Submitted to the Annals of Statistics

CONVERGENCE OF EIGENVECTOR EMPIRICAL SPECTRAL DISTRIBUTION OF SAMPLE COVARIANCE MATRICES

By Haokai Xi [†], Fan Yang [†,‡] and Jun Yin [†,‡,*]

University of Wisconsin-Madison[†] *and University of California, Los Angeles*[‡]

The *eigenvector empirical spectral distribution* (VESD) is a useful tool in studying the limiting behavior of eigenvalues and eigenvectors of covariance matrices. In this paper, we study the convergence rate of the VESD of sample covariance matrices to the deformed Marčenko-Pastur (MP) distribution. Consider sample covariance matrices of the form $\Sigma^{1/2}XX^*\Sigma^{1/2}$, where $X = (x_{ij})$ is an $M \times N$ random matrix whose entries are independent random variables with mean zero and variance N^{-1}, and Σ is a deterministic positive-definite matrix. We prove that the Kolmogorov distance between the *expected VESD* and the deformed MP distribution is bounded by $N^{-1+\epsilon}$ for any fixed $\epsilon > 0$, provided that the entries $\sqrt{N}x_{ij}$ have uniformly bounded 6th moments and $|N/M - 1| \geq \tau$ for some constant $\tau > 0$. This result improves the previous one obtained in [44], which gave the convergence rate $O(N^{-1/2})$ assuming *i.i.d.* X entries, bounded 10th moment, $\Sigma = I$ and $M < N$. Moreover, we also prove that under the finite 8th moment assumption, the convergence rate of the VESD is $O(N^{-1/2+\epsilon})$ almost surely for any fixed $\epsilon > 0$, which improves the previous bound $N^{-1/4+\epsilon}$ in [44].

1. Introduction and main results. Sample covariance matrices are fundamental objects in multivariate statistics. The population covariance matrix of a centered random vector $\mathbf{y} \in \mathbb{R}^M$ is $\Sigma = \mathbb{E}\mathbf{y}\mathbf{y}^*$. Given N independent samples $(\mathbf{y}_1, \cdots, \mathbf{y}_N)$ of $\mathbf{y}$, then the sample covariance matrix $Q := N^{-1}\sum_i \mathbf{y}_i\mathbf{y}_i^*$ is the simplest estimator for Σ. In fact, if M is fixed, then Q converges almost surely to Σ as $N \to \infty$. However, in many modern applications, the advance of technology has led to high dimensional data where M is comparable to or even larger than N. In this setting, Σ cannot be estimated through Q directly, but some properties of Σ can be inferred from the eigenvalue and eigenvector statistics of Q. The large dimensional covariance matrices have more and more applications in various fields, such

[*]Supported by NSF Career Grant DMS-1552192 and Sloan fellowship.

MSC 2010 subject classifications: Primary 15B52, 62E20; secondary 62H99

Keywords and phrases: Sample covariance matrix, Empirical spectral distribution, Eigenvector empirical spectral distribution, Marčenko-Pastur distribution

as statistics [13, 21, 22, 24], economics [33] and population genetics [34].

In this paper, we consider sample covariance matrices of the form $Q_1 := \Sigma^{1/2} X X^* \Sigma^{1/2}$, where $X = (x_{ij})$ is an $M \times N$ real or complex data matrix whose entries are independent (but not necessarily identically distributed) random variables satisfying

$$\mathbb{E}x_{ij} = 0, \quad \mathbb{E}|x_{ij}|^2 = N^{-1}, \quad 1 \le i \le M, \ 1 \le j \le N, \tag{1.1}$$

and the population covariance matrix $\Sigma := \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_M)$ is a deterministic positive-definite matrix. If the entries of X are complex, then we assume in addition that

$$\mathbb{E}x_{ij}^2 = 0, \quad 1 \le i \le M, \ 1 \le j \le N. \tag{1.2}$$

Define the aspect ratio $d_N := N/M$. In this paper, we are interested in the high dimensional case with $\lim_{N\to\infty} d_N = d \in (0, \infty)$. We will also consider the $N \times N$ matrix $Q_2 := X^* \Sigma X$, which share the same nonzero eigenvalues with Q_1.

A simple but important type of covariance matrix is the one with $\Sigma = \sigma^2 I$ (i.e. the null case). In applications of spectral analysis of large dimensional random matrices, one important problem is the convergence rate of the empirical spectral distributions (ESD). It is well-known that the ESD $F_{XX^*}^{(M)}$ of XX^* converges weakly to the Marčenko-Pastur (MP) law F_{MP} [31]. Moreover, one can use the Kolmogorov distance

$$\|F_{XX^*}^{(M)} - F_{MP}\| := \sup_x |F_{XX^*}^{(M)}(x) - F_{MP}(x)|$$

to measure the convergence rate of the ESD. The convergence rate bound for sample covariance matrices was first established in [2], and later improved in [19] to $O(N^{-1/2})$ in probability under the finite 8th moment condition. In [36], the authors proved an almost optimal bound that $\|F_{XX^*}^{(M)} - F_{MP}\| = O(N^{-1+\epsilon})$ with high probability for any fixed $\epsilon > 0$ under the sub-exponential decay assumption.

The research on the asymptotic properties of eigenvectors of large dimensional random matrices is generally harder and much less developed. However, the eigenvectors play an important role in high dimensional statistics. In particular, the principal component analysis (PCA) is now favorably recognized as a powerful technique for dimensionality reduction, and the eigenvectors corresponding to the largest eigenvalues are the directions of the principal components. The earlier work on the properties of eigenvectors goes back to Anderson [1], where the author proved that the eigenvectors of

the Wishart matrix are asymptotically normal and isotropic when M is fixed and $N \to \infty$. For the high dimensional case, Johnstone [23] proposed the spiked model to test the existence of principal components. Then Paul [35] studied the directions of eigenvectors corresponding to spiked eigenvalues. In [30], Ma proposed an iterative thresholding approach to estimate sparse principal subspaces in the setting of a high dimensional spiked covariance model. Using a reduction scheme which reduces the sparse PCA problem to a high-dimensional multivariate regression problem, [11] established the optimal rates of convergence for estimating the principal subspace for a large class of spiked covariance matrices. One can see the references in [11, 30] for more literatures on sparse PCA and spiked covariance matrices.

For the test of the existence of spiked eigenvalues, we first need to study the properties of the eigenmatrices in the null case. If $\Sigma = \sigma^2 I$, then the eigenmatrix is expected to be asymptotically Haar distributed (i.e. uniformly distributed over the unitary group). However, formulating the terminology "asymptotically Haar distributed" is far from trivial since the dimension M is increasing. Following the approach in [38, 39, 3, 43, 44], we will use the *eigenvector empirical spectral distribution* (VESD) to characterize the asymptotical Haar property. Suppose

$$\Sigma^{1/2} X = \sum_{1 \le k \le N \wedge M} \sqrt{\lambda_k} \xi_k \zeta_k^* \tag{1.3}$$

is a singular value decomposition of $\Sigma^{1/2} X$, where

$$\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_{N \wedge M} \ge 0 = \lambda_{N \wedge M + 1} = \ldots = \lambda_{N \vee M},$$

$\{\xi_k\}_{k=1}^M$ are the left-singular vectors, and $\{\zeta_k\}_{k=1}^N$ are the right-singular vectors. Then for deterministic unit vectors $\mathbf{u} \in \mathbb{C}^M$ and $\mathbf{v} \in \mathbb{C}^N$, we define the VESD of $Q_{1,2}$ as

$$F^{(M)}_{Q_1,\mathbf{u}}(x) = \sum_{k=1}^{M} |\langle \xi_k, \mathbf{u} \rangle|^2 \mathbf{1}_{\{\lambda_k \le x\}}, \quad F^{(N)}_{Q_2,\mathbf{v}}(x) = \sum_{k=1}^{N} |\langle \zeta_k, \mathbf{v} \rangle|^2 \mathbf{1}_{\{\lambda_k \le x\}}. \tag{1.4}$$

Now we apply the above formulations to the null case. Adopting the ideas of [38, 39], we define the stochastic process as

$$X_{M,\mathbf{u}}(t) := \sqrt{\frac{M}{2}} \sum_{k=1}^{\lfloor Mt \rfloor} \left(|\langle \xi_k, \mathbf{u} \rangle|^2 - M^{-1} \right).$$

If the eigenmatrix of XX^* is Haar distributed, then the vector $\mathbf{y} := (\langle \xi_k, \mathbf{u} \rangle)_{k=1}^M$ is uniformly distributed over the unit sphere, and $X_{M,\mathbf{u}}(t)$ would converge

to a Brownian bridge by Donsker's theorem. Thus the convergence of $X_{M,\mathbf{u}}$ to a Brownian bridge characterizes the asymptotical Haar property of the eigenmatrix. For convenience, we can consider the time transformation

$$X_{M,\mathbf{u}}(F^{(M)}_{XX^*}(x)) = \sqrt{\frac{M}{2}}\left(F^{(M)}_{XX^*,\mathbf{u}}(x) - F^{(M)}_{XX^*}(x)\right).$$

Thus the problem is reduced to the study of the difference between the VESD and the ESD. It was already proved in [3, 8] that $F^{(M)}_{XX^*,\mathbf{u}}$ also converges weakly to the MP law for any sequence of unit vectors $\mathbf{u} \in \mathbb{R}^M$. On the other hand, compared with ESD, much less has been known about the convergence rate of the VESD. The best result so far was obtained in [44], where the authors proved that if $d_N > 1$ and the entries of X are *i.i.d.* centered random variables, then $\|\mathbb{E}F^{(M)}_{XX^*,\mathbf{u}} - F_{MP}\| = O(N^{-1/2})$ under the finite 10th moment assumption, and $\|F^{(M)}_{XX^*,\mathbf{u}} - F_{MP}\| = O(N^{-1/4+\epsilon})$ almost surely under the finite 8th moment assumption. However, we find that both of these bounds are far away from being optimal, and can be improved with a different method. This is one purpose of this paper.

We will also extend the above formulation to include sample covariance matrices with general population Σ. For a non-scalar Σ, the eigenmatrix of Q_1 is not asymptotically Haar distributed anymore. For its distribution, we conjecture that the eigenvectors of Q_1 are asymptotically independent, and each ξ_k is asymptotically normal with covariance matrix given by some $\mathbf{D}_k$. In fact, our results in this paper suggest that $\mathbf{D}_k$ takes the form $\mathbf{F}_{1c}(\gamma_k) - \mathbf{F}_{1c}(\gamma_{k+1})$, where γ_k is defined in (1.14) to denote the classical location for λ_k, and $\mathbf{F}_{1c}$ is a matrix-valued function defined in (1.17) with the property that $\langle \mathbf{u}, \mathbf{F}_{1c}\mathbf{u}\rangle$ is the asymptotic distribution of the VESD $F_{Q_1,\mathbf{u}}$ for any $\mathbf{u} \in \mathbb{C}^M$. Again, since the dimension M increases to ∞, the above property is hard to formulate. One way is to consider the finite-dimensional restriction in the following sense: if we fix an $m \in \mathbb{N}$, then for any fixed unit vector $\mathbf{u} \in \mathbb{C}^M$ and $\{i_1, \cdots, i_m\} \subseteq \{1, \cdots, N \wedge M\}$, we should have asymptotically

$$(\langle \xi_{i_1}, \mathbf{u}\rangle, \cdots, \langle \xi_{i_m}, \mathbf{u}\rangle) \sim \mathcal{N}_m\left(0, \langle \mathbf{u}, \mathbf{D}_{i_1}\mathbf{u}\rangle, \ldots, \langle \mathbf{u}, \mathbf{D}_{i_m}\mathbf{u}\rangle\right). \tag{1.5}$$

(In fact, for a nice choice of Σ in the sense of Definition 1.2, $\langle \mathbf{u}, \mathbf{D}_k\mathbf{u}\rangle$ is typically of order N^{-1}.) We can also adopt the approach as above, that is to investigate the stochastic process

$$X^{\Sigma}_{M,\mathbf{u}}(t) := \sqrt{\frac{M}{2}} \sum_{k=1}^{\lfloor Mt \rfloor} \left(|\langle \xi_k, \mathbf{u}\rangle|^2 - \langle \mathbf{u}, \mathbf{D}_k\mathbf{u}\rangle\right). \tag{1.6}$$

If $M < N$, we conjecture that $X^{\Sigma}_{M,\mathbf{u}}(t)$ converges to the following process for $0 \le t \le 1$:

$$\mathbf{B}^{\Sigma}(t) := \int_0^t \langle \mathbf{u}, \mathbf{F}_{1c}\mathbf{u}\rangle \circ F_{1c}^{-1} dB_t \quad \text{conditioning on} \quad \mathbf{B}_{\Sigma}(1) = 0,$$

where B_t denotes the standard Brownian motion, F_{1c} is the asymptotic ESD of Q_1 defined in (1.11), and F_{1c}^{-1} denotes the quantile function. As before, we can study the above process through the time transformaton $X^{\Sigma}_{M,\mathbf{u}}(F_{Q_1}(x))$, where F_{Q_1} is the ESD of Q_1. Due to the rigidity of eigenvalues in Theorem 2.7 of this paper, we have for all x,

$$\sqrt{\frac{2}{M}} X^{\Sigma}_{M,\mathbf{u}}(F_{Q_1}(x)) = F_{Q_1,\mathbf{u}}(x) - \langle \mathbf{u}, \mathbf{F}_{1c}(x)\mathbf{u}\rangle + O(N^{-1+\epsilon})$$

with very high proability for any fixed $\epsilon > 0$. Thus we need to study the convergence rate of $F_{Q_1,\mathbf{u}}$ to $\langle \mathbf{u}, \mathbf{F}_{1c}\mathbf{u}\rangle$, and this is our main goal. In fact, we will prove that the convergence rate of $\mathbb{E}F_{Q_1,\mathbf{u}}$ is $O(N^{-1+\epsilon})$ for any fixed $\epsilon > 0$, which shows that the limiting process is centered, and the convergence rate of $F_{Q_1,\mathbf{u}}$ is $O(N^{-1/2+\epsilon})$, which partially verify the $\sqrt{M}$ scaling.

We remark that great progress has been made in other directions of the research on eigenvector statistics. For example, one can refer to [17, 8] for the delocalization and isotropic delocalization of eigenvectors, [25, 41] for the universality of eigenvectors, [10] for the local quantum unique ergodicity of eigenvectors and [9] for the eigenvectors of principal components.

1.1. *Main results.* We consider the sample covariance matrices with a general Σ, whose empirical spectral distribution is denoted by

$$\pi \equiv \pi_M := \frac{1}{M}\sum_{i=1}^{M} \delta_{\sigma_i}. \tag{1.7}$$

We assume that there exists a small constant $\tau > 0$ such that

$$\sigma_1 \le \tau^{-1} \quad \text{and} \quad \pi_M([0,\tau]) \le 1 - \tau \ \text{ for all } M. \tag{1.8}$$

The first condition means that the operator norm of Σ is bounded, and the second condition means that the spectrum of Σ cannot concentrate at zero. If π_M converges weakly to some distribution $\hat{\pi}$ as $M \to \infty$, then it was shown in [31] that the ESD of Q_2 converges in probability to some deterministic distribution, which is often referred to as the *deformed Marčenko-Pastur*

law. For any fixed N, we describe the deformed MP law $F_{2c}^{(N)}$ through its Stieltjes transform

$$m_{2c}^{(N)}(z) := \int_{\mathbb{R}} \frac{dF_{2c}^{(N)}(x)}{x-z}, \quad z = E + i\eta \in \mathbb{C}_+.$$

We define $m_{2c}^{(N)}$ as the unique solution to the self-consistent equation

$$\frac{1}{m_{2c}^{(N)}(z)} = -z + d_N^{-1} \int \frac{t}{1 + m_{2c}^{(N)}(z)t} \pi(dt), \tag{1.9}$$

subject to the conditions that $\operatorname{Im} m_{2c}^{(N)}(z) \geq 0$ and $\operatorname{Im} z m_{2c}^{(N)}(z) \geq 0$ for $z \in \mathbb{C}_+$. It is well known that the functional equation (1.9) has a unique solution that is uniformly bounded on $\mathbb{C}_+$ under the assumption (1.8) [31]. Letting $\eta \searrow 0$, we can recover the asymptotic eigenvalue density $\rho_{2c}^{(N)}$ (which in turn gives $F_{2c}^{(N)}$) with the inverse formula

$$\rho_{2c}^{(N)}(E) = \lim_{\eta \searrow 0} \frac{1}{\pi} \operatorname{Im} m_{2c}^{(N)}(E + i\eta). \tag{1.10}$$

Since Q_1 share the same nonzero eigenvalues with Q_2 and has $M - N$ more (or $N - M$ less) zero eigenvalues, we can obtain the asymptotic ESD for Q_1:

$$F_{1c}^{(M)} = d_N F_{2c}^{(N)} + (1 - d_N)\mathbf{1}_{[0,\infty)}. \tag{1.11}$$

In the rest of this paper, we will often omit the super-index N from our notations. The properties of m_{2c} and ρ_{2c} have been studied extensively; see e.g. [4, 5, 7, 20, 27, 37, 40]. The following Lemma 1.1 describes the basic structure of ρ_{2c}. For its proof, one can refer to [27, Appendix A].

LEMMA 1.1 (Support of the deformed MP law). The density ρ_{2c} is a disjoint union of connected components:

$$\operatorname{supp} \rho_{2c} \cap (0, \infty) = \bigcup_{k=1}^{L} [a_{2k}, a_{2k-1}] \cap (0, \infty), \tag{1.12}$$

where $L \in \mathbb{N}$ depends only on π_M. Moreover, $N \int_{a_{2k}}^{a_{2k-1}} \rho_{2c}(x)dx$ is an integer for any $k = 1, \ldots, L$, which give the classical number of eigenvalues in the bulk component $[a_{2k}, a_{2k-1}]$.

We shall call a_k the edges of ρ_{2c}. For any $1 \leq k \leq 2L$, we define

$$N_k := \sum_{2l \leq k} N \int_{a_{2l}}^{a_{2l-1}} \rho_{2c}(x)dx. \tag{1.13}$$

Then we define the classical locations γ_j for the eigenvalues of $\mathcal{Q}_2$ through

$$1 - F_{2c}(\gamma_j) = \frac{j - 1/2}{N}, \quad 1 \le j \le K, \tag{1.14}$$

where we abbreviate $K := M \wedge N$. Note that (1.14) is well-defined since the N_k's are integers. For convenience, we also denote $\gamma_0 := +\infty$ and $\gamma_{K+1} := 0$.

To establish our main result, we need to make some extra assumptions on Σ and π_M, which takes the form of the following regularity conditions.

Definition 1.2 (Regularity). (i) Fix a (small) constant $\tau > 0$. We say that the edge a_k, $k = 1, \dots, 2L$, is τ-regular if

$$a_k \ge \tau, \quad \min_{l \ne k} |a_k - a_l| \ge \tau, \quad \min_i |1 + m_{2c}(a_k)\sigma_i| \ge \tau, \tag{1.15}$$

where $m_{2c}(a_k) := m_{2c}(a_k + i0_+)$.

(ii) We say that the bulk components $[a_{2k}, a_{2k-1}]$ is regular if for any fixed $\tau' > 0$ there exists a constant $c \equiv c_{\tau'} > 0$ such that the density of ρ_{2c} in $[a_{2k} + \tau', a_{2k-1} - \tau']$ is bounded from below by c.

Remark 1.3. The edge regularity conditions (i) has previously appeared (in slightly different forms) in several works on sample covariance matrices [6, 15, 20, 27, 28, 32]. The condition (1.15) ensures a regular square-root behavior of ρ_{2c} near a_k, and in particular rules out outliers. The bulk regularity condition (ii) was introduced in [27], and it imposes a lower bound on the density of eigenvalues away from the edges. These conditions are satisfied by quite general classes of Σ; see e.g. [27, Examples 2.8 and 2.9].

For any $\mathbf{u} \in \mathbb{C}^M$ and $z \in \mathbb{C}_+$, we define

$$m^{(M)}_{1c,\mathbf{u}}(z) := -\langle \mathbf{u}, z^{-1}(1 + m_{2c}(z)\Sigma)^{-1}\mathbf{u}\rangle. \tag{1.16}$$

Then $m^{(M)}_{1c,\mathbf{u}}$ is the Stieltjes transform of a distribution, which we shall denote by $F^{(M)}_{1c,\mathbf{u}}$. From (1.16), it is easy to see that there exists a matrix-valued function $\mathbf{F}^{(M)}_{1c}$ depending on Σ such that $F^{(M)}_{1c,\mathbf{u}} = \langle \mathbf{u}, \mathbf{F}^{(M)}_{1c}\mathbf{u}\rangle$, i.e., we have

$$m^{(M)}_{1c,\mathbf{u}}(z) = \int_{\mathbb{R}} \frac{dF^{(M)}_{1c,\mathbf{u}}(x)}{x - z} = \langle \mathbf{u}, \int_{\mathbb{R}} \frac{d\mathbf{F}^{(M)}_{1c}(x)}{x - z}\mathbf{u}\rangle. \tag{1.17}$$

(Note that if $\Sigma = \sigma^2 I$, then $\mathbf{F}^{(M)}_{1c}(x)$ is a scalar matrix for each x.) It was already proved in [27] that for any sequence of unit vectors $\mathbf{u} \in \mathbb{C}^M$ and $\mathbf{v} \in \mathbb{C}^N$, $F^{(M)}_{Q_1,\mathbf{u}}$ converges weakly to $F_{1c,\mathbf{u}}$ and $F^{(N)}_{Q_2,\mathbf{v}}(x)$ converges weakly to F_{2c}. Now we are ready to state the main results on the convergence rates of the VESD. We first give the main assumptions.

ASSUMPTION 1.4. Fix a (small) constant $\tau > 0$.

(i) $\tau \le d_N \le \tau^{-1}$ and $|d_N - 1| \ge \tau$.

(ii) $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_M)$ is a deterministic positive-definite matrix. We assume that (1.8) holds, all the edges of ρ_{2c} are τ-regular, and all the bulk components of ρ_{2c} are regular in the sense of Definition 1.2.

(iii) $X = (x_{ij})$ is an $M \times N$ real or complex matrix whose entries are independent random variables that satisfy the following moment conditions: there exist constants $C_0, c_0 > 0$ such that for all $1 \le i \le M$, $1 \le j \le N$,

$$|\mathbb{E}x_{ij}| \le C_0 N^{-2-c_0}, \tag{1.18}$$

$$\left|\mathbb{E}|x_{ij}|^2 - N^{-1}\right| \le C_0 N^{-2-c_0}, \tag{1.19}$$

$$\left|\mathbb{E}x_{ij}^2\right| \le C_0 N^{-2-c_0}, \quad \text{if } x_{ij} \text{ is complex}, \tag{1.20}$$

$$\mathbb{E}|x_{ij}|^4 \le C_0 N^{-2}. \tag{1.21}$$

Note that (1.18)-(1.20) are slightly more general than (1.1) and (1.2).

Our main result is stated as the following theorem.

THEOREM 1.5. *Suppose d_N, X and Σ satisfy the Assumption 1.4. Suppose there exist constants $C_1, \phi > 0$ such that*

$$\max_{1\le i\le M, 1\le j\le N} |x_{ij}| \le C_1 N^{-\phi}. \tag{1.22}$$

Let $\mathbf{u} \equiv \mathbf{u}_M \in \mathbb{C}^M$ and $\mathbf{v} \equiv \mathbf{v}_N \in \mathbb{C}^N$ denote sequences of deterministic unit vectors. Then for any fixed (small) $\epsilon > 0$ and (large) $D > 0$, we have

$$\|\mathbb{E}F^{(M)}_{Q_1,\mathbf{u}} - F^{(M)}_{1c,\mathbf{u}}\| + \|\mathbb{E}F^{(N)}_{Q_2,\mathbf{v}} - F^{(N)}_{2c}\| \le N^{-1+\epsilon} \tag{1.23}$$

for sufficiently large N, and

$$\mathbb{P}\left(\|F^{(M)}_{Q_1,\mathbf{u}} - F^{(M)}_{1c,\mathbf{u}}\| + \|F^{(N)}_{Q_2,\mathbf{v}} - F^{(N)}_{2c}\| \ge N^{\epsilon}\left(N^{-2\phi} + N^{-1/2}\right)\right) \le N^{-D}. \tag{1.24}$$

As an immediate corollary of Theorem 1.5, we have the following result.

COROLLARY 1.6. *Suppose d_N and Σ satisfy the Assumption 1.4. Let $X = (x_{ij})$ be an $M \times N$ random matrix whose entries are independent and satisfy (1.1) and (1.2). Suppose there exist constants $a, A > 0$ such that*

$$\limsup_{s\to\infty} s^a \max_{i,j} \mathbb{P}\left(|\sqrt{N}x_{ij}| \ge s\right) \le A \tag{1.25}$$

for all N. Let $\mathbf{u} \equiv \mathbf{u}_M \in \mathbb{C}^M$ and $\mathbf{v} \equiv \mathbf{v}_N \in \mathbb{C}^N$ denote sequences of deterministic unit vectors. Then for any fixed $\epsilon > 0$, if $a \geq 6$, we have

$$\|\mathbb{E}F_{Q_1,\mathbf{u}}^{(M)} - F_{1c,\mathbf{u}}^{(M)}\| + \|\mathbb{E}F_{Q_2,\mathbf{v}}^{(N)} - F_{2c}^{(N)}\| \leq N^{-1+\epsilon} \tag{1.26}$$

for sufficiently large N; if $a \geq 8$, we have

$$\mathbb{P}\left(\limsup_{N\to\infty} N^{1/2-\epsilon}\left(\|F_{Q_1,\mathbf{u}}^{(M)} - F_{1c,\mathbf{u}}^{(M)}\| + \|F_{Q_2,\mathbf{v}}^{(N)} - F_{2c}^{(N)}\|\right) \leq 1\right) = 1. \tag{1.27}$$

Proof of Corollary 1.6. We use a standard cutoff argument. We fix $a > 4$ and choose a constant $\phi > 0$ small enough such that $\left(N^{1/2-\phi}\right)^a \geq N^{2+\omega}$ for some constant $\omega > 0$. Then we introduce the following truncation

$$\tilde{X} := 1_\Omega X, \quad \Omega := \left\{|x_{ij}| \leq N^{-\phi} \text{ for all } 1 \leq i \leq M, 1 \leq j \leq N\right\}.$$

By the tail condition (1.25), we have

$$\mathbb{P}(\tilde{X} \neq X) = O(N^{2-a/2+a\phi}). \tag{1.28}$$

Moreover, we have

$$\begin{aligned}
\mathbb{P}(\tilde{X} \neq X \text{ i.o.}) &= \lim_{k\to\infty} \mathbb{P}\left(\cup_{N=k}^\infty \cup_{i=1}^M \cup_{j=1}^N \left\{|x_{ij}| \geq N^{-\phi}\right\}\right) \\
&= \lim_{k\to\infty} \mathbb{P}\left(\cup_{t=k}^\infty \cup_{N\in[2^t,2^{t+1})} \cup_{i=1}^M \cup_{j=1}^N \left\{|x_{ij}| \geq N^{-\phi}\right\}\right) \\
&\leq C \lim_{k\to\infty} \sum_{t=k}^\infty \left(2^{t+1}\right)^2 \left(2^{t(1/2-\phi)}\right)^{-a} \leq C \lim_{k\to\infty} \sum_{t=k}^\infty 2^{-\omega t} = 0,
\end{aligned} \tag{1.29}$$

i.e. $\tilde{X} = X$ almost surely as $N \to \infty$. Here in the above derivation, we regard $M = N/d_N$ as a function depending on N.

Using (1.25) and integration by parts, it is easy to verify that

$$\mathbb{E}\,|x_{ij}|\,1_{|x_{ij}|>N^{-\phi}} = O(N^{-2-\omega/2}), \quad \mathbb{E}\,|x_{ij}|^2\,1_{|x_{ij}|>N^{-\phi}} = O(N^{-2-\omega/2}),$$

which imply that

$$|\mathbb{E}\tilde{x}_{ij}| = O(N^{-2-\omega/2}), \quad \mathbb{E}|\tilde{x}_{ij}|^2 = N^{-1} + O(N^{-2-\omega/2}),$$

and

$$\left|\mathbb{E}\tilde{x}_{ij}^2\right| = O(N^{-2-\omega/2}), \quad \text{if } x_{ij} \text{ is complex.}$$

Moreover, we trivially have

$$\mathbb{E}|\tilde{x}_{ij}|^4 \leq \mathbb{E}|x_{ij}|^4 = O(N^{-2}).$$

Hence $\tilde{X}$ is a random matrix satisfying Assumption 1.4. Then using (1.23) and (1.28) with $a = 6$ and $\phi = \epsilon/6$, we conclude (1.26); using (1.24) and (1.29) with $\phi = (1-\epsilon)/4$ and $a = 8$, we conclude (1.27). □

REMARK 1.7. The estimates (1.26) and (1.27) improve the bounds obtained in [44], and relax the assumptions on moments and Σ as well. The convergence rates in (1.26) and (1.27) are optimal up to an N^{ϵ} factor. In fact, it was proved in [3] that for an analytic function f,

$$\sqrt{N}\int f(x)d\left(F_{Q_1,\mathbf{u}}(x)-F_{1c,\mathbf{u}}(x)\right)\to\mathcal{N}(0,\sigma_{f,\mathbf{u}}), \tag{1.30}$$

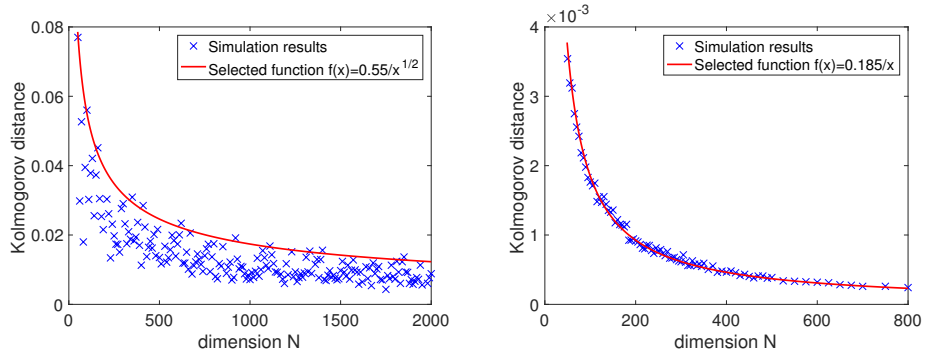
where $\mathcal{N}(0,\sigma_{f,\mathbf{u}})$ denotes the Gaussian distribution with mean zero and variance $\sigma_{f,\mathbf{u}}$. This shows that the fluctuation of $F_{Q_1,\mathbf{u}}(x)$ is of order $N^{-1/2}$ and suggests the bound in (1.27). Taking expectation of (1.30), one can see that the order of $|\mathbb{E}F_{Q_1,\mathbf{u}}(x)-F_{1c,\mathbf{u}}(x)|$ should be even smaller. Moreover, the fluctuation of eigenvalues on the microscopic scale will lead to an error of order at least N^{-1} by the universality of eigenvalues [6, 28, 36]. This shows that the bound (1.26) should be close to being optimal. We check the bounds (1.26) and (1.27) below with some numerical simulations; see Fig. 1.

REMARK 1.8. In [44], the authors can only handle the $M<N$ (i.e. $d_N>1$) case for Q_1, while our proof works for both the $d_N>1$ and $d_N<1$ cases. However, in the case with $d_N\to 1$, we will encounter some difficulties near the leftmost edge a_{2L}, which converges to 0 as $N\to\infty$ and violates the regularity condition (1.15). Also the regularity assumption (1.15) has ruled out the spiked models that have outliers. In future works, we will try to relax this assumption.

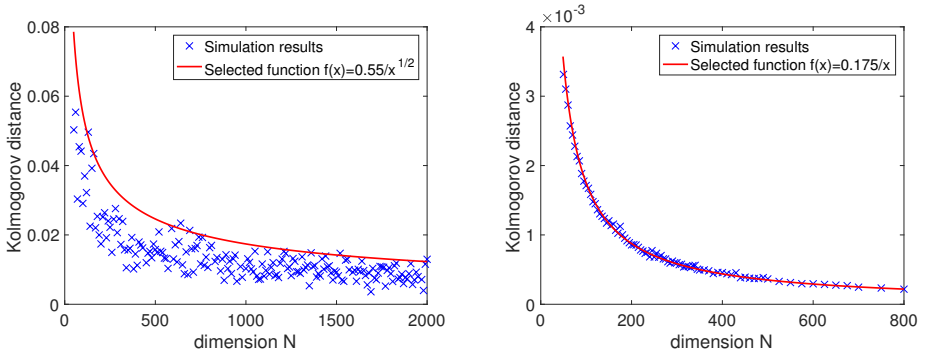
1.2. *Simulations.* Now we present some numerical simulations to illustrate the convergence of the VESD to the deformed MP law. The simulations are performed under the following setting: $M=2N$, i.e. $d_N=0.5$; the entries $\sqrt{N}x_{ij}$ are drawn from a distribution ξ with mean zero, variance 1 and tail $\mathbb{P}(|\xi|\geq s)\sim s^{-6}$ for large s; the unit vector $\mathbf{v}$ is randomly chosen for each N. In Fig. 1, we plot the Kolmogorov distances $\|F_{Q_2,\mathbf{v}}-F_{2c}\|$ and $\|\mathbb{E}F_{Q_2,\mathbf{v}}-F_{2c}\|$ for the following two choices of Σ: $\Sigma=I$ with ESD $\pi=\delta_1$, and

$$\Sigma=\operatorname{diag}(\underbrace{1,\cdots,1}_{M/2},\underbrace{4,\cdots,4}_{M/2}),\quad\text{with ESD } \pi=0.5\delta_1+0.5\delta_4. \tag{1.31}$$

For each N, we take an average over 10 repetitions to represent $F^{(N)}_{Q_2,\mathbf{v}}$ and an average over $4N^2$ repetitions to approximate $\mathbb{E}F^{(N)}_{Q_2,\mathbf{v}}$. Under each setting, we choose an appropriate function $f(x)$ to fit the simulation data. It is easy to observe that the convergence rate of the VESD is bounded by $O(N^{-1/2})$,



(a) $\pi = \delta_1$



(b) $\pi = 0.5\delta_1 + 0.5\delta_4$

FIG 1. *The left figures of (a) and (b) plot* $\|F^{(N)}_{Q_2,\mathbf{v}} - F^{(N)}_{2c}\|$ *as* N *increases from 50 to 2000, and we choose* f *to fit the upper envelope of the data. The right figures plot* $\|EF^{(N)}_{Q_2,\mathbf{v}} - F^{(N)}_{2c}\|$ *as* N *increases from 50 to 800.*

while the convergence rate of the expected VESD has order N^{-1}. This verifies the results in Corollary 1.6.

As discussed before, the convergence of $F_{Q_2,\mathbf{v}}$ to F_{2c} for any sequence of deterministic unit vectors $\mathbf{v}$ can be used to characterize the asymptotical Haar property of the eigenmatrix of $Q_2 = X^*\Sigma X$ (which also implies the asymptotical Haar property of the eigenmatrix of Q_1 when $\Sigma = \sigma^2 I$). Thus the bounds in Corollary 1.6 for the VESD of large sample covariance matrices can assist us in better studying spiked covariance matrices as assumed in [11, 30] among many others. On the other hand, for a general Σ, the eigenmatrix of Q_1 is not asymptotically Haar distributed anymore and the VESD of Q_1 depends on $\mathbf{v}$. But (1.16) gives an explicit dependence of $\mathbf{F}_{1c}$ on Σ that is of interest to statistical applications. For instance, one can use the VESD to

detect the variance structure of Σ. In Fig. 2, we plot $F_{Q_1,\mathbf{v}}$ for Σ in (1.31) and different choices of $\mathbf{v}$. One can observe a transition of $F_{Q_1,\mathbf{v}}$ when $\mathbf{v}$ changes from the direction corresponding to the smaller eigenvalues of Σ to the direction corresponding to the larger eigenvalues of Σ.

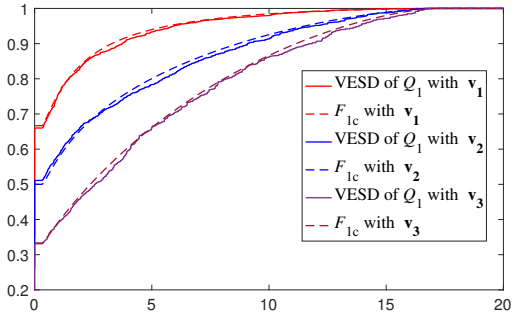


FIG 2. *The plots for $F_{Q_1,\mathbf{v}_i}$ and $F_{1c,\mathbf{v}_i}$ with $N = 2000$, Σ as in (1.31), and $\mathbf{v}_1 = \sqrt{2/M}(\underbrace{1,\cdots,1}_{M/2},0,\cdots,0)$, $\mathbf{v}_3 = \sqrt{2/M}(0,\cdots,0,\underbrace{1,\cdots,1}_{M/2})$, $\mathbf{v}_2 = (\mathbf{v}_1+\mathbf{v}_3)/\sqrt{2}$. All the other settings are the same as the ones in Fig. 1.*

The rest of this paper is organized as follows. We prove Theorem 1.5 in Section 2 using Stieltjes transforms. In the proof, we mainly use Theorems 2.4-2.6, which give the desired anisotropic local laws for the resolvents of Q_1 and Q_2. Theorem 2.5 constitutes the main novelty of this paper, and its proof will be given in Section 3. The proofs of Theorem 2.4 and Theorem 2.6 will be given in the supplementary material.

2. Proof of Theorem 1.5. For definiteness, we will focus on *real* sample covariance matrices during the proof. However, our proof also applies, after minor changes, to the *complex* case if we include the extra assumption (1.2) or (1.20).

2.1. *Anisotropic local Marčenko-Pastur law.* A basic tool for the proof is the Stieltjes transform. For any $z = E + i\eta \in \mathbb{C}_+$, we define the resolvents (the Green functions) of Q_1 and Q_2 as

$$\mathcal{G}_1(X,z) := (Q_1 - z)^{-1}, \quad \mathcal{G}_2(X,z) := (Q_2 - z)^{-1}. \tag{2.1}$$

Then the Stieltjes transforms of the ESD of $Q_{1,2}$ are equal to

$$m_1(X,z) := M^{-1}\mathrm{Tr}\,\mathcal{G}_1(X,z), \quad m_2(X,z) := N^{-1}\mathrm{Tr}\,\mathcal{G}_2(X,z),$$

and the Stieltjes transforms of $F^{(M)}_{Q_1,\mathbf{u}}$ and $F^{(N)}_{Q_2,\mathbf{v}}$ are equal to $\langle \mathbf{u}, \mathcal{G}_1(X,z)\mathbf{u}\rangle$ and $\langle \mathbf{v}, \mathcal{G}_2(X,z)\mathbf{v}\rangle$, respectively. The main goal of this subsection is to establish the following asymptotic estimate for $z \in \mathbb{C}_+$:

$$\langle \mathbf{u}, \mathcal{G}_1(X,z)\mathbf{u}\rangle \approx m_{1c,\mathbf{u}}(z), \quad \langle \mathbf{v}, \mathcal{G}_2(X,z)\mathbf{v}\rangle \approx m_{2c}(z). \tag{2.2}$$

By taking the imaginary part, it is easy to see that a control of the Stieltjes transforms $\langle \mathbf{u}, \mathcal{G}_1(X,z)\mathbf{u}\rangle$ and $\langle \mathbf{v}, \mathcal{G}_2(X,z)\mathbf{v}\rangle$ yields a control of the VESD on the scale of order $\operatorname{Im} z$ around E. An *anisotropic local law* is an estimate of the form (2.2) for all $\operatorname{Im} z \gg N^{-1}$. Such local law was first established in [26, 8, 27] for sample covariance matrices and generalized Wigner matrices, assuming that the matrix entries have arbitrarily high moments. In Section 2.2, we will finish the proof of Theorem 1.5 with the (almost) optimal anisotropic local laws for $\mathcal{G}_1$ and $\mathcal{G}_2$.

Our anisotropic local law can be stated in a simple and unified fashion using the following $(N+M)\times(N+M)$ self-adjoint matrix H:

$$H := \begin{pmatrix} 0 & \Sigma^{1/2}X \\ (\Sigma^{1/2}X)^* & 0 \end{pmatrix}. \tag{2.3}$$

We define the resolvent of H as

$$G(X,z) := \begin{pmatrix} -I_{M\times M} & \Sigma^{1/2}X \\ (\Sigma^{1/2}X)^* & -zI_{N\times N} \end{pmatrix}^{-1}, \quad z \in \mathbb{C}_+. \tag{2.4}$$

Using Schur complement formula, it is easy to check that

$$G = \begin{pmatrix} z\mathcal{G}_1 & \mathcal{G}_1(\Sigma^{1/2}X) \\ (\Sigma^{1/2}X)^*\mathcal{G}_1 & \mathcal{G}_2 \end{pmatrix} = \begin{pmatrix} z\mathcal{G}_1 & (\Sigma^{1/2}X)\mathcal{G}_2 \\ \mathcal{G}_2(\Sigma^{1/2}X)^* & \mathcal{G}_2 \end{pmatrix}. \tag{2.5}$$

Thus a control of G yields directly a control of the resolvents $\mathcal{G}_1$ and $\mathcal{G}_2$. For simplicity of notations, we define the index sets

$$\mathcal{I}_1 := \{1, ..., M\}, \quad \mathcal{I}_2 := \{M+1, ..., M+N\}, \quad \mathcal{I} := \mathcal{I}_1 \cup \mathcal{I}_2.$$

We shall consistently use the latin letters $i, j \in \mathcal{I}_1$, greek letters $\mu, \nu \in \mathcal{I}_2$, and $a, b \in \mathcal{I}$. Then we label the indices of X according to

$$X = (X_{i\mu} : i \in \mathcal{I}_1, \mu \in \mathcal{I}_2).$$

We will use the following notion of stochastic domination, which was first introduced in [16] and subsequently used in many works on random matrix theory, such as [8, 9, 27]. It simplifies the presentation of the results and their proofs by systematizing statements of the form "ξ is bounded with high probability by ζ up to a small power of N".

DEFINITION 2.1 (Stochastic domination). (i) Let

$$\xi = \left(\xi^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}\right), \quad \zeta = \left(\zeta^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}\right)$$

be two families of nonnegative random variables, where $U^{(N)}$ is a possibly N-dependent parameter set. We say ξ is stochastically dominated by ζ, uniformly in u, if for any (small) $\epsilon > 0$ and (large) $D > 0$,

$$\sup_{u \in U^{(N)}} \mathbb{P}\left[\xi^{(N)}(u) > N^{\epsilon}\zeta^{(N)}(u)\right] \leq N^{-D}$$

for large enough $N \geq N_0(\epsilon, D)$.

(ii) If ξ is stochastically dominated by ζ, uniformly in u, we use the notation $\xi \prec \zeta$. Moreover, if for some complex family ξ we have $|\xi| \prec \zeta$, we also write $\xi \prec \zeta$ or $\xi = O_{\prec}(\zeta)$.

(iii) We say that an event Ξ holds with high probability if for any constant $D > 0$, $\mathbb{P}(\Xi) \geq 1 - N^{-D}$ for large enough N.

The following lemma collects basic properties of stochastic domination, which will be used tacitly throughout the proof .

LEMMA 2.2 (Lemma 3.2 in [8]). Let ξ and ζ be families of nonnegative random variables.

(i) Suppose that $\xi(u, v) \prec \zeta(u, v)$ uniformly in $u \in U$ and $v \in V$. If $|V| \leq N^C$ for some constant C, then $\sum_{v \in V} \xi(u, v) \prec \sum_{v \in V} \zeta(u, v)$ uniformly in u.

(ii) If $\xi_1(u) \prec \zeta_1(u)$ and $\xi_2(u) \prec \zeta_2(u)$ uniformly in $u \in U$, then $\xi_1(u)\xi_2(u) \prec \zeta_1(u)\zeta_2(u)$ uniformly in $u \in U$.

(iii) Suppose that $\Psi(u) \geq N^{-C}$ is deterministic and $\xi(u)$ satisfies $\mathbb{E}\xi(u)^2 \leq N^C$ for all u. Then if $\xi(u) \prec \Psi(u)$ uniformly in u, we have $\mathbb{E}\xi(u) \prec \Psi(u)$ uniformly in u.

DEFINITION 2.3 (Bounded support condition). We say a random matrix X satisfies the *bounded support condition* with q, if

$$\max_{i \in \mathcal{I}_1, \mu \in \mathcal{I}_2} |X_{i\mu}| \prec q. \tag{2.6}$$

Here $q \equiv q(N)$ is a deterministic parameter and usually satisfies $N^{-1/2} \leq q \leq N^{-\phi}$ for some (small) constant $\phi > 0$. Whenever (2.6) holds, we say that X has support q. Moreover, if the entries of X satisfy (1.22), then X trivially satisfies the bounded support condition with $q = N^{-\phi}$.

Throughout the rest of this paper, we will consistently use the notation $E+i\eta$ for the spectral parameter z. In the following proof, we always assume that z lies in the spectral domain

$$\mathbf{D}(\omega, N) := \{z \in \mathbb{C}_+ : \omega \le E \le 2\gamma_1, N^{-1+\omega} \le \eta \le \omega^{-1}\}, \tag{2.7}$$

for some small constant $\omega > 0$, unless otherwise indicated. Recall the condition (1.15), we can take ω to be sufficiently small such that $\omega \le \gamma_K/2$. Define the distance to the spectral edges as $\kappa := \min_{1\le k\le 2L} |E - a_k|$. Then we have the following estimates for m_{2c}:

$$|m_{2c}(z)| \sim 1, \quad \operatorname{Im} m_{2c}(z) \sim \begin{cases} \eta/\sqrt{\kappa+\eta}, & \text{if } E \notin \operatorname{supp}\rho_{2c} \\ \sqrt{\kappa+\eta}, & \text{if } E \in \operatorname{supp}\rho_{2c} \end{cases}, \tag{2.8}$$

$$\max_{i\in\mathcal{I}_1} |(1 + m_{2c}(z)\sigma_i)^{-1}| = O(1). \tag{2.9}$$

for $z \in \mathbf{D}$. The reader can refer to [27, Appendix A] for the proof.

We define the deterministic limit

$$\Pi(z) := \begin{pmatrix} -(1 + m_{2c}(z)\Sigma)^{-1} & 0 \\ 0 & m_{2c}(z) I_{N\times N} \end{pmatrix}, \tag{2.10}$$

and the control parameter

$$\Psi(z) := \sqrt{\frac{\operatorname{Im} m_{2c}(z)}{N\eta}} + \frac{1}{N\eta}. \tag{2.11}$$

Note that by (2.8) and (2.9), we always have

$$\|\Pi\| = O(1), \quad \Psi \gtrsim N^{-1/2}, \quad \Psi^2 \lesssim (N\eta)^{-1}, \tag{2.12}$$

for $z \in \mathbf{D}$. Now we are ready to state the local laws for the resolvent $G(X, z)$.

THEOREM 2.4 (Local MP law). *Suppose d_N, X and Σ satisfy the Assumption 1.4. Suppose X is real and satisfies the bounded support condition (2.6) with $q \le N^{-\phi}$ for some constant $\phi > 0$. Then the following estimates hold for $z \in \mathbf{D}$:*

(1) the averaged local law:

$$|m_2(X, z) - m_{2c}(z)| + \left|M^{-1} \sum_{i\in\mathcal{I}_1} \sigma_i (G_{ii} - \Pi_{ii})\right| \prec (N\eta)^{-1}; \tag{2.13}$$

(2) the anisotropic local law: for deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}}$,

$$|\langle \mathbf{u}, G(X,z)\mathbf{v}\rangle - \langle \mathbf{u}, \Pi(z)\mathbf{v}\rangle| \prec q + \Psi(z); \tag{2.14}$$

(3) for deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}_1}$ *or* $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}_2}$,

$$|\langle \mathbf{u}, G(X,z)\mathbf{v}\rangle - \langle \mathbf{u}, \Pi(z)\mathbf{v}\rangle| \prec q^2 + (N\eta)^{-1/2}. \tag{2.15}$$

All of the above estimates are uniform in the spectral parameter z *and the deterministic vectors* $\mathbf{u}, \mathbf{v}$.

The proof for Theorem 2.5 will be given in the supplementary material. Here we make some brief comments on it.

If we assume (1.1) (instead of (1.18) and (1.19)) and $q = N^{-1/2}$, then (2.13) and (2.14) have been proved in [27]. If we have (1.1) and $q \le N^{-\phi}$, then it was proved in Lemma 3.11 and Theorem 3.14 of [14] that the averaged local law (2.13) and the entrywise local law

$$\max_{a,b\in\mathcal{I}} |G_{ab}(X,z) - \Pi_{ab}(z)| \prec q + \Psi(z) \tag{2.16}$$

hold uniformly in $z \in \mathbf{D}$. With (2.16) and the moment assumption (1.21), one can repeat the arguments in [8, Section 5] or [42, Section 5] to get the anisotropic local law (2.14). The main novelty of this theorem is the bound (2.15). In the proof, we will first establish the following version of the entrywise local law for the upper left and lower right blocks of $G(X,z)$:

$$\max_{r=1,2}\max_{a,b\in\mathcal{I}_r} |G_{ab}(X,z) - \Pi_{ab}(z)| \prec q^2 + (N\eta)^{-1/2}, \tag{2.17}$$

which can be proved with the help of (2.16). Then using (2.17) and (1.21), we will extend the arguments in [8, Section 5] to conclude the anisotropic local law (2.15). Finally, if the variance assumption in (1.1) is relaxed to the one in (1.19), we can repeat the previous arguments to get the desired estimates (2.13)-(2.15). In fact, it is easy to check that the $O(N^{-2-c_0})$ term leads to a negligible error at each step, and the whole proof remains unchanged. The relaxation of the mean zero assumption in (1.1) to the assumption (1.18) can be handled with the centralization Lemma 3.4.

After taking expectation, we have the following crucial improvement from (2.15) to (2.18), which is the main reason why we can improve the bound in [44] to the almost optimal one in (1.23). In fact, the leading order terms of $(\langle \mathbf{u}, \mathcal{G}_1\mathbf{u}\rangle - m_{1c,\mathbf{u}})$ and $(\langle \mathbf{v}, \mathcal{G}_2\mathbf{v}\rangle - m_{2c})$ vanish after taking expectation, and hence leads to a bound that is one order smaller than the one in (2.15). The proof of Theorem 2.5 will be given in Sections 3, which constitutes the main novelty of this paper.

THEOREM 2.5. *Suppose the assumptions in Theorem 2.4 hold. Then we have*

$$|\mathbb{E}\langle \mathbf{u}, G(X,z)\mathbf{v}\rangle - \langle \mathbf{u}, \Pi(z)\mathbf{v}\rangle| \prec q^4 + (N\eta)^{-1} \tag{2.18}$$

uniformly in $z \in \mathbf{D}$ *and the deterministic unit vectors* $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}_1}$ *or* $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}_2}$.

If $q = N^{-1/4}$, then (2.15) and (2.18) already give that

$$|\langle \mathbf{u}, \mathcal{G}_1 \mathbf{u}\rangle - m_{1c,\mathbf{u}}| + |\langle \mathbf{v}, \mathcal{G}_2 \mathbf{v}\rangle - m_{2c}| \prec (N\eta)^{-1/2},$$
$$|\mathbb{E}\langle \mathbf{u}, \mathcal{G}_1 \mathbf{u}\rangle - m_{1c,\mathbf{u}}| + |\mathbb{E}\langle \mathbf{v}, \mathcal{G}_2 \mathbf{v}\rangle - m_{2c}| \prec (N\eta)^{-1},$$

which are sufficient to conclude Theorem 1.5. However, we observe that the second bound on the expected VESD is still valid under a much weaker support assumption. More specifically, we have the following theorem, whose proof will be given in the supplementary material. The main strategy is a resolvent comparison method that was developed in [29].

THEOREM 2.6. *Suppose the assumptions in Theorem 2.4 hold. Then we have*

$$|\mathbb{E}\langle \mathbf{u}, G(X,z)\mathbf{v}\rangle - \langle \mathbf{u}, \Pi(z)\mathbf{v}\rangle| \prec (N\eta)^{-1}, \tag{2.19}$$

uniformly in $z \in \mathbf{D}$ *and the deterministic unit vectors* $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}_1}$ *or* $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}_2}$.

As a corollary of (2.13), we have the following rigidity result for the extreme eigenvalues λ_1 and λ_K. The reader can refer to [27, Theorem 3.12] for the proof. Recall the notations in (1.13) and (1.14).

THEOREM 2.7 (Rigidity of eigenvalues). *Suppose Theorem 2.4 and the regularity condition (1.15) hold. Then for* $\gamma_j \in [a_{2k}, a_{2k-1}]$, *we have*

$$|\lambda_j - \gamma_j| \prec [(N_{2k} + 1 - j) \wedge (j + 1 - N_{2k-1})]^{-1/3} N^{-2/3}. \tag{2.20}$$

2.2. *Convergence rate of the VESD.* In this subsection, we finish the proof of Theorem 1.5 using Theorems 2.4-2.7. The following arguments have been used previously to control the Kolmogorov distance between the ESD of a random matrix and the limiting law. For example, the reader can refer to [18, Lemma 6.1] and [36, Lemma 8.1]. By the remark below (2.7), we can choose the constant $\omega > 0$ such that $\gamma_K/2 > \omega$. Also for simplicity, we will only prove the bounds for $\|\mathbb{E}F_{Q_2,\mathbf{v}} - F_{2c}\|$ and $\|F_{Q_2,\mathbf{v}} - F_{2c}\|$. The bounds for $\|\mathbb{E}F_{Q_1,\mathbf{u}} - F_{1c,\mathbf{u}}\|$ and $\|F_{Q_1,\mathbf{u}} - F_{1c,\mathbf{u}}\|$ can be proved in the same way.

PROOF OF (1.23). The key inputs are the bounds (2.19) and (2.20). Suppose $\langle \mathbf{v}, \mathcal{G}_2(X,z)\mathbf{v}\rangle$ is the Stieltjes transform of $\hat{\rho}_{\mathbf{v}}$. Then we define

$$\hat{n}_{\mathbf{v}}(E) := \int \mathbf{1}_{[0,E]}(x)\hat{\rho}_{\mathbf{v}}dx, \quad n_c(E) := \int \mathbf{1}_{[0,E]}(x)\rho_{2c}dx, \tag{2.21}$$

and $\rho_{\mathbf{v}} := \mathbb{E}\hat{\rho}_{\mathbf{v}}$, $n_{\mathbf{v}} := \mathbb{E}\hat{n}_{\mathbf{v}}$. Hence we would like to bound

$$\|\mathbb{E}F_{Q_2,\mathbf{v}} - F_{2c}\| = \sup_E |n_{\mathbf{v}}(E) - n_c(E)|.$$

For simplicity, we denote $\Delta\rho := \rho_{\mathbf{v}} - \rho_{2c}$ and its Stieltjes transform by

$$\Delta m(z) := \mathbb{E}\langle \mathbf{v}, \mathcal{G}_2(X,z)\mathbf{v}\rangle - m_{2c}(z).$$

Let $\chi(y)$ be a smooth cutoff function with support in $[-1,1]$, with $\chi(y) = 1$ for $|y| \le 1/2$ and with bounded derivatives. Fix $\eta_0 = N^{-1+\omega}$ and $3\gamma_K/4 \le E_1 < E_2 \le 3\gamma_1/2$. Let $f \equiv f_{E_1,E_2,\eta_0}$ be a smooth function supported in $[E_1-\eta_0, E_2+\eta_0]$ such that $f(x) = 1$ if $x \in [E_1+\eta_0, E_2-\eta_0]$, and $|f'| \le C\eta_0^{-1}$, $|f''| \le C\eta_0^{-2}$ if $|x - E_i| \le \eta_0$. Using the Helffer-Sjöstrand calculus (see e.g. [12]), we have

$$f(E) = \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{iyf''(x)\chi(y) + i(f(x) + iyf'(x))\chi'(y)}{E - x - iy}dxdy.$$

Then we obtain that

$$\begin{aligned}
&\left|\int f(E)\Delta\rho(E)dE\right| \\
&\le C\int_{\mathbb{R}^2} \left(|f(x)| + |y||f'(x)|\right)|\chi'(y)||\Delta m(x+iy)|dxdy \qquad (2.22)\\
&+ C\sum_i \left|\int_{|y|\le\eta_0}\int_{|x-E_i|\le\eta_0} yf''(x)\chi(y)\mathrm{Im}\,\Delta m(x+iy)dxdy\right| \qquad (2.23)\\
&+ C\sum_i \left|\int_{|y|\ge\eta_0}\int_{|x-E_i|\le\eta_0} yf''(x)\chi(y)\mathrm{Im}\,\Delta m(x+iy)dxdy\right|. \qquad (2.24)
\end{aligned}$$

By (2.19) with $\eta = \eta_0$, we have

$$\eta_0 \mathrm{Im}\,\mathbb{E}\langle \mathbf{v}, \mathcal{G}_2(X, E+i\eta_0)\mathbf{v}\rangle \prec N^{-1+\omega}. \tag{2.25}$$

Since $\eta\mathrm{Im}\,\mathbb{E}\langle \mathbf{v}, \mathcal{G}_2(X, E+i\eta)\mathbf{v}\rangle$ and $\eta\mathrm{Im}\,m_{2c}(E+i\eta)$ are increasing with η, we obtain that

$$\eta\,|\mathrm{Im}\,\Delta m(E+i\eta)| \prec N^{-1+\omega} \quad \text{for all } 0 \le \eta \le \eta_0. \tag{2.26}$$

Moreover, since $G(X,z)^* = G(X,\bar{z})$, the estimates (2.19) and (2.26) also hold for $z \in \mathbb{C}_-$.

Now we bound the terms (2.22), (2.23) and (2.24). Using (2.19) and that the support of χ' is in $1 \geq |y| \geq 1/2$, the term (2.22) can be bounded by

$$\int_{\mathbb{R}^2} \left(|f(x)| + |y||f'(x)|\right) |\chi'(y)||\Delta m(x+iy)|dxdy \prec N^{-1}. \tag{2.27}$$

Using $|f''| \leq C\eta_0^{-2}$ and (2.26), we can bound the terms in (2.23) by

$$\left| \int_{|y|\leq\eta_0} \int_{|x-E_i|\leq\eta_0} yf''(x)\chi(y)\mathrm{Im}\,\Delta m(x+iy)dxdy \right| \prec N^{-1+\omega}. \tag{2.28}$$

Finally, we integrate the term (2.24) by parts first in x, and then in y (and use the Cauchy-Riemann equation $\partial \mathrm{Im}(\Delta m)/\partial x = -\partial \mathrm{Re}(\Delta m)/\partial y$) to get that

$$\begin{aligned}
&\int_{y\geq\eta_0} \int_{|x-E_i|\leq\eta_0} yf''(x)\chi(y)\mathrm{Im}\,\Delta m(x+iy)dxdy \\
&= \int_{y\geq\eta_0} \int_{|x-E_i|\leq\eta_0} yf'(x)\chi(y)\frac{\partial \mathrm{Re}\,\Delta m(x+iy)}{\partial y}dxdy \\
&= -\int_{|x-E_i|\leq\eta_0} \eta_0\chi(\eta_0)f'(x)\mathrm{Re}\,\Delta m(x+i\eta_0)dx \qquad (2.29) \\
&\quad -\int_{y\geq\eta_0} \int_{|x-E_i|\leq\eta_0} \left(y\chi'(y)+\chi(y)\right) f'(x)\mathrm{Re}\,\Delta m(x+iy)dxdy. \qquad (2.30)
\end{aligned}$$

We bound the term in (2.29) by $O_\prec(N^{-1})$ using (2.19) and $|f'| \leq C\eta_0^{-1}$. The first term in (2.30) can be estimated by $O_\prec(N^{-1})$ as in (2.27). For the second term in (2.30), we again use (2.19) and $|f'| \leq C\eta_0^{-1}$ to get that

$$\left| \int_{y\geq\eta_0} \int_{|x-E_i|\leq\eta_0} \chi(y)f'(x)\mathrm{Re}\,\Delta m(x+iy)dxdy \right| \prec \int_{\eta_0}^1 \frac{1}{Ny}dy \prec N^{-1}.$$

Combining the above estimates, we obtain that

$$\left| \int_{y\geq\eta_0} \int_{|x-E_i|\leq\eta_0} yf''(x)\chi(y)\mathrm{Im}\,\Delta m(x+iy)dxdy \right| \prec N^{-1}.$$

Obviously, the same estimate also holds for the $y \leq -\eta_0$ part. Together with (2.27) and (2.28), we conclude that

$$\left| \int f(E)\Delta\rho(E)dE \right| \prec N^{-1+\omega}. \tag{2.31}$$

For any interval $I := [E-\eta_0, E+\eta_0]$ with $E \in [\gamma_K/2, 2\gamma_1]$, we have

$$\begin{aligned} \hat{n}_{\mathbf{v}}(E+\eta_0) - \hat{n}_{\mathbf{v}}(E-\eta_0) &= \sum_{\lambda_k \in (E-\eta_0, E+\eta_0]} |\langle \zeta_k, \mathbf{v}\rangle|^2 \\ &\le 2\eta_0 \sum_{k=1}^{N} \frac{|\langle \zeta_k, \mathbf{v}\rangle|^2 \eta_0}{(\lambda_k - E)^2 + \eta_0^2} = 2\eta_0 \mathrm{Im}\, \langle \mathbf{v}, \mathcal{G}_2(X, E+i\eta_0)\mathbf{v}\rangle, \end{aligned} \tag{2.32}$$

where we used the spectral decomposition

$$\mathcal{G}_2(X, E+i\eta) = \sum_{k=1}^{N} \frac{\zeta_k \zeta_k^*}{\lambda_k - E - i\eta},$$

which follows from (1.3). Then by (2.25) and Lemma 2.2, we get that

$$n_{\mathbf{v}}(E+\eta_0) - n_{\mathbf{v}}(E-\eta_0) \prec N^{-1+\omega}. \tag{2.33}$$

On the other hand, since ρ_{2c} is bounded, we trivially have

$$n_c(E+\eta_0) - n_c(E-\eta_0) \le C\eta_0 = CN^{-1+\omega}. \tag{2.34}$$

Now we set $E_2 = 3\gamma_1/2$. With (2.31), (2.33) and (2.34), we get that for any $E \in [3\gamma_K/4, E_2]$,

$$|(n_{\mathbf{v}}(E_2) - n_{\mathbf{v}}(E)) - (n_c(E_2) - n_c(E))| \prec N^{-1+\omega}. \tag{2.35}$$

Note that by (2.20), the eigenvalues of Q_2 are inside $\{0\} \cup [3\gamma_K/4, E_2]$ with high probability. Hence we have that with high probability,

$$\hat{n}_{\mathbf{v}}(E_2) = n_c(E_2) = 1, \quad \hat{n}_{\mathbf{v}}(3\gamma_K/4) = \hat{n}_{\mathbf{v}}(0). \tag{2.36}$$

Together with (2.35), we get that

$$\sup_{E \ge 0} |n_{\mathbf{v}}(E) - n_c(E)| \prec N^{-1+\omega}. \tag{2.37}$$

This concludes (1.23) since ω can be arbitrarily small. □

Proof of (1.24). The proof for (1.24) is similar except that we shall use the estimate (2.15) instead of (2.19). By (2.15), we have for any $\mathbf{v} \in \mathbb{C}^{\mathcal{I}_2}$,

$$|\langle \mathbf{v}, \mathcal{G}_2(X, z)\mathbf{v}\rangle - m_{2c}(z)| \prec N^{-2\phi} + (N\eta)^{-1/2} \tag{2.38}$$

uniformly in $z \in \mathbf{D}$. Then we would like to bound (recall (2.21))

$$\|F^{(M)}_{Q_2, \mathbf{v}} - F_{2c}\| = \sup_E |\hat{n}_{\mathbf{v}}(E) - n_c(E)|,$$

where $\hat{n}_{\mathbf{v}}$ is defined in (2.21). We denote

$$\Delta\hat{\rho} := \hat{\rho}_{\mathbf{v}} - \rho_{1c}, \quad \Delta\hat{m} := \langle \mathbf{v}, \mathcal{G}_2(X,z)\mathbf{v}\rangle - m_{2c}(z).$$

Then for f_{E_1,E_2,η_0} defined above, we can repeat the Helffer-Sjöstrand argument with the estimate (2.38) to get that

$$\sup_{E_1,E_2} \left| \int f_{E_1,E_2,\eta_0}(E)\Delta\hat{\rho}(E)dE \right| \prec N^{-2\phi} + N^{-1/2}, \tag{2.39}$$

which, together with (2.32) and (2.36), implies that

$$\sup_{E\geq 0} |\hat{n}_{\mathbf{v}}(E) - n_c(E)| \prec N^{-2\phi} + N^{-1/2}.$$

This concludes (1.24) by the Definition 2.1. □

3. Proof of Theorem 2.5.

3.1. *Resolvent estimates.* In this subsection, we collect some useful identities from linear algebra and some simple resolvent estimates. For simplicity, we denote $Y := \Sigma^{1/2}X$.

Definition 3.1 (Minors). For $\mathbb{T} \subseteq \mathcal{I}$, we define the minor $H^{(\mathbb{T})} := (H_{ab} : a,b \in \mathcal{I}\setminus\mathbb{T})$ obtained by removing all rows and columns of H indexed by $a,b \in \mathbb{T}$. Note that we keep the names of indices when defining $H^{(\mathbb{T})}$, i.e. $(H^{(\mathbb{T})})_{ab} = \mathbf{1}_{\{a,b\notin\mathbb{T}\}}H_{ab}$. Correspondingly, we define the Green function

$$G^{(\mathbb{T})} := (H^{(\mathbb{T})})^{-1} = \begin{pmatrix} z\mathcal{G}_1^{(\mathbb{T})} & \mathcal{G}_1^{(\mathbb{T})}Y^{(\mathbb{T})} \\ (Y^{(\mathbb{T})})^*\mathcal{G}_1^{(\mathbb{T})} & \mathcal{G}_2^{(\mathbb{T})} \end{pmatrix} = \begin{pmatrix} z\mathcal{G}_1^{(\mathbb{T})} & Y^{(\mathbb{T})}\mathcal{G}_2^{(\mathbb{T})} \\ \mathcal{G}_2^{(\mathbb{T})}(Y^{(\mathbb{T})})^* & \mathcal{G}_2^{(\mathbb{T})} \end{pmatrix},$$

and the partial traces

$$m_1^{(\mathbb{T})} := \frac{1}{M}\operatorname{Tr}\mathcal{G}_1^{(\mathbb{T})} = \frac{1}{Mz}\sum_{i\in\mathcal{I}_1} G_{ii}^{(\mathbb{T})}, \quad m_2^{(\mathbb{T})} := \frac{1}{N}\operatorname{Tr}\mathcal{G}_2^{(\mathbb{T})} = \frac{1}{N}\sum_{\mu\in\mathcal{I}_2} G_{\mu\mu}^{(\mathbb{T})},$$

where we adopt the convention that $G_{ab}^{(\mathbb{T})} = 0$ if $a \in \mathbb{T}$ or $b \in \mathbb{T}$. For simplicity, we will abbreviate $(\{a\}) \equiv (a)$ and $(\{a,b\}) \equiv (ab)$.

Lemma 3.2 (Resolvent identities). (i) For $i \in \mathcal{I}_1$ and $\mu \in \mathcal{I}_2$, we have

$$\frac{1}{G_{ii}} = -1 - \left(YG^{(i)}Y^*\right)_{ii}, \quad \frac{1}{G_{\mu\mu}} = -z - \left(Y^*G^{(\mu)}Y\right)_{\mu\mu}. \tag{3.1}$$

(ii) For $i \neq j \in \mathcal{I}_1$ and $\mu \neq \nu \in \mathcal{I}_2$, we have

$$G_{ij} = G_{ii} G^{(i)}_{jj} \left(Y G^{(ij)} Y^* \right)_{ij}, \tag{3.2}$$

$$G_{\mu\nu} = G_{\mu\mu} G^{(\mu)}_{\nu\nu} \left(Y^* G^{(\mu\nu)} Y \right)_{\mu\nu}. \tag{3.3}$$

(iii) For $a \in \mathcal{I}$ and $b, c \in \mathcal{I} \setminus \{a\}$,

$$G_{bc} = G^{(a)}_{bc} + \frac{G_{ba} G_{ac}}{G_{aa}}, \quad \frac{1}{G_{bb}} = \frac{1}{G^{(a)}_{bb}} - \frac{G_{ba} G_{ab}}{G_{bb} G^{(a)}_{bb} G_{aa}}. \tag{3.4}$$

(iv) All of the above identities hold for $G^{(\mathbb{T})}$ instead of G for $\mathbb{T} \subset \mathcal{I}$.

PROOF. These identities can be proved using Schur complement formula. The reader can refer to e.g. [8, Lemmas 3.6 and 3.8] or [27, Lemma 4.4]. □

LEMMA 3.3. Suppose $\tilde{\Phi}(z)$ is a deterministic function on $\mathbf{D}$ satisfying $N^{-1/2} \leq \tilde{\Phi}(z) \leq N^{-c}$ for some constant $c > 0$. Suppose $|G_{ab}(z) - \Pi_{ab}(z)| \prec \tilde{\Phi}(z)$ uniformly in $a, b \in \mathcal{I}$ and $z \in \mathbf{D}$. Then for any $\mathbb{T} \subseteq \mathcal{I}$ with $|\mathbb{T}| = O(1)$, we have

$$\max_{a,b \in \mathcal{I} \setminus \mathbb{T}} \left| G_{ab}(z) - G^{(\mathbb{T})}_{ab}(z) \right| \prec \tilde{\Phi}^2(z), \tag{3.5}$$

uniformly in $z \in \mathbf{D}$.

PROOF. The bound (3.5) can be proved by repeatedly applying the first resolvent expansion in (3.4) with respect to the indices in $\mathbb{T}$ and using the entrywise local law. □

For X satisfying the assumptions in Theorem 2.4, we write $X = X_1 + B$, where $X_1 := X - \mathbb{E}X$ is a real random matrix satisfying (1.19), (1.21) and

$$\mathbb{E}(X_1)_{i\mu} = 0, \quad i \in \mathcal{I}_1, \ \mu \in \mathcal{I}_2, \tag{3.6}$$

and $B := \mathbb{E}X$ is a deterministic matrix such that

$$\max_{i,\mu} |B_{i\mu}| \leq C_0 N^{-2-c_0}. \tag{3.7}$$

The next lemma shows that $G(X, z)$ is very close to $G(X_1, z)$ in the sense of anisotropic local law. Its proof will be given in the supplementary material.

LEMMA 3.4. If (2.14) holds for $G(X_1, z)$, then we have

$$|\langle \mathbf{u}, G(X, z) \mathbf{v} \rangle - \langle \mathbf{u}, G(X_1, z) \mathbf{v} \rangle| \prec (N\eta)^{-1} \tag{3.8}$$

uniformly in $z \in \mathbf{D}$ and deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}}$.

3.2. *Sketch of the proof for Theorem 2.5.* In this subsection, we start proving our main resolvent estimate (2.18). For simplicity, we denote $\Phi := q^2 + (N\eta)^{-1/2}$. By Lemma 3.4, we can assume that the entries of X are centered without loss of generality. We will only prove (2.18) for $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}_2}$, while the proof in the case of $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}_1}$ is exactly the same. Also by polarization, it suffices to prove the following estimate

$$|\mathbb{E}\langle \mathbf{v}, \mathcal{G}_2(X,z)\mathbf{v}\rangle - m_{2c}(z)| \prec q^4 + (N\eta)^{-1}, \quad \mathbf{v} \in \mathbb{C}^{\mathcal{I}_2}. \tag{3.9}$$

In fact, we can obtain the more general bound (2.18) by applying (3.9) to the vectors $\mathbf{u}+\mathbf{v}$ and $\mathbf{u}+i\mathbf{v}$, respectively. Note that (2.15) gives the a priori bound

$$\left| \sum_{\mu,\nu} \bar{v}_\mu v_\nu \mathbb{E}\left(\mathcal{G}_2\right)_{\mu\nu} - m_{2c} \right| \prec \Phi.$$

We will show that after taking expectation, the leading order term in $(\mathcal{G}_2)_{\mu\nu} - m_{2c}\delta_{\mu\nu}$ vanishes and leads to the better estimate (3.9). We deal with the diagonal and off-diagonal parts separately:

$$\sum_{\mu} |v_\mu|^2 \left[\mathbb{E}(\mathcal{G}_2)_{\mu\mu} - m_{2c}\right], \quad \sum_{\mu\neq\nu} \bar{v}_\mu v_\nu \mathbb{E}\left(\mathcal{G}_2\right)_{\mu\nu}.$$

For any $\mathbb{T} \subseteq \mathcal{I}$, we define the Z variables

$$Z_\mu^{(\mathbb{T})} := (1-\mathbb{E}_\mu)(G^{(\mathbb{T})})^{-1}_{\mu\mu} = \frac{1}{N}\sum_{i\in\mathcal{I}_1} \sigma_i G_{ii}^{(\mathbb{T}\mu)} - (Y^* G^{(\mathbb{T}\mu)} Y)_{\mu\mu}, \quad \mu \notin \mathbb{T}, \tag{3.10}$$

where $\mathbb{E}_\mu[\cdot] := \mathbb{E}[\cdot|H^{(\mu)}]$, i.e. it is the partial expectation in the randomness of the μ-th row and column of H, and we used (3.1) in the second step. If $\mathbb{T} = \emptyset$, we shall abbreviate $Z_i \equiv Z_i^{(\emptyset)}$. Note that by (2.17), (3.5) (with $\tilde{\Phi} = q + \Psi$ by (2.14)), and Lemma 2.2, we have

$$Z_\mu^{(\mathbb{T})} := (1-\mathbb{E}_\mu)\left[(G^{(\mathbb{T})})^{-1}_{\mu\mu} - m_{2c}^{-1}\right] \prec \Phi, \tag{3.11}$$

for any $\mathbb{T} \subseteq \mathcal{I}$ with $|\mathbb{T}| = O(1)$. Then using (3.1) we get that

$$\begin{aligned}
\mathbb{E}G_{\mu\mu} - m_{2c} &= \mathbb{E}\frac{1}{-z - N^{-1}\sum_i \sigma_i \Pi_{ii} - N^{-1}\sum_i \sigma_i (G_{ii}^{(\mu)} - \Pi_{ii}) + Z_\mu} - m_{2c} \\
&= -m_{2c}^2 \mathbb{E}Z_\mu + O_\prec\left(\Phi^2 + \frac{1}{N\eta}\right) = O_\prec\left(\Phi^2\right),
\end{aligned}$$

where in the second step we used (2.13), (3.5), (3.11), and

$$-z - N^{-1}\sum_i \sigma_i \Pi_{ii} = m_{2c}^{-1}, \tag{3.12}$$

which follows from (2.10) and (1.9). So we can bound the diagonal part by

$$\sum_{\mu} |v_\mu|^2 \left[\mathbb{E}(\mathcal{G}_2)_{\mu\mu} - m_{2c}(z)\right] = \sum_{\mu} |v_\mu|^2 [\mathbb{E}G_{\mu\mu} - m_{2c}(z)] \prec q^4 + \frac{1}{N\eta}. \tag{3.13}$$

For the off-diagonal part, we claim that for $\mu \neq \nu \in \mathcal{I}_2$,

$$\left|\mathbb{E}\left(\mathcal{G}_2\right)_{\mu\nu}\right| \prec N^{-1}\Phi^2. \tag{3.14}$$

Then using (3.14) and $\|\mathbf{v}\|_1 \leq \sqrt{N}$, we obtain that

$$\left|\sum_{\mu\neq\nu} \bar{v}_\mu v_\nu \mathbb{E}\left(\mathcal{G}_2\right)_{\mu\nu}\right| \prec \|\mathbf{v}\|_1^2 N^{-1}\Phi^2 \leq C\left(q^4 + \frac{1}{N\eta}\right).$$

This concludes (3.9) together with (3.13).

To prove (3.14), we extend the arguments in [8, Section 5] and [42, Section 5]. We illustrate the basic idea with some simplified calculations. Using the resolvent identities (3.3) and (3.4), we get

$$\begin{aligned} \mathbb{E}G_{\mu\nu} &= \mathbb{E}G_{\mu\mu}G^{(\mu)}_{\nu\nu}\left(Y^*G^{(\mu\nu)}Y\right)_{\mu\nu} \\ &= \mathbb{E}G^{(\nu)}_{\mu\mu}G^{(\mu)}_{\nu\nu}\left(Y^*G^{(\mu\nu)}Y\right)_{\mu\nu} + \mathbb{E}\frac{G_{\mu\nu}G_{\nu\mu}}{G_{\nu\nu}}G^{(\mu)}_{\nu\nu}\left(Y^*G^{(\mu\nu)}Y\right)_{\mu\nu}. \end{aligned} \tag{3.15}$$

We now focus on the first term. Applying (3.1) gives that

$$\begin{aligned} \mathbb{E}G^{(\nu)}_{\mu\mu}G^{(\mu)}_{\nu\nu}\left(Y^*G^{(\mu\nu)}Y\right)_{\mu\nu} &= \mathbb{E}\frac{\left(Y^*G^{(\mu\nu)}Y\right)_{\mu\nu}}{\left[-z - (Y^*G^{(\mu\nu)}Y)_{\mu\mu}\right]\left[-z - (XG^{(\mu\nu)}X^*)_{\nu\nu}\right]} \\ &= \mathbb{E}\frac{\left(Y^*G^{(\mu\nu)}Y\right)_{\mu\nu}}{\left(m_{2c}^{-1} + \epsilon_\mu\right)\left(m_{2c}^{-1} + \epsilon_\nu\right)}. \end{aligned} \tag{3.16}$$

where we have

$$\epsilon_\mu := \frac{1}{N}\sum_{i\in\mathcal{I}_1}\sigma_i\Pi_{ii} - (Y^*G^{(\mu\nu)}Y)_{\mu\mu} = \frac{1}{N}\sum_{i\in\mathcal{I}_1}\sigma_i(\Pi_{ii} - G^{(\mu\nu)}_{ii}) + Z^{(\nu)}_\mu \prec \Phi \tag{3.17}$$

by (3.12), (2.13), (3.5) (with $\tilde{\Phi} = q + \Psi$) and (3.11). We now expand the fractions in (3.16) in order to take the expectation. Note that the $G^{(\mu\nu)}$ entries are independent of the X entries in the μ,ν-th rows and columns. Thus to attain a nonzero expectation, each X entry must appear at least

twice in the expression. Due to this reason, the leading and next-to-leading order terms in the expansion vanish. The "real" leading order term is

$$\begin{aligned}\mathbb{E}m_{2c}^4\epsilon_\mu\epsilon_\nu\left(Y^*G^{(\mu\nu)}Y\right)_{\mu\nu} &= m_{2c}^4\mathbb{E}(Y^*G^{(\mu\nu)}Y)_{\mu\mu}(Y^*G^{(\mu\nu)}Y)_{\nu\nu}(Y^*G^{(\mu\nu)}Y)_{\mu\nu}\\ &= m_{2c}^4\sum_{\mu,\nu}\frac{C_{i,j}}{N^3}\mathbb{E}G_{ii}^{(\mu\nu)}G_{jj}^{(\mu\nu)}G_{ij}^{(\mu\nu)}\\ &= m_{2c}^4\sum_{i\neq j}\frac{C_{i,j}}{N^3}\Pi_{ii}\Pi_{jj}\mathbb{E}G_{ij}^{(\mu\nu)}+O_\prec(N^{-1}\Phi^2),\end{aligned}\tag{3.18}$$

where the constants $C_{i,j}$ depend on σ_i, σ_j and the 3rd moments of $X_{i\mu}$ and $X_{j\mu}$ (recall (1.21)). Here in the last step, we used $|G_{ii}^{(\mu\nu)}-\Pi_{ii}|\prec\Phi$ (by (2.15) and (3.5)) and $|\Pi_{ii}|=O(1)$ (by (2.9)), and bounded the $i=j$ terms by $O_\prec(N^{-2})=O_\prec(N^{-1}\Phi^2)$. Now applying (3.2) to $G_{ij}^{(\mu\nu)}$, we get that

$$\begin{aligned}\mathbb{E}G_{ij}^{(\mu\nu)} &= \mathbb{E}G_{ii}^{(\mu\nu)}G_{jj}^{(i\mu\nu)}\left(YG^{(ij\mu\nu)}Y^*\right)_{ij}\\ &= \Pi_{ii}\Pi_{jj}\mathbb{E}\left(YG^{(ij\mu\nu)}Y^*\right)_{ij}+O_\prec(\Phi^2)=O_\prec(\Phi^2),\end{aligned}\tag{3.19}$$

where in the second step we used $|G_{ii}^{(\mu\nu)}-\Pi_{ii}|+|G_{jj}^{(i\mu\nu)}-\Pi_{jj}|\prec\Phi$ and

$$\left(YG^{(ij\mu\nu)}Y^*\right)_{ij}=G_{ij}^{(\mu\nu)}\left(G_{ii}^{(\mu\nu)}G_{jj}^{(i\mu\nu)}\right)^{-1}\prec\Phi,$$

which follow easily from (2.15) and (3.5), and in the last step the leading order term vanishes since the two X entries are independent for $i\neq j$. Then with (3.19), the terms in (3.18) can be bounded by $O_\prec(N^{-1}\Phi^2)$.

In general, after the expansion of the two fractions in (3.16), we get a summation of terms of the form

$$A_{m,n}:=\mathbb{E}\epsilon_\mu^m\epsilon_\nu^n(Y^*G^{(\mu\nu)}Y)_{\mu\nu},\quad \mu\neq\nu,$$

up to some deterministic coefficients of order $O(1)$. Since $|\epsilon_{\mu,\nu}|\prec\Phi\lesssim N^{-\omega/2}$ for $z\in\mathbf{D}$ (we can take ω small enough such that $N^{-\omega/2}\geq q^2$), we only need to include the terms with $m+n\leq 2+2/\omega$ and the tail terms will be smaller than $N^{-1}\Phi^2$. Note that in $A_{m,n}$, the $X_{*\mu}$ entries, $X_{*\nu}$ entries and $G^{(\mu\nu)}$ entries are mutually independent. Moreover, both the number of $X_{*\mu}$ entries and the number of $X_{*\nu}$ entries are odd. Thus to attain a nonzero expectation, we must pair the X entries such that there are products of the forms $X_{i\mu}^{n_1}$ and $X_{j\nu}^{n_2}$ for some $n_1,n_2\geq 3$. As a result, we lose $(n_1-2)/2+(n_2-2)/2\geq 1$ free indices, and this contributes an N^{-1} factor. On the other hand, for the

product of G entries, we have the following three cases: (1) if there are at least 2 off-diagonal G entries, then we bound them with $O_{\prec}(\Phi^2)$; (2) if there is only 1 off-diagonal G entry, then we can use the trick in (3.18) and the bound (3.19); (3) if there is no off-diagonal G entry, then we lose one more free index and get an extra N^{-1} factor. This leads to the estimate (3.14) for the term in (3.16).

For the second term in (3.15), we again use Lemma 3.2 to expand the $G_{\mu\nu}$, $G_{\nu\mu}$ and $G^{-1}_{\nu\nu}$ entries. Our goal is to expand all the G entries into polynomials of the random variables

$$S_{\alpha\beta} := (Y^* G^{(\mu\nu)} Y)_{\alpha\beta}, \quad \alpha, \beta \in \{\mu, \nu\}, \tag{3.20}$$

so that the X entries and $G^{(\mu\nu)}$ entries are independent in the resulting expression. In particular, the *maximally expanded* terms (see (3.21)) can be expanded into $S_{\alpha\beta}$ variables directly through (3.1) and (3.3). However, *non-maximally expanded* terms are also created along the expansions in (3.3) and (3.4). Then we need to further expand these newly appeared terms. In general, this process will not terminate. However, we will show in Lemma 3.8 that after sufficiently many expansions, the resulting expression either has enough off-diagonal terms, or is maximally expanded. In the former case, it suffices to bound each off-diagonal term by $O_{\prec}(\Phi)$. In the latter case, the expression will only consist of $S_{\alpha\beta}$ variables. Following the argument in the previous paragraph, the expectation over the X entries produces an N^{-1} factor, while the expectation over the G entries produces a Φ^2 factor.

In the next two subsections, we give a rigorous proof based on the above arguments.

3.3. *Resolvent expansion.* To perform the resolvent expansion in a systematic way, we introduce the following notions of *string* and *string operator*.

Definition 3.5 (Strings). Let $\mathfrak{A}$ be the alphabet containing all symbols that will appear during the expansion:

$$\mathfrak{A} = \left\{ G_{\alpha\beta}, G^{-1}_{\alpha\alpha}, S_{\alpha\beta} \text{ with } \alpha, \beta \in \{\mu, \nu\} \right\} \cup \left\{ G^{(\nu)}_{\mu\mu}, G^{(\mu)}_{\nu\nu}, (G^{(\nu)}_{\mu\mu})^{-1}, (G^{(\mu)}_{\nu\nu})^{-1} \right\}.$$

We define a string $\mathbf{s}$ to be a concatenation of the symbols from $\mathfrak{A}$, and we use $[\![\mathbf{s}]\!]$ to denote the random variable represented by $\mathbf{s}$. We denote an empty string by $\emptyset$ with value $[\![\emptyset]\!] = 0$.

Remark 3.6. It is important to distinguish a string $\mathbf{s}$ from its value $[\![\mathbf{s}]\!]$. For example, "$G_{\mu\nu}$" and "$G_{\mu\mu} G^{(\mu)}_{\nu\nu} S_{\mu\nu}$" are different strings, but they represent the same random variable by (3.3).

We shall call the following symbols the *maximally expanded* symbols:

(3.21)
$$\mathfrak{A}_{\max} = \left\{ G_{\mu\nu}, G_{\nu\mu}, G^{(\nu)}_{\mu\mu}, G^{(\mu)}_{\nu\nu}, (G^{(\nu)}_{\mu\mu})^{-1}, (G^{(\mu)}_{\nu\nu})^{-1}, S_{\mu\mu}, S_{\nu\nu}, S_{\mu\nu}, S_{\nu\mu} \right\}.$$

A string $\mathbf{s}$ is said to be maximally expanded if all of its symbols are in $\mathfrak{A}_{\max}$. We shall call $G_{\mu\nu}, G_{\nu\mu}, S_{\mu\nu}, S_{\nu\mu}$ the *off-diagonal* symbols and all the other symbols *diagonal*. By (2.17) and (3.5), we have $[\![\mathbf{a}_o]\!] \prec \Phi$ if $\mathbf{a}_o$ is off-diagonal (we have $S_{\mu\nu} \prec \Phi$ using (3.3)) and $[\![\mathbf{a}_d]\!] \prec 1$ if $\mathbf{a}_d$ is diagonal. We use $\mathcal{F}_{n\text{-}max}(\mathbf{s})$ and $\mathcal{F}_{\text{off}}(\mathbf{s})$ to denote the number of non-maximally expanded symbols and the number of off-diagonal symbols, respectively, in $\mathbf{s}$.

Definition 3.7 (String operators). Let $\alpha \neq \beta \in \{\mu, \nu\}$.

(i) We define an operator τ_0 acting on a string $\mathbf{s}$ in the following sense. Find the first $G_{\alpha\alpha}$ or $G^{-1}_{\alpha\alpha}$ in $\mathbf{s}$. If $G_{\alpha\alpha}$ is found, replace it with $G^{(\beta)}_{\alpha\alpha}$; if $G^{-1}_{\alpha\alpha}$ is found, replace it with $(G^{(\beta)}_{\alpha\alpha})^{-1}$; if neither is found, set $\tau_0(\mathbf{s}) = \mathbf{s}$ and we say that τ_0 is trivial for $\mathbf{s}$.

(ii) We define an operator τ_1 acting on a string $\mathbf{s}$ in the following sense. Find the first $G_{\alpha\alpha}$ or $G^{-1}_{\alpha\alpha}$ in $\mathbf{s}$. If $G_{\alpha\alpha}$ is found, replace it with $\frac{G_{\alpha\beta}G_{\beta\alpha}}{G_{\beta\beta}}$; if $G^{-1}_{\alpha\alpha}$ is found, replace it with $-\frac{G_{\alpha\beta}G_{\beta\alpha}}{G_{\alpha\alpha}G^{(\beta)}_{\alpha\alpha}G_{\beta\beta}}$; if neither is found, set $\tau_1(\mathbf{s}) = \emptyset$ and we say that τ_1 is null for $\mathbf{s}$.

(iii) Define an operator ρ acting on a string $\mathbf{s}$ in the following sense. Replace each $G_{\alpha\beta}$ in $\mathbf{s}$ with $G_{\alpha\alpha}G^{(\alpha)}_{\beta\beta}S_{\alpha\beta}$.

By Lemma 3.2, it is clear that for any string $\mathbf{s}$,

(3.22)
$$[\![\tau_0(\mathbf{s})]\!] + [\![\tau_1(\mathbf{s})]\!] = [\![\mathbf{s}]\!], \quad [\![\rho(\mathbf{s})]\!] = [\![\mathbf{s}]\!].$$

Moreover, a string $\mathbf{s}$ is trivial under τ_0 and null under τ_1 if and only if $\mathbf{s}$ is maximally expanded. Given a string $\mathbf{s}$, we abbreviate $\mathbf{s}_0 := \tau_0(\mathbf{s})$ and $\mathbf{s}_1 := \rho(\tau_1(\mathbf{s}))$. For any sequence $w = a_1 a_2 \dots a_m$ with $a_i \in \{0, 1\}$, we denote

$$\mathbf{s}_w := \rho^{a_m}\tau_{a_m} \dots \rho^{a_2}\tau_{a_2}\rho^{a_1}\tau_{a_1}(\mathbf{s}), \quad \text{where } \rho^0 := 1.$$

Then by (3.22) we have

(3.23)
$$\sum_{|w|=m} [\![\mathbf{s}_w]\!] = [\![\mathbf{s}]\!],$$

where the summation is over all binary sequences w with length $|w| = m$.

LEMMA 3.8. Consider the string $\mathbf{s} = \text{“}G_{\mu\mu}G^{(\mu)}_{\nu\nu}S_{\mu\nu}\text{”}$. Let w be any binary sequence with $|w| = 4l_0$ and such that $\mathbf{s}_w \neq \emptyset$. Then either $\mathcal{F}_{\text{off}}(\mathbf{s}_w) \geq 2l_0$ or $\mathbf{s}_w$ is maximally expanded.

PROOF. It suffices to show that any nonempty string $\mathbf{s}_w$ with $\mathcal{F}_{\text{off}}(\mathbf{s}_w) < 2l_0$ is maximally expanded.

By Definition 3.7, a nontrivial τ_0 reduces the number of non-maximally expanded symbols by 1, and keeps the number of off-diagonal symbols the same; a $\rho\tau_1$ increases the number of non-maximally expanded symbols by 2 or 3, and increases the number of off-diagonal symbols by 2. Hence $\mathcal{F}_{\text{off}}(\mathbf{s}_w) < 2l_0$ implies that there are at most $(l_0 - 1)$ 1's in w. Those $\rho\tau_1$ operators increase $\mathcal{F}_{n\text{-}max}$ at most by $3(l_0 - 1)$ in total. On the other hand, there are at least $3l_0$ 0's in w, which is sufficient to eliminate all the non-maximally expanded symbols, whose number is at most $3(l_0-1)+1 = 3l_0-2$ in total (note that $\mathcal{F}_{n\text{-}max}(\mathbf{s}) = 1$ for the initial string). □

Now we choose $l_0 = 1 + 1/\omega$. Then using $\Phi = O(N^{-\omega/2})$, we have

$$\sum_{|w|=4l_0} [\![\mathbf{s}_w]\!] \cdot \mathbf{1}(\mathcal{F}_{\text{off}}(\mathbf{s}_w) \geq 2l_0) \prec 2^{4l_0}\Phi^{2l_0} \prec N^{-1}\Phi^2.$$

By Lemma 3.8, we see that to prove (3.14), it suffices to show that

$$|\mathbb{E}[\![\mathbf{s}_w]\!]| \prec N^{-1}\Phi^2 \tag{3.24}$$

for any maximally expanded string $\mathbf{s}_w$ with $|w| = 4l_0$.

Note that the maximally expanded string $\mathbf{s}_w$ thus obtained consists only of the symbols

$$G^{(\beta)}_{\alpha\alpha},\ (G^{(\beta)}_{\alpha\alpha})^{-1},\ S_{\alpha\beta}, \quad \text{with } \alpha \neq \beta \in \{\mu, \nu\}.$$

By (3.1), we can replace $(G^{(\beta)}_{\alpha\alpha})^{-1}$ with

$$(G^{(\beta)}_{\alpha\alpha})^{-1} = -z - S_{\alpha\alpha}. \tag{3.25}$$

Note that $|S_{\alpha\alpha} - N^{-1}\sum_i \sigma_i\Pi_{ii}| \prec \Phi$ by (3.17). Then we can expand $G^{(\beta)}_{\alpha\alpha}$ as

$$\begin{aligned} G^{(\beta)}_{\alpha\alpha} &= \frac{1}{m_{2c}^{-1} + (N^{-1}\sum_i \sigma_i\Pi_{ii} - S_{\alpha\alpha})} \\ &= m_{2c}\sum_{k=0}^{2l_0} m_{2c}^k \left(S_{\alpha\alpha} - N^{-1}\sum_i \sigma_i\Pi_{ii}\right)^k + O_{\prec}(N^{-1}\Phi^2). \end{aligned} \tag{3.26}$$

We apply the expansions (3.25) and (3.26) to the G symbols in $\mathbf{s}_w$, disregard the sufficiently small tails, and denote the resulting polynomial (in terms of the symbols $S_{\alpha\beta}$) by P_w. Then P_w can be written as a finite sum of maximally expanded strings (or monomials) consisting of the $S_{\alpha\beta}$ symbols. Moreover, the number of such monomials depends only on l_0. Hence we only need to prove that for any such monomial M_w,

$$|\mathbb{E}[\![M_w]\!]| \prec N^{-1}\Phi^2. \tag{3.27}$$

Let N_μ (N_ν) be the number of times that μ (ν) appears as a lower index of the S symbols in M_w. We have $N_\mu = N_\nu = 3$ for the initial string $\mathbf{s} =$ "$G_{\mu\mu}G^{(\mu)}_{\nu\nu}S_{\mu\nu}$". From Definition 3.7, it is easy to see that the operators τ_0, τ_1 and ρ do not change the parity of N_μ and N_ν. The expansions (3.25) and (3.26) also do not change the parity of N_μ and N_ν. This leads to the following key observation:

$$\textit{both } N_\mu \textit{ and } N_\nu \textit{ are odd in } M_w. \tag{3.28}$$

3.4. *A graphical proof.* In this subsection, we finish the proof of (3.27). Suppose $M_w = C(z)(S_{\mu\mu})^{m_1}(S_{\nu\nu})^{m_2}(S_{\mu\nu})^{m_3}(S_{\nu\mu})^{m_4}$, where $C(z)$ denotes a deterministic function of order 1 for all $z \in \mathbf{D}$. Then we write

$$\begin{aligned}[\![M_w]\!] \sim \sum_{i_*^{(*)}, j_*^{(*)} \in \mathcal{I}_1} &\prod_{a=1}^{m_1} X_{i_a^{(1)}\mu} G^{(\mu\nu)}_{i_a^{(1)} j_a^{(1)}} X_{j_a^{(1)}\mu} \prod_{b=1}^{m_2} X_{i_b^{(2)}\nu} G^{(\mu\nu)}_{i_b^{(2)} j_b^{(2)}} X_{j_b^{(2)}\nu} \\ &\prod_{c=1}^{m_3} X_{i_c^{(3)}\mu} G^{(\mu\nu)}_{i_c^{(3)} j_c^{(3)}} X_{j_c^{(3)}\nu} \prod_{d=1}^{m_4} X_{i_d^{(4)}\nu} G^{(\mu\nu)}_{i_d^{(4)} j_d^{(4)}} X_{j_d^{(4)}\mu}.\end{aligned} \tag{3.29}$$

To avoid heavy expressions, we introduce the following graphical notations. We use a connected graph (V, E) to represent the string M_w, where the vertex set V consists of the indices in (3.29) and the edge set E consists of the X and G variables. The indices μ, ν are represented by the black vertices in the graph, while the i, j indices are represented by the white vertices. The X edges are represented by the zig-zag lines and the G edges are represented by the straight lines. One can refer to Fig. 3 for an example of such a graph.

We organize the summation in (3.29) in the following way. We first partition the white vertices into blocks by requiring that any pair of white vertices take the same value if they are in the same block, and take different values otherwise. Then we take the summation over the white blocks which take values in $\mathcal{I}_2$. Finally, we sum over all possible partitions. Note that the number of different partitions depends only on the total number of S variables in M_w, which in turn depends only on l_0.

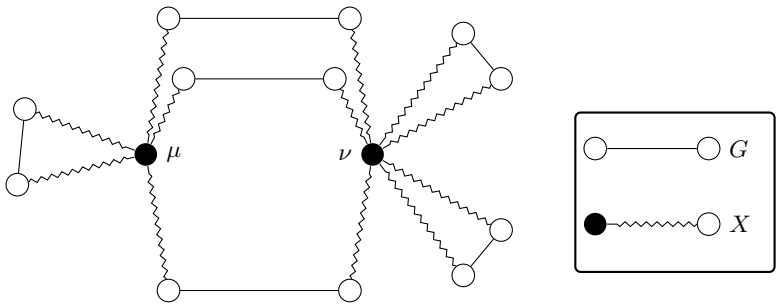


FIG 3. *The graph representing* $S_{\mu\mu}(S_{\mu\nu})^3(S_{\nu\nu})^2$.

Fix a partition Γ of the white vertices. We denote its blocks by $b_1, ..., b_k$, where k gives the number of distinct blocks in Γ. We denote by n_l^μ (n_l^ν) the number of white vertices in b_l that are connected to the vertex μ (ν). Let $G(\Gamma)$ be the product of all the G edges in the graph. Then we have

$$[\![M_w]\!] \sim \sum_\Gamma \sum_{b_1,...,b_k}^* G(\Gamma) \prod_{l=1}^k (X_{b_l\mu})^{n_l^\mu} (X_{b_l\nu})^{n_l^\nu}, \tag{3.30}$$

where $\sum^*$ denotes the summation subject to the condition that $b_1, ..., b_k$ all take distinct values. Note that k, b_l, n_l^μ and n_l^ν all depend on Γ, and we have omitted the Γ dependence for simplicity of notations.

From (3.29), it is easy to observe that the X edges are independent of $G(\Gamma)$. Thus taking expectation of (3.30) gives that

$$\begin{aligned} |\mathbb{E}[\![M_w]\!]| \le & C \sum_\Gamma \sum_{b_1,...,b_k}^* |\mathbb{E}G(\Gamma)| \prod_{l=1}^k |\mathbb{E}(X_{b_l\mu})^{n_l^\mu}| |\mathbb{E}(X_{b_l\nu})^{n_l^\nu}| \\ \le & C \sum_\Gamma \sum_{b_1,...,b_k}^* |\mathbb{E}G(\Gamma)| \prod_{l=1}^k \mathbb{E}|X_{b_l\mu}|^{n_l^\mu} \mathbb{E}|X_{b_l\nu}|^{n_l^\nu} \mathbf{1}(n_l^\mu \neq 1, n_l^\nu \neq 1). \end{aligned} \tag{3.31}$$

Note that we must have $n_l^\mu + n_l^\nu \ge 2$ for $1 \le l \le k$, because we only consider nonempty blocks. On the other hand, if all n_l^μ are even, then $N_\mu = \sum_{l=1}^k n_l^\mu$ must be even, which contradicts (3.28). Hence we can find some $1 \le l_1 \le k$ such that $n_{l_1}^\mu$ is odd and $n_{l_1}^\mu \ge 3$. Similarly, we can also find some $1 \le l_2 \le k$ such that $n_{l_2}^\nu$ is odd and $n_{l_2}^\nu \ge 3$. We abbreviate $\hat{n}_l^\mu := n_l^\mu \wedge 3$ and $\hat{n}_l^\nu := n_l^\nu \wedge 3$. From the above discussions, we see that

$$\frac{1}{2} \sum_{l=1}^k (\hat{n}_l^\mu + \hat{n}_l^\nu) \ge \frac{1}{2} \sum_{l \neq l_1, l_2}^k (\hat{n}_l^\mu + \hat{n}_l^\nu) + \frac{3}{2} + \frac{3}{2} \ge (k-2) + 3 = k+1. \tag{3.32}$$

Now using the moment assumption (1.21), we can bound (3.31) by

$$|\mathbb{E}[\![M_w]\!]| \le C \sum_{\Gamma} \sum_{b_1,\dots,b_k}^{*} |\mathbb{E}G(\Gamma)| N^{-\sum_{l=1}^{k}(\hat{n}_l^{\mu}+\hat{n}_l^{\nu})/2}. \tag{3.33}$$

Next we deal with $|\mathbb{E}G(\Gamma)|$. We consider the following 3 cases separately:

(1) there are at least 2 off-diagonal G-edges in $G(\Gamma)$;
(2) there is only 1 off-diagonal G-edge in $G(\Gamma)$;
(3) there is no off-diagonal G-edge in $G(\Gamma)$.

In case (1), we trivially have $|\mathbb{E}G(\Gamma)| \prec \Phi^2$. In case (2), we use the same trick as in (3.18). Let the off-diagonal G-edge be $G_{ij}^{(\mu\nu)}$. For each diagonal $G_{kk}^{(\mu\nu)}$, we replace it with

$$(G_{kk}^{(\mu\nu)} - \Pi_{kk}) + \Pi_{kk} = \Pi_{kk} + O_{\prec}(\Phi).$$

Plugging these expansions into $\mathbb{E}G(\Gamma)$, we obtain that

$$|\mathbb{E}G(\Gamma)| \prec \Phi^2 + \left|\mathbb{E}G_{ij}^{(\mu\nu)}\right| \prec \Phi^2,$$

where we used (3.19) in the second step.

Finally, in case (3), we have $|\mathbb{E}G(\Gamma)| \prec 1$. Moreover, $n_l^{\mu} + n_l^{\nu}$ is even for any $1 \le l \le k$. Take $1 \le l_1, l_2 \le k$ such that $n_{l_1}^{\mu}, n_{l_2}^{\nu}$ are odd and $n_{l_1}^{\mu}, n_{l_2}^{\nu} \ge 3$. If $l_1 \ne l_2$, then we must have $\hat{n}_{l_1}^{\mu} + \hat{n}_{l_1}^{\nu} \ge 4$, $\hat{n}_{l_2}^{\mu} + \hat{n}_{l_2}^{\nu} \ge 4$, and hence

$$\frac{1}{2}\sum_{l=1}^{k}\left(\hat{n}_l^{\mu} + \hat{n}_l^{\nu}\right) \ge \frac{1}{2}\sum_{l\ne l_1,l_2}^{k}(\hat{n}_l^{\mu} + \hat{n}_l^{\nu}) + 4 \ge k + 2.$$

Otherwise, if $l_1 = l_2$, then

$$\frac{1}{2}\sum_{l=1}^{k}\left(\hat{n}_l^{\mu} + \hat{n}_l^{\nu}\right) \ge \frac{1}{2}\sum_{l\ne l_1}^{k}(\hat{n}_l^{\mu} + \hat{n}_l^{\nu}) + 3 \ge k + 2.$$

Now applying the above estimates and (3.32) to (3.33), we obtain that

$$\begin{aligned} |\mathbb{E}[\![M_w]\!]| &\prec \sum_{\Gamma \text{ in Case (1), (2)}} \Phi^2 N^{k-\sum_{l=1}^{k}(\hat{n}_l^{\mu}+\hat{n}_l^{\nu})/2} + \sum_{\Gamma \text{ in Case (3)}} N^{k-\sum_{l=1}^{k}(\hat{n}_l^{\mu}+\hat{n}_l^{\nu})/2} \\ &\le C(N^{-1}\Phi^2 + N^{-2}) \le CN^{-1}\Phi^2. \end{aligned}$$

This concludes the proof of (3.27), and hence finishes the proof of (3.14).

Acknowledgements. The authors would like to thank Zongming Ma for fruitful discussions and valuable suggestions on possible statistical applications, which have significantly improved this paper. We are also grateful to the editor for carefully reading our manuscript and suggesting several improvements.

References.

[1] T. W. Anderson. Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148, 1963.

[2] Z. D. Bai. Convergence rate of expected spectral distributions of large random matrices. part II. sample covariance matrices. *Ann. Probab.*, 21(2):649–672, 1993.

[3] Z. D. Bai, B. Q. Miao, and G. M. Pan. On asymptotics of eigenvectors of large sample covariance matrix. *Ann. Probab.*, 35(4):1532–1572, 2007.

[4] Z. D. Bai and J. W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.*, 26:316–345, 1998.

[5] Z. D. Bai and J. W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*, volume 2 of *Mathematics Monograph Series*. Science Press, Beijing, 2006.

[6] Z. Bao, G. Pan, and W. Zhou. Universality for the largest eigenvalue of sample covariance matrices with general population. *Ann. Statist.*, 43:382–421, 2015.

[7] Z. G. Bao, G. M. Pan, and W. Zhou. Local density of the spectrum on the edge for sample covariance matrices with general population. *Preprint*, 2013.

[8] A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.*, 19(33):1–53, 2014.

[9] A. Bloemendal, A. Knowles, H.-T. Yau, and J. Yin. On the principal components of sample covariance matrices. *Probability Theory and Related Fields*, 164(1):459–552, 2016.

[10] P. Bourgade and H.-T. Yau. The eigenvector moment flow and local quantum unique ergodicity. *Communications in Mathematical Physics*, 350(1):231–278, 2017.

[11] T. T. Cai, Z. Ma, and Y. Wu. Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.*, 41(6):3074–3110, 2013.

[12] E. B. Davies. The functional calculus. *J. London Math. Soc. (2)*, 52:166–176, 1995.

[13] M. Dieng and C. A. Tracy. *Application of Random Matrix Theory to Multivariate Statistics*, pages 443–507. Springer New York, New York, NY, 2011.

[14] X. Ding and F. Yang. A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *arXiv:1607.06873*.

[15] N. El Karoui. Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *Ann. Probab.*, 35(2):663–714, 2007.

[16] L. Erdős, A. Knowles, and H.-T. Yau. Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré*, 14:1837–1926, 2013.

[17] L. Erdős, B. Schlein, and H.-T. Yau. Local semicircle law and complete delocalization for Wigner random matrices. *Communications in Mathematical Physics*, 287(2):641–655, 2009.

[18] L. Erdős, H.-T. Yau, and J. Yin. Universality for generalized Wigner matrices with Bernoulli distribution. *J. of Combinatorics*, 2(1):15–81, 2011.

[19] F. Götze and A. Tikhomirov. Rate of convergence in probability to the Marchenko-Pastur law. *Bernoulli*, 10(3):503–548, 2004.

[20] W. Hachem, A. Hardy, and J. Najim. Large complex correlated Wishart matrices: Fluctuations and asymptotic independence at the edges. *Ann. Probab.*, 44(3):2264–2348, 2016.

[21] I. M. Johnstone. High dimensional statistical inference and random matrices. *arXiv:0611589*.

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

[23] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327, 2001.

[24] I. M. Johnstone. Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy-Widom limits and rates of convergence. *Ann. Statist.*, 36(6):2638–2716, 2008.

[25] A. Knowles and J. Yin. Eigenvector distribution of Wigner matrices. *Probability Theory and Related Fields*, 155(3):543–582, 2013.

[26] A. Knowles and J. Yin. The isotropic semicircle law and deformation of Wigner matrices. *Communications on Pure and Applied Mathematics*, 66(11):1663–1749, 2013.

[27] A. Knowles and J. Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, pages 1–96, 2016.

[28] J. O. Lee and K. Schnelli. Tracy-Widom distribution for the largest eigenvalue of real sample covariance matrices with general population. *Ann. Appl. Probab.*, 26(6):3786–3839, 2016.

[29] J. O. Lee and J. Yin. A necessary and sufficient condition for edge universality of Wigner matrices. *Duke Math. J.*, 163:117–173, 2014.

[30] Z. Ma. Sparse principal component analysis and iterative thresholding. *Ann. Statist.*, 41(2):772–801, 2013.

[31] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1:457, 1967.

[32] A. Onatski. The tracy-widom limit for the largest eigenvalues of singular complex Wishart matrices. *Ann. Appl. Probab.*, 18(2):470–490, 2008.

[33] A. Onatski. Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479, 2009.

[34] N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLOS Genetics*, 2(12):1–20, 2006.

[35] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.

[36] N. S. Pillai and J. Yin. Universality of covariance matrices. *Ann. Appl. Probab.*, 24:935–1001, 2014.

[37] J. Silverstein and Z. Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175 – 192, 1995.

[38] J. W. Silverstein. On the eigenvectors of large dimensional sample covariance matrices. *Journal of Multivariate Analysis*, 30(1):1 – 16, 1989.

[39] J. W. Silverstein. Weak convergence of random functions defined by the eigenvectors of sample covariance matrices. *Ann. Probab.*, 18(3):1174–1194, 1990.

[40] J. W. Silverstein and S. I. Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295 – 309, 1995.

[41] T. Tao and V. Vu. Random matrices: Universal properties of eigenvectors. *Random Matrices: Theory and Applications*, 01(01):1150001, 2012.

[42] H. Xi, F. Yang, and J. Yin. Local circular law for the product of a deterministic

matrix with a random matrix. *Electron. J. Probab.*, 22:77 pp., 2017.
[43] N. Xia and Z. Bai. Convergence rate of eigenvector empirical spectral distribution of large Wigner matrices. *Statistical Papers*, pages 1–33, 2016.
[44] N. Xia, Y. Qin, and Z. Bai. Convergence rates of eigenvector empirical spectral distribution of large dimensional sample covariance matrix. *Ann. Statist.*, 41(5):2572–2607, 2013.

Department of Mathematics
University of California, Los Angeles,
Los Angeles, CA 90095,
USA
E-mail: fyang75@math.ucla.edu
jyin@math.ucla.edu

Department of Mathematics
University of Wisconsin-Madison
Madison, WI 53706
USA
E-mail: haokai@math.wisc.edu