left\{-g(\theta)\mu_k\right\}.\end{aligned}$$

The remaining equalities follow from the fact that $-g(\theta) < 0$ and the definition of μ_k.

This argument establishes the bound

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \boldsymbol{X}_j\right) \le (1-\delta)\mu_k\right\} \le k\cdot\inf_{\theta>0}\mathrm{e}^{[\theta(1-\delta)-g(\theta)]\mu_k}.$$

The right-hand side is minimized when $\theta = -\log(1-\delta)$, which gives the desired lower tail bound. □

5. Bennett and Bernstein inequalities

The classical Bennett and Bernstein inequalities use the variance or knowledge of the moments of the summands to control the probability that a sum of independent random variables deviates from its mean. In [Tro11c], matrix Bennett and Bernstein inequalities are developed for the extreme eigenvalues of self-adjoint random matrix sums. We establish that the interior eigenvalues satisfy analogous inequalities.

As in the derivation of the Chernoff inequalities of section 4, we need a measure of how concentrated the random summands are in a given subspace. Recall that the function $\Psi : \bigcup_{1\le k\le n}\mathbb{V}_k^n \to \mathbb{R}$ satisfies

$$\max\nolimits_j \lambda_{\max}\left(\boldsymbol{V}^*\boldsymbol{X}_j\boldsymbol{V}\right) \le \Psi(\boldsymbol{V}) \qquad \text{almost surely for each } \boldsymbol{V}\in\bigcup\nolimits_{1\le k\le n}\mathbb{V}_k^n. \tag{5.1}$$

The sequence $\{\boldsymbol{X}_j\}$ associated with Ψ will always be clear from context.

Theorem 5.1 (Eigenvalue Bennett Inequality)**.** *Consider a finite sequence $\{\boldsymbol{X}_j\}$ of independent, random, self-adjoint matrices with dimension n, all of which have zero mean. Given an integer $k \le n$, define*

$$\sigma_k^2 = \lambda_k\left(\sum\nolimits_j \mathbb{E}(\boldsymbol{X}_j^2)\right).$$

Choose $\boldsymbol{V}_+ \in \mathbb{V}_{n-k+1}^n$ to satisfy

$$\sigma_k^2 = \lambda_{\max}\left(\sum\nolimits_j \boldsymbol{V}_+^*\mathbb{E}(\boldsymbol{X}_j^2)\boldsymbol{V}_+\right).$$

Then, for all $t \ge 0$,

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \boldsymbol{X}_j\right) \ge t\right\} \le (n-k+1)\cdot\exp\left\{-\frac{\sigma_k^2}{\Psi(\boldsymbol{V}_+)^2}\cdot h\left(\frac{\Psi(\boldsymbol{V}_+)t}{\sigma_k^2}\right)\right\} \tag{i}$$

$$\le (n-k+1)\cdot\exp\left\{\frac{-t^2/2}{\sigma_k^2+\Psi(\boldsymbol{V}_+)t/3}\right\} \tag{ii}$$

$$\le \begin{cases}(n-k+1)\cdot\exp\left\{-\frac{3}{8}t^2/\sigma_k^2\right\} & \text{for } t \le \sigma_k^2/\Psi(\boldsymbol{V}_+)\\ (n-k+1)\cdot\exp\left\{-\frac{3}{8}t/\Psi(\boldsymbol{V}_+)\right\} & \text{for } t \ge \sigma_k^2/\Psi(\boldsymbol{V}_+),\end{cases} \tag{iii}$$

where the function $h(u) = (1+u)\log(1+u) - u$ for $u \ge 0$. The function Ψ satisfies (5.1) *above.*

Results (i) and (ii) are, respectively, matrix analogs of the classical Bennett and Bernstein inequalities. As in the scalar case, the Bennett inequality reflects a Poisson-type decay in the tails of the eigenvalues. The Bernstein inequality states that small deviations from the eigenvalues of

the expected matrix are roughly normally distributed while larger deviations are subexponential. The split Bernstein inequalities (iii) make explicit the division between these two regimes.

As stated, Theorem 5.1 estimates the probability that the eigenvalues of a sum are large. Using the identity

$$\lambda_k\left(\sum\nolimits_j \boldsymbol{X}_j\right) = -\lambda_{n-k+1}\left(-\sum\nolimits_j \boldsymbol{X}_j\right),$$

Theorem 5.1 can be applied to estimate the probability that eigenvalues of a sum are small.

To prove Theorem 5.1, we use the following lemma (Lemma 6.7 in [Tro11c]) to control the moment generating function of a random matrix with bounded maximum eigenvalue.

Lemma 5.2. *Let $\boldsymbol{X}$ be a random self-adjoint matrix satisfying $\mathbb{E}\boldsymbol{X} = \mathbf{0}$ and $\lambda_{\max}(\boldsymbol{X}) \le 1$ almost surely. Then*

$$\mathbb{E}\mathrm{e}^{\theta \boldsymbol{X}} \preceq \exp((\mathrm{e}^{\theta} - \theta - 1) \cdot \mathbb{E}(\boldsymbol{X}^2)) \quad \text{for } \theta > 0.$$

Proof of Theorem 5.1. Using homogeneity, we assume without loss that $\Psi(\boldsymbol{V}_+) = 1$. This implies that $\lambda_{\max}(\boldsymbol{X}_j) \le 1$ almost surely for all the summands. By Lemma 5.2,

$$\mathbb{E}\mathrm{e}^{\theta \boldsymbol{X}_j} \preceq \exp\left(g(\theta) \cdot \mathbb{E}(\boldsymbol{X}_j^2)\right),$$

with $g(\theta) = \mathrm{e}^{\theta} - \theta - 1$.

Theorem 3.3(i) then implies

$$\begin{aligned}
\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \boldsymbol{X}_j\right) \ge t\right\} &\le \inf_{\theta>0} \mathrm{e}^{-\theta t} \cdot \operatorname{tr}\exp\left\{g(\theta)\sum\nolimits_j \boldsymbol{V}_+^*\mathbb{E}(\boldsymbol{X}_j^2)\boldsymbol{V}_+\right\} \\
&\le (n-k+1)\cdot \inf_{\theta>0} \mathrm{e}^{-\theta t} \cdot \lambda_{\max}\left(\exp\left\{g(\theta)\sum\nolimits_j \boldsymbol{V}_+^*\mathbb{E}(\boldsymbol{X}_j^2)\boldsymbol{V}_+\right\}\right) \\
&= (n-k+1)\cdot \inf_{\theta>0} e^{-\theta t} \cdot \exp\left\{g(\theta)\cdot \lambda_{\max}\left(\sum\nolimits_j \boldsymbol{V}_+^*\mathbb{E}(\boldsymbol{X}_j^2)\boldsymbol{V}_+\right)\right\}.
\end{aligned}$$

The maximum eigenvalue in this expression equals σ_k^2, thus

$$\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \boldsymbol{X}_j\right) \ge t\right\} \le (n-k+1)\cdot \inf_{\theta>0} \mathrm{e}^{g(\theta)\sigma_k^2 - \theta t}.$$

The Bennett inequality (i) follows by substituting $\theta = \log(1 + t/\sigma_k^2)$ into the right-hand side and simplifying.

The Bernstein inequality (ii) is a consequence of (i) and the fact that

$$h(u) \ge \frac{u^2/2}{1 + u/3} \quad \text{for } u \ge 0,$$

which can be established by comparing derivatives.

The subgaussian and subexponential portions of the split Bernstein inequalities (iii) are verified through algebraic comparisons on the relevant intervals. □

Occasionally, as in the application in section 7 to the problem of covariance matrix estimation, one desires a Bernstein-type tail bound that applies to summands that do not have bounded maximum eigenvalues. In this case, if the moments of the summands satisfy sufficiently strong growth restrictions, one can extend classical scalar arguments to obtain results such as the following Bernstein bound for subexponential matrices.

Theorem 5.3 (Eigenvalue Bernstein Inequality for Subexponential Matrices)**.** *Consider a finite sequence $\{\boldsymbol{X}_j\}$ of independent, random, self-adjoint matrices with dimension n, all of which satisfy the subexponential moment growth condition*

$$\mathbb{E}(\boldsymbol{X}_j^m) \preceq \frac{m!}{2} B^{m-2} \boldsymbol{\Sigma}_j^2 \quad \text{for } m = 2, 3, 4, \ldots,$$

where B is a positive constant and $\mathbf{\Sigma}_j^2$ are positive-semidefinite matrices. Given an integer $k \leq n$, set

$$\mu_k = \lambda_k \left(\sum\nolimits_j \mathbb{E}\boldsymbol{X}_j \right).$$

Choose $\boldsymbol{V}_+ \in \mathbb{V}_{n-k+1}^n$ that satisfies

$$\mu_k = \lambda_{\max} \left(\sum\nolimits_j \boldsymbol{V}_+^* (\mathbb{E}\boldsymbol{X}_j) \boldsymbol{V}_+ \right),$$

and define

$$\sigma_k^2 = \lambda_{\max} \left(\sum\nolimits_j \boldsymbol{V}_+^* \mathbf{\Sigma}_j^2 \boldsymbol{V}_+ \right).$$

Then, for any $t \geq 0$,

$$\mathbb{P}\left\{ \lambda_k \left(\sum\nolimits_j \boldsymbol{X}_j \right) \geq \mu_k + t \right\} \leq (n-k+1) \cdot \exp\left\{ -\frac{t^2/2}{\sigma_k^2 + Bt} \right\} \tag{i}$$

$$\leq \begin{cases} (n-k+1) \cdot \exp\left\{ -\frac{1}{4} t^2/\sigma_k^2 \right\} & \text{for } t \leq \sigma_k^2/B \\ (n-k+1) \cdot \exp\left\{ -\frac{1}{4} t/B \right\} & \text{for } t \geq \sigma_k^2/B. \end{cases} \tag{ii}$$

This result is an extension of [Tro11c, Theorem 6.2], which, in turn, generalizes a classical scalar argument [DG98].

As with the other matrix inequalities, Theorem 5.3 follows from an application of Theorem 3.3 and appropriate semidefinite bounds on the moment generating functions of the summands. Thus, the key to the proof lies in exploiting the moment growth conditions of the summands to majorize their moment generating functions. The following lemma, a trivial extension of Lemma 6.8 in [Tro11c], provides what we need.

Lemma 5.4. *Let $\boldsymbol{X}$ be a random self-adjoint matrix satisfying the subexponential moment growth conditions*

$$\mathbb{E}(\boldsymbol{X}^m) \preceq \frac{m!}{2} \mathbf{\Sigma}^2 \quad \text{for } m = 2, 3, 4, \ldots.$$

Then, for any θ in $[0, 1)$,

$$\mathbb{E} \exp(\theta \boldsymbol{X}) \preceq \exp\left(\theta \mathbb{E}\boldsymbol{X} + \frac{\theta^2}{2(1-\theta)} \mathbf{\Sigma}^2 \right).$$

Proof of Theorem 5.3. We note that $\boldsymbol{X}_j$ satisfies the growth condition

$$\mathbb{E}(\boldsymbol{X}_j^m) \preceq \frac{m!}{2} B^{m-2} \mathbf{\Sigma}_j^2 \quad \text{for } m \geq 2$$

if and only if the scaled matrix $\boldsymbol{X}_j / B$ satisfies

$$\mathbb{E}\left(\frac{\boldsymbol{X}_j}{B} \right)^m \preceq \frac{m!}{2} \cdot \frac{\mathbf{\Sigma}_j^2}{B^2} \quad \text{for } m \geq 2.$$

Thus, by rescaling, it suffices to consider the case $B = 1$. We now do so.

By Lemma 5.4, the moment generating functions of the summands satisfy

$$\mathbb{E} \exp(\theta \boldsymbol{X}_j) \preceq \exp\left(\theta \mathbb{E}\boldsymbol{X}_j + g(\theta) \mathbf{\Sigma}_j^2 \right),$$

where $g(\theta) = \theta^2/(2-2\theta)$. Now we apply Theorem 3.3(i):

$$\begin{aligned}\mathbb{P}\left\{\lambda_k\left(\sum\nolimits_j \boldsymbol{X}_j\right) \geq \mu_k + t\right\} &\leq \inf_{\theta\in[0,1)} \mathrm{e}^{-\theta(\mu_k+t)} \cdot \operatorname{tr}\exp\left\{\theta\sum\nolimits_j \boldsymbol{V}_+^*(\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_+ + g(\theta)\sum\nolimits_j \boldsymbol{V}_+^*\boldsymbol{\Sigma}_j^2\boldsymbol{V}_+\right\} \\ &\leq \inf_{\theta\in[0,1)} (n-k+1)\cdot\exp\Big\{-\theta(\mu_k+t) + \theta\cdot\lambda_{\max}\left(\sum\nolimits_j \boldsymbol{V}_+^*(\mathbb{E}\boldsymbol{X}_j)\boldsymbol{V}_+\right) \\ &\qquad + g(\theta)\cdot\lambda_{\max}\left(\sum\nolimits_j \boldsymbol{V}_+^*\boldsymbol{\Sigma}_j^2\boldsymbol{V}_+\right)\Big\} \\ &= \inf_{\theta\in[0,1)} (n-k+1)\cdot\exp\left(-\theta t + g(\theta)\sigma_k^2\right).\end{aligned}$$

To achieve the final simplification, we identified μ_k and σ_k^2. Now, select $\theta = t/(t+\sigma_k^2)$. Then simplication gives the Bernstein inequality (i).

Algebraic comparisons on the relevant intervals yield the split Bernstein inequalities (ii). □

6. An application to column subsampling

As an application of our Chernoff bounds, we examine how sampling columns from a matrix with orthonormal rows affects the spectrum. This question has applications in numerical linear algebra and compressed sensing. The special cases of the maximum and minimum eigenvalues have been studied in the literature [Tro08, RV07]. The limiting spectral distributions of matrices formed by sampling columns from similarly structured matrices have also been studied: the results of [GH10] apply to matrices formed by sampling columns from any fixed orthogonal matrix, and [Far10] studies matrices formed by sampling columns and rows from the discrete Fourier transform matrix. We mention in particular [Rud99], the main result of which provides a uniform bound on the tails of all singular values of the sampled matrix. The theorem proven in this section provides bounds which reflect the differences in the tails of the individual singular values, and thus can be viewed as an elaboration of the result in [Rud99].

Let $\boldsymbol{U}$ be an $n \times r$ matrix with orthonormal rows. We model the sampling operation using a random diagonal matrix $\boldsymbol{D}$ whose entries are independent Bern(p) random variables. Then the random matrix

$$\widehat{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{D} \tag{6.1}$$

can be interpreted as a random column submatrix of $\boldsymbol{U}$ with an average of pr nonzero columns. Our goal is to study the behavior of the spectrum of $\widehat{\boldsymbol{U}}$.

Recall that the jth column of $\boldsymbol{U}$ is written $\boldsymbol{u}_j$. Consider the following coherence-like quantity associated with $\boldsymbol{U}$:

$$\tau_k = \min_{\boldsymbol{V}\in\mathbb{V}_k^n} \max\nolimits_j \|\boldsymbol{V}^*\boldsymbol{u}_j\|^2 \quad \text{for } k = 1,\dots,n. \tag{6.2}$$

There does not seem to be a simple expression for τ_k. However, by choosing $\boldsymbol{V}^*$ to be the restriction to an appropriate k-dimensional coordinate subspace, we see that τ_k always satisfies

$$\tau_k \leq \min_{|I|\leq k} \max\nolimits_j \sum\nolimits_{i\in I} u_{ij}^2.$$

The following theorem shows that the behavior of $s_k(\widehat{\boldsymbol{U}})$, the kth singular value of $\widehat{\boldsymbol{U}}$, can be explained in terms of τ_k.

Theorem 6.1 (Column Subsampling of Matrices with Orthonormal Rows)**.** *Let $\boldsymbol{U}$ be an $n \times r$ matrix with orthonormal rows, and let p be a sampling probability. Define the sampled matrix $\widehat{\boldsymbol{U}}$*

according to (6.1)*, and the numbers* $\{\tau_k\}$ *according to* (6.2)*. Then, for each* $k = 1, \dots, n$,

$$\mathbb{P}\left\{ s_k(\widehat{\boldsymbol{U}}) \geq \sqrt{(1+\delta)p} \right\} \leq (n-k+1) \cdot \left[\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}} \right]^{p/\tau_{n-k+1}} \qquad \text{for } \delta > 0$$

$$\mathbb{P}\left\{ s_k(\widehat{\boldsymbol{U}}) \leq \sqrt{(1-\delta)p} \right\} \leq k \cdot \left[\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}} \right]^{p/\tau_k} \qquad \text{for } \delta \in [0,1).$$

Proof. Observe, using (6.1), that

$$s_k(\widehat{\boldsymbol{U}})^2 = \lambda_k(\boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^*) = \lambda_k\left(\sum\nolimits_j d_j \boldsymbol{u}_j \boldsymbol{u}_j^* \right),$$

where $\boldsymbol{u}_j$ is the jth column of $\boldsymbol{U}$ and $d_j \sim \mathrm{Bern}(p)$. Compute

$$\mu_k = \lambda_k\left(\sum\nolimits_j \mathbb{E} d_j \boldsymbol{u}_j \boldsymbol{u}_j^* \right) = p \cdot \lambda_k(\boldsymbol{U}\boldsymbol{U}^*) = p \cdot \lambda_k(\mathbf{I}) = p.$$

It follows that, for *any* $\boldsymbol{V} \in \mathbb{V}_{n-k+1}^n$,

$$\lambda_{\max}\left(\sum\nolimits_j \boldsymbol{V}^*(\mathbb{E} d_j \boldsymbol{u}_j \boldsymbol{u}_j^*)\boldsymbol{V} \right) = p \cdot \lambda_{\max}\left(\boldsymbol{V}^*\boldsymbol{V}\right) = p = \mu_k,$$

so the choice of $\boldsymbol{V}_+ \in \mathbb{V}_{n-k+1}^n$ is arbitrary. Similarly, the choice of $\boldsymbol{V}_- \in \mathbb{V}_k^n$ is arbitrary. We select $\boldsymbol{V}_+$ to be an isometric embedding that achieves τ_{n-k+1} and $\boldsymbol{V}_-$ to be an isometric embedding that achieves τ_k. Accordingly,

$$\Psi(\boldsymbol{V}_+) = \max_j \|\boldsymbol{V}_+^* \boldsymbol{u}_j \boldsymbol{u}_j^* \boldsymbol{V}_+\| = \max_j \|\boldsymbol{V}_+^* \boldsymbol{u}_j\|^2 = \tau_{n-k+1}, \quad \text{and}$$
$$\Psi(\boldsymbol{V}_-) = \max_j \|\boldsymbol{V}_-^* \boldsymbol{u}_j \boldsymbol{u}_j^* \boldsymbol{V}_-\| = \max_j \|\boldsymbol{V}_-^* \boldsymbol{u}_j\|^2 = \tau_k.$$

Theorem 4.1 delivers the upper bound

$$\mathbb{P}\left\{ s_k(\hat{\boldsymbol{U}}) \geq \sqrt{(1+\delta)p} \right\} = \mathbb{P}\left\{ \lambda_k\left(\sum\nolimits_j d_j \boldsymbol{u}_j \boldsymbol{u}_j^* \right) \geq (1+\delta)p \right\} \leq (n-k+1) \cdot \left[\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}} \right]^{p/\tau_{n-k+1}}$$

for $\delta > 0$ and the lower bound

$$\mathbb{P}\left\{ s_k(\hat{\boldsymbol{U}}) \leq \sqrt{(1-\delta)p} \right\} = \mathbb{P}\left\{ \lambda_k\left(\sum\nolimits_j d_j \boldsymbol{u}_j \boldsymbol{u}_j^* \right) \leq (1-\delta)p \right\} \leq k \cdot \left[\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}} \right]^{p/\tau_k}$$

for $\delta \in [0,1)$. □

To illustrate the discriminatory power of these bounds, let $\boldsymbol{U}$ be an $n \times n^2$ matrix consisting of n rows of the $n^2 \times n^2$ Fourier matrix and choose $p = (\log n)/n$ so that, on average, sampling reduces the aspect ratio from n to $\log n$. For $n = 100$, we determine upper and lower bounds for the median value of $s_k(\widehat{\boldsymbol{U}})$ by numerically finding the value of δ where the probability bounds in Theorem 6.1 equal 1/2. Figure 1 plots the empirical median value along with the computed interval. We see that these ranges reflect the behavior of the singular values more faithfully than the simple estimates $s_k(\mathbb{E}\widehat{\boldsymbol{U}}) = p$.

7. Covariance Estimation

We conclude with an extended example that illustrates how this circle of ideas allows one to answer interesting statistical questions. Specifically, we investigate the convergence of the individual eigenvalues of sample covariance matrices, with errors measured in *relative* precision.

Covariance estimation is a basic and ubiquitious problem that arises in signal processing, graphical modeling, machine learning, and genomics, among other areas. Let $\{\boldsymbol{\eta}_j\}_{j=1}^n \subset \mathbb{R}^p$ be i.i.d. samples drawn from some distribution with zero mean and covariance matrix $\boldsymbol{C}$. Define the sample covariance matrix

$$\widehat{\boldsymbol{C}}_n = \frac{1}{n} \sum\nolimits_{j=1}^n \boldsymbol{\eta}_j \boldsymbol{\eta}_j^*.$$

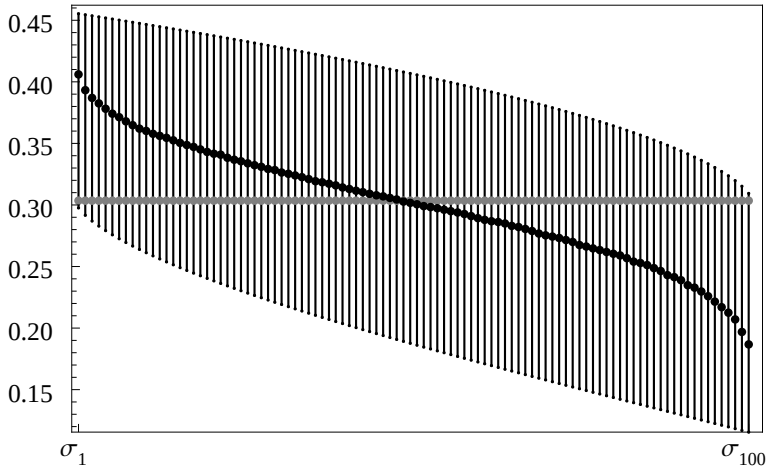

FIGURE 1. [Spectrum of a random submatrix] The matrix $\boldsymbol{U}$ is a $10^2 \times 10^4$ submatrix of the unitary DFT matrix with dimension 10^4, and the sampling probability $p = 10^{-4} \log(10^4)$. The kth vertical bar, calculated using Theorem 6.1, describes an interval containing the median value of the kth singular value of the sampled matrix $\widehat{\boldsymbol{U}}$. The black circles denote the empirical medians of the singular values of $\widehat{\boldsymbol{U}}$, calculated from 500 trials. The gray circles represent the singular values of $\mathbb{E}\widehat{\boldsymbol{U}}$.

An important challenge is to determine how many samples are needed to ensure that the empirical covariance estimator has a fixed relative accuracy in the spectral norm. That is, given a fixed ε, how large must n be so that

$$\|\widehat{\boldsymbol{C}}_n - \boldsymbol{C}\| \leq \varepsilon \|\boldsymbol{C}\|? \tag{7.1}$$

This estimation problem has been studied extensively. It is now known that for distributions with a finite second moment, $\Omega(p \log p)$ samples suffice [Rud99], and for log-concave distributions, $\Omega(p)$ samples suffice [ALPTJ11]. More broadly, Vershynin [Ver11] conjectures that, for distributions with finite fourth moment, $\Omega(p)$ samples suffice; he establishes this result to within iterated log factors. In [SV11], Srivastava and Vershynin establish that $\Omega(p)$ samples suffice for distributions which have finite $2 + \varepsilon$ moments, for some $\varepsilon > 0$, and satisfy an additional regularity condition.

Inequality (7.1) ensures that the difference between the kth eigenvalues of $\widehat{\boldsymbol{C}}_n$ and $\boldsymbol{C}$ is small, but it requires $\mathrm{O}(p)$ measurements to obtain estimates of even a few of the eigenvalues. Specifically, letting $\kappa_\ell = \lambda_1(\boldsymbol{C})/\lambda_\ell(\boldsymbol{C})$, we see that $O(\varepsilon^{-2}\kappa_\ell^2 p)$ measurements are required to obtain relative-error estimates of the dominant ℓ eigenvalues of $\boldsymbol{C}$ using the results of [ALPTJ11, Ver11, SV11]. However, it is reasonable to expect that when the spectrum of $\boldsymbol{C}$ exhibits decay and $\ell \ll p$, much fewer than $\mathrm{O}(p)$ measurements should suffice for relative-error recovery of the dominant ℓ eigenvalues.

In this section, we derive a relative approximation bound for each eigenvalue of $\boldsymbol{C}$ that allows us to confirm this intuition. For simplicity we assume the samples are drawn from a $\mathcal{N}(\mathbf{0}, \boldsymbol{C})$ distribution where $\boldsymbol{C}$ is full-rank, but the arguments can be extended to cover other distributions.

Theorem 7.1. *Assume that $\boldsymbol{C} \in \mathbb{M}_{\mathrm{sa}}^p$ is positive definite. Let $\{\boldsymbol{\eta}_j\}_{j=1}^n \subset \mathbb{R}^p$ be i.i.d. samples drawn from a $\mathcal{N}(\mathbf{0}, \boldsymbol{C})$ distribution. Define*

$$\widehat{\boldsymbol{C}}_n = \frac{1}{n} \sum\nolimits_{j=1}^n \boldsymbol{\eta}_j \boldsymbol{\eta}_j^*.$$

Write λ_k for the kth eigenvalue of $\boldsymbol{C}$, and write $\hat{\lambda}_k$ for the kth eigenvalue of $\widehat{\boldsymbol{C}}_n$. Then for $k = 1, \dots, p$,

$$\mathbb{P}\left\{\hat{\lambda}_k \geq \lambda_k + t\right\} \leq (p-k+1) \cdot \exp\left(\frac{-\mathrm{c}nt^2}{\lambda_k \sum_{i=k}^{p} \lambda_i}\right) \quad \textit{for } t \leq 4n\lambda_k,$$

and

$$\mathbb{P}\left\{\hat{\lambda}_k \leq \lambda_k - t\right\} \leq k \cdot \exp\left(\frac{-\mathrm{c}nt^2}{\lambda_1 \sum_{i=1}^{k} \lambda_i}\right) \quad \textit{for } t \leq 4n\lambda_1,$$

where the constant c *is at least* 1/32.

The following corollary provides an answer to our question about relative error estimates.

Corollary 7.2. *Let λ_k and $\hat{\lambda}_k$ be as in Theorem 7.1. Then*

$$\mathbb{P}\left\{\hat{\lambda}_k \geq (1+\varepsilon)\lambda_k\right\} \leq (p-k+1) \cdot \exp\left(\frac{-\mathrm{c}n\varepsilon^2}{\sum_{i=k}^{p} \frac{\lambda_i}{\lambda_k}}\right) \quad \textit{for } \varepsilon \leq 4n,$$

and

$$\mathbb{P}\left\{\hat{\lambda}_k \leq (1-\varepsilon)\lambda_k\right\} \leq k \cdot \exp\left(\frac{-\mathrm{c}n\varepsilon^2}{\frac{\lambda_1}{\lambda_k} \sum_{i=1}^{k} \frac{\lambda_i}{\lambda_k}}\right) \quad \textit{for } \varepsilon \in (0, 1],$$

where the constant c *is at least* 1/32.

The first bound in Corollary 7.2 tells us how many samples are needed to ensure that $\hat{\lambda}_k$ does not overestimate λ_k. Likewise, the second bound tells us how many samples ensure that $\hat{\lambda}_k$ does not underestimate λ_k.

Corollary 7.2 suggests that the relationship of $\hat{\lambda}_k$ to λ_k is determined by the spectrum of $\boldsymbol{C}$ in the following manner. When the eigenvalues below λ_k are small compared with λ_k, the quantity

$$\sum_{i=k}^{p} \lambda_i / \lambda_k$$

is small, and so $\hat{\lambda}_k$ is not likely to overestimate λ_k. Similarly, when the eigenvalues above λ_k are comparable with λ_k, the quantity

$$\frac{\lambda_1}{\lambda_k} \sum_{i=1}^{k} \lambda_i / \lambda_k$$

is small, and so $\hat{\lambda}_k$ is not likely to underestimate λ_k.

We now have everything needed to establish Theorem 1.1.

Proof of Theorem 1.1 from Corollary 7.2. From Corollary 7.2, we see that

$$\mathbb{P}\left\{\hat{\lambda}_k \leq (1-\varepsilon)\lambda_k\right\} \leq p^{-\beta} \quad \text{when} \quad n \geq 32\varepsilon^{-2}\left(\frac{\lambda_1}{\lambda_k} \sum\nolimits_{i \leq k} \frac{\lambda_i}{\lambda_k}\right)(\log k + \beta \log p).$$

Recall that $\kappa_k = \lambda_1(\boldsymbol{C})/\lambda_k(\boldsymbol{C})$. Clearly, taking $n = \Omega(\varepsilon^{-2}\kappa_\ell^2 \ell \log p)$ samples ensures that, with high probability, each of the top ℓ eigenvalues of the sample covariance matrix satisfies $\hat{\lambda}_k > (1-\varepsilon)\lambda_k$.

Likewise,

$$\mathbb{P}\left\{\hat{\lambda}_k \geq (1+\varepsilon)\lambda_k\right\} \leq p^{-\beta} \quad \text{when} \quad n \geq 32\varepsilon^{-2}\left(\sum\nolimits_{i \geq k} \frac{\lambda_i}{\lambda_k}\right)(\log(p-k+1) + \beta \log p).$$

Assuming the stated decay condition, that

$$\sum\nolimits_{i > \ell} \lambda_i = \mathrm{O}(\lambda_1),$$

we see that taking $n = \Omega(\varepsilon^{-2}(\ell + \kappa_\ell) \log p)$ samples ensures that, with high probability, each of the top ℓ eigenvalues of the sample covariance matrix satisfies $\hat{\lambda}_k < (1 + \varepsilon)\lambda_k$.

Combining these two results, we conclude that $n = \Omega(\varepsilon^{-2}\kappa_\ell^2 \ell \log p)$ ensures that the top ℓ eigenvalues of $\boldsymbol{C}$ are estimated to within relative precision $1 \pm \varepsilon$. □

Remark 7.1. The results in Theorem 7.1 and Corollary 7.2 also apply when $\boldsymbol{C}$ is rank-deficient: simply replace each occurence of the dimension p in the bounds with $\operatorname{rank}(\boldsymbol{C})$.

7.1. **Proof of Theorem 7.1.** We now prove Theorem 7.1. This result requires supporting lemmas; we defer their proofs until after a discussion of extensions to Theorem 7.1.

We study the error $|\lambda_k(\widehat{\boldsymbol{C}}_n) - \lambda_k(\boldsymbol{C})|$. To apply the methods developed in this paper, we pass to a question about the eigenvalues of a difference of two matrices. The first lemma accomplishes this goal by compressing both the population covariance matrix and the sample covariance matrix to a fixed invariant subspace of the population covariance matrix.

Lemma 7.3. *Let $\boldsymbol{X}$ be a random self-adjoint matrix with dimension p, and let $\boldsymbol{A}$ be a fixed self-adjoint matrix with dimension p. Choose $\boldsymbol{W}_+ \in \mathbb{V}_{p-k+1}^p$ and $\boldsymbol{W}_- \in \mathbb{V}_k^p$ for which*

$$\lambda_k(\boldsymbol{A}) = \lambda_{\max}\left(\boldsymbol{W}_+^* \boldsymbol{A} \boldsymbol{W}_+\right) = \lambda_{\min}\left(\boldsymbol{W}_-^* \boldsymbol{A} \boldsymbol{W}_-\right).$$

Then, for all $t > 0$,

$$\mathbb{P}\left\{\lambda_k(\boldsymbol{X}) \geq \lambda_k(\boldsymbol{A}) + t\right\} \leq \mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{W}_+^* \boldsymbol{X} \boldsymbol{W}_+\right) \geq \lambda_k(\boldsymbol{A}) + t\right\} \tag{7.2}$$

and

$$\mathbb{P}\left\{\lambda_k(\boldsymbol{X}) \leq \lambda_k(\boldsymbol{A}) - t\right\} \leq \mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{W}_-^* (-\boldsymbol{X}) \boldsymbol{W}_-\right) \geq -\lambda_k(\boldsymbol{A}) + t\right\}. \tag{7.3}$$

We apply this result with $\boldsymbol{A} = \boldsymbol{C}$ and $\boldsymbol{X} = \widehat{\boldsymbol{C}}_n$. Because $\widehat{\boldsymbol{C}}_n$ is unbounded, we apply Theorem 5.3 to handle the estimates in (7.2) and (7.3). To use this theorem, we need the following moment growth estimate for rank-one Wishart matrices.

Lemma 7.4. *Let $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{G})$. Then for any integer $m \geq 2$,*

$$\mathbb{E}\left(\boldsymbol{\xi}\boldsymbol{\xi}^*\right)^m \preceq 2^m m! (\operatorname{tr} \boldsymbol{G})^{m-1} \cdot \boldsymbol{G}.$$

With these preliminaries addressed, we prove Theorem 7.1.

Proof of upper estimate. First we consider the probability that $\hat{\lambda}_k$ overestimates λ_k. Let $\boldsymbol{W}_+ \in \mathbb{V}_{p-k+1}^p$ satisfy

$$\lambda_k(\boldsymbol{C}) = \lambda_{\max}\left(\boldsymbol{W}_+^* \boldsymbol{C} \boldsymbol{W}_+\right).$$

Then Lemma 7.3 implies

$$\begin{aligned}\mathbb{P}\left\{\lambda_k(\widehat{\boldsymbol{C}}_n) \geq \lambda_k(\boldsymbol{C}) + t\right\} &\leq \mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{W}_+^* \widehat{\boldsymbol{C}}_n \boldsymbol{W}_+\right) \geq \lambda_k(\boldsymbol{C}) + t\right\} \\ &= \mathbb{P}\left\{\lambda_{\max}\left(\textstyle\sum_j \boldsymbol{W}_+^* (\boldsymbol{\eta}_j \boldsymbol{\eta}_j^*) \boldsymbol{W}_+\right) \geq n\lambda_k(\boldsymbol{C}) + nt\right\}. \end{aligned} \tag{7.4}$$

The factor n comes from the normalization of the sample covariance matrix.

The covariance matrix of $\boldsymbol{\eta}_j$ is $\boldsymbol{C}$, so that of $\boldsymbol{W}_+^* \boldsymbol{\eta}_j$ is $\boldsymbol{W}_+^* \boldsymbol{C} \boldsymbol{W}_+$. Apply Lemma 7.4 to verify that $\boldsymbol{W}_+^* \boldsymbol{\eta}_j \boldsymbol{\eta}_j \boldsymbol{W}_+$ satisfies the subexponential moment growth bound required by Theorem 5.3 with

$$B = 2 \operatorname{tr}(\boldsymbol{W}_+^* \boldsymbol{C} \boldsymbol{W}_+) \quad \text{and} \quad \boldsymbol{\Sigma}_j^2 = 8 \operatorname{tr}(\boldsymbol{W}_+^* \boldsymbol{C} \boldsymbol{W}_+) \cdot \boldsymbol{W}_+^* \boldsymbol{C} \boldsymbol{W}_+.$$

In fact, $\boldsymbol{W}_+^* \boldsymbol{C} \boldsymbol{W}_+$ is the compression of $\boldsymbol{C}$ to the invariant subspace corresponding with its bottom $p - k + 1$ eigenvalues, so

$$B = 2 \textstyle\sum_{i=k}^p \lambda_i(\boldsymbol{C}) \quad \text{and} \quad \lambda_{\max}\left(\boldsymbol{\Sigma}_j^2\right) = 8\lambda_k(\boldsymbol{C}) \textstyle\sum_{i=k}^p \lambda_i(\boldsymbol{C}).$$

We are concerned with the maximum eigenvalue of the sum in (7.4), so we take $\boldsymbol{V}_+ = \mathbf{I}$ in the statement of Theorem 5.3 to find that

$$\sigma_1^2 = \lambda_{\max}\left(\sum\nolimits_j \boldsymbol{\Sigma}_j^2\right) = n\lambda_{\max}\left(\boldsymbol{\Sigma}_1^2\right) = 8n\lambda_k(\boldsymbol{C})\sum\nolimits_{i=k}^{p}\lambda_i(\boldsymbol{C}) \quad \text{and}$$
$$\mu_1 = \lambda_{\max}\left(\sum\nolimits_j \boldsymbol{W}_+^*\mathbb{E}(\boldsymbol{\eta}_j\boldsymbol{\eta}_j^*)\boldsymbol{W}_+\right) = n\lambda_{\max}\left(\boldsymbol{W}_+^*\boldsymbol{C}\boldsymbol{W}_+\right) = n\lambda_k(\boldsymbol{C}).$$

It follows from the subgaussian branch of the split Bernstein inequality of Theorem 5.3 that

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_j \boldsymbol{W}_+^*(\boldsymbol{\eta}_j\boldsymbol{\eta}_j^*)\boldsymbol{W}_+\right) \geq n\lambda_k(\boldsymbol{C}) + nt\right\} \leq (p-k+1)\cdot\exp\left(\frac{-nt^2}{32\lambda_k(\boldsymbol{C})\sum_{i=k}^{p}\lambda_i(\boldsymbol{C})}\right)$$

when $t \leq 4n\lambda_k(\boldsymbol{C})$. This provides the desired bound on the probability that $\lambda_k(\widehat{\boldsymbol{C}}_n)$ overestimates $\lambda_k(\boldsymbol{C})$. □

Proof of lower estimate. Now we consider the probability that $\hat{\lambda}_k$ underestimates λ_k. The proof proceeds similarly to the proof of the upper estimate. Let $\boldsymbol{W}_- \in \mathbb{V}_k^p$ satisfy

$$\lambda_k(\boldsymbol{C}) = \lambda_{\min}\left(\boldsymbol{W}_-^*\boldsymbol{C}\boldsymbol{W}_-\right).$$

Then Lemma 7.3 implies

$$\begin{aligned}\mathbb{P}\left\{\lambda_k(\widehat{\boldsymbol{C}}_n) \leq \lambda_k(\boldsymbol{C}) - t\right\} &\leq \mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{W}_-^*(-\widehat{\boldsymbol{C}}_n)\boldsymbol{W}_-\right) \geq -n\lambda_k(\boldsymbol{C}) + nt\right\} \\ &= \mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_j \boldsymbol{W}_-^*(-\boldsymbol{\eta}_j\boldsymbol{\eta}_j^*)\boldsymbol{W}_-\right) \geq -n\lambda_k(\boldsymbol{C}) + nt\right\} \qquad (7.5)\end{aligned}$$

The factor n comes from the normalization of the sample covariance matrix.

The covariance matrix of $\boldsymbol{\eta}_j$ is $\boldsymbol{C}$, so that of $\boldsymbol{W}_-^*\boldsymbol{\eta}_j$ is $\boldsymbol{W}_-^*\boldsymbol{C}\boldsymbol{W}_-$. Apply Lemma 7.4 to verify that for any integer $m \geq 2$,

$$\mathbb{E}(\boldsymbol{W}_-^*(-\boldsymbol{\eta}_j\boldsymbol{\eta}_j^*)\boldsymbol{W}_-)^m \preceq \mathbb{E}(\boldsymbol{W}_-^*\boldsymbol{\eta}_j\boldsymbol{\eta}_j^*\boldsymbol{W}_-)^m \preceq 2^m m!\,\mathrm{tr}(\boldsymbol{W}_-^*\boldsymbol{C}\boldsymbol{W}_-)^{m-1}\cdot\boldsymbol{W}_-^*\boldsymbol{C}\boldsymbol{W}_-.$$

Thus, $\boldsymbol{W}_-^*(-\boldsymbol{\eta}_j\boldsymbol{\eta}_j^*)\boldsymbol{W}_-$ satisfies the subexponential moment growth bound required by Theorem 5.3 with

$$B = 2\,\mathrm{tr}(\boldsymbol{W}_-^*\boldsymbol{C}\boldsymbol{W}_-) \quad \text{and} \quad \boldsymbol{\Sigma}_j^2 = 8\,\mathrm{tr}(\boldsymbol{W}_-^*\boldsymbol{C}\boldsymbol{W}_-)\cdot\boldsymbol{W}_-^*\boldsymbol{C}\boldsymbol{W}_-.$$

In fact, $\boldsymbol{W}_-^*\boldsymbol{C}\boldsymbol{W}_-$ is the compression of $\boldsymbol{C}$ to the invariant subspace corresponding with its top k eigenvalues, so

$$B = 2\sum\nolimits_{i=1}^{k}\lambda_i(\boldsymbol{C}) \quad \text{and} \quad \lambda_{\max}\left(\boldsymbol{\Sigma}_j^2\right) = 8\lambda_1(\boldsymbol{C})\sum\nolimits_{i=1}^{k}\lambda_i(\boldsymbol{C}).$$

We are concerned with the maximum eigenvalue of the sum in (7.5), so we take $\boldsymbol{V}_+ = \mathbf{I}$ in the statement of Theorem 5.3 to find that

$$\sigma_1^2 = \lambda_{\max}\left(\sum\nolimits_j \boldsymbol{\Sigma}_j^2\right) = n\lambda_{\max}\left(\boldsymbol{\Sigma}_1^2\right) = 8n\lambda_1(\boldsymbol{C})\sum\nolimits_{i=1}^{k}\lambda_i(\boldsymbol{C}) \quad \text{and}$$
$$\mu_1 = \lambda_{\max}\left(\sum\nolimits_j \boldsymbol{W}_-^*\mathbb{E}(-\boldsymbol{\eta}_j\boldsymbol{\eta}_j^*)\boldsymbol{W}_-\right) = n\lambda_{\max}\left(\boldsymbol{W}_-^*(-\boldsymbol{C})\boldsymbol{W}_-\right) = -n\lambda_k(\boldsymbol{C}).$$

It follows from the subgaussian branch of the split Bernstein inequality of Theorem 5.3 that

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_j \boldsymbol{W}_-^*(-\boldsymbol{\eta}_j\boldsymbol{\eta}_j^*)\boldsymbol{W}_-\right) \geq -n\lambda_k(\boldsymbol{C}) + nt\right\} \leq k\cdot\exp\left(\frac{-nt^2}{32\lambda_1(\boldsymbol{C})\sum_{i=1}^{k}\lambda_i(\boldsymbol{C})}\right)$$

when $t \leq 4n\lambda_1(\boldsymbol{C})$. This provides the desired bound on the probability that $\lambda_k(\widehat{\boldsymbol{C}}_n)$ underestimates $\lambda_k(\boldsymbol{C})$. □

7.2. **Extensions of Theorem 7.1.** Results analogous to Theorem 7.1 can be established for other distributions. If the distribution is bounded, the possibility that $\hat{\lambda}_k$ deviates above or below λ_k can be controlled using the Bernstein inequality of Theorem 5.1. If the distribution is unbounded but has matrix moments that satisfy a sufficiently nice growth condition, the probability that $\hat{\lambda}_k$ deviates below λ_k as well as the probability that it deviates above λ_k can be bounded using a Bernstein inequality analogous to that in Theorem 5.3.

Theorem 7.1 controls the error in the kth sample eigenvalue in terms of all the eigenvalues of the covariance matrix, so it is most useful when the eigenvalues of the covariance matrix satisfy decay conditions such as those given in the statement of Theorem 1.1. If such conditions are not satisfied, the results of [ALPTJ11] on the convergence of empirical covariance matrices of isotropic log-concave random vectors lead to tighter bounds on the probabilities that $\hat{\lambda}_k$ overestimates or underestimates λ_k.

To see the relevance of the results in [ALPTJ11], first observe the following consequence of the subadditivity of the maximum eigenvalue mapping:

$$\begin{aligned}\lambda_{\max}\left(\boldsymbol{W}_+^*(\boldsymbol{X}-\boldsymbol{A})\boldsymbol{W}_+\right) &\geq \lambda_{\max}\left(\boldsymbol{W}_+^*\boldsymbol{X}\boldsymbol{W}_+\right) - \lambda_{\max}\left(\boldsymbol{W}_+^*\boldsymbol{A}\boldsymbol{W}_+\right)\\ &= \lambda_{\max}\left(\boldsymbol{W}_+^*\boldsymbol{X}\boldsymbol{W}_+\right) - \lambda_k(\boldsymbol{A}).\end{aligned}$$

In conjunction with (7.2), this gives us the following control on the probability that $\lambda_k(\boldsymbol{X})$ overestimates $\lambda_k(\boldsymbol{A})$:

$$\mathbb{P}\left\{\lambda_k(\boldsymbol{X}) \geq \lambda_k(\boldsymbol{A}) + t\right\} \leq \mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{W}_+^*(\boldsymbol{X}-\boldsymbol{A})\boldsymbol{W}_+\right) \geq t\right\}.$$

In our application, $\boldsymbol{X}$ is the empirical covariance matrix and $\boldsymbol{A}$ is the actual covariance matrix. The spectral norm dominates the maximum eigenvalue, so

$$\begin{aligned}\mathbb{P}\left\{\lambda_k(\widehat{\boldsymbol{C}}_n) \geq \lambda_k(\boldsymbol{C}) + t\right\} &\leq \mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{W}_+^*(\widehat{\boldsymbol{C}}_n - \boldsymbol{C})\boldsymbol{W}_+\right) \geq t\right\}\\ &\leq \mathbb{P}\left\{\|\boldsymbol{W}_+^*(\widehat{\boldsymbol{C}}_n - \boldsymbol{C})\boldsymbol{W}_+\| \geq t\right\} = \mathbb{P}\left\{\|\boldsymbol{W}_+^*\widehat{\boldsymbol{C}}_n\boldsymbol{W}_+ - \boldsymbol{S}^2\| \geq t\right\},\end{aligned}$$

where $\boldsymbol{S}$ is the square root of $\boldsymbol{W}_+^*\boldsymbol{C}\boldsymbol{W}_+$. Now factor out $\boldsymbol{S}^2$ and identify $\lambda_k(\boldsymbol{C}) = \|\boldsymbol{S}^2\|$ to obtain

$$\begin{aligned}\mathbb{P}\left\{\lambda_k(\widehat{\boldsymbol{C}}) \geq \lambda_k(\boldsymbol{C}) + t\right\} &\leq \mathbb{P}\left\{\|\boldsymbol{S}^{-1}\boldsymbol{W}_+^*\widehat{\boldsymbol{C}}_n\boldsymbol{W}_+\boldsymbol{S}^{-1} - \mathbf{I}\|\|\boldsymbol{S}^2\| \geq t\right\}\\ &= \mathbb{P}\left\{\|\boldsymbol{S}^{-1}\boldsymbol{W}_+^*\widehat{\boldsymbol{C}}_n\boldsymbol{W}_+\boldsymbol{S}^{-1} - \mathbf{I}\| \geq t/\lambda_k(\boldsymbol{C})\right\}.\end{aligned}$$

Note that if $\boldsymbol{\eta}$ is drawn from a $\mathcal{N}(\mathbf{0}, \boldsymbol{C})$ distribution, then the covariance matrix of the transformed sample $\boldsymbol{S}^{-1}\boldsymbol{W}_+^*\boldsymbol{\eta}$ is the identity:

$$\mathbb{E}\left(\boldsymbol{S}^{-1}\boldsymbol{W}_+^*\boldsymbol{\eta}\boldsymbol{\eta}^*\boldsymbol{W}_+\boldsymbol{S}^{-1}\right) = \boldsymbol{S}^{-1}\boldsymbol{W}_+^*\boldsymbol{C}\boldsymbol{W}_+\boldsymbol{S}^{-1} = \mathbf{I}.$$

Thus $\boldsymbol{S}^{-1}\boldsymbol{W}_+^*\widehat{\boldsymbol{C}}_n\boldsymbol{W}_+\boldsymbol{S}^{-1}$ is the empirical covariance matrix of a standard Gaussian vector in $\mathbb{R}^{p-k+1}$. By Theorem 1 of [ALPTJ11], it follows that $\hat{\lambda}_k$ is unlikely to overestimate λ_k in relative error when the number n of samples is $\Omega(p-k+1)$. A similar argument shows that $\hat{\lambda}_k$ is unlikely to underestimate λ_k in relative error when $n = \Omega(\kappa_p^2 k)$.

Similarly, for more general distributions, the bounds on the probability of $\hat{\lambda}_k$ overestimating or underestimating λ_k can be tightened beyond those suggested in Theorem 7.1 by using the results in [ALPTJ11] or [Ver11]. Note, however, that one cannot use knowledge of spectral decay to sharpen the results obtained from [ALPTJ11] and [Ver11] into estimates like those given in Theorem 1.1.

Finally, we note that the techniques developed in the proof of Theorem 7.1 can be used to investigate the spectrum of the error matrices $\widehat{\boldsymbol{C}}_n - \boldsymbol{C}$.

7.3. **Proofs of the supporting lemmas.** We now establish the lemmas used in the proof of Theorem 7.1.

Proof of Lemma 7.3. The probability that $\lambda_k(\boldsymbol{X})$ overestimates $\lambda_k(\boldsymbol{A})$ is controlled with the sequence of inequalities

$$\begin{aligned}\mathbb{P}\left\{\lambda_k(\boldsymbol{X}) \geq \lambda_k(\boldsymbol{A}) + t\right\} &= \mathbb{P}\left\{\inf_{\boldsymbol{W}\in\mathbb{V}_{p-k+1}^p} \lambda_{\max}\left(\boldsymbol{W}^*\boldsymbol{X}\boldsymbol{W}\right) \geq \lambda_k(\boldsymbol{A}) + t\right\} \\ &\leq \mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{W}_+^*\boldsymbol{X}\boldsymbol{W}_+\right) \geq \lambda_k(\boldsymbol{A}) + t\right\}.\end{aligned}$$

We use a related approach to study the probability that $\lambda_k(\boldsymbol{X})$ underestimates $\lambda_k(\boldsymbol{A})$. Our choice of $\boldsymbol{W}_-$ implies that

$$\begin{aligned}\mathbb{P}\left\{\lambda_k(\boldsymbol{X}) \leq \lambda_k(\boldsymbol{A}) - t\right\} &= \mathbb{P}\left\{\max_{\boldsymbol{W}\in\mathbb{V}_k^p} \lambda_{\min}\left(\boldsymbol{W}^*\boldsymbol{X}\boldsymbol{W}\right) \leq \lambda_k(\boldsymbol{A}) - t\right\} \\ &\leq \mathbb{P}\left\{\lambda_{\min}\left(\boldsymbol{W}_-^*\boldsymbol{X}\boldsymbol{W}_-\right) \leq \lambda_k(\boldsymbol{A}) - t\right\} \\ &= \mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{W}_-^*(-\boldsymbol{X})\boldsymbol{W}_-\right) \geq -\lambda_k(\boldsymbol{A}) + t\right\}.\end{aligned}$$

This establishes the bounds on the probabilities of $\lambda_k(\boldsymbol{X})$ deviating above or below $\lambda_k(\boldsymbol{A})$. □

Proof of Lemma 7.4. Factor the covariance matrix of $\boldsymbol{\xi}$ as $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^*$ where $\boldsymbol{U}$ is orthogonal and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ is the matrix of eigenvalues of $\boldsymbol{G}$. Let $\boldsymbol{\gamma}$ be a $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ random variable. Then $\boldsymbol{\xi}$ and $\boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\gamma}$ are identically distributed, so

$$\begin{aligned}\mathbb{E}(\boldsymbol{\xi}\boldsymbol{\xi}^*)^m &= \mathbb{E}\left[(\boldsymbol{\xi}^*\boldsymbol{\xi})^{m-1}\boldsymbol{\xi}\boldsymbol{\xi}^*\right] = \mathbb{E}\left[(\boldsymbol{\gamma}^*\boldsymbol{\Lambda}\boldsymbol{\gamma})^{m-1}\boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\gamma}\boldsymbol{\gamma}^*\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^*\right] \\ &= \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\mathbb{E}\left[(\boldsymbol{\gamma}^*\boldsymbol{\Lambda}\boldsymbol{\gamma})^{m-1}\boldsymbol{\gamma}\boldsymbol{\gamma}^*\right]\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^*. \qquad (7.6)\end{aligned}$$

Consider the (i, j) entry of the bracketed matrix in (7.6):

$$\mathbb{E}\left[(\boldsymbol{\gamma}^*\boldsymbol{\Lambda}\boldsymbol{\gamma})^{m-1}\gamma_i\gamma_j\right] = \mathbb{E}\left[\left(\sum\nolimits_{\ell=1}^p \lambda_\ell\gamma_\ell^2\right)^{m-1}\gamma_i\gamma_j\right]. \qquad (7.7)$$

From this expression, and the independence of the Gaussian variables $\{\gamma_i\}$, we see that this matrix is diagonal.

To bound the diagonal entries, use a multinomial expansion to further develop the sum in (7.7) for the (i, i) entry:

$$\mathbb{E}\left[(\boldsymbol{\gamma}^*\boldsymbol{\Lambda}\boldsymbol{\gamma})^{m-1}\gamma_i^2\right] = \sum\nolimits_{\ell_1+\cdots+\ell_p=m-1}\binom{m-1}{\ell_1, \ldots, \ell_p}\lambda_1^{\ell_1}\cdots\lambda_p^{\ell_p}\mathbb{E}\left[\gamma_1^{2\ell_1}\cdots\gamma_p^{2\ell_p}\gamma_i^2\right].$$

Denote the L_r norm of a random variable X by

$$\|X\|_r = \left(\mathbb{E}|X|^r\right)^{1/r}.$$

Since $\ell_1, \ldots, \ell_p$ are nonnegative integers summing to $m-1$, the generalized AM-GM inequality justifies the first of the following inequalities:

$$\begin{aligned}\mathbb{E}\gamma_1^{2\ell_1}\cdots\gamma_p^{2\ell_p}\gamma_i^2 &\leq \mathbb{E}\left(\frac{\ell_1|\gamma_1| + \cdots + \ell_p|\gamma_p| + |\gamma_i|}{m}\right)^{2m} = \left\|\frac{1}{m}\left(|\gamma_i| + \sum\nolimits_{j=1}^p \ell_j|\gamma_j|\right)\right\|_{2m}^{2m} \\ &\leq \left(\frac{1}{m}\left(\|\gamma_i\|_{2m} + \sum\nolimits_{j=1}^p \ell_j\|\gamma_j\|_{2m}\right)\right)^{2m} \\ &= \left(\frac{1+\ell_1+\ldots+\ell_p}{m}\right)^{2m}\|g\|_{2m}^{2m} = \mathbb{E}(g^{2m}).\end{aligned}$$

The second inequality is the triangle inequality for L_r norms. Now we reverse the multinomial expansion to see that the diagonal terms satisfy the inequality

$$\begin{aligned}\mathbb{E}\left[(\boldsymbol{\gamma}^*\boldsymbol{\Lambda}\boldsymbol{\gamma})^{m-1}\gamma_i^2\right] &\le \sum\nolimits_{\ell_1+\cdots+\ell_p=m-1}\binom{m-1}{\ell_1,\ldots,\ell_p}\lambda_1^{\ell_1}\cdots\lambda_p^{\ell_p}\mathbb{E}(g^{2m}) \\ &= (\lambda_1+\ldots+\lambda_p)^{m-1}\mathbb{E}(g^{2m}) = \operatorname{tr}(\boldsymbol{G})^{m-1}\mathbb{E}(g^{2m}). \qquad (7.8)\end{aligned}$$

Estimate $\mathbb{E}(g^{2m})$ using the fact that $\Gamma(x)$ is increasing for $x \ge 1$:

$$\mathbb{E}\left(g^{2m}\right) = \frac{2^m}{\sqrt{\pi}}\Gamma(m+1/2) < \frac{2^m}{\sqrt{\pi}}\Gamma(m+1) = \frac{2^m}{\sqrt{\pi}}m! \quad \text{for } m \ge 1.$$

Combine this result with (7.8) to see that

$$\mathbb{E}\left[(\boldsymbol{\gamma}^*\boldsymbol{\Lambda}\boldsymbol{\gamma})^{m-1}\boldsymbol{\gamma}\boldsymbol{\gamma}^*\right] \preceq \frac{2^m}{\sqrt{\pi}}m!\operatorname{tr}(\boldsymbol{G})^{m-1}\cdot\mathbf{I}.$$

Complete the proof by using this estimate in (7.6).

□

References

[Ach03] D. Achlioptas, *Database-friendly random projections: Johnson-Lindenstrauss with binary coins*, J. Comput. System Sci. **66** (2003), 671–687.

[AKV02] N. Alon, M. Krivelevich, and V. H. Vu, *On the concentration of eigenvalues of random symmetric matrices*, Israel J. Math. **131** (2002), 259–267.

[ALPTJ11] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann, *Sharp bounds on the rate of convergence of the empirical covariance matrix*, C. R. Math. Acad. Sci. Paris **349** (2011), 195–200.

[AM01] D. Achlioptas and F. McSherry, *Fast Computation of Low Rank Matrix Approximations*, Proc. 33rd ACM Symposium on Theory of Computing (STOC 2001), 2001, pp. 611–618.

[AM07] ———, *Fast Computation of Low Rank Matrix Approximations*, J. ACM **54** (2007), no. 2.

[Arn71] L. Arnold, *On Wigner's semicircle law for the eigenvalues of random matrices*, Z. Wahrsch. Verw. Gebiete **19** (1971), 191–198.

[AW02] R. Ahlswede and A. Winter, *Strong converse for identification via quantum channels*, IEEE Trans. Inform. Theory **48** (2002), no. 3, 569–579.

[Bai93a] Z. D. Bai, *Convergence rate of expected spectral distributions of large random matrices. Part I. Wigner matrices*, Ann. Probab. **21** (1993), 625–648.

[Bai93b] ———, *Convergence rate of expected spectral distributions of large random matrices. Part II. Sample covariance matrices*, Ann. Probab. **21** (1993), 649–672.

[BHPZ11] Z. D. Bai, J. Hu, G. Pan, and W. Zhou, *A note on rate of convergence in probability to semicircular law*, Preprint, arXiv:1105.3056, 2011.

[BK99] M. V. Berry and J. P. Keating, *The Riemann Zeros and Eigenvalue Asymptotics*, SIAM Rev. **41** (1999), 236–266.

[BMT97] Z. D. Bai, B. Q. Miao, and J. Tsay, *A note on the convergence rate of the spectral distributions of large dimensional random matrices*, Statist. Probab. Lett. **34** (1997), 95–102.

[BMT99] ———, *Remarks on the convergence rate of the spectral distributions of Wigner matrices.*, J. Theoret. Probab. **12** (1999), 301–311.

[BMT02] ———, *Convergence rates of the spectral distributions of large Wigner matrices*, Int. Math. J. **1** (2002), 65–90.

[BMY03] Z. D. Bai, B. Q. Miao, and J. F. Yao, *Convergence rates of spectral distributions of large sample covariance matrices.*, SIAM J. Matrix Anal. Appl. **25** (2003), 105–127.

[BSY88] Z. D. Bai, J. W. Silverstein, and Y. Q. Yin, *A note on the largest eigenvalue of a large dimensional sample covariance matrix*, J. Multivariate Anal. **26** (1988), 166–168.

[BY88a] Z. D. Bai and Y. Q. Yin, *A convergence to the semicircle law*, Ann. Probab. **16** (1988), 863–875.

[BY88b] ———, *Necessary and sufficient conditions for the almost sure convergence of the largest eigenvalue of Wigner matrices*, Ann. Probab. **16** (1988), 1729–1741.

[CD05] Z. Chen and J. J. Dongarra, *Condition numbers of Gaussian random matrices*, SIAM J. Matrix Anal. Appl. **27** (2005), 603–620.

[CM08] D. Christofides and K. Markström, *Expansion properties of random Cayley graphs and vertex transitive graphs via matrix martingales*, Random Structures Algorithms **32** (2008), 88–100.

[Dei07] P. Deift, *Universality for mathematical and physical systems*, International Congress of Mathematicians. Vol. I, Eur. Math. Soc., Zürich, 2007.

[DG98] V. De la Peña and E. Giné, *Decoupling: From Dependence to Independence*, Probability and its Applications, Springer, 1998.

[DM05] P. Drineas and M. W. Mahoney, *On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning*, J. Mach. Learn. Res. **6** (2005), 2153–2175.

[DM10] ———, *Effective Resistances, Statistical Leverage, and Applications to Linear Equation Solving*, Preprint, arXiv:1005.3097, 2010.

[DMMS11] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, *Faster least squares approximation*, Numer. Math. **117** (2011), 219–249.

[El 07] N. El Karoui, *Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices*, Ann. Probab. **35** (2007), 663–714.

[El 08] N. El Karoui, *Spectrum estimation for large dimensional covariance matrices using random matrix theory*, Ann. Statist. **36** (2008), 2757–2790.

[EYY10] L. Erdös, H. Yau, and J. Yin, *Rigidity of Eigenvalues of Generalized Wigner Matrices*, Preprint, arXiv:1007.4652, 2010.

[Far10] B. Farrell, *Limiting Empirical Singular Value Distribution of Restrictions of Discrete Fourier Transform Matrices*, J. Fourier Anal. Appl. (2010), 1–21.

[Gem80] S. Gemon, *A limit theorem for the norm of random matrices*, Ann. Probab. **8** (1980), 252–261.

[GH10] S. Gurevich and R. Hadani, *The statistical restricted isometry property and the Wigner semicircle distribution of incoherent dictionaries*, Submitted to Appl. Comput. Harmon. Anal., 2010.

[GMGW98] T. Guhr, A. Müller-Groeling, and H. A. Weidenmüller, *Random matrix theories in quantum physics: common concepts*, Phys. Rep. **299** (1998), 189–425.

[Gre63] U. Grenander, *Probabilities on Algebraic Structures*, John Wiley & Sons, Inc., 1963.

[Gro11] D. Gross, *Recovering low-rank matrices from few coefficients in any basis*, IEEE Trans. Inform. Theory **57** (2011), 1548–1566.

[Gus05] J. Gustavsson, *Gaussian fluctuations of eigenvalues in the GUE*, Ann. Inst. Henri Poincaré Probab. Stat. **41** (2005), 151–178.

[HKZ11] D. Hsu, S. M. Kakade, and T. Zhang, *Dimension-free tail inequalities for sums of random matrices*, Preprint, arXiv:1104.1672, 2011.

[HMT11] N. Halko, P. G. Martinsson, and J. A. Tropp, *Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions*, SIAM Rev. **53** (2011), 217–288.

[Joh01] I. M. Johnstone, *On the distribution of the largest eigenvalue in principal components analysis*, Ann. Statist. **29** (2001), 295–327.

[Joh07] ———, *High-dimensional statistical inference and random matrices*, 2007 Proc. of the International Congress of Mathematicians (ICM 2007), 2007, pp. 307–333.

[Lat05] R. Latała, *Some estimates of norms of random matrices*, Proc. Amer. Math. Soc. **133** (2005), 1273–1282.

[Lie73] E. H. Lieb, *Convex trace functions and the Wigner–Yanase–Dyson conjecture*, Adv. Math. **11** (1973), no. 3, 267–288.

[LM93] J. Lindenstrauss and V. D. Milman, *Handbook of convex geometry*, vol. B, ch. The Local Theory of Normed Spaces and Its Applications to Convexity, pp. 1149–1220, Elsevier Science Publishers B.V., 1993.

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

[LPP91] F. Lust-Piquard and G. Pisier, *Noncommutative Khintchine and Paley Inequalities*, Ark. Mat. **29** (1991), 241–260.

[LS05] E. H. Lieb and R. Seiringer, *Stronger subadditivity of entropy*, Phys. Rev. A **71** (2005), no. 6.

[Mah11] M. W. Mahoney, *Randomized algorithms for matrices and data*, Preprint, arXiv:1104.5557, 2011.

[Mec04] M. W. Meckes, *Concentration of norms and eigenvalues of random matrices*, J. Funct. Anal. **211** (2004), 508–524.

[Meh04] M. L. Mehta, *Random Matrices*, Third ed., Academic Press, 2004.

[MP67] V. A. Marčenko and L. A. Pastur, *Distributions of eigenvalues for some sets of random matrices*, Math. USSR Sb. **1** (1967), 457–483.

[MP06] S. Mendelson and A. Pajor, *On singular values of matrices with independent rows*, Bernoulli **12** (2006), 761–773.

[Nem07] A. Nemirovski, *Sums of random symmetric matrices and quadratic optimization under orthogonality constraints*, Mathem. Program. **109** (2007), 283–317.

[Oli09] R. I. Oliveira, *Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges*, Preprint, arXiv:0911.0600, 2009.

[Oli10] ———, *Sums of random Hermitian matrices and an inequality due to Rudelson*, Elect. Comm. in Probab. **15** (2010), 203–212.

[Rec09] B. Recht, *A simpler approach to matrix completion*, To appear in J. Mach. Learn. Res., 2009.

[RS96] Z. Rudnick and P. Sarnack, *Zeros of principal L-functions and random matrix theory*, Duke Math. J. **81** (1996), 269–322.

[Rud99] M. Rudelson, *Random Vectors in the Isotropic Position*, J. Funct. Anal. **164** (1999), no. 1, 60–72.

[RV07] M. Rudelson and R. Vershynin, *Sampling from large matrices: An approach through geometric functional analysis*, J. Assoc. Comput. Mach. **54** (2007), no. 4.

[RV08] ———, *The least singular value of a random square matrix is $O(n^{-1/2})$*, C. R. Math. Acad. Sci. Paris **346** (2008), 893–896.

[SD01] S. J. Szarek and K. R. Davidson, *Handbook on the Geometry of Banach spaces*, ch. Local operator theory, random matrices, and Banach spaces, pp. 317–366, North Holland, 2001.

[So09] A. M. So, *Moment inequalities for sums of random matrices and their applications in optimization*, Math. Program. (2009), 1–27.

[SS08] D. A. Spielman and N. Srivastava, *Graph sparsification by effective resistances*, Proc. 40th ACM Symposium on Theory of Computing (STOC 2008), 2008.

[ST06] J. W. Silverstein and A. M. Tulino, *Theory of Large Dimensional Random Matrices for Engineers*, 2006 IEEE Ninth International Symposium on Spread Spectrum Techniques and Applications, 2006, pp. 458–464.

[SV11] N. Srivastava and R. Vershynin, *Covariance Estimation for Distributions with $2 + \varepsilon$ Moments*, Preprint, arXiv:1106.2775, 2011.

[Sza90] S. J. Szarek, *Spaces with large distance to ℓ_∞^n and random matrices*, Amer. J. Math. **112** (1990), 899–942.

[Tal95] M. Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Inst. Hautes Études Sci. Publ. Math. **81** (1995), 73–205.

[Tal05] ———, *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*, Springer, 2005.

[Tro08] J. A. Tropp, *On the conditioning of random subdictionaries*, Appl. Comput. Harmon. Anal. **25** (2008), no. 1, 1–24.

[Tro11a] ———, *Freedman's inequality for matrix martingales*, Electron. Comm. Probab. **16** (2011), 262–270.

[Tro11b] ———, *User-Friendly Tail Bounds for Matrix Martingales*, Tech. report, California Institute of Technology, 2011, `http://www.acm.caltech.edu/~jtropp/reports/Tro10-User-Friendly-Martingale-TR.pdf`, retrieved June 13, 2011.

[Tro11c] ———, *User-Friendly Tail Bounds for Sums of Random Matrices*, Accepted to Found. Comput. Math., June 2011. Preprint, arXiv:1004.4389, 2011.

[TV04] A. M. Tulino and S. Verdu, *Random Matrix Theory and Wireless Communication*, Now Publishers Inc., 2004.

[TV10a] T. Tao and V. Vu, *Random Matrices: Localization of the eigenvalues and the necessity of four moments*, Preprint, arXiv:1005.2901, 2010.

[TV10b] ———, *Random Matrices: Universality of ESDs and the circular law*, Ann. Probab. **38** (2010), 2023–2065, With an appendix by M. Krishnapur.

[TV11] ———, *Random Matrices: Universality of local eigenvalue statistics*, Acta Math. **206** (2011), 127–204.

[Vem04] S. S. Vempala, *The Random Projection Method*, AMS, 2004.

[Ver10] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, `http://www-personal.umich.edu/~romanv/papers/non-asymptotic-rmt-plain.pdf`, 2010, retrieved June 13, 2011.

[Ver11] ———, *How close is the sample covariance matrix to the actual covariance matrix?*, J. Theoret. Probab. (2011), 1–32.

[Wig55] E. P. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, Annals Math. **62** (1955), 548–564.

[YBK88] Y. Q. Yin, Z. D. Bai, and P. R. Krishnaiah, *On the limit of the largest eigenvalue of the large dimensional sample covariance matrix*, Probab. Theory Relat. Fields **78** (1988), 509–521.