Universality for the Conjugate Gradient and MINRES Algorithms on Sample Covariance Matrices

ELLIOT PAQUETTE
McGill University

THOMAS TROGDON
University of Washington

Abstract

We present a probabilistic analysis of two Krylov subspace methods for solving linear systems. We prove a central limit theorem for norms of the residual vectors that are produced by the conjugate gradient and MINRES algorithms when applied to a wide class of sample covariance matrices satisfying some standard moment conditions. The proof involves establishing a four-moment theorem for the so-called spectral measure, implying, in particular, universality for the matrix produced by the Lanczos iteration. The central limit theorem then implies an almost-deterministic iteration count for the iterative methods in question.

Contents

1 Introduction

Sample covariance matrices constitutes one of the oldest classes of random matrices. One can trace their theory at least back to the seminal work of Wishart [44].

1991 *Mathematics Subject Classification.* 65F10, 60B20.

Key words and phrases. Sample covariance matrices, conjugate gradient, MINRES, Wishart distribution.

Communications on Pure and Applied Mathematics, Vol. LXXVI, 1085–1136 (2023)

Specifically, Wishart considered matrices of the form

$$W = \frac{1}{M} XX^T \tag{1.1}$$

where X is an $N \times M$ matrix whose entries are independent and identically distributed (iid) standard normal random variables. Such matrices provide an estimator for the covariance matrix of the columns of X, and the Wishart distribution can play the role of the null distribution in covariance estimation. Wishart matrices arise in other settings too, and particularly relevant to this paper, they appear in the seminal work of Goldstine and von Neumman [23] on the numerical inversion of matrices.

Recently, there has been increasing interest in understanding how algorithms from numerical linear algebra and beyond act on random matrices. Specifically, this allows one to give a precise average-case analysis of the algorithms, replacing the standard worst-case estimates/bounds. For noniterative methods such as Gaussian elimination, one looks for average-case bounds on rounding errors (see [36], for example). For iterative methods, more questions can be asked, the most basic of which is the question, "In exact arithmetic, how many iterations are required, on average, to solve a problem?" The simplex method from linear programming was addressed in this context by many authors [5, 39, 40]. In these works, the notion of average case is typically restricted to one ensemble, or distribution. Indeed, the natural criticism of a simple average-case analysis is that the outcome could be ensemble-dependent, and thus it only has predictive power for a small subset of real-world phenomena.

So, in the context of average-case analysis, it becomes important to show that any arbitrary modeling choices made in defining the ensemble have a limited effect. In the probability literature, this concept is called *universality*, and it has been studied extensively for many years. The most famous example of universality is the central limit theorem which states that for sufficiently large M, the sums

$$S_M = \frac{1}{M} \sum_{j=1}^{M} X_j$$

for iid $(X_j)_{j \geq 1}$ concentrate on the mean of X_1 (and hence X_j for every j) and have small fluctuations of size $M^{-1/2}$ about this mean that are asymptotically normally distributed. This is true, as soon as the random variables have a finite second moment, and more to the point, it does not depend on any further information about the distribution beyond its first two moments. It can be argued that this particular universality explains the peculiar prevalence and usefulness of the normal distribution in statistics and nature.

Universality has been featured as a particularly important central feature of random matrix theory, especially in the last 20 years. Many quantities, such as the largest eigenvalue of W, are universal—they have fluctuations that are independent of the distribution on entries of W, with some mild moment conditions. The

specific statement for the largest eigenvalue $\lambda_1(W)$ of W is

$$\lim_{M\to\infty} \mathbb{P}\big(c_d N^{2/3}\big(\lambda_1(W) - (1+\sqrt{\mathfrak{d}})^2\big) \le t\big) = F_1(t), \quad \mathfrak{d} = \frac{N}{M}, \tag{1.2}$$

where $F_1(t)$ is the cumulative distribution function for the Tracy–Widom ($\beta = 1$) distribution (see [1], for example). Here we suppose that

$$\mathfrak{d} \xrightarrow{M\to\infty} d$$

where $0 < d < \infty$. If we choose X to have complex entries ($W = \frac{1}{M} XX^*$) then we would arrive at the Tracy–Widom ($\beta = 2$) distribution. Specifying real versus complex through $\beta = 1$ versus $\beta = 2$ is common practice in the random matrix literature and we continue this practice in the current work.

Universality was first combined with the average-case analysis of algorithms in [35], then expanded in [8], with rigorous results presented in [10, 11]. See [12] for a review. Here we summarize a result found in [10] concerning the power method. The power method itself is the simple iteration

$$\begin{aligned} \boldsymbol{y}_k &= W\boldsymbol{x}_{k-1}, \\ \nu_k &= \boldsymbol{y}_k^T x_{k-1}, \qquad k = 1, 2, \ldots, \\ \boldsymbol{x}_k &= \boldsymbol{y}_k / \|y_k\|_2, \end{aligned}$$

where $\boldsymbol{x}_0$ is a starting unit vector that is often, in practice, chosen randomly. If, for example, W is positive definite, then $\nu_k \to \lambda_1(W)$ as $k \to \infty$. A relevant question is to understand how many iterations are required to properly approximate $\lambda_1(W)$. Given the halting time

$$T(W, \boldsymbol{x}_0, \epsilon) = \min\{k : |\nu_k - \nu_{k-1}| < \epsilon^2\},$$

a result from [10] gives the distributional limit

$$\lim_{N\to\infty} \mathbb{P}\left(\frac{T(W, \boldsymbol{x}_0, \epsilon)}{\widetilde{c}_d N^{2/3}\big(\log\epsilon - \frac{2}{3}\log N\big)} \le t\right) = F_\beta^{\mathrm{gap}}(t), \quad \epsilon \le N^{-5/3-\sigma}, \tag{1.3}$$

for $t \ge 0$, $\sigma > 0$ and a constant $\widetilde{c}_d$. Here $F_\beta^{\mathrm{gap}}(t)$ can be expressed in terms of the limiting distribution of

$$\frac{1}{N^{2/3}(\lambda_1(W) - \lambda_2(W))}.$$

But, more importantly, $F_\beta^{\mathrm{gap}}(t)$ only depends on β and not on the precise distribution on the entries of X. One may also consider the distribution of $\nu_k - \nu_{k-1}$ as $M \to \infty$ and ask whether it is universal.

The purpose of this article is threefold.

- We present a full derivation of distributional formulae for the conjugate gradient algorithm (CGA) and the MINRES algorithm applied to linear systems $W\boldsymbol{x} = \mathbf{b}$ where W is distributed as in (1.1), addressing both the real and complex cases. A formula for the CGA applied to the normal

equations $W\boldsymbol{x} = \frac{X}{\sqrt{M}}\mathbf{b}$ is also given. This elementary derivation pulls on many well-known results at the intersection of numerical linear algebra and random matrix theory. In particular, the derivation involves many algorithms that are well-known to the applied mathematics community: the QR factorization, Golub–Kahan bidiagonalization, singular value decomposition, Lanczos iteration, and Cholesky factorization.

- We then show how universality theorems for the so-called anisotropic local law [26] can be upgraded to give universality theorems for the moments of discrete measures that arise in the Lanczos and conjugate gradient algorithms. This is the key component in showing that the behavior determined in the asymptotic analysis of the formulae in the case of Gaussian matrices indeed persists for a wide class of non-Gaussian matrices giving universality for the norms of residual and error vectors for the CGA and MINRES algorithms. In the well-conditioned case

$$\mathfrak{d} \xrightarrow{M\to\infty} d \in (0,1),$$

  the number of iterations of the algorithm to achieve a tolerance ϵ (i.e., the halting time) is almost deterministic.
- Because the calculations are so explicit and the estimates are so exact, this work can be viewed as a benchmark for the average-case analysis of an algorithm. This shows that it is indeed possible to completely analyze an algorithm, in a specific regime, applied to wide class of random matrix distributions.

Currently, $\mathfrak{d} \approx 1$ behavior of the CGA and MINRES algorithms on Wishart matrices is open. By this, we are referring to determining the (asymptotic) distribution on the number of iterations required to achieve a tolerance of ϵ. Numerical experiments indicate that a universality statement analogous to (1.3) holds for the CGA provided M and N are scaled appropriately [8], the limiting distribution is conjectured to be Gaussian [9], and the leading-order behavior is conjectured in [28].

So, in this paper we focus on fixed ϵ while running the algorithms $O(1)$ steps. The leading-order analysis along these lines was completed for Gaussian entries in [13]. This confirmed that the deterministic analysis of Beckermann and Kuijlaars [4] (see also [27]) holds in the random setting with overwhelming probability. In this paper we improve upon and simplify the results in [13] in many respects. In particular, our exact distributional formulae (see Theorem 1.2) can be used to establish many, but not all, of the results in [13]. We then prove that the leading-order results in [13] are universal and provide the universal distributional limit (after rescaling) for the fluctuations. This also provides a universal, almost-deterministic halting time (see Remarks 1.6 and 1.7). Such almost-deterministic halting times for the CGA were first observed in [9] and proved in [13] in the Gaussian case. See [34] for similar results in the case of gradient descent.

While our analysis for the CGA and MINRES algorithms is focused on sample covariance matrices of the form (1.1), many other distributions should be analyzable. One example would be $I + \gamma G$, $G = \frac{X+X^T}{\sqrt{2N}}$ where X is an $N \times N$ iid Gaussian matrix. This is the shifted Gaussian orthogonal ensemble. For a definite and well-conditioned problem, one should choose $\gamma < 1/2$. Another interesting case is for sample covariance matrices $T^{1/2} X X^T T^{1/2}$, for deterministic positive definite matrix T, which correspond to sample covariance matrices with nonidentity covariance. But in either of these cases, one can run the Lanczos iteration on it and ask about the distribution on the tridiagonalization that results. The leading-order behavior is implied by [42]. And indeed, as we discuss, this fact is qualitatively implied by the fact that the entries in the Lanczos matrix are differentiable functions of the moments of an associated spectral measure.

The paper is laid out as follows. In this section we fix notation, introduce the Gaussian distributions from which we perturb and discuss the algorithms that we will analyze. We present our main results in Theorems 1.1, 1.2, 1.4 and 1.5. The section closes with a numerical demonstration of the theorems. In Section 2 we introduce the notion of sample covariance matrices and the moment matching condition and discuss properties of basic algorithms applied to Gaussian matrices. Section 3 gives some properties of orthogonal polynomials that are critical in our calculations. Section 4 gives a deterministic description of the CGA and MINRES algorithm along with the derivation of formulae for the errors that result from the algorithms. The main probabilistic contribution of the paper is in Section 5. It comes in the form of a "four moment theorem" for the spectral measure. Lastly, Section 6 completes the proofs of our main theorems.

We also point out that subsequent to the current work, the papers [14, 15] have extended our results in various ways. In [15] the current results were largely extended to the case of spiked sample covariance matrices with nontrivial covariance. Fluctuations in this paper were shown to be universal but not specifically identified as Gaussian. This gap was filled in [14] and this work allows k, the number of steps in the CGA to depend on N in a nontrivial way and allowed for multiple intervals of support for the limiting eigenvalue distribution.

1.1 Notation

Throughout this article we use boldface, e.g., $\boldsymbol{y}$, to denote vectors. The norm $\|\boldsymbol{y}\|_2^2 = \boldsymbol{y}^* \boldsymbol{y}$ gives the usual 2-norm. The expression $W > 0$ indicates that W is a real-symmetric or complex-Hermitian positive definite matrix. And W then induces an important norm $\|\boldsymbol{y}\|_W^2 = \boldsymbol{y}^* W \boldsymbol{y}$. We then use $\lambda_1(W) \geq \lambda_2(W) \geq \cdots \lambda_N(W)$ to denote the eigenvalues of W.

The notation $\mathcal{N}_\beta(\mu, \sigma^2)$ refers to a real ($\beta = 1$) or complex ($\beta = 2$) normal random variable with mean μ and variance σ^2 and the symbol $\stackrel{\mathscr{L}}{=}$ refers to equality in law. The notation

$$x_M \xrightarrow[M\to\infty]{(d)} y$$

denotes convergence in distribution, or weak convergence. Additionally, since we will be using $\mathbf{e}_k$ to denote error vectors arising in the approximate solution of linear systems, we use $\boldsymbol{f}_1, \dots, \boldsymbol{f}_n$ to denote the standard basis of $\mathbb{R}^n$ where n is inferred from context. The notation $\chi_{\beta k}$ is used to denote the chi distribution with βk degrees of freedom parameterized[1] by

$$\chi_{\beta k} \overset{\mathscr{L}}{=} \left(\sum_{j=1}^{k} |X_j|^2 \right)^{1/2},$$

where $(X_j)_{j=1}^k$ are iid $\mathcal{N}_\beta(0,1)$ random variables.

We also encounter settings where the size of a random matrix or vector is increasing as a parameter $M \to \infty$. We say that, for example,

$$(x_j)_{j=1}^M =: \boldsymbol{x}_M \xrightarrow[M\to\infty]{(d)} \boldsymbol{y},$$

$\boldsymbol{y} = (y_j)_{j=1}^\infty$ in the sense of convergence of finite-dimensional marginals if for any finite set S of integers

$$(x_j)_{j\in S} \xrightarrow[M\to\infty]{(d)} (y_j)_{j\in S}.$$

This notion is very convenient as it allows one to bypass dimension mismatches between processes. Lastly, we will use subblock notation $X_{i:k,j:\ell}$ to denote the subblock of the matrix X that contains rows i through k and columns j through ℓ.

1.2 The Wishart distributions

Suppose X is an $N \times M$ matrix of iid $\mathcal{N}_\beta(0,1)$ normal random variables. Then we say that $X \overset{\mathscr{L}}{=} \mathcal{G}_\beta(N,M)$, and we say $W = XX^*/M$ has the β-Wishart distribution and write $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N,M)$. The β-Wishart distributions in the cases[2] $\beta = 1, 2$ has many important properties that we will use extensively. In addition, classical algorithms from numerical linear algebra act on these matrices in a way that allows for explicit (distributional) calculations.

1.3 The conjugate gradient and MINRES algorithms

The CGA [24] is an iterative method to solve a linear system $W\boldsymbol{x} = \mathbf{b}$ where $W > 0$. Supposing exact arithmetic, the algorithm is simplest to characterize in its varational form. Define the Krylov subspace

$$\mathcal{K}_k = \operatorname{span}\{\mathbf{b}, W\mathbf{b}, \dots, W^{k-1}\mathbf{b}\}. \tag{1.4}$$

[1] Parameterizing a distribution is expressing it as a transformation of well-understood random variables.

[2] The case $\beta = 4$ can be introduced using quarternions.

Then the k^{th} iterate, x_k, of the CGA satisfies[3]

$$\boldsymbol{x}_k = \operatorname{argmin}_{\boldsymbol{y}\in\mathcal{K}_k} \|\boldsymbol{x}-\boldsymbol{y}\|_W.$$

In Section 4 the algorithm that is often used to compute $\boldsymbol{x}_k$ effectively is presented but since our analysis assumes exact arithmetic, this algorithm is not needed to perform the analysis.

The MINRES algorithm (see Algorithm 3 below) is another iterative method that works with $\mathcal{K}_k$ by again producing a sequence sequence of vectors

$$\boldsymbol{x}_1 \to \cdots \to \boldsymbol{x}_k,$$

but for the MINRES algorithm each vector $\boldsymbol{x}_k$ solves

$$\boldsymbol{x}_k = \operatorname{argmin}_{\boldsymbol{y}\in\mathcal{K}_k} \|\boldsymbol{b}-W\boldsymbol{y}\|_2.$$

For both the CGA and the MINRES algorithm we use the notation $\mathbf{r}_k(W,\mathbf{b}) := \mathbf{b}-W\boldsymbol{x}_k$ and $\mathbf{e}_k(W,\mathbf{b}) := \boldsymbol{x}-\boldsymbol{x}_k$ to denote the residual and error vectors, respectively.

1.4 Main results

We first establish some deterministic formulae. The result for $\|\mathbf{r}_k\|_2$ in the CG algorithm is entirely classical as it encapsulates a well-known relation between b_{k-1} in Algorithm 2 below and the entries in the matrix generated by the Lanczos procedure (see [29], for example). The proof is found in Sections 6.1, 6.2, and 6.3. See Algorithm 1 and the surrounding text for a discussion of the Lanczos iteration.

THEOREM 1.1 (Deterministic formulae). *Consider the Lanczos iteration applied to the pair* $(W,\mathbf{b})$ *with* $W>0$ *and* $\|\mathbf{b}\|_2=1$. *Suppose the iteration terminates at step* $n\le N$ *producing a matrix* $T=T(W,\mathbf{b})$. *Let* $T=HH^T$ *be the Cholesky factorization* (*see Algorithm* 4 *below*) *of* T *where*

$$H = \begin{bmatrix} \alpha_0 & & & & \\ \beta_0 & \alpha_1 & & & \\ & \beta_1 & \alpha_2 & & \\ & & \ddots & \ddots & \\ & & & \beta_{n-2} & \alpha_{n-1} \end{bmatrix}.$$

(a) *For the CGA on* $W\boldsymbol{x}=\mathbf{b}$ *with* $\boldsymbol{x}_0=0$, *for* $k<n$,

$$\|\mathbf{r}_k(W,\mathbf{b})\|_2 = \prod_{j=0}^{k-1}\frac{\beta_j}{\alpha_j},$$

$$\|\mathbf{e}_k(W,\mathbf{b})\|_W = \|\mathbf{r}_k(W,\mathbf{b})\|_2\sqrt{\boldsymbol{f}_1^*(L_kL_k^T)^{-1}\boldsymbol{f}_1}, \quad L_k = H_{k+1:n,k+1:n}.$$

[3] Here we are characterizing the CGA with $\boldsymbol{x}_0=0$.

(b) *For the MINRES algorithm on* $W\boldsymbol{x} = \mathbf{b}$*, for* $k < n$*,*

$$\|\mathbf{r}_k(W,\mathbf{b})\|_2 = \left(\sum_{j=0}^{k}\prod_{\ell=0}^{j-1}\frac{\alpha_\ell^2}{\beta_\ell^2}\right)^{-1/2}.$$

And $\mathbf{r}_n = 0$*.*

THEOREM 1.2 (CG and MINRES on $\mathcal{W}_\beta(N,M)$). *Suppose* $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N,M)$ *with* $N \le M$ *and* $\mathbf{b} \in \mathbb{R}^N$ $(\beta = 1)$ *or* $\mathbf{b} \in \mathbb{C}^N$ $(\beta = 2)$ *nonzero. Let* $\alpha_j \overset{\mathscr{L}}{=} \chi_{\beta(M-j)}$*,* $\beta_j \overset{\mathscr{L}}{=} \chi_{\beta(N-j-1)}$*,* $j = 0,1,\dots,$ *be independent and* $k < N$*.*

(a) *For the CGA applied to* $W\boldsymbol{x} = \mathbf{b}$ *with* $\boldsymbol{x}_0 = 0$*,*

$$\|\mathbf{r}_k(W,\mathbf{b})\|_2 \overset{\mathscr{L}}{=} \|\mathbf{b}\|_2 \prod_{j=0}^{k-1}\frac{\beta_j}{\alpha_j},$$

$$\|\mathbf{e}_k(W,\mathbf{b})\|_W \overset{\mathscr{L}}{=} \Sigma_k^{-1}\|\mathbf{r}_k\|_2, \quad \Sigma_k^{-1} \overset{\mathscr{L}}{=} \frac{\sqrt{\beta M}}{\chi_{\beta(M-N+1)}},$$

where Σ_k^{-1} *is independent of* α_j, β_j*,* $j = 0,1,\dots,k-1$*, but dependent on* α_j, β_j*,* $j \ge k$*.*

(b) *For the MINRES algorithm applied to*[4] $W\boldsymbol{x} = \mathbf{b}$*,*

$$\|\mathbf{r}_k(W,\mathbf{b})\|_2 \overset{\mathscr{L}}{=} \left(\sum_{j=0}^{k}\prod_{\ell=0}^{j-1}\frac{\alpha_\ell^2}{\beta_\ell^2}\right)^{-1/2}\|\mathbf{b}\|_2.$$

(c) *Now suppose* $\mathbf{b} \in \mathbb{R}^M$ $(\beta = 1)$ *or* $\boldsymbol{b} \in \mathbb{C}^M$ $(\beta = 2)$ *is nonzero, and* $X \overset{\mathscr{L}}{=} \mathcal{G}_\beta(N,M)$*,* $N \le M$*. For the CGA applied to* $W\boldsymbol{x} = \frac{X}{\sqrt{M}}\mathbf{b}$*,* $W = \frac{XX^*}{M}$*,*

$$\left\|\mathbf{e}_k\left(W, \frac{X}{\sqrt{M}}\mathbf{b}\right)\right\|_W \overset{\mathscr{L}}{=} \Delta_{N,M}\left(\sum_{j=0}^{k}\prod_{\ell=0}^{j-1}\frac{\alpha_\ell^2}{\beta_\ell^2}\right)^{-1/2}\|\mathbf{b}\|_2,$$

where $\Delta_{N,M} \overset{\mathscr{L}}{=} \frac{\chi^2_{\beta N}}{\chi^2_{\beta M}}$ *may have nontrivial correlations with* α_j, β_j*,* $j = 0,1,2,\dots,$ *but does not depend on* k*.*

Remark 1.3. From Theorem 1.2(c) we obtain a complete parametrization of the relative errors

$$\frac{\left\|\mathbf{e}_k\left(W, \frac{X}{\sqrt{M}}\mathbf{b}\right)\right\|_W}{\left\|\mathbf{e}_0\left(W, \frac{X}{\sqrt{M}}\mathbf{b}\right)\right\|_W} \overset{\mathscr{L}}{=} \left(\sum_{j=0}^{k}\prod_{\ell=0}^{j-1}\frac{\alpha_\ell^2}{\beta_\ell^2}\right)^{-1/2}.$$

To state the next couple results, we define the parameter $\mathfrak{d} = N/M$.

[4] We use the convention that $\prod_{\ell=0}^{-1} \equiv 1$.

THEOREM 1.4 (Universality to leading order). *Let* $W = XX^*$ *where* X *is an* $N \times M$ *random matrix* $N \le M$,

$$\mathfrak{d} \xrightarrow{M\to\infty} d \in (0,1],$$

with independent real ($\beta = 1$) *or complex* ($\beta = 2$) *entries. Suppose, in addition, that there exists constants* $\{C_p\}_1^\infty$ *so that* all *entries of* X *satisfy, for nonnegative integers* ℓ, p,

$$\begin{aligned} &\mathbb{E}(\Re X_{ij})^\ell(\Im X_{ij})^p = \mathbb{E}(\Re Y)^\ell(\Im Y)^p, \\ &Y \overset{\mathscr{L}}{=} \mathcal{N}_\beta(0, 1/M), \quad \ell + p \le 2, \\ &\mathbb{E}|\sqrt{M}X_{ij}|^p \le C_p \quad \textit{for all } p \in \mathbb{N}. \end{aligned} \tag{1.5}$$

For any sequence $\mathbf{b} = \mathbf{b}_N$ *of unit vectors, in the sense of convergence of finite-dimensional marginals*:

(a) *For the CGA*[5]

$$\begin{aligned} &\left(\|\mathbf{e}_k(W,\mathbf{b})\|_W^2\right)_{k\ge 0} \xrightarrow[M\to\infty]{(d)} \left(\frac{d^k}{1-d}\right)_{k\ge 0}, \ d \ne 1, \\ &\left(\|\mathbf{r}_k(W,\mathbf{b})\|_2^2\right)_{k\ge 1} \xrightarrow[M\to\infty]{(d)} (d^k)_{k\ge 1}. \end{aligned}$$

(b) *For the MINRES algorithm*

$$\left(\|\mathbf{r}_k(W,\mathbf{b})\|_2^2\right)_{k\ge 1} \xrightarrow[M\to\infty]{(d)} \left(d^k\frac{1-d}{1-d^{k+1}}\right)_{k\ge 1}.$$

(c) *For the CGA applied to the normal equations*

$$\left(\left\|\mathbf{e}_k\left(W, \frac{X}{\sqrt{M}}\mathbf{b}\right)\right\|_W^2\right)_{k\ge 0} \xrightarrow[M\to\infty]{(d)} \left(d^{k+1}\frac{1-d}{1-d^{k+1}}\right)_{k\ge 0}.$$

The case $d = 1$ in Theorem 1.4 is treated by continuity, $d \uparrow 1$. To state our last limit theorem, we must define the limit processes. Let $\mathcal{G} = (Z_k)_{k=1}^\infty$ be a process of independent $\mathcal{N}_1(0,1)$ random variables. Define three new processes, $\mathcal{G}^{\mathbf{e}} = (Z_k^{\mathbf{e}})_{k=0}^\infty$, $\mathcal{G}^{\mathbf{r},\mathrm{CG}} = (Z_j^{\boldsymbol{r},\mathrm{CG}})_{j=1}^\infty$, and $\mathcal{G}^{\mathbf{r},\mathrm{MINRES}} = (Z_j^{\boldsymbol{r},\mathrm{MINRES}})_{j=1}^\infty$ via

$$\begin{aligned} Z_k^{\mathbf{e}} = \frac{d^k}{1-d}\Bigg[&\sum_{j=k}^\infty d^{j-k}(Z_{2j}/\sqrt{d} - Z_{2j+1}) \\ &+ \sum_{j=1}^{k-1}(Z_{2j}/\sqrt{d} - Z_{2j-1}) - Z_{2k-1}\Bigg], \end{aligned}$$

[5] We do not discuss $\|\mathbf{r}_0\|_2$ here because $\mathbf{r}_0 = \mathbf{b}$.

$$Z_k^{\mathbf{r},\mathrm{CG}} = d^k \left[\sum_{j=0}^{k-1} \left(Z_{2j+2}/\sqrt{d} - Z_{2j+1} \right) \right], \quad k > 0, \quad Z_0^{\mathbf{r},\mathrm{CG}} = 0,$$

$$Z_k^{\mathbf{r},\mathrm{MINRES}} = \left(\frac{1-d}{1-d^{k+1}} \right)^2 \sum_{j=0}^{k} d^{2(k-j)} Z_j^{\mathbf{r},\mathrm{CG}}.$$

THEOREM 1.5 (Universality of the fluctuations). *Let $W = XX^*$ where X is an $N \times M$ random matrix, $N \leq M$,*

$$\mathfrak{d} \xrightarrow{M\to\infty} d \in (0,1]$$

with independent real ($\beta = 1$) or complex ($\beta = 2$) entries. Suppose, in addition, that there exists constants $\{C_p\}_1^\infty$ so that all *entries of X satisfy, for nonnegative integers ℓ, p,*

$$\begin{aligned} &\mathbb{E}(\Re X_{ij})^\ell (\Im X_{ij})^p = \mathbb{E}(\Re Y)^\ell (\Im Y)^p, \\ &Y \overset{\mathscr{L}}{=} \mathcal{N}_\beta(0, 1/M), \quad \ell + p \leq 4, \\ &\mathbb{E}|\sqrt{M} X_{ij}|^p \leq C_p \quad \text{for all } p \in \mathbb{N}. \end{aligned} \tag{1.6}$$

For any sequence $\mathbf{b} = \mathbf{b}_N$ of unit vectors, in the sense of convergence of finite-dimensional marginals:

(a) *For the CGA*

$$\sqrt{\frac{\beta M}{2}} \left(\|\mathbf{e}_k(W,\mathbf{b})\|_W^2 - \frac{\mathfrak{d}^k}{1-\mathfrak{d}} \right)_{k\geq 0} \xrightarrow[M\to\infty]{(d)} \mathcal{G}^{\mathbf{e}}, \quad d \neq 1,$$

$$\sqrt{\frac{\beta M}{2}} \left(\|\mathbf{r}_k(W,\mathbf{b})\|_2^2 - \mathfrak{d}^k \right)_{k\geq 1} \xrightarrow[M\to\infty]{(d)} \mathcal{G}^{\mathbf{r},\mathrm{CG}}.$$

(b) *For the MINRES algorithm (the case $\mathfrak{d} = 1$ obtained using continuity)*

$$\sqrt{\frac{\beta M}{2}} \left(\|\mathbf{r}_k(W,\mathbf{b})\|_2^2 - \mathfrak{d}^k \frac{1-\mathfrak{d}}{1-\mathfrak{d}^{k+1}} \right)_{k\geq 1} \xrightarrow[M\to\infty]{(d)} \mathcal{G}^{\mathbf{r},\mathrm{MINRES}}.$$

(c) *For the CGA applied to the normal equations (the case $\mathfrak{d} \to 1$ obtained using continuity)*

$$\sqrt{\frac{\beta M}{2}} \left(\frac{\left\| \mathbf{e}_k\left(W, \frac{X}{\sqrt{M}}\mathbf{b}\right) \right\|_W^2}{\left\| \mathbf{e}_0\left(W, \frac{X}{\sqrt{M}}\mathbf{b}\right) \right\|_W^2} - \mathfrak{d}^k \frac{1-\mathfrak{d}}{1-\mathfrak{d}^{k+1}} \right)_{k\geq 1} \xrightarrow[M\to\infty]{(d)} \mathcal{G}^{\mathbf{r},\mathrm{MINRES}}.$$

The proofs of the previous theorems can be roughly summarized as follows. Modulo some technical issues in dealing with correlations, Theorem 1.2 can be directly used, with the asymptotics of independent chi random variables, to prove Theorem 1.4 and 1.5 in the case

$$\sqrt{M} X \overset{\mathscr{L}}{=} \mathcal{G}_\beta(N, M).$$

Asymptotic correlations are addressed in Proposition 5.11. Associated to $(W, \mathbf{b})$, $W > 0$, $\|\mathbf{b}\|_2 = 1$ is a weighted empirical spectral measure (see (2.3) below). The orthogonal polynomials with respect to this measure satisfy a three-term recurrence which when assembled into a Jacobi matrix coincides with the output $T_n(W, \mathbf{b})$ of the Lanczos iteration (see Proposition 3.1 below). Then the well-known fact that the entries in the three-term recurrence Jacobi matrix can be recovered as algebraic functions of the moments of the measure is used (see (3.2)). This means that the entries in the Cholesky factorization of $T_n(W, b)$ are (generically) differentiable functions of the moments of the weighted empirical spectral measure. Then Theorem 5.15 establishes universality for the moments and hence for the entries in the Cholesky factorization. More specifically, this implies that Proposition 5.11 holds in the non-Gaussian case, implying our theorems.

Some important remarks are in order.

Remark 1.6. Let W, $d < 1$, and $\mathbf{b}$ be as in Theorem 1.4. Define two CGA halting times

$$t^{\mathbf{e}}(W, \mathbf{b}, \epsilon) = \min\{k : \|e_k(W, \mathbf{b})\|_W < \epsilon\},$$
$$t^{\mathbf{r}}(W, \mathbf{b}, \epsilon) = \min\{k : \|\mathbf{r}_k(W, \mathbf{b})\|_2 < \epsilon\}.$$

If $\epsilon^2 \neq d^k/(1-d)$ for all k

$$\lim_{N\to\infty} \mathbb{P}\left(t^{\mathbf{e}}(W, \mathbf{b}, \epsilon) = \left\lceil \frac{\log \epsilon^2(1-d)}{\log d} \right\rceil\right) = 1,$$

and if $\epsilon^2 = d^k/(1-d)$ for some k, then

$$\lim_{N\to\infty} \mathbb{P}\left(t^{\mathbf{e}}(W, \mathbf{b}, \epsilon) = \left\lceil \frac{\log \epsilon^2(1-d)}{\log d} \right\rceil\right) = \frac{1}{2},$$
$$\lim_{N\to\infty} \mathbb{P}\left(t^{\mathbf{e}}(W, \mathbf{b}, \epsilon) = 1 + \left\lceil \frac{\log \epsilon^2(1-d)}{\log d} \right\rceil\right) = \frac{1}{2}.$$

Similarly, if $\epsilon^2 \neq d^k$ for all k

$$\lim_{N\to\infty} \mathbb{P}\left(t^{\mathbf{r}}(W, \mathbf{b}, \epsilon) = \left\lceil \frac{2\log \epsilon)}{\log d} \right\rceil\right) = 1,$$

and if $\epsilon = d^k$ for some k, then

$$\lim_{N\to\infty} \mathbb{P}\left(t^{\mathbf{r}}(W, \boldsymbol{b}, \epsilon) = \left\lceil \frac{2\log \epsilon}{\log d} \right\rceil\right)$$
$$= \frac{1}{2} = \lim_{N\to\infty} \mathbb{P}\left(t^{\mathbf{r}}(W, \mathbf{b}, \epsilon) = 1 + \left\lceil \frac{2\log \epsilon}{\log d} \right\rceil\right).$$

Remark 1.7. Let W, $d < 1$, and $\mathbf{b}$ be as in Theorem 1.4. Define the MINRES halting time

$$t^{\text{MINRES}}(W, \mathbf{b}, \epsilon) = \min\{k : \|r_k(W, \mathbf{b})\|_2 < \epsilon\}.$$

Then if $\epsilon^2 \neq d^k \frac{1-d}{1-d^{k+1}}$ for all k

$$\lim_{N\to\infty} \mathbb{P}\left(t^{\text{MINRES}}(W, \mathbf{b}, \epsilon) = \left\lceil \frac{\log \frac{\epsilon^2}{1-d+\epsilon^2 d}}{\log d} \right\rceil \right) = 1,$$

and if $\epsilon^2 = d^k \frac{1-d}{1-d^{k+1}}$ for some k then

$$\lim_{N\to\infty} \mathbb{P}\left(t^{\text{MINRES}}(W, \mathbf{b}, \epsilon) = \left\lceil \frac{\log \frac{\epsilon^2}{1-d+\epsilon^2 d}}{\log d} \right\rceil \right) = \frac{1}{2},$$

$$\lim_{N\to\infty} \mathbb{P}\left(t^{\text{MINRES}}(W, \mathbf{b}, \epsilon) = 1 + \left\lceil \frac{\log \frac{\epsilon^2}{1-d+\epsilon^2 d}}{\log d} \right\rceil \right) = \frac{1}{2}.$$

And so, the MINRES algorithm, using the halting criterion $\|r_k\|_2 < \epsilon$ will run for approximately $\frac{\log(1-d+\epsilon^2 d)}{\log d}$ fewer steps than the CGA.

Remark 1.8. Let W, $d < 1$, and $\mathbf{b}$ be as in Theorem 1.5. For fixed k

$$\begin{aligned} \sqrt{\frac{\beta M}{2}} \left(\|\mathbf{e}_k(W, \mathbf{b})\|_W^2 - \frac{d^k}{1-d} \right) &\xrightarrow[M\to\infty]{(d)} \mathcal{N}_1(0, \sigma_{k,\mathbf{e}}^2), \quad d \neq 1, \\ \sigma_{k,\mathbf{e}}^2 &= \frac{d^{2k}}{(1-d)^2} \left[\frac{1}{d(1-d)} + (k-1)\left(1 + \frac{1}{d}\right) + 1 \right], \\ \sqrt{\frac{\beta M}{2}} \left(\|\mathbf{r}_k(W, \mathbf{b})\|_2^2 - d^k \right) &\xrightarrow[M\to\infty]{(d)} \mathcal{N}_1(0, \sigma_{k,\mathbf{r}}^2), \\ \sigma_{k,\mathbf{r}}^2 &= k d^{2k} \left(1 + \frac{1}{d}\right). \end{aligned} \tag{1.7}$$

Remark 1.9. The expression for $Z_k^{\mathbf{r},\text{MINRES}}$ can be written as

$$Z_k^{\mathbf{r},\text{MINRES}} = d^{2k} \left(\frac{1-d}{1-d^{k+1}} \right)^2 \sum_{\ell=0}^{k} \frac{d^{-k} - d^{-\ell}}{1-d} \left(Z_{2\ell+2}/\sqrt{d} - Z_{2\ell+1} \right), \quad k > 0.$$

Let W, $d < 1$, and $\mathbf{b}$ be as in Theorem 1.5. For fixed k it then follows that

$$\begin{aligned} \sqrt{\frac{\beta M}{2}} \left(\|\mathbf{r}_k(W, \mathbf{b})\|_2^2 - \mathfrak{d}^k \right) &\xrightarrow[M\to\infty]{(d)} \mathcal{N}_1(0, \widehat{\sigma}_{k,\mathbf{r}}^2), \\ \widehat{\sigma}_{k,\mathbf{r}}^2 &= \frac{(1-d)d^{2k-1}\left(2d^{k+1} + 2d^{k+2} - d^{2k+2} - d^2(k+1) - 2d + k\right)}{\left(1-d^{k+1}\right)^4}. \end{aligned} \tag{1.8}$$

Remark 1.10. Additionally, one obtains the formulae for the CGA applied to $W\boldsymbol{x} = \mathbf{b}$, $\|\mathbf{b}\|_2 = 1$, $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N, M)$, $N \leq M$:

$$\begin{aligned}\mathbb{E}\|\mathbf{r}_k(W,\mathbf{b})\|_2 &= \prod_{j=0}^{k-1} \frac{\Gamma\left(\frac{\beta(N-j-1)+1}{2}\right)}{\Gamma\left(\frac{\beta(N-j-1)}{2}\right)} \frac{\Gamma\left(\frac{\beta(M-j)-1}{2}\right)}{\Gamma\left(\frac{\beta(M-j)}{2}\right)},\\ \mathbb{E}\|\mathbf{e}_k(W,\mathbf{b})\|_W &= \sqrt{\frac{\beta M}{2}} \frac{\Gamma\left(\frac{\beta(M-N+1)-1}{2}\right)}{\Gamma\left(\frac{\beta(M-N+1)}{2}\right)} \prod_{j=0}^{k-1} \frac{\Gamma\left(\frac{\beta(N-j-1)+1}{2}\right)}{\Gamma\left(\frac{\beta(N-j-1)}{2}\right)} \frac{\Gamma\left(\frac{\beta(M-j)-1}{2}\right)}{\Gamma\left(\frac{\beta(M-j)}{2}\right)},\end{aligned} \tag{1.9}$$

where $\Gamma(z)$ is the Gamma function [31]. For even moderately large M, one needs to use the Beta function to compute these ratios and avoid underflow/overflow.

Remark 1.11. For $\mathfrak{d} \to 1$, the CGA applied to $W\boldsymbol{x} = \mathbf{b}$ gives

$$\frac{\|\mathbf{r}_k(W,\mathbf{b})\|_2}{\|\mathbf{r}_0(W,\mathbf{b})\|_2} \xrightarrow[M\to\infty]{(d)} 1.$$

Thus the number of iterations required to hit a tolerance ϵ increases without bound as $M \to \infty$. On the other hand, for the MINRES algorithm,

$$\frac{\|\mathbf{r}_k(W,\mathbf{b})\|_2}{\|\mathbf{r}_0(W,\mathbf{b})\|_2} \xrightarrow[M\to\infty]{(d)} \frac{1}{\sqrt{k+1}}.$$

And so, one expects $k \approx \epsilon^{-2} - 1$ iterations to achieve $\|\mathbf{r}_k(W,\mathbf{b})\|_2 < \epsilon$. The same statement holds for the CGA applied to the normal equations when $\mathfrak{d} \to 1$, when one considers the ratio

$$\frac{\left\|\mathbf{e}_k\left(W, \frac{X}{\sqrt{M}}\mathbf{b}\right)\right\|_W}{\left\|\mathbf{e}_0\left(W, \frac{X}{\sqrt{M}}\mathbf{b}\right)\right\|_W}.$$

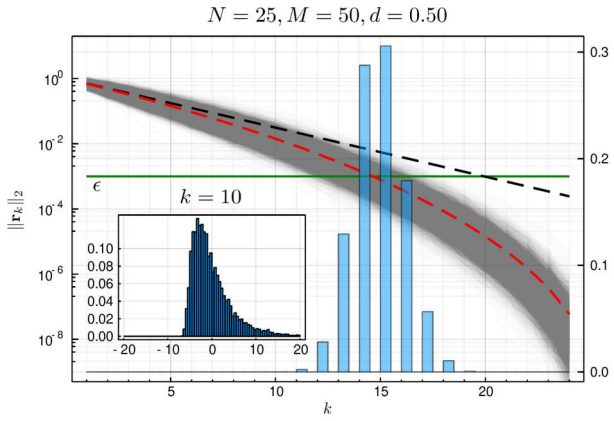
Remark 1.12. If $\mathbf{b} = \boldsymbol{c}/\|\boldsymbol{c}\|_2$ where $\boldsymbol{c}$ has iid, mean-zero entries with a finite (nonzero) variance then one expects (1.5) to be sufficient for Theorem 1.5 to hold—the moment matching to order 2 is sufficient if the right-hand side vector is "sufficiently" random.

We demonstrate the essential aspects of Theorem 1.5(a) for $\|\boldsymbol{r}_k\|_2$ in Figures 1.1 and 1.2. In these figures we compare the CGA applied to $W\boldsymbol{x} = \boldsymbol{f}_1$ with $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N, M)$ and $W = XX^*/M$ where X has iid entries with

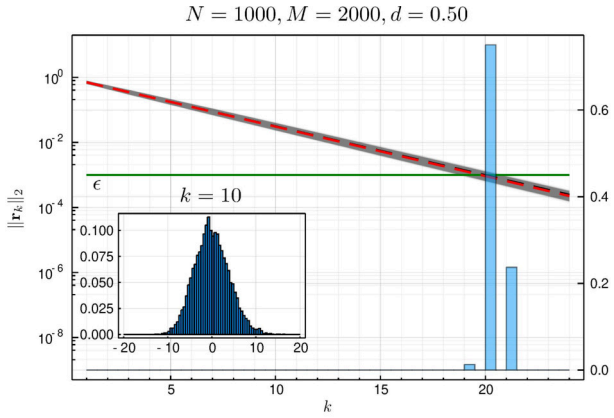
$$\mathbb{P}(X_{ij} = 0) = 2/3, \quad \mathbb{P}(X_{ij} = \pm\sqrt{3}) = 1/6. \tag{1.10}$$

This discrete distribution, which we refer to as the moment matching distribution, is chosen so that the first four moments of X_{ij} coincide with that of $\mathcal{N}_1(0,1)$. The figures demonstrate that $\|r_k\|_2$ concentrates heavily as M increases.

The essential aspects of Theorem 1.5(b) are shown in Figures 1.3 and 1.4. These figures again give the behavior of the MINRES algorithm and CGA applied to the $\beta = 1$ Wishart distribution and the moment matching distribution.

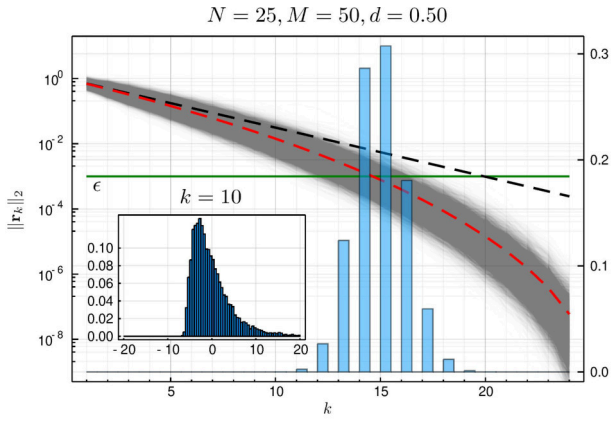


(a)

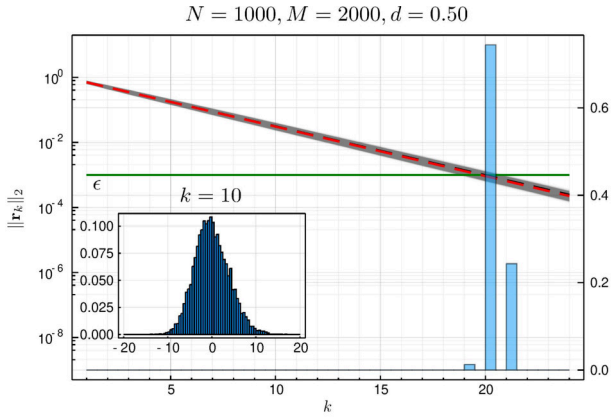


(b)

FIGURE 1.1. The CGA applied to $W\boldsymbol{x} = \boldsymbol{f}_1$ were $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N, M)$, $N/M \xrightarrow{M\to\infty} d$. The dashed black curve indicates the large M limit for the error $\|\boldsymbol{r}_k\|_2$ at step k and the dashed red curve gives $\mathbb{E}\|\boldsymbol{r}_k\|_2$ at step k. The shaded gray area is an ensemble of 10000 runs of the method, displaying the norms that resulted. The overlaid histogram shows the rescaled fluctuations in the error at $k = 10$. As $M \to \infty$ this approaches a Gaussian density. Lastly, the histogram in the main frame gives the halting distribution for $\epsilon = 0.001$ (green line). It is highly concentrated when $N = 1000, M = 2000$. With these parameters, Remark 1.6 implies that for M large, the algorithm will run for approximately $\lceil 2\frac{\log \epsilon}{\log d} \rceil = 20$ iterations.

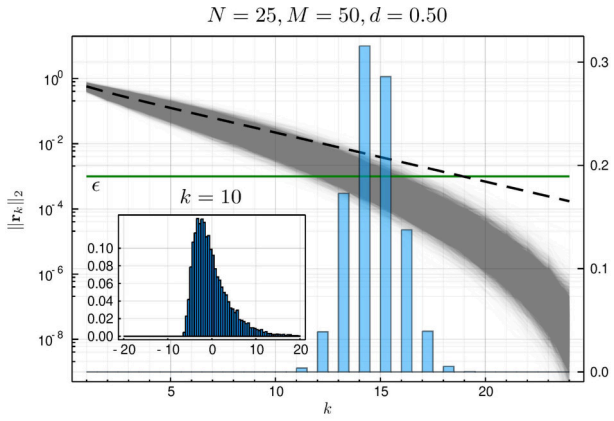


(a)

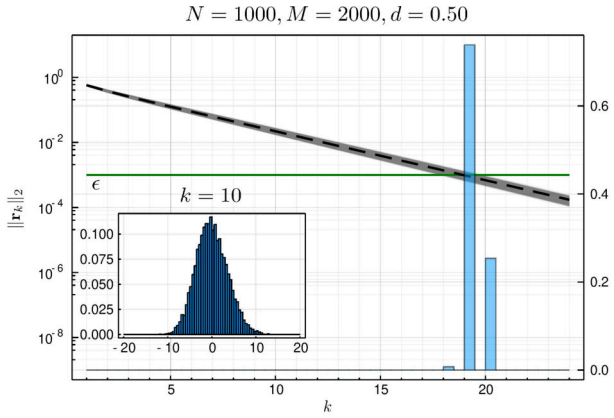


(b)

FIGURE 1.2. The CGA applied to $W\boldsymbol{x} = M^{-1}XX^*\boldsymbol{x} = f_1$ where X has iid entries with $\mathbb{P}(X_{ij} = 0) = 2/3, \mathbb{P}(X_{ij} = \pm\sqrt{3}) = 1/6$. The black dashed curve indicates the large M limit for the error $\|\boldsymbol{r}_k\|_2$ at step k the dashed red curve gives $\mathbb{E}\|\boldsymbol{r}_k\|_2$ at step k in the case of $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N, M)$, for comparison. The shaded gray area is an ensemble of 10000 runs of the method, displaying the errors that resulted. The overlaid histogram shows the rescaled fluctuations in the error at $k = 10$. As $M \to \infty$ this approaches a Gaussian density. Lastly, the histogram in the main frame gives the halting distribution for $\epsilon = 0.001$ (green line). With these parameters, Remark 1.6 implies that for M large, the algorithm will run for approximately $\lceil 2\frac{\log \epsilon}{\log d} \rceil = 20$ iterations.



(a)

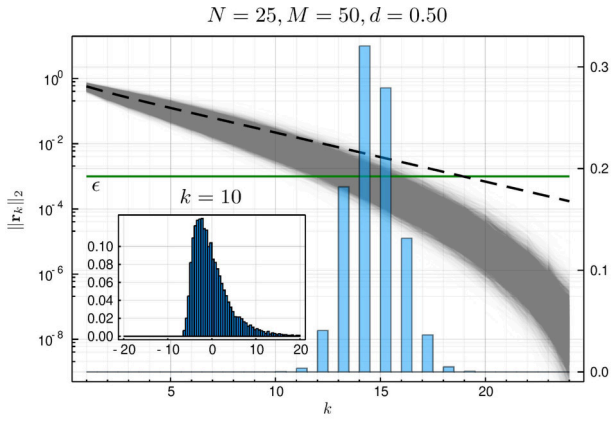


(b)

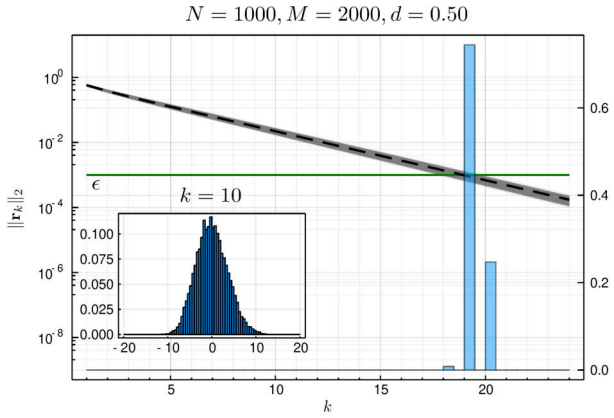
FIGURE 1.3. The MINRES algorithm applied to $W\boldsymbol{x} = \boldsymbol{f}_1$ where $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N, M)$. The dashed curve indicates the large M limit for the error $\|\boldsymbol{r}_k\|_2$ at step k. The shaded gray area is an ensemble of 10000 runs of the method, displaying the errors that resulted. The overlaid histogram shows the rescaled fluctuations in the error at $k = 10$. As $M \to \infty$ this approaches a Gaussian density. Lastly, the histogram in the main frame gives the halting distribution for $\epsilon = 0.001$ (green line). With these parameters, Remark 1.7 implies that for M large, the algorithm will run for approximately $\left\lceil \frac{\log \frac{\epsilon^2}{1-d+\epsilon^2 d}}{\log d} \right\rceil = 19$ iterations.

Lastly, in Figure 1.5, for the CGA, we compare the statistics of

$$\sqrt{M}\left(\frac{\|\mathbf{r}_k(W, \boldsymbol{f}_1)\|_2}{\langle \|\mathbf{r}_k(W, \boldsymbol{f}_1)\|_2 \rangle} - 1\right), \tag{1.11}$$



(a)



(b)

FIGURE 1.4. The MINRES algorithm applied to $W\boldsymbol{x} = M^{-1}XX^*\boldsymbol{x} = f_1$ were X has iid entries with $\mathbb{P}(X_{ij} = 0) = 2/3, \mathbb{P}(X_{ij} = \pm\sqrt{3}) = 1/6$. The shaded gray area is an ensemble of 10000 runs of the method, displaying the errors that resulted. The overlaid histogram shows the rescaled fluctuations in the error at $k = 10$. As $M \to \infty$ this approaches a Gaussian density. Lastly, the histogram in the main frame gives the halting distribution for $\epsilon = 0.001$ (green line). With these parameters, Remark 1.7 implies that for M large, the algorithm will run for approximately $\left\lceil \frac{\log \frac{\epsilon^2}{1-d+\epsilon^2 d}}{\log d} \right\rceil = 19$ iterations.

where $\langle Z \rangle$ represents the sample average of Z over 50,000 samples. Note that if (1.6) holds then

$$\sqrt{M}\left(\frac{\|\mathbf{r}_k(W, \boldsymbol{f}_1)\|_2}{\langle \|\mathbf{r}_k(W, \boldsymbol{f}_1)\|_2 \rangle} - 1\right) \approx \mathcal{N}_1(0, \sigma_{k,d}/2),$$

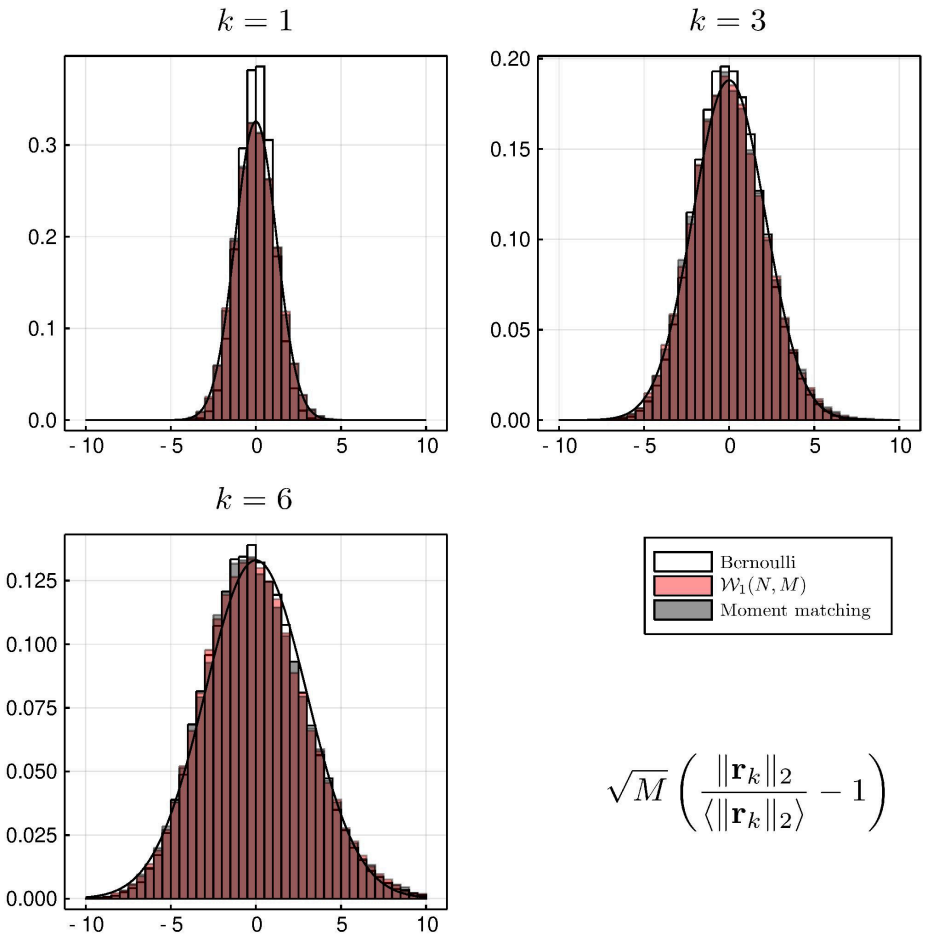


FIGURE 1.5. A comparison of the rescaled statistics (1.11) across three distributions. Since the Bernoulli ensemble fails to match the moments in (1.6), we see that it does not match the variance (1.7),

and we therefore compare the density for $\mathcal{N}_1(0, \sigma_{k,d}/2)$ with (1.11) in Figure 1.5. In this figure we also include computations with the Bernoulli ensemble: $W = XX^*/M$, X_{ij} iid, $\mathbb{P}(X_{ij} = \pm 1) = 1/2$ which fails to satisfy (1.6).

In Table 1.1 we display sample variance of (1.11) for the three different distributions: Wishart, moment matching and Bernoulli. In the case of the Wishart and moment matching distributions, the variance is close to the large M limit. In the case of Bernoulli, the variance is quite different. This indicates that the moment matching condition is a necessary condition for the limiting the variance to be given by (1.7).

k	$k/2(1+1/d)$	Wishart	Moment matching (see (1.10))	Bernoulli
1	1.5	1.493	1.48	1.003
2	3.0	3.002	2.997	2.511
3	4.5	4.532	4.519	4.036
4	6.0	6.040	6.039	5.527
5	7.5	7.576	7.54	7.004
6	9.0	9.135	9.054	8.547

TABLE 1.1. A numerical demonstration of the necessity of the moment matching condition (1.6). This table gives the sample variance of (1.11) across three different distributions for $N = 500, d = 1/2$ and 50,000 samples. Presumably, the values in the last column differ from the values in the other columns for a reason other than a lack of samples.

2 Sample Covariance Matrices and Classical Numerical Linear Algebra

A fundamental property of a matrix $X \overset{\mathscr{L}}{=} \mathcal{G}_\beta(N, M)$ is its orthogonal ($\beta = 1$) or unitary ($\beta = 2$) invariance. That is, let Q be an $N \times N$ fixed orthogonal matrix then

$$QWQ^* \overset{\mathscr{L}}{=} W, \quad W = XX^*.$$

If $\beta = 2$, then Q can be a complex unitary matrix. Furthermore, this is true even if Q is random, provided it is independent of X.

Let $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N, M)$ and perform an eigenvalue decomposition $W = U\Lambda U^*$, $U^*U = I$. It follows directly from the invariance of the Wishart distribution that the vector

$$\boldsymbol{\omega} = \begin{bmatrix} |U_{11}|^2 \\ \vdots \\ |U_{1n}|^2 \end{bmatrix} \quad \text{where } \left[U_{ij}\right]_{1 \le i,j \le n} = U$$

can be parameterized by

$$\boldsymbol{\omega} \overset{\mathscr{L}}{=} \frac{\boldsymbol{\nu}}{\|\boldsymbol{\nu}\|_1}, \tag{2.1}$$

where $\boldsymbol{\nu}$ is a vector of iid χ^2_β random variables. This fact is discussed in detail in [13, app. A].

The eigenvalues of the Wishart distributions

The global asymptotic eigenvalue distribution of the Wishart distributions is the same, regardless of the choice of $\beta = 1, 2$. The classical setup is the following.

For $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N, M)$, define the (random) empirical spectral measure

$$\mu_{\mathrm{em}}(\mathrm{d}\lambda; W) = \frac{1}{N}\sum_{j=1}^{N} \delta_{\lambda_j(W)}(\mathrm{d}\lambda).$$

The eigenvalues $\{\lambda_j(W) : 1 \le j \le N\}$ are well–known to have the distribution of the Laguerre orthogonal ensemble or the Laguerre unitary ensemble, respectively, according to whether $\beta = 1, 2$. Recall the parameter $\mathfrak{d} = N/M$.

DEFINITION 2.1. Define the *Marchenko–Pastur* law *for all* $d > 0$ by

$$\varrho_d(\mathrm{d}x) = \frac{1}{2\pi d}\sqrt{\frac{[(x-\gamma_-)(\gamma_+ - x)]_+}{x^2}}\mathrm{d}x + \left[1 - \frac{1}{d}\right]_+ \delta_0(\mathrm{d}x), \tag{2.2}$$
$$\text{where } \gamma_\pm = (1 \pm \sqrt{d})^2$$

are the spectral edges. The notation $[\cdot]_+$ refers to the positive part of $(\cdot)$.

The following gives the global eigenvalue distribution (see [1], for example):

THEOREM 2.2. *Suppose that* $\mathfrak{d} \xrightarrow{M\to\infty} d \in (0, 1]$. *Then*

$$\mu_{\mathrm{em}}(\mathrm{d}\lambda; \mathcal{W}_\beta(N, M)) \xrightarrow[N\to\infty]{(d)} \varrho_d(\mathrm{d}\lambda),$$

almost surely.

Historically, the behavior of individual eigenvalues, and gaps between eigenvalues, have been studied extensively. In the analysis we present it is not necessary to use such detailed microscopic results. Instead, we need finer results about global properties of the matrix. One such example is the so-called central limit theorem for linear statistics.

The Bai-Silverstein [3] central limit theorem for linear statistics of sample covariance matrices shows that for sufficiently smooth functions f,

$$\sum_{j=1}^{N} f(\lambda_j) - N\int f(x)\varrho_{\mathfrak{d}}(\mathrm{d}x)$$
$$= N\int f(x)(\mu_{\mathrm{em}}(\mathrm{d}x) - \varrho_{\mathfrak{d}}(\mathrm{d}x)) \xrightarrow[N\to\infty]{(d)} \mathcal{N}_1\big(\mu_f, \sigma_f^2\big).$$

The standard deviation σ_f can be understood as a weighted Sobolev-1/2 norm of f, restricted to the support of the Marchenko-Pastur law. Other related central limit theorems for linear spectral statistics of sample covariance matrices include [17, 25, 37].

But the classical central limit theorem for linear statistics involves the empirical spectral measure $\mu_{\mathrm{em}}(\mathrm{d}\lambda; W)$ which rarely arises in a numerical or computational

context. What is much more likely to arise is the weighted empirical spectral measure: for $\boldsymbol{b} \in \mathbb{C}^N$, $\|\boldsymbol{b}\|_2 = 1$ and $W = W^* \in \mathbb{C}^{N\times N}$ the weighted empirical spectral measure is given by

$$\mu = \mu_{\boldsymbol{b}} = \sum_{j=1}^{N} \omega_j \delta_{\lambda_j}, \quad (\omega_j)_{j=1}^N = |U^*\boldsymbol{b}|^2, \quad W = U\Lambda U^*,$$
$$U^*U = I, \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_N).$$

We refer to this as the spectral measure associated to the pair $(W, \boldsymbol{b})$.

We show in Section 4 that for polynomials p and a sample covariance matrix W with identity covariance and for which $\mathfrak{d} \to d$,

$$\sqrt{M} \int f(x)(\mu_{\boldsymbol{b}}(\mathrm{d}x) - \varrho_{\mathfrak{d}}(dx)) \xrightarrow[N\to\infty]{(d)} \mathcal{N}_1(0, \widehat{\sigma_f}^2).$$

Note that the rate of the central limit theorem changes dramatically from the case of the central limit theorem for linear statistics. Although we will not need it, the variance $\widehat{\sigma_f}^2$ can be expressed as $c_{\beta,d} \int f^2(x)\varrho_d(dx)$. Similar theorems have been proven before, most notably by [33] who prove a more general statement in the case that $\boldsymbol{b}$ is a coordinate vector. There is also [32] in which the analogous statement is made for Wigner matrices. We also mention [18] and [19] which prove related theorems for Gaussian cases.

While it is natural to assume these statements extend to other classes of test functions beyond polynomials, we will not need them (except for the specific case of $f(x) = 1/x$, which we handle by other means – note that the extension to analytic functions in a neighborhood of the Marchenko-Pastur law does not need new ideas beyond what is necessary for the polynomial case)

2.1 Sample covariance matrices with independence

In the current work, we use a restricted definition of a sample covariance matrix.

DEFINITION 2.3. A real ($\beta = 1$) or complex ($\beta = 2$) sample covariance matrix is given by $W \overset{\mathscr{L}}{=} XX^*$ where X is an $N \times M$ random matrix with independent entries satisfying

$$\mathbb{E}X_{ij} = 0, \ \mathbb{E}(\Re X_{ij})(\Im X_{ij}) = 0, \ \mathbb{E}(\Re X_{ij})^2 = \frac{1}{\beta M},$$
$$\mathbb{E}|X_{ij}|^2 = \frac{1}{M}, \quad \text{and} \quad \mathbb{E}|\sqrt{M}X_{ij}|^p \le C_p, \quad \text{for all } p \in \mathbb{N}.$$

In some cases, we will need restrictions on the first four generalized moments.

DEFINITION 2.4. A sample covariance matrix satisfies the $\beta = 1, 2$ moment matching condition if

$$\mathbb{E}(\Re X_{ij})^\ell (\Im X_{ij})^p = \mathbb{E}(\Re Y)^\ell (\Im Y)^p$$

where $Y \overset{\mathscr{L}}{=} \mathcal{N}_\beta(0, 1/M)$, for all choices of non-negative integers ℓ, p such that $\ell + p \leq 4$.

Remark 2.5. To see the necessity of the moment matching condition consider a sample covariance matrix $W' = XX/M$ where X' is $N \times M$, with $X'_{ij} = \pm 1$ with equal probability and $W \overset{\mathscr{L}}{=} \mathcal{W}_1(N, M)$. Then consider the first moments of the spectral measures μ and μ' associated to $(W, \boldsymbol{f}_1)$ and $(W', \boldsymbol{f}_1)$, respectively:

$$\int \lambda \mu(\mathrm{d}\lambda) = \frac{1}{M} \boldsymbol{f}_1^T X X^T \boldsymbol{f}_1 \overset{\mathscr{L}}{=} \frac{\chi^2_M}{M},$$
$$\int \lambda \mu'(\mathrm{d}\lambda) = \frac{1}{M} \boldsymbol{f}_1^T X' X'^T \boldsymbol{f}_1 = 1.$$

2.2 The Golub–Kahan bidiagonalization algorithm

DEFINITION 2.6. A Jacobi matrix is given by

$$T = \begin{bmatrix} a_0 & b_0 & & \\ b_0 & a_1 & b_1 & \\ & b_1 & a_2 & \ddots \\ & & \ddots & \ddots \end{bmatrix}.$$

It may be finite or semi-infinite. The entries are real and $b_j > 0$ for $j \geq 0$. If $b_j = 0$ for some j, the matrix T is called a degenerate Jacobi matrix.

A reduction of $W = XX^*$ to a (possibly degenerate) Jacobi matrix can be obtained via the Golub–Kahan bidiagonalization procedure. The distributional action of this algorithm on the Wishart ensembles $\mathcal{W}_\beta(N, M)$ is given in [16]. Specifically, if $W = M^{-1} XX^* \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N, M)$, $X \overset{\mathscr{L}}{=} \mathcal{G}_\beta(N, M)$ then there exists unitary matrices U_1, U_2 such that

$$U_1 X U_2 \overset{\mathscr{L}}{=} \sqrt{\beta} X_{\mathrm{GK}},$$
$$\sqrt{\beta} X_{\mathrm{GK}} \overset{\mathscr{L}}{=} \left[\begin{array}{ccccccc|c} \chi_{\beta M} & & & & & & & \\ \chi_{\beta(N-1)} & \chi_{\beta(M-1)} & & & & & & \\ & \chi_{\beta(N-2)} & \chi_{\beta(M-2)} & & & & & \\ & & \ddots & \ddots & & & & 0 \\ & & & & \chi_\beta & \chi_{\beta(M-N+1)} & & \end{array}\right], \tag{2.3}$$

where all entries are independent. Therefore the law of the entries of the tridiagonal matrix $U_1 W U_1^* = M^{-1} U_1 X U_2 U_2^* X^* U_1^* = \beta X_{\mathrm{GK}} X_{\mathrm{GK}}^T$ is completely parameterized.

2.3 The Lanczos iteration

The Lanczos iteration is another algorithm for obtaining a tridiagonal reduction of a matrix.

Algorithm 1: Lanczos Iteration

(1) $\boldsymbol{q}_1$ is the initial vector. Suppose $\|\boldsymbol{q}_1\|_2^2 = \boldsymbol{q}_1^* \boldsymbol{q}_1 = 1$, $W^* = W$.
(2) Set $b_{-1} = 1$, $\boldsymbol{q}_0 = 0$.
(3) For $k = 1, 2, \ldots, n$
  (a) Compute $a_{k-1} = (W\boldsymbol{q}_k - b_{k-2}\boldsymbol{q}_{k-1})^* \boldsymbol{q}_k$.
  (b) Set $\boldsymbol{v}_k = W\boldsymbol{q}_k - a_{k-1}\boldsymbol{q}_k - b_{k-2}\boldsymbol{q}_{k-1}$.
  (c) Compute $b_{k-1} = \|\boldsymbol{v}_k\|_2$ and if $b_{k-1} \neq 0$, set $\boldsymbol{q}_{k+1} = \boldsymbol{v}_k / b_{k-1}$, otherwise terminate.

The Lanczos algorithm at step $k \leq N$ produces a matrix T_k and orthogonal vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k$

$$Q_k = \begin{bmatrix} \boldsymbol{q}_1 & \boldsymbol{q}_2 & \cdots & \boldsymbol{q}_k \end{bmatrix}, \quad T_k = T_k(W, \boldsymbol{y}_1) = \begin{bmatrix} a_0 & b_0 & & \\ b_0 & a_1 & \ddots & \\ & \ddots & \ddots & b_{k-2} \\ & & b_{k-2} & a_{k-1} \end{bmatrix},$$

such that

$$WQ_k = Q_k T_k + b_{k-1} \boldsymbol{q}_{k+1} \boldsymbol{f}_k^* \tag{2.4}$$

We use the notation $T = T(W, \boldsymbol{q}_1) = T_n(W, \boldsymbol{q}_1)$ for the matrix produced when the Lanczos iteration is run for its maximum of n steps.

The following is entirely classical [41].

LEMMA 2.7. *Suppose W is a symmetric matrix. And suppose that the Lanczos iteration does not terminate before step $n \leq N$. For $k = 1, 2, \ldots, n$,*

$$\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k$$

is an orthonormal basis for the Krylov subspace $\mathcal{K}_k = \operatorname{span}\{\boldsymbol{q}_1, W\boldsymbol{q}_1, \ldots, W^{k-1}\boldsymbol{q}_1\}$.

The following result gives us the distribution of T_k throughout the Lanczos iteration applied to a Wishart matrix and it is a direct consequence of the invariance of the Wishart distributions.

THEOREM 2.8. *Suppose $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N, M)$. For any given $\boldsymbol{q}_1 \in \mathbb{R}^n$ with $\|\boldsymbol{q}_1\|_2 = 1$(or $\mathbb{C}^n$ for $\beta = 2$) with probability one, the Lanczos iteration does not terminate if $k < n := \min\{N, M\}$. And the distribution on a_k, b_k, $k = 0, 2, \ldots, n-1$ does not depend on $\boldsymbol{q}_1$. In a distributional sense it suffices to take $\boldsymbol{q}_1 = \boldsymbol{f}_1$ and therefore the distribution is determined by the Householder tridiagonalization of W, i.e., the Golub–Kahan bidiagonalization of X.*

Every $N \times N$ Jacobi matrix T produces a probability measure

$$\mu_T = \sum_{j=1}^{N} \omega_j \delta_{\lambda_j}$$

where λ_j's are the eigenvalues of T and ω_j is the squared modulus of the first component of the normalized eigenvector associated to λ_j. The spectral measure μ_T, $T = T(W, \boldsymbol{b})$ coincides with the spectral measure associated to the pair $(W, \boldsymbol{b})$ whenever $\boldsymbol{b}$ is a unit vector. This mapping is a bijection between probability measures supported on n points and Jacobi matrices of dimension n [7].

3 Theory of Orthogonal Polynomials

Let μ be a Borel probability measure on $\mathbb{R}$ with finite moments. The orthonormal polynomials $(p_n)_{n\geq 0}$, $p_n(\lambda) = p_n(\lambda; \mu)$ are constructed by applying the Gram–Schmidt process to the sequence of functions

$$\{\lambda \mapsto 1, \lambda \mapsto \lambda, \lambda \mapsto \lambda^2, \ldots\}.$$

If the support of μ contains at least N points, then one is guaranteed to be able to construct $(p_0, p_1, \ldots, p_{N-1})$.

3.1 Hankel determinants, moments, and the three-term recurrence

We now recall the classical fact that the coefficients in a three-term recurrence relation can be recovered as an algebraic function of the moments of the associated spectral measure. For a given sequence of orthonormal polynomials, $(p_j(\lambda))_{j\geq 0} = (p_j(x; \mu))_{j\geq 0}$ with respect to a measure[6] μ, we have the associated three-term recurrence

$$\lambda p_n(\lambda) = b_n p_{n+1}(\lambda) + a_n p_n(\lambda) + b_{n-1} p_{n-1}(\lambda), \quad n \geq 0, \quad b_n > 0, \tag{3.1}$$

with the convention $p_{-1}(\lambda) = 0$ and $b_{-1} = 0$. Here $b_n = b_n(\mu)$, $a_n = a_n(\mu)$ are called the recurrence coefficients. We will use the following proposition in a critical way to translate any discussion of the output of the Lanczos iteration to a discussion of orthogonal polynomials.

PROPOSITION 3.1. *The three-term recurrence coefficients generated by the spectral measure associated to the pair* $(W, \mathbf{b})$, $W > 0, \|\mathbf{b}\|_2 = 1$ *coincide with the entries of the Lanczos matrix* $T(W, \mathbf{b})$.

We write $p_n(\lambda) = \ell_n \lambda^n + s_n \lambda^{j-1} + \cdots$ and find by equating coefficients that

$$\begin{aligned} \ell_n &= b_n \ell_{n+1}, \\ a_n \ell_n &= b_n s_{n+1}. \end{aligned}$$

Define D_n and $D_n(\lambda)$ by the determinants

$$\begin{aligned} D_n &= \det M_n, \quad (M_n)_{ij} = m_{i+j-2}, \quad 1 \leq i, j \leq n+1, \\ m_j(\mu) = m_j &= \int \lambda^j \mu(\mathrm{d}\lambda), \quad D_n(\lambda) = \det M_n(\lambda), \end{aligned}$$

[6] For our purposes it suffices to assume that μ has compact support.

and $M_n(\lambda)$ is formed by replacing the last row of M_n with the row vector $[1\ \lambda\ \lambda^2 \cdots \lambda^n]$. Then, it is well-known that [7]

$$p_n(\lambda) = \frac{D_n(\lambda)}{\sqrt{D_n D_{n-1}}},$$

and therefore

$$\ell_n = \sqrt{\frac{D_{n-1}}{D_n}}, \quad s_n = \det \widetilde{M}_n, \tag{3.2}$$

where $\widetilde{M}_n$ is the matrix formed by removing the last row and second-to-last column of M_n. This shows that $a_n/\sqrt{D_{n-1}}$ and b_n^2 are rational functions of determinants of matrices involving only the moments of μ up to order $2n$.

Associated to the three-term recurrence (3.1) is the Jacobi matrix

$$T = \begin{bmatrix} a_0 & b_0 & & \\ b_0 & a_1 & b_1 & \\ & b_1 & a_2 & \ddots \\ & & \ddots & \ddots \end{bmatrix}.$$

Let T_n denote the upper-left $n \times n$ subblock of T. It follows immediately that T_n is a differentiable function of $(m_0, m_1, \ldots, m_{2n})$ on the open subset of $\mathbb{R}^{2n+1}$ where all $D_k > 0$ for $1 \le k \le n$. We also note that

$$\boldsymbol{f}_1^* T^k \boldsymbol{f}_1 = \int \lambda^k \mu(\mathrm{d}\lambda). \tag{3.3}$$

This can be seen by a direct calculation if T is a finite-dimensional matrix. If T is semi-infinite, then this fact follows from [7, (2.25)].

3.2 Monic polynomials and Stieltjes transforms

The monic orthogonal polynomials associated to a measure μ are given by

$$\pi_n(\lambda; \mu) = \pi_n(\lambda) = p_n(\lambda)/\ell_n = \lambda^n + \cdots. \tag{3.4}$$

We will also need the Stieltjes transform of the monic polynomials

$$c_n(z; \mu) = c_n(z) = \int_{\mathbb{R}} \frac{\pi_n(\lambda)}{\lambda - z} \mu(\mathrm{d}\lambda). \tag{3.5}$$

With the convention that $b_0 = 1$, $\pi_{-1} \equiv 0$, and $c_{-1} \equiv -1$, it is elementary that the following recurrences are satisfied for $n = 0, 1, 2, \ldots,$

$$\pi_{n+1}(\lambda) = (\lambda - a_n)\pi_n(\lambda) - b_{n-1}^2 \pi_{n-1}(\lambda), \quad \pi_0(\lambda) = 1,$$

$$c_{n+1}(z) = (z - a_n)c_n(z) - b_{n-1}^2 c_{n-1}(z), \quad c_0(z) = \int_{\mathbb{R}} \frac{\mu(\mathrm{d}\lambda)}{\lambda - z}.$$

4 The Conjugate Gradient Algorithm and the MINRES Algorithm

In this section we discuss three algorithms: the CGA, the CGA applied to the normal equations, and the MINRES algorithm.

4.1 The CGA

The actual CGA is given by the following.

Algorithm 2: Conjugate Gradient Algorithm

(1) $\boldsymbol{x}_0$ is the initial guess.
(2) Set $\mathbf{r}_0 = \mathbf{b} - W\boldsymbol{x}_0$, $\boldsymbol{p}_0 = \mathbf{r}_0$.
(3) For $k = 1, 2, \ldots, n$
  (a) Compute $a_{k-1} = \dfrac{\mathbf{r}_{k-1}^* \boldsymbol{r}_{k-1}}{\mathbf{r}_{k-1}^* W \boldsymbol{p}_{k-1}}$.
  (b) Set $\boldsymbol{x}_k = \boldsymbol{x}_{k-1} + a_{k-1}\boldsymbol{p}_{k-1}$.
  (c) Set $\mathbf{r}_k = \mathbf{r}_{k-1} - a_{k-1} W \boldsymbol{p}_{k-1}$.
  (d) Compute $b_{k-1} = -\dfrac{\mathbf{r}_k^* \mathbf{r}_k}{\boldsymbol{r}_{k-1}^* \mathbf{r}_{k-1}}$.
  (e) Set $\boldsymbol{p}_k = \mathbf{r}_k - b_{k-1}\boldsymbol{p}_{k-1}$.

As noted previously, a remarkable fact is that the iterates x_k of the CGA applied to the linear system $W\boldsymbol{x} = \mathbf{b}$ are given by the solution of the minimization problem (1.4) [24]. From this, we see that $\boldsymbol{y} \in \mathcal{K}_k$ can be written as

$$\boldsymbol{y} = \sum_{j=0}^{k-1} c_j W^j \mathbf{b} \quad \Rightarrow \quad \boldsymbol{x} - \boldsymbol{y} = W^{-1}\left(\mathbf{b} - \sum_{j=0}^{k-1} c_j W^{j+1}\mathbf{b}\right) = W^{-1} q_{\boldsymbol{y}}(W)\mathbf{b},$$

for a polynomial $q_{\boldsymbol{y}}$ of degree at most k and it satisfies $q_{\boldsymbol{y}}(0) = 1$. Then, computing further,

$$\|\boldsymbol{x} - \boldsymbol{y}\|_W^2 = \mathbf{b}^* q_{\boldsymbol{y}}(W)^* W^{-1} q_{\boldsymbol{y}}(W)\mathbf{b}.$$

And setting $W = U\Lambda U^*$, we find

$$\|\boldsymbol{x} - \boldsymbol{y}\|_W^2 = \sum_{j=1}^{N} \frac{|q_{\boldsymbol{y}}(\lambda_j)|^2}{\lambda_j} |(U^*\mathbf{b})_j|^2 = \int \frac{|q_{\boldsymbol{y}}(\lambda)|^2}{\lambda} \mu_T(\mathrm{d}\lambda), \quad T = T(W, \mathbf{b}).$$

Now, all directional derivatives of this, when $\boldsymbol{y} = \boldsymbol{x}_k$, with respect to coefficients of the polynomial must vanish identically. This gives a characterization of $q_{\boldsymbol{x}_k}$: Let δq_k be a polynomial of degree at most k that satisfies $\delta q_k(0) = 0$, and we must have

$$0 = \int q_{\boldsymbol{x}_k}(\lambda) \frac{\delta q_k(\lambda)}{\lambda} \mu_T(\mathrm{d}\lambda).$$

This implies that $q_{\boldsymbol{x}_k}(\lambda)$ is orthogonal to all lower-degree polynomials, with respect to μ_T: It is given by

$$q_{\boldsymbol{x}_k}(\lambda) = \frac{\pi_k(\lambda;\mu_T)}{\pi_k(0;\mu_T)}.$$

PROPOSITION 4.1. *Let $\boldsymbol{x}_k$ be the computed solution at step k of the CGA applied to $W\boldsymbol{x} = \mathbf{b}$. For any $k \in \mathbb{N}$, with $T = T(W,\mathbf{b})$,*

$$\|\mathbf{e}_k\|_W^2 = \frac{c_k(0;\mu_T)}{\pi_k(0;\mu_T)} \quad \textit{and} \quad \|\mathbf{r}_k\|_2^2 = \frac{\prod_{j=0}^{k-1} b_j(\mu_T)^2}{\pi_k(0;\mu_T)^2}.$$

PROOF. By orthogonality

$$\begin{aligned}\|\mathbf{e}_k\|_W^2 &= \int_{\mathbb{R}} \frac{\pi_k(\lambda;\mu_T)^2}{\lambda\pi_k(0;\mu_T)^2}\mu_T(d\lambda)\\ &= \int_{\mathbb{R}} \frac{\pi_k(\lambda;\mu_T)\left(\pi_k(0;\mu_T)\lambda^{-1} + \sum_{j=1}^{k} c_j\lambda^{k-1}\right)}{\pi_k(0;\mu_T)^2}\mu_T(d\lambda)\\ &= \int_{\mathbb{R}} \frac{\pi_k(\lambda;\mu_T)}{\lambda\pi_k(0;\mu_T)}\mu_T(d\lambda) = \frac{c_k(0;\mu_T)}{\pi_k(0;\mu_T)}.\end{aligned}$$

For the $\mathbf{r}_k$ equation, by definition of the polynomials $\{p_n\}$, we have that

$$\begin{aligned}\int_{\mathbb{R}} \pi_k(\lambda;\mu_T)^2\mu_T(d\lambda) &= \frac{1}{\ell_k^2}\int_{\mathbb{R}} p_k(\lambda;\mu_T)^2\mu_T(d\lambda)\\ &= \frac{1}{\ell_k^2} = \prod_{j=0}^{k-1} b_j(\mu_T)^2.\end{aligned} \tag{4.1}$$

□

4.2 MINRES

The MINRES algorithm, at iteration k, gives the solution of

$$\boldsymbol{x}_k = \operatorname{argmin}_{\boldsymbol{y}\in\mathcal{K}_k} \|\mathbf{b} - W\boldsymbol{y}\|_2.$$

More explicitly, the algorithm is given by

Algorithm 3: MINRES Algorithm for $W\boldsymbol{x} = \mathbf{b}$

(1) Suppose $W = W^* \in \mathbb{C}^{N\times N}, \epsilon > 0$.
(2) Set $\boldsymbol{q}_1 = \mathbf{b}/\|\mathbf{b}\|_2$.
(3) For $k = 1, 2, \ldots, n, n \leq N$
  (a) Compute $a_{k-1} = (W\boldsymbol{q}_k - b_{k-2}\boldsymbol{q}_{k-1})^*\boldsymbol{q}_k$.
  (b) Set $\boldsymbol{v}_k = W\boldsymbol{q}_k - a_{k-1}\boldsymbol{q}_k - b_{k-2}\boldsymbol{q}_{k-1}$.
  (c) Compute $b_{k-1} = \|\boldsymbol{v}_k\|_2$ and if $b_{k-1} \neq 0$, set $\boldsymbol{q}_{k+1} = \boldsymbol{v}_k/b_{k-1}$.
  (d) Form

$$\widetilde{T}_k = \begin{bmatrix} a_0 & b_0 & & \\ b_0 & a_1 & \ddots & \\ & \ddots & \ddots & b_{k-2} \\ & & b_{k-2} & a_{k-1} \\ & & & b_{k-1} \end{bmatrix}.$$

  (e) Compute $z_k = \operatorname{argmin}_{z\in\mathbb{C}^k} \|\widetilde{T}_k z - \|\mathbf{b}\|_2 \boldsymbol{f}_1\|_2$.
  (f) If $\|\widetilde{T}_k z_k - \|\mathbf{b}\|_2 \boldsymbol{f}_1\|_2 < \epsilon$, return $\boldsymbol{x}_k = \begin{bmatrix}\boldsymbol{q}_1 & \cdots & \boldsymbol{q}_k\end{bmatrix} z_k$.

Following the same prescription as in the previous section we are led to the problem of finding the polynomial $r_{\boldsymbol{x}_k}$ of degree less than or equal to k satisfying $r_{\boldsymbol{x}_k}(0) = 1$ that minimizes

$$\|\mathbf{b} - W\boldsymbol{y}\|_2^2 = \sum_{j=1}^{N} |r_{\boldsymbol{y}}(\lambda_j)|^2 |(U^*\mathbf{b})_j|^2 = \int |r_{\boldsymbol{y}}(\lambda)|^2 \mu_T(\mathrm{d}\lambda), \quad T = T(W, \mathbf{b}),$$

among all such polynomials. We then must have

$$0 = \int r_{\boldsymbol{x}_k}(\lambda)\delta r_k(\lambda)\mu_T(\mathrm{d}\lambda)$$

for all polynomials δr_k of degree less than or equal to k with $\delta r_k(0) = 0$. So, write $r_{x_k}(\lambda) = \sum_{j=0}^{k} c_j p_j(\lambda; \mu_T)$. And choosing $\delta r_k(\lambda) = p_\ell(\lambda; \mu_T) - p_\ell(0; \mu_T)$ we find

$$0 = \int \left(\sum_{j=0}^{k} c_j p_j(\lambda; \mu_T)\right) (p_\ell(\lambda; \mu_T) - p_\ell(0; \mu_T))\, \mu_T(\mathrm{d}\lambda) \Leftrightarrow c_\ell = p_\ell(0; \mu_T)c_0.$$

From this, we obtain

$$r_{\boldsymbol{x}_k}(\lambda; \mu_T) = \frac{\sum_{j=0}^{k} p_j(0; \mu_T) p_j(\lambda; \mu_T)}{\sum_{j=0}^{k} p_j^2(0; \mu_T)}. \tag{4.2}$$

PROPOSITION 4.2. *Let $\boldsymbol{x}_k$ be the computed solution at step k of the MINRES algorithm applied to $W\boldsymbol{x} = \mathbf{b}$. For any $k \in \mathbb{N}$, with $T = T(W, \mathbf{b})$*

$$\begin{aligned}\|\mathbf{r}_k\|_2^2 &= \frac{1}{\sum_{j=0}^{k} p_j^2(0;\mu_T)} \\ &= \frac{1}{b_k(\mu_T)^2\left[p'_{k+1}(0;\mu_T)p_k(0;\mu_T) - p'_k(0;\mu_T)p_{k+1}(0;\mu_T)\right]} \\ &= \frac{\prod_{j=0}^{k-1} b_j(\mu_T)^2}{\pi'_{k+1}(0)\pi_k(0;\mu_T) - \pi'_k(0;\mu_T)\pi_{k+1}(0;\mu_T)}.\end{aligned}$$

PROOF. Integrating (4.2)

$$\|\mathbf{b} - W\boldsymbol{x}_k\|_2^2 = \frac{1}{\sum_{j=0}^{k} p_j^2(0;\mu_T)}.$$

Employing the Christoffel-Darboux formula,

$$\begin{aligned}\sum_{j=0}^{k} p_j^2(0;\mu_T) &= \frac{\ell_k}{\ell_{k+1}}\left[p'_{k+1}(0;\mu_T)p_k(0;\mu_T) - p'_k(0;\mu_T)p_{k+1}(0;\mu_T)\right] \\ &= b_k(\mu_T)^2\left[p'_{k+1}(0;\mu_T)p_k(0;\mu_T) - p'_k(0;\mu_T)p_{k+1}(0;\mu_T)\right].\end{aligned}$$

Then using (4.1)

$$p_k(\lambda;\mu_T) = \left(\prod_{j=0}^{k-1} b_j(\mu_T)^{-1}\right)\pi_k(\lambda;\mu_T)$$

we find the alternate expression

$$\begin{aligned}\sum_{j=0}^{k} p_j^2(0;\mu_T) = &\left(\prod_{j=0}^{k-1} b_j(\mu_T)^{-2}\right) \\ &\cdot\left[\pi'_{k+1}(0;\mu_T)\pi_k(0;\mu_T) - \pi'_k(0;\mu_T)\pi_{k+1}(0;\mu_T)\right]. \quad \square\end{aligned}$$

4.3 The CGA on the normal equations

Next, for $X \in \mathbb{C}^{N\times M}$, $N \le M$, consider solving the normal equations $XX^*\boldsymbol{x} = X\mathbf{b}$ with the CGA. The appearance of X on the right-hand side changes the minimization problem one has to consider. With $W = XX^*$, the CGA will solve

$$\boldsymbol{x}_k = \operatorname{argmin}_{\boldsymbol{y}\in\mathcal{K}_k}\|\boldsymbol{x} - \boldsymbol{y}\|_W, \quad \mathcal{K}_k = \{X\mathbf{b}, WX\mathbf{b}, \ldots, W^{k-1}X\mathbf{b}\}.$$

As before, we express

$$\boldsymbol{x} - \boldsymbol{y} = W^{-1}q_{\boldsymbol{y}}(W)X\mathbf{b}.$$

Using the singular value decomposition $X = U\Sigma V^*$ where U, V are square matrices, we write

$$\begin{aligned}\|\boldsymbol{x} - \boldsymbol{y}\|_W^2 &= \mathbf{b}^* V \Sigma^* U^* q_{\boldsymbol{y}}(W)^* W^{-1} q_{\boldsymbol{y}}(W) U \Sigma V^* \mathbf{b}, \\ &= \mathbf{b}^* V \Sigma^* q_{\boldsymbol{y}}(\Lambda)^* \Lambda^{-1} q_{\boldsymbol{y}}(\Lambda) \Sigma V^* \mathbf{b}\end{aligned}$$

where $\Lambda = \Sigma\Sigma^*$. Since Σ has its last $M - N$ columns being identically zero, we use the notation $\Sigma = \begin{bmatrix}\Sigma_0 & 0\end{bmatrix}$ and find $\Lambda = \Sigma_0^2$. Thus

$$\|\boldsymbol{x} - \boldsymbol{y}\|_W^2 = \boldsymbol{c}^* \Sigma_0 q_{\boldsymbol{y}}(\Lambda)^* \Lambda^{-1} q_{\boldsymbol{y}}(\Lambda) \Sigma_0 \boldsymbol{c}, \quad \boldsymbol{c} = \begin{bmatrix} I & 0\end{bmatrix} V^* \mathbf{b}.$$

The techniques used in the case of MINRES directly apply.

PROPOSITION 4.3. *Let $\boldsymbol{x}_k$ be the computed solution at step k of applying the CGA to the normal equations $XX^*\boldsymbol{x} = X\mathbf{b}$, $X \in \mathbb{C}^{N\times M}$, $N \le M$. For any $k \in \mathbb{N}$,*

$$\|\mathbf{e}_k\|_W^2 = \frac{\prod_{j=0}^{k-1} b_j(\nu)^2}{\pi'_{k+1}(0;\nu)\pi_k(0;\nu) - \pi'_k(0;\nu)\pi_{k+1}(0;\nu)} = \frac{1}{\sum_{j=0}^{k} p_j^2(0;\nu)},$$

where

$$\nu = \sum_{j=1}^{N} \omega_j \delta_{\lambda_j}, \quad \omega_j = |(V^*\mathbf{b})_j|^2, \tag{4.3}$$

$X = U\Sigma V^*$ is the singular value decomposition of X and $\lambda_1, \dots, \lambda_N$ are the eigenvalues of XX^*.

5 Universality

5.1 Bidiagonal central limit theorem, Gaussian case

Throughout the asymptotic analysis that follows d will be a fixed positive real number and $\mathfrak{d} = N/M \xrightarrow{M\to\infty} d$. Taking the entrywise limit in (2.3), using that the final $M - N$ columns of X_{GK} are zero and the notation

$$\frac{1}{\sqrt{\beta M}} X_{\mathrm{GK}} = \begin{bmatrix} H & 0\end{bmatrix},$$

$$H \overset{\mathscr{L}}{=} \frac{1}{\beta M} \begin{bmatrix} \chi_{\beta M} & & & & & \\ \chi_{\beta(N-1)} & \chi_{\beta(M-1)} & & & & \\ & \chi_{\beta(N-2)} & \chi_{\beta(M-2)} & & & \\ & & \ddots & \ddots & & \\ & & & \chi_{\beta} & \chi_{\beta(M-N+1)} \end{bmatrix}, \tag{5.1}$$

it follows that

$$H \xrightarrow[N\to\infty]{(d)} \mathbb{H}_d = \begin{bmatrix} 1 & & & \\ \sqrt{d} & 1 & & \\ & \sqrt{d} & 1 & \\ & & \ddots & \ddots \end{bmatrix}.$$

This limit is in the sense of weak convergence of the finite-dimensional marginals of a random infinite bidiagonal matrix.

Furthermore, for a χ_k random variable

$$\chi_k - \sqrt{k} \xrightarrow[k\to\infty]{(d)} \mathcal{N}_1(0, 1/2),$$

and so by independence, for iid standard normals $\{Z_j\}_1^\infty$,

$$\sqrt{2\beta M}\,(H - \mathbb{H}_\mathfrak{d}) \xrightarrow[N\to\infty]{(d)} \mathbb{G} = \begin{bmatrix} Z_1 & & & \\ Z_2 & Z_3 & & \\ & Z_4 & Z_5 & \\ & & \ddots & \ddots \end{bmatrix}. \tag{5.2}$$

From here, it follows immediately that the Jacobi matrix produced by the Lanczos algorithm applied to $\mathcal{W}_\beta(N, M)$ has a limit, in the same sense of finite-dimensional marginal convergence, to an infinite tridiagonal matrix.

DEFINITION 5.1. Given a positive-definite Jacobi matrix T we define φ to be the function that gives the Cholesky factorization of T. That is, $\varphi(T) = H$ where H is a lower-triangular bidiagonal matrix with all nonnegative entries and $HH^* = T$.

The Cholesky factorization $\varphi(T)$ is unique for $T > 0$ and φ is differentiable (see [20]). The actual algorithm to compute it is given as follows:

Algorithm 4: Jacobi matrix Cholesky factorization

(1) Suppose T is an $N \times N$ positive-definite Jacobi matrix, set $H = T$.
(2) For $k = 1, 2, \ldots, N-1$
   (a) Set $H_{k+1,k+1} = H_{k+1,k+1} - \dfrac{H_{k+1,k}^2}{H_{kk}}$.
   (b) Set $H_{k:k+1,k} = H_{k:k+1,k}/\sqrt{H_{k,k}}$.
(3) Set $H_{N,N} = \sqrt{H_{N,N}}$.
(4) Return $\varphi(T) = H$.

The following is immediate.

PROPOSITION 5.2. *Let* $W \overset{\mathcal{L}}{=} \mathcal{W}_\beta(N, M)$, $N \le M$. *For any sequence of unit vectors* $\mathbf{b} = \mathbf{b}_N$ *of length* N,

$$\sqrt{2\beta M}\,(\varphi(T(W, \mathbf{b}_N)) - \mathbb{H}_\mathfrak{d}) \xrightarrow[M\to\infty]{(d)} \mathbb{G}.$$

Now, define

$$\mathbb{H}_d \mathbb{H}_d^* = \mathbb{T}_d := \begin{bmatrix} 1 & \sqrt{d} & & \\ \sqrt{d} & 1+d & \sqrt{d} & \\ & \sqrt{d} & 1+d & \ddots \\ & & \ddots & \ddots \end{bmatrix}.$$

PROPOSITION 5.3. *Let $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N, M)$ for $N \leq M$ where $\mathfrak{d} \xrightarrow{M\to\infty} d \in (0, 1]$. Then for any sequence of unit vectors $\mathbf{b} = \mathbf{b}_N$ of length N, with $T = T(W, \mathbf{b})$, the vector*

$$\left(\sqrt{\beta M}\, \boldsymbol{f}_1^*(T^k - \mathbb{T}_{\mathfrak{d}}^k)\boldsymbol{f}_1\right)_{k\geq 1} = \left(\sqrt{\beta M}\int_{\mathbb{R}} x^k(\mu_T(\mathrm{d}x) - \varrho_{\mathfrak{d}}(\mathrm{d}x))\right)_{k\geq 1},$$

converges in the sense of finite-dimensional marginals to a centered Gaussian random vector $\mathcal{G} = (G_1)_{k\geq 1}$.

PROOF. The equality follows using (3.3). The proposition then follows using (5.2) because, for each k, $\sqrt{M}\, \boldsymbol{f}_1^*(T^k - \mathbb{T}_{\mathfrak{d}}^k)\boldsymbol{f}_1$ depends only on a finite number of elements of T. □

5.2 Contour integral reformulation of the moments

Let Γ be a simple curve that encloses the nonzero spectrum of a symmetric tridiagonal matrix T. Then

$$m_k(\mu_T) = \frac{1}{2\pi i}\oint_\Gamma z^k c_0(z; \mu_T)\mathrm{d}z.$$

Now, let $\Gamma = \Gamma_d$ be a smooth simple contour that properly encloses the support of the Marchenko–Pastur law (2.2).

We denote the Stieltjes transform $s_d(z)$ of (2.2) by

$$s_d(z) = \int_{\mathbb{R}} \frac{\varrho_d(\mathrm{d}\lambda)}{\lambda - z}. \tag{5.3}$$

There are many classical references for the following two results.

THEOREM 5.4 (Global eigenvalue bounds; see, e.g., [6,22,38,43]). *For the eigenvalues $\lambda_N \leq \cdots \leq \lambda_1$ of $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N, M)$, $N \leq M$, and $t > 0$*

$$\mathbb{P}\left(1 - \sqrt{\frac{N}{M}} - t \leq \lambda_N^{1/2} \leq \lambda_1^{1/2} \leq 1 + \sqrt{\frac{N}{M}} + t\right) \xrightarrow{M\to\infty} 1.$$

THEOREM 5.5 (Global eigenvalue bounds; see, e.g., [2,45]). *For the eigenvalues $\lambda_N \leq \cdots \leq \lambda_1$ of a sample covariance matrix W and $t > 0$. Then if $\mathfrak{d} \xrightarrow{M\to\infty} d \in (0, 1)$,*

$$\mathbb{P}\left(1 - \sqrt{d} - t \leq \lambda_N^{1/2} \leq \lambda_1^{1/2} \leq 1 + \sqrt{d} + t\right) \xrightarrow{M\to\infty} 1.$$

Hence with probability tending to 1 as $M \to \infty$, $\mathfrak{d} \xrightarrow{M\to\infty} d$, the support of μ_T, $T = T(W, \mathbf{b})$ is contained within Γ_d. As a corollary, we have:

COROLLARY 5.6. *Let* $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N, M)$ *for* $N \le M$ *where* $\mathfrak{d} \xrightarrow{M\to\infty} d \in (0, 1]$. *Then for any sequence of unit vectors* $\mathbf{b} = \mathbf{b}_N$ *of dimension* N, *with* $T = T(W, \mathbf{b})$, *the vector*

$$\left(\frac{\sqrt{M}}{2\pi i}\oint_{\Gamma_d} z^k (c_0(z; \mu_T) - s_{\mathfrak{d}}(z))\mathrm{d}z\right)_{k\ge 1} \xrightarrow[N\to\infty]{(d)} \mathcal{G},$$

in the sense of finite-dimensional marginals, where $\mathcal{G}$ *is the same process as in Proposition* 5.3.

We also need to treat the case of $k = -1$. Suppose $T = HH^T$ where H is real, square, and lower-triangular and given by

$$H = \begin{bmatrix} \alpha_0 & & & \\ \beta_0 & \alpha_1 & & \\ & \beta_1 & \alpha_2 & \\ & & \ddots & \ddots \end{bmatrix} \tag{5.4}$$

Then $T_{11} = \alpha_0^2$ and $T_{jj} = \alpha_{j-1}^2 + \beta_{j-2}^2$ for $j > 1$. Let $\widetilde{H}$ be the matrix formed by removing the first row and column of H and let $\widetilde{T} = \widetilde{H}\widetilde{H}^T$. Then it follows by Cramer's rule that

$$\begin{aligned} \boldsymbol{f}_1^* T^{-1} \boldsymbol{f}_1 = \frac{\det(\beta_0^2 \boldsymbol{f}_1 \boldsymbol{f}_1^* + \widetilde{T})}{\det T} &= \frac{(\det \widetilde{T})(1 + \beta_0^2 \boldsymbol{f}_1^* \widetilde{T}^{-1} \boldsymbol{f}_1)}{\det T} \\ &= \frac{1}{\alpha_0^2}(1 + \beta_0^2 \boldsymbol{f}_1^* \widetilde{T}^{-1} \boldsymbol{f}_1). \end{aligned} \tag{5.5}$$

From this expression, one obtains

$$\boldsymbol{f}_1^* T^{-1} \boldsymbol{f}_1 = \frac{1}{\alpha_0^2}\left(1 + \sum_{j=1}^{N-1} \prod_{k=1}^{j} \left(\frac{\beta_{k-1}^2}{\alpha_k^2}\right)\right).$$

Following [30, theorem 3.2.12]:

PROPOSITION 5.7. *Let* $\boldsymbol{y}$ *be random vector in* $\mathbb{C}^N$ *that does not vanish a.s. Let* X *be an* $N \times M$ *matrix with independent* $\mathcal{N}_\beta(0, 1)$ *entries independent of* $\boldsymbol{y}$. *Then*

$$\frac{\boldsymbol{y}^* \boldsymbol{y}}{\boldsymbol{y}(XX^*)^{-1}\boldsymbol{y}} \overset{\mathscr{L}}{=} \beta^{-1} \chi^2_{\beta(M-N+1)}$$

and therefore

$$\boldsymbol{f}_1^* T^{-1} \boldsymbol{f}_1 \overset{\mathscr{L}}{=} \frac{\beta M}{\chi^2_{\beta(M-N+1)}}.$$

PROOF. The first claim can be established using the QR factorization of X. The second claim for $\boldsymbol{f}_1^* T^{-1} \boldsymbol{f}_1$ follows from the first once we realize

$$T = T(M^{-1} XX^*, \mathbf{b}),$$

then $\boldsymbol{f}_1^* T^{-1} \boldsymbol{f}_1 = M \mathbf{b}^* (XX^*)^{-1} \mathbf{b}$. □

We can also apply the same proposition to an $(N-1) \times (M-1)$ matrix of normals, and conclude

$$\boldsymbol{f}_1^* \widetilde{T}^{-1} \boldsymbol{f}_1 \overset{\mathscr{L}}{=} \frac{\beta M}{\chi^2_{\beta(M-N+1)}}.$$

Using (5.5) this provides a remarkable identity in law involving chi-square distributions:

PROPOSITION 5.8. *For any integers $\ell \geq 0$ and $M \geq N \geq 1$*

$$\frac{\beta M}{\chi^2_{\beta(M-N+1)}} \overset{\mathscr{L}}{=} \frac{1}{\chi^2_{\beta(M-\ell)}} \left(1 + \chi^2_{\beta(N-\ell-1)} \frac{\beta M}{\chi^2_{\beta(M-N+1)}} \right)$$

where the chi-squared variables on the right-hand side are mutually independent.

But more importantly, iterating (5.5) ℓ times and applying and using Proposition 5.7 to describe the remainder, we have:

PROPOSITION 5.9. *Suppose H is distributed as in* (5.1)*. Then for $0 < \ell < N$*

$$\boldsymbol{f}_1^* T^{-1} \boldsymbol{f}_1 = \frac{1}{\alpha_0^2} \left(1 + \sum_{j=1}^{\ell} \prod_{k=1}^{j} \frac{\beta_{k-1}^2}{\alpha_k^2} + \left(\beta_\ell^2 \frac{\beta M}{\chi^2_{\beta(M-N+1)}} \right) \prod_{k=1}^{\ell} \frac{\beta_{k-1}^2}{\alpha_k^2} \right)$$

where $\chi_{\beta(M-N+1)}$ depends only on $H_{\ell+1:N,\ell+1:N}$.

The following notation is convenient.

DEFINITION 5.10. We write $X_M = c_M + Y_M + o(M^{-1/2})$ if

$$\sqrt{M}(X_M - c_M) \quad \text{and} \quad \sqrt{M}(Y_M),$$

converge, in distribution, to the same distribution as $M \to \infty$.

Let ℓ be fixed. We use the approximation in distribution (5.2),

$$\alpha_j = 1 + Z_{2j+1}/\sqrt{2\beta M} + o(M^{-1/2}),$$
$$\beta_j = \sqrt{\mathfrak{d}} + Z_{2j+2}/\sqrt{2\beta M} + o(M^{-1/2})$$

to find

$$\alpha_j^2 = 1 + \frac{\sqrt{2}}{\sqrt{\beta M}} Z_{2j+1} + o(M^{-1/2}), \tag{5.6}$$

$$\beta_j^2 = \mathfrak{d} + \frac{\sqrt{2\mathfrak{d}}}{\sqrt{\beta M}} Z_{2j+2} + o(M^{-1/2}), \tag{5.7}$$

$$\frac{\beta M}{\chi^2_{\beta(M-N+1)}} = \frac{1}{1-\mathfrak{d}}\left(1 + \frac{\sqrt{2}}{\sqrt{1-\mathfrak{d}}\sqrt{\beta M}} Z_0\right) + o(M^{-1/2}), \quad Z_0 \overset{\mathscr{L}}{=} \mathcal{N}_1(0,1),$$

and compute as $N \to \infty$

$$\begin{aligned}
&1 + \sum_{j=1}^{\ell} \prod_{k=1}^{j} \left(\frac{\beta^2_{k-1}}{\alpha^2_k}\right) \\
&\quad = 1 + \sum_{j=1}^{\ell} \prod_{k=1}^{j} \mathfrak{d}\left[1 + \frac{\sqrt{2}}{\sqrt{\beta M}}\left(Z_{2k}/\sqrt{d} - Z_{2k+1}\right)\right] \\
&\quad = \frac{1-\mathfrak{d}^{\ell+1}}{1-\mathfrak{d}} + \frac{\sqrt{2}}{\sqrt{\beta M}} \sum_{k=1}^{\ell} \mathfrak{d}^k \frac{1-\mathfrak{d}^{\ell-k+1}}{1-\mathfrak{d}}\left(Z_{2k}/\sqrt{\mathfrak{d}} - Z_{2k+1}\right) + o(M^{-1/2}).
\end{aligned}$$

Thus

$$\begin{aligned}
&\sqrt{\beta M}\left(\boldsymbol{f}_1^* T^{-1} \boldsymbol{f}_1 - \frac{1}{1-\mathfrak{d}}\right) \xrightarrow[M\to\infty]{(d)} \\
&\qquad \sqrt{2}\frac{Z_1}{d-1} + \sqrt{2}\sum_{k=1}^{\infty} \frac{d^k}{1-d}\left(Z_{2k}/\sqrt{d} - Z_{2k+1}\right).
\end{aligned}$$

We arrive at the following proposition.

PROPOSITION 5.11. *Suppose H is distributed as in* (5.1) *where the entries are labeled according to* (5.4) *and $T = HH^*$. Let $\mathcal{Z} = [Z_1, Z_2, \ldots]^T$ be a vector of iid standard normal random variables. Then if $N \le M$, $\mathfrak{d} \xrightarrow{M\to\infty} d \in (0,1)$,*

$$\begin{aligned}
&\sqrt{\beta M}\left(\boldsymbol{f}_1^* T^{-1} \boldsymbol{f}_1 - \frac{1}{1-\mathfrak{d}}\right) \xrightarrow[M\to\infty]{(d)} Z_{-1}, \\
&Z_{-1} := -\sqrt{2}\frac{Z_1}{1-d} + \sqrt{2}\sum_{k=1}^{\infty} \frac{d^k}{1-d}\left(Z_{2k}/\sqrt{d} - Z_{2k+1}\right).
\end{aligned}$$

Additionally,

$$\sqrt{\beta M}\left(\begin{bmatrix} \boldsymbol{f}_1^* T^{-1} \boldsymbol{f}_1 \\ \alpha_0 \\ \beta_0 \\ \alpha_1 \\ \beta_1 \\ \vdots \end{bmatrix} - \begin{bmatrix} \frac{1}{1-\mathfrak{d}} \\ 1 \\ \sqrt{\mathfrak{d}} \\ 1 \\ \sqrt{\mathfrak{d}} \\ \vdots \end{bmatrix}\right) \xrightarrow[M\to\infty]{(d)} \begin{bmatrix} Z_{-1} \\ \mathcal{Z}/\sqrt{2} \end{bmatrix}$$

in the sense of convergence of finite-dimensional marginals.

5.3 Universality for the moment fluctuations of the spectral measure

We now generalize Corollary 5.6 to general distributions. Let $R(z) = R(z; X) = (XX^* - zI)^{-1}$ denote the resolvent of XX^* and define

$$G(z) = G(z; X) = \begin{bmatrix} -I & X^* \\ X & -zI \end{bmatrix}^{-1}.$$

The following is a direct consequence of [26, theorems 3.6 and 3.7].

PROPOSITION 5.12. *Suppose X is a sample covariance matrix with*

$$\mathfrak{d} = N/M \xrightarrow{M\to\infty} d \in (0,1) \cup (1,\infty).$$

For any $\delta, \epsilon > 0$ and for any $R, D > 0$ there is a constant C so that for all $M \in \mathbb{N}$,

$$\sup_{z\in\Gamma} \sup_{\boldsymbol{v},\boldsymbol{w}\in\mathbb{C}^{N+M}} \mathbb{P}\left[\left|\boldsymbol{v}^* G(z)\boldsymbol{w} - \boldsymbol{v}^*\Pi_{\mathfrak{d}}(z)\boldsymbol{w}\right| \geq \|\boldsymbol{v}\|\|\boldsymbol{w}\| M^{\epsilon-1/2}\right] \leq CM^{-D},$$

$$\Pi_{\mathfrak{d}}(z) = \begin{bmatrix} -(1+s_{\mathfrak{d}}(z))^{-1} I_M & 0 \\ 0 & s_{\mathfrak{d}}(z) I_N \end{bmatrix},$$

and therefore

$$\sup_{z\in\Gamma} \sup_{\boldsymbol{v},\boldsymbol{w}\in\mathbb{C}^{N}} \mathbb{P}\left[\left|\boldsymbol{v}^* R(z)\boldsymbol{w} - s_{\mathfrak{d}}(z)\boldsymbol{v}^*\boldsymbol{w}\right| \geq \|\boldsymbol{v}\|\|\boldsymbol{w}\| M^{\epsilon-1/2}\right] \leq CM^{-D},$$

where Γ is any bounded simple closed curve that does not intersect the support of ϱ_d.

Remark 5.13. For our results, we will need an analogous result to Proposition 5.12 for $d = 1$. One such result of this type is found in [15, lemma 6.1].

DEFINITION 5.14. Let $\Phi : \mathbb{C}^n \to \mathbb{R}$ be bounded. Suppose, in addition, that for any multi-index $\alpha = (\alpha_1, \ldots, \alpha_n)$, $1 \leq |\alpha| \leq 5$,and for any $\epsilon' > 0$ sufficiently small, we have

$$\max\{|\partial^\alpha \Phi(x_1, \ldots, x_n)| : \max_j |x_j| \leq M^{\epsilon'}\} \leq M^{C_0\epsilon'},$$

for $C_0 > 0$. Then Φ is called an admissible test function.

THEOREM 5.15 (Comparison). *Let $W = XX^*$ and $\widetilde{W} = YY^*$ be two sample covariance matrices such that*

$$\mathbb{E}(\Re X_{ij})^\ell (\Im X_{ij})^p = \mathbb{E}(\Re Y_{ij})^\ell (\Im Y_{ij})^p, \quad \ell + p \leq 4,\ 1 \leq i \leq N,\ 1 \leq j \leq M.$$

For each j, let $\Gamma_j = \partial\Omega_j$, $\Omega_j = \overline{\Omega}_j$ be a simple, smooth, positively oriented curve that is uniformly bounded away from support of the Marchenko–Pastur law ϱ_d. Suppose that $f_1, f_2, \ldots, f_n$ is a finite collection of functions that are analytic

in a neighborhood of Ω. *Then for any admissible test function* $\Phi : \mathbb{C}^n \to \mathbb{R}$ *we have for* $T = T(W, \mathbf{b})$, $\widetilde{T} = T(\widetilde{W}, \mathbf{b})$,

$$\left| \mathbb{E}\Phi\left(\frac{\sqrt{M}}{2\pi i} \oint_{\Gamma_1} f_1(z)(c_0(z;\mu_T) - s_{\mathfrak{d}}(z))\mathrm{d}z, \ldots, \frac{\sqrt{M}}{2\pi i} \oint_{\Gamma_n} f_n(z)(c_0(z;\mu_T) - s_{\mathfrak{d}}(z))\mathrm{d}z \right) \right.$$
$$\left. - \mathbb{E}\Phi\left(\frac{\sqrt{M}}{2\pi i} \oint_{\Gamma_1} f_1(z)(c_0(z;\mu_{\widetilde{T}}) - s_{\mathfrak{d}}(z))\mathrm{d}z, \ldots, \frac{\sqrt{M}}{2\pi i} \oint_{\Gamma_n} f_n(z)(c_0(z;\mu_{\widetilde{T}}) - s_{\mathfrak{d}}(z))\mathrm{d}z \right) \right|$$
$$\leq CM^{-\sigma}$$

for some $C, \sigma > 0$. *Here* C *will depend on* n, *the constants* C_p *in Definition* 2.3, Φ, $\Gamma_1, \ldots, \Gamma_n$, *and* $f_1, \ldots, f_n$, *and* σ *will depend on the constant* C_0 *in Definition* 5.14.

Remark 5.16. Note that in Theorem 5.15, if $\mathfrak{d}$ is bounded uniformly away from 1, a contour Γ_j could just encircle $z = 0$. And if $\mathfrak{d} \to d \in (0, 1]$, the only nontrivial case is where the contour Γ_j encircles the entire support of ϱ_d.

This gives immediate corollaries.

COROLLARY 5.17. *Suppose* W *is a sample covariance matrix satisfying the moment matching condition* (*Definition* 2.4) *with*

$$\mathfrak{d} = N/M \xrightarrow{M\to\infty} d \in (0, \infty).$$

Then for any sequence of unit vectors $\mathbf{b} = \mathbf{b}_N$ *of length* N, *with* $T = T(W, \mathbf{b})$, *the vector*

$$\left(\frac{\sqrt{M}}{2\pi i} \oint_{\Gamma} z^k (c_0(z;\mu_T) - s_{\mathfrak{d}}(z))\mathrm{d}z \right)_{k\geq 1} \xrightarrow[N\to\infty]{(d)} \mathcal{G},$$

in the sense of finite-dimensional marginals, where $\mathcal{G}$ *is the same process as in Proposition* 5.3.

COROLLARY 5.18. *Suppose* W *is a sample covariance matrix satisfying the moment matching condition* (*Definition* 2.4) *with*

$$\mathfrak{d} = N/M \xrightarrow{M\to\infty} d \in (0, \infty).$$

Then for any sequence of unit vectors $\mathbf{b} = \mathbf{b}_N$ *of length* N, *with* $T = T(W, \mathbf{b})$, *let* H *be given by the Cholesky factorization of* T, $H = \varphi(T)$, *and label the entries of* H *as in* (5.4). *Then Proposition* 5.11 *holds for* H.

PROOF. Fix k. For all $N > k$, the Hankel matrix of moments

$$(m_{j+r-2}(\mu_T))_{j,r=1}^{k}$$

is positive definite almost surely. On this set, the mapping to $(m_j(\mu_T) : 0 \leq j \leq 2k) \mapsto T_k(W, \mathbf{b})$ is differentiable. It follows that H_k, the upper-left $k \times k$ subblock of H, is also a differentiable function $m_j(\mu_T)$, $j = 0, 1, \ldots, 2k$. Then the corollary follows directly from Theorem 5.15. □

Before we prove Theorem 5.15, we establish some intermediate results.

LEMMA 5.19. *For an $N \times M$ matrix X and $\Im z \neq 0$*

$$\begin{bmatrix} -I & X^* \\ X & -zI \end{bmatrix}^{-1} = \begin{bmatrix} (z^{-1}XX^* - I)^{-1} & (X^*X - zI)^{-1}X^* \\ X(X^*X - zI)^{-1} & (XX^* - zI)^{-1}, \end{bmatrix} \tag{5.8}$$

$$\left\| \begin{bmatrix} -I & X^* \\ X & -zI \end{bmatrix}^{-1} \right\| \leq (|z|+1)|\Im z|^{-1} + 2\sqrt{|\Im z|^{-1} + |z||\Im z|^{-2}} \tag{5.9}$$

Recall that $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots$ denotes the standard basis and we use the notation

$$\widehat{\boldsymbol{u}} = \begin{bmatrix} 0 \\ \boldsymbol{u} \end{bmatrix} \in \mathbb{C}^{N+M}$$

for $\boldsymbol{u} \in \mathbb{C}^N$.

LEMMA 5.20 (Resolvent expansion with leading-order correction). *Let X be an iid matrix satisfying the assumptions of Definition 2.3. Let Q be the matrix that is equal to X with the exception of one entry that is set to 0 so that $X = Q + X_{ij}\boldsymbol{f}_i\boldsymbol{f}_j^*$ for some $1 \leq i \leq N$, $1 \leq j \leq M$. For two unit vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^N$*

$$\widehat{\boldsymbol{u}}^* S(z;X)\widehat{\boldsymbol{v}} = \widehat{\boldsymbol{u}}^* S(z;Q)\widehat{\boldsymbol{v}} + \sum_{k=1}^{3} M^{-k/2} J_k + M^{-5/2} J_4,$$

$$S(z;X) = \sqrt{M}\left[G(z;X) - \Pi_d(z)\right],$$

and for every $\epsilon > 0$ and $D > 0$ there exists $C > 0$ such that J_4 satisfies

$$\mathbb{P}(|J_4| > M^{\epsilon}) \leq CM^{-D}.$$

In addition, J_k for $k < 4$ is a finite sum of the form

$$J_k = \sum_{\ell} f_{k,\ell} g_{k,\ell},$$

where $g_{k,\ell}$ is a monomial in $X_{ij}\sqrt{M}$ and $\overline{X_{ij}}\sqrt{M}$ with degree at most $k+1$ and $f_{k,\ell}$ is independent of X_{ij} satisfying that for every $\epsilon > 0$ and $D > 0$ there exists $C > 0$ such that

$$\mathbb{P}(|f_{k,\ell}| > M^{\epsilon}) \leq CM^{-D}.$$

PROOF. Write $V := X_{ij}\boldsymbol{f}_{i+M}\boldsymbol{f}_j^* + \overline{X_{ij}}\boldsymbol{f}_j\boldsymbol{f}_{i+M}^*$. Consider for a diagonal matrix D

$$\begin{aligned} &\widehat{\boldsymbol{u}}^* DVD\widehat{\boldsymbol{v}} \\ &\quad = X_{ij}(\widehat{\boldsymbol{u}}^* D\boldsymbol{f}_{i+M})(\boldsymbol{f}_j^* D\widehat{\boldsymbol{v}}) + \overline{X_{ij}}(\widehat{\boldsymbol{u}}^* D\boldsymbol{f}_j)(\boldsymbol{f}_{i+M}^* D\widehat{\boldsymbol{v}}) \\ &\quad = X_{ij}(\widehat{\boldsymbol{u}}^* D_{i+M,i+M}\boldsymbol{f}_{i+M})(\boldsymbol{f}_j^* D_{jj}\widehat{\boldsymbol{v}}) \\ &\qquad + \overline{X_{ij}}(\widehat{\boldsymbol{u}}^* D_{jj}\boldsymbol{f}_j)(\boldsymbol{f}_{i+M}^* D_{i+M,i+M}\widehat{\boldsymbol{v}}) = 0. \end{aligned} \tag{5.10}$$

This is because $1 \leq j \leq M$ and $\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}$ must have zeros in their first M entries.

We then consider the expansion of

$$S(z;X) = S(z;Q) - \sqrt{M}G(z;Q)VG(z;Q) + \cdots + \sqrt{M}[G(z;Q)V]^4G(z;Q) - \sqrt{M}[G(z;Q)V]^5G(z;X).$$

We write

$$\hat{\boldsymbol{u}}^*\sqrt{M}G(z;Q)VG(z;Q)\hat{\boldsymbol{v}} = \sqrt{M}\hat{\boldsymbol{u}}^*\Pi_d(z)V\Pi_d(z)\hat{\boldsymbol{v}} + E_{ij}.$$

From (5.10) the first term vanishes. Explicitly,

$$\begin{aligned}
E_{ij} &= \hat{\boldsymbol{u}}^*\Big[S(z;Q)VG(z;Q) + G(z;Q)VS(z;Q) + M^{-1/2}S(z;Q)VS(z;Q)\Big]\hat{\boldsymbol{v}},\\
&= X_{ij}(\hat{\boldsymbol{u}}^*S(z;Q)\boldsymbol{f}_{j+M})(\boldsymbol{f}_i^*G(z,Q)\hat{\boldsymbol{v}})\\
&\quad + \overline{X_{ij}}(\hat{\boldsymbol{u}}^*G(z;Q)\boldsymbol{f}_i)(\boldsymbol{f}_{j+M}^*S(z,Q)\hat{\boldsymbol{v}})\\
&\quad + X_{ij}(\hat{\boldsymbol{u}}^*G(z;Q)\boldsymbol{f}_{j+M})(\boldsymbol{f}_i^*S(z,Q)\hat{\boldsymbol{v}}) + \overline{X_{ij}}(\hat{\boldsymbol{u}}^*S(z;Q)\boldsymbol{f}_i)(\boldsymbol{f}_{j+M}^*G(z,Q)\hat{\boldsymbol{v}})\\
&\quad + M^{-1/2}X_{ij}(\hat{\boldsymbol{u}}^*S(z;Q)\boldsymbol{f}_{j+M})(\boldsymbol{f}_i^*S(z,Q)\hat{\boldsymbol{v}})\\
&\quad + M^{-1/2}\overline{X_{ij}}(\hat{\boldsymbol{u}}^*S(z;Q)\boldsymbol{f}_i)(\boldsymbol{f}_{j+M}^*S(z,Q)\hat{\boldsymbol{v}}).
\end{aligned}$$

Observe that this is a linear function of $X_{ij}, \overline{X_{ij}}$ with coefficients that are independent of X_{ij} and controlled by Proposition 5.12.

Then consider

$$\hat{\boldsymbol{u}}^*(G(z;Q)V)^jG(z;Q)\hat{\boldsymbol{v}}.$$

With the notation $a_1 = X_{ij}$, $a_2 = \overline{X_{ij}}$, $\boldsymbol{v}_1 = \boldsymbol{f}_{j+M}$, $\boldsymbol{v}_2 = \boldsymbol{f}_i$, and $\boldsymbol{w}_1 = \boldsymbol{f}_i$, $\boldsymbol{w}_2 = \boldsymbol{f}_{j+M}$, one has for $\ell = 2, 3, 4$

$$\begin{aligned}
&\sqrt{M}\hat{\boldsymbol{u}}^*(G(z;Q)V)^\ell G(z;Q)\hat{\boldsymbol{v}}\\
&\quad = \sqrt{M}\sum_{p\in\{1,2\}^\ell}\Bigg[\Bigg(\prod_{k=1}^{\ell}a_{p_k}\Bigg)(\hat{\boldsymbol{u}}^*G(z;Q)\boldsymbol{v}_{p_1})(\boldsymbol{w}_{p_\ell}^*G(z;Q)\hat{\boldsymbol{v}})\\
&\qquad\qquad \cdot\prod_{k=1}^{\ell-1}(\boldsymbol{w}_{p_k}^*G(z;Q)\boldsymbol{v}_{p_{k+1}})\Bigg] := P_{ij}^{(\ell)}
\end{aligned}$$

and set

$$\begin{aligned}
P_{ij}^{(5)} := \sqrt{M}\sum_{p\in\{1,2\}^5}\Bigg[&\Bigg(\prod_{k=1}^{5}a_{p_k}\Bigg)(\hat{\boldsymbol{u}}^*G(z;Q)\boldsymbol{v}_{p_1})(\boldsymbol{w}_{p_\ell}^*G(z;X)\hat{\boldsymbol{v}})\\
&\cdot\prod_{k=1}^{4}(\boldsymbol{w}_{p_k}^*G(z;Q)\boldsymbol{v}_{p_{k+1}})\Bigg].
\end{aligned}$$

Whenever two vectors are orthogonal because they have disjoint support, we can replace $G(z)$ with $S(z)/\sqrt{M}$. When ℓ is odd, suppose that for a choice of $p \in \{1,2\}^\ell$ no two vectors are orthogonal in such a way. Then $p_1 = 1$ so that $\hat{\boldsymbol{u}}$ is

not orthogonal to $\boldsymbol{v}_{p_1}$. And then $\boldsymbol{w}_i$ and $\boldsymbol{v}_j$ are not orthogonal if $i \neq j$, so then $p_2 = 2$, $p_3 = 1$, and so on. This implies that $p_\ell = 1$ because ℓ is odd. But then $\widehat{\boldsymbol{v}}$ is orthogonal to $\boldsymbol{v}_{p_\ell}$. This implies that the order of the odd terms is actually one less than is immediately apparent. Write

$$\widehat{\boldsymbol{u}}^* S(z;X)\widehat{\boldsymbol{v}} = \widehat{\boldsymbol{u}}^* S(z;Q)\widehat{\boldsymbol{v}} + \sum_{k=1}^{3} M^{-k/2} J_k + M^{-5/2} J_4 = \widehat{\boldsymbol{u}}^* S(z;Q)\widehat{\boldsymbol{v}} + \xi,$$

$$J_1 = M^{1/2}(E_{ij} + P_{ij}^{(2)}), \quad J_2 = M P_{ij}^{(3)},$$
$$J_3 = M^{3/2} P_{ij}^{(4)}, \qquad J_4 = M^{5/2} P_{ij}^{(5)}.$$

□

PROPOSITION 5.21 (Green's function replacement). *Suppose* Φ *is an admissible test function. Suppose further that* X *and* Y *are two matrices satisfying assumptions in Definition* 2.3 *and that*

$$\mathbb{E} X_{ij}^{\ell} \overline{X_{ij}}^{p} = \mathbb{E} Y_{ij}^{\ell} \overline{Y_{ij}}^{p},$$

for all choices of $\ell, p \in \mathbb{N}$, $\ell + p \leq 4$, *and* $1 \leq i \leq N$, $1 \leq j \leq M$. *Then for any* $\epsilon > 0$, *any families of unit vectors* $\{\boldsymbol{q}_j\}_{j=1}^n$, $\{\boldsymbol{p}_j\}_{j=1}^n$, *and any collection of points* $\{z_j\}_{j=1}^n$ *bounded uniformly away from the support of the Marchenko–Pastur law* ϱ_d *and bounded away from the real axis by* $M^{-\delta}$, $1 > \delta > 0$, *we have*

$$\begin{aligned}\Big|\mathbb{E}\Phi\big(\widehat{\boldsymbol{q}}_1^* S(z_1,X)\widehat{\boldsymbol{p}}_1, \ldots, \widehat{\boldsymbol{q}}_n^* S(z_n,X)\widehat{\boldsymbol{p}}_n\big) \\ - \mathbb{E}\Phi\big(\widehat{\boldsymbol{q}}_1^* S(z_1,Y)\widehat{\boldsymbol{p}}_1, \ldots, \widehat{\boldsymbol{q}}_n^* S(z_n,Y)\widehat{\boldsymbol{p}}_n\big)\Big| \\ \leq C n^5 M^{-1/2+C'\epsilon},\end{aligned}$$

where $C' > 0$ *depends only on* C_0 *in Definition* 5.14.

PROOF. The following proof is adapted from [21, theorem 16.1] and [26]. Let $\phi : [\![1, MN]\!] \to [\![1, N]\!] \times [\![1, M]\!]$ be a bijection.[7] For $\gamma \in [\![1, MN]\!]$ define X_γ by

$$(X_\gamma)_{\phi(\ell)} = \begin{cases} Y_{\phi(\ell)} & \ell \leq \gamma, \\ X_{\phi(\ell)} & \ell > \gamma. \end{cases}$$

Note that $X_0 = X$ and $X_{MN} = Y$ and also that X_γ and $X_{\gamma+1}$ differ only in the $\phi(\gamma+1)$ entry. Define Q_γ by $(Q_\gamma)_{\phi(\ell)} = (X_{\gamma+1})_{\phi(\ell)}$ if $\ell \neq \gamma+1$ and $(Q_\gamma)_{\phi(\gamma+1)} = 0$, so that Q_γ has a zero in the exact entry where X_γ and $X_{\gamma+1}$ differ. We then compare X_γ to Q_γ using Lemma 5.20 and a fifth-order Taylor expansion of Φ:

$$\begin{aligned}&\Phi(\widehat{\boldsymbol{q}}_1^* S(z_1,X_\gamma)\widehat{\boldsymbol{p}}_1, \ldots, \widehat{\boldsymbol{q}}_n^* S(z_n,X_\gamma)\widehat{\boldsymbol{p}}_n) \\ &\quad = \Phi(\widehat{\boldsymbol{q}}_1^* S(z_1,Q_\gamma)\widehat{\boldsymbol{p}}_1 + \xi_1, \ldots, \widehat{\boldsymbol{q}}_n^* S(z_n,Q_\gamma)\widehat{\boldsymbol{p}}_n + \xi_n) \\ &\quad = \Phi(\widehat{\boldsymbol{q}}_1^* S(z_1,Q_\gamma)\widehat{\boldsymbol{p}}_1, \ldots, \widehat{\boldsymbol{q}}_n^* S(z_n,Q_\gamma)\widehat{\boldsymbol{p}}_n) +\end{aligned}$$

[7] Here $[\![1, N]\!] = \{1, 2, \ldots, N\}$.

$$+\sum_{k=1}^{4}\sum_{|\alpha|=k}\partial^{\alpha}\Phi(\widehat{\boldsymbol{q}}_1^{*}S(z_1,Q_\gamma)\widehat{\boldsymbol{p}}_1,\dots,\widehat{\boldsymbol{q}}_n^{*}S(z_n,Q_\gamma)\widehat{\boldsymbol{p}}_n)\frac{\boldsymbol{\xi}^{\alpha}}{\alpha!}$$

$$+\sum_{|\alpha|=5}\partial^{\alpha}\Phi(\widehat{\boldsymbol{q}}_1^{*}S(z_1,Q_\gamma)\widehat{\boldsymbol{p}}_1+c\xi_1,\dots,\widehat{\boldsymbol{q}}_n^{*}S(z_n,Q_\gamma)\widehat{\boldsymbol{p}}_n+c\xi_n)\widehat{\boldsymbol{p}}_n)\frac{\boldsymbol{\xi}^{\alpha}}{\alpha!},$$

for some $0\le c\le 1$. Here $\boldsymbol{\xi}=(\xi_1,\dots,\xi_n)$ and $\xi_j=\sum_{k=1}^{5}M^{-k/2}J_{k,j}$, $J_{5,j}=0$ represents the ξ-term in Lemma 5.20 applied to $\widehat{\boldsymbol{q}_j}$, $\widehat{\boldsymbol{p}_j}$, z_j, and X_γ. We rewrite this expansion by collecting powers of $M^{1/2}$

$$\Phi(\widehat{\boldsymbol{q}}_1^{*}S(z_1,X_\gamma)\widehat{\boldsymbol{p}}_1,\dots,\widehat{\boldsymbol{q}}_n^{*}S(z_n,X_\gamma)\widehat{\boldsymbol{p}}_n)$$

$$=\Phi(\widehat{\boldsymbol{q}}_1^{*}S(z_1,Q_\gamma)\widehat{\boldsymbol{p}}_1,\dots,\widehat{\boldsymbol{q}}_n^{*}S(z_n,Q_\gamma)\widehat{\boldsymbol{p}}_n)+\sum_{k=1}^{4}M^{-k/2}T_{k,\gamma}.$$

By independence $\mathbb{E}[T_k]$ for $k\le 4$ decomposes into a sum of terms that are a product of a quantity depending only on moments $X_{\phi(\gamma+1)}$, $\mathbb{E}M^{(l+p)/2}X^{\ell}_{\phi(\gamma+1)}\overline{X}^{p}_{\phi(\gamma+1)}$, $p+\ell\le 4$, and a quantity depending on other variables. Then, an estimate is needed for $\mathbb{E}T_k$.

For $\epsilon>0$ and $D>0$, let $\mathcal{E}_{Q_\gamma}$ be the event where

$$\max_{k,\ell,j}\big[\big|\boldsymbol{v}_k^{*}G(z_\ell,Q_\gamma)\boldsymbol{w}_j\big|+\big|\boldsymbol{v}_k^{*}S(z_\ell,Q_\gamma)\boldsymbol{w}_j\big|\big]>M^{\epsilon},$$

and the families of vectors $\{\boldsymbol{v}_k\}$ and $\{\boldsymbol{w}_k\}$ are given by the union of the families $\{\widehat{\boldsymbol{q}}_k\}$ and $\{\widehat{\boldsymbol{p}}_k\}$ with the standard basis vectors, respectively. Then there exists a constant $C>0$, independent of γ, such that the probability of this event is bounded above by CM^{-D}. Also, let $\mathcal{X}_\gamma$ be the event where

$$\sqrt{M}\,|X_{\phi(\gamma)}|>M^{\epsilon}.$$

We use the a priori bound $\|G(z_\ell,X_\gamma)\|\le CM^{\delta}$ (see (5.9)) and that

$$\begin{aligned}\boldsymbol{v}_k^{*}G(z_\ell;X_\gamma)\boldsymbol{w}_j&=\boldsymbol{v}_k^{*}G(z_\ell;Q_\gamma)\boldsymbol{w}_j\\&\quad-\boldsymbol{v}_k^{*}G(z_\ell;Q_\gamma)(X_\gamma-Q_\gamma)G(z;Q_\gamma)\boldsymbol{w}_j\\&\quad+\boldsymbol{v}_k^{*}G(z_\ell;Q_\gamma)(X_\gamma-Q_\gamma)G(z_\ell;Q_\gamma)(X_\gamma-Q_\gamma)G(z_\ell;X_\gamma)\boldsymbol{w}_j.\end{aligned}$$

On the event $\mathcal{E}^{c}_{Q_\gamma}\cap\mathcal{X}^{c}_{\gamma}$

$$\big|\boldsymbol{v}_k^{*}G(z_\ell;X_\gamma)\boldsymbol{w}_j\big|\le M^{\epsilon}+2M^{3\epsilon-1/2}+4CM^{5\epsilon-1+\delta}.$$

Using an expansion to the next order, one obtains

$$\big|\boldsymbol{v}_k^{*}S\big(z_\ell;X_\gamma\big)\boldsymbol{w}_j\big|\le 2M^{3\epsilon}+4M^{5\epsilon-1/2}+8CM^{7\epsilon-1+\delta}$$

Provided that $4\epsilon-1+\delta\le 0$, we have that

$$\max_{k,\ell,j}\big[\big|\boldsymbol{v}_k^{*}G(z_\ell,X_\gamma)\boldsymbol{w}_j\big|+\big|\boldsymbol{v}_k^{*}S(z_\ell,X_\gamma)\boldsymbol{w}_j\big|\big]\le C'M^{3\epsilon},$$

for a new constant C'.

Now, consider

$$\mathbb{E}\Phi = \mathbb{E}\Phi(\mathbb{1}_{\mathcal{X}_\gamma} + \mathbb{1}_{\mathcal{X}_\gamma^c})(\mathbb{1}_{\mathcal{E}_{Q_\gamma}} + \mathbb{1}_{\mathcal{E}_{Q_\gamma}^c}),$$

where $|\Phi| \leq 1$, without loss of generality. Then for every $D > 0$ there exists $C > 0$ such that

$$|\mathbb{E}\Phi\mathbb{1}_{\mathcal{X}_\gamma}\mathbb{1}_{\mathcal{E}_{Q_\gamma}^c} + \mathbb{E}\Phi\mathbb{1}_{\mathcal{X}_\gamma^c}\mathbb{1}_{\mathcal{E}_{Q_\gamma}} + \mathbb{E}\Phi\mathbb{1}_{\mathcal{X}_\gamma^c}\mathbb{1}_{\mathcal{E}_{Q_\gamma}^c}| \leq CM^{-D}.$$

We need to consider

$$\mathbb{E}T_{k,\gamma}\mathbb{1}_{\mathcal{X}_\gamma^c}\mathbb{1}_{\mathcal{E}_{Q_\gamma}^c}.$$

First,

$$\left|\mathbb{E}T_{5,\gamma}\mathbb{1}_{\mathcal{E}_{Q_\gamma}^c}\mathbb{1}_{\mathcal{E}_{X_\gamma,M}^c}\right| \leq 1024n^5 M^{3(C_0+8)\epsilon-5/2} \max_{1\leq k\leq 25} \mathbb{E}\left|\sqrt{M}X_{\phi(\gamma)}\right|^k$$

where $M^{3C_0\epsilon}$ is the upper bound on all derivatives of Φ, and $1024n^5$ is a bound on the number of terms in the Taylor expansion. For $T_{k,\gamma}$ we note that for any $D > 0$ there exists a constant $C > 0$ such that

$$\left|\mathbb{E}T_{k,\gamma}\mathbb{1}_{\mathcal{E}_{Q_\gamma}^c}\mathbb{1}_{\mathcal{E}_{X_\gamma,M}^c} - \mathbb{E}T_{k,\gamma}\mathbb{1}_{\mathcal{E}_{Q_\gamma}^c}\right| \leq CM^{-D}.$$

So, we can write

$$\left|\mathbb{E}\Phi(\widehat{\boldsymbol{q}}_1^* S(z_1, X_\gamma)\widehat{\boldsymbol{p}}_1, \ldots, \widehat{\boldsymbol{q}}_n^* S(z_n, X_\gamma)\widehat{\boldsymbol{p}}_n) - \sum_{k=1}^{4} M^{-k/2}L_k\right| \leq Cn^5 M^{3(C_0+8)\epsilon-5/2}$$

where L_k depends only on Q_γ and the moments of $X_{\phi(\gamma)}$ up to order 4. The proposition follows using

$$\begin{aligned}
&\mathbb{E}\Phi(\widehat{\boldsymbol{q}}_1^* S(z_1, X)\widehat{\boldsymbol{p}}_1, \ldots, \widehat{\boldsymbol{q}}_n^* S(z_n, X)\widehat{\boldsymbol{p}}_n) \\
&\quad - \mathbb{E}\Phi(\widehat{\boldsymbol{q}}_1^* S(z_1, Y)\widehat{\boldsymbol{p}}_1, \ldots, \widehat{\boldsymbol{q}}_n^* S(z_n, Y)\widehat{\boldsymbol{p}}_n) \\
&\quad = \sum_{\gamma=1}^{NM} \mathbb{E}\Phi(\widehat{\boldsymbol{q}}_1^* S(z_1, X_\gamma)\widehat{\boldsymbol{p}}_1, \ldots, \widehat{\boldsymbol{q}}_n^* S(z_n, X_\gamma)\widehat{\boldsymbol{p}}_n) \\
&\qquad - \sum_{\gamma=1}^{NM} \mathbb{E}\Phi(\widehat{\boldsymbol{q}}_1^* S(z_1, X_{\gamma+1})\widehat{\boldsymbol{p}}_1, \ldots, \widehat{\boldsymbol{q}}_n^* S(z_n, X_{\gamma+1})\widehat{\boldsymbol{p}}_n).
\end{aligned}$$

□

We recall well-known important facts about the trapezoidal rule applied to approximate contour integrals on smooth closed curves. Suppose Γ is such a curve of length 1 with arc length parametrization $\ell : [0, 1] \to \Gamma$. We choose ℓ so that

$\ell(0), \ell(1/2) \in \mathbb{R}$ and $\ell(0) < \ell(1/2)$. With m points, the trapezoidal rule can be used at the nodes $t_j = j/m$ for $j = 0, 1, \ldots, m$. In our case, however, we wish to avoid evaluating on the real axis and we choose $s_j^{(m)} = s_j = (t_j + t_{j+1})/2 = (2j+1)/(2m)$, $j = 0, 1, \ldots, m$, with the convention that $s_m = s_0$. Consider

$$\oint_\Gamma f(z)\mathrm{d}z = \int_0^1 f(\ell(s))\ell'(s)\mathrm{d}s \approx \sum_{j=0}^{m-1} f(\ell(s_j))\frac{\ell'(s_j)}{m} = \sum_{j=0}^{m-1} f(z_j)w_j,$$

$$z_j^{(m)} = z_j = \ell(s_j), \quad w_j^{(m)} = w_j = \frac{\ell'(s_j)}{m}.$$

Using the Euler–Maclaurin formula, for every $D > 0$ there exists $C_D > 0$ such that

$$\left| \oint_\Gamma f(z)dz - \sum_{j=0}^{m-1} f(z_j)w_j \right| \le C_D(\Gamma)\| f^{(D)} \|_\infty m^{-D}.$$

PROOF OF THEOREM 5.15. We prove the proposition for $\Gamma_j = \Gamma$ for all j. The arguments easily extend to the general case. Let $\Phi : \mathbb{C}^n \to \mathbb{R}$ be an admissible test function. We approximate

$$\frac{\sqrt{M}}{2\pi i}\oint_\Gamma f_j(z)(c_0(z;\mu_T) - s_\mathfrak{d}(z))\mathrm{d}z$$

using the trapezoidal rule and consider

$$\Delta_{M,m} := \Phi\left(\frac{\sqrt{M}}{2\pi i}\oint_\Gamma f_1(z)(c_0(z;\mu_T) - s_\mathfrak{d}(z))\mathrm{d}z, \ldots, \frac{\sqrt{M}}{2\pi i}\oint_\Gamma f_n(z)(c_0(z;\mu_T) - s_\mathfrak{d}(z))\mathrm{d}z \right)$$

The choice of m is critical. Examining how the conclusion of Proposition 5.21 depends on n, we need $m^5 < M^{1/2}$. So, we choose $m = M^{1/20}$.

Because Φ is bounded, for $\delta > 0$ we can restrict to the event $\mathcal{L}_\delta = \{\lambda_N \ge \gamma_- - \delta, \lambda_1 \le \gamma_+ + \delta\}$, and there exists C_D such that $\mathbb{P}(\mathcal{L}_\delta) \ge 1 - C_D M^{-D}$ for all $D > 0$. Furthermore, we choose δ so that $[\gamma_- - \delta, \gamma_+ + \delta] \subset \Omega$. By fixing δ, on this event the integrands and all their derivatives up to order E are bounded by $\sqrt{M}c_E$ for some $c_E > 0$. Then, for example, on the event $\mathcal{L}_\delta$,

$$\left| \frac{\sqrt{M}}{2\pi i}\oint_\Gamma f_1(z)(c_0(z;\mu_T) - s_\mathfrak{d}(z))\mathrm{d}z - \frac{\sqrt{M}}{2\pi i}\sum_{j=1}^{m} f_1(z_j)(c_0(z_j;\mu_T) - s_\mathfrak{d}(z_j))w_j \right| \le \frac{C_E(\Gamma)c_E\sqrt{M}}{m^E}.$$

Since E can be chosen arbitrarily large, we then find

$$|\mathbb{E}\Delta_{M,m}\mathbb{1}_{\mathcal{L}_\delta}| \leq CM^{-D}$$

for any $D > 0$. Therefore, it suffices to consider

$$\begin{aligned}\tilde{\Delta}_{M,m} := \Phi\Bigg(&\sqrt{M}\sum_{j=1}^m f_1(z_j)(c_0(z_j;\mu_T) - s_{\mathfrak{d}}(z_j))w_j, \dots,\\ &\sqrt{M}\sum_{j=1}^m f_n(z_j)(c_0(z_j;\mu_T) - s_{\mathfrak{d}}(z_j))w_j\Bigg)\\ -\Phi\Bigg(&\sqrt{M}\sum_{j=1}^m f_1(z_j)(c_0(z_j;\mu_{\widetilde{T}}) - s_{\mathfrak{d}}(z_j))w_j, \dots,\\ &\sqrt{M}\sum_{j=1}^m f_n(z_j)(c_0(z_j;\mu_{\widetilde{T}}) - s_{\mathfrak{d}}(z_j))w_j\Bigg).\end{aligned}$$

And, we are led to consider the function $\Psi : \mathbb{C}^m \to \mathbb{R}$

$$\Psi(x_1, x_2, \dots, x_m) = \Phi\left(\sum_{j=1}^m f_1(z_j)\frac{w_j}{2\pi i}x_j, \dots, \sum_{j=1}^m f_n(z_j)\frac{w_j}{2\pi i}x_j\right). \tag{5.11}$$

Define $W \in \mathbb{C}^{n\times m}$ by $W_{\ell j} = f_\ell(z_j)\frac{w_j}{2\pi i}$. It follows that

$$\begin{aligned}&\partial_{x_{j_1}x_{j_2}\cdots x_{j_q}}\Psi(x_1,\dots,x_m)\\ &\quad= \sum_{k_1,k_2,\dots,k_q=1}^n \partial_{y_{k_1}y_{k_2}\cdots y_{k_p}}\Phi(y_1,\dots,y_n)\left(\prod_{p=1}^q W_{k_p,j_p}\right),\\ &y_k = \sum_{j=1}^m f_k(z_j)\frac{w_j}{2\pi i}x_j.\end{aligned}$$

From this, we are able to estimate

$$\begin{aligned}&\left|\partial_{x_{j_1}x_{j_2}\cdots x_{j_q}}\Psi(x_1,\dots,x_m)\right|\\ &\quad\leq \max_{k_1,k_2,\dots,k_1}|\partial_{y_{k_1}y_{k_2}\cdots y_{k_p}}\Phi(y_1,\dots,y_n)|\sum_{k_1,k_2,\dots,k_q=1}^n\prod_{p=1}^q|W_{k_p,j_p}|\\ &\quad\leq \max_{k_1,k_2,\dots,k_1}|\partial_{y_{k_1}y_{k_2}\cdots y_{k_p}}\Phi(y_1,\dots,y_n)|\max_j\|f_j\|_\infty^q\left(\frac{C}{2\pi}\right)^q,\end{aligned}$$

where $C > 0$ is such that $\sum_j |w_j| \leq C$. Note that C can be chosen independent of m. Now let $\epsilon > 0$ be sufficiently small so that

$$|\partial_{x_{j_1}x_{j_2}\cdots x_{j_q}}\Phi(x_1,\dots,x_n)| \leq M^{C_0\epsilon} \quad \text{for } \max_j|x_j| \leq M^\epsilon.$$

All arguments for Φ in (5.11) are uniformly bounded by M^{ϵ} for

$$\max_j |x_j| \le M^{\epsilon} / \left(\frac{C}{2\pi} \max_j \| f_j \|_{\infty} \right).$$

Thus

$$|\partial_{y_{k_1} y_{k_2} \cdots y_{k_p}} \Psi(x_1, \ldots, x_\alpha)| \le M^{C_0 \epsilon} \max_j \| f_j \|_{\infty}^p \left(\frac{C}{2\pi} \right)^p$$

By setting $L = \frac{C}{2\pi} \max_j \| f_j \|_{\infty}$ we find that

$$\widetilde{\Psi}(x_1, x_2, \ldots, x_m) = \Phi \left(L^{-1} \sum_{j=1}^{m} f_1(z_j) \frac{w_j}{2\pi i} x_j, \ldots, L^{-1} \sum_{j=1}^{m} f_n(z_j) \frac{w_j}{2\pi i} x_j \right).$$

is admissible with the same constant C_0. Applying Proposition 5.21 to $\widetilde{\Psi}$ establishes the proposition. □

We also remark that these arguments, without the use of Proposition 5.21, can be used to show the following:

PROPOSITION 5.22. *Suppose W is a sample covariance matrix, $N/M \xrightarrow{M \to \infty} d \in (0, 1)$ and $T = T(W, \mathbf{b})$ for a sequence $\mathbf{b} = \mathbf{b}_N \in \mathbb{C}^N$ of nontrivial vectors. Then*

$$\left(\int \lambda^k \mu_T(\mathrm{d}\lambda) \right)_k \xrightarrow[M \to \infty]{(d)} \left(\int \lambda^k \varrho_{\mathfrak{d}}(\mathrm{d}\lambda) \right)_k,$$

in the sense of convergence of finite-dimensional marginals where $k \ge 0$ if $d = 1$ and $k \in \mathbb{Z}$ if $d < 1$.

6 Analysis of the Algorithms

The important fact that we use to prove Theorems 1.4 and 1.5 is that the entries in the Cholesky factorization of the three-term recurrence matrix associated to a measure μ are (generically) differentiable functions of the moments of the measure. This implies that the leading-order behavior (Theorem 1.4) is the same as in the Gaussian case and that, with the moment matching condition (Definition 2.4), the fluctuations must be the same as in the Gaussian case (Theorem 1.5). So, it suffices to prove Theorem 1.5 in the case of X having $\mathcal{N}_\beta(0, 1/M)$ entries. The following three sections do just this.

6.1 Proofs for the conjugate gradient algorithm

The basis for our analysis is Proposition 4.1 and Theorem 2.8. In this section we suppose $W \overset{\mathscr{L}}{=} \mathcal{W}_\beta(N, M)$, $N \le M$, and $\mathbf{b} = \mathbf{b}_N \in \mathbb{C}^N$ (or $\mathbb{R}^N$ if $\beta = 1$). And we recall the notation that $\boldsymbol{x}_k = \boldsymbol{x}_k(W, \mathbf{b})$ is the k^{th} iterate of the CGA applied to $W\boldsymbol{x} = \mathbf{b}$ and $\mathbf{r}_k = \mathbf{b} - W\boldsymbol{x}_k$, $\mathbf{e}_k = \boldsymbol{x} - \boldsymbol{x}_k$.

Nonasymptotic calculations

Using the notation (5.4), with $T = T(W, \mathbf{b}) = HH^T$, it follows that

$$\pi_k(0;\mu_T) = (-1)^{k+1}\prod_{j=0}^{k-1}\alpha_j^2, \quad T_{j,j+1} = \alpha_j\beta_j,$$

and therefore

$$\|\mathbf{r}_k\|_2^2 = \prod_{j=0}^{k-1}\frac{\beta_j^2}{\alpha_j^2}, \tag{6.1}$$

where the chi squared random variables are all mutually independent. This formula lends itself easily to asymptotic analysis.

Deriving a distributional expression for $\|\mathbf{e}_k\|_W^2$ is more involved. With the convention that $b_{-1} = 1$

$$\begin{bmatrix}\pi_{k+1}(x;\mu_T)\\ \pi_k(0;\mu_T)\end{bmatrix} = \begin{bmatrix}x-a_k & -b_{k-1}^2\\ 1 & 0\end{bmatrix}\begin{bmatrix}x-a_{k-1} & -b_{k-2}^2\\ 1 & 0\end{bmatrix}\cdots\begin{bmatrix}x-a_0 & -b_{-1}^2\\ 1 & 0\end{bmatrix}\begin{bmatrix}1\\ 0\end{bmatrix},$$

Then define the complementary polynomials

$$\begin{bmatrix}\widetilde{\pi}_{k+1}(x;\mu_T)\\ \widetilde{\pi}_k(x;\mu_T)\end{bmatrix} = \begin{bmatrix}x-a_k & -b_{k-1}^2\\ 1 & 0\end{bmatrix}\begin{bmatrix}x-a_{k-1} & -b_{k-2}^2\\ 1 & 0\end{bmatrix}\cdots\begin{bmatrix}x-a_0 & -b_{-1}^2\\ 1 & 0\end{bmatrix}\begin{bmatrix}0\\ 1\end{bmatrix}.$$

Decompose

$$c_k(0;\mu_T) = c_0(0;\mu_T)\pi_k(0;\mu_T) - \widetilde{\pi}_k(0;\mu_T).$$

Then

$$\widetilde{\pi}_k(0;\mu_T) = (-1)^{k+1}\sum_{\ell=0}^{k-1}\left(\prod_{j=1}^{\ell}\beta_{j-1}^2\right)\left(\prod_{j=\ell+1}^{k-1}\alpha_j^2\right),$$

giving

$$\frac{\widetilde{\pi}_k(0;\mu_T)}{\pi_k(0;\mu_T)} = \frac{1}{\alpha_0^2}\sum_{\ell=0}^{k-1}\prod_{j=1}^{\ell}\frac{\beta_{j-1}^2}{\alpha_j^2},$$

where the empty product returns 1. From Proposition 5.9

$$c_0(0,\mu_T) = \frac{1}{\alpha_0^2}\left(1 + \sum_{\ell=1}^{k-1}\prod_{j=1}^{\ell}\frac{\beta_{j-1}^2}{\alpha_j^2} + \left(\beta_{k-1}^2\Sigma_k^{-2}\right)\prod_{j=1}^{k-1}\frac{\beta_{j-1}^2}{\alpha_j^2}\right),$$

where $\Sigma_k \overset{\mathscr{L}}{=} \frac{\chi_{\beta(M-N+1)}}{\sqrt{\beta M}}$ is independent of $(\alpha_j, \beta_j)_{j=0}^{k-1}$. We find

$$\|\mathbf{e}_k\|_W^2 = \Sigma_k^{-2}\prod_{j=0}^{k-1}\frac{\beta_j^2}{\alpha_j^2}. \tag{6.2}$$

where the chi squared random variables are all mutually independent. This establishes Theorem 1.2(a), and Theorem 1.1 follows as well.

Asymptotic calculations

PROOF OF THEOREMS 1.4(A) AND 1.5(A) WHEN $\sqrt{M}X \overset{\mathscr{L}}{=} \mathcal{G}_\beta(N, M)$. Decompose

$$c_0(0;\mu_T) = \frac{1}{1-\mathfrak{d}} + \frac{\sqrt{2}}{\sqrt{\beta M}} R(\mu_T), \quad T = T(W, \mathbf{b}),$$

using Proposition 5.11. In the notation of this proposition $R(\mu_T) \xrightarrow[M\to\infty]{(d)} Z_{-1}/\sqrt{2}$. Then using the complementary polynomials

$$\frac{c_k(0;\mu_T)}{\pi_k(0;\mu_T)} = \frac{(1-\mathfrak{d})^{-1}\pi_k(0;\mu_T) - \widetilde{\pi}_k(0;\mu_T)}{\pi_k(0;\mu_T)} + \frac{\sqrt{2}}{\sqrt{\beta M}} R(\mu_T).$$

We write $T = HH^T$, again using the notation (5.4). Using the distributional limit described in Proposition 5.11 one can compute the large N behavior. Specifically, we use (5.6) and (7.7) extensively. Using the same process $(Z_j)_{j\geq 1}$ write

$$A_k := \begin{bmatrix} -a_k & -b_{k-1}^2 \\ 1 & 0 \end{bmatrix} = \hat{E} + \sqrt{\frac{2}{\beta M}}\check{E}_k + o(M^{-1/2}),$$

$$\hat{E} = \begin{bmatrix} -1-\mathfrak{d} & -\mathfrak{d} \\ 1 & 0 \end{bmatrix},$$

$$\check{E}_k = \begin{bmatrix} -Z_{2k+1} - \sqrt{\mathfrak{d}}Z_{2k} & -\mathfrak{d}Z_{2k-1} - \sqrt{\mathfrak{d}}Z_{2k} \\ 0 & 1 \end{bmatrix}.$$

We compute the asymptotics of the quantity

$$\begin{bmatrix} 1 & 0 \end{bmatrix} A_{k-1}A_{k-2}\cdots A_1 \begin{bmatrix} -\alpha_0^2 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{1-\mathfrak{d}} \\ -1 \end{bmatrix}$$

using that

$$\hat{E} = V\Lambda V^{-1}, \quad V = \begin{bmatrix} -1 & -\mathfrak{d} \\ 1 & 1 \end{bmatrix}, \quad V^{-1} = \frac{1}{1-\mathfrak{d}} \begin{bmatrix} -1 & -\mathfrak{d} \\ 1 & 1 \end{bmatrix},$$

$$\Lambda = \operatorname{diag}(-1, -\mathfrak{d}).$$

We find that

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \hat{E}^{k-j-1}\check{E}_j\hat{E}^{j-1} \begin{bmatrix} -1 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{1-\mathfrak{d}} \\ -1 \end{bmatrix}$$

$$= (-1)^{k+1}\frac{\mathfrak{d}^j - \mathfrak{d}^k}{(1-\mathfrak{d})^2}\left[Z_{2j+1} + \sqrt{\mathfrak{d}}\, Z_{2j} - Z_{2j-1} - Z_{2j}/\sqrt{\mathfrak{d}}\right].$$

Similarly,

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \hat{E}^{k-1} \begin{bmatrix} -Z_1 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{1-\mathfrak{d}} \\ -1 \end{bmatrix} = (-1)^k \frac{1-\mathfrak{d}^k}{(1-\mathfrak{d})^2} Z_1.$$

Therefore it remains to analyze

$$\prod_{j=0}^{k-1}\alpha_j^2 = 1 + \frac{\sqrt{2}}{\sqrt{\beta M}}\sum_{j=0}^{k-1} Z_{2j+1} + o(M^{-1/2}),$$
$$\prod_{j=0}^{k-1}\beta_j^2 = \mathfrak{d}^k + \frac{\sqrt{2}}{\sqrt{\beta M}}\sum_{j=0}^{k-1} \mathfrak{d}^{k-1/2} Z_{2j+2} + o(M^{-1/2}).$$

The distributional limit of $R(\mu_T)$ is provided by Proposition 5.11. So, our final expressions become

$$\begin{aligned}\|\mathbf{e}_k(W,\mathbf{b})\|_W^2 &= \frac{\mathfrak{d}^k}{1-\mathfrak{d}}\Bigg(1 + \frac{\sqrt{2}}{\sqrt{\beta M}}\Bigg[\sum_{j=k}^{\infty}\mathfrak{d}^{j-k}(Z_{2j}/\sqrt{\mathfrak{d}} - Z_{2j+1}) \\ &\qquad + \sum_{j=1}^{k-1}(Z_{2j}/\sqrt{\mathfrak{d}} - Z_{2j-1}) - Z_{2k-1}\Bigg]\Bigg) + o(M^{-1/2}),\end{aligned}$$

$$\|\mathbf{r}_k(W,\mathbf{b})\|_2^2 = \mathfrak{d}^k\left(1 + \frac{\sqrt{2}}{\sqrt{\beta M}}\left[\sum_{j=0}^{k-1}\left(Z_{2j+2}/\sqrt{\mathfrak{d}} - Z_{2j+1}\right)\right]\right) + o(M^{-1/2}).$$

The theorem follows. □

6.2 Proofs for the MINRES algorithm

Nonasymptotic calculations

The proof of Theorem 1.1(b) is immediate from the simple formula

$$p_j(0,\mu_T) = \frac{\det(-T_j)}{\prod_{\ell=0}^{j-1} b_j^2} = (-1)^j\prod_{\ell=0}^{j-1}\frac{\beta_\ell}{\alpha_\ell},$$

using (5.4). Then using (5.1), Theorem 1.2(b) follows.

Asymptotic calculations

PROOFS OF THEOREMS 1.4(B) AND 1.5(B) WHEN $\sqrt{M}X \overset{\mathscr{L}}{=} \mathcal{G}_\beta(N,M)$. It suffices to prove Theorem 1.5(b) in this case. From Theorem 1.5(b) we have that

$$\prod_{\ell=0}^{j-1}\frac{\beta_\ell^2}{\alpha_\ell^2} = \mathfrak{d}^j\left(1 + \frac{\sqrt{2}}{\sqrt{\beta M}}\widetilde{Z}_j^{\mathbf{r},\mathrm{CG}}\right) + o(M^{-1/2}), \quad \widetilde{Z}_j^{\mathbf{r},\mathrm{CG}} = d^{-j}Z_j^{\mathbf{r},\mathrm{CG}}.$$

This implies

$$\sum_{j=0}^{k}\prod_{\ell=0}^{j-1}\frac{\alpha_\ell^2}{\beta_\ell^2} = \sum_{j=0}^{k}\mathfrak{d}^{-j} - \frac{\sqrt{2}}{\sqrt{\beta M}}\sum_{j=1}^{k}\mathfrak{d}^{-j}\widetilde{Z}_j^{\mathbf{r},\mathrm{CG}} + o(M^{-1/2}).$$

And this gives

$$\left(\sum_{j=0}^{k}\prod_{\ell=0}^{j-1}\frac{\alpha_\ell^2}{\beta_\ell^2}\right)^{-1} = \frac{1-\mathfrak{d}^{-1}}{1-\mathfrak{d}^{-k-1}} + \left(\frac{1-\mathfrak{d}^{-1}}{1-\mathfrak{d}^{-k-1}}\right)^2 \frac{\sqrt{2}}{\sqrt{\beta M}}\sum_{j=1}^{k}\mathfrak{d}^{-j}\,\widetilde{Z}_j^{\mathbf{r},\mathrm{CG}} + o(M^{-1/2}).$$

In writing,

$$\frac{1-\mathfrak{d}^{-1}}{1-\mathfrak{d}^{-k-1}} = \mathfrak{d}^k\frac{1-\mathfrak{d}}{1-\mathfrak{d}^{k+1}},$$

we establish the theorem. □

6.3 Proofs for the conjugate gradient algorithm applied to the normal equations

First, observe that for $\alpha > 0$

$$p_j(\lambda;\alpha\mu) = \frac{p_j(\lambda;\mu)}{\sqrt{\alpha}}. \tag{6.3}$$

Consider the distribution of the measure ν as defined in (4.3), and in particular, the distribution on the absolute value of the vector $V^*\mathbf{b}$ where $X = U\Sigma V^*$ is the singular value decomposition of X. We know that V can be taken to be Haar distributed on either the orthogonal ($\beta = 1$) or unitary ($\beta = 2$) group [20]. By invariance, if $\|\mathbf{b}\|_2 = 1$ then $\mathbf{b}$ can be replaced with $\boldsymbol{f}_1$. From this it follows that for $T = T(W,\mathbf{a})$

$$\mu_T \overset{\mathscr{L}}{=} \sum_{j=1}^{N}\omega_j\delta_{\lambda_j}, \quad \omega_j \overset{\mathscr{L}}{=} \frac{\chi_{\beta,j}^2}{\sum_{\ell=1}^{N}\chi_{\beta,\ell}^2}, \quad j = 1,2,\ldots,N,$$

and $(\lambda_1,\ldots,\lambda_N)$ are the eigenvalues of W which are independent of $(\omega_1,\ldots,\omega_N)$. So, we find that, in the notation of Theorem 1.2(c)

$$\nu \overset{\mathscr{L}}{=} \underbrace{\left(\frac{\sum_{\ell=1}^{N}\chi_{\beta,\ell}^2}{\sum_{\ell=1}^{M}\chi_{\beta,\ell}^2}\right)}_{\Delta_{N,M}}\sum_{j=1}^{N}\omega_j\delta_{\lambda_j}.$$

Combined with (6.3), this gives the proof of Theorem 1.2(c). And then Theorem 1.5(c) and Theorem 1.4(c), in the Gaussian case, follow.

Acknowledgment. The authors would like to thank Xuicai Ding for helping clarify when Proposition 5.12 applies. This work was supported in part by NSF Grants DMS-1753185 and DMS-1945652 (TT)

Bibliography

[1] Bai, Z.; Silverstein, J. W. *Spectral analysis of large dimensional random matrices.* Second edition. Springer Series in Statistics. Springer, New York, 2010. doi:10.1007/978-1-4419-0661-8

[2] Bai, Z.; Yin, Y. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab.* **21** (1993), no. 3, 1275–1294.

[3] Bai, Z. D.; Silverstein, J. W. CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32** (2004), no. 1A, 553–605. doi:10.1214/aop/1078415845

[4] Beckermann, B.; Kuijlaars, A. B. J. Superlinear convergence of conjugate gradients. *SIAM J. Numer. Anal.* **39** (2001), no. 1, 300–329. doi:10.1137/S0036142999363188

[5] Borgwardt, K. H. *The simplex method: A probabilistic analysis*. Springer, Berlin-Heidelberg, 1987.

[6] Davidson, K.; Szarek, S. J. Local operator theory, random matrices and Banach spaces. *Handbook of the geometry of Banach spaces, Vol. I*, 317-–366. North-Holland, Amsterdam, 2001. doi:10.1016/S1874-5849(01)80010-3

[7] Deift, P. A. *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach.* Courant Lecture Notes in Mathematics, 3. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 1999.

[8] Deift, P. A.; Menon, G.; Olver, S.; Trogdon, T. Universality in numerical computations with random data. *Proc. Natl. Acad. Sci. USA* **111** (2014), no. 42, 14973-–14978. doi:10.1073/pnas.1413446111

[9] Deift, P. A.; Menon, G.; Trogdon, T. On the condition number of the critically-scaled Laguerre unitary ensemble. *Discrete Contin. Dyn. Syst.* **36** (2016), no. 8, 4287-–4347. doi:10.3934/dcds.2016.36.4287

[10] Deift, P.; Trogdon, T. Universality for eigenvalue algorithms on sample covariance matrices. *SIAM J. Numer. Anal.* **55** (2017), no. 6, 2835-–2862. doi:10.1137/17M1110900

[11] Deift, P.; Trogdon, T. Universality for the Toda algorithm to compute the largest eigenvalue of a random matrix. *Comm. Pure Appl. Math.* **71** (2018), no. 3, 505-–536. doi:10.1002/cpa.21715

[12] Deift, P.; Trogdon, T. Universality in numerical computation with random data: case studies, analytical results and some speculations. *Computation and combinatorics in dynamics, stochastics and control*, 221-–231. Abel Symp., 13. Springer, Cham, 2018.

[13] Deift, P.; Trogdon, T. The conjugate gradient algorithm on well-conditioned Wishart matrices is almost deterministic. *Quart. Appl. Math* **79** (2020), no. 1, 125–161. doi:10.1090/qam/1574

[14] Ding, X.; Trogdon, T. A Riemann-Hilbert approach to the perturbation theory for orthogonal polynomials: Applications to numerical linear algebra and random matrix theory. Preprint, 2021. arXiv:2112.12354 [math.PR]

[15] Ding, X.; Trogdon, T. The conjugate gradient algorithm on a general class of spiked covariance matrices. *Quart. Appl. Math.* **80** (2021), no. 1, 99–155. doi:10.1090/qam/1605

[16] Dumitriu, I.; Edelman, A. Matrix models for beta ensembles. *J. Math. Phys.* **43** (2002), no. 11, 5830–5847. doi:10.1063/1.1507823

[17] Dumitriu, I.; Edelman, A. Global spectrum fluctuations for the β-Hermite and β-Laguerre ensembles via matrix models. *J. Math. Phys.* **47** (2006), no. 6, 063302, 36 pp. doi:10.1063/1.2200144

[18] Duy, T. K. On spectral measures of random Jacobi matrices. *Osaka J. Math.* **55** (2018), no. 4, 595–617.

[19] Duy, T. K.; Shirai, T. The mean spectral measures of random Jacobi matrices related to Gaussian beta ensembles. *Electron. Commun. Probab.* **20** (2015), no. 68, 13 pp. doi:10.1214/ECP.v20-4252

[20] Edelman, A.; Rao, N. R. Random matrix theory. *Acta Numer.* **14** (2005), 233–297. doi:10.1017/S0962492904000236

[21] Erdős, L.; Yau, H.-T. *A dynamical approach to random matrix theory.* Courant Lecture Notes in Mathematics, 28. Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 2017.

[22] Geman, S. A limit theorem for the norm of random matrices. *Ann. Probab* **8** (1980), no. 2, 252–261.

[23] Goldstine, H. H.; von Neumann, J. Numerical inverting of matrices of high order. II. *Proc. Amer. Math. Soc.* **2** (1951), no. 2, 188–202. doi:10.2307/2032484

[24] Hestenes, M.; Steifel, E. Method of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards* **20** (1952), 409–436 (1953).

[25] Johansson, K. On fluctuations of eigenvalues of random Hermitian matrices. *Duke Math. J.* **91** (1998), no. 1, 151–204. doi:10.1215/S0012-7094-98-09108-6

[26] Knowles, A.; Yin, J. Anisotropic local laws for random matrices. *Probab. Theory Related Fields* **169** (2017), no. 1-2, 257–352. doi:10.1007/s00440-016-0730-4

[27] Kuijlaars, A. B. J. Convergence analysis of Krylov subspace iterations with methods from potential theory. *SIAM Rev.* **48** (2006), no. 1, 3–40. doi:10.1137/S0036144504445376

[28] Menon, G.; Trogdon, T. Smoothed analysis for the conjugate gradient algorithm. *SIGMA Symmetry Integrability Geom. Methods Appl.* **12** (2016), Paper No. 109, 22 pp. doi:10.3842/SIGMA.2016.109

[29] Meurant, G. On prescribing the convergence behavior of the conjugate gradient algorithm. *Numer. Algorithms* **84** (2020), no. 4, 1353-–1380. doi:10.1007/s11075-019-00851-2

[30] Muirhead, R. J. *Aspects of multivariate statistical Theory*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1982. doi:10.1002/9780470316559

[31] Olver, F. W. J.; Lozier, D. W.; Boisvert, R. F.; Clark, C. W. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[32] O'Rourke, S.; Renfrew, D.; Soshnikov, A. On fluctuations of matrix entries of regular functions of Wigner matrices with non-identically distributed entries. *J. Theoret. Probab.* **26** (2013), no. 3, 750–780. doi:10.1007/s10959-011-0396-x

[33] O'Rourke, S.; Renfrew, D.; Soshnikov, A. Fluctuations of matrix entries of regular functions of sample covariance random matrices. *Theory Probab. Appl.* **58** (2014), no. 4, 615–639. doi:10.1137/S0040585X97986801

[34] Paquette, C.; van Merriënboer, B.; Paquette, E.; Pedregosa, F. Halting time is predictable for large models: a universality property and average-case analysis. Preprint, 2020. arXiv:2006.04299 [math.OC]

[35] Pfrang, C. W.; Deift, P.; Menon, G. How long does it take to compute the eigenvalues of a random symmetric matrix? *Random matrix theory, interacting particle systems, and integrable systems*, 411–442. Math. Sci. Res. Inst. Publ., 65. Cambridge Univ. Press, New York, 2014.

[36] Sankar, A.; Spielman, D. A.; Teng, S.-H. Smoothed analysis of the condition numbers and growth factors of matrices. *SIAM J. Matrix Anal. Appl.* **28** (2006), no. 2, 446–476. doi:10.1137/S0895479803436202

[37] Shcherbina, M. Central limit theorem for linear eigenvalue statistics of the Wigner and sample covariance random matrices. *Zh. Mat. Fiz. Anal. Geom.* **7** (2011), no. 2, 176–192, 197, 199.

[38] Silverstein, J. W. The smallest eigenvalue of a large dimensional Wishart matrix. *Ann. Probab.* **13** (1985), no. 4, 1364–1368.

[39] Smale, S. On the average number of steps of the simplex method of linear programming. *Math. Programming* **27** (1983), no. 3, 241–262. doi:10.1007/BF02591902

[40] Spielman, D.; Teng, S.-H. Smoothed analysis of algorithms. *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing - STOC '01*, 296–305. ACM Press, New York, 2001.

[41] Trefethen, L. N.; Bau III, D. *Numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997. doi:10.1137/1.9780898719574

[42] Vargas, J. G.; Kulkarni, A. The Lanczos algorithm under few iterations: concentration and location of the Ritz values. Preprint, 2019. arXiv:1904.06012 [math.NA]

[43] Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing*, 210–268. Cambridge University Press, Cambridge, 2009.

[44] Wishart, J. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika* **20A** (1928), no. 1-2, 32–52.

[45] Yin, Y. Q.; Bai, Z. D.; Krishnaiah, P. R. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probab. Theory Related Fields* **78** (1988), no. 4, 509–521. doi:10.1007/BF00353874

ELLIOT PAQUETTE
McGill University
Department of Mathematics
and Statistics
805 Rue Sherbrooke O
Montréal, QC H3A 2K6
CANADA
E-mail: elliot.paquette@mcgill.ca

THOMAS TROGDON
University of Washington
Department of Applied Mathematics
Seattle, WA 98195-3925
USA
E-mail: trogdon@uw.edu

Received January 2022.