The Annals of Statistics
2012, Vol. 40, No. 3, 1737–1763
DOI: 10.1214/12-AOS1022

EDGE UNIVERSALITY OF CORRELATION MATRICES

BY NATESH S. PILLAI[1] AND JUN YIN[2]

Harvard University and University of Wisconsin–Madison

Let $\widetilde{X}_{M\times N}$ be a rectangular data matrix with independent real-valued entries $[\widetilde{x}_{ij}]$ satisfying $\mathbb{E}\widetilde{x}_{ij}=0$ and $\mathbb{E}\widetilde{x}_{ij}^2=\frac{1}{M}$, $N, M\to\infty$. These entries have a subexponential decay at the tails. We will be working in the regime $N/M=d_N$, $\lim_{N\to\infty}d_N\neq 0,1,\infty$. In this paper we prove the edge universality of correlation matrices $X^\dagger X$, where the rectangular matrix X (called the standardized matrix) is obtained by normalizing each column of the data matrix $\widetilde{X}$ by its Euclidean norm. Our main result states that asymptotically the k-point ($k\geq 1$) correlation functions of the extreme eigenvalues (at both edges of the spectrum) of the correlation matrix $X^\dagger X$ converge to those of the Gaussian correlation matrix, that is, Tracy–Widom law, and, thus, in particular, the largest and the smallest eigenvalues of $X^\dagger X$ after appropriate centering and rescaling converge to the Tracy–Widom distribution. The asymptotic distribution of extreme eigenvalues of the Gaussian correlation matrix has been worked out only recently. As a corollary of the main result in this paper, we also obtain that the extreme eigenvalues of Gaussian correlation matrices are asymptotically distributed according to the Tracy–Widom law. The proof is based on the comparison of Green functions, but the key obstacle to be surmounted is the strong dependence of the entries of the correlation matrix. We achieve this via a novel argument which involves comparing the moments of product of the entries of the standardized data matrix to those of the raw data matrix. Our proof strategy may be extended for proving the edge universality of other random matrix ensembles with dependent entries and hence is of independent interest.

1. Introduction.

The aim of this paper is to prove the edge universality of correlation matrices. The data matrix $\widetilde{X}=(\widetilde{x}_{ij})$ is an $M\times N$ matrix with independent centered real-valued entries. The entries in each column j all are assumed to be identically distributed:

$$\widetilde{x}_{ij}=M^{-1/2}q_{ij},\qquad \mathbb{E}q_{ij}=0,\qquad \mathbb{E}q_{ij}^2=\sigma_j^2,\qquad 1\leq i\leq M. \tag{1.1}$$

Furthermore, the entries q_{ij} have a subexponential decay, that is, there exists a constant $\vartheta>0$ such that for $u>1$,

$$\mathbb{P}(|q_{ij}|>u\sigma_j)\leq\vartheta^{-1}\exp(-u^\vartheta). \tag{1.2}$$

Received February 2012; revised May 2012.

[1] Supported by NSF Grant DMS-11-07070.

[2] Supported by NSF Grant DMS-10-01655.

MSC2010 subject classifications. 15B52, 82B44.

Key words and phrases. Covariance matrix, correlation matrix, Marcenko–Pastur law, universality.

We will be working the regime

$$d = d_N = N/M, \qquad \lim_{N\to\infty} d \neq 0, 1, \infty. \tag{1.3}$$

Thus, without loss of generality, henceforth we will assume that for some small constant θ, for all $N \in \mathbb{N}$,

$$\theta < d_N < \theta^{-1} \quad \text{and} \quad \theta < |d_N - 1|.$$

Notice that all our constants may depend on θ and ϑ, but we will subsume this dependence in the notation.

For a Euclidean vector $a \in \mathbb{R}^M$, define the ℓ_2 norm

$$\|a\|_2 := \left(\sum_{i=1}^M a_i^2\right)^{1/2}.$$

The matrix $\widetilde{X}^\dagger \widetilde{X}$ is the usual covariance matrix. The jth column of $\widetilde{X}$ is denoted by $\widetilde{\mathbf{x}}_j$. Define the matrix $M \times N$ matrix $X = (x_{ij})$

$$x_{ij} := \widetilde{x}_{ij}/\|\widetilde{\mathbf{x}}_j\|_2. \tag{1.4}$$

The $(N \times N)$ matrix $X^\dagger X$ is called the correlation matrix.[3] Using the identity $\mathbb{E}x_{ij}^2 = \frac{1}{M}\mathbb{E}\sum_i x_{ij}^2$, we have

$$\mathbb{E}x_{ij}^2 = M^{-1}.$$

Since we are mainly interested in correlation matrices, without loss of generality, henceforth we will assume that

$$\sigma_j^2 = 1, \qquad 1 \le j \le N.$$

Covariance matrices are ubiquitous in modern multivariate statistics where the advance of technology has led to a profusion of high-dimensional data sets. See [17–19, 24] and the references therein for motivation and applications in a wide variety of fields. Correlation matrices are sometimes preferred in certain statistical applications. For instance, the classic exploratory method Principal Component Analysis (PCA) is not invariant to change of scale in the matrix entries. Therefore, it is often recommended first to standardize the matrix entries and then perform PCA on the resulting correlation matrix [17].

Recent progress in random matrix theory has led to a wealth of techniques for proving universality of various matrix ensembles (see [3–13, 16, 20, 21, 26, 27] and the references therein). Here the word universality refers to the phenomenon that the asymptotic distributions of various functionals of covariance/correlation matrices (such as eigenvalues, eigenvector, etc.) are identical to those Gaussian

[3] Some authors prefer to call this the standardized covariance matrix, but we chose this terminology from the statistical literature [17].

covariance/correlation matrices. Thus, harnessing these methods to obtain universality results in statistical problems is an important step, since these results let us calculate the exact asymptotic distributions of various test statistics without having restrictive distributional assumptions of the matrix entries. For instance, an important consequence of universality is that in some cases one can perform various hypothesis tests under the assumption that the matrix entries are *not* normally distributed but use the same test statistic as in the Gaussian case.

In this context, in a recent paper [24] we studied the asymptotic distribution of the eigenvalues of the covariance matrix $\widetilde{X}^\dagger \widetilde{X}$ under the assumptions of (1.1) and (1.2). In [24], we proved that the Stieltjes transform of the empirical eigenvalue distribution of the sample covariance matrix is given by the Marcenko–Pastur law [22] uniformly up to the edges of the spectrum with an error of order $(N\eta)^{-1}$, where η is the imaginary part of the spectral parameter in the Stieltjes transform. From this strong local Marcenko–Pastur law, we derived the following results: (1) rigidity of eigenvalues (2) delocalization of eigenvectors (3) universality of eigenvalues in the bulk and (4) universality of eigenvalues at the edges. Furthermore, in our proof of edge universality of eigenvalues for covariance matrices (see Theorem 7.5 of [24]), we gave a sufficient criterion for checking whether two matrices of form $Q^\dagger Q$ (Q is a data matrix) have the same asymptotic eigenvalue distribution at the edge (see Section 3 for details). Here $Q^\dagger Q$ could be quite general, including covariance and correlation matrices.

Verifying the above criteria for correlation matrices is much more complicated, owing to the fact that even if it has the same form $X^\dagger X$ as above, the matrix entries of X are not independent. Fortunately in [24], as a byproduct, we also proved the strong Marcenko–Pastur law, the rigidity of eigenvalues and delocalization of eigenvectors of correlation matrices (see Lemma 2.3 in Section 2 below or Theorem 1.5 of [24]). In this paper, we complete the research program initiated in [24] by proving the edge universality of correlation matrices. There are not many papers which study the asymptotics of the correlation matrices as compared to the relatively large literature on covariance matrices. The asymptotic distribution of the largest (appropriately rescaled) eigenvalue of the Gaussian correlation matrix was only very recently established by [1]. As will be explained below, we also obtain this result as a special case of our main result and, more importantly, *we do not need this result in our proof* (see Remark 1.3). The almost sure convergence of the largest and smallest eigenvalues of the correlation matrix was established in [15]. The very recent paper [1], relying on our results in [24], shows that the asymptotic distribution of the largest or smallest eigenvalue of the correlation matrix is given by the Tracy–Widom law, under the assumption that the data matrix X satisfies (1.1) and its entries have *symmetric* distributions. In particular, the authors in [1] use the above mentioned sufficiency criteria for edge universality developed in [24]. Furthermore, the assumption that the matrix entries are symmetric is very restrictive and not natural in statistical applications. In this paper we will build on our previous work [24] and prove edge universality of correlation matrices just under the assumptions (1.1) and (1.2). Furthermore, we believe that all of our main results should hold if one replaces the subexponential tail decay of the matrix entries by a uniform bound on the pth moment ($p > 4$) of the matrix entries (e.g., $p = 13$ will suffice), as proved in [3] for Wigner matrices.

The central ideas in this paper are based on the general machinery for proving universality established in a series of recent papers [3–13, 20, 21], where the authors Yau, Erdős et al. study the distribution of eigenvalues and eigenvectors by studying the Green's functions (resolvent) of the random matrices.

The proof of this paper is based on the comparison of Green's functions first initiated in [12], but, as mentioned earlier, the key obstacle to be surmounted is the strong dependence of the entries of the correlation matrix. We achieve this via a novel argument which involves comparing the moments of the product of the entries of the standardized data matrix to those of the raw data matrix (see Section 3 for a summary of the key ideas). Our proof strategy may be extended for proving the edge universality of other random matrix ensembles with dependent entries and hence is of independent interest. Furthermore, it will be interesting to see if bulk universality of correlation matrices can be established using the methods developed in this paper.

Let us state the main result now. We denote λ_i, $1 \le i \le N$, as the eigenvalues of $X^\dagger X$ and $\lambda_\alpha = 0$ for $\min\{N, M\} + 1 \le \alpha \le \max\{N, M\}$. We order them as

$$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_{\max\{M,N\}} \ge 0.$$

Analogously, let $\widetilde{\lambda}_\alpha$ denote the eigenvalues values of the matrix $\widetilde{X}^\dagger \widetilde{X}$.

The following is the main result of this paper. It shows that the largest and smallest k eigenvalues of the correlation matrix, after appropriate centering and rescaling, converge in distribution to those of the corresponding covariance matrix.

THEOREM 1.1 (Edge universality). *Let X and $\widetilde{X}$, respectively, denote the correlation and covariance matrix as defined in* (1.1)–(1.4). *For any fixed $k \in \mathbb{N}$, there exists $\varepsilon > 0$ and $\delta > 0$ such that for any $\{s_1, s_2, \dots, s_k\} \in \mathbb{R}$ (which may depend on N), there exists $N_0 \in \mathbb{N}$ independent of $s_1, s_2, \dots, s_k$ such that for all $N \ge N_0$, we have*

$$\begin{aligned}
&\mathbb{P}\big(N^{2/3}(\widetilde{\lambda}_1 - \lambda_+) \le s_1 - N^{-\varepsilon}, \dots, N^{2/3}(\widetilde{\lambda}_k - \lambda_+) \le s_k - N^{-\varepsilon}\big) - N^{-\delta} \\
&\quad \le \mathbb{P}\big(N^{2/3}(\lambda_1 - \lambda_+) \le s_1, \dots, N^{2/3}(\lambda_k - \lambda_+) \le s_k\big) \\
&\quad \le \mathbb{P}\big(N^{2/3}(\widetilde{\lambda}_1 - \lambda_+) \le s_1 + N^{-\varepsilon}, \dots, N^{2/3}(\widetilde{\lambda}_k - \lambda_+) \le s_k + N^{-\varepsilon}\big) \\
&\qquad + N^{-\delta}.
\end{aligned} \tag{1.5}$$

An analogous result holds for the k smallest eigenvalues.

In [14, 23] and [25], Peche, Soshnikov and Sodin proved that for some covariance matrices (including the Wishart matrix), the largest and smallest k eigenvalues after appropriate centering and rescaling converge in distribution to the Tracy–Widom law[4] whose density is a smooth function. Combining with our recent result on the universality of covariance matrices in [24], we have the following immediate corollary for Theorem 1.1:

COROLLARY 1.2. *Let X denote the correlation matrix as defined in* (1.1)–(1.4). *For any fixed $k > 0$, we have*

$$\begin{aligned}&\biggl(\frac{M\lambda_1 - (\sqrt{N}+\sqrt{M})^2}{(\sqrt{N}+\sqrt{M})(1/\sqrt{N}+1/\sqrt{M})^{1/3}}, \dots,\\&\qquad\frac{M\lambda_k - (\sqrt{N}+\sqrt{M})^2}{(\sqrt{N}+\sqrt{M})(1/\sqrt{N}+1/\sqrt{M})^{1/3}}\biggr)\\&\qquad\longrightarrow \mathrm{TW}_1,\end{aligned}$$

where TW_1 *denotes the Tracy–Widom distribution. An analogous statement holds for the k-smallest* (*nontrivial*) *eigenvalues.*

REMARK 1.3. Thus, as a special case, we also obtain the TW law for the Gaussian correlation matrices.

Although the current paper builds on our recent work [24], it is mostly self-contained and for the reader's convenience, we will recall all of the needed results from [24]. The rest of the paper is organized as follows. In Section 2, after establishing some notation, we give the key results establishing the strong Marcenko–Pastur law and rigidity of eigenvalues for correlation matrices, as obtained from [24]. In Section 3 we give a brief proof sketch illustrating the key ideas. In Section 4 we give the proof of the main results and in Section 5 we prove some technical lemmas which constitute the key ingredients in the proof of the main result. For the rest of the paper the letter C will denote a generic constant whose value might change from one line to the next, but will be independent of everything else. The notation $O_\varepsilon(N^a)$ will be used to denote $O(N^{a+C\varepsilon})$.

2. Preliminaries. We will adopt the notation used in this paper from [24]. Define the Green function of $X^\dagger X$ by

$$G_{ij}(z) = \left(\frac{1}{X^\dagger X - z}\right)_{ij}, \qquad z = E + i\eta, \qquad E \in \mathbb{R}, \eta > 0. \tag{2.1}$$

[4] Here we use the term Tracy–Widom law as in [25].

The Stieltjes transform of the empirical eigenvalue distribution of $X^\dagger X$ is given by

$$m(z) := \frac{1}{N}\sum_j G_{jj}(z) = \frac{1}{N}\operatorname{Tr}\frac{1}{X^\dagger X - z}. \tag{2.2}$$

Recall that $d = N/M$ from (1.3) and define

$$\lambda_\pm := \bigl(1 \pm \sqrt{d}\bigr)^2. \tag{2.3}$$

The Marcenko–Pastur (henceforth abbreviated by MP) law is given by

$$\varrho_W(x) = \frac{1}{2\pi d}\sqrt{\frac{[(\lambda_+ - x)(x-\lambda_-)]_+}{x^2}}. \tag{2.4}$$

We define $m_W(z)$, $z \in \mathbb{C}$, as the Stieltjes transform of ϱ_W, that is,

$$m_W(z) = \int_{\mathbb{R}} \frac{\varrho_W(x)}{(x-z)}\,dx. \tag{2.5}$$

The function m_W depends on d and has the closed form solution

$$m_W(z) = \frac{1-d-z+i\sqrt{(z-\lambda_-)(\lambda_+-z)}}{2dz}, \tag{2.6}$$

where $\sqrt{\ }$ denotes the square root on a complex plane whose branch cut is the negative real line. We also define the classical location of the eigenvalues with ρ_W as follows:

$$\int_{\gamma_j}^{\lambda_+} \varrho_W(x)\,dx = \int_{\gamma_j}^{+\infty} \varrho_W(x)\,dx = j/N. \tag{2.7}$$

Define the parameter

$$\varphi := (\log N)^{\log\log N}. \tag{2.8}$$

DEFINITION 2.1 (High probability events). Let $\zeta > 0$. We say that an event Ω holds with *ζ-high probability* if there exists a constant $C > 0$ such that

$$\mathbb{P}(\Omega^c) \le N^C \exp(-\varphi^\zeta) \tag{2.9}$$

for large enough N.

Let us first give the following large deviation lemma for independent random variables (see [12], Appendix B for a proof).

LEMMA 2.2 (Large deviation lemma). *Suppose, for $1 \le i \le M$, a_i are independent, mean 0 complex variables, with $\mathbb{E}|a_i|^2 = \sigma^2$ and have a subexponential* *decay as in* (1.2). *Then there exists a constant* $\rho \equiv \rho(\vartheta) > 1$ *such that*, *for any* $\zeta > 0$ *and for any* $A_i \in \mathbb{C}$ *and* $B_{ij} \in \mathbb{C}$, *the bounds*

$$\sum_{i=1}^{M} a_i A_i \leq (\log M)^{\rho\zeta} \sigma \|A\|, \tag{2.10}$$

$$\left|\sum_{i=1}^{M} \bar{a}_i B_{ii} a_i - \sum_{i=1}^{M} \sigma^2 B_{ii}\right| \leq (\log M)^{\rho\zeta} \sigma^2 \left(\sum_{i=1}^{M} |B_{ii}|^2\right)^{1/2}, \tag{2.11}$$

$$\left|\sum_{i\neq j} \bar{a}_i B_{ij} a_j\right| \leq (\log M)^{\rho\zeta} \sigma^2 \left(\sum_{i\neq j} |B_{ij}|^2\right)^{1/2} \tag{2.12}$$

hold with ζ*-high probability.*

It can be easily seen that for any fixed $j \leq N$, the random variables defined by $a_i = x_{ij}$, $1 \leq i \leq M$, satisfy the large deviation bounds (2.10), (2.11) and (2.12), for any $A_i \in \mathbb{C}$ and $B_{ij} \in \mathbb{C}$ and $\zeta > 0$.

Thus, the main result of [24] (see Theorem 1.5 of [24]) is applicable for the correlation matrix X, yielding the following strong local MP law and rigidity of eigenvalues:

LEMMA 2.3 (Strong local Marcenko–Pastur law and rigidity of the eigenvalues of the correlation matrix). *Let* $X = [x_{ij}]$ *be the correlation matrix given by* (1.4). *Then for any* $\zeta > 0$ *there exists a constant* C_ζ *such that the following events hold with* ζ*-high probability.*

(i) *The Stieltjes transform of the empirical eigenvalue distribution of* $X^\dagger X$ *satisfies*

$$\bigcap_{z\in mS(C_\zeta)} \left\{ |m(z) - m_W(z)| \leq \varphi^{C_\zeta} \frac{1}{N\eta} \right\}, \tag{2.13}$$

where $mS(C_\zeta)$ *defined as the set*

$$mS(C_\zeta) := \{z \in \mathbb{C} : \mathbf{1}_{d>1}(\lambda_-/5) \leq E \leq 5\lambda_+, \varphi^{C_\zeta} N^{-1} \leq \eta \leq 10(1+d)\}.$$

(ii) *The individual matrix elements of the Green function satisfy*

$$\bigcap_{z\in mS(C_\zeta)} \left\{ |G_{ij}(z) - m_W(z)\delta_{ij}| \leq \varphi^{C_\zeta} \left(\sqrt{\frac{\Im m_W(z)}{N\eta}} + \frac{1}{N\eta} \right) \right\}. \tag{2.14}$$

(iii) *The smallest nonzero and largest eigenvalues of* $X^\dagger X$ *satisfy*

$$\lambda_- - N^{-2/3}\varphi^{C_\zeta} \leq \min_{j\leq \min\{M,N\}} \lambda_j \leq \max_j \lambda_j \leq \lambda_+ + N^{-2/3}\varphi^{C_\zeta}. \tag{2.15}$$

(iv) *Rigidity of the eigenvalues*: *recall* γ_j *in* (2.7). *For any* $1 \le j \le \min\{M, N\}$, *let* $\widetilde{j} = \min\{\min\{N, M\} + 1 - j, j\}$. *Then*

$$|\lambda_j - \gamma_j| \le \varphi^{C_\zeta} N^{-2/3} \widetilde{j}^{-1/3}. \tag{2.16}$$

We conclude this section with the following theorem quoted from [24] (see Theorem 1.7 in [24]) on edge universality of covariance matrices, which is also needed for our proof of the edge universality of the correlation matrix. Define two independent matrices $\widetilde{X}^{\mathbf{v}} = [\widetilde{x}_{ij}^{\mathbf{v}}]$, $\widetilde{X}^{\mathbf{w}} = [\widetilde{x}_{ij}^{\mathbf{w}}]$ with the entries $\widetilde{x}_{ij}^{\mathbf{v}}, \widetilde{x}_{ij}^{\mathbf{w}}$ satisfying (1.1) and (1.2) and the entries $\widetilde{x}_{ij}^{\mathbf{v}}, \widetilde{x}_{ij}^{\mathbf{w}}$ are mutually independent. Henceforth, we will write $\mathbb{E}^{\mathbf{v}}, \mathbb{P}^{\mathbf{v}}$ ($\mathbb{E}^{\mathbf{w}}, \mathbb{P}^{\mathbf{w}}$) to indicate that the expectation and probability are computed for the ensemble $\widetilde{X}^{\mathbf{v}}$, ($\widetilde{X}^{\mathbf{w}}$).

THEOREM 2.4 (Universality of extreme eigenvalues of covariance matrices). *There exists* $\varepsilon > 0$ *and* $\delta > 0$ *such that for any* $s \in \mathbb{R}$ (*which may depend on* N) *there exists* $N_0 \in \mathbb{N}$ *independent of* s *such that for all* $N \ge N_0$, *we have*

$$\begin{aligned} &\mathbb{P}^{\mathbf{v}}\big(N^{2/3}\big(\widetilde{\lambda}_1^{\mathbf{v}} - \lambda_+\big) \le s - N^{-\varepsilon}\big) - N^{-\delta} \\ &\quad \le \mathbb{P}^{\mathbf{w}}\big(N^{2/3}\big(\widetilde{\lambda}_1^{\mathbf{w}} - \lambda_+\big) \le s\big) \\ &\quad \le \mathbb{P}^{\mathbf{v}}\big(N^{2/3}\big(\widetilde{\lambda}_1^{\mathbf{v}} - \lambda_+\big) \le s + N^{-\varepsilon}\big) + N^{-\delta}. \end{aligned} \tag{2.17}$$

An analogous result holds for the smallest eigenvalues $\widetilde{\lambda}_{\min\{M,N\}}^{\mathbf{v}}$ *and* $\widetilde{\lambda}_{\min\{M,N\}}^{\mathbf{w}}$.

As remarked in [24], Theorem 2.4 can be extended to finite correlation functions of extreme eigenvalues as follows:

$$\begin{aligned} &\mathbb{P}^{\mathbf{v}}\big(N^{2/3}\big(\widetilde{\lambda}_1^{\mathbf{v}} - \lambda_+\big) \le s_1 - N^{-\varepsilon}, \dots, N^{2/3}\big(\widetilde{\lambda}_k^{\mathbf{v}} - \lambda_+\big) \le s_k - N^{-\varepsilon}\big) - N^{-\delta} \\ &\quad \le \mathbb{P}^{\mathbf{w}}\big(N^{2/3}\big(\widetilde{\lambda}_1^{\mathbf{w}} - \lambda_+\big) \le s_1, \dots, N^{2/3}\big(\widetilde{\lambda}_k^{\mathbf{w}} - \lambda_+\big) \le s_k\big) \\ &\quad \le \mathbb{P}^{\mathbf{v}}\big(N^{2/3}\big(\widetilde{\lambda}_1^{\mathbf{v}} - \lambda_+\big) \le s_1 + N^{-\varepsilon}, \dots, N^{2/3}\big(\widetilde{\lambda}_k^{\mathbf{v}} - \lambda_+\big) \le s_k + N^{-\varepsilon}\big) \\ &\qquad + N^{-\delta} \end{aligned} \tag{2.18}$$

for all k fixed and sufficiently large N. We remark that edge universality is usually formulated in terms of joint distributions of edge eigenvalues as in (2.18) with fixed parameters $s_1, s_2, \dots$ etc. However, we note that Theorem 2.4 holds uniformly in these parameters, and thus they may depend on N.

3. Key ideas and proof sketch. Our basic strategy is the so-called "Green function comparison" method initiated in a recent series of papers including [11–13] for proving universality for (generalized) Wigner matrices. The Green function comparison method has subsequently been applied to proving the spectral universality of adjacency matrices of random graphs [3, 4], the universality of eigenvectors of Wigner matrices [20], as well as the the spectrum of additive finite-rank deformations of Wigner matrices and the isotropic local semicircle law [21].

In this paper, we will show that (2.17) and (2.18) still hold with $\widetilde{X}^{\mathbf{v}}$ and $\widetilde{X}^{\mathbf{w}}$ replaced by the correlation matrix X and the corresponding covariance matrix $\widetilde{X}$, that is, Theorem 1.1. To show this result, we introduce a sufficient criteria for (2.17) and (2.18) derived in [24] (see Theorem 7.5 of [24]).

Consider two matrix ensembles $X^{\mathbf{v}}, X^{\mathbf{w}}$ (could be covariance, correlation or more general matrix[5]) and let their respective Green functions and empirical Stieltjes transforms [see (2.1) and (2.2)] be denoted by $G^{\mathbf{v}}, G^{\mathbf{w}}$ and $m^{\mathbf{v}}, m^{\mathbf{w}}$. To prove that the asymptotic distribution of the extreme eigenvalues of the matrix ensembles $X^{\mathbf{v}}, X^{\mathbf{w}}$ are identical in the sense of (2.17) and (2.18), it suffices to show the following [24]:

(i) The matrices $X^{\mathbf{v}}, X^{\mathbf{w}}$ satisfy the strong Marcenko–Pastur law and the rigidity of eigenvalues as given in Lemma 2.3.

(ii) The difference of the expectation of smooth functionals of the corresponding Green functions ($G^{\mathbf{v}}, G^{\mathbf{w}}$ and $m^{\mathbf{v}}, m^{\mathbf{w}}$) evaluated at the spectral edge must vanish asymptotically. More precisely, as pointed out in [24], it suffices to establish Theorems 3.1 and 3.2 below for the matrices $X^{\mathbf{v}}, X^{\mathbf{w}}$.

THEOREM 3.1 (Green function comparison theorem on the edge). *Let $F:\mathbb{R}\to\mathbb{R}$ be a function whose derivatives satisfy*

$$\max_x |F^{(\alpha)}(x)|(|x|+1)^{-C_1} \le C_1, \qquad \alpha = 1,2,3,4, \tag{3.1}$$

for some constant $C_1 > 0$. Then there exist $\varepsilon_0 > 0$, $N_0 \in \mathbb{N}$ and $\delta > 0$ depending only on C_1 such that for any $\varepsilon < \varepsilon_0$, $N \ge N_0$ and real numbers E, E_1 and E_2 satisfying

$$|E - \lambda_+| \le N^{-2/3+\varepsilon}, \qquad |E_1 - \lambda_+| \le N^{-2/3+\varepsilon}, \qquad |E_2 - \lambda_+| \le N^{-2/3+\varepsilon}$$

and $\eta_0 = N^{-2/3-\varepsilon}$, we have

$$\big|\mathbb{E}^{\mathbf{v}} F\big(N\eta_0 \Im m^{\mathbf{v}}(z)\big) - \mathbb{E}^{\mathbf{w}} F\big(N\eta_0 \Im m^{\mathbf{w}}(z)\big)\big| \le CN^{-\delta+C\varepsilon}, \qquad z = E + i\eta_0, \tag{3.2}$$

and

$$\begin{aligned}&\left|\mathbb{E}^{\mathbf{v}} F\left(N\int_{E_1}^{E_2} dy\, \Im m^{\mathbf{v}}(y+i\eta_0)\right) - \mathbb{E}^{\mathbf{w}} F\left(N\int_{E_1}^{E_2} dy\, \Im m^{\mathbf{w}}(y+i\eta_0)\right)\right| \\ &\quad \le CN^{-\delta+C\varepsilon}\end{aligned} \tag{3.3}$$

for some constant C.

[5] Notice that throughout the paper we use X for the correlation matrix and $\widetilde{X}$ for the covariance matrix. This is the only instance we denote a generic matrix by X for compactness of notation.

THEOREM 3.2. *Fix any $k \in \mathbb{N}_+$ and let $F : \mathbb{R}^k \to \mathbb{R}$ be a smooth, bounded function with bounded derivatives. Then there exist $\varepsilon_0 > 0$, $N_0 \in \mathbb{N}$ and $\delta > 0$ such that for any $\varepsilon < \varepsilon_0$, $N \geq N_0$ and sequence of real numbers $E_k < \cdots < E_1 < E_0$ with $|E_j - \lambda_+| \leq N^{-2/3+\varepsilon}$, $j = 0, 1, \ldots, k$ and $\eta_0 = N^{-2/3-\varepsilon}$, we have*

$$\begin{aligned}\biggl|\mathbb{E}^{\mathbf{v}} F\biggl(N \int_{E_1}^{E_0} dy\, \Im m^{\mathbf{v}}(y + i\eta_0), \ldots, N \int_{E_k}^{E_0} dy\, \Im m^{\mathbf{v}}(y + i\eta_0)\biggr)& \\ - \mathbb{E}^{\mathbf{w}} F(m^{\mathbf{v}} \to m^{\mathbf{w}})\biggr|& \\ \leq N^{-\delta},&\end{aligned} \tag{3.4}$$

where the second term in the left-hand side above is obtained by changing the arguments of F in the first term from $m^{\mathbf{v}}$ to $m^{\mathbf{w}}$ and keeping all the other parameters fixed.

REMARK 3.3. Theorems 3.1 and 3.2 yield the edge universality of the k-point correlation functions at the edge for $k = 1$ and $k \geq 1$, respectively.

Thus, to complete the proof of Theorem 1.1, by the Green function comparison method it suffices to show (i) and (ii) above for

$$X^{\mathbf{v}} = X, \qquad X^{\mathbf{w}} = \widetilde{X},$$

where $X^\dagger X$ denotes the correlation matrix and $\widetilde{X}^\dagger \widetilde{X}$ is the corresponding covariance matrix. Here condition (i) is guaranteed by Theorem 2.3.

Verifying condition (ii) entails the heart of this paper. In previous works mentioned earlier, the authors use a Lindeberg replacement strategy, as in [2, 27]. These proofs proceed via showing that the distribution of some smooth functional of the Green function (e.g., G_{ii}, m and $\langle \mathbf{x}_1, G\mathbf{x}_1 \rangle$) of the two matrix ensembles is identical asymptotically provided that the first two (in some cases up to four) moments of all matrix elements of these two ensembles are identical. For instance, if one needs to show the edge universality of two covariance matrices $\widetilde{X}^{\mathbf{v}}$ and $\widetilde{X}^{\mathbf{w}}$, the basic strategy is to express

$$\mathbb{E}F(\widetilde{G}^{\mathbf{v}}) - \mathbb{E}F(\widetilde{G}^{\mathbf{w}}) = \sum_{\gamma=1}^{MN} \mathbb{E}F(\widetilde{G}_\gamma) - \mathbb{E}F(\widetilde{G}_{\gamma-1}), \tag{3.5}$$

where F is a smooth function and $\widetilde{G}_\gamma$ denotes the Green function of the ensemble $\widetilde{X}_\gamma$ (with $\widetilde{X}_0 = \widetilde{X}^{\mathbf{v}}$) which is obtained from $\widetilde{X}_{\gamma-1}$ by replacing the distribution of the ijth entry of $\widetilde{X}_{\gamma-1}[ij]$ with $\widetilde{X}^{\mathbf{w}}[ij]$ [here $\gamma = i + (j-1)M$] so that $\widetilde{X}_{MN} = \widetilde{X}^{\mathbf{w}}$. The next step is to obtain an estimate

$$\mathbb{E}F(\widetilde{G}_\gamma) - \mathbb{E}F(\widetilde{G}_{\gamma-1}) = o(N^{-2}) \tag{3.6}$$

for each of the N^2 terms in the sum (3.5). Usually (3.6) is obtained by resolvent expansions, perturbation theory and the fact that $\widetilde{X}_\gamma$ and $\widetilde{X}_{\gamma-1}$ differ by a single entry and the first few moments of these two distributions are the same.

But clearly the above method does not work in our case, since the entries within the same column are not independent and, therefore, one cannot replace the distribution of a single entry of a column without changing the distribution of all the other $M-1$ entries. To circumvent this, in [24] a new telescoping argument consisting of $O(N)$ ensembles was used for the comparison of Green functions. The idea is that instead of replacing entries one at a time, one can replace the entries of the data matrix column by column and thus require only $O(N)$ ensembles. This argument from [24] is adapted here along with new insights for dealing with nonindependence of the entries and is outlined below.

Now we set $X^{\mathbf{v}} = X$, $X^{\mathbf{w}} = \widetilde{X}$. For $1 \le \gamma \le N$, let X_γ denote the random matrix whose jth column is the same as that of $X^{\mathbf{v}}$ if $j > \gamma$ and that of $X^{\mathbf{w}}$ otherwise. In particular, we can choose $X_0 = X^{\mathbf{v}} = X$ and $X_N = X^{\mathbf{w}} = \widetilde{X}$, where X is correlation matrix and $\widetilde{X}$ the corresponding covariance matrix of X. As before, we define

$$m_\gamma(z) = \frac{1}{N}\operatorname{Tr} G_\gamma(z), \qquad G_\gamma(z) = \left(X_\gamma^\dagger X_\gamma - z\right)^{-1},$$

so that we have telescoping sum

$$\begin{aligned}&\mathbb{E}^{\mathbf{w}} F\left(N\eta_0 \Im m^{\mathbf{w}}(z)\right) - \mathbb{E}^{\mathbf{v}} F\left(N\eta_0 \Im m^{\mathbf{v}}(z)\right)\\ &\quad = \sum_{\gamma=1}^{N} \mathbb{E} F\left(N\eta_0 \Im m_\gamma(z)\right) - \mathbb{E} F\left(N\eta_0 \Im m_{\gamma-1}(z)\right).\end{aligned} \tag{3.7}$$

Clearly, (3.2) will follow from (3.7) and the following estimate:

$$\left|\mathbb{E} F\left(N\eta_0 \Im m_\gamma(z)\right) - \mathbb{E} F\left(N\eta_0 \Im m_{\gamma-1}(z)\right)\right| \le O_\varepsilon\left(N^{-1-\delta}\right) \tag{3.8}$$

for some $\delta > 0$. Our strategy to obtain (3.8) is the following. First notice that

$$\begin{aligned}&\mathbb{E} F\left(N\eta_0 \Im m_\gamma(z)\right) - \mathbb{E} F\left(N\eta_0 \Im m_{\gamma-1}(z)\right)\\ &\quad = \mathbb{E} F\left(\eta_0 \Im \operatorname{Tr} G_\gamma(z)\right) - \mathbb{E} F\left(\eta_0 \Im \operatorname{Tr} G_{\gamma-1}(z)\right).\end{aligned}$$

Let $X^{(\gamma)}$ be the $M \times (N-1)$ matrix obtained by removing the γth column of X_γ, which has the same distribution of the $M \times (N-1)$ matrix obtained by removing the γth column of $X_{\gamma-1}$. Define

$$G^{(\gamma)} = \left(\left(X^{(\gamma)}\right)^\dagger \left(X^{(\gamma)}\right) - z\right)^{-1}, \qquad \mu = \eta_0 \Im \operatorname{Tr} G^{(\gamma)} - \Im \frac{\eta_0}{z}. \tag{3.9}$$

In Lemma 4.1 we will establish (3.8) by showing that

$$\begin{aligned}&\left(\mathbb{E} F(\eta_0 \Im \operatorname{Tr} G_\gamma) - \mathbb{E} F(\mu)\right) - \left(\mathbb{E} F(\eta_0 \Im \operatorname{Tr} G_{\gamma-1}) - \mathbb{E} F(\mu)\right)\\ &\quad = O_\varepsilon\left(N^{-7/6}\right).\end{aligned} \tag{3.10}$$

Once (3.8) is verified, the main result follows by virtue of Theorems 3.1 and 3.2 as mentioned in the beginning of this section. Notice that since the columns of the data matrix $X^{\mathbf{v}}$, $X^{\mathbf{w}}$ are assumed to be independent, μ is independent of the γth column of $X^{\mathbf{v}}$, $X^{\mathbf{w}}$ or, equivalently, the γth column of X_γ, $X_{\gamma-1}$.

Thus, it boils down to establishing (3.10) in the case $X_0 = X^{\mathbf{v}} = X$ and $X_N = X^{\mathbf{w}} = \widetilde{X}$. Our proof relies on the key observation that even if the entries of the γth column vector $\mathbf{x}_\gamma$ are not independent, the difference between the moments of the entries of the standardized vector $\mathbf{x}_\gamma$ and its unnormalized counterpart $\widetilde{\mathbf{x}}_\gamma$ is at least an order of magnitude smaller than those of $\widetilde{\mathbf{x}}_\gamma$. For instance, since $\mathbf{x}_{i\gamma} = O(N^{-1/2})$ for $1 \le i \le M$, for two independent ensembles of covariance matrices $\widetilde{X}^{\mathbf{v}}$ and $\widetilde{X}^{\mathbf{w}}$ satisfying (1.1) and (1.2), we have the bound

$$\mathbb{E}(\widetilde{\mathbf{x}}^{\mathbf{v}}_{i\gamma})^3 - \mathbb{E}(\widetilde{\mathbf{x}}^{\mathbf{w}}_{i\gamma})^3 = O(N^{-3/2}). \tag{3.11}$$

On the other hand, if $\widetilde{\mathbf{x}}_\gamma$ is the unnormalized counterpart of $\mathbf{x}_\gamma$, as shown in Lemma 5.5,

$$\mathbb{E}(\widetilde{\mathbf{x}}_{i\gamma})^3 - \mathbb{E}(\mathbf{x}_{i\gamma})^3 = O(N^{-5/2}). \tag{3.12}$$

The above observation combined with a resolvent expansion—detailed in Lemmas 4.3, 5.4 and 5.5—gives (3.10).

4. Proof of the main result.

In this section we will prove (3.10) in the case $X_0 = X^{\mathbf{v}} = X$ and $X_N = X^{\mathbf{w}} = \widetilde{X}$. As discussed above, it implies (3.2) in Theorem 3.1. Similarly, one can prove (3.3) and (3.4) in Theorems 3.1 and 3.2, which complete the proof of Theorem 1.1, the main result of this paper.

It is easy to see that (3.10) is a direct consequence of the following lemma.

LEMMA 4.1. *Let X be a $M \times N$ random matrix whose columns satisfy the large deviation bounds* (2.10), (2.11) *and* (2.12), *for any $A_i \in \mathbb{C}$ and $B_{ij} \in \mathbb{C}$ and for any $\zeta > 0$. The columns of X are assumed to be mutually independent. Furthermore, assume that the first column is given by*

$$X_{i1} = \frac{\widetilde{x}_{i1}}{\|\widetilde{\mathbf{x}}_1\|_2}, \qquad 1 \le i \le M, \tag{4.1}$$

where $\widetilde{x}_{i1}$ are i.i.d. random variables with mean zero and variance M^{-1} and have an exponentially decay in the tails as given by (1.2).

Let $\widetilde{X}$ be the random matrix whose entries have the same distribution as X except for the first column, and the first column of $\widetilde{X}$ is given by

$$\widetilde{X}_{i1} = \widetilde{x}_{i1},$$

where $\widetilde{x}_{i1}$ are as in (4.1). *The columns of $\widetilde{X}$ are also assumed to be mutually independent. Let $m, \widetilde{m}$ denote the empirical Stieltjes transforms of $X^\dagger X$, $\widetilde{X}^\dagger \widetilde{X}$.*

Then for any function F satisfying (3.1), *there exists $\delta > 0$, $\varepsilon_0 > 0$ depending only on C_1 such that for any $\varepsilon < \varepsilon_0$ and for any real number E satisfying*

$$|E - \lambda_+| \le N^{-2/3+\varepsilon}, \qquad \eta_0 = N^{-2/3-\varepsilon}, \tag{4.2}$$

we have

$$\bigl|\mathbb{E}F\bigl(N\eta_0 \Im m(z)\bigr) - \mathbb{E}F\bigl(N\eta_0 \Im \widetilde{m}(z)\bigr)\bigr| \le O_\varepsilon\bigl(N^{-1-\delta}\bigr), \qquad z = E + i\eta_0. \tag{4.3}$$

Note: In this lemma X and $\widetilde{X}$ are neither pure correlation nor pure covariance matrices, but their respective first columns are distributed according to the standardized data matrix and raw data matrix.

REMARK 4.2. Under condition (4.2) (see [24]), we have the bound

$$C^{-1} \le \bigl|m_W(z)\bigr| \le C, \qquad \Im m_W(z) = O_\varepsilon\bigl(N^{-1/3}\bigr), \qquad z = E + i\eta_0. \tag{4.4}$$

First we collect some properties on submatrices of a *generic* $M \times N$ matrix Q which can be proved using standard results from linear algebra. Let $Q^{(1)}$ be the $M \times (N-1)$ matrix obtained by removing the first column of Q. Define

$$G_Q^{(1)} = \bigl(\bigl(Q^{(1)}\bigr)^\dagger\bigl(Q^{(1)}\bigr) - z\bigr)^{-1}, \qquad \mathcal{G}_Q^{(1)} = \bigl(\bigl(Q^{(1)}\bigr)\bigl(Q^{(1)}\bigr)^\dagger - z\bigr)^{-1}. \tag{4.5}$$

Then by definition, $G_Q^{(1)}$ is a $(N-1) \times (N-1)$ matrix, $\mathcal{G}_Q^{(1)}$ is a $M \times M$ matrix and we have the identity

$$\operatorname{Tr} G_Q^{(1)}(z) - \operatorname{Tr} \mathcal{G}_Q^{(1)}(z) = \frac{M - N + 1}{z}. \tag{4.6}$$

Using the Cauchy interlacing theorem (see Equation (8.5) of [10]), it can be shown that

$$\operatorname{Tr} G_Q^{(1)}(z) - \operatorname{Tr} G_Q(z) = O\bigl(\eta^{-1}\bigr), \qquad \eta = \Im z. \tag{4.7}$$

PROOF OF LEMMA 4.1. First we note that from Theorem 1.5 of [24], the conclusions of Theorem 2.3 hold for both X and $\widetilde{X}$.

Let $X^{(1)}$ be the $M \times (N-1)$ matrix obtained by removing the first column of X. Define

$$G^{(1)} = \bigl(\bigl(X^{(1)}\bigr)^\dagger\bigl(X^{(1)}\bigr) - z\bigr)^{-1}, \qquad \mathcal{G}^{(1)} = \bigl(\bigl(X^{(1)}\bigr)\bigl(X^{(1)}\bigr)^\dagger - z\bigr)^{-1} \tag{4.8}$$

and as in (3.9) set

$$\mu = \eta_0 \Im \operatorname{Tr} G^{(1)} - \Im \frac{\eta_0}{z}. \tag{4.9}$$

We will first verify that

$$\begin{aligned} &\mathbb{E}F(\eta_0 \Im \operatorname{Tr} G) - \mathbb{E}F(\mu) \\ &\quad = \mathbb{E}F^{(1)}(\mu)(\Im y_1 + \Im y_2 + \Im y_3) + \mathbb{E}F^{(2)}(\mu)\bigl(\tfrac{1}{2}(\Im y_1)^2 + \Im y_1 \Im y_2\bigr) \\ &\qquad + \mathbb{E}F^{(3)}(\mu)\bigl(\tfrac{1}{6}(\Im y_1)^3\bigr) + O_\varepsilon\bigl(N^{-4/3}\bigr), \end{aligned} \tag{4.10}$$

where $F^{(s)}$ denotes the sth derivative of F and y_k's are defined as

$$y_k := \eta_0 z m_W (-B)^{k-1}\bigl(\mathbf{x}_1, (\mathcal{G}^{(1)})^2 \mathbf{x}_1\bigr), \tag{4.11}$$

where $\mathbf{x}_1$ denotes the first column of X. Define the quantity

$$B := -z m_W \biggl[(\mathbf{x}_1, \mathcal{G}^{(1)}(z)\mathbf{x}_1) - \biggl(\frac{-1}{z m_W(z)} - 1\biggr)\biggr]. \tag{4.12}$$

First, recall the following identity (see (6.23) of [24]):

$$\begin{aligned} \operatorname{Tr} G - \operatorname{Tr} G^{(1)} + z^{-1} &= (G_{11} + z^{-1}) + \frac{(\mathbf{x}_1, X^{(1)} G^{(1)} G^{(1)} X^{(1)\dagger}\mathbf{x}_1)}{-z - z(\mathbf{x}_1, \mathcal{G}^{(1)}(z)\mathbf{x}_1)} \\ &= z G_{11}\bigl(\mathbf{x}_1, (\mathcal{G}^{(1)})^2(z)\mathbf{x}_1\bigr). \end{aligned} \tag{4.13}$$

Furthermore, as proved in Lemma 2.5 of [24],

$$\begin{aligned} G_{11}(z) &= \frac{1}{-z - z(\mathbf{x}_1, \mathcal{G}^{(1)}(z)\mathbf{x}_1)} \quad \text{that is} \\ (\mathbf{x}_1, \mathcal{G}^{(1)}(z)\mathbf{x}_1) &= \frac{-1}{z G_{11}(z)} - 1. \end{aligned} \tag{4.14}$$

From (4.12) and (4.14) we obtain that

$$B = -z m_W \biggl[\biggl(\frac{-1}{z G_{11}(z)} - 1\biggr) - \biggl(\frac{-1}{z m_W(z)} - 1\biggr)\biggr] = \frac{m_W - G_{11}}{G_{11}}.$$

Fix $\zeta > 0$. From (2.14), Remark 4.2 and the bound $|G_{11}| \le |m_W| + O(1)$, it follows that for $z = E + i\eta_0$,

$$|B| = \frac{|m_W - G_{11}|}{|G_{11}|} \le O_\varepsilon(N^{-1/3}) \ll 1 \tag{4.15}$$

with ζ-high probability (see Definition 2.1). Therefore, with ζ-high probability, we have the identity

$$G_{11} = \frac{m_W}{B+1} = m_W \sum_{k \ge 0} (-B)^k. \tag{4.16}$$

Define y to be the l.h.s. of (4.13) multiplied by η_0, that is,

$$y = \eta_0(\operatorname{Tr} G - \operatorname{Tr} G^{(1)} + z^{-1}),$$

so that using (4.13) and (4.16), we obtain

$$y = \eta_0 z G_{11}\bigl(\mathbf{x}_1, (\mathcal{G}^{(1)})^2 \mathbf{x}\bigr) = \sum_{k=1}^{\infty} y_k.$$

Since $\mathbf{x}_1$ satisfies (2.10), (2.11) and (2.12), and $\mathcal{G}^{(1)}$ is independent of $\mathbf{x}_1$, using Lemma 2.2, we infer that for some $C_\zeta > 0$

$$\big|(\mathbf{x}_1, (\mathcal{G}^{(1)})^2\mathbf{x}_1)\big| \le \frac{1}{M}\operatorname{Tr}(\mathcal{G}^{(1)})^2 + \frac{\varphi^{C_\zeta}}{M}\sqrt{\operatorname{Tr}|\mathcal{G}^{(1)}|^4} \tag{4.17}$$

with ζ-high probability. Using its definition, we bound $\operatorname{Tr}(\mathcal{G}^{(1)})^2$ as

$$\begin{aligned}\big|\operatorname{Tr}(\mathcal{G}^{(1)})^2\big| &\le \operatorname{Tr}|\mathcal{G}^{(1)}|^2 = \frac{\Im \operatorname{Tr}\mathcal{G}^{(1)}}{\eta_0}\\ &= O_\varepsilon(N^{4/3}) + \frac{\Im\operatorname{Tr} G}{\eta_0} = O_\varepsilon(N^{4/3}),\end{aligned} \tag{4.18}$$

where for the last two inequalities we have used (4.6), (4.7), (2.13) and (4.4). Similarly, we bound the last term of (4.17) with

$$\operatorname{Tr}|\mathcal{G}^{(1)}|^4 \le \eta_0^{-2}\operatorname{Tr}|\mathcal{G}^{(1)}|^2 \le O_\varepsilon(N^{8/3}) \tag{4.19}$$

and obtain that

$$\big|(\mathbf{x}_1, (\mathcal{G}^{(1)})^2\mathbf{x}_1)\big| \le O_\varepsilon(N^{1/3}).$$

Equation (4.15) and the fact $|z| + |m_W(z)| = O(1)$ yields that

$$|y_k| \le O_\varepsilon(N^{-k/3}) \quad \text{and} \quad |y| \le O_\varepsilon(N^{-1/3}) \tag{4.20}$$

holds with ζ-high probability. Consequently, using (3.1) and (4.13), we see that the expansion

$$F(\eta_0\Im\operatorname{Tr} G) - F(\mu) = \sum_{k=1}^{3}\frac{1}{k!}F^{(k)}(N\eta_0\Im\widetilde{m}^{(1)}(z))(\Im y)^k + O_\varepsilon(N^{-4/3}) \tag{4.21}$$

holds with ζ-high probability. From the bounds on y_k's obtained above, equation (4.10) follows.

Now we estimate $\widetilde{G}$, which is defined as

$$\widetilde{G} = (\widetilde{X}^\dagger\widetilde{X} - z)^{-1}.$$

Let $\widetilde{X}^{(1)}$ be the $M \times (N-1)$ matrix obtained by removing the first column of $\widetilde{X}$ and $\widetilde{\mathbf{x}}_1$ denote its first column. Proceeding as in the previous calculations,

$$\begin{aligned}&\mathbb{E}F(\eta_0\Im\operatorname{Tr}\widetilde{G}) - \mathbb{E}F(\mu)\\ &\quad= \mathbb{E}F^{(1)}(\mu)(\Im\widetilde{y}_1 + \Im\widetilde{y}_2 + \Im\widetilde{y}_3) + \mathbb{E}F^{(2)}(\mu)\big(\tfrac{1}{2}(\Im\widetilde{y}_1)^2 + \Im\widetilde{y}_1\Im\widetilde{y}_2\big)\\ &\qquad + \mathbb{E}F^{(3)}(\mu)\big(\tfrac{1}{6}(\Im\widetilde{y}_1)^3\big) + O_\varepsilon(N^{-4/3}),\end{aligned} \tag{4.22}$$

where

$$\begin{aligned}\widetilde{y}_k &= \eta_0 z m_W(-\widetilde{B})^{k-1}(\widetilde{\mathbf{x}}_1, (\mathcal{G}^{(1)})^2\widetilde{\mathbf{x}}_1),\\ \widetilde{B} &= -z m_W\Big[(\widetilde{\mathbf{x}}_1, \mathcal{G}^{(1)}(z)\widetilde{\mathbf{x}}_1) - \Big(\frac{-1}{z m_W(z)} - 1\Big)\Big].\end{aligned}$$

Notice that μ appears in (4.22) because the entries of $\widetilde{X}^{(1)}$ and $X^{(1)}$ are assumed to be identically distributed.

Define the matrices

$$Y = \left(\mathcal{G}^{(1)}\right)^2, \qquad Z = \mathcal{G}^{(1)}. \tag{4.23}$$

The symmetric matrices Y and Z are independent of $\mathbf{x}_1$ and $\widetilde{\mathbf{x}}_1$. Clearly, $YZ = ZY$. Therefore, using the fact that $z, m_W \sim 1$, we can write

$$y_k = \eta_0 \sum_{0 \le n < k} C_{k,n}(\mathbf{x}_1, Y\mathbf{x}_1)(\mathbf{x}_1, Z\mathbf{x}_1)^n,$$

where $C_{k,n} = O(1)$. Let $\mathcal{Y} = (\mathbf{x}_1, Y\mathbf{x}_1)$ and $\mathcal{Z} = (\mathbf{x}_1, Z\mathbf{x}_1)$. Then (4.10) can be written as

$$\begin{aligned}
&\mathbb{E}F(\eta_0 \Im \operatorname{Tr} G) - \mathbb{E}F(\mu) \\
&\quad = \mathbb{E}F^{(1)}(\mu)\Im\biggl(\eta_0 \sum_{0 \le n < k \le 3} C_{k,n}\mathcal{Y}\mathcal{Z}^n\biggr) \\
&\qquad + \mathbb{E}F^{(2)}(\mu)\eta_0^2\biggl(\frac{1}{2}(\Im(C_{1,0}\mathcal{Y}))^2 + \Im(C_{1,0}\mathcal{Y})\Im(C_{2,0}\mathcal{Y}) \\
&\qquad\qquad + \Im(C_{1,0}\mathcal{Y})\Im(C_{2,1}\mathcal{Y}\mathcal{Z})\biggr) \\
&\qquad + \mathbb{E}F^{(3)}(\mu)\eta_0^3\biggl(\frac{1}{6}(\Im(C_{1,0}\mathcal{Y}))^3\biggr) + O_\varepsilon(N^{-4/3}).
\end{aligned} \tag{4.24}$$

Define $\widetilde{\mathcal{Y}} = (\widetilde{\mathbf{x}}_1, Y\widetilde{\mathbf{x}}_1)$ and $\widetilde{\mathcal{Z}} = (\widetilde{\mathbf{x}}_1, Z\widetilde{\mathbf{x}}_1)$. Using (4.22) and proceeding similarly as before, we obtain that (4.24) also holds for the case when G, $\mathcal{Y}$ and $\mathcal{Z}$ are replaced with $\widetilde{G}$, $\widetilde{\mathcal{Y}}$ and $\widetilde{\mathcal{Z}}$, respectively. The following is the key technical lemma of this paper whose proof is deferred to the next section.

LEMMA 4.3. *Let $f : \mathbb{R} \to \mathbb{R}$ be a function satisfying*

$$\max_x |f(x)|\left(|x| + 1\right)^{-C} \le C \tag{4.25}$$

for some constant C. Let $\mathcal{A}$ be of the form

$$\eta_0^a \prod_{i=1}^{a}(\mathbf{x}, Y_i\mathbf{x}) \prod_{j=1}^{b}(\mathbf{x}, Z_j\mathbf{x}), \tag{4.26}$$

where $Y_i = Y$ or Y^* and $Z_j = Z$ or Z^* with Y, Z as defined in (4.23) and a, b are integers with $1 \le a \le 3$, $1 \le a+b \le 3$. Then, under the assumptions of Lemma 4.1, we have

$$\left|\mathbb{E}(f(\mu)\mathcal{A}) - \mathbb{E}(f(\mu)\widetilde{\mathcal{A}})\right| \le O_\varepsilon(N^{-7/6}), \tag{4.27}$$

where $\widetilde{\mathcal{A}}$ is obtained by replacing $\mathbf{x}$ with $\widetilde{\mathbf{x}}$ in (4.26).

Taking the difference of (4.24) and the equation obtained by replacing (4.24) with $\widetilde{G}$, $\widetilde{\mathcal{Y}}$ and $\widetilde{\mathcal{Z}}$, we deduce that the difference

$$\mathbb{E}F(\eta_0\Im\operatorname{Tr}G)-\mathbb{E}F(\eta_0\Im\operatorname{Tr}\widetilde{G})$$

can be approximated by the sum of $O(1)$ number of terms of the form $\mathbb{E}(f(\mu)\mathcal{A})-\mathbb{E}(f(\mu)\widetilde{\mathcal{A}})$, where $\mathcal{A}$ is as in (4.26) and f is equal to $F^{(1)}$, $F^{(2)}$ and $F^{(3)}$. Therefore, by applying Lemma 4.3, we conclude that Lemma 4.1 holds with any $\delta < 1/6$ and the proof is finished. □

Finally, we are ready to give the proof of the main result of this paper:

PROOF OF THEOREM 1.1. By the Green function comparison theorem discussed in Section 3, it only remains to prove that Theorems 3.1 and 3.2 hold for the case

$$X^{\mathbf{v}}=X, \qquad X^{\mathbf{w}}=\widetilde{X}.$$

For simplicity, we will only prove (3.2) of Theorem 3.1; the rest can be proved using almost identical arguments.

For $1\le\gamma\le N$, let X_γ denote the random matrix whose jth column is the same as that of $X^{\mathbf{v}}$ if $j\ge\gamma$ and that of $X^{\mathbf{w}}$ otherwise; in particular, $X_0=X^{\mathbf{v}}$ and $X_N=X^{\mathbf{w}}$. As before, we define

$$m_\gamma(z)=\frac{1}{N}\operatorname{Tr}G_\gamma(z), \qquad G_\gamma(z)=\left(X_\gamma^\dagger X_\gamma-z\right)^{-1}.$$

We have the telescoping sum,

$$\begin{aligned}&\mathbb{E}^{\mathbf{w}}F\left(N\eta_0\Im m^{\mathbf{w}}(z)\right)-\mathbb{E}^{\mathbf{v}}F\left(N\eta_0\Im m^{\mathbf{v}}(z)\right)\\&\quad=\sum_{\gamma=1}^{N}\mathbb{E}F\left(N\eta_0\Im m_\gamma(z)\right)-\mathbb{E}F\left(N\eta_0\Im m_{\gamma-1}(z)\right).\end{aligned}\tag{4.28}$$

Applying Lemma 4.1 on X_γ and $X_{\gamma-1}$ gives the estimate

$$\left|\mathbb{E}F\left(N\eta_0\Im m_\gamma(z)\right)-\mathbb{E}F\left(N\eta_0\Im m_{\gamma-1}(z)\right)\right|\le O_\varepsilon\left(N^{-1-\delta}\right)\tag{4.29}$$

for some $\delta>0$. Now (3.2) follows from (4.28) and (4.29) and the proof is finished. □

5. Moment computations.

In this section we prove Lemma 4.3. For notational convenience, let us denote $\mathbf{x}=\mathbf{x}_1$, $\widetilde{\mathbf{x}}=\widetilde{\mathbf{x}}_1$. We will also write

$$\mathbf{x}(k)=x_{k1}, \qquad \widetilde{\mathbf{x}}(k)=\widetilde{x}_{k1}, \qquad 1\le k\le M.$$

Recall μ from (4.9). For the rest of this section, a, b will denote two integers with

$$1\le a\le 3, \qquad 1\le a+b\le 3.$$

Before stating the key results of this section, let us first give some definitions.

DEFINITION 5.1 [$\mathcal{I}(A,\mathbf{k})$]. For any partition A of the set $\{1,2,\dots,2a+2b\}$, and a vector $\mathbf{k}=\{k_1,k_2,\dots,k_{2a+2b}\}$, $k_i\in\{1,2,\dots,M\}$, define the binary function $\mathcal{I}(A,\mathbf{k})$ as follows. The function $\mathcal{I}(A,\mathbf{k})$ is equal to 1 if (1) for any i, j in the same block of A we have $k_i=k_j$, (2) if i, j are in different blocks of A, we have $k_i\neq k_j$; otherwise $\mathcal{I}(A,\mathbf{k})=0$.

EXAMPLE 5.2. If

$$A=\{\{1\},\{2,4\},\{3,5,6\}\} \tag{5.1}$$

and $a+b=3$, then

$$\mathcal{I}(A,\mathbf{k})=\mathbf{1}(k_2=k_4)\mathbf{1}(k_3=k_5=k_6)\mathbf{1}(k_1\neq k_2)\mathbf{1}(k_2\neq k_3)\mathbf{1}(k_1\neq k_3).$$

DEFINITION 5.3 [$\mathcal{N}(A,1),\mathcal{N}(A,2)$ and $\mathbf{I}_{(A,3)}$]. Given a partition A of the set $\{1,2,\dots,2a+2b\}$, let $\mathcal{N}(A,1)$ be the number of the blocks in A that contain only one element of the set $\{1,2,\dots,2a+2b\}$. Let $\mathcal{N}(A,2)$ be the number of the blocks in A of the form $\{k_{2i-1},k_{2i}\}$ with $i>a$. Note that $\mathcal{N}(A,2)$ depends on a and b in addition to A. Let $\mathbf{I}_{(A,3)}$ be equal to one if and only if $a+b=3$ and A is composed of 2 blocks with three elements in each block.

The proof of Lemma 4.3 relies on Lemmas 5.4 and 5.5 stated below and proved at the end of this section.

LEMMA 5.4. *Recall the matrices Y, Z from* (4.23). *Then for any $\varepsilon>0$ the following estimate*

$$\sum_{k_1,k_2,\dots,k_{2a+2b}=1}^{M}\mathcal{I}(A,\mathbf{k})\eta_0^a(Y_{k_1k_2}\cdots Y_{k_{2a-1}k_{2a}})(Z_{k_{2a+1}k_{2a+2}}\cdots Z_{k_{2a+2b-1}k_{2a+2b}})$$
$$=O_\varepsilon\big((N^{2/3})^{a+b}(N^{1/2})^{\mathcal{N}(A,1)+\mathbf{I}_{(A,3)}}(N^{1/3})^{\mathcal{N}(A,2)}\big)$$

holds with ζ-high probability for any fixed $\zeta>0$. The result also holds if any of the Y, Z are replaced by their complex conjugates Y^*, Z^*, respectively.

LEMMA 5.5. *Let $\widetilde{y}_i$ be i.i.d. random variables such that*

$$\mathbb{E}\widetilde{y}_i=0,\qquad \mathbb{E}(\widetilde{y}_i)^2=M^{-1},\qquad 1\le i\le M,$$

and have a subexponential decay as in (1.2). *Let A be a partition of the set $\{1,2,\dots,2a+2b\}$ and let*

$$y_i:=\frac{\widetilde{y}_i}{(\sum_j\widetilde{y}_j^2)^{1/2}}.$$

Then for any vector $\mathbf{k} = (k_1, k_2, \dots, k_{2a+2b})$ *and for any* $\varepsilon > 0$, *we have*

$$\mathbb{E}\left(\mathcal{I}(A, \mathbf{k}) \prod_{i=1}^{2a+2b} y_{k_i}\right) - \mathbb{E}\left(\mathcal{I}(A, \mathbf{k}) \prod_{i=1}^{2a+2b} \tilde{y}_{k_i}\right) = O_\varepsilon\bigl(N^{-(a+b)-\max\{\mathcal{N}(A,1),1\}}\bigr). \tag{5.2}$$

With the above two lemmas in hand, we are now ready to give the proof of Lemma 4.3.

PROOF OF LEMMA 4.3. We will only prove the case when

$$Y_i = Y, \qquad Z_i = Z \tag{5.3}$$

for all i and, thus,

$$\mathcal{A} = \eta_0^a (\mathbf{x}, Y\mathbf{x})^a (\mathbf{x}, Z\mathbf{x})^b.$$

The other cases can be proved similarly. First, let us write (4.26) as

$$\begin{aligned} &\eta_0^a (\mathbf{x}, Y\mathbf{x})^a (\mathbf{x}, Z\mathbf{x})^b \\ &\quad = \sum_A \sum_{k_1, k_2, \dots, k_{2a+2b}=1}^{M} \eta_0^a \mathcal{I}(A, \mathbf{k}) \prod_{i=1}^{2a+2b} \mathbf{x}(k_i)(Y_{k_1 k_2} \cdots Y_{k_{2a-1} k_{2a}}) \\ &\qquad \times (Z_{k_{2a+1} k_{2a+2}} \cdots Z_{k_{2a+2b-1} k_{2a+2b}}), \end{aligned}$$

where the summation index A ranges over all the partitions of the set $\{1, 2, \dots, 2a + 2b\}$. Taking expectations, and using the fact that $\mathbf{x}$ is independent of Y, Z and μ, leads to

$$\begin{aligned} \mathbb{E} f(\mu) \mathcal{A} &= \sum_A \sum_{k_1, k_2, \dots, k_{2a+2b}=1}^{M} \mathbb{E}\Biggl(\eta_0^a \mathcal{I}(A, \mathbf{k}) \\ &\qquad \times \prod_{i=1}^{2a+2b} \mathbf{x}(k_i)(Y_{k_1 k_2} \cdots Y_{k_{2a-1} k_{2a}}) \\ &\qquad \times (Z_{k_{2a+1} k_{2a+2}} \cdots Z_{k_{2a+2b-1} k_{2a+2b}})\Biggr) \\ &= \sum_A \left(\mathbb{E}\mathcal{I}(A, \mathbf{k}) \prod_{i=1}^{2a+2b} \mathbf{x}(k_i)\right) \\ &\qquad \times \Biggl(\mathbb{E} f(\mu) \sum_{k_1, k_2, \dots, k_{2a+2b}=1}^{M} \mathcal{I}(A, \mathbf{k}) \eta_0^a (Y_{k_1 k_2} \cdots Y_{k_{2a-1} k_{2a}}) \\ &\qquad \times (Z_{k_{2a+1} k_{2a+2}} \cdots Z_{k_{2a+2b-1} k_{2a+2b}})\Biggr), \end{aligned} \tag{5.4}$$

where the last inequality follows from the fact that $(\mathbb{E}\mathcal{I}(A,\mathbf{k})\prod_{i=1}^{2a+2b}\mathbf{x}(k_i))$ is independent of Y, Z. Combining (5.4), Lemmas 5.4 and 5.5, we deduce that

$$
\begin{aligned}
&|\mathbb{E}(f(\mu)\mathcal{A})-\mathbb{E}(f(\mu)\widetilde{\mathcal{A}})| \\
&\quad\le \sum_A O_\varepsilon\bigl((N^{-1/3})^{a+b}(N^{1/2})^{\mathcal{N}(A,1)+\mathbf{I}_{(A,3)}}(N^{-1})^{\max\{\mathcal{N}(A,1),1\}}(N^{1/3})^{\mathcal{N}(A,2)}\bigr) \\
&\quad\le \sum_A O_\varepsilon\bigl((N^{-a/3})(N^{1/2})^{\mathcal{N}(A,1)+\mathbf{I}_{(A,3)}}(N^{-1})^{\max\{\mathcal{N}(A,1),1\}}(N^{1/3})^{\mathcal{N}(A,2)-b}\bigr).
\end{aligned}
\tag{5.5}
$$

Now we claim that the terms in the r.h.s. of (5.5) are bounded by $O_\varepsilon(N^{-7/6})$. Indeed, note that $\mathcal{N}(A,1)>0$ implies $\mathbf{I}_{(A,3)}=0$. Therefore, the worse case scenario is the case in which

$$a=1,\qquad b=\mathcal{N}(A,2)\quad\text{and}\quad\mathcal{N}(A,1)=1,$$

since by definition we have $\mathcal{N}(A,2)\le b$. But it is easy to see the above scenario cannot occur, since if the first two conditions hold, then it follows that $\mathcal{N}(A,1)=0$ or 2. Thus, we have finished the proof of Lemma 4.3. □

PROOF OF LEMMA 5.4. Note that all of the bounds in this lemma hold with ζ-high probability, not in expectation. For simplicity, we will subsume this in the notation.

First let us prove a slightly different result. Define the binary function $\widetilde{\mathcal{I}}(A,\mathbf{k})$ [similar to $\mathcal{I}(A,\mathbf{k})$] as follows. $\widetilde{\mathcal{I}}(A,\mathbf{k})$ is equal to 1 in the following scenarios: (1) for any i, j in the same block of A we have $k_i=k_j$, (2) if i, j are in different blocks of A, we have $k_i\ne k_j$ except that if one of the indices i, j is in the block of A which contains exactly two elements, then k_i is allowed to be equal to k_j. In all other instances $\widetilde{\mathcal{I}}(A,\mathbf{k})=0$. For instance, in the previous example (5.1), we have

$$\widetilde{\mathcal{I}}(A,\mathbf{k})=\mathbf{1}(k_2=k_4)\mathbf{1}(k_3=k_5=k_6)\mathbf{1}(k_1\ne k_3).$$

We first claim that

$$
\begin{aligned}
&\sum_{k_1,k_2,\ldots,k_{2a+2b}=1}^{M}\widetilde{\mathcal{I}}(A,\mathbf{k})\eta_0^a(Y_{k_1k_2}\cdots Y_{k_{2a-1}k_{2a}})(Z_{k_{2a+1}k_{2a+2}}\cdots Z_{k_{2a+2b-1}k_{2a+2b}}) \\
&\qquad= O_\varepsilon\bigl((N^{2/3})^{a+b}(N^{1/2})^{\mathcal{N}(A,1)+\mathbf{I}_{(A,3)}}(N^{1/3})^{\mathcal{N}(A,2)}\bigr).
\end{aligned}
\tag{5.6}
$$

Let us first prove (5.6) when $\mathbf{I}_{(A,3)}=0$. Define the functions

$$g_1(m):=\operatorname{Tr}|Z^m|,\qquad g_2(m):=\sqrt{(\operatorname{Tr}|Z|^{2m})},\qquad 1\le m\le 2a+b.$$

We will show that the

$$\text{l.h.s. of (5.6)}\le O_\varepsilon\Bigl(\eta_0^a(N^{1/2})^{\mathcal{N}(A,1)}\prod_i g_{\alpha_i}(m_i)\Bigr),\tag{5.7}$$

where $\alpha_i\in\{1,2\}$ and $m_i\le 2a+b$.

To this end, we will use the following 2–1–3 rule:

- 2: If the index i appears in a block of A which contains exactly two elements, first sum up over the index k_i. Then estimate the remaining terms with absolute sum. For example, let $A = \{\{1\}, \{2, 3\}, \{4\}\}$. Recall that $Y = Z^2$,

$$\left|\sum_{k_1,k_2,k_3} \widetilde{\mathcal{I}}(A, \mathbf{k}) Y_{k_1k_2} Z_{k_2k_4}\right| = \left|\sum_{k \neq l} (YZ)_{kl}\right| \leq \sum_{kl} |(YZ)_{kl}| = \sum_{kl} |(Z^3)_{kl}|.$$

- 1: Next do the summation over the index k_i if i appears in the block of A which contains only one element as follows:

$$\sum_l |(Z^m)_{kl}| \leq CN^{1/2}\sqrt{(|Z|^{2m})_{kk}}, \qquad \sum_{kl} |(Z^m)_{kl}| \leq CN\sqrt{\operatorname{Tr}|Z|^{2m}}.$$

  In the above inequalities, we have used the Cauchy–Schwarz and the fact that Z is a symmetric matrix. Note that each summation of the above kind brings an extra $N^{1/2}$ factor.

- 3: Finally, sum up over the other indices. After the first two steps, (5.6) will be reduced to the product of following terms:

$$(N^{1/2})^{\mathcal{N}(A,1)}, \qquad |\operatorname{Tr} Z^r|, \qquad \sqrt{\operatorname{Tr}|Z|^{2r}}, \qquad r \leq 2a + b,$$

  and terms of the form

$$\sum_k \prod_{i=1}^m |(Z^{m_i})_{kk}| \prod_{j=1}^n \sqrt{(|Z|^{2n_j})_{kk}}, \qquad 2 \leq m + n. \tag{5.8}$$

  If $m + n = 2$, then using the Cauchy–Schwarz inequality, (5.8) can be estimated as

$$\prod_{i=1}^m \prod_{j=1}^n \sum_k |(Z^{m_i})_{kk}| \sqrt{(|Z|^{2n_j})_{kk}} \leq \prod_{i=1}^m \prod_{j=1}^n \sqrt{\operatorname{Tr}|Z|^{2m_i}} \sqrt{\operatorname{Tr}|Z|^{2n_j}}. \tag{5.9}$$

  For $m + n > 2$, we bound $m + n - 2$ of them $[|(Z^{m_i})_{kk}|$ or $\sqrt{(|Z|^{2n_j})_{kk}}]$ by the maximum as follows:

$$|(Z^{m_i})_{kk}| \leq \max_k |(Z^{m_i})_{kk}| \leq \sqrt{\operatorname{Tr}|Z|^{2m_i}},$$

$$\sqrt{(|Z|^{2n_j})_{kk}} \leq \max_k \sqrt{(|Z|^{2n_j})_{kk}} \leq \sqrt{\operatorname{Tr}|Z|^{2n_j}},$$

  to reduce to the case of $m + n = 2$ and use the bound (5.9).

Let us give an example in the case $a = 1$, $b = 2$ and $A = \{\{1\}, \{2, 3\}, \{4, 5, 6\}\}$. Then the term (5.6) in this case reduces to

$$\sum_{k_1k_2k_4} \eta_0 Y_{k_1k_2} Z_{k_2k_4} Z_{k_4k_4} \leq \sum_{k_1k_4} \eta_0 |(Z^3)_{k_1k_4}| |Z_{k_4k_4}|,$$

where the above inequality is obtained by applying rule 2. Next, applying rule 1 yields

$$\leq \sum_{k_4} \eta_0 N^{1/2} \sqrt{(|Z|^6)_{k_4k_4}} |Z_{k_4k_4}|$$

and, finally, applying rule 3 leads to the bound

$$\leq \sum_{k_4} \eta_0 N^{1/2} \sqrt{\mathrm{Tr}|Z|^6} \sqrt{\mathrm{Tr}|Z|^2}.$$

Using this 2–1–3 rule described above, we obtain (5.7). By the definition of the 2–1–3 rule, it is easy to see that

$$\sum_i m_i = 2a + b. \tag{5.10}$$

Recall $\eta_0 = N^{-2/3-\varepsilon}$. Using (4.18) and (4.19), we deduce that if $\alpha_i m_i \neq 1$, then

$$g_{\alpha_i}(m_i) \leq O_\varepsilon(N^{2m_i/3}).$$

For $\alpha_i m_i = 1$, using (4.6), (4.7), (2.13) and $m_W = O(1)$, we see that $g_1(1) = O_\varepsilon(N)$. Thus,

$$g_{\alpha_i}(m_i) \leq O_\varepsilon(N^{2m_i/3})(N^{1/3})^{\mathbf{1}(\alpha_i m_i = 1)}. \tag{5.11}$$

Combining equations (5.7)–(5.11), we have the

$$\begin{aligned} &\text{l.h.s. of equation (5.6)} \\ &\quad = O_\varepsilon(N^{1/2})^{\mathcal{N}(A,1)} N^{2a/3+2b/3} (N^{1/3})^{\#\{i : a_i m_i = 1\}}. \end{aligned} \tag{5.12}$$

Now notice that by the definition, the term $g_1(1)$ in (5.7) can only be created during the first step of the 2–1–3 rule, that is, the 2 rule, and, therefore, we deduce that

$$\mathcal{N}(A, 2) = \#\{i : \alpha_i m_i = 1\},$$

which completes the proof of the claim made in (5.6) for the case $\mathbf{I}_{(A,3)} = 0$.

Now consider the case $\mathbf{I}_{(A,3)} = 1$. Using the fact that Y, Z are symmetric matrices and the relation $Y = Z^2$, we deduce that the term

$$\sum \widetilde{\mathcal{I}}(A, \mathbf{k})(Y_{k_1k_2} \cdots Y_{k_{2a-1}k_{2a}})(Z_{k_{2a+1}k_{2a+2}} \cdots Z_{k_{2a+2b-1}k_{2a+2b}})$$

reduces to one of the following situations:

$$\begin{aligned} &\sum_{k_1,k_2,\dots,k_{2a+2b}} \widetilde{\mathcal{I}}(A, \mathbf{k})(Y_{k_1k_2} \cdots Y_{k_{2a-1}k_{2a}})(Z_{k_{2a+1}k_{2a+2}} \cdots Z_{k_{2a+2b-1}k_{2a+2b}}) \\ &\quad = \begin{cases} \displaystyle\sum_{k_1,k_2=1}^{M} Z_{k_1k_1}^{m_1} Z_{k_1k_2}^{m_2} Z_{k_2k_2}^{m_3}, \\ \displaystyle\sum_{k_1,k_2=1}^{M} Z_{k_1k_2}^{m_1} Z_{k_1k_2}^{m_2} Z_{k_1k_2}^{m_3}, \end{cases} \end{aligned} \tag{5.13}$$

for $m_i \in \{1, 2\}, i \in \{1, 2, 3\}$. We bound the first scenario above as

$$\begin{aligned} &\Big|\sum_{k_1k_2}(Z^{m_1})_{k_1k_1}(Z^{m_2})_{k_1k_2}(Z^{m_3})_{k_2k_2}\Big| \\ &\quad\le \sum_{k_1k_2}\big|(Z^{m_1})_{k_1k_1}(Z^{m_2})_{k_1k_2}\big|\max_k\big|(Z^{m_3})_{kk}\big| \\ &\quad\le \sum_{k_1k_2}\big|(Z^{m_1})_{k_1k_1}(Z^{m_2})_{k_1k_2}\big|\sqrt{\operatorname{Tr}|Z|^{2m_3}}. \end{aligned} \tag{5.14}$$

Using rule 1 and rule 3 above yields

$$\sum_{k_1k_2}\big|(Z^{m_1})_{k_1k_1}(Z^{m_2})_{k_1k_2}\big| \le CN^{1/2}\sqrt{\operatorname{Tr}|Z|^{2m_1}}\sqrt{\operatorname{Tr}|Z|^{2m_2}}$$

and, thus,

$$\begin{aligned} \eta_a^0\sum_{k_1,k_2=1}^{M}\big|Z_{k_1k_1}^{m_1}Z_{k_1k_2}^{m_2}Z_{k_2k_2}^{m_3}\big| &\le C\eta_a^0N^{1/2}\sqrt{\operatorname{Tr}|Z|^{2m_1}}\sqrt{\operatorname{Tr}|Z|^{2m_2}}\sqrt{\operatorname{Tr}|Z|^{2m_3}} \\ &= O_\varepsilon\big(N^{-2a/3+1/2}\big)O_\varepsilon\big(N^{2/3(m_1+m_2+m_3)}\big) \\ &= O_\varepsilon\big(N^{2/3(a+b)+1/2}\big), \end{aligned}$$

where in the last inequality we have used the fact that $\sum_i m_i = 2a + b$. For the second case in (5.13), first we note

$$\max_{kl}\big|(Z^m)_{kl}\big| \le \sqrt{\operatorname{Tr}|Z|^{2m}}.$$

Now using the Cauchy–Schwarz inequality,

$$\sum_{k_1,k_2}\big|(Z^{m_1})_{k_1k_2}(Z^{m_2})_{k_1k_2}(Z^{m_3})_{k_1k_2}\big| \le \sqrt{\operatorname{Tr}|Z|^{2m_1}}\sqrt{\operatorname{Tr}|Z|^{2m_2}}\sqrt{\operatorname{Tr}|Z|^{2m_3}}$$

and, thus,

$$\begin{aligned} \eta_a^0\sum_{k_1,k_2=1}^{M}\big|Z_{k_1k_2}^{m_1}Z_{k_1k_2}^{m_2}Z_{k_1k_2}^{m_3}\big| &\le C\eta_a^0\sqrt{\operatorname{Tr}|Z|^{2m_1}}\sqrt{\operatorname{Tr}|Z|^{2m_2}}\sqrt{\operatorname{Tr}|Z|^{2m_3}} \\ &= O_\varepsilon\big(N^{-2a/3}\big)O_\varepsilon\big(N^{2/3(m_1+m_2+m_3)}\big) \\ &= O_\varepsilon\big(N^{2/3(a+b)}\big). \end{aligned}$$

Summarizing the above computations, and noticing that $\mathcal{N}(A, 1) = \mathcal{N}(A, 2) = 0$ when $\mathbf{I}_{(A,3)} = 1$, we obtain the bound

$$\begin{aligned} &\eta_0^a\widetilde{\mathcal{I}}(A,\mathbf{k})\big|(Y_{k_1k_2}\cdots Y_{k_{2a-1}k_{2a}})(Z_{k_{2a+1}k_{2a+2}}\cdots Z_{k_{2a+2b-1}k_{2a+2b}})\big| \\ &\quad= O_\varepsilon\big(N^{2/3(a+b)+1/2}\big) \\ &\quad= O_\varepsilon\big(\big(N^{2/3}\big)^{a+b}\big(N^{1/2}\big)^{\mathbf{I}_{(A,3)}}\big), \end{aligned}$$

proving the claim (5.6) when $\mathbf{I}_{(A,3)} = 1$.

Now we return to prove Lemma 5.4. One can see that for any partition A of the set $\{1, 2, \dots, 2a + 2b\}$ and a vector $\mathbf{k}$, the function $\mathcal{I}(A, \mathbf{k})$ can be written as linear combinations of the functions $\widetilde{\mathcal{I}}(A_i, \mathbf{k})$ for some partitions A_i's of the set $\{1, 2, \dots, 2a + 2b\}$ such that

$$\mathcal{N}(A_i, 1) \le \mathcal{N}(A, 1), \qquad \mathcal{N}(A_i, 2) \le \mathcal{N}(A, 2), \qquad \mathbf{I}_{A_i,3} = \mathbf{I}_{(A,3)}.$$

For instance, for A given in (5.1),

$$\widetilde{\mathcal{I}}(A, \mathbf{k}) = \mathbf{1}(k_2 = k_4)\mathbf{1}(k_3 = k_5 = k_6)\mathbf{1}(k_1 \neq k_3),$$

we have the identity

$$\mathcal{I}(A, \mathbf{k}) = \widetilde{\mathcal{I}}(A, \mathbf{k}) - \widetilde{\mathcal{I}}(A_1, \mathbf{k}) - \widetilde{\mathcal{I}}(A_2, \mathbf{k}),$$

where $A_1 = \{\{1\}, \{2, 3, 4, 5, 6\}\}$ and $A_2 = \{\{1, 2, 4\}, \{3, 5, 6\}\}$. Now the lemma follows from (5.6) and the proof is finished. □

PROOF OF LEMMA 5.5. For any $k_1, k_2, \dots, k_m \in \{1, 2, \dots, M\}$ and $m \in \mathbb{N}$, by definition we have

$$\begin{aligned}\mathbb{E}\mathcal{I}(A, \mathbf{k}) \prod_{i=1}^{m} y_{k_i} &= \mathbb{E}\mathcal{I}(A, \mathbf{k}) \frac{\prod_{i=1}^{m} \widetilde{y}_{k_i}}{(\sum_j \widetilde{y}_j^2)^{m/2}} \\ &= \mathbb{E}\mathcal{I}(A, \mathbf{k}) \prod_{i=1}^{m} \widetilde{y}_{k_i} \Biggl[1 - \sum_{j=1}^{M} \Bigl(\frac{1}{M} - \widetilde{y}_j^2\Bigr)\Biggr]^{-m/2}.\end{aligned} \tag{5.15}$$

Using large deviation bounds, it is easy to see that for any $\varepsilon > 0$

$$\sum_{j=1}^{M} \Bigl(\frac{1}{M} - \widetilde{y}_j^2\Bigr) = O_\varepsilon(N^{-1/2}). \tag{5.16}$$

Therefore, by the Taylor expansion,

$$\begin{aligned}&\mathbb{E}\mathcal{I}(A, \mathbf{k}) \prod_{i=1}^{2a+2b} y_{k_i} - \mathbb{E}\mathcal{I}(A, \mathbf{k}) \prod_{i=1}^{2a+2b} \widetilde{y}_{k_i} \\ &\quad = \sum_{n=1}^{\infty} C_n \mathbb{E}\Biggl[\mathcal{I}(A, \mathbf{k}) \Biggl(\prod_{i=1}^{2a+2b} \widetilde{y}_{k_i}\Biggr) \Biggl(\sum_{r_1, r_2, \dots, r_n = 1}^{M} \prod_{j=1}^{n} \Bigl(\frac{1}{M} - \widetilde{y}_{r_j}^2\Bigr)\Biggr)\Biggr],\end{aligned} \tag{5.17}$$

where $C_n = C_{a,b,n}$ is a combinatorial factor. Using (5.16), the r.h.s. of equation (5.17) may be expressed as

$$\begin{aligned}&= \sum_{n=1}^{n_0} C_n \mathbb{E}\Biggl[\mathcal{I}(A, \mathbf{k}) \Biggl(\prod_{i=1}^{2a+2b} \widetilde{y}_{k_i}\Biggr) \Biggl(\sum_{r_1, r_2, \dots, r_n = 1}^{M} \prod_{j=1}^{n} \Bigl(\frac{1}{M} - \widetilde{y}_{r_j}^2\Bigr)\Biggr)\Biggr] \\ &\quad + O_\varepsilon\bigl((N^{-1/2})^{2a+2b+n_0}\bigr)\end{aligned} \tag{5.18}$$

for some fixed $n_0 \in \mathbb{N}$ (say, $n_0 = 20$).

Since $n_0, a, b = O(1)$, the combinatorial factors do not increase with N, that is, $C_n = O(1)$, and, thus, we can bound

$$\mathbb{E}\left[\mathcal{I}(A, \mathbf{k})\left(\prod_{i=1}^{2a+2b} \widetilde{y}_{k_i}\right)\left(\prod_{j=1}^{n}\left(\frac{1}{M} - \widetilde{y}_{r_j}^2\right)\right)\right] \tag{5.19}$$

as follows. Notice that the number of distinct indices k_i in (5.19) is equal to the number of blocks in the partition A. Thus, for a given set of values for the indices $r_1, r_2, \dots, r_n$, the term (5.19) is nonzero only if at least $\mathcal{N}(A, 1)$ of the indices r_j belong to the set $\{k_1, k_2, \dots, k_{2a+2b}\}$. The above observation also implies that for (5.19) to be nonzero we must have

$$n \geq \mathcal{N}(A, 1). \tag{5.20}$$

Furthermore, the indices r_j which do not belong to the set $\{k_1, k_2, \dots, k_{2a+2b}\}$ must appear more than once since $\mathbb{E}(1/M - y_{r_j}^2) = 0$. This crucial observation implies that, if the term (5.19) is nonzero and

$$\mathcal{N}(A, 1) = 0 \qquad \text{then } n \geq 2. \tag{5.21}$$

Therefore, the number of nonzero terms in the sum

$$\sum_{r_1, r_2, \dots, r_n = 1}^{M} \mathbb{E}\left[\mathcal{I}(A, \mathbf{k})\left(\prod_{i=1}^{2a+2b} \widetilde{y}_{k_i}\right)\left(\prod_{j=1}^{n}\left(\frac{1}{M} - \widetilde{y}_{r_j}^2\right)\right)\right] \tag{5.22}$$

is $O((N^{1/2})^{n-\mathcal{N}(A,1)})$, and each of these terms are of the size $O_\varepsilon(N^{-(a+b)-n})$, yielding

$$\begin{aligned}\sum_{r_1, r_2, \dots, r_n = 1}^{M} & \mathbb{E}\left[\mathcal{I}(A, \mathbf{k})\left(\prod_{i=1}^{2a+2b} \widetilde{y}_{k_i}\right)\left(\prod_{j=1}^{n}\left(\frac{1}{M} - \widetilde{y}_{r_j}^2\right)\right)\right] \\ & \leq O_\varepsilon\big(N^{-(a+b)-n/2-\mathcal{N}(A,1)/2}\big).\end{aligned} \tag{5.23}$$

Combining (5.23) with (5.20) and the observation made in (5.21), we obtain that

$$\begin{aligned}\sum_{r_1, r_2, \dots, r_n = 1}^{M} & \mathbb{E}\left[\mathcal{I}(A, \mathbf{k})\left(\prod_{i=1}^{2a+2b} \widetilde{y}_{k_i}\right)\left(\prod_{j=1}^{n}\left(\frac{1}{M} - \widetilde{y}_{r_j}^2\right)\right)\right] \\ & \leq O_\varepsilon\big(N^{-(a+b)-\max\{\mathcal{N}(A,1),1\}}\big),\end{aligned} \tag{5.24}$$

obtaining (5.2), and the proof is finished. □

Acknowledgments. The authors would like to thank Jiefeng Jiang, two anonymous referees, the Associate Editor and the Editor for very useful comments.

REFERENCES

[1] BAO, Z. G., PAN, G. M. and ZHOU, W. (2011). Tracy–Widom law for the extreme eigenvalues of sample correlation matrices. Preprint. Available at arXiv:1110.5208.

[2] CHATTERJEE, S. (2006). A generalization of the Lindeberg principle. *Ann. Probab.* **34** 2061–2076. MR2294976

[3] ERDŐS, L., KNOWLES, A., YAU, H. T. and YIN, J. (2011). Spectral statistics of Erdós–Rényi graphs II: Eigenvalue spacing and the extreme eigenvalues. Preprint. Available at arXiv:1103.3869.

[4] ERDŐS, L., KNOWLES, A., YAU, H. T. and YIN, J. (2012). Spectral statistics of Erdós–Rényi graphs I: Local semicircle law. *Ann. Probab.* To appear. Available at arXiv:1103.1919.

[5] ERDŐS, L., PÉCHÉ, S., RAMÍREZ, J. A., SCHLEIN, B. and YAU, H.-T. (2010). Bulk universality for Wigner matrices. *Comm. Pure Appl. Math.* **63** 895–925. MR2662426

[6] ERDŐS, L., SCHLEIN, B. and YAU, H.-T. (2009). Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices. *Ann. Probab.* **37** 815–852. MR2537522

[7] ERDŐS, L., SCHLEIN, B. and YAU, H.-T. (2009). Local semicircle law and complete delocalization for Wigner random matrices. *Comm. Math. Phys.* **287** 641–655. MR2481753

[8] ERDŐS, L., SCHLEIN, B. and YAU, H.-T. (2010). Wegner estimate and level repulsion for Wigner random matrices. *Int. Math. Res. Not. IMRN* **3** 436–479. MR2587574

[9] ERDŐS, L., SCHLEIN, B. and YAU, H.-T. (2011). Universality of random matrices and local relaxation flow. *Invent. Math.* **185** 75–119. MR2810797

[10] ERDŐS, L., SCHLEIN, B., YAU, H.-T. and YIN, J. (2012). The local relaxation flow approach to universality of the local statistics for random matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1–46. MR2919197

[11] ERDŐS, L., YAU, H.-T. and YIN, J. (2011). Universality for generalized Wigner matrices with Bernoulli distribution. *J. Comb.* **2** 15–81. MR2847916

[12] ERDŐS, L., YAU, H. T. and YIN, J. (2012). Bulk universality for generalized Wigner matrices. *Probab. Theory Related Fields.* To appear. Available at arXiv:1001.3453.

[13] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Rigidity of eigenvalues of generalized Wigner matrices. *Adv. Math.* **229** 1435–1515. MR2871147

[14] FELDHEIM, O. N. and SODIN, S. (2010). A universality result for the smallest eigenvalues of certain sample covariance matrices. *Geom. Funct. Anal.* **20** 88–123. MR2647136

[15] JIANG, T. (2004). The limiting distributions of eigenvalues of sample correlation matrices. *Sankhyā* **66** 35–48. MR2082906

[16] JOHANSSON, K. (2012). Universality for certain Hermitian Wigner matrices under weak moment conditions. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 47–79. MR2919198

[17] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

[18] JOHNSTONE, I. M. (2007). High dimensional statistical inference and random matrices. In *International Congress of Mathematicians. Vol. I* 307–333. Eur. Math. Soc., Zürich. MR2334195

[19] JOHNSTONE, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *Ann. Statist.* **36** 2638–2716. MR2485010

[20] KNOWLES, A. and YIN, J. (2011). Eigenvector distribution of Wigner matrices. Preprint. Available at arXiv:1102.0057.

[21] KNOWLES, A. and YIN, J. (2011). The isotropic semicircle law and deformation of Wigner matrices. Preprint. Available at arXiv:1110.6449.

[22] MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb.* (*N.S.*) **72** 507–536. MR0208649

[23] PÉCHÉ, S. (2009). Universality results for the largest eigenvalues of some sample covariance matrix ensembles. *Probab. Theory Related Fields* **143** 481–516. MR2475670

[24] PILLAI, N. S. and YIN, J. (2011). Universality of covariance matrices. Preprint. Available at arXiv:1110.2501.

[25] SOSHNIKOV, A. (2002). A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *J. Stat. Phys.* **108** 1033–1056. Dedicated to David Ruelle and Yasha Sinai on the occasion of their 65th birthdays. MR1933444

[26] TAO, T. and VU, V. (2010). Random matrices: Universality of local eigenvalue statistics up to the edge. *Comm. Math. Phys.* **298** 549–572. MR2669449

[27] TAO, T. and VU, V. (2012). Random covariance matrices: Universality of local statistics of eigenvalues. *Ann. Probab.* **40** 1285–1315.

DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
1 OXFORD STREET
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: pillai@stat.harvard.edu

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF WISCONSIN–MADISON
480 LINCOLN DR.
MADISON, WISCONSIN 53706
USA
E-MAIL: jyin@math.wisc.edu