The Spectrum of Random Inner-product Kernel Matrices

Xiuyuan Cheng and Amit Singer

Princeton University,
Fine Hall, Washington Rd.,
Princeton, NJ, 08544
e-mail: xiuyuanc@math.princeton.edu

amits@math.princeton.edu

Abstract:
We consider n-by-n matrices whose (i,j)th entry is $f(X_i^T X_j)$, where $X_1, \ldots, X_n$ are i.i.d. standard Gaussian in $\mathbb{R}^p$, and f is a real-valued function. The weak limit of the eigenvalue distribution of these random matrices is studied at the limit when $p, n \to \infty$ and $p/n = \gamma$ which is a constant. We show that, under certain conditions on the kernel function f, the limiting spectral density exists and is dictated by a cubic equation involving its Stieltjes transform. The parameters of this cubic equation are decided by a Hermite-like expansion of the rescaled kernel function. While the case that f is differentiable at the origin has been previously resolved by [12], our result is applicable to non-smooth f, such as the Sign function and the hard thresholding operator of sample covariance matrices. For this larger class of kernel functions, we obtain a new family of limiting densities, which includes the Marčenko-Pastur distribution and Wigner's semi-circle distribution as special cases.

AMS 2000 subject classifications: Primary 62H10; secondary 60F99.
Keywords and phrases: kernel matrices, limiting spectrum, Stieltjes transform, Hermite polynomials.

1. Introduction

In recent years there has been significant progress in the development and application of kernel methods in machine learning and statistical analysis of high-dimensional data [16]. These methods include kernel PCA (Principal Component Analysis), the "kernel trick" in SVM (Support Vector Machine), and non-linear dimensionality reduction [5, 7], to name a few. In such kernel methods, the input is a set of n high-dimensional data points $X_1, \ldots, X_n$ from which an n-by-n matrix is constructed, where its (i,j)-th entry is a symmetric function of X_i and X_j. Whenever the function depends merely on the inner-product $X_i^T X_j$, it is called an inner-product kernel matrix.

In this paper we study the spectral properties of an $n \times n$ symmetric random kernel matrix A whose construction is as follows. Let $X_1, \ldots, X_n$ be n i.i.d Gaussian random vectors in $\mathbb{R}^p$, where $X_i \sim \mathcal{N}(0, p^{-1} I_p)$ and I_p is the $p \times p$ identity matrix. That is, the np-many coordinates $\{(X_i)_j, 1 \le i \le n, 1 \le j \le p\}$ are i.i.d Gaussian random variables with mean 0 and variance p^{-1}. The entries

of A are defined as

$$A_{ij} = \begin{cases} f(X_i^T X_j; p), & i \neq j, \\ 0, & i = j, \end{cases} \tag{1.1}$$

where $f(\xi; p)$ is a real-valued function possibly depending on p. We will later consider another model where X_i are drawn from the uniform distribution over the unit sphere S^{p-1} in $\mathbb{R}^p$.

The study of the spectrum of large random matrices, since Wigner's semi-circle law, has been an active research area motivated by applications such as quantum physics, signal processing, numerical linear algebra, statistical inference, among others. An important result is the Marčenko-Pastur (M.P.) law [15] for the spectrum of random matrices of the form $S = XX^T$ (also known as Wishart matrices), where X is a p-by-n (complex or real) matrix with i.i.d Gaussian entries. In the "large p, large n" limit, i.e. $p, n \to \infty$ and $p/n = \gamma$ $(0 < \gamma < \infty)$, the spectral density of S converges to a deterministic limit, known as the Marčenko-Pastur distribution, which has γ as its only parameter. We refer the reader to [2], [21] and [4, Chapters 1-3] for an introduction of these topics. Notice that Wishart matrices share the non-zero eigenvalues with their corresponding Gram matrices $G = X^T X$, the latter of which, neglecting the difference at the diagonal entries, can be considered as a kernel matrix as in Eqn. (1.1) with the linear kernel function $f(\xi; p) = \xi$. Thus, the M.P. law and other results involving Wishart matrices can be translated to the Gram matrix case.

The spectrum of inner-product random kernel matrices with kernel functions that are locally smooth at the origin has been studied in [12]. It was shown that, in the limit $p, n \to \infty$ and $p/n = \gamma$,

(1) whenever f is locally C^3, the non-linear kernel matrix converges asymptotically in spectral norm to a linear kernel matrix;
(2) with less regularity of f (locally C^2), the weak convergence of the spectral density is established.

We refer to [12] and references therein for more details, including a complete review of the origins of this problem. The problem we study here is similar to the one considered in [12], except that we allow the kernel function f to belong to a much larger class of functions, in particular, f can be discontinuous at the origin.

Our main result, Thm 3.4, establishes the convergence of the spectral density of random kernel matrices under the condition that the kernel function belongs to a weighted L^2 space, is properly normalized and satisfies some additional technical conditions. The limiting spectral density is characterized by an algebraic equation, Eqn. (3.5), of its Stieltjes transform. The equation involves only three parameters, namely ν, a and γ. The parameter ν is the limit of $p \cdot Var(f(X_i^T X_j))$ and simply scales the limiting spectral density. The parameter a is the limiting coefficient of the linear term ξ in the expansion of $f(\xi; p)$ into rescaled Hermite polynomials, and has some non-trivial effect on the shape of the limiting spectral density. The result concerning the weak convergence of

the spectral density in [12], for the case where the random vectors are standard Gaussian i.e. the covariance matrix is an identity matrix, can be regarded as a special case of our result. Specifically, [12] proves that for a locally smooth kernel function, the limiting spectral density is dictated by its first-order Taylor expansion. The linear term in our rescaled Hermite expansion asymptotically coincides with the first-order term of the Taylor expansion. See also Remark 3.8 after Thm. 3.4.

Notice that the entries of the random kernel matrix are dependent. For example, the triplet of entries (i,j), (j,k) and (k,i) are mutually dependent. In the literature of random matrix theory (RMT), random matrices with dependent entries have received some attention. For example, the spectral distribution of random matrices with "finite-range" dependency among entries is studied in [3]. However, we do not find studies of this sort to be readily applicable to the analysis of the random inner-product kernel matrices considered here. We emphasize that our result only addresses the weak limit of the spectral density, while leaving many other questions about random kernel matrices unanswered. These include the analysis of the local statistics of the eigenvalues, the limiting distribution of the largest eigenvalue, and universality type questions with respect to different probability distributions for the data points.

The rest of the paper is organized as follows: in Sec. 2 we review the properties of the Stieltjes transform and the proof of the M.P. law. Sec. 3 includes the statement of our main theorem, Thm. 3.4, three examples of the kernel function and some numerical results. The proof of Thm. 3.4 is in Sec. 4. Finally, the concluding remarks, discussion and open problems are provided in Sec. 5.

Notations: For a vector X, we denote by $|X|$ its l^2 norm, i.e. for $X = (X_1, \cdots, X_p)^T$ in $\mathbb{R}^p$, $|X| = \sqrt{X_1^2 + \cdots + X_p^2}$. We write $x = \mathcal{O}(1)p^\alpha$ to indicate that $|x| \leq Cp^\alpha$ for some positive constant C and large enough p (which also implies large enough n since $p/n = \gamma$). Also, $\mathcal{O}_a(1)$ means that the constant C depends on the quantity a, and the latter is often independent of p. Throughout the paper, ζ stands for a random variable observing the standard normal distribution.

2. Review of the Stieltjes Transform and the M.P. Law

2.1. The Stieltjes Transform

For a probability measure $d\mu$ on $\mathbb{R}$, its Stieltjes transform (also known as the Cauchy transform) is defined as (see, e.g. Appendix B of [4])

$$m(z) = \int_{\mathbb{R}} \frac{1}{t-z} d\mu(t), \quad \Im(z) > 0,$$

and hence $\Im(m) > 0$. The probability density function can be recovered from its Stieltjes transform via the "inversion formula"

$$\lim_{b \to 0+} \frac{1}{\pi} \Im(m(t+ib)) = \frac{d\mu}{dt}(t), \tag{2.1}$$

where the convergence is in the weak sense.

Point-wise convergence of the Stieltjes transform implies weak convergence of the probability density (see e.g. Thm. B.9 in [4]). This is the fundamental tool that we use to establish the main result in our paper. For the n-by-n random kernel matrix A, its empirical spectral density (ESD) is defined as

$$ESD_A = \frac{1}{n}\sum_{i=1}^{n}\delta_{\lambda_i(A)}(\lambda), \tag{2.2}$$

where $\{\lambda_i(A), i = 1, \cdots, n\}$ are the n (real) eigenvalues of A. Considering ESD_A as a random probability measure on $\mathbb{R}$, we have its Stieltjes transform as

$$m_A(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda_i(A) - z} = \frac{1}{n}\mathbf{Tr}(A - zI)^{-1}, \quad \Im(z) > 0. \tag{2.3}$$

To show the convergence of ESD_A, in expectation (or in a.s. sense), it suffices to show that, for every fixed z above the real axis, $m_A(z)$ converges to the Stieltjes transform of the limiting density in expectation (or in a.s. sense).

Another convenience brought by fixed z is the uniform boundedness of many quantities. Specifically, for $z = u + iv$, $v > 0$,

$$|m_A(z)| \leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|\lambda_i(A) - z|} \leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{v} = \frac{1}{v}.$$

Also,

$$\left|\left((A - zI)^{-1}\right)_{ii}\right| \leq \frac{1}{v}, \quad 1 \leq i \leq n, \tag{2.4}$$

which follows from the spectral decomposition of A.

2.2. Proving the M.P. Law using the Stieltjes Transform

Thm. 2.1 is the version of the M.P. law for random kernel matrices with a linear kernel function. The version for Wishart matrices is well known and its proof can be found in many places, see e.g. [4, Chapter 3.3]. In this section we reproduce the proof of Thm. 2.1 in order to make the paper self-contained, and because the proof is presented in a way that will later proceed to prove our main theorem.

Theorem 2.1 (the M.P. law [15], [4]). *Suppose that $X_i \sim \mathcal{N}(0, p^{-1}I_p)$. Let A be the random kernel matrix as in Eq. (1.1) with the kernel function $f(\xi) = a\xi$ where a is a constant. As $p, n \to \infty$ and $p/n \to \gamma$, ESD_A to $\rho_I(t)$ whose Stieltjes transform satisfies the following quadratic equation*

$$-\frac{1}{m(z)} = z + a\left(1 - \frac{1}{1 + \frac{a}{\gamma}m(z)}\right). \tag{2.5}$$

The convergence of ESD_A is in the weak sense, almost surely. It has been shown that Eq. (2.5) has a unique solution with positive imaginary part (Lemma 3.11 of [4]).

Remark 2.2. The limiting spectral density $\rho_I(t)$ has an explicit expression as

$$\rho_I(t) = \frac{1}{a}\rho_{M.P.}\left(\frac{t+a}{a}; \frac{1}{\gamma}\right), \tag{2.6}$$

where $\rho_{M.P.}$ is the famous Marčenko-Pastur density

$$\rho_{M.P.}(t;y) = \left(1-\frac{1}{y}\right)^+ \delta_0(t) + \frac{\sqrt{(b(y)-t)^+(t-a(y))^+}}{2\pi y t}, \tag{2.7}$$

with $(x)^+ = \max\{x, 0\}$, $b(y) = (1+\sqrt{y})^2$, and $a(y) = (1-\sqrt{y})^2$. In literature Eq. (2.5) is sometimes called the M.P. equation. In Eq. (2.6), the rescaling by a is due to the constant a in front of the inner-product, and the shifting by a is due to our setting diagonal entries to be zero.

Remark 2.3. The limiting distribution of the largest eigenvalue is well-known for Wishart matrices, and thus applies to Gram matrices. Following the similar result for Wigner matrices by Tracy and Widom [23], the Tracy-Widom Law was established for Wishart matrices [13], and was shown to be universal for sample covariance matrices with non-Gaussian entries, see e.g. [20, 10]. As a corollary, the largest eigenvalue of the matrices studied in Thm. 2.1 converges almost surely to $b(\gamma^{-1})$, which is the right end of the support of the limiting spectral density. Since the smallest eigenvalue of a Gram matrix is always non-negative, this implies that as $p, n \to \infty, p/n = \gamma$, almost surely the spectral norm $s(A) < b(\gamma^{-1}) + \epsilon$ for any $\epsilon > 0$, which is an $\mathcal{O}_\gamma(1)$ bound.

Proof of Thm. 2.1. In two steps it can be shown that $m_A(z)$, as defined in Eq. (2.3), converges almost surely to the solution of Eq. (2.5). Without loss of generality, let $a = 1$.

Step 1. Reduce a.s. convergence to convergence of $\mathbb{E}m_A(z)$.

Lemma 2.4 (concentration of m_A at $\mathbb{E}m_A$)**.** *For the n-by-n random kernel matrix A as in Eq. (1.1), where X_i's are independent random vectors, and a fixed complex number z with $\Im(z) > 0$, we have that as $n \to \infty$,*

$$m_A(z) - \mathbb{E}m_A(z) \to 0$$

almost surely, and also

$$\mathbb{E}|m_A - \mathbb{E}m_A| \leq \mathcal{O}(1)n^{-1/2}. \tag{2.8}$$

The above lemma relies on that $\Im(z) > 0$ and that the X_i's are independent, while there is no restriction on the specific form of the kernel function, nor on the distribution of X_i. The proof (left to Appendix B) uses Burkholder's inequality and the interlacing law of eigenvalues of the minor of a symmetric matrix, combined with the observation that among all the entries of A only the k-th column/row depend on X_k.

Step 2. Convergence of $\mathbb{E}m_A(z)$. Observe that

$$\begin{aligned}\mathbb{E}m_A(z) &= \mathbb{E}\frac{1}{n}\mathbf{Tr}(A-zI)^{-1}\\ &= \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\left((A-zI)^{-1}\right)_{ii}\\ &= \mathbb{E}\left((A-zI)^{-1}\right)_{nn},\end{aligned}$$

where the last equality follows from that the rows/columns of A are exchangeable and so are those of $(A-zI)^{-1}$. By Schur complement,

$$((A-zI)^{-1})_{nn} = \frac{1}{(A_{nn}-z) - A_{\cdot,n}^T(A^{(n)}-zI_{n-1})^{-1}A_{\cdot,n}}, \tag{2.9}$$

where $A^{(n)}$ is the top left $(n-1)\times(n-1)$ minor of A, i.e. the matrix A is written in blocks as

$$A = \begin{bmatrix} A^{(n)} & A_{\cdot,n} \\ A_{\cdot,n}^T & A_{nn}\end{bmatrix},$$

and I_{n-1} is the $(n-1)\times(n-1)$ identity matrix. Notice that since $\Im(z)>0$, both $A-zI$ and $A^{(n)}-zI_{n-1}$ are invertible. Formula (2.9) can be verified by elementary linear algebra manipulation.

By Eq. (2.9) (recall that $A_{nn}=0$ from Eq. (1.1)),

$$\mathbb{E}m_A(z) = \mathbb{E}\left((A-zI)^{-1}\right)_{nn} = \mathbb{E}\frac{1}{-z - A_{\cdot,n}^T(A^{(n)}-zI_{n-1})^{-1}A_{\cdot,n}}. \tag{2.10}$$

To proceed, we condition on the choice of X_n, and write

$$X_i = \eta_i(X_n)_0 + \tilde{X}_i, \quad 1\le i\le n-1, \tag{2.11}$$

where $(X_n)_0 = \frac{X_n}{|X_n|}$ is the unit vector in the same direction of X_n, and $\tilde{X}_i$ lie in the $(p-1)$-dimensional subspace orthogonal to X_n. Due to the orthogonal invariance of the standard multivariate Gaussian distribution and that $X_1,\cdots,X_n$ are independent, we know that $\eta_i\sim\mathcal{N}(0,p^{-1})$, $\tilde{X}_i\sim\mathcal{N}(0,p^{-1}I_{p-1})$, and they are independent. Now we have

$$X_i^TX_n = \eta_i|X_n|, \quad 1\le i\le n-1, \tag{2.12}$$

and

$$X_i^TX_j = \eta_i\eta_j + \tilde{X}_i^T\tilde{X}_j, \quad 1\le i,j\le n-1, i\ne j. \tag{2.13}$$

Define $\eta=(\eta_1,\cdots,\eta_{n-1})^T$, $D_\eta = diag\{\eta_1^2,\cdots,\eta_{n-1}^2\}$ which is a diagonal matrix. Also, define

$$\tilde{A}_{ij}^{(n)} = \begin{cases}\tilde{X}_i^T\tilde{X}_j, & i\ne j,\\ 0, & i=j,\end{cases} \quad 1\le i,j\le n-1. \tag{2.14}$$

Then

$$\begin{aligned} A_{\cdot,n}^T(A^{(n)} - zI_{n-1})^{-1}A_{\cdot,n} &= |X_n|^2\eta^T(\eta\eta^T - D_\eta + \tilde{A}^{(n)} - zI_{n-1})^{-1}\eta \\ &= |X_n|^2 \cdot \frac{\eta^T(\tilde{A}^{(n)} - D_\eta - zI_{n-1})^{-1}\eta}{1 + \eta^T(\tilde{A}^{(n)} - D_\eta - zI_{n-1})^{-1}\eta} \\ &= |X_n|^2\left(1 - \frac{1}{1 + \eta^T(\tilde{A}^{(n)} - D_\eta - zI_{n-1})^{-1}\eta}\right), \end{aligned} \tag{2.15}$$

where to get the 2nd line we use the *Sherman-Morrison* formula

$$q^T(pq^T + M - zI)^{-1} = \frac{q^T(M - zI)^{-1}}{1 + q^T(M - zI)^{-1}p}, \quad \forall p, q.$$

By showing that the denominator in Eq. (2.15) is asymptotically concentrating at the value of $\mathbb{E}\tilde{m}(z)$, where $\tilde{m}(z) := \frac{1}{n}\mathbf{Tr}(\tilde{A}^{(n)} - zI_{n-1})^{-1}$, we end up with

$$\mathbb{E}\left|m_A(z) - \left(-z - \left(1 - \left(1 + \frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right)\right)^{-1}\right| \to 0.$$

The detailed derivation is left to Appendix (Lemma B.1). Notice that the probability law of η_i and $\tilde{X}_i^T\tilde{X}_j$ do not depend on the position of X_n, so we omit the conditioning on X_n when computing the probabilities and expectations. Furthermore, by Lemma B.6,

$$\mathbb{E}|m_A(z) - \tilde{m}(z)| \to 0, \tag{2.16}$$

thus

$$\mathbb{E}\tilde{m}(z) - \left(-z - \left(1 - \left(1 + \frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right)\right)^{-1} \to 0. \tag{2.17}$$

Since the quadratic Eq. (2.5) has a unique solution $m_I(z)$ with positive imaginary part, Eq. (2.17) means that

$$\mathbb{E}\tilde{m}(z) \to m_I(z).$$

At last, by Eq. (2.16), $m_I(z)$ is the limit of $\mathbb{E}m_A(z)$. □

3. Random Inner-product Kernel Matrices

3.1. Model and Notations

Let $X_1, \cdots, X_n$ be i.i.d random vectors in $\mathbb{R}^p$ and assume that $X_i \sim \mathcal{N}(0, p^{-1}I_p)$. The random kernel matrix A is defined in Eq. (1.1) with the kernel function $f(\xi; p)$, and we define

$$k(x; p) = \sqrt{p}f(\frac{x}{\sqrt{p}}; p). \tag{3.1}$$

In many cases of interest $f(\xi;p)$ does not depend on p, or the dependency is in the form of some rescaling or normalization. However, we formulate our result in a general form, keeping the dependency of $k(x;p)$ on p, and require $k(x;p)$ to satisfy certain conditions. We will see that those conditions are often satisfied in the cases of interest (Remark 3.2 and Remark 3.3).

Let X and Y be two independent random vectors distributed as $\mathcal{N}(0,p^{-1}I_p)$, and define $\xi_p = \sqrt{p}X^TY$. Denote the probability density of ξ_p by $q_p(x)$, and the L^2 spaces $\mathcal{H}_p = L^2(\mathbb{R}, q_p(x)dx)$. Let $\{P_{l,p}(x), l = 0,1,\cdots\}$ be a set of orthonormal polynomials in $\mathcal{H}_p$, that is

$$\int_{\mathbb{R}} P_{l_1,p}(x)P_{l_2,p}(x)q_p(x)dx = \delta_{l_1,l_2},$$

where $\delta_{l,k}$ equals 1 when $l = k$ and 0 otherwise. We define $P_{l,p}$ $(l \geq 0)$ using the Gram-Schmidt procedure on the monomials $\{1, x, x^2, \ldots\}$, so that $P_{0,p} = 1$, $P_{1,p} = x$ (notice that $\mathbb{E}\xi_p^2 = 1$), and $P_{l,p}$ is a polynomial of degree l. Notice that by the Central Limit Theorem, $\xi_p \to \mathcal{N}(0,1)$ in distribution as $p \to \infty$. We define $\mathcal{H}_{\mathcal{N}} = L^2(\mathbb{R}, q(x)dx)$ where $q(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. It can be shown (Lemma 4.1) that for any finite degree l, the coefficients of the polynomial $P_{l,p}(x)$ converge to those of the normalized l-degree Hermite polynomial, the latter being an orthonormal basis of $\mathcal{H}_{\mathcal{N}}$.

We formally expand $k(x;p)$ as

$$\begin{aligned} k(x;p) &= \sum_{l=0}^{\infty} a_{l,p}P_{l,p}(x), \\ a_{l,p} &= \int_{\mathbb{R}} k(x;p)P_{l,p}(x)q_p(x)dx, \end{aligned} \tag{3.2}$$

and will later explain how to understand this formal expansion. Corresponding to the l-th term in Eq. (3.2), we define the random kernel matrix A_l to be

$$(A_l)_{ij} = \begin{cases} f_l(X_i^TX_j;p), & i \neq j, \\ 0, & i = j, \end{cases} \tag{3.3}$$

where $f_l(\xi;p) = \frac{a_{l,p}}{\sqrt{p}}P_{l,p}(\sqrt{p}\xi)$.

3.2. Statement of the Main Theorem

Our main result is stated in Thm. 3.4, which establishes the weak convergence of the spectrum of random inner-product kernel matrices. The following conditions are required for $k(x;p)$:

1. **(C.Variance)** For all p, $k(x;p) \in \mathcal{H}_p$, and as $p \to \infty$, $Var(k(\xi_p;p)) = \nu_p \to \nu$ which is a finite non-negative number. We also assume that $a_{0,p} = \mathbb{E}k(\xi_p;p) = 0$ (Remark 3.5).

2. **(C.p-Uniform)** The expansion in Eq. (3.2) converges in $\mathcal{H}_p$ uniformly in p. Equivalently, let
$$k_L(x;p) = \sum_{l=0}^{L} a_{l,p} P_{l,p}(x),$$
then for any $\epsilon > 0$, there exist L and p_0 such that $\sum_{l=L+1}^{\infty} a_{l,p}^2 < \epsilon$ for $p > p_0$.
3. **(C.a_1)** As $p \to \infty$, $a_{1,p} \to a$ which is a constant.

Remark 3.1. By condition **(C.Variance)**, the integrals in Eq. (3.2) are well-defined. The requirement $\nu_p \to \nu$ can be fulfilled as long as $k(x;p) \in \mathcal{H}_p$ and is properly scaled. Notice that $\nu_p = Var(k(\xi_p;p)) = \sum_{l=1}^{\infty} a_{l,p}^2$, thus in condition **(C.a_1)**, $a^2 \leq \nu$.

Remark 3.2. When $k(x;p) = k(x)$, and if (1) $k(x) \in \mathcal{H}_{\mathcal{N}}$, and $\mathbb{E}k(\zeta) = 0$ where $\zeta \sim \mathcal{N}(0,1)$, and (2) $k(x)$ satisfies
$$\int_{\mathbb{R}} k(x)^2 |q_p(x) - q(x)| dx \to 0, \quad p \to \infty, \tag{3.4}$$
then the three conditions are satisfied (more precisely, by setting $k(x;p) = k(x) - \mathbb{E}k(\xi_p)$) and $\nu_p \to \nu_{\mathcal{N}} := \mathbb{E}k(\zeta)^2$, and $a_{1,p} \to a_{\mathcal{N}} := \mathbb{E}\zeta k(\zeta)$ (Lemma C.2). Eq. (3.4) holds as long as the singularity in the integral, say at $x = \infty$ or $k(x) = \infty$, can be controlled p-uniformly. This is the case, for example, when $k(x)$ is bounded, or when $k(x)$ is bounded on $|x| \leq R$ for any $R > 0$ and $k(x)^2$ is p-uniformly integrable at $x \to \infty$ (Lemma C.5). It is also possible for $k(x)$ to be unbounded. See Sec. 3.3 for an example of $k(x)$ that diverges at $x = 0$.

Remark 3.3. For the case that $f(\xi,p) = f(\xi)$ and $f(\xi)$ is C^1 at $\xi = 0$, see Remark 3.8.

Theorem 3.4 (the limiting spectrum of random inner-product kernel matrices)**.** *Suppose that $X_1, \cdots, X_n \sim \mathcal{N}(0, p^{-1}I_p)$ are i.i.d., and $k(x;p)$ satisfies conditions* **(C.Variance)**, **(C.p-Uniform)** *and* **(C.a_1)**. *Then, as $p, n \to \infty$ with $p/n = \gamma$, ESD_A (the empirical spectral density of the random kernel matrix A, defined in Eq. (2.2)) converges weakly to a continuous probability measure on $\mathbb{R}$ in the almost sure sense. The Stieltjes transform of the limiting spectral density is the solution of the following algebraic equation*
$$-\frac{1}{m(z)} = z + a\left(1 - \frac{1}{1 + \frac{a}{\gamma}m(z)}\right) + \frac{\nu - a^2}{\gamma} m(z), \tag{3.5}$$
which is at most cubic, and involves three parameters: ν (defined in **(C.Variance)***), a (defined in* **(C.a_1)***) and γ. Eq. (3.5) has a unique solution $m(z)$ with positive imaginary part (Lemma A.1), and the explicit formula of*
$$y(u) := \lim_{v \to 0+} \Im(m(u + iv)) \tag{3.6}$$
is given in Appendix A.

Remark 3.5. Without loss of generality we assume that $a_{0,p} = 0$. Otherwise, it results in adding to the kernel matrix a perturbation of $\frac{1}{\sqrt{p}} a_{0,p}(\mathbf{1}_n \mathbf{1}_n^T - I_n)$, where $\mathbf{1}_n$ is the all-ones vector of length n and I_n is the identity matrix. The term involving $\mathbf{1}_n \mathbf{1}_n^T$ does not change the limiting spectrum, as the limiting spectral density of a sequence of Hermitian matrices with growing size ($n \to \infty$) is invariant to a finite-rank perturbation where the rank does not depend on n, see Thm. A.43 in [4]. The term of $\frac{1}{\sqrt{p}} a_{0,p} I_n$ shifts the ESD by $a_{0,p}/\sqrt{p}$.

Remark 3.6. Recall the definition of A_l in Eq. (3.3). The limiting spectral density of A_1 is the M.P. distribution. For this case, $f(\xi; p) = a\xi$, or equivalently $k(x; p) = ax$, for some constant a. Then, the expansion in Eq. (3.2) has one term, $a_{1,p} = a$, $\nu_p = a^2$, and Eq. (3.5) is reduced to Eq. (2.5).

Remark 3.7. The limiting spectral density of A_l ($l \geq 2$) is a semi-circle. Moreover, the limiting density of any partial sum (finite or infinite) of $A_2, A_3, \cdots$ is a semi-circle, whose squared radius equals the sum of the squared radii of the semi-circle of each A_l.

Remark 3.8. For random kernel matrices where $f(\xi; p) = f(\xi)$ and is differentiable at $\xi = 0$, the limiting spectral density is the M.P. distribution. Specifically, the result in the theorem holds and $a^2 = \nu = (f'(0))^2$ (Lemma C.3). In other words, the linear term in Eq. (3.2) determines the limiting spectral density, in consistance with the result in [12].

3.3. Numerical Experiments

In this subsection we consider three examples of non-smooth kernel functions which are covered by Thm. 3.4. The first example involves the Sign function. In the limit $n \to \infty$ and p fixed, with points X_i's sampled from a fixed manifold, the results in [14] give the limiting spectrum to be that of the integral operator on the manifold associated with the kernel function. Our result shows that when the kernel is fed with high-dimensional Gaussian vectors instead of points drawn from a fixed manifold, and the dimension p increases with n, the spectrum of the matrices observes a limiting law quantified in Thm. 3.4. The second example, the hard-thresholding function, is motivated by the usage of thresholding in covariance estimation, see e.g. [6, 11] among others. The third example, kernel functions that are divergent at zero, shows the validity of the theoretical result for such singular functions.

In what follows, $k(x)$ is the rescaled and re-normalized kernel function, as defined in Eq. (3.1), and does not depend on p in all examples. In the subsequent plots, we compare the eigenvalue histogram and the theoretical limiting spectral density numerically. The eigenvalues that produce the empirical histogram are computed by MATLAB's eig function and correspond to a single realization of the random kernel matrix. The "theoretical curve" is calculated using the expression of the limiting density $y(u; a, \nu, \gamma)$ defined in Eq. (3.6) based on Thm. 3.4.

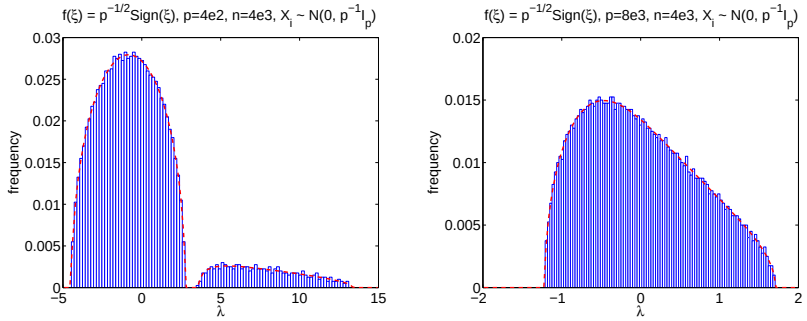

FIG 1. *Random kernel matrix with the Sign kernel, and* $X_i \sim \mathcal{N}(0, p^{-1}I_p)$. *(Left)* $p = 4\times 10^2$, $n = 4 \times 10^3$, $\gamma = p/n = 0.1$. *(Right)* $p = 8 \times 10^3$, $n = 4 \times 10^3$, $\gamma = p/n = 2$. *The blue-boundary bars are the empirical eigenvalue histograms, and the red broken-line curves are the theoretical prediction of the eigenvalue densities by Thm. 3.4.*

3.3.1. Example: $k(x) = Sign(x)$

$\text{Sign}(x) = 1$ for $x > 0$, $\text{Sign}(x) = -1$ for $x < 0$ and $\text{Sign}(x) = 0$ for $x = 0$. $k(x)$ is bounded, and according to Remark 3.2, by Lemma C.2 and Lemma C.5, $k(x)$ satisfies conditions **(C.Variance)**, **(C.p-Uniform)** and **(C.a_1)**. Meanwhile, $a = \mathbb{E}|\zeta| = \sqrt{2/\pi}$, and $\nu_p = 1$ for all p, thus $\nu = 1$.

Fig. 1 is for $X_i \sim \mathcal{N}(0, p^{-1}I_p)$. Notice that for the sign kernel, the two models $X_i \sim \mathcal{N}(0, p^{-1}I_p)$ and $X_i \sim \mathcal{U}(S^{p-1})$ result in the same probability law of the random kernel matrix. This is due to the fact that $Sign(X_i^T X_j) = Sign((X_i/|X_i|)^T(X_j/|X_j|))$ and that if $X_i \sim \mathcal{N}(0, p^{-1}I_p)$ then $X_i/|X_i| \sim \mathcal{U}(S^{p-1})$. As such, the results for $X_i \sim \mathcal{U}(S^{p-1})$ are omitted.

The following serves as a motivation for the sign kernel matrix. Consider a network of n "subjects" represented by $X_1, \ldots, X_n$ lying in $\mathbb{R}^p$. Subjects i and j have a friendship relationship if they are positively correlated, i.e., if $X_i^T X_j > 0$, and a non-friendship relationship if $X_i^T X_j < 0$. The off-diagonal entries of the n-by-n kernel matrix A are all ± 1 representing the friendship/non-friendship relationships. This model has the merit that if i and j are friends, and j and k are also friends, then chances are greater that i and k are also friends. When the X_i's are i.i.d uniformly distributed on the unit sphere in $\mathbb{R}^p$ and p is fixed, according to [14], as n grows to infinity the top p eigenvectors of the kernel matrix A converge, up to a multiplying constant and a global rotation, to the coordinates of the n data points. In this case, the eigenvalues of the sign kernel matrix converges to those of the integral operator on the manifold, and the eigenvectors recover the positioning of the subjects in the whole community from their pairwise relationships.

On the other hand, Thm. 3.4 deals with the case where the data are distributed as p-dimensional standard Gaussian, and the dimension p increases with n where $p/n \to \gamma$. For this case, the spectrum of the random kernel matrices observes a limiting law, where the dependence of the kernel function involves only the parameters ν and a.

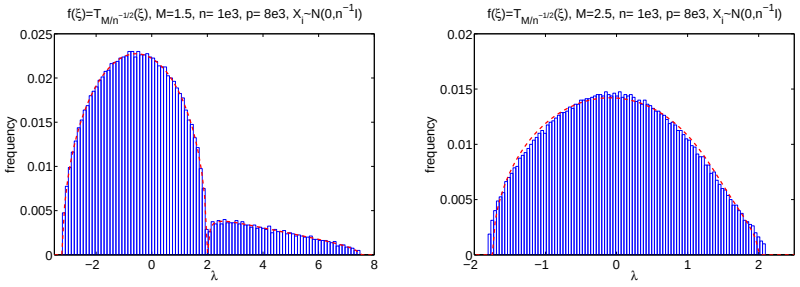

FIG 2. *Random kernel matrix where* $k(x) = T_M(x) = x\mathbf{1}_{\{|x|>M\}}$*, with* $M = 1.5$ *(left) and* $M = 2.5$ *(right).* $X_i \sim \mathcal{N}(0, n^{-1}I_p)$*, and* $p = 8 \times 10^3$*,* $n = 1 \times 10^3$*. (*p *and* n *switched from the rest of this paper.)*

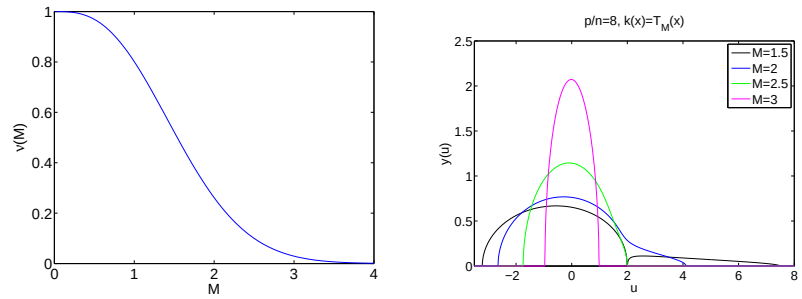

FIG 3. *(Left) The function* $\nu(M)$ *as in Eq. (3.7). (Right) The limiting spectral density of random kernel matrices where* $k(x) = T_M(x)$ *for different values of* M*. When* $M = 0$ *the density is the M.P. density, which is not shown.*

3.3.2. Example: $k(x) = T_M(x) = x\mathbf{1}_{\{|x|>M\}}$

Due to the convention that the sample covariance matrices is p-by-p, we switch the notion of n and p in this example. We consider standard normal X_i's, so the distribution of the vectors is again normal after switching the rows with the columns.

Let $X_i, \cdots, X_p \sim \mathcal{N}(0, n^{-1}I_n)$, and the kernel function be the hard-thresholding function

$$k(x) = T_M(x) = x\mathbf{1}_{\{|x|>M\}},$$

where M is a positive constant. The off-diagonal entries of the random kernel matrix A equals

$$A_{ij} = T_{\frac{M}{\sqrt{n}}}(X_i^T X_j), \quad i \neq j,$$

and A equals a sample covariance matrix of n standard Gaussian vectors hard-thresholded at $t = \frac{M}{\sqrt{n}}$, except for the diagonal entries. For this model, the analysis of the diagonal entries can be separated from that of the off-diagonal

ones, by the fact that the diagonal entries of the sample covariance matrix asymptotically concentrate at 1. One may use a union bound argument and Lemma 4.4 to show that the limiting spectral density is the same after replacing the diagonal part to be an identity matrix. This results in shifting the limiting density to the right by 1. The details of extending the analysis to including the diagonal entries are omitted here.

According to Remark 3.2, and the arguments in Lemma C.2 and Lemma C.5, $T_M(x)$ satisfies the three conditions, and

$$\begin{aligned}\nu = \nu(M) &= \mathbb{E}T_M(\zeta)^2 \\ &= 2\int_M^\infty x^2 \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \\ &= \sqrt{\frac{2}{\pi}} M e^{-M^2/2} + 2(1-\Phi(M)),\end{aligned} \tag{3.7}$$

where $\Phi(x) = \int_{-\infty}^x \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$. For this case $a = \nu$. Eq. (3.7) implies that $\nu(M) < \sqrt{(M+1)e^{-M^2/2}}$, which means that as M grows, typically beyond 2, $a = \nu \to 0$ exponentially fast. A plot of $\nu(M)$ is shown on the left of Fig. 3. This implies that the size of the support interval of the limiting density is rapidly decreasing to zero.

To be more specific, as M increases, the limiting density tends to have a semi-circle shape due to that $a^2 = \nu^2 \ll \nu$, and the radius of the semi-circle is $2\sqrt{\frac{p}{n}(\nu(M) - \nu(M)^2)} \lesssim 2\sqrt{\frac{p}{n}\nu(M)}$. Using the crucial bound that $s(A+B) \leq s(A)+s(B)$, where $s(\cdot)$ is the operator norm of the matrix, instead of considering the free convolution, we have that the right end of the support of the limiting spectral density is bounded by

$$2\sqrt{\frac{p}{n}\nu(M)} + \left(1 + \sqrt{\frac{p}{n}}\right)^2 \nu(M).$$

Based on the above analysis of the limiting density, we conjecture that, for general p and n, when $\frac{p}{n}(1+M)e^{-M^2/2} \ll 1$, the operator norm of A is bounded by

$$2\sqrt{\frac{p}{n}(M+1)e^{-M^2/2}} + \left(1 + \sqrt{\frac{p}{n}}\right)^2 (M+1)e^{-M^2/2}$$

with overwhelming probability.

The comparison of the empirical eigenvalues and the limiting density are shown in Fig. 2, where $n = 1 \times 10^3$, $p = 8 \times 10^3$, and $M = 1.5$ and 2.5 respectively. The limiting densities for different values of M are shown on the right of Fig. 3. According to Thm. 2.1, without thresholding, the right edge of the limiting density is at $(1 + \sqrt{p/n})^2 - 1 = 13.6569$. The right edge decreases to about 1 when thresholded at $M = 3$.

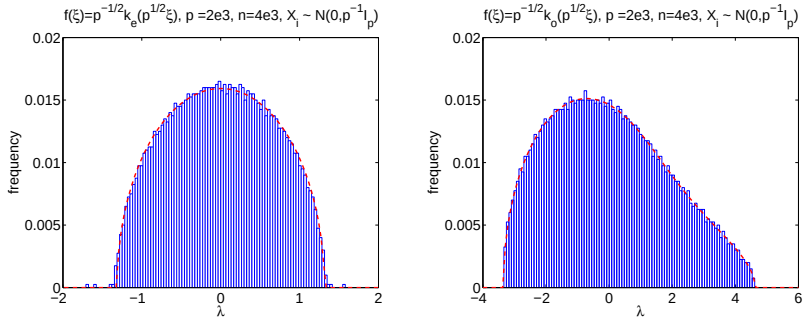

FIG 4. *Random kernel matrix where* $k(x) = k_e(x) = |x|^{-1/4} - \mathbb{E}|\zeta|^{-1/4}$ *(left) and* $k_o(x) = Sign(x)|x|^{-1/4}$ *(right).* $X_i \sim \mathcal{N}(0, p^{-1}I_p)$*, and* $p = 2 \times 10^3$*,* $n = 4 \times 10^3$*,* $\gamma = p/n = 0.5$*.*

3.3.3. Example: $k(x) = |x|^{-r}$ $(r < 1/2)$

As examples of unbounded kernel functions, we study the even function

$$k_e(x) = |x|^{-r} - \mathbb{E}|\zeta|^{-r}$$

and the odd function

$$k_o(x) = \text{Sign}(x)|x|^{-r},$$

where $r < 1/2$ so as to guarantee the integrability of $k(x)^2$ at $x = 0$.

Notice that for both cases, $|k(x)|$ is bounded on $\{|x| > R\}$ for any $R > 0$, and diverge at $x = 0$. Meanwhile, $k(x)^2 = |x|^{-2r}$ is integrable at $x = 0$, and with the fact that $q_p(x) \le q_p(0) \to q(0) = 1/\sqrt{2\pi}$, Eq. (3.4) still holds. Thus by Lemma C.2, Thm. 3.4 applies to both k_e and k_o. By

$$\mathbb{E}|\zeta|^{-r} = \sqrt{\frac{2}{\pi}} 2^{-(r+1)/2} \Gamma(\frac{1-r}{2})$$

where $\Gamma(\cdot)$ is the Gamma function, and similarly for $\mathbb{E}|\zeta|^{-2r}$, the constants ν and a for both k_e and k_o can be explicitly computed. For k_e, $\nu = Var(|\zeta|^{-r})$ and $a = 0$. For k_o, $\nu = |\zeta|^{-2r}$, and

$$a = \mathbb{E}|\zeta|^{1-r} = \sqrt{\frac{2}{\pi}} 2^{-r/2} \Gamma(1 - \frac{r}{2}).$$

The numerical results for $r = 1/4$ with $X_i \sim \mathcal{N}(0, p^{-1}I_p)$ are shown in Fig. 4. The empirical histograms for $X_i \sim \mathcal{U}(S^{p-1})$ look almost identical and are therefore omitted. In the left panel of Fig. 4, the empirical spectral density is close to a semi-circle, as our theory predicts.

4. Proof of the Main Theorem

The model and the notations are the same as in Sec. 3.1. The proof of Thm. 3.4 is provided in Sec. 4.3. Prior to the proof, in Sec. 4.1 we review some useful

properties of Hermite polynomials, and in Sec. 4.2 we introduce an asymptotic upper bound for the expected value of the spectral norm of random kernel matrices. The other model $X_i \sim \mathcal{U}(S^{p-1})$ is analyzed in Sec. 4.4, where it is shown that the result of Thm. 3.4 still holds.

4.1. Orthonormal Polynomials

4.1.1. $\mathcal{H}_{\mathcal{N}}$ and normalized Hermite polynomials

Define the normalized Hermite polynomials as

$$h_l(x) = \frac{1}{\sqrt{l!}} H_l(x), \quad l = 0, 1, \cdots \tag{4.1}$$

where $H_l(x)$ is the l-degree Hermite polynomial, satisfying

$$\int_{\mathbb{R}} H_{l_1}(x) H_{l_2}(x) q(x) dx = \delta_{l_1, l_2} \cdot l_1!.$$

Thus, $\{h_l(x), l = 0, 1, \cdots\}$ form an orthonormal basis of $\mathcal{H}_{\mathcal{N}}$. The explicit formula of H_l is [1]

$$H_l(x) = l! \sum_{k=0}^{\lfloor l/2 \rfloor} (-\frac{1}{2})^k \frac{1}{k!(l-2k)!} x^{l-2k}.$$

Also, the derivative of $H_l(x)$ satisfies the recurrence relation $H_l{}'(x) = l H_{l-1}(x)$ for $l \geq 1$, and as a result,

$$h_l{}'(x) = \sqrt{l} h_{l-1}(x). \tag{4.2}$$

4.1.2. $\mathcal{H}_p$ and $P_{l,p}(x)$

Recall that the random variable ξ_p converges in distribution to $\mathcal{N}(0,1)$ as $p \to \infty$. Meanwhile, the moments of ξ_p approximate those of $\mathcal{N}(0,1)$:

$$\mathbb{E}\xi_p^k = \begin{cases} (k-1)!! + \mathcal{O}_k(1) p^{-1}, & k \text{ even}; \\ 0, & k \text{ odd}. \end{cases} \tag{4.3}$$

Eq. (4.3) is verified by directly computing the moments of ξ_p using the model, i.e. $\xi_p = \sqrt{p} X^T Y$ and X and Y are independently distributed as $\mathcal{N}(0, p^{-1} I_p)$. Eq. (4.3) implies the asymptotic consistency between $P_{l,p}$ and h_l, which is the statement of the following lemma.

Lemma 4.1 (convergence of $P_{l,p}$ to h_l)**.** *Let $\{P_{l,p}, l = 0, 1, \cdots\}$ be the orthonormal polynomials of $L^2(\mathbb{R}, d\mu_p)$, where μ_p is a sequence of probability measures.*

Suppose that the moments of μ_p approximate those of $\mathcal{N}(0,1)$ in the sense that, for every fixed k,

$$\int_{\mathbb{R}} x^k d\mu_p(x) = \mathbb{E}\zeta^k + \mathcal{O}_k(1)p^{-1}.$$

Then, for every fixed degree l,

$$P_{l,p}(x) = h_l(x) + \sum_{j=0}^{l} (\delta_{l,p})_j x^j,$$

where $(\delta_{l,p})_j$ satisfy

$$\max_{0\le j\le l} |(\delta_{l,p})_j| < \mathcal{O}_l(1)p^{-1}.$$

The proof of Lemma 4.1 follows from the fact that the coefficients of the l-degree orthogonal polynomials are decided by up to the first $2l$ moments.

One consequence of Lemma 4.1 is that as $p \to \infty$

$$|P_{l,p}(x)| \le \mathcal{O}_l(1)M^l, \quad |x| \le M, \tag{4.4}$$

as the coefficients of $P_{l,p}(x)$ for each l converge to those of $h_l(x)$. Also, Eq. (4.2) leads to

$$\begin{aligned} P'_{l,p}(x) &= \sqrt{l}P_{l-1,p}(x) + \mathcal{O}_l(1)M^{l-1}p^{-1}, \\ P''_{l,p}(x) &= \sqrt{l(l-1)}P_{l-2,p}(x) + \mathcal{O}_l(1)M^{l-2}p^{-1}, \end{aligned} \quad l \ge 2. \tag{4.5}$$

Another consequence is the "asymptotic consistency between the $P_{l,p}$-expansion and the Hermite-expansion" in their first finite-many terms (Lemma C.1). This further implies that conditions **(C.Variance)**, **(C.p-Uniform)** and **(C.a_1)** are satisfied by a large class of kernel functions (Remark 3.2).

4.2. Spectral Norm Bound

The following lemma gives an upper bound for the expectation of the spectral norm of random kernel matrices whose rescaled kernel function $k(x;p)$ is $P_{l,p}(x)$ (defined in Sec. 4.1) for some l. The method is by analyzing the 4th moment of the random matrix.

Lemma 4.2 (bounding mean spectral norm by the 4th moment)**.** *Let A be the random kernel matrix defined in Eq. (1.1) with the kernel function $f(\xi;p) = p^{-1/2}P_{l,p}(\sqrt{p}\xi)$, $l \ge 1$, where $P_{l,p}$ is defined as in Sec. 4.1. X_i's are i.i.d. distributed as $\mathcal{N}(0, p^{-1}I_p)$. Then, as $p, n \to \infty$, $p/n = \gamma$,*

$$\mathbb{E}s(A) \le \mathcal{O}_{l,\gamma}(1)n^{1/4}.$$

Remark 4.3. The asymptotic concentration of the largest eigenvalue at its mean value is quantified by the Tracy-Widom Law for Gaussian ensembles [23] (see also [2, Chapter 3]) and a large class of Wigner-type matrices (see e.g. [19],

[22] and references therein), as well as Wishart-type matrices (Remark 2.3). For random kernel matrices studied in Lemma 4.2, the spectral norm is conjectured to be $\mathcal{O}(1)$, and see more in Sec. 5. However, the bound provided by Lemma 4.2, though not tight, is sufficient for the proof of our main theorem.

Proof of Lemma 4.2 . Let $\{\lambda_i, 1 \leq i \leq n\}$ be the eigenvalues of A. Since

$$s(A)^4 \leq \sum_{i=1}^{n} \lambda_i^4 = \mathbf{Tr}(A^4) = \sum_{i,j,k,l} A_{ij}A_{jk}A_{kl}A_{li},$$

we have

$$\mathbb{E}s(A) \leq (\mathbb{E}s(A)^4)^{1/4} \leq (\sum_{i,j,k,l} \mathbb{E}A_{ij}A_{jk}A_{kl}A_{li})^{1/4}. \tag{4.6}$$

In Eq. (4.6) for $\mathbb{E}A_{ij}A_{jk}A_{kl}A_{li}$ to be non-zero, in $\{i,j,k,l\}$ the neighboring indices must differ since $A_{ii} = 0$. Meanwhile, by Eq. (4.3) and Lemma 4.1, for any fixed l,

$$\mathbb{E}P_{l,p}(\sqrt{p}\xi_{12})^4 = \mathcal{O}_l(1).$$

We have the following cases:

1. $i = k$, $j = l$:

$$\begin{aligned}\mathbb{E}A_{ij}A_{jk}A_{kl}A_{li} &= \mathbb{E}A_{12}^4 \\ &= p^{-2}\mathbb{E}P_{l,p}(\sqrt{p}\xi_{12})^4 \\ &= \mathcal{O}_l(1)p^{-2},\end{aligned}$$

2. $i = k$, $j \neq l$ or $i \neq k$, $j = l$:

$$\begin{aligned}\mathbb{E}A_{ij}A_{jk}A_{kl}A_{li} &= \mathbb{E}A_{12}^2A_{13}^2 \\ &= p^{-2}\mathbb{E}P_{l,p}(\sqrt{p}\xi_{12})^2P_{l,p}(\sqrt{p}\xi_{13})^2 \\ &\leq p^{-2}\mathbb{E}P_{l,p}(\sqrt{p}\xi_{12})^4 \quad \text{(by Cauchy-Schwarz inequality)} \\ &= \mathcal{O}_l(1)p^{-2}.\end{aligned}$$

3. $i \neq k$, $l \neq j$: when $l = 1$, $\mathbb{E}A_{ij}A_{jk}A_{kl}A_{li} = p^{-3}$. When $l \geq 2$, we have the following estimate (Lemma D.1)

$$\mathbb{E}A_{ij}A_{jk}A_{kl}A_{li} = \mathcal{O}_l(1)p^{-4}.$$

As a result, when $l = 1$,

$$\begin{aligned}&\sum_{i,j,k,l} \mathbb{E}A_{ij}A_{jk}A_{kl}A_{li} \\ &\leq n^2\mathcal{O}(1)p^{-2} + 2n^3\mathcal{O}(1)p^{-2} + n^4p^{-3} \\ &= \mathcal{O}_\gamma(1)n + \mathcal{O}_\gamma(1),\end{aligned}$$

and when $l \geq 2$,

$$\begin{aligned}&\sum_{i,j,k,l} \mathbb{E}A_{ij}A_{jk}A_{kl}A_{li}\\ &\leq n^2\mathcal{O}_l(1)p^{-2} + 2n^3\mathcal{O}_l(1)p^{-2} + n^4\mathcal{O}_l(1)p^{-4}\\ &= \mathcal{O}_{l,\gamma}(1)n + \mathcal{O}_{l,\gamma}(1).\end{aligned}$$

Combining the above estimates with Eq. (4.6) leads to the bound wanted. □

4.3. Proof of Thm. 3.4

Proof of Thm. 3.4. Same as in Sec. 2.2, it suffices to show the mean convergence of the Stieltjes transform. Specifically, we want to show that for a fixed $z = u+iv$, $\mathbb{E}m_A(z)$ converges to the unique solution of Eq. (3.5). Recall that the expansion Eq. (3.2) converges p-uniformly in $\mathcal{H}_p$, and we first reduce the general case to that where the expansion has finite many terms.

Step 1. Reduction to the case of finite expansion up to order L.

Denote the truncated kernel function up to finite order L by $f_L(\xi;p) = p^{-1/2}k_L(\sqrt{p}\xi;p)$ where (recall that $a_{0,p} = 0$ by Remark 3.5)

$$k_L(x;p) = \sum_{l=1}^{L} a_{l,p}P_{l,p}(x).$$

Let $m_A(z)$ and $m_L(z)$ be the Stieltjes transforms of the random kernel matrix with the kernel function $f(\xi;p)$ and $f_L(\xi;p)$, respectively. For a fixed z, define

$$RHS(m;a,\nu) = \left(-z - a\left(1 - \frac{1}{1+\frac{a}{\gamma}m}\right) - \frac{\nu - a^2}{\gamma}m\right)^{-1}. \tag{4.7}$$

The goal is to show that, as $p,n \to \infty$ with $p/n = \gamma$, $\mathbb{E}m_A$ converges to the solution of Eq. (3.5) which can be rewritten as $m = RHS(m;a,\nu)$, and it suffices to show that

$$|\mathbb{E}m_A - RHS(\mathbb{E}m_A;a,\nu)| \to 0. \tag{4.8}$$

We need the following lemma, whose proof is left to Appendix D:

Lemma 4.4 (stability of the Stieltjes transform to L^2 perturbation in the kernel function)**.** *Suppose that X_i $(i = 1,\cdots,n)$ are i.i.d random vectors, and the two functions $f_A(\xi;p)$ and $f_B(\xi;p)$ satisfy that with large p*

$$\mathbb{E}(f_A(X^TY;p) - f_B(X^TY;p))^2 \leq \epsilon p^{-1},$$

where X and Y are two independent random vectors distributed in the same way as X_i's, and ϵ is some positive constant. Let A be the n-by-n random kernel matrix with the kernel function $f_A(\xi;p)$, and B with $f_B(\xi;p)$. Also, let m_A and m_B be the Stieltjes Transforms of A and B respectively. Then for a fixed z,

$$\mathbb{E}|m_A(z) - m_B(z)| \leq \mathcal{O}(1)\sqrt{\epsilon}.$$

By condition **(C.p-Uniform)**, for arbitrary $\epsilon > 0$, there exists some $L = L(\epsilon)$, so that $\mathbb{E}(k(\xi_p; p) - k_{L(\epsilon)}(\xi_p; p))^2 \leq \epsilon^2$ for all p, and then

$$\mathbb{E}(f(X^T Y; p) - f_{L(\epsilon)}(X^T Y; p))^2 \leq \epsilon^2 p^{-1}.$$

By Lemma 4.4,

$$|\mathbb{E}m_A(z) - \mathbb{E}m_{L(\epsilon)}(z)| \leq \mathbb{E}|m_A(z) - m_{L(\epsilon)}(z)| \leq \mathcal{O}(1)\epsilon.$$

If in addition we can show that, for any fixed L and some sequence of $a_L(p)$ and $\nu_L(p)$,

$$\begin{aligned} &|\mathbb{E}m_L - RHS(\mathbb{E}m_L; a_L(p), \nu_L(p))| \to 0, \\ &a_L(p) \to a, \quad \nu_L(p) \to \nu, \end{aligned} \tag{4.9}$$

then Eq. (4.8) holds asymptotically.

Step 2. Convergence of $\mathbb{E}m_L(z)$ for finite L.

With slight abuse of notation, we denote the random kernel matrix with kernel function $f_L(\xi; p)$ by A. Its Stieltjes transform is denoted by $m_L(z)$. In what follows we sometimes drop the dependence on p and write $f_L(\xi; p)$ as $f_L(\xi)$, and similar for other functions.

Recall that

$$\begin{aligned} \mathbb{E}m_L(z) &= \mathbb{E}\left((A - zI)^{-1}\right)_{nn} \\ &= \mathbb{E}(-z - A_{\cdot,n}^T (A^{(n)} - zI_{n-1})^{-1} A_{\cdot,n})^{-1}. \end{aligned} \tag{4.10}$$

Notations as in Eq. (2.11, 2.12, 2.13), we have

$$\begin{aligned} &A_{\cdot,n} = f_{(1)} + f_{(2)}, \\ &f_{(1)} := a_{1,p}|X_n|\eta, \\ &f_{(2)} := (f_{>1}(\xi_{1n}), \cdots, f_{>1}(\xi_{n-1,n}))^T, \end{aligned} \tag{4.11}$$

where $\xi_{in} = |X_n|\eta_i$ for $1 \leq i \leq n-1$, $\eta := (\eta_1, \cdots, \eta_{n-1})^T$, and $f_{>1}(\xi) := \frac{1}{\sqrt{p}} \sum_{l=2}^{L} a_{l,p} P_{l,p}(\sqrt{p}\xi)$. The off-diagonal entries of $A^{(n)}$ are

$$A_{ij}^{(n)} = f_L(X_i^T X_j) = f_L(\eta_i \eta_j + \tilde{\xi}_{ij}), \quad 1 \leq i, j \leq n-1, i \neq j,$$

where $\tilde{\xi}_{ij} = \tilde{X}_i^T \tilde{X}_j$.

The typical magnitude of η_i and $\tilde{\xi}_{ij}$ is $p^{-1/2}$, and specifically, we have the large probability set Ω_δ defined as

$$\Omega_\delta = \{|\eta_i| < \delta, |\tilde{\xi}_{ij}| < \delta, \left||X_n|^2 - 1\right| < \sqrt{2}\delta, 1 \leq i, j \leq n-1, i \neq j\}, \tag{4.12}$$

where $\delta = \frac{M}{\sqrt{p}}$, $M = \sqrt{20 \ln p}$. By Lemma D.3, $\mathbf{Pr}(\Omega_\delta^c) \leq \mathcal{O}(1)p^{-7}$. On Ω_δ,

$$\begin{aligned} f_L(\eta_i \eta_j + \tilde{\xi}_{ij}) = a_{1,p}\eta_i \eta_j + a_{1,p}\tilde{\xi}_{ij} \\ + f_{>1}(\tilde{\xi}_{ij}) + f_{>1}'(\tilde{\xi}_{ij})\eta_i \eta_j + t_{ij}, \end{aligned}$$

where

$$t_{ij} = \frac{1}{2} f_{>1}^{''}(\theta_{ij})(\eta_i \eta_j)^2.$$

Recall that $f_{>1}(\xi) = \frac{1}{\sqrt{p}} \sum_{l=2}^{L} a_{l,p} P_{l,p}(\sqrt{p}\xi)$, and by Eq. (4.5),

$$f_{>1}^{'}(\xi) = \sum_{l=2}^{L} a_{l,p}(\sqrt{l} P_{l-1,p}(\sqrt{p}\xi) + \mathcal{O}_l(1) M^{l-1} p^{-1}), \tag{4.13}$$

and

$$f_{>1}^{''}(\xi) = \sqrt{p} \sum_{l=2}^{L} a_{l,p}(\sqrt{l(l-1)} P_{l-2,p}(\sqrt{p}\xi) + \mathcal{O}_l(1) M^{l-2} p^{-1}). \tag{4.14}$$

We define

$$\begin{aligned} \tilde{A}_{ij}^{(n)} &= a_{1,p}\tilde{\xi}_{ij} + f_{>1}(\tilde{\xi}_{ij}), \\ \tilde{F}_{ij} &= \frac{1}{\sqrt{p}} \sum_{l=2}^{L} a_{l,p} \sqrt{l} P_{l-1,p}(\sqrt{p}\tilde{\xi}_{ij}), \end{aligned} \qquad i \neq j,$$

and set the diagonal entries to be zeros for both $\tilde{A}^{(n)}$ and $\tilde{F}$, then

$$A^{(n)} = \tilde{A}^{(n)} + a_{1,p}(\eta\eta^T - D_\eta) + \sqrt{p} W \tilde{F} W + T,$$

where T is Hermitian, $T_{ii} = 0$ and for $i \neq j$, $T_{ij} = t_{ij} + \eta_i \eta_j (f_{>1}^{'}(\tilde{\xi}_{ij}) - \sqrt{p}\tilde{F}_{ij})$; $W := \mathrm{diag}\{\eta_1, \cdots, \eta_{n-1}\}$. We have (recall that $\sum_{l=1}^{L} a_{l,p}^2$ is bounded by some $\mathcal{O}(1)$ constant for all p, by Remark 3.1)

1. Since θ_{ij} is between $\tilde{\xi}_{ij}$ and $\tilde{\xi}_{ij} + \eta_i\eta_j$, and both $\tilde{\xi}_{ij}$ and η_i are bounded in magnitude by $\delta = p^{-1/2}M$, then $|\theta_{ij}| \leq \delta + \delta^2 \leq 1.01\delta = p^{-1/2}1.01M$. Thus, by Eq. (4.14, 4.4), $|f_{>1}^{''}(\theta_{ij})| \leq \sqrt{p}\mathcal{O}_L(1)M^{L-2}$, and then $|t_{ij}| \leq \mathcal{O}_L(1)M^{L+2}p^{-3/2}$. Meanwhile, Eq. (4.13) implies that

   $$|f_{>1}^{'}(\tilde{\xi}_{ij}) - \sqrt{p}\tilde{F}_{ij}||\eta_i||\eta_j| \leq \mathcal{O}_L(1)M^{L-1}p^{-1} \cdot M^2 p^{-1} = \mathcal{O}_L(1)M^{L+1}p^{-2}.$$

   Thus

   $$\begin{aligned} |T_{ij}| &\leq \mathcal{O}_L(1)M^{L+2}p^{-3/2} + \mathcal{O}_L(1)M^{L+1}p^{-2} \\ &= \mathcal{O}_L(1)M^{L+2}p^{-3/2}. \end{aligned} \tag{4.15}$$

   As a result,

   $$\begin{aligned} s(T - a_{1,p}D_\eta) \cdot \mathbf{1}_{\Omega_\delta} &\leq s(T) \cdot \mathbf{1}_{\Omega_\delta} + |a_{1,p}|\delta^2 \\ &= \mathcal{O}_L(1)M^{L+2}p^{-1/2} + \mathcal{O}(1)M^2p^{-1} \\ &= \mathcal{O}_L(1)M^{L+2}p^{-1/2}. \end{aligned} \tag{4.16}$$

2. $\tilde{F}$ can be written as $\sum_{l=1}^{L-1} a_{l,p}\sqrt{l}\tilde{F}_l$, where Lemma 4.2 applies to each $\tilde{F}_l$, and the coefficients $a_{l,p}$ for $1 \le l \le L-1$ are uniformly bounded by some constant since $\sum_{l=1}^{L} a_{l,p}^2 = \nu_p \to \nu$ by Condition **(C.Variance)**. Thus we have

$$\mathbb{E}s(\tilde{F}) \le \sum_{l=2}^{L} \sqrt{l}\mathcal{O}_l(1)p^{1/4} = \mathcal{O}_L(1)p^{1/4}, \tag{4.17}$$

and as a result,

$$\mathbb{E}s(\sqrt{p}W\tilde{F}W)\cdot \mathbf{1}_{\Omega_\delta} \le M^2 p^{-1/2}\mathbb{E}s(\tilde{F}) \le \mathcal{O}_L(1)M^2 p^{-1/4}. \tag{4.18}$$

Now we break the quantity $A_{\cdot,n}^T(A^{(n)}-zI_{n-1})^{-1}A_{\cdot,n}$ into the following pieces: define $\hat{A}^{(n)} = a_{1,p}\eta\eta^T + \tilde{A}^{(n)}$, and recall that $A_{\cdot,n} = f_{(1)} + f_{(2)}$ as defined in Eq. (4.11),

$$\begin{aligned} A_{\cdot,n}^T(A^{(n)} - zI_{n-1})^{-1}A_{\cdot,n} =& A_{\cdot,n}^T(\hat{A}^{(n)} - zI_{n-1})^{-1}A_{\cdot,n} - A_{\cdot,n}^T(A^{(n)} - zI_{n-1})^{-1} \\ & \cdot(\sqrt{p}W\tilde{F}W + T - a_{1,p}D_\eta)(\hat{A}^{(n)} - zI_{n-1})^{-1}A_{\cdot,n} \\ =& f_{(1)}^T(\hat{A}^{(n)} - zI_{n-1})^{-1}f_{(1)} \\ & + f_{(2)}^T(\hat{A}^{(n)} - zI_{n-1})^{-1}f_{(2)} + r_2 - r_1 \end{aligned} \tag{4.19}$$

where

$$\begin{aligned} r_2 &= 2f_{(1)}^T(\hat{A}^{(n)} - zI_{n-1})^{-1}f_{(2)}, \\ r_1 &= A_{\cdot,n}^T(A^{(n)} - zI_{n-1})^{-1}(\sqrt{p}W\tilde{F}W + T - a_{1,p}D_\eta) \\ &\quad \cdot(\hat{A}^{(n)} - zI_{n-1})^{-1}A_{\cdot,n}. \end{aligned} \tag{4.20}$$

For r_2,

$$\begin{aligned} r_2 &= 2a_{1,p}|X_n|\eta^T(\hat{A}^{(n)} - zI_{n-1})^{-1}f_{(2)} \\ &= 2a_{1,p}f_{(2)}^T(\hat{A}^{(n)} - zI_{n-1})^{-1}(|X_n|\eta) \\ &= 2a_{1,p}\{f_{(2)}^T(\tilde{A}^{(n)} - zI_{n-1})^{-1}(|X_n|\eta) \\ &\quad - f_{(2)}^T(\tilde{A}^{(n)} - zI_{n-1})^{-1}a_{1,p}\eta\eta^T(\hat{A}^{(n)} - zI_{n-1})^{-1}(|X_n|\eta)\} \\ &:= 2a_{1,p}(r_{2,1} - r_{2,2}), \end{aligned} \tag{4.21}$$

and by moment method we can show that (Lemma D.4)

$$\mathbb{E}|r_2|\cdot\mathbf{1}_{\Omega_\delta} \le \mathcal{O}_L(1)M^2p^{-1/2}. \tag{4.22}$$

To bound r_1, we restrict ourselves to Ω_δ where $||A_{\cdot,n}||^2 = \sum_{i=1}^{n-1} f_L(\xi_{in})^2 \le \mathcal{O}_L(1)M^L$, and with Eq. (4.18, 4.16)

$$\begin{aligned} \mathbb{E}|r_1|\cdot\mathbf{1}_{\Omega_\delta} &\le \mathbb{E}(s(\sqrt{p}W\tilde{F}W) + s(T - a_{1,p}D_\eta))||A_{\cdot,n}||^2\cdot\mathbf{1}_{\Omega_\delta} \\ &\le \mathcal{O}_L(1)M^L\mathbb{E}(s(\sqrt{p}W\tilde{F}W) + s(T - a_{1,p}D_\eta)) \\ &= \mathcal{O}_L(1)M^L(\mathcal{O}_L(1)M^2p^{-1/4} + \mathcal{O}_L(1)M^{L+2}p^{-1/2}) \\ &= \mathcal{O}_L(1)M^{2L+2}p^{-1/4}. \end{aligned} \tag{4.23}$$

Furthermore, as in Sec. 2.2, we can compute the first term in Eq. (4.19):

$$\begin{aligned} f_{(1)}^T(\hat{A}^{(n)} - zI_{n-1})^{-1} f_{(1)} &= |X_n|^2 a_{1,p}^2 \eta^T (\hat{A}^{(n)} - zI_{n-1})^{-1}\eta \\ &= |X_n|^2 a_{1,p}\left(1 - (1 + a_{1,p}\eta^T(\tilde{A}^{(n)} - zI_{n-1})^{-1}\eta)^{-1}\right) \\ &= |X_n|^2 a_{1,p}\left(1 - (1 + a_{1,p}(\gamma^{-1}\mathbb{E}\tilde{m}(z) + \gamma^{-1}\tilde{r} + r_{(1),2}))^{-1}\right), \end{aligned}$$

where $\tilde{m}(z) = \frac{1}{n-1}\mathbf{Tr}(\tilde{A}^{(n)} - zI_{n-1})^{-1}$, and

1. $\tilde{r} = \tilde{m}(z) - \mathbb{E}\tilde{m}(z)$, $\mathbb{E}|\tilde{r}| \leq \mathcal{O}(1)n^{-1/2}$ by Lemma 2.4;
2. The term
$$r_{(1),2} = \eta^T(\tilde{A}^{(n)} - zI_{n-1})^{-1}\eta - \frac{1}{p}\mathbf{Tr}(\tilde{A}^{(n)} - zI_{n-1})^{-1}$$
is similar to r_2 in Lemma B.1 and satisfies $\mathbb{E}|r_{(1),2}| \leq \mathcal{O}(1)p^{-1/2}$.

Going through a process similar to that in Lemma B.1 to bound the denominators, including

1. introducing a large probability set
$$\Omega_{(1)} := \{|\tilde{r}| \leq p^{-1/4}, |r_{(1),2}| \leq p^{-1/4}\}, \quad \mathbf{Pr}(\Omega_{(1)}^c) \leq \mathcal{O}(1)p^{-1/4},$$
so as to bound $|(1 + a_{1,p}\gamma^{-1}\mathbb{E}\tilde{m}(z))^{-1}|$ on $\Omega_\delta \cap \Omega_{(1)}$ by $\mathcal{O}(1)M^2$,
2. making use of that $|(1 + a_{1,p}\eta^T(\tilde{A}^{(n)} - zI_{n-1})^{-1}\eta)^{-1}|$ on Ω_δ is bounded by $\mathcal{O}(1)M^2$,

we have

$$f_{(1)}^T(\hat{A}^{(n)} - zI_{n-1})^{-1} f_{(1)} = a_{1,p}\left(1 - (1 + \frac{a_{1,p}}{\gamma}\mathbb{E}\tilde{m}(z))^{-1}\right) + r_{(1)}, \tag{4.24}$$

where

$$\mathbb{E}|r_{(1)}| \cdot \mathbf{1}_{\Omega_\delta \cap \Omega_{(1)}} \leq \mathcal{O}(1)M^4 p^{-1/2}. \tag{4.25}$$

We turn to compute the second term in Eq. (4.19). We have

$$\begin{aligned} f_{(2)}^T(\hat{A}^{(n)} - zI_{n-1})^{-1} f_{(2)} &= f_{(2)}^T(\tilde{A}^{(n)} - zI_{n-1})^{-1} f_{(2)}^T \\ &\quad - f_{(2)}^T(\tilde{A}^{(n)} - zI_{n-1})^{-1} a_{1,p}\eta\eta^T(\hat{A}^{(n)} - zI_{n-1})^{-1} f_{(2)} \\ &= \frac{\nu_{>1,p}}{\gamma}\mathbb{E}\tilde{m}(z) + \frac{\nu_{>1,p}}{\gamma}\tilde{r} + r_{(2),2} - r_{(2),3} \end{aligned} \tag{4.26}$$

where

$$\nu_{>1,p} = \mathbb{E}(f_{(2)})_i^2 = \mathbb{E}f_{>1}(\xi_{in})^2 = \nu_p - a_{1,p}^2,$$

and

$$\begin{aligned} r_{(2),2} &= f_{(2)}^T(\tilde{A}^{(n)} - zI_{n-1})^{-1} f_{(2)}^T - \frac{\nu_{>1,p}}{p}\mathbf{Tr}(\tilde{A}^{(n)} - zI_{n-1})^{-1}, \\ r_{(2),3} &= f_{(2)}^T(\tilde{A}^{(n)} - zI_{n-1})^{-1} a_{1,p}\eta\eta^T(\hat{A}^{(n)} - zI_{n-1})^{-1} f_{(2)} \\ &= a_{1,p}(\eta^T(\hat{A}^{(n)} - zI_{n-1})^{-1} f_{(2)}) r_{2,1}. \end{aligned}$$

For $r_{(2),2}$, by a moment method argument similar to the first part in the proof of Lemma D.4, we have

$$\mathbb{E}|r_{(2),2}| \leq \mathcal{O}_L(1)p^{-1/2}. \tag{4.27}$$

To bound $r_{(2),3}$, we restrict ourselves to Ω_δ, where

$$|f_{(2)}(\xi_{in})| \leq \mathcal{O}_L(1)M^L p^{-1/2}, \quad |\eta_i| \leq Mp^{-1/2}, \quad 1 \leq i \leq n-1,$$

thus

$$\begin{aligned}
&|a_{1,p}\eta^T(\hat{A}^{(n)} - zI_{n-1})^{-1}f_{(2)}| \cdot \mathbf{1}_{\Omega_\delta} \\
&\leq \mathcal{O}(1)s((\hat{A}^{(n)} - zI_{n-1})^{-1})||\eta|| \cdot ||f_{(2)}|| \\
&\leq \frac{\mathcal{O}(1)}{v}\sqrt{\mathcal{O}(1)M^2}\sqrt{\mathcal{O}_L(1)M^{2L}} = \mathcal{O}_L(1)M^{L+1},
\end{aligned}$$

and then

$$\begin{aligned}
\mathbb{E}|r_{(2),3}| \cdot \mathbf{1}_{\Omega_\delta} &= \mathbb{E}|r_{2,1}||a_{1,p}(\eta^T(\hat{A}^{(n)} - zI_{n-1})^{-1}f_{(2)})| \cdot \mathbf{1}_{\Omega_\delta} \\
&\leq \mathcal{O}_L(1)M^{L+1}\mathbb{E}|r_{2,1}| \\
&\leq \mathcal{O}_L(1)M^{L+1}p^{-1/2}.
\end{aligned} \tag{4.28}$$

Now puting Eq. (4.19,4.23,4.22,4.24,4.25,4.26,4.27,4.28) together, we have

$$\begin{aligned}
&|\mathbb{E}m_L(z) - \mathrm{RHS}(\mathbb{E}\tilde{m}(z), a_{1,p}, \nu_p)| \\
&\leq \frac{2}{v}\mathbf{Pr}\{(\Omega_\delta \cap \Omega_{(1)})^c\} \\
&\quad + \left|\mathbb{E}\frac{1}{-z - A^T_{\cdot,n}(A^{(n)} - zI_{n-1})^{-1}A_{\cdot,n}} - \mathrm{RHS}(\mathbb{E}\tilde{m}(z), a_{1,p}, \nu_p)\right| \cdot \mathbf{1}_{\Omega_\delta \cap \Omega_{(1)}} \\
&\leq \frac{2}{v}\mathbf{Pr}\{(\Omega_\delta \cap \Omega_{(1)})^c\} \\
&+ \frac{2}{v}\mathbb{E}(|r_1| + |r_2| + |r_{(1)}| + |\nu_{>1,p}\gamma^{-1}\tilde{r}| + |r_{(2),2}| + |r_{(2),3}|) \cdot \mathbf{1}_{\Omega_\delta \cap \Omega_{(1)}} \\
&\leq \mathcal{O}(1)p^{-1/4} + \mathcal{O}(1)M^{2L+2}p^{-1/4} \to 0.
\end{aligned} \tag{4.29}$$

Meanwhile, similar to the proof of Lemma B.6 (making use of the fact that $\mathbb{E}s(\sqrt{p}W\tilde{F}W + T) \cdot \mathbf{1}_{\Omega_\delta} \leq \mathcal{O}_L(1)M^2p^{-1/4}$ and the inequality that $\mathbf{Tr}(AB) \leq n \cdot s(A)s(B)$ for n-by-n Hermitian matrices A and B), it can be shown that

$$\mathbb{E}|m_L(z) - \tilde{m}(z)| \to 0.$$

With Eq. (4.29), we have (dropping the dependence on z)

$$|\mathbb{E}\tilde{m} - RHS(\mathbb{E}\tilde{m}; a_{1,p}, \nu_p)| \to 0,$$

and thus

$$|\mathbb{E}m_L - RHS(\mathbb{E}m_L; a_{1,p}, \nu_p)| \to 0.$$

At last, by condition **(C.Variance)** and **(C.a_1)**, $a_{1,p} \to a$ and $\nu_p \to \nu$. Thus Eq. (4.9) is verified if we set $a_L(p) = a_{1,p}$ and $\nu_L(p) = \nu_p$. □

4.4. Model $X_i \sim \mathcal{U}(S^{p-1})$

We also consider the model where the random vectors X_i's are i.i.d. uniformly distributed on a high-dimensional sphere. For this model, the marginal distribution of the inner-product $\xi_{ij} = X_i^T X_j$ has probability density $Q_p^{'}(u) = A_p(1-u^2)^{(p-3)/2}$, where A_p is a normalization constant. Let $\xi_p^{'}$ have the same distribution as $\sqrt{p}\xi_{ij}$, whose probability density is $q_p^{'}(x) = \frac{1}{\sqrt{p}}Q_p^{'}(\frac{x}{\sqrt{p}})$, and let $\mathcal{H}_p^{'} = L^2(\mathbb{R}, q_p^{'}(x)dx)$. By Lemma D.6,

$$\mathbb{E}(\xi_p^{'})^k = \begin{cases} (k-1)!! + \mathcal{O}_k(1)p^{-1}, & k \text{ even;} \\ 0, & k \text{ odd,} \end{cases}$$

which echos Eq. (4.3). As a result, by Lemma 4.1, the orthonormal polynomials of $\mathcal{H}_p^{'}$ are asymptotically consistent with the Hermite polynomials. If we expand $k(x;p)$ into the orthonormal polynomials of $\mathcal{H}_p^{'}$, and require the conditions **(C.Variance)**, **(C.p-Uniform)** and **(C.a_1)** accordingly, the result in Thm. 3.4 still holds.

One way of showing this is sketched as follows:

Condition on the draw of X_n, and without loss of generality let $X_n = (1, 0, \cdots, 0)^T$. Then

$$X_i = (u_i, \sqrt{1-u_i^2}\tilde{X}_i^T)^T, \quad 1 \le i \le n-1,$$

where u_i's are i.i.d distributed, and $\tilde{X}_i$'s are i.i.d. uniformly distributed on the unit sphere in $\mathbb{R}^{p-1}$ independently from u_i's. As a result, let $\xi_{ij} = X_i^T X_j$ and $\tilde{\xi}_{ij} = \tilde{X}_i^T \tilde{X}_j$, then

$$\xi_{ij} = u_i u_j + \sqrt{1-u_i^2}\sqrt{1-u_j^2}\tilde{\xi}_{ij}, \quad 1 \le i, j \le n-1, i \ne j,$$

which is different from before. However, on the large probability set

$$\Omega_\delta = \{|u_i| \le \delta, |\tilde{\xi}_{ij}| \le \delta, 1 \le i, j \le n-1, i \ne j, \delta = p^{-1/2}M, M = \sqrt{20 \ln p}\},$$

it can be shown that

$$\xi_{ij} = u_i u_j + \tilde{\xi}_{ij} + r_{ij}, \quad |r_{ij}| \le \delta^3.$$

Thus, the Taylor expansion can be carried out in the same way, where the contribution of the extra r_{ij} term is put into T_{ij} and the bound Eq. (4.15) remains true.

We still need the mean spectral norm bound to show that Eq. (4.17) holds, and to use the bound given by the 4th moment (Lemma 4.2), it suffices to establish the bound in Lemma D.1. Notice that Gegenbauer polynomials [1] are orthogonal in the space $L^2([-1,1], Q_p^{'}(u)du)$. Gegenbauer polynomials are related to the p-spherical harmonics $\{\phi_j, j \in J\}$, which form an orthonormal

basis of $L^2(S^{p-1}, dP)$. $J = \cup_{l=0}^{\infty} J_l$, and $\{\phi_j(X), j \in J_l\}$ are p-spherical harmonics of degree l, which are homogeneous harmonic polynomials restricted to the surface of the unit sphere. The Gegenbauer polynomial of degree l as a function of X^TY, $X, Y \in S^{p-1}$, up to a multiplicative constant, equals

$$Z_{l,X}(Y) = \sum_{j \in J_l} \phi_j(X)\phi_j(Y)$$

which is named "the l-degree zonal harmonic function with axis X". We thus define $G_{l,p}(\xi)$ to be

$$G_{l,p}(X^TY) = \sum_{j \in J_l} \phi_j(X)\phi_j(Y). \tag{4.30}$$

Notice that $G_{1,p}(X^TY) = pX^TY$, and by convention $G_{0,p} = 1$. $G_{l,p}(\xi)$ is a polynomial of degree l for all l, and

$$\int_{S^{p-1}} \int_{S^{p-1}} G_{l,p}(X^TY) G_{k,p}(X^TY) dP(X) dP(Y) = \delta_{l,k}|J_l|.$$

$|J_l|$ is the number of p-spherical harmonics of degree l, $|J_1| = p$, and for $l \geq 2$

$$|J_l| = \begin{pmatrix} p+l-1 \\ l \end{pmatrix} - \begin{pmatrix} p+l-3 \\ l-2 \end{pmatrix} = \left(\frac{1}{l!} + \frac{\mathcal{O}_l(1)}{p} \right) p^l.$$

Thus, the orthonormal polynomials $P_{l,p}(x)$ of the space $\mathcal{H}_p^{'}$ can be written as

$$P_{l,p}(x) = \frac{1}{\sqrt{|J_l|}} G_{l,p}(\frac{x}{\sqrt{p}}).$$

By Eq. (4.30), we have

$$\int_{S^{p-1}} G_{l,p}(X_1^T X_2) G_{l,p}(X_2^T X_3) dP(X_2) = G_{l,p}(X_1^T X_3), \quad X_1, X_2, X_3 \in S^{p-1},$$

which gives that (define $\xi_{ij} = X_i^T X_j$)

$$\mathbb{E}[P_{l,p}(\sqrt{p}\xi_{12}) P_{l,p}(\sqrt{p}\xi_{23}) | X_1, X_3] = \frac{1}{\sqrt{|J_l|}} P_{l,p}(\sqrt{p}\xi_{13}).$$

As a result, $\mathbb{E} P_{l,p}(\sqrt{p}\xi_{12}) P_{l,p}(\sqrt{p}\xi_{23}) P_{l,p}(\sqrt{p}\xi_{34}) P_{l,p}(\sqrt{p}\xi_{41})$ is bounded by

$$\frac{1}{|J_l|} = \mathcal{O}_l(1) p^{-l},$$

which is stronger than the estimate in Lemma D.1. We comment that carrying out this analysis to higher order moments gives a moment-method proof of the convergence to semi-circle law of the ESD of random kernel matrices where $k(x; p) = P_{l,p}(x)$ for $l \geq 2$, under the model $X_i \sim \mathcal{U}(S^{p-1})$.

To continue to show the result in Thm. 3.4, the mechanism in Sec. 4.3 applies to what follows in almost the same way.

Another way of extending to the model where $X_i \sim \mathcal{U}(S^{p-1})$ is by comparing to the standard Gaussian case. That is, to replace the X_i by $X_i/|X_i|$ in the model $X_i \sim \mathcal{N}(0, p^{-1}I_p)$ and to bound the difference resulted in $m_A(z)$ (reducing to the finite expansion case $k = k_L$ first). This "comparison" argument can be used to extend the result in Thm. 3.4 to other models of the distribution of X_i's, but we do not develop this idea any further here.

5. Summary and Discussion

The main theorem, Thm. 3.4, establishes the convergence of the spectral density of random kernel matrices in the limit $p, n \to \infty$, $p/n = \gamma$, under the assumption that the random vectors are standard Gaussian. The theorem and the proofs also hold under the condition that $p/n \to \gamma$. Our proof is based on analyzing the Stieltjes transform of the random kernel matrix, and uses the expansion of the kernel function into orthonormal Hermite-like polynomials. The limiting spectral density holds for a larger class of kernel functions than the cases studied in [12], which are smooth kernels.

The assumption that the random vectors are standard Gaussian can be weakened. We showed that the result extends to the case that they are uniformly distributed over the unit sphere. Numerical simulations (not reported here) suggest that the limiting spectral density holds for other non-Gaussian random vectors, including the "Bernoulli" case, where X_i's are uniformly sampled from the 2^p vertices of the hypercube $\{-p^{-1/2}, p^{-1/2}\}^p$ (the value of the sign kernel and the divergent kernel at x = 0 is set to be 0). We conjecture that the result of our main theorem extends to more general distribution of X_i's. The validity of our result for X_i's whose p entries are independent copies of a random variable $x/\sqrt{p}$ and x has finite k-th moments for all k (which includes the "Bernoulli" case) is proved recently by Do and Vu (see Thm. 3 in [9]). However, the universality of the limiting spectral density is however beyond the scope of this paper.

While our paper mainly focused on the limiting spectral density, another question of practical importance concerns the statistics of the largest eigenvalue of random kernel matrices. This include studying the mean, variance, limiting distribution, as well as large deviation bounds for the largest eigenvalue. As discussed in Remark 4.3, the bound in Lemma 4.2 for the expected value of the spectral norm is far from being sharp. Numerical simulations (not reported here) have shown that for the models studied in this paper, the largest/smallest eigenvalue lies at the right/left end of the support of the limiting spectral density, and thus both of them are conjectured to be $\mathcal{O}(1)$ almost surely. We are not aware of any result concerning the limiting probability law of the largest eigenvalue of random kernel matrices, except for the one in [12] where the kernel function is assumed to have strong (C^3) regularity. Many other interesting questions can be asked from the RMT point of view, e.g. the "eigenvalue spacing" problem, namely the "local law" of eigenvalues. If the asymptotic concentration

of the eigenvalues at the "local level" could be established, one consequence would be that the top eigenvalue can be shown to concentrate at the right end of the limiting spectral density.

There are several interesting extensions of the inner-product kernel matrix model. The first possible extension is to distance kernel functions of the form $f(X_i, X_j) = f(|X_i - X_j|)$, which are popular in machine learning applications. Due to the relation

$$|X_i - X_j|^2 = |X_i|^2 + |X_j|^2 - 2X_i^T X_j,$$

for the model where $X_i \sim \mathcal{U}(S^{p-1})$, where $|X_i| \equiv 1$, distance kernels can be regarded as inner-product kernels. However, for the model where $X_i \sim \mathcal{N}(0, p^{-1}I_p)$, the fluctuations in $|X_i|$'s do seem to make a difference, and so far we have not been able to draw any conclusion about the limiting spectrum.

Another extension is to kernels that are of more general forms, neither an inner-product kernel nor a distance one. For example, a complex-valued kernel has been used in [18] for a dataset of tomographic images. Every pair of images is brought into in-plane rotational alignment. The modulus of the kernel function corresponds to the similarity of the images when they are optimally aligned, while the phase of the kernel is the optimal in-plane alignment angle. Notice that this kernel is discontinuous, since a small perturbation in the images may lead to a completely different phase. Similar kernels with discontinuity have also been used for dimensionality reduction [17] and sensor network localization [8]. In many senses, these applications have been the motivation for the analysis presented in this paper.

Finally, it is also possible to extend the study to non-Hermitian matrices as follows. Suppose that $X_1, \cdots, X_m$ are m i.i.d random vectors in $\mathbb{R}^p$, and $Y_1, \cdots, Y_n$ are n i.i.d random vectors in $\mathbb{R}^p$, independent from the X_i's. The m-by-n matrix A is constructed as $A_{ij} = f(X_i^T Y_j)$ where f is some function. The distribution of the singular values of A in the limit $p, m, n \to \infty$ and $p/n = \gamma_1, p/m = \gamma_2$ is conjectured to converge to a certain limiting density.

Acknowledgements

The authors were partially supported by by Award Number DMS-0914892 from the NSF and by Award Number R01GM090200 from the NIGMS. A. Singer was partially supported by Award Numbers FA9550-12-1-0317 and FA9550-13-1-0076 from AFOSR, and by Award Number LTR DTD 06-05-2012 from the Simons Foundation.

References

[1] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover publications.

[2] Anderson, G. W., Guionnet, A. and Zeitouni, O. (2010). *An Introduction to Random Matrices. Cambridge Studies in Advanced Mathematics* **118**. Cambridge Univ. Press, Cambridge, UK.
[3] Anderson, G. W. and Zeitouni, O. (2008). A law of large numbers for finite-range dependent random matrices. *Comm. Pure Appl. Math.* **61** 1118–1154.
[4] Bai, Z. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. Springer, New York.
[5] Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15** 1373–1396.
[6] Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics* **36** 2577–2604.
[7] Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Appl. Comput. Harmon. Anal.* **21** 5–30.
[8] Cucuringu, M., Lipman, Y. and Singer, A. (2012). Sensor network localization by eigenvector synchronization over the Euclidean group. *ACM Transactions on Sensor Networks (TOSN)* **8** 19.
[9] Do, Y. and Vu, V. (2013). The spectrum of random kernel matrices: universality results for rough and varying kernels. *Random Matrices: Theory and Applications*.
[10] El Karoui, N. (2007). Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability* **35** 663–714.
[11] El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics* 2717–2756.
[12] El Karoui, N. (2010). The spectrum of kernel random matrices. *The Annals of Statistics* **38** 1–50.
[13] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics* **29** 295–327.
[14] Koltchinskii, V. and Giné, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli* **6** 113–167.
[15] Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik* **114** 507–536.
[16] Schölkopf, B. and Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge MA.
[17] Singer, A. and Wu, H. T. (2012). Vector diffusion maps and the connection laplacian. *Communications on Pure and Applied Mathematics* **65** 1067–1144.
[18] Singer, A., Zhao, Z., Shkolnisky, Y. and Hadani, R. (2011). Viewing angle classification of cryo-electron microscopy images using eigenvectors. *SIAM Journal on Imaging Sciences* **4** 723–759.
[19] Soshnikov, A. (1999). Universality at the Edge of the Spectrum in Wigner Random Matrices. *Comm. Math. Phys.* **207** 697–733.

[20] Soshnikov, A. (2002). A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *Journal of Statistical Physics* **108** 1033–1056.
[21] Tao, T. (2012). *Topics in Random Matrix Theory* **132**. AMS Bookstore.
[22] Tao, T. and Vu, V. (2010). Random matrices: Universality of local eigenvalue statistics up to the edge. *Comm. Math. Phys.* **298** 549–572.
[23] Tracy, C. A. and Widom, H. (1996). On orthogonal and symplectic matrix ensembles. *Comm. Math. Phys.* **177** 727–754.

Appendix A: Solution of the Equation of $m(z)$

We rewrite Eq. (3.5) as

$$\frac{a(\nu-a^2)}{\gamma}m^3+(\nu+az)m^2+(a+\gamma z)m+\gamma=0,\quad \Im(z)>0,\Im(m)>0, \tag{A.1}$$

where $a^2\le\nu$. When $a=0$ ($a^2=\nu$) the equation corresponds to the semicircle distribution (M.P. distribution), and the existence and uniqueness of the solution with positive imaginary part are known. We consider the case where $0<a^2<\nu$, thus the cubic term in Eq. (A.1) does not vanish.

Lemma A.1. *For every z with $\Im(z)>0$, there exists a unique m with $\Im(m)>0$ for which Eq. (A.1) holds.*

Proof. It can be verified that whenever a,ν,γ are real and $\Im(z)>0$, the solution m must not be real. Define the domain $\mathcal{D}:=\{(a,\nu,\gamma,z),\gamma>0,0<a^2<\nu,\Im(z)>0\}$ which has two connected components $\mathcal{D}_+=\mathcal{D}\cap\{a>0\}$ and $\mathcal{D}_-=\mathcal{D}\cap\{a<0\}$. The three solutions of the cubic equation depend continuously on the coefficients, thus if we let (a,ν,γ,z) vary continuously in $\mathcal{D}_+$, the imaginary parts of the three solutions never change sign, and similarly for $\mathcal{D}_-$. As a result, it suffices to show that for one choice of $(a,\nu,\gamma,z)\in\mathcal{D}_+$ and one choice in $\mathcal{D}_-$, there is a unique solution with positive imaginary part. This can be done, for example, by choosing $a=\pm1/2$, $\nu=1$, $\gamma=1$ and $z=i$. □

The explicit expression for $y(u)$ defined in Eq. (3.6) is given by

$$y(u;a,\nu,\gamma)=\begin{cases}0, & D\le0,\\ \frac{\sqrt{3}}{2}((\sqrt{D}+R)^{\frac{1}{3}}+(\sqrt{D}-R)^{\frac{1}{3}}), & D>0,\end{cases} \tag{A.2}$$

where

$$\begin{aligned}D&=Q^3+R^2,\\ R&=(9\alpha_2\alpha_1-27\alpha_0-2\alpha_2^3)/54,\\ Q&=(3\alpha_1-\alpha_2^2)/9,\end{aligned}$$

and

$$m^3+\alpha_2m^2+\alpha_1m+\alpha_0=0$$

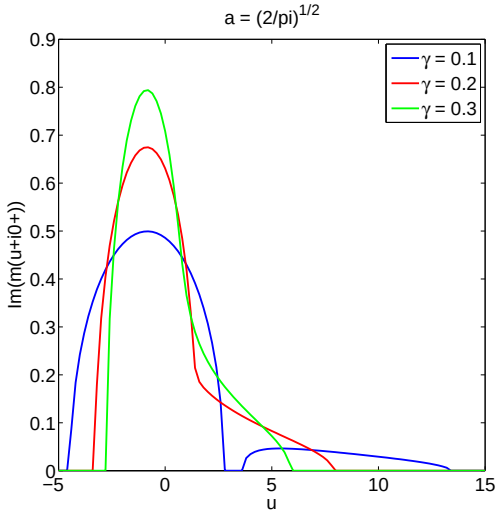

FIG 5. *Function $y(u; a, \nu, \gamma)$ as in Eq. (A.2).*

is derived from Eq. (A.1) by multiplying $(\frac{a\nu}{\gamma})^{-1}$on both sides. Explicitly,

$$\begin{aligned}\alpha_2 &= \tfrac{(\nu+au)\gamma}{a(\nu-a^2)},\\ \alpha_1 &= \tfrac{(a+\gamma u)\gamma}{a(\nu-a^2)},\\ \alpha_0 &= \tfrac{\gamma^2}{a(\nu-a^2)}.\end{aligned}$$

So all of α_2, α_1, α_0, and thus R, Q and D are real numbers. D is the "discriminant" of cubic equation, where D turning from negative to positive signals the emergence of a pair of complex solutions. The function $y(u; a, \nu, \gamma)$ is plotted in Fig. 5 where $\nu = 1$, $a = \sqrt{2/\pi}$ and $\gamma = 0.1, 0.2, 0.3$. Notice the invariance of Eq. (A.1) under the transformation

$$\nu c^2 \to \nu,\ ac \to a,\ zc \to z,\ m/c \to m$$

where c is any positive constant, which corresponds to multiplying the kernel function by c.

Appendix B: Lemma in Sec. 2

Proof of Lemma 2.4. We need the Burkholder's Inequality (Lemma 2.12. of [4]), which says that for $\{\gamma_k, 1 \le k \le n\}$ being a (complex-valued) martingale difference sequence, for $\beta > 1$,

$$\mathbb{E}|\sum_{k=1}^{n}\gamma_k|^\beta \le K_\beta \mathbb{E}\left(\sum_{k=1}^{n}|\gamma_k|^2\right)^{\beta/2}, \tag{B.1}$$

where K_β is a positive constant depending on β. Using the i.i.d. random vectors $\{X_i, 1 \le i \le n\}$, we will define the martingale to be

$$M_k = \mathbb{E}(\mathbf{Tr}(A - zI)^{-1} | \sigma\{X_{k+1}, \cdots, X_n\}) := \mathbb{E}_k \mathbf{Tr}(A - zI)^{-1}, \quad 0 \le k \le n,$$

where $\sigma\{X_{k+1}, \cdots, X_n\} := \mathcal{F}_{n-k}$ denotes the σ-algebra generated by $\{X_i, k+1 \le i \le n\}$ and $\mathbb{E}(\cdot|\mathcal{G})$ the conditional expectation with respect to the sub-σ-algebra $\mathcal{G}$. We have $M_n = \mathbb{E}\mathbf{Tr}(A - zI)^{-1}$ and $M_0 = \mathbf{Tr}(A - zI)^{-1}$, and $M_n, \cdots, M_0$ form an martingale with respect to the filtration $\{\mathcal{F}_t, t = 0, \cdots, n\}$. The martingale difference

$$\begin{aligned}
\gamma_k &= M_{k-1} - M_k \\
&= \mathbb{E}_{k-1}\mathbf{Tr}(A - zI)^{-1} - \mathbb{E}_k \mathbf{Tr}(A - zI)^{-1} \\
&= \mathbb{E}_k(\mathbf{Tr}(A - zI)^{-1} - \mathbf{Tr}(A^{(k)} - zI)^{-1}) \\
&\quad - \mathbb{E}_{k-1}(\mathbf{Tr}(A - zI)^{-1} - \mathbf{Tr}(A^{(k)} - zI)^{-1})
\end{aligned} \tag{B.2}$$

where $A^{(k)}$ is an $(n-1)$-by-$(n-1)$ matrix that is obtained from the matrix A by eliminating its k-th column and k-th row. Notice that $A^{(k)}$ is independent of X_k, $\mathbb{E}_{k-1}\mathbf{Tr}(A^{(k)} - zI)^{-1} = \mathbb{E}_k \mathbf{Tr}(A^{(k)} - zI)^{-1}$, which verifies the last line of Eq. (B.2). At the same time, we have

$$|\mathbf{Tr}(A - zI)^{-1} - \mathbf{Tr}(A^{(k)} - zI)^{-1}| \le \frac{4}{v}, \tag{B.3}$$

where $v = \Im(z) > 0$, using an argument similar to that in Sec. 2.4. of [21] (see Eq. (2.96)). The way to show Eq. (B.3) is by making use of (1) that the ordered $n-1$ eigenvalues of a minor of a symmetric (or Hermitian) matrix A "interlace" the ordered n eigenvalues of A, which follows from the Courant-Fischer theorem (see, for example, Exercise 1.3.14 of [21]), and (2) that for fixed z both real and imaginary parts of $(t - z)^{-1}$ as functions of t have bounded total variation. As a result,

$$\begin{aligned}
|\gamma_k| &\le |\mathbb{E}_k(\mathbf{Tr}(A - zI)^{-1} - \mathbf{Tr}(A^{(k)} - zI)^{-1})| \\
&\quad + |\mathbb{E}_{k-1}(\mathbf{Tr}(A - zI)^{-1} - \mathbf{Tr}(A^{(k)} - zI)^{-1})| \\
&\le 2\frac{4}{v} := C,
\end{aligned}$$

and then with Eq. (B.1), choosing $\beta = 4$,

$$\begin{aligned}
\mathbb{E}|m_A - \mathbb{E}m_A|^4 &= \frac{1}{n^4}\mathbb{E}|\sum_{k=1}^n \gamma_k|^4 \\
&\le \frac{1}{n^4} K_4 \left(\sum_{k=1}^n |\gamma_k|^2\right)^2 \\
&\le \frac{1}{n^4} K_4 (nC^2)^2 = \mathcal{O}(1) n^{-2}.
\end{aligned}$$

This implies the almost sure convergence of $m_A - \mathbb{E}m_A$ to 0 by Borel-Cantelli lemma. Also, Eq. (2.8) follows by Jensen's inequality. □

Lemma B.1. *Notations as in Sec. 2.2*

$$\mathbb{E}\left|m_A(z)-\left(-z-\left(1-\left(1+\frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right)\right)^{-1}\right|\to 0.$$

Remark B.2. The proof provided below can be replaced by a simpler one. The reason we give this proof is that it contains many of the techniques that are used in showing the main result.

Proof. Continue from Eq. (2.15). We first observe that when p is large, $|X_n|^2$ concentrates at 1, and specifically, with p large enough

$$\mathbf{Pr}\left[||X_n|^2-1|>\sqrt{\frac{40\ln p}{p}}\right]<p^{-9}, \tag{B.4}$$

which can be verified by standard large deviation inequality techniques. However, at this stage the following moment bound will be enough for our purpose:

$$\mathbb{E}\left||X_n|^2-1\right|\le\sqrt{\mathbb{E}\left(|X_n|^2-1\right)^2}=\sqrt{\frac{2}{p}}\to 0. \tag{B.5}$$

We then write the denominator in Eq. (2.15) as

$$\begin{aligned}\eta^T(\tilde{A}^{(n)}-D_\eta-zI_{n-1})^{-1}\eta&=\frac{1}{p}\mathbf{Tr}(\tilde{A}^{(n)}-zI_{n-1})^{-1}+r\\&=\frac{1}{\gamma}\mathbb{E}\tilde{m}(z)+\frac{1}{\gamma}\tilde{r}+r,\end{aligned} \tag{B.6}$$

where $r=\eta^T(\tilde{A}^{(n)}-D_\eta-zI_{n-1})^{-1}\eta-\frac{1}{p}\mathbf{Tr}(\tilde{A}^{(n)}-zI_{n-1})^{-1}$, $\tilde{m}(z):=\frac{1}{n}\mathbf{Tr}(\tilde{A}^{(n)}-zI_{n-1})^{-1}$, and $\tilde{r}:=(\tilde{m}(z)-\mathbb{E}\tilde{m}(z))$. We have that

1. $\mathbb{E}|\tilde{r}|\le\mathcal{O}(1)n^{-1/2}$ as $n\to\infty$: Because $\tilde{A}^{(n)}$ is itself an $(n-1)\times(n-1)$ kernel matrix by Eq. (2.14), Lemma 2.4 applies.
2. r splits into two terms

$$\begin{aligned}r&=\left(\eta^T(\tilde{A}^{(n)}-D_\eta-zI_{n-1})^{-1}\eta-\eta^T(\tilde{A}^{(n)}-zI_{n-1})^{-1}\eta\right)\\&\quad+\left(\eta^T(\tilde{A}^{(n)}-zI_{n-1})^{-1}\eta-\frac{1}{p}\mathbf{Tr}(\tilde{A}^{(n)}-zI_{n-1})^{-1}\right)\\&:=r_1+r_2,\end{aligned}$$

where (1) $\mathbb{E}|r_2|\le\mathcal{O}(1)p^{-1/2}$, by Lemma B.4; (2) $|r_1|\mathbf{1}_{\Omega_\delta}\le\mathcal{O}(1)p^{-1/2}$, where Ω_δ is a large probability set depending on p, defined as

$$\Omega_\delta=\{|\eta_i|<\delta,\,1\le i\le n-1,\,\delta=\frac{M}{\sqrt{p}}\},\quad M=\sqrt{20\ln p},$$

by Lemma B.3. Notice that $M=o(p^\epsilon)$ for any $\epsilon>0$.

Back to Eq. (2.10). By Eqs. (2.15) and (B.6), we have

$$\begin{aligned}\mathbb{E}m_A(z) &= \mathbb{E}\left((A-zI)^{-1}\right)_{nn} \\ &= \mathbb{E}\left(-z-|X_n|^2\left(1-(1+\frac{1}{p}\mathbf{Tr}(\tilde{A}^{(n)}-zI_{n-1})^{-1}+r)^{-1}\right)\right)^{-1}.\end{aligned}$$

The following bounds (1) - (4) can be verified:

(1) (Lemma B.8) On Ω_δ, $|\eta^T(A^{(n)}-zI_{n-1})^{-1}\eta|$ and $|(1+\eta^T(\tilde{A}^{(n)}-D_\eta-zI_{n-1})^{-1}\eta)^{-1}|$ are both bounded by $M' = 1+\mathcal{O}(1)M^2$, $M' = o(p^\epsilon)$ for any $\epsilon > 0$.

(2) (Lemma B.7) On $\Omega_r \cap \Omega_\delta$, $\left|\left(1+\frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right| \le 2M'$, where we define

$$\Omega_r = \{|\tilde{r}| < p^{-1/4}, |r_2| < p^{-1/4}\},$$

and by Markov inequality, we have

$$\mathbf{Pr}(\Omega_r^c) \le p^{1/4}\mathbb{E}|\tilde{r}| + p^{1/4}\mathbb{E}|r_2| \le \mathcal{O}(1)p^{-1/4}$$

when p is large.

(3) $\left|\left((A-zI)^{-1}\right)_{nn}\right| \le \frac{1}{v}$, which is Eq. (2.4).

(4) $\left|\left(-z-\left(1-\left(1+\frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right)\right)^{-1}\right| \le \frac{1}{v}$: By $\Im\left(-\left(1+\frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right)$ equals a positive number times $\Im\left(\frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)$ which is also positive, one verifies that

$$\Im\left(-z-\left(1-\left(1+\frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right)\right) < \Im(-z) = -v,$$

so

$$\left|-z-\left(1-\left(1+\frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right)\right| > v.$$

With (1) and (2), we have

$$\begin{aligned}&\mathbb{E}\left|\left(1+\eta^T(\tilde{A}^{(n)}-D_\eta-zI_{n-1})^{-1}\eta\right)^{-1}-\left(1+\frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right|\cdot\mathbf{1}_{\Omega_\delta\cap\Omega_r} \\ &\le \mathbb{E}(M'\cdot 2M')(|r|+\frac{1}{\gamma}|\tilde{r}|)\cdot\mathbf{1}_{\Omega_\delta\cap\Omega_r} \\ &\le 2M'^2(\mathbb{E}|r_2|+\mathbb{E}|r_1|\cdot\mathbf{1}_{\Omega_\delta}+\gamma^{-1}\mathbb{E}|\tilde{r}|) \\ &\le 2M'^2(\mathcal{O}(1)p^{-1/2}+\mathcal{O}(1)p^{-1/2}+\mathcal{O}(1)n^{-1/2}) \\ &= \mathcal{O}(1)M'^2p^{-1/2}. \qquad \text{(B.7)}\end{aligned}$$

Using bounds (1)-(4), together with Eqs. (B.7) and (B.5), we have

$$
\begin{aligned}
&\mathbb{E}\left|m_A(z)-\left(-z-\left(1-\left(1+\frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right)\right)^{-1}\right| \\
&=\mathbb{E}\left|\left(-z-|X_n|^2\eta^T(A^{(n)}-zI_{n-1})^{-1}\eta\right)^{-1}-\left(-z-\left(1-\left(1+\frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right)\right)^{-1}\right| \\
&\leq \frac{2}{v}\left(\mathbf{Pr}(\Omega_\delta^c)+\mathbf{Pr}(\Omega_r^c)\right) \\
&\quad+\mathbb{E}\left|\left(-z-|X_n|^2\eta^T(A^{(n)}-zI_{n-1})^{-1}\eta\right)^{-1}-\left(-z-\left(1-\left(1+\frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right)\right)^{-1}\right|\cdot\mathbf{1}_{\Omega_\delta\cap\Omega_r} \\
&\leq \frac{2}{v}\left(\mathbf{Pr}(\Omega_\delta^c)+\mathbf{Pr}(\Omega_r^c)\right) \\
&\quad+\mathbb{E}\frac{1}{v^2}\left||X_n|^2-1\right|\cdot|\eta^T(A^{(n)}-zI_{n-1})^{-1}\eta|\cdot\mathbf{1}_{\Omega_\delta\cap\Omega_r} \\
&\quad+\mathbb{E}\frac{1}{v^2}\left|\left(1+\eta^T(\tilde{A}^{(n)}-D_\eta-zI_{n-1})^{-1}\eta\right)^{-1}-\left(1+\frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right|\cdot\mathbf{1}_{\Omega_\delta\cap\Omega_r} \\
&\leq \frac{2}{v}\left(\mathbf{Pr}(\Omega_\delta^c)+\mathbf{Pr}(\Omega_r^c)\right)+\mathbb{E}\frac{1}{v^2}\left||X_n|^2-1\right|M'\mathbf{1}_{\Omega_\delta\cap\Omega_r}+\frac{1}{v^2}\mathcal{O}(1)M'^2p^{-1/2} \\
&\leq \mathcal{O}(1)p^{-9}+\mathcal{O}(1)p^{-1/4}+M'\mathcal{O}(1)p^{-1/2}+\mathcal{O}(1)M'^2p^{-1/2} \\
&= o(p^{-1/2+\epsilon}),
\end{aligned}
$$

for any $\epsilon>0$, which proves the statement. □

Lemma B.3. *Notations as in Lemma B.1,*

$$|r_1|\mathbf{1}_{\Omega_\delta}\leq\mathcal{O}(1)p^{-1/2}$$

Proof. By

$$
\begin{aligned}
\mathbf{Pr}[|\eta_i|>\delta] &= 2\int_M^\infty\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du \\
&\leq \frac{1}{\sqrt{2}}e^{-\frac{M^2}{2}} \\
&= \frac{1}{\sqrt{2}}p^{-10},\quad 1\leq i\leq n-1,
\end{aligned}
\tag{B.8}
$$

and the union bound, we have

$$\mathbf{Pr}(\Omega_\delta^c)\leq(n-1)\mathbf{Pr}[|\eta_i|>\delta]\leq\mathcal{O}(1)p^{-9}.$$

Now (recall that $s(\cdot)$ denotes the magnitude of the largest singular value/spectral

norm of a matrix)

$$\begin{aligned}|r_1| &= \left|\eta^T(\tilde{A}^{(n)} - D_\eta - zI_{n-1})^{-1}\eta - \eta^T(\tilde{A}^{(n)} - zI_{n-1})^{-1}\eta\right| \\ &= \left|\eta^T(\tilde{A}^{(n)} - D_\eta - zI_{n-1})^{-1}D_\eta(\tilde{A}^{(n)} - zI_{n-1})^{-1}\eta\right| \\ &\leq s\left((\tilde{A}^{(n)} - D_\eta - zI_{n-1})^{-1}D_\eta(\tilde{A}^{(n)} - zI_{n-1})^{-1}\right)|\eta|^2.\end{aligned}$$

Notice that on Ω_δ

$$s(D_\eta) \leq \max_{1\leq i\leq n-1}\eta_i^2 \leq \delta^2,$$

also $|\eta|^2 \leq (n-1)\delta^2$. At the same time both $s((\tilde{A}^{(n)} - D_\eta - zI_{n-1})^{-1})$ and $s((\tilde{A}^{(n)} - zI_{n-1})^{-1})$ is bounded by $\frac{1}{v}$ an absolute constant. Adding together (for Hermitian matrices A and B, $s(AB) \leq s(A)s(B)$) we have

$$|r_1|\mathbf{1}_{\Omega_\delta} \leq \frac{1}{v^2}\delta^2 \cdot (n-1)\delta^2 = \frac{M^4(n-1)}{v^2p^2} < \mathcal{O}(1)p^{-1/2}. \tag{B.9}$$

□

Lemma B.4. *Notations as in Sec. 2.2,*

$$\mathbb{E}|r_2| \leq \mathcal{O}(1)p^{-1/2}.$$

Remark B.5. The technique is similar to the moment bound method in [4, Chapter 3.3], where the main observation is that $\tilde{A}^{(n)}$ is independent of the vector η.

Proof. Define $(\tilde{A}^{(n)} - zI_{n-1})^{-1}$ as $\tilde{B}$ which is Hermitian, we have

$$\begin{aligned}\mathbb{E}|r_2|^2 &= \mathbb{E}\left|\sum_{i=1}^{n-1}\left(\eta_i^2 - \frac{1}{p}\right)\tilde{B}_{ii} + \sum_{i_1\neq i_2}\eta_{i_1}\eta_{i_2}\tilde{B}_{i_1i_2}\right|^2 \\ &= \mathbb{E}\sum_{i,i'}\left(\eta_i^2 - \frac{1}{p}\right)\left(\eta_{i'}^2 - \frac{1}{p}\right)\tilde{B}_{ii}\overline{\tilde{B}_{i'i'}} \\ &\quad + \mathbb{E}\sum_i\sum_{i_1\neq i_2}\left(\eta_i^2 - \frac{1}{p}\right)\eta_{i_1}\eta_{i_2}\left(\tilde{B}_{ii}\overline{\tilde{B}_{i_1i_2}} + \overline{\tilde{B}_{ii}}\tilde{B}_{i_1i_2}\right) \\ &\quad + \mathbb{E}\sum_{i_1\neq i_2}\sum_{i_1'\neq i_2'}\eta_{i_1}\eta_{i_2}\eta_{i_1'}\eta_{i_2'}\tilde{B}_{i_1i_2}\overline{\tilde{B}_{i_1'i_2'}}.\end{aligned}$$

By taking expectation over η_i's first, we see many terms vanish due to the

independence of η_{i_1} and η_{i_2} for $i_1 \neq i_2$, and what remains gives

$$\begin{aligned}\mathbb{E}|r_2|^2 &\leq \mathbb{E}\left(\sum_i \mathbb{E}\left(\eta_i^2 - \frac{1}{p}\right)^2 |\tilde{B}_{ii}|^2 + \sum_{i_1 \neq i_2} 2\frac{1}{p^2}|\tilde{B}_{i_1 i_2}|^2\right) \\ &= \mathbb{E}\frac{2}{p^2}\left(\sum_i |\tilde{B}_{ii}|^2 + \sum_{i_1 \neq i_2} |\tilde{B}_{i_1 i_2}|^2\right) \\ &= \mathbb{E}\frac{2}{p^2}\mathbf{Tr}(\bar{\tilde{B}}^T \tilde{B}).\end{aligned}$$

Observe that

$$\mathbf{Tr}(\bar{\tilde{B}}^T \tilde{B}) = \sum_{i=1}^{n-1} \frac{1}{|\tilde{\lambda}_i - z|^2} \leq \sum_{i=1}^{n-1} \frac{1}{v^2} = \frac{n-1}{v^2},$$

where $v = \Im(z) > 0$ and $\tilde{\lambda}_i$ are the eigenvalues of $\tilde{A}^{(n)}$. Then

$$\frac{2}{p^2}\mathbf{Tr}(\bar{\tilde{B}}^T \tilde{B}) \leq \frac{2}{v^2}\frac{n-1}{p^2} \leq \frac{2}{v^2\gamma} \cdot \frac{1}{p},$$

which means that

$$\mathbb{E}|r_2|^2 \leq \frac{\mathcal{O}(1)}{p},$$

so we have $\mathbb{E}|r_2| \leq \sqrt{\mathbb{E}|r_2|^2} \leq \mathcal{O}(1)p^{-1/2}$. □

Lemma B.6. *Notations as in Sec. 2.2,*

$$\mathbb{E}|m_A(z) - \tilde{m}(z)| \to 0.$$

Proof. First, $|m_A(z) - m_{A^{(n)}}(z)| \leq \frac{4}{v} \cdot n^{-1} \to 0$, due to Eq. (B.3). Second, we show that $\mathbb{E}|m_{A^{(n)}} - m_{\tilde{A}^{(n)}}| \to 0$. By

$$\begin{aligned}&\mathbf{Tr}(A^{(n)} - zI_{n-1})^{-1} - \mathbf{Tr}(\tilde{A}^{(n)} - zI_{n-1})^{-1} \\ &= \mathbf{Tr}(-(A^{(n)} - zI_{n-1})^{-1}(\eta\eta^T - D_\eta)(\tilde{A}^{(n)} - zI_{n-1})^{-1}) \\ &= -\eta^T(A^{(n)} - zI_{n-1})^{-1}(\tilde{A}^{(n)} - zI_{n-1})^{-1}\eta \\ &\quad + \mathbf{Tr}((A^{(n)} - zI_{n-1})^{-1}D_\eta(\tilde{A}^{(n)} - zI_{n-1})^{-1}),\end{aligned}$$

and using a similar argument as before, we can show that on Ω_δ

$$\left|\eta^T(A^{(n)} - zI_{n-1})^{-1}(\tilde{A}^{(n)} - zI_{n-1})^{-1}\eta\right| \leq \frac{1}{v^2}|\eta|^2 \leq \frac{1}{v^2}(n-1)\delta^2 \leq \mathcal{O}(1)M^2,$$

and

$$\begin{aligned}&\left|\mathbf{Tr}((A^{(n)} - zI_{n-1})^{-1}D_\eta(\tilde{A}^{(n)} - zI_{n-1})^{-1})\right| \\ &\leq (n-1)s((A^{(n)} - zI_{n-1})^{-1}D_\eta(\tilde{A}^{(n)} - zI_{n-1})^{-1}) \\ &\leq \frac{1}{v^2}(n-1)\delta^2 = \mathcal{O}(1)M^2.\end{aligned}$$

As a result,

$$\begin{aligned}\mathbb{E}|m_{A^{(n)}} - m_{\tilde{A}^{(n)}}| &= \frac{2}{v}\mathbf{Pr}(\Omega_\delta^c) + \mathbb{E}|m_{A^{(n)}} - m_{\tilde{A}^{(n)}}| \cdot \mathbf{1}_{\Omega_\delta} \\ &\leq \mathcal{O}(1)p^{-9} + \frac{1}{n}\mathcal{O}(1)M^2\end{aligned}$$

which goes to 0 as $n, p \to \infty$ with $p/n = \gamma$. □

Lemma B.7. *Notations as in Sec. 2.2, on* $\Omega_r \cap \Omega_\delta$,

$$\left|\left(1 + \frac{1}{\gamma}\mathbb{E}\tilde{m}(z)\right)^{-1}\right| \leq 2M^{'}.$$

Proof. On $\Omega_r \cap \Omega_\delta$, with Eq. (B.9) $|r_1| < \mathcal{O}(1)p^{-1/2}$ thus $|r| \leq |r_1| + |r_2|$ is bounded by $\mathcal{O}(1)p^{-1/4}$,

$$\begin{aligned}\left|\frac{1}{1 + \frac{1}{\gamma}\mathbb{E}\tilde{m}(z)}\right| &= \left|\frac{1}{1 + \eta^T(\tilde{A}^{(n)} - D_\eta - zI_{n-1})^{-1}\eta - r - \frac{1}{\gamma}\tilde{r}}\right| \\ &\leq \frac{1}{\left|1 + \eta^T(\tilde{A}^{(n)} - D_\eta - zI_{n-1})^{-1}\eta\right| - |r| - \frac{1}{\gamma}|\tilde{r}|} \\ &\leq 2M^{'}\end{aligned}$$

as $\left|1 + \eta^T(\tilde{A}^{(n)} - D_\eta - zI_{n-1})^{-1}\eta\right| \geq 1/M^{'} \gg (|r| + \frac{1}{\gamma}|\tilde{r}|)$, where the latter is bounded by $\mathcal{O}(1)p^{-1/4}$. □

Lemma B.8. *Notation as in Sec. 2.2, on* Ω_δ, *both* $|\eta^T(A^{(n)} - zI_{n-1})^{-1}\eta|$ *and* $|(1 + \eta^T(\tilde{A}^{(n)} - D_\eta - zI_{n-1})^{-1}\eta)^{-1}|$ *are bounded by* $M^{'}$.

Proof. On Ω_δ, $|\eta^T(A^{(n)} - zI_{n-1})^{-1}\eta| \leq s((A^{(n)} - zI_{n-1})^{-1})|\eta|^2 \leq \frac{1}{v}\delta^2(n-1) = \mathcal{O}(1)M^2$, and also

$$\begin{aligned}&\left|\left(1 + \eta^T(\tilde{A}^{(n)} - D_\eta - zI_{n-1})^{-1}\eta\right)^{-1}\right| \\ &= \left|1 - \eta^T(\eta\eta^T + \tilde{A}^{(n)} - D_\eta - zI_{n-1})^{-1}\eta\right| \\ &\leq 1 + |\eta^T(A^{(n)} - zI_{n-1})^{-1}\eta| \\ &\leq 1 + \mathcal{O}(1)M^2 := M^{'}.\end{aligned}$$ □

Appendix C: Lemma in Sec. 3

Lemma C.1. *Model and notations as in Sec. 3.1. Due to Eqn. (4.3), the result in Lemma 4.1 holds.*

Suppose that $k(x;p)$ *is in* $\mathcal{H}_\mathcal{N}$ *and* $\mathcal{H}_p$ *for all* p, *and satisfies*

$$\int_{\mathbb{R}} k(x;p)^2|q_p(x) - q(x)|dx \to 0, \quad p \to \infty.$$

Let

$$b_{l,p} = \int_{\mathbb{R}} k(x;p) h_l(x) q(x) dx,$$
$$a_{l,p} = \int_{\mathbb{R}} k(x;p) P_{l,p}(x) q_p(x) dx,$$

for $l = 0, 1, \cdots$. *Then for each* l, $|b_{l,p} - a_{l,p}| \to 0$ *as* $p \to \infty$.

Proof.

$$\begin{aligned}
&|b_{l,p} - a_{l,p}| \\
&= |\int_{\mathbb{R}} k h_l (q - q_p) dx + \int_{\mathbb{R}} k(h_l - P_{l,p}) q_p dx| \\
&\leq \int |k h_l| |q - q_p| dx + \int |k| |h_l - P_{l,p}| q_p dx \\
&:= (1) + (2).
\end{aligned}$$

For (1), by Cauchy-Schwarz inequality

$$(1)^2 \leq (\int k^2 |q - q_p| dx)(\int h_l^2 |q - q_p| dx),$$

where

$$\int h_l^2 |q - q_p| dx \leq \int h_l^2 q dx + \int h_l^2 q_p dx = 1 + (1 + \mathcal{O}_l(1) p^{-1}),$$

which is bounded as $p \to \infty$, and $\int k^2 |q - q_p| dx \to 0$, thus $(1) \to 0$. For (2),

$$(2)^2 \leq (\int k^2 q_p dx)(\int (h_l - P_{l,p})^2 q_p dx),$$

where $\int k^2 q_p dx \to \int k^2 q dx$ which is bounded, and by Lemma 4.1

$$h_l(x) - P_{l,p}(x) = \sum_{j=0}^{l} (\delta_{l,p})_j x^j, \quad \max_{0 \leq j \leq l} |(\delta_{l,p})_j| < \mathcal{O}_l(1) p^{-1},$$

thus

$$\left(\int (h_l - P_{l,p})^2 q_p dx \right)^{1/2} \leq \mathcal{O}_l(1) p^{-1},$$

so $(2) \to 0$. □

Lemma C.2. *Model and notations as in Sec. 3.2, and suppose that* $k(x)$ *is as in Remark 3.2. Eqn. (3.4) implies that* $\mathbb{E}k(\xi_p)^2 \to \mathbb{E}k(\zeta)^2 = \nu_{\mathcal{N}}$. *Without loss of generality,* $k(x)$ *is in* $\mathcal{H}_p$ *for all* p. *Define* $b_{l,p}$ *and* $a_{l,p}$ *as in Lemma C.1, and notice that since* $k(x)$ *does not depend on* p, $b_{l,p} = b_l$ *independent of* p.

Then conditions **(C.Variance)**, **(C.p-Unform)** *and* **(C.α_1)** *are satisfied by* $k(x;p) = k(x) - a_{0,p}$. *Also,* $\nu_p \to \nu_{\mathcal{N}}$, *and* $a_{1,p} \to a_{\mathcal{N}} = b_1$.

Proof. By definition $\mathbb{E}k(\xi_p;p)=0$. In this case,

$$\nu_p=\mathbb{E}k(\xi_p;p)^2=\mathbb{E}k(\xi_p)^2-a_{0,p}^2.$$

Since Lemma C.1 applies to $k(x)$, we know that

$$a_{0,p}\to b_0=\mathbb{E}k(\zeta)=0.$$

Together with the fact that $\mathbb{E}k(\xi_p)^2\to\mathbb{E}k(\zeta)^2=\nu_{\mathcal{N}}$, we know that $\nu_p\to\nu_{\mathcal{N}}$ as $p\to\infty$. Thus **(C.Variance)** is satisfied.

Also $a_{1,p}\to b_1$ which is a constant, thus **(C.α_1)** holds.

For **(C.p-Unform)** to be satisfied, it suffices to show that $\sum_{l=L+1}^{\infty}a_{l,p}^2$ can be made p-uniformly small. Notice that

$$\sum_{l=0}^{\infty}a_{l,p}^2=\mathbb{E}k(\xi_p)^2\to\nu_{\mathcal{N}},$$
$$\sum_{l=0}^{\infty}b_l^2=\nu_{\mathcal{N}},$$

and meanwhile for each l, $a_{l,p}\to b_l$ by Lemma C.1, thus for any finite L

$$\sum_{l=L+1}^{\infty}a_{l,p}^2=\mathbb{E}k(\xi_p)^2-\sum_{l=0}^{L}a_{l,p}^2$$
$$\to\nu_{\mathcal{N}}-\sum_{l=0}^{L}b_l^2=\sum_{l=L+1}^{\infty}b_l^2,$$

which can be made small by choosing L large independently of p. □

Lemma C.3. *Notations as in Sec. 3.2. If $f(\xi;p)=f(\xi)$ is C^1 at $\xi=0$, then the theorem applies and $a^2=\nu$. Specifically, $a=f'(0)$.*

Proof. We first truncate $f(\xi)$ to be $\hat{f}(\xi;p)=f(\xi)\mathbf{1}_{\{|\xi|\le\delta\}}$, where $\delta=\delta(p)=\frac{M}{\sqrt{p}}$, $M=\sqrt{20\ln p}$. Using a similar argument as in Lemma D.3, we have

$$\mathbf{Pr}[\exists i\neq j,\ |X_i^TX_j|>\delta]\le\mathcal{O}(1)p^{-7}.$$

Thus, if we denote $\hat{A}$ as the random kernel matrix with kernel function $\hat{f}$, then for fixed $z=u+iv$

$$\mathbb{E}|m_A(z)-m_{\hat{A}}(z)|\le\frac{2}{v}\mathbf{Pr}[\exists i\neq j,\ |X_i^TX_j|>\delta]\to0,$$

where $m_A(z)$ and $m_{\hat{A}}(z)$ are the Stieltjes transforms of A and $\hat{A}$ respectively. Since the convergence of $\mathbb{E}m_A(z)$ implies the convergence of the spectral density, if suffices to show that the claim in the lemma holds for $\hat{f}(\xi;p)$.

Since $f(\xi)$ is C^1 at $\xi = 0$, for any $\epsilon > 0$ there exists a neighborhood $[-R, R]$ on which

$$f(\xi) = f(0) + f^{'}(0)\xi + r(\xi), \quad |r(\xi)| \leq \epsilon|\xi|.$$

Since $\delta \to 0$ when $p \to \infty$, we assume that p is large enough so that $\delta < R$. Let $k(x;p) = \sqrt{p}\hat{f}(x/\sqrt{p})$, and assume that $f(0) = 0$ since it only contributes to $\mathbb{E}k(\xi_p, p) = a_{0,p}$, we have

$$\begin{aligned} k(x;p) &= \left(f^{'}(0)x + \sqrt{p}r(\frac{x}{\sqrt{p}}) \right) \mathbf{1}_{\{|x|\leq M\}} \\ &:= k_1 + k_2, \end{aligned}$$

where $|k_2(x;p)| \leq \epsilon|x|$, so $\mathbb{E}k_2(\xi_p;p)^2 \leq \epsilon^2$. Thus, the L^2 norm of k_2 is arbitrarily small in $\mathcal{H}_p$, and $\nu_p = Var(k(\xi_p;p))$ and $a_{1,p} = \mathbb{E}\xi_p(k(\xi_p;p) - \mathbb{E}k(\xi_p;p))$ are decided by k_1. For $k_1(x;p) = f^{'}(0)x\mathbf{1}_{\{|x|\leq M\}}$, $\mathbb{E}k_1(\xi_p;p) = 0$, and since $M \to \infty$ as $p \to \infty$, $\mathbb{E}k_1(\xi_p;p)^2 \to (f^{'}(0))^2$ and $\mathbb{E}\xi_p k_1(\xi_p;p) \to f^{'}(0)$. Thus $\nu_p \to (f^{'}(0))^2 = \nu$, and $a_{1,p} \to f^{'}(0) = a$. □

Lemma C.4. *Let ξ_p be as in Sec. 3.2, and equivalently $\xi_p = p^{-1/2}\sum_{i=1}^{p} x_i y_i$ where x_i and y_i i.i.d.$\sim \mathcal{N}(0,1)$. Then for $p > 2$,*

$$\mathbf{Pr}[|\xi_p| > R] \leq (2e)e^{-R}.$$

Proof. Since for $|t| < \sqrt{p}$, $\mathbb{E}e^{t\frac{x_1 y_1}{\sqrt{p}}} = (1 - t^2/p)^{-1/2}$, by choosing $t = 1$ we have

$$\begin{aligned} \mathbf{Pr}[\xi_p > R] &\leq e^{-M}(\mathbb{E}e^{\frac{x_1 y_1}{\sqrt{p}}})^p \\ &= e^{-M}(1 - \frac{1}{p})^{-p/2} \\ &\leq e^{-M}e, \end{aligned}$$

where the last line is due to that $x = 1/p$ satisfies $\log(1-x)/x > -2$ when $0 < x < 1/2$. The argument for bounding $\mathbf{Pr}[\xi_p < -R]$ is similar. □

Lemma C.5. *Notations as in Sec. 3.2. Suppose $k(x)$ is (Case 1) bounded, or (Case 2) in $\mathcal{H}_\mathcal{N}$ and $\mathcal{H}_p$ for all p, is bounded on $|x| \leq R$ for any $R > 0$, and satisfies*

$$\int_{|x|>R} k(x)^2 q_p(x)dx \to 0, \quad R \to \infty$$

uniformly in p, then Eqn. (3.4) holds. When $k(x)$ is any polynomial, it belongs to (Case 2).

Proof. First, we reduce (Case 2) to (Case 1). Notice that

$$\begin{aligned} &\int_{\mathbb{R}} k(x)^2|q_p(x) - q(x)|dx \\ \leq &\int_{|x|\leq R} k(x)^2|q_p(x) - q(x)|dx + \int_{|x|>R} k(x)^2 q_p(x)dx \\ &+ \int_{|x|>R} k(x)^2 q(x)dx. \end{aligned}$$

The last two terms can be made arbitrarily small independently of p by choosing R large, and for fixed R, the first term goes to 0 given that (Case 1) is proved.

To show the claim for (Case 1), it suffices to show that $\int |q_p - q| dx \to 0$. Since ξ_p converge in distribution to $\mathcal{N}(0,1)$, we know that for any finite R, $\int_{|x|<R} |q_p(x) - q(x)| dx \to 0$. Thus, it suffices to show that

$$\int_{|x|>R} q_p(x) dx \to 0, \quad R \to \infty$$

uniformly in p. This follows from the large deviation bound given in Lemma C.4.

(Case 2) includes all the polynomials. The p-uniform integrability is verified by Cauchy-Schwarz inequality, combined with 1) the fact that all (even) moments of ξ_p are finite and converge to those of the standard Gaussian (Eq. (4.3)), thus are p-uniformly bounded, and 2) the bound given in Lemma C.4.

□

Appendix D: Lemma in Sec. 4

Lemma D.1. *Let X_1, X_2, X_3, X_4 be i.i.d distributed as $\mathcal{N}(0, p^{-1} I_p)$, and $P_{l,p}(x)$ is the degree-l Hermite-like polynomial as defined in Sec. 4.1, $l \geq 2$. Then*

$$\mathbb{E} P_{l,p}(\sqrt{p}\xi_{12}) P_{l,p}(\sqrt{p}\xi_{23}) P_{l,p}(\sqrt{p}\xi_{34}) P_{l,p}(\sqrt{p}\xi_{41}) = \mathcal{O}_l(1) p^{-2},$$

where $\xi_{ij} = X_i^T X_j$.

Lemma D.2. *Suppose that Z is a random variable, and for positive integer l and $\epsilon > 0$,*

$$|\mathbb{E} Z^k - 1| < \epsilon, \quad 0 \leq k \leq l.$$

Let $q(x) = \sum_{k=0}^{l} b_k x^k$ is a polynomial of degree l, then

$$|\mathbb{E} q(Z)| \leq |q(1)| + \epsilon \sum_{j=0}^{l} |b_j|.$$

The proof is elementary and is omitted.

Proof of Lemma D.1. There exists an orthogonal transform P_1 that depends on X_1 so that

$$P_1 X_1 = (|X_1|, 0, \cdots, 0)^T,$$

and let

$$P_1 X_i = (\eta_i, \tilde{X}_i)^T, \quad i = 2, 3, 4.$$

Since $X_1, \cdots, X_4$ i.i.d. $\sim \mathcal{N}(0, p^{-1} I_p)$, $|X_1| \sim \chi(p)/\sqrt{p}$, $\eta_i \sim \mathcal{N}(0, p^{-1})$, $\tilde{X}_i \sim \mathcal{N}(0, p^{-1} I_{p-1})$ for $i = 2, 3, 4$ are independent. Also, there exists an orthogonal transform P_3 which applies to the 2-to-p coordinates so that

$$P_3 P_1 X_3 = (\eta_3, |\tilde{X}_3|, 0, \cdots, 0)^T,$$

and let

$$P_3P_1X_i = (\eta_i, \tilde{\eta}_i, \cdots)^T, \quad i = 2, 4.$$

By the independence of $X_1, \cdots, X_4$ and that P_3 only depends on X_1, X_3, $|\tilde{X}_3| \sim \chi(p-1)/\sqrt{p}$, $\tilde{\eta}_i \sim \mathcal{N}(0, p^{-1})$ for $i = 2, 4$, and $|X_1|$, η_i, $i = 2, 3, 4$, $|\tilde{X}_3|$ and $\tilde{\eta}_i$, $i = 2, 4$ are jointly independent.

Thus,

$$\xi_{12} = |X_1|\eta_2,\ \xi_{14} = |X_1|\eta_4,\ \xi_{23} = \eta_2\eta_3 + |\tilde{X}_3|\tilde{\eta}_2,\ \xi_{34} = \eta_3\eta_4 + |\tilde{X}_3|\tilde{\eta}_4,$$

and define

$$\zeta_2 := \sqrt{p}\eta_2,\ \zeta_4 := \sqrt{p}\eta_4,\ \tilde{\zeta}_2 := \sqrt{p}\tilde{\eta}_2,\ \tilde{\zeta}_4 := \sqrt{p}\tilde{\eta}_4,$$

which are i.i.d. distributed as $\mathcal{N}(0, 1)$, we have that

$$\sqrt{p}\xi_{12} = |X_1|\zeta_2,\ \sqrt{p}\xi_{14} = |X_1|\zeta_4,\ \sqrt{p}\xi_{23} = \eta_3\zeta_2 + |\tilde{X}_3|\tilde{\zeta}_2,\ \sqrt{p}\xi_{34} = \eta_3\zeta_4 + |\tilde{X}_3|\tilde{\zeta}_4.$$

Since $P_{l,p}(x)$ is a polynomial of degree l,

$$P_{l,p}(x_1 + x_2) \quad = \quad \sum_{k=0}^{l} \frac{P_{l,p}^{(k)}(x_2)}{k!} x_1^k,$$

thus

$$\begin{aligned} & \mathbb{E}P_{l,p}(\sqrt{p}\xi_{12})P_{l,p}(\sqrt{p}\xi_{23})P_{l,p}(\sqrt{p}\xi_{34})P_{l,p}(\sqrt{p}\xi_{41}) \\ = & \sum_{j,k=0}^{l} \mathbb{E}\eta_3^{k+j}\mathbb{E}\left(\zeta_2^k P_{l,p}(|X_1|\zeta_2)\zeta_4^j P_{l,p}(|X_1|\zeta_4)\right) \\ & \cdot \frac{\mathbb{E}\left(P_{l,p}^{(k)}(|\tilde{X}_3|\tilde{\zeta}_2)P_{l,p}^{(j)}(|\tilde{X}_3|\tilde{\zeta}_4)\right)}{k!j!}. \end{aligned} \tag{D.1}$$

Meanwhile, let

$$P_{l,p}(x) = \sum_{j=0}^{l} (c_{l,p})_j x^j,$$

and recall that by Lemma 4.1 and Eq. (4.2),

$$\begin{aligned} P_{l,p}(x) \quad &= \quad h_l(x) + \sum_{j=0}^{l} (r_{l,p})_j x^j, \\ P_{l,p}^{'}(x) \quad &= \quad \sqrt{l}h_{l-1}(x) + \sum_{j=0}^{l-1} (r_{l,p})_{j+1}(j+1)x^j, \\ \max_{0\le j\le l} |(r_{l,p})_j| \quad &= \quad \mathcal{O}_l(1)p^{-1}, \end{aligned}$$

and as a result

$$\max_{0\le j\le l} |(c_{l,p})_j| = \mathcal{O}_l(1).$$

For the claim in the lemma to hold, it suffices to show that each term in Eq. (D.1) is $\mathcal{O}_l(1)p^{-2}$. Notice that when $k+j$ is odd, $\mathbb{E}\eta_3^{k+j}$ vanishes. Then we have the following cases:

1. $k = j = 0$: In $\mathbb{E}\left(P_{l,p}(|X_1|\zeta_2)P_{l,p}(|X_1|\zeta_4)\right)$, since ζ_2, ζ_4 and $|X_1|$ are independent, we can take expectation with respect to ζ_2, ζ_4 first, which gives that
$$\mathbb{E}\left(P_{l,p}(|X_1|\zeta_2)P_{l,p}(|X_1|\zeta_4)\right) = \mathbb{E}q_{0l}(|X_1|)^2,$$
where
$$q_{0l}(r) = \sum_{j=0}^{l}(c_{l,p})_j\mathbb{E}\zeta^j \cdot r^j, \quad \zeta \sim \mathcal{N}(0,1),$$
and $q_{0l}(1) = \sum_{j=0}^{l}(c_{l,p})_j\mathbb{E}\zeta^j = \mathbb{E}h_l(\zeta) + \sum_{j=0}^{l}(r_{l,p})_j\mathbb{E}\zeta^j$. Because $l \geq 2$, $\mathbb{E}h_l(\zeta) = 0$, and then
$$|q_{0l}(1)| \leq \sum_{j=0}^{l}|(r_{l,p})_j|\mathbb{E}\zeta^j \leq \mathcal{O}_l(1)p^{-1}.$$
Applying Lemma D.2 to $q = q_{0l}$ and $Z = |X_1|$, together with Lemma D.5, we have that
$$\mathbb{E}q_{0l}(|X_1|)^2 \leq (\mathcal{O}_l(1)p^{-1})^2 + \max_{0\leq k\leq 2l}|\mathbb{E}|X_1|^k - 1| \cdot \mathcal{O}_l(1) = \mathcal{O}_l(1)p^{-1}.$$
For same reason, we have that $\left|\mathbb{E}\left(P_{l,p}(|\tilde{X}_3|\tilde{\zeta}_2)P_{l,p}(|\tilde{X}_3|\tilde{\zeta}_4)\right)\right| = \mathcal{O}_l(1)p^{-1}$. This implies the $\mathcal{O}_l(1)p^{-2}$ bound for the term of $j = k = 0$.
2. $k = 0, j = 2$ or $k = 2, j = 0$: $\mathbb{E}\eta_3^2 = p^{-1}$. Taking $k = 0, j = 2$ as an example, for $\mathbb{E}\left(P_{l,p}(|X_1|\zeta_2)\zeta_4^2P_{l,p}(|X_1|\zeta_4)\right)$ and $\mathbb{E}\left(P_{l,p}(|\tilde{X}_3|\tilde{\zeta}_2)P_{l,p}^{(2)}(|\tilde{X}_3|\tilde{\zeta}_4)\right)$, a similar argument as above gives a bound of $\mathcal{O}_l(1)p^{-1}$ for each of them. Then the term of $k = 0, j = 2$ is bounded by $\mathcal{O}_l(1)p^{-3}$.
3. $k = 1, j = 1$: $\mathbb{E}\eta_3^2 = p^{-1}$. For $\mathbb{E}\left(\zeta_2P_{l,p}(|X_1|\zeta_2)\zeta_4P_{l,p}(|X_1|\zeta_4)\right)$, as $l \geq 2$, $\mathbb{E}\zeta h_l(\zeta) = 0$, and a similar argument as above gives a bound of $\mathcal{O}_l(1)p^{-1}$. For $\mathbb{E}\left(P_{l,p}'(|\tilde{X}_3|\tilde{\zeta}_2)P_{l,p}'(|\tilde{X}_3|\tilde{\zeta}_4)\right)$, as $\mathbb{E}h_{l-1}(\zeta) = 0$, we have a abound of $\mathcal{O}_l(1)p^{-1}$. Then the term of $k = 1, j = 1$ is bounded by $\mathcal{O}_l(1)p^{-3}$.
4. $2 \leq k, j \leq l$ (and $k+j$ is even): $\mathbb{E}\eta_3^{k+j} = (k+j-1)!!p^{-(k+j)/2}$ which is at least $\mathcal{O}_l(1)p^{-2}$. For similar reason as above, both $\mathbb{E}\left(\zeta_2^kP_{l,p}(|X_1|\zeta_2)\zeta_4^kP_{l,p}(|X_1|\zeta_4)\right)$ and $\mathbb{E}\left(P_{l,p}^{(k)}(|\tilde{X}_3|\tilde{\zeta}_2)P_{l,p}^{(j)}(|\tilde{X}_3|\tilde{\zeta}_4)\right)$ can be shown to be $\mathcal{O}_l(1)$. Then the term is bounded by $\mathcal{O}_l(1)p^{-2}$.

□

Proof of Lemma 4.4. We have

$$\begin{aligned} & m_A(z) - m_B(z) \\ &= \frac{1}{n}\left(\mathbf{Tr}((A-zI)^{-1}) - \mathbf{Tr}((B-zI)^{-1})\right) \\ &= \frac{1}{n}\mathbf{Tr}((A-zI)^{-1}(B-A)(B-zI)^{-1}), \end{aligned}$$

thus

$$\begin{aligned} & \mathbb{E}|m_A(z) - m_B(z)|^2 \\ &= \mathbb{E}\frac{1}{n^2}(\mathbf{Tr}((A-zI)^{-1}(B-A)(B-zI)^{-1}))^2 \\ &\leq \mathbb{E}\frac{1}{n^2}\mathbf{Tr}(((B-zI)^{-1}(A-zI)^{-1})^2)\mathbf{Tr}((B-A)^2) \\ &\leq \mathbb{E}\frac{1}{n^2}\frac{n}{v^4}\sum_{i,j=1}^{n}|A_{ij}-B_{ij}|^2, \\ &\leq \frac{1}{v^4 n}\sum_{i,j=1}^{n}\mathbb{E}(f_A(X_i^T X_j;p) - f_B(X_i^T X_j;p))^2 \\ &\leq \frac{1}{v^4 n}n^2 p^{-1}\epsilon = \mathcal{O}(1)\epsilon. \square \end{aligned}$$

Lemma D.3. *Let Ω_δ be defined as in Eq. (4.12),*

$$\mathbf{Pr}(\Omega_\delta^c) \leq \mathcal{O}(1)p^{-7}.$$

Proof. For η_i we have the concentration inequality Eq. (B.8); For each $\tilde{\xi}_{ij}$, we write it as

$$\tilde{\xi}_{ij} = |\tilde{X}_i|\tilde{\eta}_{ij},$$

where $\tilde{\eta}_{ij}$ has marginal distribution $\mathcal{N}(0, p^{-1})$ and is independent of $|\tilde{X}_i|$. With inequality Eq. (B.4) which also holds for $|X_i|$ in place of $|X_n|$, we have

$$\begin{aligned} \mathbf{Pr}[|\tilde{X}_i||\tilde{\eta}_{ij}| > \delta] &\leq \mathbf{Pr}\left[|\tilde{X}_i|^2 > 1 + \sqrt{\frac{40\ln p}{p}}\right] \\ &\quad + \mathbf{Pr}\left[|\tilde{X}_i||\tilde{\eta}_{ij}| > \delta, |\tilde{X}_i|^2 < 1 + \sqrt{\frac{40\ln p}{p}}\right] \\ &\leq p^{-9} + \mathbf{Pr}[|\tilde{\eta}_{ij}| > \frac{\delta}{1.01}] \\ &\leq p^{-9} + \frac{1}{\sqrt{2}}p^{-9}, \end{aligned}$$

thus

$$\mathbf{Pr}[|\tilde{\xi}_{ij}| > \delta] < \mathcal{O}(1)p^{-9}.$$

Then, a union bound gives

$$\begin{aligned}\mathbf{Pr}(\Omega_\delta^c) &\le (n-1)\mathbf{Pr}[|\eta_i| > \delta] \\ &\quad + \frac{(n-1)(n-2)}{2}\mathbf{Pr}[|\tilde{\xi}_{ij}| > \delta] + \mathbf{Pr}[\big||X_n|^2 - 1\big| > \sqrt{2}\delta] \\ &\le \mathcal{O}(1)p^{-9} + \mathcal{O}(1)p^{-7} + p^{-9} = \mathcal{O}(1)p^{-7}.\square\end{aligned}$$

Lemma D.4. *Notation as in Sec. 4.3. r_2 defined in Eq. (4.20) satisfies*

$$\mathbb{E}|r_2| \cdot \mathbf{1}_{\Omega_\delta} \le \mathcal{O}_L(1)M^2 p^{-1/2}.$$

Proof. From Eq. (4.21), firstly,

$$r_{2,1} = f_{(2)}^T(\tilde{A}^{(n)} - zI_{n-1})^{-1}(|X_n|\eta)$$

satisfies $\mathbb{E}|r_{2,1}| \le \mathcal{O}_L(1)p^{-1/2}$ by a moment bound: recall the definition of ξ_{in} as in Eq. (2.12), and that $f_{>1}(\xi)$ is a linear combination of rescaled and renormalized Hermite-like polynomials of degree ≥ 2. Also, $\mathbb{E}|X_n|^{2m} = 1 + \mathcal{O}_m(1)p^{-1}$ (Eq. (D.3)), and $|X_n|$ is independent from η_i's and $\tilde{X}_i$'s. Denote $\tilde{B} = (\tilde{A}^{(n)} - zI_{n-1})^{-1}$. By taking expectation over $|X_n|$ first and then over η_i's, we have

$$\begin{aligned}\mathbb{E}|r_{2,1}|^2 &= \mathbb{E}\left|\sum_{i_1,i_2=1}^{n-1} f_{>1}(\xi_{i_1 n})\xi_{i_2 n}\tilde{B}_{i_1 i_2}\right|^2 \\ &= \mathbb{E}\sum_{i_1,i_2}\sum_{i_1',i_2'} f_{>1}(\xi_{i_1 n})\xi_{i_2 n} f_{>1}(\xi_{i_1' n})\xi_{i_2' n}\tilde{B}_{i_1 i_2}\overline{\tilde{B}_{i_1' i_2'}} \\ &= \{i_1 = i_2 = i_1' = i_2'\} + \{i_1, i_2 = i_1' = i_2', \text{or } i_1' \text{as } i_1\} \\ &\quad + \{i_2 = i_2', i_1, i_1'\} + \{i_1 = i_2, i_1' = i_2', \text{or } i_1' \text{as } i_1\} + \{i_1 = i_1', i_2 = i_2'\} \\ &= \mathcal{O}_L(1)p^{-1} + \nu_{>1,p}p^{-2}\mathbb{E}\mathbf{Tr}(\overline{\tilde{B}}^T\tilde{B}) \\ &\le \mathcal{O}_L(1)p^{-1} + \mathcal{O}(1)\cdot p^{-2}\frac{n}{v^2} = \mathcal{O}_L(1)p^{-1}.\end{aligned}$$

By $\{i_1, i_2 = i_1' = i_2'\}$ we denote the term in summation where the last three indices take the same value while i_1 is distinct from them, and similar for others.

Secondly,

$$\begin{aligned}r_{2,2} &= (f_{(2)}^T(\tilde{A}^{(n)} - zI_{n-1})^{-1}(|X_n|\eta))(a_1(p)\eta^T(\hat{A}^{(n)} - zI_{n-1})^{-1}\eta) \\ &= r_{2,1}(a_1(p)\eta^T(\hat{A}^{(n)} - zI_{n-1})^{-1}\eta),\end{aligned}$$

where

$$\begin{aligned}|a_1(p)\eta^T(\hat{A}^{(n)} - zI_{n-1})^{-1}\eta| \cdot \mathbf{1}_{\Omega_\delta} &\le \mathcal{O}(1)s((\hat{A}^{(n)} - zI_{n-1})^{-1})||\eta||^2 \cdot \mathbf{1}_{\Omega_\delta} \\ &\le \mathcal{O}(1)M^2 = \mathcal{O}(1)M^2,\end{aligned}$$

thus

$$\mathbb{E}|r_{2,2}| \cdot \mathbf{1}_{\Omega_\delta} \leq \mathcal{O}(1)M^2\mathbb{E}|r_{2,1}| \leq \mathcal{O}(1)M^2 \cdot \mathcal{O}_L(1)p^{-1/2} = \mathcal{O}_L(1)M^2p^{-1/2}.$$

Then

$$\mathbb{E}|r_2| \cdot \mathbf{1}_{\Omega_\delta} \leq \mathcal{O}(1)(\mathbb{E}|r_{2,1}| + \mathbb{E}|r_{2,2}| \cdot \mathbf{1}_{\Omega_\delta}) \leq \mathcal{O}_L(1)M^2p^{-1/2}. \qquad \square$$

Lemma D.5. *Let $X \sim \mathcal{N}(0, p^{-1}I_p)$, then for any positive integer k,*

$$\mathbb{E}|X|^k = 1 + \mathcal{O}_k(1)p^{-1}, \quad p \to \infty. \tag{D.2}$$

Proof of Lemma D.5. When $k = 2$, $\mathbb{E}|X|^2 = 1$.

When $k = 1$,

$$\mathbb{E}|X| \leq \sqrt{\mathbb{E}|X|^2} = 1.$$

At the same time, notice that $\sqrt{p}|X| \sim \chi(p)$, thus by the expression of the expectation of the Chi distribution we have that

$$\mathbb{E}|X| = \sqrt{\frac{2}{p}}\frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})}.$$

When p is even, Eq. (D.2) is verified by the formula

$$\Gamma(m + \frac{1}{2}) = \frac{(2m-1)!!\sqrt{\pi}}{2^m}$$

and the Stirling Formula

$$n! = \sqrt{2\pi n}(n/e)^n(1 + \mathcal{O}(n^{-1})).$$

When p is odd,

$$1 \geq \mathbb{E}|X| = \mathbb{E}\frac{\chi(p)}{\sqrt{p}} \geq \mathbb{E}\frac{\chi(p-1)}{\sqrt{p}},$$

and thus is reduced to the case when p is even.

When $k = 2m - 1, m \geq 2$,

$$1 = (\mathbb{E}|X|^2)^{(2m-1)/2} \leq \mathbb{E}|X|^{2m-1} \leq \sqrt{\mathbb{E}|X|^2\mathbb{E}|X|^{4m-4}} = \sqrt{\mathbb{E}|X|^{4m-4}}$$

and thus is reduced to the case where $k \geq 4$ and is even, which is analyzed below.

When $k = 2m, m \geq 2$, let

$$y := |X|^2 = \frac{1}{p}\sum_{j=1}^{p} x_j^2, \quad x_1, \cdots, x_p \text{ i.i.d. } \sim \mathcal{N}(0,1),$$

and the characteristic function of y is

$$\phi(t) := \mathbb{E}e^{ty} = (\mathbb{E}e^{tx_1^2/p})^p = \left(1 - \frac{2t}{p}\right)^{-p/2}, \quad 0 < \frac{t}{p} < \frac{1}{2}.$$

This gives that

$$\begin{aligned}\mathbb{E}|X|^{2m} = \mathbb{E}y^m &= \frac{d^m}{dt^m}\phi(t)\Big|_{t=0} \\ &= \prod_{j=1}^{m-1}\left(1+\frac{2j}{p}\right) \\ &= 1+\frac{m(m-1)}{p}+\mathcal{O}_m(1)p^{-2}.\square \end{aligned} \tag{D.3}$$

Lemma D.6. *Notations as in Sec. 4.4,*

$$\mathbb{E}(\xi_p')^k = \begin{cases}(k-1)!!+\mathcal{O}_k(1)p^{-1}, & k \text{ even;}\\ 0, & k \text{ odd.}\end{cases}$$

Proof. The odd moments vanish since the distribution of ξ_p' is symmetric with respect to 0. For even moments, let $k=2m$. Let $\xi_p = \sqrt{p}X^TY$ where X and Y are i.i.d $\mathcal{N}(0,p^{-1}I_p)$, and we have that ξ_p and $\xi_p'|X||Y|$ observe the same probability distribution. Notice that ξ_p', $|X|$ and $|Y|$ are independent, so

$$\mathbb{E}\xi_p^{2m} = \mathbb{E}|X|^{2m}\mathbb{E}|Y|^{2m}\mathbb{E}(\xi_p')^{2m} = (\mathbb{E}|X|^{2m})^2\mathbb{E}(\xi_p')^{2m}.$$

The claim follows by Eq. (4.3) and Eq. (D.3). □