The Tracy-Widom law for the Largest Eigenvalue of F Type Matrices

X. Han, G. M. Pan and B. Zhang

Division of Mathematical Sciences

School of Physical and Mathematical Sciences

Nanyang Technological University, Singapore

Abstract

Let $\mathbf{A}_p = \frac{\mathbf{Y}\mathbf{Y}^*}{m}$ and $\mathbf{B}_p = \frac{\mathbf{X}\mathbf{X}^*}{n}$ be two independent random matrices where $\mathbf{X} = (X_{ij})_{p\times n}$ and $\mathbf{Y} = (Y_{ij})_{p\times m}$ respectively consist of real (or complex) independent random variables with $\mathbb{E}X_{ij} = \mathbb{E}Y_{ij} = 0$, $\mathbb{E}|X_{ij}|^2 = \mathbb{E}|Y_{ij}|^2 = 1$. Denote by λ_1 the largest root of the determinantal equation $\det(\lambda\mathbf{A}_p - \mathbf{B}_p) = 0$. We establish the Tracy-Widom type universality for λ_1 under some moment conditions on X_{ij} and Y_{ij} when p/m and p/n approach positive constants as $p \to \infty$.

KEYWORDS: Tracy-Widom distribution, largest eigenvalue, sample covariance matrix, F matrix.

1 Introduction

High-dimensional data now commonly arise in many scientific fields such as genomics, image processing, microarray, proteomics and finance, to name but a few. It is well-known that the classical theory of multivariate statistical analysis for the fixed dimension p and large sample size n may lose its validity when handling high-dimensional data. A popular tool in analyzing large covariance matrices and hence high-dimensional data is random matrix theory. The spectral analysis of high-dimensional sample covariance matrices has attracted considerable interests among statisticians, probabilitists and mathematicians since the seminal work of Marcenko and Pastur [17] about the limiting spectral distribution for a class of sample covariance matrices. One can refer to the monograph of Bai and Silverstein [1] for a comprehensive summary and references therein.

The largest eigenvalue of covariance matrices plays an important role in multivariate statistical analysis such as principle component analysis (PCA), multivariate analysis of variance (MANOVA) and discriminant analysis. One may refer to [18] for more details. In this paper we focus on the

largest eigenvalue of the F type matrices. Suppose that

$$\mathbf{A}_p = \frac{\mathbf{Y}\mathbf{Y}^*}{m}, \quad \mathbf{B}_p = \frac{\mathbf{X}\mathbf{X}^*}{n} \tag{1.1}$$

are two independent random matrices where $\mathbf{X} = (X_{ij})_{p\times n}$ and $\mathbf{Y} = (Y_{ij})_{p\times m}$ respectively consist of real (or complex) independent random variables with $\mathbb{E}X_{ij} = \mathbb{E}Y_{ij} = 0$ and $\mathbb{E}|X_{ij}|^2 = \mathbb{E}|Y_{ij}|^2 = 1$. Consider the determinantal equation

$$\det(\lambda\mathbf{A}_p - \mathbf{B}_p) = 0. \tag{1.2}$$

When $\mathbf{A}_p$ is invertible, the roots to (1.2) are the eigenvalues of a F matrix

$$\mathbf{A}_p^{-1}\mathbf{B}_p, \tag{1.3}$$

referred to as a Fisher matrix in the literature. The determinantal equation (1.2) is closely connected with the generalized eigenproblem

$$\det[\lambda(\mathbf{A}_p + \mathbf{B}_p) - \mathbf{B}_p] = 0. \tag{1.4}$$

We illustrate this in the next section. Many classical multivariate statistical tests are based on the roots of (1.2) or (1.4). For instance, one may use them to test the equality of two covariance matrices and the general linear hypothesis. In the framework of multivariate analysis of variance (MANOVA), $\mathbf{A}_p$ represents the within group covariance matrix while $\mathbf{B}_p$ means the between groups covariance matrix. A one-way MANOVA can be used to examine the hypothesis of equality of the mean vectors of interest.

Tracy and Widom in [24, 25] first discovered the limiting distributions of the largest eigenvalue for the large Gaussian Wigner ensemble, thus named as Tracy-Widom's law. Since their pioneer work study toward the largest eigenvalues of large random matrices becomes flourishing. To name a few we mention [11], [12], [6], [10] and [21]. Among them we would mention El Karoui [6] which handled the largest eigenvalue of Wishart matrices for the nonnull population covariance matrix and provided a kind of condition on the population covariance matrix to ensure the Tracy-Widow law (see (4.41) below).

A follow-up to the above results is to establish the so-called universality property for generally distributed large random matrices. Specifically speaking, the universality property states that the limiting behavior of an eigenvalue statistic usually is not dependent on the distribution of the matrix entries. Indeed, the Tracy-Widom law has been established for the general sample covariance matrices under very general assumptions on the distributions of the entries of $\mathbf{X}$. The readers can refer to [22], [23], [8], [9], [19], [27], [3], [16], [15] for some representative developments on this topic. When proving universality an important tool is the Lindeberg comparison strategy (see Tao and Vu in [22] and Erdos, Yau and Yin [8]) and an important input when applying Lindeberg's

comparison strategy is the strong local law developed by Erdos, Schlein and Yau in [7] and Erdos, Yau and Yin in [8].

Johnstone in [13] proved that the largest root of (1.1) converges to Tracy and Widom's distribution of type one after appropriate centering and scaling when the dimension p of the matrices $\mathbf{A}_p$ and $\mathbf{B}_p$ is even, $\lim_{p\to\infty} p/m < 1$ and $\mathbf{B}_p$ and $\mathbf{A}_p$ are both Wishart matrices. It is believed that the limiting distribution should not be affected by the dimension p. Indeed, numerical investigations both in [13] and [14] suggest that the Tracy and Widom approximation in the odd dimension case works as well as in the even dimension case. Besides, as it can be guessed, the Tracy and Widom approximation should not rely on the Gaussian assumption. However, theoretical support for these remains open. Furthermore, when $\mathbf{A}_p$ is not invertible the limiting distribution of the largest root to (1.1) is unknown yet even under the gaussian assumption.

In this paper, we prove the universality of the largest root of (1.2) by imposing some moment conditions on $\mathbf{A}_p$ and $\mathbf{B}_p$. Specifically speaking we prove that the largest root of (1.2) converges in distribution to the Tracy and Widom law for the general distributions of the entries of $\mathbf{X}$ and $\mathbf{Y}$ no matter what the dimension p is, even or odd. Moreover the result holds when $\lim_{p\to\infty} p/m < 1$ or $\lim_{p\to\infty} p/m > 1$, corresponding to invertible $\mathbf{A}_p$ and non-invertible $\mathbf{A}_p$. This result also implies the asymptotic distribution of the largest root of (1.4).

At this point it is also appropriate to mention some related work about the roots of (1.2). The limiting spectral distribution of the roots was derived by [26] and [1]. One may also find the limits of the largest root and the smallest root in [1]. Central limit theorem about linear spectral statistics was established in [29]. Very recently, the so-called spiked F model has been investigated by [5] and [28]. We would like to point out that they prove the local asymptotic normality or asymptotic normality for the largest eigenvalue of the spiked F model, which is completely different from our setting.

We conclude this section by outlining some ideas in the proof and presenting the structure of the rest of the paper. When $\mathbf{A}_p$ is invertible, the roots to (1.2) become those of the F matrix $\mathbf{A}_p^{-1}\mathbf{B}_p$ so that we may work on $\mathbf{A}_p^{-1}\mathbf{B}_p$. Roughly speaking, $\mathbf{A}_p^{-1}\mathbf{B}_p$ can be viewed as a kind of general sample covariance matrix $\mathbf{T}_n^{1/2}\mathbf{X}\mathbf{X}^*\mathbf{T}_n^{1/2}$ with $\mathbf{T}_n$ being a population covariance matrix by conditioning on $\mathbf{B}_p$. Denote the largest root of (1.2) by λ_1. The key idea is to break λ_1 into a sum of two parts as follows

$$\lambda_1 - \mu_p = (\lambda_1 - \hat{\mu}_p) + (\hat{\mu}_p - \mu_p), \tag{1.5}$$

where $\hat{\mu}_p$ is an appropriate value when $\mathbf{B}_p$ is given and μ_p is an appropriate value when $\mathbf{B}_p$ is not given (their definitions are given in the later sections). However we can not condition on $\mathbf{B}_p$ directly. Instead we first construct an appropriate event so that we can handle the first term on the right hand of (1.5) on the event to apply the earlier results about $\mathbf{T}_n^{1/2}\mathbf{X}\mathbf{X}^*\mathbf{T}_n^{1/2}$. Particularly we need to verify the condition (4.41) below. Once this is done, the next step is to prove that the second term on the right hand of (1.5) after scaling converges to zero in probability. This approach

is different from that used in the literature in proving universality for the local eigenvalue statistics.

Unfortunately, when $\mathbf{A}_p$ is not invertible we can not work on F matrices $\mathbf{A}^{-1}\mathbf{B}_p$ anymore. To overcome the difficulty we instead start from the determinantal equation (1.2). It turns out that the largest root λ_1 can then be linked to the largest root of some F matrix when $\mathbf{X}$ consists of Gaussian random variables. Therefore the result about F matrices $\mathbf{A}^{-1}\mathbf{B}_p$ is applicable. For general distributions we find that it is equivalent to working on such a "covariance-type" matrix

$$\mathbf{D}^{-\frac{1}{2}}\mathbf{U}_1\mathbf{X}(\mathbf{I}-\mathbf{X}^*\mathbf{U}_2^*(\mathbf{U}_2\mathbf{X}\mathbf{X}^*\mathbf{U}_2^*)^{-1}\mathbf{U}_2\mathbf{X})\mathbf{X}^*\mathbf{U}_1^*\mathbf{D}^{-\frac{1}{2}}. \tag{1.6}$$

The definitions of $\mathbf{D}$ and $\mathbf{U}_j, j=1,2$ are given in the later section. This matrix is much more complicated than general sample covariance matrices. To deal with (1.6) we construct a 3×3 block linearization matrix

$$\mathbf{H}=\mathbf{H}(\mathbf{X})=\begin{pmatrix} -z\mathbf{I} & 0 & \mathbf{D}^{-1/2}\mathbf{U}_1\mathbf{X} \\ 0 & 0 & \mathbf{U}_2\mathbf{X} \\ \mathbf{X}^T\mathbf{U}_1^T\mathbf{D}^{-1/2} & \mathbf{X}^T\mathbf{U}_2^T & -\mathbf{I} \end{pmatrix}, \tag{1.7}$$

where $z=E+i\eta$ is a complex number with a positive imaginary part. It turns out that the upper left block of the 3×3 block matrix $\mathbf{H}^{-1}$ is the Stieltjes transform of (1.6) by simple calculations. We next develop the strong local law around the right end support μ_p by using a type of Lindeberg's comparison strategy raised in [15] and then use it to prove edge universality by adapting the approach used in [8] and [3].

The paper is organized as follows. Section 2 is to give the main results. A statistical application and Tracy-Widom approximation will be discussed in Section 3. Section 4 is devoted to proving the main result when $\mathbf{A}_p$ is invertible. In section 5 we will show the equivalence between the asymptotic means and asymptotic variances respectively given by [13] and by this paper. Sections 6 and 7 will prove the main result when $\mathbf{A}_p$ is not invertible.

2 The main results

Throughout the paper we make the following conditions.

Condition 1. *Assume that $\{Z_{ij}\}$ are independent random variables with $\mathbb{E}Z_{ij}=0, \mathbb{E}|Z_{ij}|^2=1$. For all $k\in N$, there is a constant C_k such that $\mathbb{E}|Z_{ij}|^k\le C_k$. In addition, if $\{Z_{ij}\}$ are complex, then $\mathbb{E}Z_{ij}^2=0$.*

We say that a random matrix $\mathbf{Z}=(Z_{ij})$ satisfies Condition 1 if its entries $\{Z_{ij}\}$ satisfy Condition 1.

Condition 2. *Assume that random matrices $\mathbf{X}=(\mathbf{X}_{ij})_{p,n}$ and $\mathbf{Y}=(\mathbf{Y}_{ij})_{p,m}$ are independent.*

Condition 3. *Set $m = m(p)$ and $n = n(p)$. Suppose that*

$$\lim_{p\to\infty}\frac{p}{m} = d_1 > 0, \quad \lim_{p\to\infty}\frac{p}{n} = d_2 > 0, \quad 0 < \lim_{p\to\infty}\frac{p}{m+n} < 1.$$

To present the main results uniformly we define $\breve{m} = \max\{m, p\}$, $\breve{n} = \min\{n, m+n-p\}$ and $\breve{p} = \min\{m, p\}$. Moreover let

$$\sin^2(\gamma/2) = \frac{\min\{\breve{p},\breve{n}\} - 1/2}{\breve{m}+\breve{n}-1}, \quad \sin^2(\psi/2) = \frac{\max\{\breve{p},\breve{n}\} - 1/2}{\breve{m}+\breve{n}-1}. \tag{2.1}$$

$$\mu_{J,p} = \tan^2(\frac{\gamma+\psi}{2}), \quad \sigma^3_{J,p} = \mu^3_{J,p}\frac{16}{(\breve{m}+\breve{n}-1)^2}\frac{1}{\sin(\gamma)\sin(\psi)\sin^2(\gamma+\psi)}. \tag{2.2}$$

Formulas (2.2) can be found in [13] when $d_1 < 1$.

We below present alternative expressions of $\mu_{J,p}$ and $\sigma_{J,p}$. To this end, define a modified density of the Marchenko-Pastur law [17] (MP law) by

$$\varrho_p(x) = \frac{1}{2\pi x\frac{\breve{p}}{\breve{m}}}\sqrt{(b_p - x)(x - a_p)}\mathbf{I}(a_p \le x \le b_p), \tag{2.3}$$

where $a_p = (1 - \sqrt{\frac{\breve{p}}{\breve{m}}})^2$ and $b_p = (1 + \sqrt{\frac{\breve{p}}{\breve{m}}})^2$. Let $\gamma_1 \ge \gamma_2 \ge \cdots \ge \gamma_p$ satisfy

$$\int_{\gamma_j}^{+\infty}\varrho_p(x)dx = \frac{j}{p}, \tag{2.4}$$

with $\gamma_0 = b_p$ and $\gamma_p = a_p$. Moreover suppose that $c_p \in [0, a_p)$ satisfies the equation

$$\int_{-\infty}^{+\infty}(\frac{c_p}{x - c_p})^2\varrho_p(x)dx = \frac{n}{p}. \tag{2.5}$$

One may easily check the existence and uniqueness of c_p. Define

$$\mu_p = \frac{1}{c_p}(1 + \frac{p}{n}\int_{-\infty}^{+\infty}(\frac{c_p}{x - c_p})\varrho_p(x)dx) \tag{2.6}$$

and

$$\frac{1}{\sigma_p^3} = \frac{1}{c_p^3}(1 + \frac{p}{n}\int_{-\infty}^{+\infty}(\frac{c_p}{x - c_p})^3\varrho_p(x)dx). \tag{2.7}$$

It turns out that (2.2) and (2.6)-(2.7) are equivalent subject to some scaling, which is verified in Section 5.

We also need the following moment match condition.

Definition 1 (moment matching)**.** *Let $\mathbf{X}^1 = (x^1_{ij})_{M\times N}$ and $\mathbf{X}^0 = (x^0_{ij})_{M\times N}$ be two matrices satsfing Condition 1 . We say that $\mathbf{X}^1$ matches $\mathbf{X}^0$ to order q, if for the integers i,j,l and k satisfing $1 \le i \le M$, $1 \le j \le N$,$0 \le l, k$ and $l + k \le q$, they have the relationship*

$$\mathbb{E}\left[(\Im x^1_{ij})^l(\Re x^1_{ij})^k\right] = \mathbb{E}\left[(\Im x^0_{ij})^l(\Re x^0_{ij})^k\right] + O(\exp(-(\log p)^C)), \tag{2.8}$$

where C is some positive constant bigger than one, $\Re x$ is the real part and $\Im x$ is the imaginary part of x.

Throughout the paper we use $\mathbf{X}^0$ to stand for the random matrix consisting of independent Gaussian random variables with mean zero and variance one.

Denote the type-i Tracy-Widom distribution by F_i, i=1, 2(see [25]). Set $\mathbf{B}_p = \frac{\mathbf{X}\mathbf{X}^*}{\breve{n}}$ and $\mathbf{A}_p = \frac{\mathbf{Y}\mathbf{Y}^*}{\breve{m}}$. We are now in a position to state the main results about F type matrices.

Theorem 2.1. *Suppose that the real random matrices* $\mathbf{X}$ *and* $\mathbf{Y}$ *satisfy Conditions 1-3. Moreover suppose that* $0 < d_2 < \infty$. *Denote the largest root of* $\det(\lambda\mathbf{A}_p - \mathbf{B}_p) = 0$ *by* λ_1.

(i) If $0 < d_1 < 1$, *then*

$$\lim_{p\to\infty} P(\frac{\frac{\breve{n}}{\breve{m}}\lambda_1 - \mu_{J,p}}{\sigma_{J,p}} \le s) = F_1(s). \tag{2.9}$$

(ii) If $d_1 > 1$ *and* $\mathbf{X}$ *matches the standard* $\mathbf{X}^0$ *to order 3, then (2.9) still holds.*

Remark 1. *When* $\mathbf{X}$ *and* $\mathbf{Y}$ *are complex random matrices, Theorem 2.1 still holds but the Tracy-Widom distribution* $F_1(s)$ *should be replaced by* $F_2(s)$.

If $0 < d_1 < 1$, *then* $\mathbf{A}_p$ *is invertible. In this case the largest eigenvalue* λ_1 *is that of F matrices* $\mathbf{A}_p^{-1}\mathbf{B}_p$. *If* $d_1 > 1$, *then* $\mathbf{A}_p$ *is not invertible.*

Remark 2. *Theorem 2.1 immediately implies the distribution of the largest root of* $\det(\lambda(\mathbf{B}_p + \mathbf{A}_p) - \mathbf{B}_p) = 0$. *In fact the largest root of* $\det(\lambda(\mathbf{B}_p + \mathbf{A}_p) - \mathbf{B}_p) = 0$ *is* $\frac{\lambda_1}{1+\lambda_1}$ *if* λ_1 *is the largest root of the F matrices* $\mathbf{B}_p\mathbf{A}_p^{-1}$ *in Theorem 2.1 when* $0 < d_1 < 1$.

When $d_1 > 1$ *the largest root of* $\det\Big(\lambda(\mathbf{B}_p + \mathbf{A}_p) - \mathbf{B}_p\Big) = 0$ *is one with multiplicity* $(p - m)$. *We instead consider the* $(p - m + 1)$*th largest root of* $\det\Big(\lambda(\mathbf{B}_p + \mathbf{A}_p) - \mathbf{B}_p\Big) = 0$. *It turns out that the* $(p - m + 1)$*th largest root of* $\det\Big(\lambda(\mathbf{B}_p + \mathbf{A}_p) - \mathbf{B}_p\Big) = 0$ *is* $\frac{\lambda_1}{1+\lambda_1}$ *if* λ_1 *is the largest root of* $det(\lambda\mathbf{A}_p - \mathbf{B}_p) = 0$.

Moreover note the equality

$$(\mathbf{B}_p + \mathbf{A}_p)^{-1}\mathbf{B}_p + (\mathbf{B}_p + \mathbf{A}_p)^{-1}\mathbf{A}_p = I.$$

If $\mathbf{Y}$ *matches* $\mathbf{X}^0$ *to order 3, then the smallest positive root of* $\det(\lambda(\mathbf{B}_p + \mathbf{A}_p) - \mathbf{B}_p) = 0$ *also tends to type-1 Tracy-Widom distribution after appropriate centralizing and rescaling by Theorem 2.1 when* $d_1 > 1$ *and* $d_2 > 1$.

We would like to point out that Johnstone [13] proved part (i) of Theorem (2.1) when p is even, $\mathbf{A}_p$ and $\mathbf{B}_p$ are both Wishart matrices. Part (ii) of Theorem (2.1) is new even if $\mathbf{A}_p$ and $\mathbf{B}_p$ are both Wishart matrices. When proving Theorem 2.1 we have indeed obtained different asymptotic mean and variance. Precisely we have proved that

$$\lim_{p\to\infty} P(\sigma_p\breve{n}^{2/3}(\lambda_1 - \mu_p) \le s) = F_1(s) \tag{2.10}$$

and that

$$|\frac{\breve{m}}{\breve{n}}\mu_{J,p} - \mu_p| = O(p^{-1}), \quad \lim_{p\to\infty} \sigma_p\frac{\breve{m}}{\breve{n}^{1/3}}\sigma_{J,p} = 1. \tag{2.11}$$

(2.10) and (2.11) imply Theorem 2.1.

3 Application and Simulations

This section is to discuss some applications of our universality results in high-dimensional statistical inference and conduct simulations to check the quality of the approximations of our limiting law.

3.1 Equality of two covariance matrices

Consider the model of the following form

$$\mathbf{Z}_1 = \mathbf{\Sigma}_1^{\frac{1}{2}}\mathbf{X}, \quad \mathbf{Z}_2 = \mathbf{\Sigma}_2^{\frac{1}{2}}\mathbf{Y},$$

where $\mathbf{X}$ and $\mathbf{Y}$ are $p \times n$ and $p \times m$ random matrices satisfying the conditions of Theorem 2.1, $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are $p \times p$ invertible population covariance matrices. We are interested in testing whether $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. Formally, we focus on the following hypothesis testing problem

$$\mathbf{H}_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 \quad vs. \quad \mathbf{H}_1 : \mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2.$$

Under the null hypothesis we have

$$det(\lambda \frac{\mathbf{Z}_2\mathbf{Z}_2^*}{\breve{m}} - \frac{\mathbf{Z}_1\mathbf{Z}_1^*}{\breve{n}}) = 0 \Longleftrightarrow det(\lambda \frac{\mathbf{Y}\mathbf{Y}^*}{\breve{m}} - \frac{\mathbf{X}\mathbf{X}^*}{\breve{n}}) = 0,$$

which implies that we can apply our theoretical result to the largest root of $det(\lambda \frac{\mathbf{Z}_2\mathbf{Z}_2^*}{\breve{m}} - \frac{\mathbf{Z}_1\mathbf{Z}_1^*}{\breve{n}}) = 0$ under the null hypothesis. By Theorem 2.1 we see that λ_1 tends to Tracy-Widom's distribution after centralizing and rescalling.

3.2 Simulations

We conduct some numerical simulations to check the accuracy of the distributional approximations in Theorem 2.1 under various settings of $(p, m, , n)$ and the distribution of $\mathbf{X}$. We also study the power for the testing of equality of two covariance matrices.

As in [13] we below use $\ln(\lambda_1)$ to run simulations. To do so we first give its distribution. By [13] and (2.10) we can find that

$$\lambda_1 = \mu_p + \frac{Z}{\sigma_p \breve{n}^{2/3}} + o_p(\breve{n}^{-2/3}), \tag{3.1}$$

where $Z = F_1^{-1}(U)$ and U is a $U(0,1)$ random variable. By Taylor's expansion we then have

$$\ln(\lambda_1) = \ln(\mu_p) + \frac{Z}{\mu_p \sigma_p \breve{n}^{2/3}} + o_p(\breve{n}^{-2/3}). \tag{3.2}$$

Recall $|\frac{m}{n}\mu_{J,p} - \mu_p| = O(p^{-1})$ and $\lim_{p\to\infty} \sigma_p \frac{m}{n^{1/3}} \sigma_{J,p} = 1$ in Section 2. Summarizing the above we can find

$$\lim_{p\to\infty} P(\sigma_{pln}(\ln(\lambda_1) - \mu_{pln}) \leq s) = F_1(s), \tag{3.3}$$

where

$$\mu_{pln} = \ln(\frac{\breve{m}}{\breve{n}}\mu_{J,p}), \quad \sigma_{pln} = \frac{\mu_{J,p}}{\sigma_{J,p}}. \tag{3.4}$$

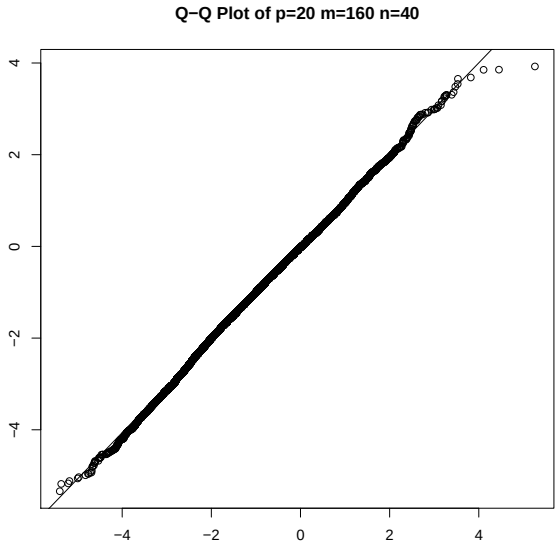
3.2.1 Accuracy of approximations for TW laws and size

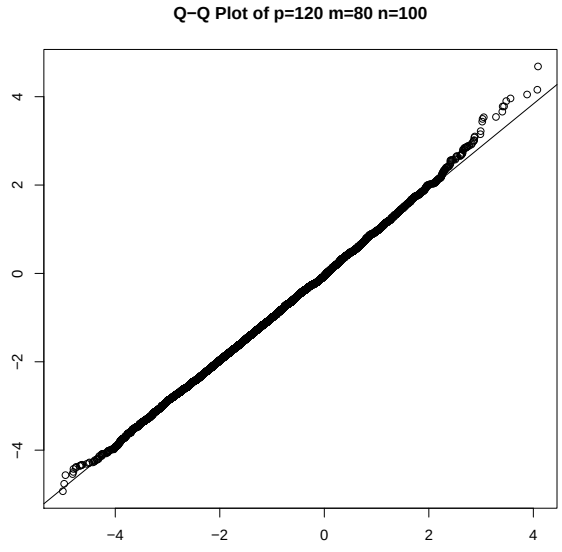
We conduct some numerical simulations to check the accuracy of the distributional approximations in Theorem 2.1, which include the size of the test as well.

Table 1: Standard quantiles for several triples (p,m,n): Gaussian case

		Initial triple M_0=(5,40,10)				Initial triple M_1=(30,20,25)				
Percentile	TW	M_0	$2M_0$	$3M_0$	$4M_0$	M_1	$2M_1$	$3M_1$	$4M_1$	2*SE
-3.9	0.01	0.0208	0.0133	0.0124	0.0115	0.0017	0.0035	0.0048	0.0060	0.002
-3.18	0.05	0.0680	0.0601	0.0562	0.0582	0.0210	0.0276	0.0327	0.0370	0.004
-2.78	0.1	0.1176	0.1120	0.1088	0.1095	0.0608	0.0712	0.0808	0.0842	0.006
-1.91	0.3	0.3154	0.3030	0.3080	0.3084	0.2641	0.2744	0.2864	0.2909	0.009
-1.27	0.5	0.5139	0.5070	0.5051	0.5082	0.4839	0.4904	0.4960	0.4964	0.01
-0.59	0.7	0.7073	0.7154	0.7012	0.7111	0.7055	0.7031	0.7019	0.7005	0.009
0.45	0.9	0.9083	0.9058	0.9047	0.9090	0.9040	0.9010	0.9016	0.9003	0.006
0.98	0.95	0.9561	0.9544	0.9517	0.9557	0.9489	0.9530	0.9504	0.9498	0.004
2.02	0.99	0.9919	0.9909	0.9913	0.9919	0.9878	0.9887	0.9897	0.9901	0.002

Table 1 is done by R. We set two initial triples (p, m, n) of $M_0 = (5, 40, 10)$ and $M_1 = (30, 20, 25)$ and then consider $2M_i, 3M_i$ and $4M_i$, i=1,2. The triples M_0 and M_1 correspond to invertible $\mathbf{YY}^*$ and noninvertible $\mathbf{YY}^*$ respectively. For each case we generate 10000 ($\mathbf{X}$,$\mathbf{Y}$) whose entries follow standard normal distribution. We calculate the largest root of $det(\lambda\frac{\mathbf{Z}_2\mathbf{Z}_2^*}{\breve{m}} - \frac{\mathbf{Z}_1\mathbf{Z}_1^*}{\breve{n}}) = 0$ to get $\ln(\lambda_1)$ and renormalize it with μ_{pln} and σ_{pln}. In the "Pecentile column", the quantiles of TW_1 law corresponding to the "TW" column are listed. We state the values of the empirical distributions of the renormalized λ_1 for various triples at the corresponding quantiles in columns 3-10 and the standard errors based on binomial sampling are listed in the last column. QQ-plots corresponding to the triples $(20, 160, 40)$ and $(120, 80, 100)$ are also stated below.

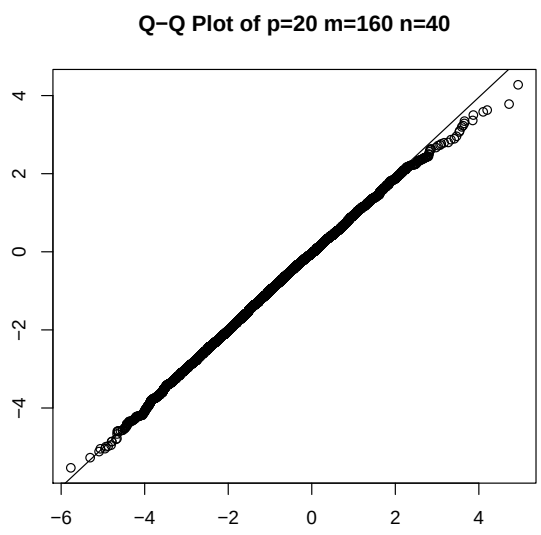




The next two tables and graphs are the same as table 1 and the corresponding graphs except that that we replace the gaussian distribution by the some discrete distribution and uniform distribution.

Table 2: Standard quantiles for several triples (p,m,n): Discrete distribution with the probability mass function P($\mathbf{x} = \sqrt{3}$)=P($\mathbf{x} = -\sqrt{3}$)=1/6 and P($\mathbf{x}$=0)=2/3.

		Initial triple M_0=(5,40,10)				Initial triple M_1=(30,20,25)				
Percentile	TW	M_0	$2M_0$	$3M_0$	$4M_0$	M_1	$2M_1$	$3M_1$	$4M_1$	2*SE
-3.9	0.01	0.0192	0.0132	0.0136	0.0123	0.0006	0.0031	0.0046	0.0047	0.002
-3.18	0.05	0.0637	0.0581	0.0571	0.0573	0.0216	0.0302	0.0321	0.0356	0.004
-2.78	0.1	0.1147	0.1101	0.1099	0.1088	0.0626	0.0733	0.0757	0.0824	0.006
-1.91	0.3	0.3100	0.2966	0.3060	0.3029	0.2665	0.2721	0.2808	0.2827	0.009
-1.27	0.5	0.5000	0.4959	0.4969	0.4996	0.4841	0.4834	0.4985	0.4899	0.01
-0.59	0.7	0.7025	0.7013	0.7099	0.7018	0.6990	0.6992	0.7109	0.6975	0.009
0.45	0.9	0.9107	0.9061	0.9071	0.9036	0.9014	0.9040	0.9059	0.9001	0.006
0.98	0.95	0.9566	0.9546	0.9538	0.9546	0.9503	0.9527	0.9526	0.9512	0.004
2.02	0.99	0.9929	0.994	0.9903	0.9914	0.9890	0.9908	0.9901	0.9894	0.002



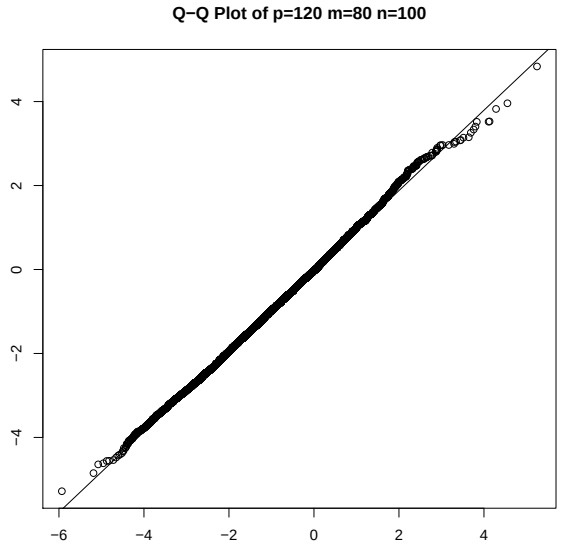
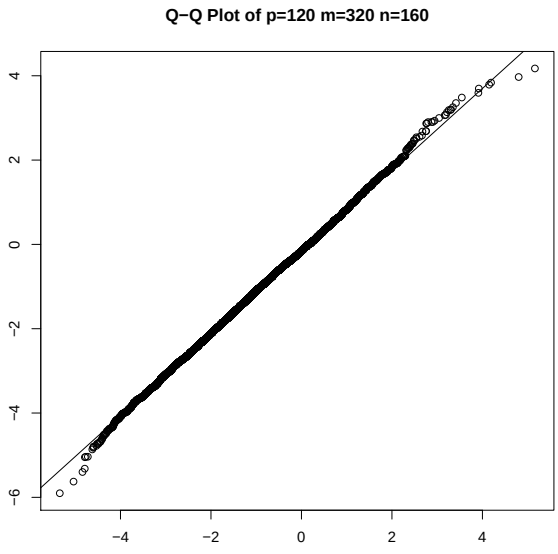
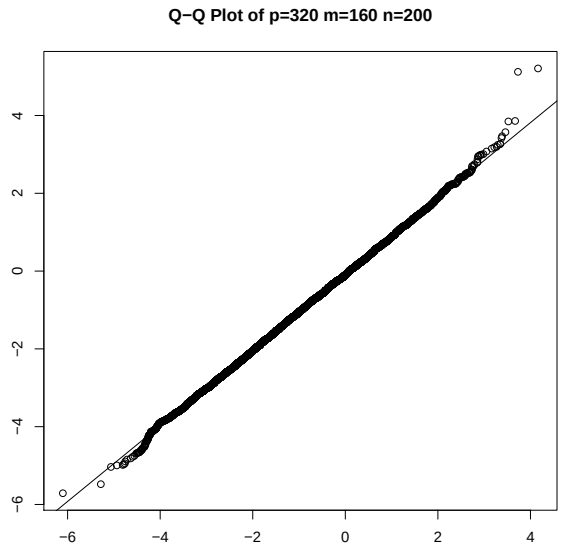


Table 3: Standard quantiles for several triples (p,m,n): Continuous uniform distribution $U(-\sqrt{3},\sqrt{3})$

		Initial triple M_0=(30,80,40)				Initial triple M_1=(80,40,50)				
Percentile	TW	M_0	$2M_0$	$3M_0$	$4M_0$	M_1	$2M_1$	$3M_1$	$4M_1$	2*SE
-3.9	0.01	0.0098	0.0117	0.0122	0.0120	0.0101	0.0087	0.0092	0.0096	0.002
-3.18	0.05	0.0612	0.0632	0.0606	0.0592	0.0514	0.0462	0.0492	0.0482	0.004
-2.78	0.1	0.1205	0.1243	0.1208	0.1197	0.1023	0.0942	0.1033	0.0992	0.006
-1.91	0.3	0.3644	0.3542	0.351	0.3432	0.3132	0.2946	0.3101	0.3017	0.009
-1.27	0.5	0.5767	0.5575	0.5563	0.5496	0.516	0.5073	0.5151	0.5069	0.01
-0.59	0.7	0.7728	0.7540	0.7443	0.7440	0.7182	0.7123	0.714	0.7171	0.009
0.45	0.9	0.9397	0.9243	0.9181	0.9202	0.9141	0.9068	0.9071	0.9059	0.006
0.98	0.95	0.9722	0.9672	0.9599	0.9614	0.9584	0.9538	0.9556	0.9534	0.004
2.02	0.99	0.9959	0.9941	0.993	0.9922	0.9932	0.9912	0.9919	0.9916	0.002





When considering the test of equality of two population covariance matrices since $\mathbf{\Sigma}_1$ is assumed to be invertible in the null case $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, without loss of generality, we may assume that $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}$. Therefore one may refer to Table one as well for the size of the test for the nominal significant levels.

3.2.2 Power

We study the power of the test and consider the alternative case

$$\mathbf{Z}_1 = \mathbf{\Sigma}\mathbf{X}, \quad \mathbf{Z}_2 = \mathbf{Y},$$

where $\mathbf{\Sigma} \neq \mathbf{I}$.

When $\mathbf{Y}\mathbf{Y}^*$ is invertible we choose $\mathbf{\Sigma} = \mathbf{I} + \tau \frac{\frac{p}{m} - r}{1 - \frac{p}{m}} \mathbf{e}_1 \mathbf{e}_1^T$, where $r = \sqrt{\frac{p}{m} + \frac{p}{n} - \frac{p^2}{mn}}$. The reason why we choose the factor $\frac{\frac{p}{m} - r}{1 - \frac{p}{m}}$ is that when $\tau > 1$ it is a spiked F matrix and the largest eigenvalue converges to normal distribution weakly by Proposition 11 of [5].

When $\mathbf{Y}\mathbf{Y}^*$ is not invertible by Theorem 1.2 of [2] we can find out that the smallest non-zero eigenvalue of $\frac{1}{m}\mathbf{\Sigma}^{-1/2}\mathbf{Y}\mathbf{Y}^*\mathbf{\Sigma}^{-1/2}$ is not spiked for the above $\mathbf{\Sigma}$. So it is hard to get a spiked F

matrix. Therefore we use another matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & & & & & & \\ & \omega & & & & & \\ & & 1 & & & & \\ & & & \omega & & & \\ & & & & \ddots & & \\ & & & & & 1 & \\ & & & & & & \omega \end{pmatrix}.$$

In Tables 4-6 the data $\mathbf{X}$ and $\mathbf{Y}$ are generated as in Tables 1-3 and the nominal significant level of our test is 5%.

Table 4: Power of several triples(p,m,n): Gaussian distribution

	Initial triple M_0=(5,40,10)					Initial triple M_1=(30,20,25)			
τ	M_0	$2M_0$	$3M_0$	$4M_0$	ω	M_1	$2M_1$	$3M_1$	$4M_1$
0.5	0.0672	0.0585	0.0563	0.0593	0.3	0.2178	0.4934	0.7071	0.8419
2	0.2763	0.3801	0.4551	0.5067	0.6	0.0574	0.1332	0.2241	0.3106
4	0.6291	0.816	0.9072	0.9567	2	0.1037	0.2166	0.3463	0.5029
6	0.8162	0.9543	0.988	0.9967	3	0.2242	0.5521	0.8156	0.9537

Table 5: Power of several triples(p,m,n): Discrete distribution

	Initial triple M_0=(5,40,10)					Initial triple M_1=(30,20,25)			
τ	M_0	$2M_0$	$3M_0$	$4M_0$	ω	M_1	$2M_1$	$3M_1$	$4M_1$
0.5	0.0674	0.0573	0.0576	0.0595	0.3	0.2101	0.4883	0.7024	0.8425
2	0.3045	0.397	0.4561	0.5171	0.6	0.057	0.1382	0.2176	0.3078
4	0.647	0.8137	0.8984	0.9478	2	0.1055	0.2232	0.3504	0.4974
6	0.8147	0.943	0.9813	0.9936	3	0.2254	0.5487	0.8211	0.9529

Table 6: Power of several triples(p,m,n): Continuous uniform distribution $U(-\sqrt{3}, \sqrt{3})$

	Initial triple M_0=(30,80,40)					Initial triple M_1=(80,40,50)			
τ	M_0	$2M_0$	$3M_0$	$4M_0$	ω	M_1	$2M_1$	$3M_1$	$4M_1$
0.5	0.2283	0.3188	0.3977	0.4662	0.3	0.9965	1.0000	1.0000	1.0000
2	1.0000	1.0000	1.0000	1.0000	0.6	0.7112	0.9623	0.9964	0.9999
4	1.0000	1.0000	1.0000	1.0000	2	0.9257	1.0000	1.0000	1.0000
6	1.0000	1.0000	1.0000	1.0000	3	1.0000	1.0000	1.0000	1.0000

In Tables 4-6 we can find that when $\tau = 0.5 < 1$ $(\mathbf{\Sigma}^{-1/2}\mathbf{Y}\mathbf{Y}^*\mathbf{\Sigma}^{-1/2})\mathbf{X}\mathbf{X}^*$ is not a spiked F matrix and the power is poor. When $\tau > 1$ it is a spiked F matrix and the power increases with the dimension and τ. This phenomenon is due to the fact that it may not cause significant change to the largest eigenvalue of F matrix when finite rank perturbation is weak enough. This phenomenon has been widely discussed for sample covariance matrices, see [10] and [3]. For the spiked F matrix one can refer to [5] and [28]. For the non-invertible case when $\mathbf{\Sigma}$ is far away from $\mathbf{I}(\omega = 0.3$ or $3)$ the power becomes better. This is because when the empirical spectral distribution (ESD) of $\mathbf{\Sigma}$ is very different from the M-P law λ_1 may tend to another point $\mu_{\mathbf{\Sigma}}$ instead of μ_p. Then we may gain good power because $n^{2/3}(\mu_{\mathbf{\Sigma}} - \mu_p)$ may tend to infinity.

4 Proof of Part(i) of Theorem 2.1

4.1 Two key Lemmas

This subsection is to first prove two key lemmas for proving part(i) of Theorem 2.1. We begin with some notation and definitions. Throughout the paper we use $M, M_0, M_0', M_0'', M_1, M_1''$ to denote some generic positive constants whose values may differ from line to line. We also use D to denote sufficiently large positive constants whose values may differ from line to line. We say that an event Λ holds with high probability if for any big positive constant D

$$P(\Lambda^c) \leq n^{-D},$$

for sufficiently large n. Recall the definition of γ_j in (2.4). Let $c_{p,0} \in [0, a_p)$ satisfy

$$\frac{1}{p}\sum_{j=1}^{p}(\frac{c_{p,0}}{\gamma_j - c_{p,0}})^2 = \frac{n}{p}. \tag{4.1}$$

Existence of $c_{p,0}$ will be verified in Lemma 1 below. Moreover define

$$\mu_{p,0} = \frac{1}{c_{p,0}}(1 + \frac{1}{n}\sum_{j=1}^{p}(\frac{c_{p,0}}{\gamma_j - c_{p,0}})), \quad \frac{1}{\sigma_{p,0}^3} = \frac{1}{c_{p,0}^3}(1 + \frac{1}{n}\sum_{j=1}^{p}(\frac{c_{p,0}}{\gamma_j - c_{p,0}})^3). \tag{4.2}$$

Set $\mathbf{A}_p = \frac{1}{m}\mathbf{Y}\mathbf{Y}^*$ and $\mathbf{B}_p = \frac{1}{n}\mathbf{X}\mathbf{X}^*$. Rank the eigenvalues of the matrix $\mathbf{A}_p$ as $\hat{\gamma}_1 \geq \hat{\gamma}_2 \geq \cdots \geq \hat{\gamma}_p$. Let $\hat{c}_p \in [0, \hat{\gamma}_p)$ satisfy

$$\frac{1}{p}\sum_{j=1}^{p}(\frac{\hat{c}_p}{\hat{\gamma}_j - \hat{c}_p})^2 = \frac{n}{p}. \tag{4.3}$$

The existence of $\hat{c}_p$ with high probability will be given in Lemma 2 below. Moreover set

$$\hat{\mu}_p = \frac{1}{\hat{c}_p}(1 + \frac{1}{n}\sum_{j=1}^{p}(\frac{\hat{c}_p}{\hat{\gamma}_j - \hat{c}_p})), \quad \frac{1}{\hat{\sigma}_p^3} = \frac{1}{\hat{c}_p^3}(1 + \frac{1}{n}\sum_{j=1}^{p}(\frac{\hat{c}_p}{\hat{\gamma}_j - \hat{c}_p})^3). \tag{4.4}$$

We now discuss the properties of $c_p, c_{p,0}, \hat{c}_p, \mu_p, \mu_{p,0}, \hat{\mu}_p, \sigma_p, \sigma_{p,0}$ defined (2.5)-(2.7), (4.1)- (4.4) in the next two lemmas. These lemmas are crucial to the proof strategy which transforms F matrices into an appropriate sample covariance matrix.

Lemma 1. *Under the conditions in Theorem 2.1, there exists a constant M_0 such that*

$$\sup_p\{\frac{c_p}{a_p-c_p}\}\le M_0,\quad \sup_p\{\frac{c_{p,0}}{a_p-c_{p,0}}\}\le M_0, \tag{4.5}$$

$$\lim_{p\to\infty} n^{2/3}|\mu_p-\mu_{p,0}|=0, \tag{4.6}$$

$$\lim_{p\to\infty}\frac{\sigma_p}{\sigma_{p,0}}=1,\quad \limsup_p \frac{c_{p,0}}{a_p}<1. \tag{4.7}$$

Proof. The exact expression of c_p in (2.5) can be figured out under the conditions in Theorem 2.1 (see Section 5). In fact, when $n=p$, from (5.9) below we have

$$c_p=\frac{(m-p)^2}{2(m+p)m}.$$

Recall the definition of a_p in (2.3). It follows that

$$\frac{c_p}{a_p}=\frac{(\sqrt{m}+\sqrt{p})^2}{2(m+p)},$$

which further implies that

$$\limsup_p \frac{c_p}{a_p}<1. \tag{4.8}$$

In view of this, there are two constants $M_0>0$ and $M_0''>0$ such that

$$\sup_p\{\frac{c_p}{a_p-c_p}\}\le M_0,\quad \inf_p\{c_p\}\ge M_0''. \tag{4.9}$$

When $n\neq p$, from (5.7) below we have

$$c_p=\frac{n(m+p)(m+n-p)-(m+2n-p)\sqrt{mnp(m+n-p)}}{m(n-p)(m+n)}.$$

Using the above expression for c_p one may similarly obtain (4.8)-(4.9) as well but with tedious calculations and we ignore details here.

Now we define a function $f_1(x)$ by

$$f_1(x)=\frac{1}{p}\sum_{j=1}^{p}(\frac{x}{\gamma_j-x})^2. \tag{4.10}$$

We claim that there exists $c_{p,0}\in(0,a_p)$ so that

$$f_1(c_{p,0})=\frac{n}{p}. \tag{4.11}$$

Indeed, due to (4.8) we obtain

$$f_1(c_p)=\frac{1}{p}\sum_{j=1}^{p}(\frac{c_p}{\gamma_j-c_p})^2\ge\sum_{j=1}^{p}\int_{\gamma_j}^{\gamma_{j-1}}(\frac{c_p}{x-c_p})^2\varrho_p(x)dx=\int_{\gamma_p}^{\gamma_0}(\frac{c_p}{x-c_p})^2\varrho_p(x)dx.$$

This, together with (2.3) and (2.5), implies that

$$f_1(c_p) \geq \frac{n}{p} \tag{4.12}$$

and

$$\frac{n}{p} = \int_{\gamma_p}^{\gamma_0} (\frac{c_p}{x - c_p})^2 \varrho_p(x) dx \geq \frac{1}{p} \sum_{j=1}^{p} (\frac{c_p}{\gamma_{j-1} - c_p})^2. \tag{4.13}$$

Note that $f_1(x)$ is a continuous function on $(0, a_p)$ and $f_1(0) = 0$. These, together with (4.12), ensure that there exists $c_{p,0} \in (0, c_p]$ so that (4.11) holds, as claimed.

We next develop an upper bound for the difference between $c_{p,0}$ and c_p. It follows from (4.12) and (4.13) that

$$|f_1(c_p) - \frac{n}{p}| = \Big|\frac{1}{p} \sum_{j=1}^{p} (\frac{c_p}{\gamma_j - c_p})^2 - \int_{\gamma_p}^{\gamma_0} (\frac{c_p}{x - c_p})^2 \varrho_p(x) dx\Big|$$

$$\leq \frac{1}{p} \Big| \sum_{j=1}^{p} \Big((\frac{c_p}{\gamma_j - c_p})^2 - (\frac{c_p}{\gamma_{j-1} - c_p})^2 \Big) \Big| \leq \frac{2c_p^2(b_p - c_p)}{p(a_p - c_p)^4} \sum_{j=1}^{p} |\gamma_j - \gamma_{j-1}| \leq \frac{2(M_0)^4 b_p (b_p - a_p)}{(M_0'')^2 p},$$

where the last inequality uses (4.8)-(4.9). With $M_1' = \frac{2(M_0)^4 b_p (b_p - a_p)}{(M_0'')^2}$ the above inequality becomes

$$|f_1(c_p) - \frac{n}{p}| \leq \frac{M_1'}{p}. \tag{4.14}$$

Moreover taking derivative of $f(x)$ in (4.10) yields

$$f_1'(x) = \frac{1}{p} \sum_{j=1}^{p} (\frac{2x^2}{(\gamma_j - x)^3} + \frac{2x}{(\gamma_j - x)^2}). \tag{4.15}$$

When $0 < x < c_p$ (smaller than a_p) and $f_1(x) \geq \frac{n}{2p}$,

$$f_1'(x) > \frac{1}{p} \sum_{j=1}^{p} \frac{2x}{(\gamma_j - x)^2} \geq \frac{n}{px} \geq \frac{n}{pa_p}. \tag{4.16}$$

When $c_{p,0} < x \leq c_p$ we always have $f_1(x) \geq \frac{n}{p}$ via (4.11) because $f_1'(x) > 0$ by (4.15). Via (4.14) and (4.16) we then obtain from the mean value theorem that

$$|c_{p,0} - c_p| \leq \frac{M_1' a_p}{n}. \tag{4.17}$$

This, together with (4.9), implies that there is a constant $M_1 > 0$ such that when p is big enough,

$$M_1 < c_{p,0} \leq c_p. \tag{4.18}$$

We conclude from (2.6), (4.2), (4.9), (4.17) and (4.18) that

$$|\mu_p - \mu_{p,0}| \leq |\frac{1}{c_p} - \frac{1}{c_{p,0}}| + \frac{1}{n} \sum_{j=1}^{p} \max\{(|\frac{1}{\gamma_j - c_{p,0}} - \frac{1}{\gamma_j - c_p}|, |\frac{1}{\gamma_j - c_{p,0}} - \frac{1}{\gamma_{j-1} - c_p}|\}$$

$$\leq \frac{|c_p - c_{p,0}|}{M_1^2} + \frac{1}{n}\sum_{j=1}^{p} \frac{(|\gamma_j - \gamma_{j-1}| + |c_p - c_{p,0}|)M_0^2}{M_1^2}$$

$$\leq \frac{M_1' a_p}{nM_1^2} + \frac{b_p - a_p}{nM_1^2} + \frac{pM_0^2 M_1' a_p}{n^2 M_1^2} = O(\frac{1}{p}).$$

Similarly one can prove that

$$|\frac{1}{\sigma_p^3} - \frac{1}{\sigma_{p,0}^3}| = O(\frac{1}{p}). \tag{4.19}$$

(4.6) and the first result in (4.7) then follow. From (4.8) and (4.17) one can also obtain (4.5) and the second result in (4.7). □

Lemma 2. *Under the conditions in Theorem 2.1, for any $\zeta > 0$ there exists a constant $M_\zeta \geq M_0$ such that*

$$\sup_p \{\frac{\hat{c}_p}{\hat{\gamma}_p - \hat{c}_p}\} \leq M_\zeta, \quad \limsup_p \frac{\hat{c}_p}{\hat{\gamma}_p} < 1, \tag{4.20}$$

and

$$\lim_{p\to\infty} n^{2/3}|\hat{\mu}_p - \mu_{p,0}| = 0, \quad \lim_{p\to\infty} \frac{\hat{\sigma}_p}{\sigma_{p,0}} = 1. \tag{4.21}$$

hold with high probability. Indeed (4.20) and (4.21) hold on the event S_ζ defined by

$$S_\zeta = \{\forall j, 1 \leq j \leq p, |\hat{\gamma}_j - \gamma_j| \leq p^\zeta p^{-2/3}\tilde{j}^{-1/3}\}, \tag{4.22}$$

where ζ is a sufficiently small positive constant and $\tilde{j} = \min\{\min\{m, p\} + 1 - j, j\}$.

Proof. Define a function $\hat{f}(x)$ by

$$\hat{f}(x) = \frac{1}{p}\sum_{j=1}^{p}(\frac{x}{\hat{\gamma}_j - x})^2. \tag{4.23}$$

From (4.1) and (4.10) we have

$$f_1(c_{p,0}) = \frac{n}{p}. \tag{4.24}$$

The first aim is to find $\hat{c}_p \in [0, \hat{\gamma}_p)$ to satisfy

$$\hat{f}(\hat{c}_p) = \frac{n}{p}. \tag{4.25}$$

When ζ is small enough we conclude from (4.23), (4.24) and (4.35) that on the event S_ζ

$$|\hat{f}(c_{p,0}) - \frac{n}{p}| = |\hat{f}(c_{p,0}) - f_1(c_{p,0})| = |\frac{1}{p}\sum_{j=1}^{p}((\frac{c_{p,0}}{\hat{\gamma}_j - c_{p,0}})^2 - (\frac{c_{p,0}}{\gamma_j - c_{p,0}})^2)|$$

$$\leq \frac{c_{p,0}^2}{p}\max_j\{\frac{|\hat{\gamma}_j + \gamma_j - 2c_{p,0}|}{(\hat{\gamma}_j - c_{p,0})^2(\gamma_j - c_{p,0})^2}\}\sum_{j=1}^{p}|\hat{\gamma}_j - \gamma_j|$$

$$\leq \frac{c_{p,0}^2}{p}\max_j\{\frac{|p^\zeta p^{-2/3}\tilde{j}^{-1/3}| + 2|\gamma_j - c_{p,0}|}{(-p^\zeta p^{-2/3}\tilde{j}^{-1/3} + \gamma_j - c_{p,0})^2(\gamma_j - c_{p,0})^2}\}\sum_{j=1}^{p}p^\zeta p^{-2/3}\tilde{j}^{-1/3} = O(p^{\zeta-1}), \tag{4.26}$$

where the last step uses the fact that via (4.5) and (4.18)

$$\max_j \{\frac{|p^{\zeta}p^{-2/3}\tilde{j}^{-1/3} + 2\gamma_j - 2c_{p,0}|}{(-p^{\zeta}p^{-2/3}\tilde{j}^{-1/3} + \gamma_j - c_{p,0})^2(\gamma_j - c_{p,0})^2}\} \leq M.$$

Taking derivative of (4.23) yields

$$\hat{f}'(x) = \frac{1}{p}\sum_{j=1}^{p}(\frac{2x^2}{(\hat{\gamma}_j - x)^3} + \frac{2x}{(\hat{\gamma}_j - x)^2}). \tag{4.27}$$

When $0 < c_{p,0} - p^{-1/2} < x < c_{p,0} + p^{-1/2}$ from (4.5) and (4.18) we have on the event S_ζ

$$\begin{aligned} |\hat{f}(c_{p,0}) - \hat{f}(x)| &= \frac{1}{p}|\sum_{j=1}^{p}\frac{c_{p,0}^2(\hat{\gamma}_j - x)^2 - x^2(\hat{\gamma}_j - c_{p,0})^2}{(\hat{\gamma}_j - x)^2(\hat{\gamma}_j - c_{p,0})^2}| \\ &= \frac{1}{p}|\sum_{j=1}^{p}\frac{(c_{p,0} - x)\hat{\gamma}_j[c_{p,0}(\hat{\gamma}_j - x) + x(\hat{\gamma}_j - c_{p,0})]}{(\hat{\gamma}_j - x)^2(\hat{\gamma}_j - c_{p,0})^2}| = O(p^{-1/2}). \end{aligned} \tag{4.28}$$

When $0 < x < \hat{\gamma}_p$ we have

$$\hat{f}'(x) > \frac{1}{p}\sum_{j=1}^{p}\frac{2x}{(\hat{\gamma}_j - x)^2} = \frac{2}{x}\hat{f}(x) > \frac{2}{(c_{p,0} + p^{-1/2})}\hat{f}(x).$$

In view of this, (4.26) and (4.28) there exists $M_2 > 0$ so that

$$\hat{f}'(x) > M_2, \tag{4.29}$$

for sufficiently large p when $0 < x < \hat{\gamma}_p$. On the event S_ζ, applying the mean value theorem yields

$$\hat{f}(c_{p,0} - p^{-1/2}) < \hat{f}(c_{p,0}) - M_2 p^{-1/2}$$

and

$$\hat{f}(c_{p,0} + p^{-1/2}) > \hat{f}(c_{p,0}) + M_2 p^{-1/2}.$$

It follows from (4.26) that when p is large enough,

$$\hat{f}(c_{p,0} - p^{-1/2}) < \frac{n}{p} < \hat{f}(c_{p,0} + p^{-1/2}).$$

Since $\hat{f}(x)$ is continuous on $(0, \hat{\gamma}_p)$ there is $\hat{c}_p \in [0, \hat{\gamma}_p)$ ($c_{p,0} \leq c_p < a_p = \gamma_p$ by Lemma 1) so that (4.25) holds and

$$c_{p,0} - p^{-1/2} < \hat{c}_p < c_{p,0} + p^{-1/2}.$$

From (4.26), (4.25) and (4.29) we have

$$|c_{p,0} - \hat{c}_p| = O(p^{\zeta - 1}). \tag{4.30}$$

Recall $a_p = \gamma_p$. The second inequality in (4.20) holds on the event S_ζ due to (4.7), (4.22) and (4.30). Likewise on the event S_ζ in view of (4.5) and (4.30) there exists a constant $M_\zeta \geq M_0$ such that

$$\sup_p \{\frac{\hat{c}_p}{\hat{\gamma}_p - \hat{c}_p}\} \leq M_\zeta, \tag{4.31}$$

the first inequality in (4.20).

Due to $\hat{c}_p < \hat{\gamma}_p$ and the definition of $\hat{f}(x)$ in (4.23) we have

$$(\frac{\hat{c}_p}{\hat{\gamma}_p - \hat{c}_p})^2 \geq \frac{n}{p},$$

which implies that

$$\hat{c}_p \geq \frac{\sqrt{\frac{n}{p}}\hat{\gamma}_p}{1 + \sqrt{\frac{n}{p}}}. \tag{4.32}$$

It follows from (4.2) and (4.4) that

$$\begin{aligned} |\mu_{p,0} - \hat{\mu}_p| &\leq |\frac{1}{c_{p,0}} - \frac{1}{\hat{c}_p}| + \frac{1}{n}\sum_{j=1}^{p} |\frac{1}{\gamma_j - c_{p,0}} - \frac{1}{\hat{\gamma}_j - \hat{c}_p}| \\ &\leq \frac{|c_{p,0} - \hat{c}_p|}{c_{p,0}\hat{c}_p} + \frac{1}{n}\frac{\sum_{j=1}^{p}(|\gamma_j - \hat{\gamma}_j| + |c_{p,0} - \hat{c}_p|)}{(\gamma_p - c_{p,0})(\hat{\gamma}_p - \hat{c}_p)}. \end{aligned}$$

We then conclude from (4.30)-(4.32) that on the event S_ζ

$$|\mu_{p,0} - \hat{\mu}_p| = O(p^{\zeta - 1}). \tag{4.33}$$

It's similar to prove that

$$|\frac{1}{\hat{\sigma}_p^3} - \frac{1}{\sigma_{p,0}^3}| = O(p^{\zeta-1}). \tag{4.34}$$

(4.21) then holds on the event S_ζ. Moreover, by Theorem 3.3 of [19], for any small $\zeta > 0$ and any $D > 0$,

$$P(S_\zeta^c) \leq p^{-D}. \tag{4.35}$$

The proof is therefore complete.

□

4.2 Proof of Part (i) of Theorem 2.1

Proof. Recall the definition of the matrices $\mathbf{A}_p$ and $\mathbf{B}_p$ above (4.3). Define a F matrix $\mathbf{F} = \mathbf{A}_p^{-1}\mathbf{B}_p$ whose largest eigenvalue is λ_1 according to the definition of λ_1 in Theorem 2.1. It then suffices to find the asymptotic distribution of λ_1 to prove Theorem 2.1.

Recalling the definition of the event S_ζ in (4.22) we may write

$$P(\sigma_p n^{2/3}(\lambda_1 - \mu_p) \leq s) = P\Big((\sigma_p n^{2/3}(\lambda_1 - \mu_p) \leq s) \bigcap S_\zeta\Big) + P\Big((\sigma_p n^{2/3}(\lambda_1 - \mu_p) \leq s) \bigcap S_\zeta^c\Big).$$

This, together with (4.35), implies that (2.10) is equivalent to

$$\lim_{p\to\infty} P\Big((\sigma_p n^{2/3}(\lambda_1-\mu_p)\le s)\bigcap S_\zeta\Big)=F_1(s). \tag{4.36}$$

Write

$$\sigma_p n^{2/3}(\lambda_1-\mu_p)=\frac{\sigma_p}{\hat\sigma_p}\hat\sigma_p n^{2/3}(\lambda_1-\hat\mu_p)+\sigma_p n^{2/3}(\hat\mu_p-\mu_p). \tag{4.37}$$

(see (4.3) and (4.4) for $\hat\sigma_p$ and $\hat\mu_p$). Note that the eigenvalues of $\mathbf{A}_p^{-1}$ are $\frac{1}{\hat\gamma_1}\le\frac{1}{\hat\gamma_2}\le\cdots\le\frac{1}{\hat\gamma_p}$. Rewrite (4.3) as

$$\frac{1}{p}\sum_{j=1}^{p}\Big(\frac{\frac{1}{\hat\gamma_j}\hat c_p}{1-\frac{1}{\hat\gamma_p}\hat c_p}\Big)^2=\frac{n}{p}. \tag{4.38}$$

Also recast (4.4) as

$$\hat\mu_p=\frac{1}{\hat c_p}\Big(1+\frac{p}{n}\frac{1}{p}\sum_{j=1}^{p}\frac{\frac{1}{\hat\gamma_j}\hat c_p}{1-\frac{1}{\hat\gamma_p}\hat c_p}\Big),\quad \frac{1}{\hat\sigma_p^3}=\frac{1}{\hat c_p^3}\Big(1+\frac{p}{n}\frac{1}{p}\sum_{j=1}^{p}\frac{\frac{1}{\hat\gamma_j}\hat c_p}{1-\frac{1}{\hat\gamma_p}\hat c_p}\Big)^3. \tag{4.39}$$

Up to this stage the result about the largest eigenvalue of the sample covariance matrices $\mathbf{ZZ}^*\mathbf{\Sigma}$ with $\mathbf{\Sigma}$ being the population covariance matrix comes into play where $\mathbf{Z}$ is of size $p\times n$ satisfying Condition 1 and $\mathbf{\Sigma}$ is of size $p\times p$. A key condition to ensure Tracy-Widom's law for the largest eigenvalue is that if $\rho\in(0,1/\sigma_1)$ is the solution to the equation

$$\int\Big(\frac{t\rho}{1-t\rho}\Big)^2 dF^{\mathbf{\Sigma}}(t)=\frac{n}{p} \tag{4.40}$$

then

$$\limsup_p \rho\sigma_1<1, \tag{4.41}$$

(one may see [6], Conditions 1.2 and 1.4 and Theorem 1.3 [3], Conditions 2.21 and 2.22 and Theorem 2.18 of [15]). Here $F^{\mathbf{\Sigma}}(t)$ denotes the empirical spectral distribution of Σ and σ_1 means the largest eigenvalue of $\mathbf{\Sigma}$. Now given $\mathbf{A}_p$, if we treat $\mathbf{A}_p^{-1}$ as $\mathbf{\Sigma}$, then (4.41) is satisfied on the event S_ζ due to (4.3) and (4.20) in Lemma 2. It follows from Theorem 1.3 of [3] and Theorem 2.18 of [15] that

$$\lim_{p\to\infty} P\Big((\hat\sigma_p n^{2/3}(\lambda_1-\hat\mu_p)\le s)\bigcap S_\zeta|\mathbf{A}_p\Big)=F_1(s), \tag{4.42}$$

which implies that

$$\lim_{p\to\infty} P\Big((\hat\sigma_p n^{2/3}(\lambda_1-\hat\mu_p)\le s)\bigcap S_\zeta\Big)=F_1(s). \tag{4.43}$$

Moreover by Lemmas 1 and 2 we obtain on the event S_ζ

$$\lim_{p\to\infty}\frac{\sigma_p}{\hat\sigma_p}=1 \tag{4.44}$$

and

$$\lim_{p\to\infty}\sigma_p n^{2/3}(\hat\mu_p-\mu_p)=0. \tag{4.45}$$

(4.36) then follows from (4.37), (4.42)-(4.45) and Slutsky's theorem. The proof is complete. □

5 Proof of (2.11)

Proof. This section is to verify (2.11) and give an exact expressions of c_p, μ_p and σ_p in (2.5)-(2.7) at the mean time. We first introduce the following notation. Let $\breve{m} = \max\{m, p\}$, $\breve{n} = \min\{n, m+n-p\}$ and $\breve{p} = \min\{m, p\}$. Choose $0 < \alpha_p < \frac{\pi}{2}$ and $0 < \beta_p < \frac{\pi}{2}$ to satisfy

$$\sin^2(\alpha_p) = \frac{\breve{p}}{\breve{m}+\breve{n}}, \quad \sin^2(\beta_p) = \frac{\breve{n}}{\breve{m}+\breve{n}}. \tag{5.1}$$

Define

$$\mu_p = \frac{\breve{m}}{\breve{n}} \tan^2(\alpha_p + \beta_p) \tag{5.2}$$

and

$$\frac{1}{\sigma_p^3} = \mu_p^3 \frac{16\breve{n}^2}{(\breve{m}+\breve{n})^2} \frac{1}{\sin(2\beta_p)\sin(2\alpha_p)\sin^2(2\beta_p+2\alpha_p)}. \tag{5.3}$$

We below first verify the equivalence between (2.6)-(2.7) and (5.2)-(5.3). For definiteness, consider $p < m$ in what follows and the case $p > m$ can be discussed similarly. Denote by $s(z)$ the Stieltjes transform of the MP law $\rho_p(x)$

$$s(z) = \int \frac{\rho_p(x)}{x-z} dx, \quad Im(z) > 0$$

and set

$$g_p(x) = \frac{1 - \frac{p}{m} - x - \sqrt{(x-1-\frac{p}{m})^2 - 4\frac{p}{m}}}{2\frac{p}{m}x}, \tag{5.4}$$

which is the function obtained from $s(z)$ by replacing z with x (one may see (3.3.2) of [1]). Evidently, the derivative of $s(z)$ is

$$s'(z) = \int \frac{\rho_p(x)}{(x-z)^2} dx.$$

Note that c_p is outside the support of the MP law (see Lemma 1). In view of the above and (2.5) we obtain

$$c_p^2 g_p'(c_p) = \frac{n}{p}, \tag{5.5}$$

which further implies that

$$\sqrt{(c_p - 1 - \frac{p}{m})^2 - 4\frac{p}{m}} = \frac{(1-\frac{p}{m})^2 - (1+\frac{p}{m})c_p}{\frac{2n}{m} + 1 - \frac{p}{m}}. \tag{5.6}$$

When $n \neq p$, solving (5.6) and disregarding one of the solutions bigger than a_p we have

$$\begin{aligned} c_p &= \frac{(\frac{m+p}{m})(\frac{p}{m}+\frac{p}{n}-\frac{p^2}{mn}) - \sqrt{(1+\frac{p}{m})^2(\frac{p}{m}+\frac{p}{n}-\frac{p^2}{mn})^2 + (1-\frac{p}{m})^2(\frac{p}{m}+\frac{p}{n}-\frac{p^2}{mn}) - (\frac{p}{n}-1)(\frac{p}{m}+\frac{p}{n})}}{(1-\frac{p}{n})(\frac{p}{m}+\frac{p}{n})} \\ &= \frac{n(m+p)(m+n-p) - (m+2n-p)\sqrt{mnp(m+n-p)}}{m(n-p)(m+n)}. \end{aligned} \tag{5.7}$$

This, together with (2.6), yields

$$\mu_p = \frac{1}{c_p} + \frac{p}{n} g_p(c_p) = \frac{1}{c_p} \frac{2(m+n-p)}{m+2n-p} - \frac{(n-p)m}{(m+2n-p)n}$$

$$= \frac{m}{n} \frac{(n-p)(n(m+n-p) + \sqrt{mnp(m+n-p)})}{n(m+p)(m+n-p) - (m+2n-p)\sqrt{mnp(m+n-p)}}$$

$$= \frac{m}{n} \frac{(\sqrt{(m+n-p)n} + \sqrt{mp})^2}{(\sqrt{m(m+n-p)} - \sqrt{np})^2}.$$

By (5.1) one may obtain

$$\frac{\sqrt{(m+n-p)n} + \sqrt{mp}}{\sqrt{m(m+n-p)} - \sqrt{np}} = \frac{\frac{\sqrt{(m+n-p)n}+\sqrt{mp}}{m+n}}{\frac{\sqrt{m(m+n-p)}-\sqrt{np}}{m+n}}$$

$$= \frac{\cos\alpha_p \sin\beta_p + \sin\alpha_p \cos\beta_p}{\cos\alpha_p \cos\beta_p - \sin\alpha_p \sin\beta_p} = \tan(\alpha_p + \beta_p).$$

It follows that

$$\mu_p = \frac{m}{n} \tan^2(\alpha_p + \beta_p), \tag{5.8}$$

which is (5.2).

Using (5.1), (5.2) and the second derivative of $g_p''(x)$ at c_p (2.7) can be rewritten as

$$\frac{1}{\sigma_p^3} = \frac{1}{c_p^3} + \frac{p}{2n} g_p''(c_p)$$

$$= \frac{1}{c_p^3} + \frac{p}{2n}\Big(\frac{1-\frac{p}{m}}{\frac{p}{m}c_p^3} - \frac{\sqrt{(c_p - 1 - \frac{p}{m})^2 - 4\frac{p}{m}}}{\frac{p}{m}c_p^3} - \frac{1+\frac{p}{m}}{\frac{p}{m}c_p^2\sqrt{(c_p - 1 - \frac{p}{m})^2 - 4\frac{p}{m}}} +$$

$$\frac{1}{2\frac{p}{m}c_p\sqrt{(c_p - 1 - \frac{p}{m})^2 - 4\frac{p}{m}}} + \frac{(c_p - \frac{p}{m} - 1)^2}{2\frac{p}{m}c_p((c_p - 1 - \frac{p}{m})^2 - 4\frac{p}{m})^{3/2}}\Big)$$

$$= \cos^2(\beta_p)\cot^3(\beta_p)\csc(\alpha_p)\sec(\alpha_p)\sec^4(\beta_p+\alpha_p)\tan^4(\beta_p+\alpha_p)$$

$$= 16\cos^4(\beta_p)\cot^2(\beta_p)\csc(2\beta_p)\csc(2\alpha_p)\csc^2(2\beta_p+2\alpha_p)\tan^6(\beta_p+\alpha_p)$$

$$= 16\frac{m^2}{(m+n)^2}\frac{m}{n}\frac{1}{\sin(2\beta_p)\sin(2\alpha_p)\sin^2(2\beta_p+2\alpha_p)}\tan^6(\beta_p+\alpha_p)$$

$$= \mu_p^3 \frac{16n^2}{(m+n)^2}\frac{1}{\sin(2\beta_p)\sin(2\alpha_p)\sin^2(2\beta_p+2\alpha_p)},$$

which is (5.7).

When $n = p$, solving (5.6) yields

$$c_p = \frac{(1-\frac{p}{m})^2}{2(1+\frac{p}{m})} = \frac{(m-p)^2}{2(m+p)m}. \tag{5.9}$$

From (5.1) one may conclude that $\alpha_p = \beta_p$. Since

$$\mu_p = \frac{1}{c_p} + \frac{p}{n} g_p(c_p) = \frac{1}{c_p}\frac{2(m+n-p)}{m+2n-p} = \frac{4m^2}{(m-p)^2}$$

and

$$\frac{2\sqrt{mp}}{(m-p)} = \frac{2\frac{\sqrt{mp}}{m+p}}{\frac{m-p}{m+p}} = \frac{\sin(2\alpha_p)}{\cos(2\alpha_p)} = \tan(\alpha_p + \beta_p)$$

we have

$$\mu_p = \frac{m}{n}\tan^2(\alpha_p + \beta_p). \tag{5.10}$$

It's similar to prove

$$\frac{1}{\sigma_p^3} = \mu_p^3 \frac{16n^2}{(m+n)^2} \frac{1}{\sin(2\beta_p)\sin(2\alpha_p)\sin^2(2\beta_p + 2\alpha_p)}. \tag{5.11}$$

The above implies the equivalence between (2.6)-(2.7) and (5.2)-(5.3).

It is straightforward to verify that $|\frac{m}{n}\mu_{J,p} - \mu_p| = O(p^{-1})$ and $\lim_{p\to\infty} \sigma_p \frac{m}{n^{1/3}} \sigma_{J,p} = 1$ according to (5.1)-(5.3) and (2.2). □

6 Proof of Part (ii) of Theorem 2.1: Standard Gaussian Distribution

This section is to consider the case when $\{X_{ij}\}$ follow normal distribution with mean zero and variance one. We below first introduce more notation. Let $\mathbf{A} = (A_{ij})$ be a matrix. We define the following norms

$$\|\mathbf{A}\| = \max_{|\mathbf{x}|=1} |\mathbf{A}\mathbf{x}|, \quad \|\mathbf{A}\|_\infty = \max_{i,j} |A_{ij}|, \quad \|\mathbf{A}\|_F = \sqrt{\sum_{ij} |A_{ij}|^2},$$

where $|\mathbf{x}|$ represents the Euclidean norm of a vector $\mathbf{x}$. Notice that we have a simple relationship among these norms

$$\|\mathbf{A}\|_\infty \le \|\mathbf{A}\| \le \|\mathbf{A}\|_F.$$

We also need the following commonly used definition about stochastic domination to simplify the statements.

Definition 2. *(Stochastic domination) Let*

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

be two families of random variables, where $U^{(n)}$ is a n-dependent parameter set (or independent of n). If for sufficiently small positive ϵ and sufficiently large σ,

$$\sup_{u\in U^{(n)}} \mathbb{P}\left[|\xi^{(n)}(u)| > n^\epsilon |\zeta^{(n)}(u)|\right] \le n^{-\sigma}$$

for large enough $n \ge n(\epsilon, \sigma)$, then we say that ζ stochastically dominates ξ uniformly in u. We denote this relationship by $|\xi| \prec \zeta$ and also write it as $\xi = O_\prec(\zeta)$. Furthermore we also write it as $|x| \prec y$ if x and y are both nonrandom and $|x| \le n^\epsilon |y|$ for sufficiently small positive ϵ.

Proof. We start the proof by reminding readers that $m < p$ and $m + n > p$. Since $m < p$ the limit of the empirical distribution function of $\frac{1}{p}\mathbf{Y}^*\mathbf{Y}$ is the MP law and we denote its density by $\rho_{pm}(x)$. We define $\gamma_{m,1} \geq \gamma_{m,2} \geq \cdots \geq \gamma_{m,m}$ to satisfy

$$\int_{\gamma_{m,j}}^{+\infty} \rho_{pm} dx = \frac{j}{m}, \tag{6.1}$$

with $\gamma_{m,0} = (1 + \sqrt{\frac{m}{p}})^2$,$\gamma_{m,m} = (1 - \sqrt{\frac{m}{p}})^2$. Correspondingly denote the eigenvalues of $\frac{1}{p}\mathbf{Y}^*\mathbf{Y}$ by $\hat{\gamma}_{m,1} \geq \hat{\gamma}_{m,2} \geq \cdots \geq \hat{\gamma}_{m,m}$. Here we would remind the readers that $\rho_{pm}(x), \gamma_{m,j}, \hat{\gamma}_{m,1}$ are similar to those in (2.3), below (2.3) and above (4.3) except that we are interchanging the role of p and m because we are considering $\frac{1}{p}\mathbf{Y}^*\mathbf{Y}$ rather than $\frac{1}{m}\mathbf{Y}\mathbf{Y}^*$. Moreover as in (4.35) and (4.22) for any sufficiently small $\zeta > 0$ and big $D > 0$ there exists an event S_ζ (here with a bit abuse of notion S_ζ) such that

$$S_\zeta = \{\forall j, 1 \leq j \leq m, |\hat{\gamma}_{m,j} - \gamma_{m,j}| \leq p^{\zeta - 2/3}\tilde{j}^{-1/3}\} \tag{6.2}$$

and

$$P(S_\zeta^c) \leq p^{-D}. \tag{6.3}$$

Note that $\frac{1}{p}\mathbf{Y}\mathbf{Y}^*$ and $\frac{1}{p}\mathbf{Y}^*\mathbf{Y}$ have the same nonzero eigenvalues. To simplify notation let $m_p = m + n - p$. Write

$$\frac{1}{p}\mathbf{Y}\mathbf{Y}^* = \mathbf{U}^* \begin{pmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{pmatrix} \mathbf{U}, \tag{6.4}$$

with $\mathbf{D} = diag\{\hat{\gamma}_{m,1}, \hat{\gamma}_{m,2}, \cdots, \hat{\gamma}_{m,m}\}$ and $\mathbf{U}$ is an orthogonal matrix. Then $\det(\lambda\frac{\mathbf{Y}\mathbf{Y}^*}{p} - \frac{\mathbf{X}\mathbf{X}^*}{m_p}) = 0$ is equivalent to

$$\det\left(\lambda \begin{pmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{pmatrix} - \frac{1}{m_p} U^*\mathbf{X}\mathbf{X}^*U\right) = 0.$$

Moreover, since $\{\mathbf{X}_{ij}\}$ are independent standard normal random variables and $\mathbf{U}$ is an orthogonal matrix we have $U\mathbf{X} \overset{d}{=} \mathbf{X}$ so that it suffices to consider the following determinant

$$\det\left(\lambda \begin{pmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{pmatrix} - \frac{1}{m_p}\mathbf{X}\mathbf{X}^*\right) = 0. \tag{6.5}$$

Here $\overset{d}{=}$ means having the identical distribution.

Now rewrite $\mathbf{X}$ as $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$, where $\mathbf{X}_1$ is a $m \times n$ matrix and $\mathbf{X}_2$ is a $(p-m) \times n$ matrix. It follows that

$$\mathbf{X}\mathbf{X}^* = \begin{pmatrix} \mathbf{X}_1\mathbf{X}_1^* & \mathbf{X}_1\mathbf{X}_2^* \\ \mathbf{X}_2\mathbf{X}_1^* & \mathbf{X}_2\mathbf{X}_2^* \end{pmatrix} \triangleq \begin{pmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{X}_{21} & \mathbf{X}_{22} \end{pmatrix}. \tag{6.6}$$

(6.5) can be rewritten as

$$\det\begin{pmatrix} \frac{1}{m_p}\mathbf{X}_{11} - \lambda\mathbf{D} & \frac{1}{m_p}\mathbf{X}_{12} \\ \frac{1}{m_p}\mathbf{X}_{21} & \frac{1}{m_p}\mathbf{X}_{22} \end{pmatrix} = 0.$$

Since $m+n>p$, $\mathbf{X}_{22}$ is invertible. (6.5) is further equivalent to

$$\det(\frac{1}{m_p}\mathbf{X}_{11} - \lambda\mathbf{D} - \frac{1}{m_p}\mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21}) = 0. \tag{6.7}$$

Moreover,

$$\mathbf{X}_{11} - \mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21} = \mathbf{X}_1\mathbf{X}_1^* - \mathbf{X}_1\mathbf{X}_2^*(\mathbf{X}_2\mathbf{X}_2^*)^{-1}\mathbf{X}_2\mathbf{X}_1^* = \mathbf{X}_1(I_n - \mathbf{X}_2^*(\mathbf{X}_2\mathbf{X}_2^*)^{-1}\mathbf{X}_2)\mathbf{X}_1^*.$$

Since $rank(I_n - \mathbf{X}_2^*(\mathbf{X}_2\mathbf{X}_2^*)^{-1}\mathbf{X}_2) = m+n-p = m_p$ we can write

$$\mathbf{I}_n - \mathbf{X}_2^*(\mathbf{X}_2\mathbf{X}_2^*)^{-1}\mathbf{X}_2 = \mathbf{V}\begin{pmatrix} \mathbf{I}_{m_p} & 0 \\ 0 & 0 \end{pmatrix}\mathbf{V}^*.$$

where $\mathbf{V}$ is an orthogonal matrix. In view of the above we can construct a $m \times m_p$ matrix $\mathbf{Z} = (Z_{ij})_{m,m_p}$ consisting of independent standard normal random variables so that

$$\mathbf{X}_{11} - \mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21} \stackrel{d}{=} \mathbf{Z}\mathbf{Z}^*. \tag{6.8}$$

It follows that (6.7) and hence (6.5) are equivalent to

$$\det(\frac{1}{m_p}\mathbf{Z}\mathbf{Z}^* - \lambda\mathbf{D}) = 0. \tag{6.9}$$

It then suffices to consider the largest eigenvalue of $\frac{1}{m_p}\mathbf{D}^{-1}\mathbf{Z}\mathbf{Z}^*$. Denote by λ_1 the largest eigenvalue of $\frac{1}{m_p}D^{-1}ZZ^*$. As in (4.3) and (4.4) define $\hat{c}_m \in [0, \hat{\gamma}_{m,m})$ to satisfy

$$\frac{1}{m}\sum_{j=1}^{m}(\frac{\hat{c}_m}{\hat{\gamma}_{m,j} - \hat{c}_m})^2 = \frac{m_p}{m} \tag{6.10}$$

and $\hat{\mu}_p$ and $\hat{\sigma}_p$ by

$$\hat{\mu}_m = \frac{1}{\hat{c}_m}(1 + \frac{1}{m_p}\sum_{j=1}^{m}(\frac{\hat{c}_m}{\hat{\gamma}_{m,j} - \hat{c}_m})), \quad \frac{1}{\hat{\sigma}_m^3} = \frac{1}{\hat{c}_m^3}(1 + \frac{1}{m_p}\sum_{j=1}^{m}(\frac{\hat{c}_m}{\hat{\gamma}_{m,j} - \hat{c}_m})^3).$$

From Lemma 2 we have on the event S_ζ

$$\limsup_p \frac{\hat{c}_m}{\hat{\gamma}_{m,m}} < 1, \tag{6.11}$$

which implies condition (4.41). It follows from Theorem 1.3 of [3] and Theorem 2.18 of [15] that

$$\lim_{p\to\infty} P(\hat{\sigma}_m(m+n-p)^{2/3}(\lambda_1 - \hat{\mu}_m) \le s) = F_1(s). \tag{6.12}$$

As in the proof of Theorem 2.1, by Lemmas 1 and 2 one may further conclude that

$$\lim_{p\to\infty} P(\sigma_p(m+n-p)^{2/3}(\lambda_1 - \mu_p) \le s) = F_1(s). \tag{6.13}$$

□

7 Proof of Part (ii) of Theorem 2.1: General distributions

The aim of this section is to relax the gaussian assumption on $\mathbf{X}$. We below assume that $\mathbf{X}$ and $\mathbf{Y}$ are real matrices. The complex case can be handled similarly and hence we omit it here. In the sequel, we absorb $\frac{1}{\sqrt{m+n-p}}$ and $\frac{1}{\sqrt{p}}$ into $\mathbf{X}$ and $\mathbf{Y}$ respectively(i.e. $Var(X_{ij}) = \frac{1}{m+n-p}$, $Var(Y_{st}) = \frac{1}{p}$) for convenience.

In terms of the notation in this section ($Var(\mathbf{Y}_{st}) = \frac{1}{p}$), (6.4) can be rewritten as

$$\mathbf{Y}\mathbf{Y}^* = \mathbf{U}^* \begin{pmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{pmatrix} \mathbf{U}.$$

Break $\mathbf{U}$ as $\begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix}$ where $\mathbf{U}_1$ and $\mathbf{U}_2$ are $m \times p$ and $(p-m) \times p$ respectively. By (6.4)-(6.7) (note that here we can not omit $\mathbf{U}$ by $\mathbf{U}\mathbf{X} \overset{d}{=} \mathbf{X}$), the maximum eigenvalue of $\det(\lambda \mathbf{Y}\mathbf{Y}^* - \mathbf{X}\mathbf{X}^*)$ is equivalent to that of the following matrix

$$\begin{aligned} & \mathbf{A} = \mathbf{D}^{-\frac{1}{2}} \mathbf{U}_1 \mathbf{X} (I - \mathbf{X}^T \mathbf{U}_2^T (\mathbf{U}_2 \mathbf{X}\mathbf{X}^T U_2^T)^{-1} U_2 \mathbf{X}) \mathbf{X}^T \mathbf{U}_1^T \mathbf{D}^{-\frac{1}{2}} \\ \triangleq \; & \mathbf{D}^{-\frac{1}{2}} \mathbf{U}_1 \mathbf{X} (I - P_{\mathbf{X}^T U_2^T}) \mathbf{X}^T \mathbf{U}_1^T \mathbf{D}^{-\frac{1}{2}}, \end{aligned} \tag{7.1}$$

where $\mathbf{P}_{\mathbf{X}^T \mathbf{U}_2^T}$ is the projection matrix. It is not necessary to assume that $\mathbf{U}_2 \mathbf{X}\mathbf{X}^T \mathbf{U}_2^T$ is invertible since $P_{\mathbf{X}^T U_2^T}$ is unique even if $(\mathbf{U}_2 \mathbf{X}\mathbf{X}^T \mathbf{U}_2^T)^-$ is the generalized inverse matrix of $\mathbf{U}_2 \mathbf{X}\mathbf{X}^T \mathbf{U}_2^T$. Moreover we indeed have the following lemma to control the smallest eigenvalue of $\mathbf{U}_2 \mathbf{X}\mathbf{X}^T \mathbf{U}_2^T$.

Lemma 3. *Suppose that $(m+n-p)^{\frac{1}{2}} \mathbf{X}$ satisfies Condition 1. Then $\mathbf{U}_2 \mathbf{X}\mathbf{X}^T \mathbf{U}_2^T$ is invertible and*

$$\|(\mathbf{U}_2 \mathbf{X}\mathbf{X}^T \mathbf{U}_2^T)^{-1}\| \leq M \tag{7.2}$$

for a large constant M with high probability. Moreover,

$$\|\mathbf{X}\mathbf{X}^*\| \leq M \tag{7.3}$$

with high probability under conditions in Theorem 2.1.

Proof. One may check that the conditions in Theorem 3.12 in [15] are satisfied when considering $\mathbf{U}_2 \mathbf{X}\mathbf{X}^T \mathbf{U}_2^T$. Applying Theorem 3.12 in [15] then yields

$$|\lambda_{\min}\Big(\mathbf{U}_2 \mathbf{X}\mathbf{X}^T \mathbf{U}_2^T\Big) - (1 - \sqrt{\frac{n}{p-m}})^2| \prec n^{-2/3},$$

where $(1 - \sqrt{\frac{n}{p-m}})^2$ can be obtained when considering the special case when the entries of $\mathbf{X}$ are Gaussian. As for (7.3) see Lemma 3.9 in [7]. □

Since the matrix in (7.1) is quite complicated we construct a linearization matrix for it

$$\mathbf{H} = \mathbf{H}(\mathbf{X}) = \begin{pmatrix} -z\mathbf{I} & 0 & \mathbf{D}^{-1/2}\mathbf{U}_1\mathbf{X} \\ 0 & 0 & \mathbf{U}_2\mathbf{X} \\ \mathbf{X}^T\mathbf{U}_1^T\mathbf{D}^{-1/2} & \mathbf{X}^T\mathbf{U}_2^T & -\mathbf{I} \end{pmatrix}. \tag{7.4}$$

The connection between $\mathbf{H}$ and the matrix in (7.1) is that the upper left block of the 3×3 block matrix $\mathbf{H}^{-1}$ is the Stieltjes transform of (7.1) by simple calculations. We next give the limit of the Stieltjes transform of (7.1) and need the following well-known result (see [1]). There exists a unique solution $m(z) : \mathcal{C}^+ \to \mathcal{C}$ such that

$$\frac{1}{m(z)} = -z + \frac{m}{m+n-p}\int \frac{t}{1+tm(z)} dH_n(t), \tag{7.5}$$

where H_n is the empirical distribution function of $\mathbf{D}^{-1}$. Moreover, we set

$$\underline{m}(z) = -Tr(z(1+m(z)\mathbf{D}^{-1}))^{-1}, \quad \rho(x) = \lim_{z \in \mathcal{C}^+ \to x} \Im m(z).$$

From the end of the last section we see that under the gaussian case (7.1) $\stackrel{d}{=} \mathbf{D}^{-1/2}\mathbf{Z}\mathbf{Z}^*\mathbf{D}^{-1/2}$. Hence it is easy to see that $\hat{\mu}_m$ defined above (6.11) is the right most end point of the support of $\rho(x)$.

For any small positive constant τ we define the domains

$$E(\tau, n) = \{z = E + i\eta \in \mathbb{C}^+ : |z| \geq \tau, |E| \leq \tau^{-1}, n^{-1+\tau} \leq \eta \leq \tau^{-1}\}, \tag{7.6}$$

$$E_+ = E_+(\tau, \tau', n) = \{z \in E(\tau, n) : E \geq \hat{\mu}_m - \tau'\}, \tag{7.7}$$

where τ' is a sufficiently small positive constant.

Set

$$\Psi = \Psi(z) = \sqrt{\frac{\Im m(z)}{n\eta}} + \frac{1}{n\eta}, \ \mathbf{G}(z) = \mathbf{H}^{-1}, \ \mathbf{\Sigma} = \Sigma(z) = z^{-1}(1+m(z)\mathbf{D}^{-1})^{-1}. \tag{7.8}$$

To calculate an explicit expression of $\mathbf{G}(z)$ we need the following well-known formula

$$\begin{pmatrix} \mathbf{K} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{D}^{-1} \end{pmatrix} + \begin{pmatrix} \mathbf{I} \\ -\mathbf{D}^{-1}\mathbf{C} \end{pmatrix} (\mathbf{K} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} \begin{pmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}. \tag{7.9}$$

We next develop the explicit expression of $\mathbf{G}(z)$. Denote the spectral decomposition of $\mathbf{A}$ by

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T = \sum_{k=1}^{m} \lambda_k \mathbf{v}_k \mathbf{v}_k^T,$$

where

$$\lambda_1 \geq ... \geq \lambda_{m+n-p} > 0 = \lambda_{m+n-p+1} = ... = \lambda_m.$$

It follows that

$$G_{ij} = \sum_{k=1}^{m} \frac{\mathbf{v}_k(i)\mathbf{v}_k(j)}{\lambda_k - z}, \quad 1 \leq i, j \leq m, \tag{7.10}$$

where G_{ij} denotes the (i,j)th entry of the matrix $\mathbf{G}(z)$ and $\mathbf{v}_k(i)$ means the ith component of the vector $\mathbf{v}_k$. We denote $(G_{ij})_{1\leq i,j\leq m}$ by $\mathbf{G}_m$, which is the same as $(\mathbf{A} - z\mathbf{I})^{-1}$, the green function of (7.1). Moreover, let

$$\mathbf{A}_2 = \begin{pmatrix} \mathbf{I} & -\mathbf{D}^{-1/2}\mathbf{U}_1\mathbf{X}\mathbf{X}^T\mathbf{U}_2^T\Gamma & \mathbf{D}^{-1/2}\mathbf{U}_1\mathbf{X}(\mathbf{I} - \mathbf{P}_{\mathbf{X}^T\mathbf{U}_2^T}) \end{pmatrix}^T,$$

and

$$\mathbf{A}_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \mathbf{\Gamma} & \Gamma\mathbf{U}_2\mathbf{X} \\ 0 & \mathbf{X}^T\mathbf{U}_2^T\Gamma & -I + P_{\mathbf{X}^T\mathbf{U}_2^T} \end{pmatrix},$$

where $\mathbf{\Gamma} = (\mathbf{U}_2\mathbf{X}\mathbf{X}^T\mathbf{U}_2^T)^{-1}$. Applying (7.9) twice implies that

$$\mathbf{G}(z) = \mathbf{A}_3 + \sum_{k=1}^{m} \frac{\mathbf{A}_2\mathbf{v}_k\mathbf{v}_k^T\mathbf{A}_2^T}{\lambda_k - z} = \mathbf{A}_3 + \mathbf{A}_2\mathbf{G}_m\mathbf{A}_2^T. \tag{7.11}$$

To control the inverse of a matrix in the projection matrix we introduce the following smooth cutoff function

$$\mathcal{X}(x) = \begin{cases} 1 & \text{if } |x| \leq M_1 n^{-2} \\ 0 & \text{if } |x| \geq 2M_1 n^{-2}, \end{cases}$$

whose derivatives satisfy $|\mathcal{X}^{(k)}| \leq Mn^{2k}$, k=1,2,... and M_1 is some positive constant. Let $\tilde{\lambda}_1 \geq ... \geq \tilde{\lambda}_{p-m}$ be the eigenvalues of $\mathbf{U}_2\mathbf{X}\mathbf{X}^T\mathbf{U}_2^T$ and $\underline{s}(z)$ be the Stieltjes transform of its ESD. Since

$$\Im(\underline{s}(in^{-2})) = (p-m)^{-1} \sum_{i=1}^{p-m} \frac{n^{-2}}{\tilde{\lambda}_i^2 + n^{-4}}, \tag{7.12}$$

we conclude that

$$\text{if } |\Im(\underline{s}(in^{-2}))| \leq M_1 n^{-2}, \text{ then } \tilde{\lambda}_{p-m} \geq \frac{M_2}{n} \tag{7.13}$$

for some positive constant M_2, which allows us to control the maximum eigenvalue of $(\mathbf{U}_2\mathbf{X}\mathbf{X}^T\mathbf{U}_2^T)^{-1}$ outside the event $\{\tilde{\lambda}_{p-m} \geq c\}$ with some positive constant c. Moreover, consider the event $\{\tilde{\lambda}_{p-m} \geq c\}$. By Lemma 3, choosing a sufficient small constant c, we have

$$1 - o(n^{-l}) = \mathbb{P}(\tilde{\lambda}_{p-m} \geq c) \leq \mathbb{P}(\Im(\underline{s}(in^{-2})) \leq M_1 n^{-2}), \quad \text{for any positive integer } l. \tag{7.14}$$

Therefore, by Lemma 3 we have

$$\mathbb{P}(\mathcal{X}(\Im(\underline{s}(in^{-2}))) \neq 1) \leq o(n^{-l}), \quad \text{for any positive integer } l. \tag{7.15}$$

Similarly, by Lemma 3, for $\|\mathbf{X}\|_F^2$, we have

$$\mathbb{P}(\mathcal{X}(n^{-3}\|\mathbf{X}\|_F^2) \neq 1) \leq o(n^{-l}), \quad \text{for any positive integer } l. \tag{7.16}$$

Set $\mathcal{T}_n(X) = \mathcal{X}(\Im(\underline{s}(in^{-2}))\mathcal{X}(n^{-3}\|\mathbf{X}\|_F^2)$, and

$$\mathbf{F}(z) = \begin{pmatrix} -\mathbf{\Sigma} & \mathbf{\Sigma}\mathbf{D}^{-1/2}\mathbf{U}_1\mathbf{X}\mathbf{X}^T\mathbf{U}_2^T\mathbf{\Gamma} & 0 \\ \mathbf{\Gamma}\mathbf{U}_2\mathbf{X}\mathbf{X}^T\mathbf{U}_1^T\mathbf{D}^{-1/2}\Sigma & \mathbf{\Gamma} - \mathbf{\Gamma}\mathbf{U}_2\mathbf{X}\mathbf{X}^T\mathbf{U}_1^T\mathbf{D}^{-1/2}\mathbf{\Sigma}\mathbf{D}^{-1/2}\mathbf{U}_1\mathbf{X}\mathbf{X}^T\mathbf{U}_2^T\mathbf{\Gamma} & \mathbf{\Gamma}\mathbf{U}_2\mathbf{X} \\ 0 & \mathbf{X}^T\mathbf{U}_2^T\Gamma & (zm(z)+1)(\mathbf{I} - \mathbf{P}_{\mathbf{X}^T\mathbf{U}_2^T}) \end{pmatrix}. \tag{7.17}$$

In fact, $\mathbf{F}(z)$ is close to $\mathbf{G}(z)$ with high probability. In view of (7.15) and (7.16) it is straight forward to see that

$$\mathcal{T}_n(X) = 1 \tag{7.18}$$

with high probability and we will use it frequently without mention.

We are now in a position to state our main result about the local law near $\hat{\mu}_m$, the right end point of the support of the limit of the ESD of $\mathbf{A}$ in (7.1).

Theorem 7.1. *(Strong local law) Suppose that $(m+n-p)^{\frac{1}{2}}\mathbf{X}$ and $p^{\frac{1}{2}}\mathbf{Y}$ satisfy the conditions of Theorem 2.1. Then*

(i) For any deterministic unit vectors $\mathbf{v}$, $\mathbf{w} \in \mathbb{R}^{p+n}$

$$\langle \mathbf{v}, (\mathbf{G}(z) - \mathbf{F}(z))\mathbf{w} \rangle \prec \Psi \tag{7.19}$$

uniformly $z \in E_+$ and

(ii)

$$|\underline{m}_n(z) - \underline{m}(z)| \prec \frac{1}{n\eta} \tag{7.20}$$

uniformly in $z \in E_+$, where $\underline{m}_n(z) = \frac{1}{m}\sum_{i=1}^m G_{ii}$.

7.1 Local law (7.19)

The aim of this subsection is to prove (7.19). Before proving (7.19) we first collect some frequently used bounds below. Recall the definition of $m(z)$ in (7.5). For $z \in E(\tau, n)$ one may verify that

$$M_2 \leq |m(z)| \leq M_1 \tag{7.21}$$

and

$$Im(m(z)) \geq M\eta. \tag{7.22}$$

(see Lemma 2.3 in [4] or Lemma 3.1 and Lemma 3.2 in [20]). Order the eigenvalues of $\mathbf{D}^{-1}$ as $d_1 \geq d_2 \geq \cdots \geq d_m$. From (6.10) and (6.11) we conclude that on the event S_ξ defined in (6.2)

$$\limsup_p \hat{c}_m d_1 < 1. \tag{7.23}$$

Here we remind the readers that d_1 corresponds to $\frac{1}{\gamma_{m,m}}$ there, validity of (7.23) does not depend on the Gaussian assumption there and we do not assume the entries of $\mathbf{Y}$ to be Gaussian in the last section. In addition, with probability one

$$\hat{c}_m = -\lim_{z \in \mathcal{C}^+ \to \hat{\mu}_m} m(z), \tag{7.24}$$

(one may see below (1.8) in [4] or [20]). It follows from (7.23) and (7.24) that for $z \in E_+$ on the event S_ξ

$$|1 + dm(z)| \geq \tau_2, \quad d \in [d_m, d_1] \tag{7.25}$$

for some positive constant τ_2 (one may also see (iv) of Lemma 2.3 of [4]). We then conclude (7.21) and (7.25) that on the event S_ξ

$$\|\mathbf{\Sigma}\| = \|\mathbf{\Sigma}(z)\| \leq M, \quad d \in [d_m, d_1], \tag{7.26}$$

where $\mathbf{\Sigma} = \mathbf{\Sigma}(z)$ is defined in (7.8). Moreover, for $z \in E_+$ it follows from Lemma 3, (7.3), and (7.21)-(7.25) that

$$\|\mathbf{F}(z)\| \prec 1, \quad \|\mathbf{A}_2\| \prec 1, \quad \|\mathbf{A}_3\| \prec 1. \tag{7.27}$$

We further introduce more notations with bold lower index

$$\mathbf{G}_{\mathbf{v}s} = \langle \mathbf{v}, \mathbf{G}\mathbf{e}_s \rangle, \quad \mathbf{G}_{\mathbf{vw}} = \langle \mathbf{v}, \mathbf{G}\mathbf{w} \rangle, \quad \text{and} \quad \mathbf{G}_{s\mathbf{v}} = \langle \mathbf{e}_s, \mathbf{G}\mathbf{v} \rangle,$$

where $\mathbf{e}_s$ is the unit vector with the s-th coordinate equal to 1. In the sequel, if the lower index of a matrix is bold, then it represents the inner product above and otherwise it means one entry of the corresponding matrix. Fix $\tau > 0$. For any $z \in E(\tau, n)$ we claim that

$$\|\mathbf{G}(z)\mathcal{T}_n(\mathbf{X})\| \leq Cn^{10}\eta^{-1}, \quad \|\partial_z \mathbf{G}(z)\mathcal{T}_n(\mathbf{X})\| \leq Cn^{10}\eta^{-2}, \tag{7.28}$$

$$\|\mathbf{G}(z)\| \prec \eta^{-1}, \quad \|\partial_z \mathbf{G}(z)\| \prec \eta^{-2}, \tag{7.29}$$

$$\sum_{i=1}^m |\mathbf{G}_{\mathbf{v}i}|^2 = \frac{\Im \mathbf{G}_{\mathbf{vv}}}{\eta}, \quad \|\mathbf{F}(z)\mathcal{T}_n(\mathbf{X})I(S_\xi)\| \leq Cn^4\eta^{-1} \tag{7.30}$$

and

$$|G_{\mathbf{vv}}|^2 \prec \frac{\Im G_{\mathbf{vv}}}{\eta} + 1, \tag{7.31}$$

where and in what follows $I(\cdot)$ denotes an indicator function. Indeed, the estimates (7.28) follow from (7.11) and the definition of $\mathcal{T}_n(\mathbf{X})$ directly. (7.29) and (7.31) about the partial order follow from Lemma 3, (7.3) and (7.11). The first equality in (7.30) is straightforward and the second one is from the definition of $\mathcal{T}_n(\mathbf{X})$ directly.

When the entries of $\mathbf{X}$ are Gaussian distributed Theorem 7.1 can be obtained from by Theorem 2.10 of [7]. Indeed, from (7.11) one can see a key observation that each block of $\mathbf{G}(z)$ can be represented as a linear combination of the blocks of (4.3) in [15] in the Gaussian case. We now demonstrate such an observation by looking at three block matrices of $\mathbf{G}(z)$ and other blocks can be checked similarly. For example, the upper left block of the 3×3 block $\mathbf{G}(z)$ is $(\mathbf{A} - z\mathbf{I})^{-1}$ or $\mathbf{G}_m$ (see (7.1) for the definition of $\mathbf{A}$). From the end of the last section we see that

$$\mathbf{A} \stackrel{d}{=} \mathbf{D}^{-1/2}\mathbf{Z}\mathbf{Z}^*\mathbf{D}^{-1/2} \tag{7.32}$$

under the gaussian case. Therefore $(\mathbf{A} - z\mathbf{I})^{-1}$ is just one block of (4.3) in [15]. A second block matrix of $\mathbf{G}(z)$ is $\mathbf{G}_m\mathbf{T}$ with $\mathbf{T} = \mathbf{D}^{-1/2}\mathbf{U}_1\mathbf{X}(\mathbf{I} - \mathbf{P}_{\mathbf{X}^T\mathbf{U}_2^T})$. From the end of the last section and (7.32) we also see that $\mathbf{G}_m = (\mathbf{T}\mathbf{T}^* - z\mathbf{I})^{-1}$ due to $(\mathbf{I} - \mathbf{P}_{\mathbf{X}^T\mathbf{U}_2^T})$ is a projection matrix so that this block is also one of the block in (4.3) in [15]. A third block matrix of $\mathbf{G}(z)$ is $\mathbf{G}_m\mathbf{D}^{-1/2}\mathbf{U}_1\mathbf{X}\mathbf{X}^T\mathbf{U}_2^T\mathbf{\Gamma}$. Note that $\mathbf{D}^{-1/2}\mathbf{U}_1\mathbf{X}\mathbf{X}^T\mathbf{U}_2^T\mathbf{\Gamma}$ is independent of $\mathbf{G}_m$ given $\mathbf{U}_2\mathbf{X}$ and $\mathbf{D}$ by noting that $(\mathbf{I} - \mathbf{P}_{\mathbf{X}^T\mathbf{U}_2^T})\mathbf{X}^T\mathbf{U}_2^T = 0$. It follows that this block can be regarded as a product of random $\mathbf{G}_m$ and a non-random matrix given $\mathbf{U}_2\mathbf{X}$ and $\mathbf{D}$. So the local law holds for this block from Theorem 2.10 of [7] by absorbing the nonrandom matrix into the fixed vector $\mathbf{v}$ or $\mathbf{w}$ (note that (7.25) is required in the conditions of Theorem 2.10 of [7]).

7.1.1 Proving (7.19) for general distributions

We next prove (7.19) for general distributions by fixing $\mathbf{Y}$ first since $\mathbf{X}$ and $\mathbf{Y}$ are independent (the dominated convergence theorem then ensures (7.19)). However to simplify notations we drop the statements about conditioning on $\mathbf{Y}$ as well as the event S_ξ. In other words, whenever we come across expectations they should be understood as conditional expectations and involve $I(S_\xi)$. For example, (7.38) below should be understood as follows

$$\mathbb{E}\Big(|F_{ab}(\mathbf{X}, z)|^{2q} I(S_\xi)\Big|\mathbf{Y}\Big) \leq (n^{24\delta}\Psi)^{2q}.$$

In order to prove Theorem 7.1, it suffices to show that for any deterministic orthogonal matrices $\mathbf{V}_1$ and $\mathbf{V}_2$, we have

$$\|\mathbf{V}_1(\mathbf{G}(z) - \mathbf{F}(z))\mathbf{V}_2^T\|_\infty \prec \Psi, \tag{7.33}$$

for all $z \in E_+$. We define S to be a ϵ-net of $E(\tau, n)$ with $\epsilon = n^{-10}$ and the cardinality of S, $|S|$, not bigger than n^{30}. Note that the function $\mathbf{D}^{1/2}(\mathbf{G}(z)-\mathbf{F}(z))\mathbf{D}^{1/2}$ is Lipschitz continuous with respect to the operation norm in E_+ and the Lipschitz constant is $Mn^2\|\mathbf{X}\mathbf{X}^*\| + Mn^2\|1/\lambda_{\min}(\mathbf{U}_2\mathbf{X})\|$. By (7.3) it then suffices to focus on S to prove Theorem 7.1 by Lemma 3.

Following [7] the main idea of the proof is an induction argument from bigger imaginary parts to smaller imaginary parts. Set δ to be a sufficient small positive constant such that $n^{24\delta}\Psi \ll 1$. For any given $\eta \geq \frac{1}{n}$, we define a sequence of numbers $\eta_0 \leq \eta_1 \leq \eta_2 ... \leq \eta_L$ with

$$\eta_l = \eta n^{l\delta}, \quad (l = 0, 1, ..., L-1), \quad \eta_L = 1, \tag{7.34}$$

where

$$L \equiv L(\eta) = \max\{l \in \mathbb{N} : \eta n^{l\delta} < n^{-\delta}\}.$$

One can see that $L \leq \delta^{-1} + 1$ by the definition. From now on we will work on the net S containing the points $E + i\eta_l \in S, \quad l = 0, ..., L$. Moreover define $S_k = \{z \in S : \Im z \geq n^{-\delta k}\}$ and sequence of properties

$$B_k = \{\|\mathbf{V}_1(\mathbf{G}(z) - \mathbf{F}(z))\mathbf{V}_2^T\|_\infty \prec 1, \quad \text{for any z } \in S_k\} \tag{7.35}$$

$$C_k = \{\|\mathbf{V}_1(\mathbf{G}(z) - \mathbf{F}(z))\mathbf{V}_2^T\|_\infty \prec n^{24\delta}\Psi, \quad \text{for any z } \in S_k\}. \tag{7.36}$$

We start the induction by considering property B_0. We claim that the property B_0 holds. Indeed we conclude from (7.11) and(7.27) that

$$\|\mathbf{V}_1(\mathbf{G}(z) - \mathbf{F}(z))\mathbf{V}_2^T\|_\infty \prec \|\mathbf{G}_m(z)\| + \|\mathbf{F}(z)\| + 1 \prec 1,$$

as claimed. Moreover it's easy to see that property C_k implies property B_k by the choice of δ such that $n^{24\delta}\Psi \ll 1$. We next prove that property B_{k-1} implies property C_k for any $1 \leq k \leq \delta^{-1}$. If this is true then the induction is complete and (7.33) holds for all $z \in S$.

To this end, we calculate the higher moments of the following function

$$F_{ab}(\mathbf{X}, z) = \left((\mathbf{J}_1\mathbf{G}(z)\mathbf{J}_2^T)_{ab} - (\mathbf{J}_1\mathbf{F}(z)\mathbf{J}_2^T)_{ab}\right)\mathcal{T}_n(\mathbf{X}), \tag{7.37}$$

where $\mathbf{J}_1, \mathbf{J}_2 \in \mathcal{L} = \{1, \mathbf{\Delta}, \mathbf{V}\}$, $\mathbf{\Delta}$ is defined in (7.51) below and $\mathbf{V}$ is any deterministic orthogonal matrix. Lemma 4 below, Markov's inequality and (7.18) then ensure that property B_{k-1} implies property C_k.

Lemma 4. *Let q be a positive constant and $k \leq \delta^{-1}$. Suppose that property B_{k-1} in (7.35) holds. Then*

$$\mathbb{E}\Big(|F_{ab}(\mathbf{X}, z)|^{2q}\Big) \leq (n^{24\delta}\Psi)^{2q}, \tag{7.38}$$

for all $1 \leq a, b \leq n + p$ and $z \in S_k$.

The proof will be complete if we prove Lemma 4. Before proceeding, we present a simple but frequently used lemma which can help us transfer the partial order of two random variables to the partial order of the expectations.

Lemma 5. *Let ζ be a random variable satisfying $\zeta \prec \nu$ where positive ν may be random or deterministic. Suppose $|\zeta| \le n^{M_0}$ for some positive constant M_0. Then*

$$\mathbb{E}\zeta \prec (E\nu + n^{M_0 - D}), \tag{7.39}$$

where D is a sufficiently large positive constant.

Proof. Since $\zeta \prec \nu$ there exists a sufficiently small positive ϵ and sufficiently large D so that

$$P(\zeta \ge n^{\epsilon}\nu) \le n^{-D}.$$

Define the event $A_\epsilon = \{\zeta \le n^{\epsilon}\nu\}$. Write

$$|\mathbb{E}\zeta| = \left|\mathbb{E}\zeta I(A_\epsilon) + \mathbb{E}\zeta I(A_\epsilon^c)\right| \le n^{\epsilon}\mathbb{E}\nu + n^{M_0}P(A_\epsilon^c) \le n^{\epsilon}\mathbb{E}\nu + n^{M_0 - D}.$$

□

We now claim that

$$\mathbb{E}\Big(|F_{ab}(\mathbf{X}^0, z)|^{2q}\Big) \le (n^{24\delta}\Psi)^{2q}, \tag{7.40}$$

if $\mathbf{X}$ in Lemma 4 is replaced by the corresponding Gaussian random matrix $\mathbf{X}^0 = (X_{i\mu}^0) = \mathbf{X}^{Gauss}$ consisting of Gaussian random variables with mean zero and variance one. Indeed, one can see that $|F_{ab}(\mathbf{X}^0, z)|^{2q} \prec \Psi^{2q}$ from the paragraph containing (7.32). To apply (7.39) to conclude the claim we need $|F_{ab}(\mathbf{X}^0, z)| \le n^{M_0}$, which follows immediately from the first estimate in (7.28) and the second estimate in (7.30).

7.1.2 Proving Lemma 4 by the interpolation method

We next finish Lemma 4 for the general distributions by the interpolation method developed by [15]. To this end we need to define the interpolation matrix $\mathbf{X}^t$ between $\mathbf{X}^1 = (X_{i\mu}^1) = \mathbf{X}$ and $\mathbf{X}^0$. For $1 \le i \le p$ and $1 \le \mu \le n$, denote the distribution function of the random variables $X_{i\mu}^u$ by $F_{i\mu}^u$ for $u = 0, 1$. For $t \in [0, 1]$, we define the interpolated distribution function by

$$F_{i\mu}^t = tF_{i\mu}^1 + (1 - t)F_{i\mu}^0. \tag{7.41}$$

Define the interpolation matrix $\mathbf{X}^t = (X_{i\mu}^t)$ with $F_{i\mu}^t$ being the distribution of $X_{i\mu}^t$ and $\{X_{i\mu}^t\}$ are independent for i, μ. We furthermore introduce the matrix

$$\mathbf{X}_{(i\mu)}^{t,\lambda} = \mathbf{X}^t + (\lambda - X_{i\mu}^t)\mathbf{e}_i\mathbf{e}_\mu^T, \tag{7.42}$$

which differs from $\mathbf{X}^t$ at the (i, μ) position only. We also define $\mathbf{G}^t(z) = \mathbf{G}(\mathbf{X}^t, z)$ and $\mathbf{G}_{(i\mu)}^{t,\lambda}(z) = \mathbf{G}(\mathbf{X}_{(i\mu)}^{t,\lambda}, z)$, the analogues of $\mathbf{G}(z)$ defined above (7.9), by replacing the random matrix $\mathbf{X}$ in $\mathbf{G}(z)$ with $\mathbf{X}^t$ and $\mathbf{X}_{(i\mu)}^{t,\lambda}$ respectively.

We now need the following interpolation formula and one may see Lemma 6.9 of [15].

Lemma 6. *For any function $F : \mathbb{R}^{p\times n} \to \mathbb{C}$, we have*

$$\mathbb{E}F(\mathbf{X}^1) - \mathbb{E}F(\mathbf{X}^0) = \int_0^1 dt \sum_{i=1}^{p}\sum_{\mu=1}^{n}\left[\mathbb{E}F(\mathbf{X}_{(i\mu)}^{t,\mathbf{X}_{(i\mu)}^1}) - \mathbb{E}F(\mathbf{X}_{(i\mu)}^{t,\mathbf{X}_{(i\mu)}^0})\right]. \tag{7.43}$$

To handle the right hand side of (7.43) we establish the following Lemma.

Lemma 7. *Fix an positive integer q and $k \le \delta^{-1}$. Suppose that property B_{k-1} holds. Then there exists some function $g_{ab}(.,z)$ such that for $t \in [0,1], u \in \{0,1\}, z \in S_k$*

$$\sum_{i=1}^{p}\sum_{\mu=1}^{n}\left[\mathbb{E}\Big(|F_{ab}(\mathbf{X}_{(i\mu)}^{t,X_{i\mu}^u}, z)|^{2q}\Big) - \mathbb{E}|g_{ab}(\mathbf{X}_{(i\mu)}^{t,0}, z)|^{2q}\right] = O((n^{24\delta}\Psi)^{2q} + \|\mathbb{E}\mathbf{L}(\mathbf{X}^t, z)\|_\infty), \tag{7.44}$$

with the matrix $\mathbf{L}(\mathbf{X}^t, z) = \Big(|F_{ab}(\mathbf{X}^t, z)|^{2q}\Big)_{1\le a,b\le n+p}$.

Lemma 7 immediately implies that for $z \in S_k$

$$\sum_{i=1}^{p}\sum_{\mu=1}^{n}\left[\mathbb{E}\Big(|F_{ab}(\mathbf{X}_{(i\mu)}^{t,X_{i\mu}^1}, z)|^{2q}\Big) - \mathbb{E}\Big(|F_{ab}(\mathbf{X}_{(i\mu)}^{t,X_{i\mu}^0}, z)|^{2q}\Big)\right] = O((n^{24\delta}\Psi)^{2q} + \|\mathbb{E}L(\mathbf{X}^t, z)\|_\infty). \tag{7.45}$$

To apply the above results we need the following Gronnwall's inequality.

Lemma 8. *Suppose that $\beta(t)$ is nonnegative and continuous and $u(t)$ is continuous. If for any $t \in \mathbb{R}$, $\alpha(t)$ is nondecreasing and $u(t)$ satisfies the following equality*

$$u(t) \le \alpha(t) + \int_0^t \beta(s)u(s)ds,$$

then

$$u(t) \le \alpha(t)\exp\Big(\int_0^t \beta(s)ds\Big).$$

To apply Gronnwall's inequality it is observed that

$$\frac{\partial}{\partial t}\Big(\max_{1\le s,t\le n+p}\mathbb{E}|F_{ab}(\mathbf{X}^t, z)|^{2q}\Big) \le \max_{1\le s,t\le n+p}\frac{\partial}{\partial t}\mathbb{E}|F_{ab}(\mathbf{X}^t, z)|^{2q}.$$

From (7.43) and (7.45) we see that

$$\frac{\partial\mathbb{E}|F_{ab}(\mathbf{X}^t, z)|^{2q}}{\partial t} = O((n^{24\delta}\Psi)^{2q} + \|\mathbb{E}\mathbf{L}(\mathbf{X}^t, z)\|_\infty),$$

if F in (7.43) is taken as $|F_{ab}(\cdot, z)|^{2q}$. Gronnwall's inequality and (7.40) imply that

$$\frac{\partial\mathbb{E}|F_{ab}(\mathbf{X}^t, z)|^{2q}}{\partial t} \le M(n^{24\delta}\Psi)^{2q} + M\Big(\max_{1\le s,t\le n+p}\mathbb{E}|F_{ab}(\mathbf{X}^0, z)|^{2q}\Big) \le M(n^{24\delta}\Psi)^{2q}$$

This, together with Lemma 6 and (7.40), implies that Lemma 4 holds. Similarly for future use we would point out that if $n^{24\delta}\Psi$ in (7.44) is replaced by $n^\delta\Psi^2$ and (7.40) is strengthened to

$$\mathbb{E}\Big(|F_{ab}(\mathbf{X}^0, z)|^{2q}\Big) \le (n^\delta\Psi^2)^{2q} \tag{7.46}$$

then

$$\mathbb{E}\Big(|F_{ab}(\mathbf{X},z)|^{2q}\Big) \le (n^{\delta}\Psi^2)^{2q}, \tag{7.47}$$

if the real part of z is outside the support.

What remains is to prove Lemma 7 and we below consider the case $u = 1$ only ($u = 0$ is similar). We first develop a crude bound below so that we may use property B_{k-1} in (7.35), which is the assumption of Lemma 7.

Lemma 9. *Suppose that property B_{k-1} holds. Then for any unit vector $\mathbf{v}$ and $\mathbf{w}$*

$$\langle \mathbf{v}, (\mathbf{G}(z) - \mathbf{F}(z))\mathbf{w}\rangle = O_{\prec}(n^{2\delta})$$

for all $z \in S_k$.

Proof. Recall the definition of η_l in (7.34). Note that $z_l = E + i\eta_l \in S_k$ for l=1,2,...,L when $z = E + i\eta \in S_{k-1}$. Hence (7.35) ensures that

$$\Im G_{\mathbf{vv}}(E + i\eta_l) \prec |\mathbf{v}|^2 + \Im\langle \mathbf{v}, \Pi(E + i\eta_l)\mathbf{v}\rangle \prec |\mathbf{v}|^2,$$

where the last $\prec$ follows from (7.27). We conclude the proof by Lemma 10 below. □

Lemma 10. *For any $z \in S$ and $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^{p+n}$, we have*

$$\langle \mathbf{x}, (\mathbf{G}(z) - \mathbf{F}(z))\mathbf{y}\rangle \prec n^{2\delta} \sum_{l=1}^{L(\eta)} (\Im G_{\mathbf{xx}}(E + i\eta_l) + \Im G_{\mathbf{yy}}(E + i\eta_l)) + |\mathbf{x}||\mathbf{y}|.$$

Proof. The proof of this lemma follows that of Lemma 6.12 in [15] closely. It follows from (7.11) and (7.27) that

$$\langle \mathbf{x}, (\mathbf{G}(z) - \mathbf{F}(z))\mathbf{y}\rangle \prec \sum_{k=1}^{m} \frac{\langle \mathbf{x}, A_2\mathbf{v}_k\rangle^2}{|\lambda_k - z|} + \sum_{k=1}^{m} \frac{\langle \mathbf{y}, A_2\mathbf{v}_k\rangle^2}{|\lambda_k - z|} + |\mathbf{x}||\mathbf{y}|.$$

We evaluate the first term below and the second term can be handled similarly. We introduce the indices subsets

$$\mathcal{C}_l = \{k : \eta_{l-1} \le |\lambda_k - E| < \eta_l\}, \quad (l = 0, 1, ..., L+1),$$

where $\eta_{-1} = 0$ and $\eta_{L+1} = \infty$ so that we can rewrite the first term as follows.

$$\sum_{k=1} \frac{\langle \mathbf{x}, A_2\mathbf{v}_k\rangle^2}{|\lambda_k - z|} = \sum_{l=0}^{L+1} \sum_{k\in\mathcal{C}_l} \frac{\langle \mathbf{x}, A_2\mathbf{v}_k\rangle^2}{|\lambda_k - z|}.$$

Consider the inner sum for $l \in \{1, 2, ..., L\}$,

$$\begin{aligned}
&\sum_{k\in U_l} \frac{\langle \mathbf{x}, A_2\mathbf{v}_k\rangle^2}{|\lambda_k - z|} \le \sum_{k\in\mathcal{C}_l} \frac{\langle \mathbf{x}, A_2\mathbf{v}_k\rangle^2\eta_l}{(\lambda_k - E)^2} \le 2\sum_{k\in\mathcal{C}_l} \frac{\langle \mathbf{x}, A_2\mathbf{v}_k\rangle^2\eta_l}{(\lambda_k - E)^2 + \eta_{l-1}^2} \\
&\le 2\frac{\eta_l}{\eta_{l-1}} \Im G_{\mathbf{xx}}(E + i\eta_{l-1}) \le 2n^{\delta}\Im G_{\mathbf{xx}}(E + i\eta_{l-1}).
\end{aligned} \tag{7.48}$$

Combining with the fact that $y\Im G_{\mathbf{xx}}(E+iy)$ is nondecreasing function of y, we have

$$\sum_{k\in U_l}\frac{\langle \mathbf{x}, A_2\mathbf{v}_k\rangle^2}{|\lambda_k - z|} \le 2n^{2\delta}\Im G_{\mathbf{xx}}(E+i\eta_{l\vee 1}).$$

Next, we consider the cases l=0 and l=L+1.

$$\sum_{k\in\mathcal{C}_0}\frac{\langle \mathbf{x}, A_2\mathbf{v}_k\rangle^2}{|\lambda_k - z|} \le 2\sum_{k\in\mathcal{C}_0}\frac{\langle \mathbf{x}, A_2\mathbf{v}_k\rangle^2\eta}{(\lambda_k-E)^2+\eta^2} \le 2\Im G_{\mathbf{xx}}(E+i\eta) \le 2n^{\delta}\Im G_{\mathbf{xx}}(E+i\eta_1).$$

$$\sum_{k\in\mathcal{C}_{L+1}}\frac{\langle \mathbf{x}, A_2\mathbf{v}_k\rangle^2}{|\lambda_k - z|} \le 2\sum_{k\in\mathcal{C}_{l+1}}\frac{\langle \mathbf{x}, A_2\mathbf{v}_k\rangle^2|\lambda_k-E|\eta_L}{(\lambda_k-E)^2+\eta_L^2} \prec \sum_{k\in\mathcal{C}_{L+1}}\frac{\langle \mathbf{x}, A_2\mathbf{v}_k\rangle^2\eta_L}{(\lambda_k-E)^2+\eta_L^2} \le \Im G_{\mathbf{xx}}(E+i\eta_L),$$

where we also use (7.3). □

It is observed that Lemma 9 holds for the interpolation random matrix $\mathbf{X}^t$ as well because from (7.41) one can see that the entries of $\mathbf{X}^t$ are independent random variables with mean zero, variance one and finite moment. Recall the definitions of $\mathbf{J}_i, i=1,2$ in (7.37). It follows that

$$\|\mathbf{J}_1(\mathbf{G}^t(z)-\mathbf{F}(z))\mathbf{J}_2^T\|_\infty \prec n^{2\delta}, \quad \text{for z in } S_k. \tag{7.49}$$

Below we further generalize it so that (7.49) still holds even if any entry $X^t_{i\mu}$ of $\mathbf{G}^t(z)$ is replaced by any other random variable of size not bigger than $n^{-1/2}$. From (7.42) write

$$\mathbf{X}^{t,\lambda_1}_{(i\mu)} - \mathbf{X}^{t,\lambda_2}_{(i\mu)} = (\lambda_1-\lambda_2)\mathbf{e}_i\mathbf{e}_\mu^T.$$

This, together with (7.4), yields that

$$\mathbf{H}(\mathbf{X}^{t,\lambda_1}_{(i\mu)}) - \mathbf{H}(\mathbf{X}^{t,\lambda_2}_{(i\mu)}) = \mathbf{\Delta}^{\lambda_1-\lambda_2}_{(i\mu)}, \tag{7.50}$$

where $\mathbf{H}(\mathbf{X}^{t,\lambda_1}_{(i\mu)})$ is obtained from $\mathbf{H}(\mathbf{X})$ in (7.4) with $\mathbf{X}$ replaced by $\mathbf{X}^{t,\lambda}_{(i\mu)}$ and

$$\mathbf{\Delta}^{\lambda}_{(i\mu)} = \lambda\Big(\mathbf{e}_{\mu+p}\mathbf{e}_i^T\mathbf{\Delta} + \mathbf{\Delta}^T\mathbf{e}_i\mathbf{e}_{\mu+p}^T\Big), \quad \mathbf{\Delta} = \begin{pmatrix} \mathbf{U}_1^T\mathbf{D}^{-1/2} & \mathbf{U}_2^T & 0\end{pmatrix}, \tag{7.51}$$

where and in the following $\mathbf{e}_{\mu+p}$ is always $(n+p)\times 1$ and $\mathbf{e}_i$ is $p\times 1$. Applying the formula $\mathbf{A}^{-1}-\mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B}-\mathbf{A})\mathbf{B}^{-1}$ repeatedly we further obtain the following resolvent formula for any $H\in\mathbb{N}$,

$$\mathbf{G}^{t,\lambda_1}_{(i\mu)} = \mathbf{G}^{t,\lambda_2}_{(i\mu)} + \sum_{h=1}^{H}(-1)^h\mathbf{G}^{t,\lambda_2}_{(i\mu)}\Big(\mathbf{\Delta}^{\lambda_1-\lambda_2}_{(i\mu)}\mathbf{G}^{t,\lambda_2}_{(i\mu)}\Big)^h + (-1)^{H+1}\mathbf{G}^{t,\lambda_1}_{(i\mu)}\Big(\mathbf{\Delta}^{\lambda_1-\lambda_2}_{(i\mu)}\mathbf{G}^{t,\lambda_2}_{(i\mu)}\Big)^{H+1}, \tag{7.52}$$

recalling the definition of $\mathbf{G}^{t,\lambda_1}_{(i\mu)}$ below (7.42). Here and below we drop the variable z when there is no confusion but one should keep in mind that $z\in S_k$.

Lemma 11. *Suppose that λ is a random variable and satisfies $|\lambda|\prec n^{-1/2}$. Then*

$$\|\mathbf{J}_1(\mathbf{G}^{t,\lambda}_{(i\mu)} - \mathbf{F})\mathbf{J}_2^T\|_\infty \prec n^{2\delta}. \tag{7.53}$$

Proof. Recall (7.27)

$$\|\mathbf{F}\| \prec 1.$$

It is easy to see that

$$\|\mathbf{\Delta}\| \leq M,$$

which implies that

$$\|\mathbf{\Delta}^{\lambda}_{(i\mu)}\| \leq M\lambda. \tag{7.54}$$

We next apply (7.52) with $\lambda_1 = \lambda$, $H = 11$ and $\lambda_2 = X^t_{i\mu}$ so that $\mathbf{G}^{t,\lambda_2}_{(i\mu)} = \mathbf{G}^t$. We conclude from (7.49) that

$$\|\mathbf{J}_1\mathbf{G}^{t,\lambda_2}_{(i\mu)}\| + \|\mathbf{G}^{t,\lambda_2}_{(i\mu)}\mathbf{J}_2^T\| \prec n^{2\delta}.$$

Note that $|\lambda_1 - \lambda_2| \prec n^{-1/2}$. Similar to the first inequality in (7.29), $\mathbf{G}^{t,\lambda_1}_{(i\mu)}$ can be bounded by the imaginary part of z, i.e. $\mathbf{G}^{t,\lambda_1}_{(i\mu)} = O_{\prec}(n)$. Summarizing the above we conclude Lemma 11. □

In order to simplify the notations, recalling (7.37) we define

$$f_{(i\mu)}(\lambda) = |F_{ab}(\mathbf{X}^{t,\lambda}_{(i\mu)})|^{2q} = \left(F_{st}(\mathbf{X}^{t,\lambda}_{(i\mu)})\overline{F_{ab}(\mathbf{X}^{t,\lambda}_{(i\mu)})}\right)^q,$$

where we omit some parameters. By Lemma 11 and (7.52) one can easily get the following Lemma.

Lemma 12. *Suppose that λ is a random variable and satisfies $|\lambda| \prec n^{-1/2}$. Then for any fixed integer k we have*

$$|f^{(k)}_{(i\mu)}(\lambda)| \prec n^{2\delta(2q+k)}, \tag{7.55}$$

where $f^{(k)}_{(i\mu)}(\lambda)$ denotes the kth derivative of $f_{(i\mu)}(\lambda)$ with respect to λ.

From Taylor's expansion and (7.55) when $|\lambda| \prec n^{-1/2}$ we have

$$f_{(i\mu)}(\lambda) = \sum_{k=0}^{8q} \frac{\lambda^k}{k!} f^{(k)}_{(i\mu)}(0) + O_{\prec}(\Psi^{2q}). \tag{7.56}$$

It follows from Lemma 12 and (7.39) that

$$\begin{aligned} &\mathbb{E}|F_{ab}(\mathbf{X}^{t,X^1_{i\mu}}_{(i\mu)})|^{2q} - \mathbb{E}|F_{ab}(\mathbf{X}^{t,0}_{(i\mu)})|^{2q} = \mathbb{E}f_{(i\mu)}(X^1_{i\mu}) - \mathbb{E}f_{(i\mu)}(0) \\ &= \frac{1}{2(m+n-p)}\mathbb{E}f^{(2)}_{(i\mu)}(0) + \sum_{k=4}^{8q}\frac{1}{k!}\mathbb{E}f^{(k)}_{(i\mu)}(0)\mathbb{E}(X^1_{i\mu})^k + O_{\prec}(\Psi^{2q}), \end{aligned} \tag{7.57}$$

where we use $\mathbb{E}(X^1_{i\mu})^k = 0, k = 1, 3$. To show (7.44), it suffices to prove that

$$n^{-k/2}\sum_{i=1}^{p}\sum_{\mu=1}^{n}\mathbb{E}f^{(k)}_{(i\mu)}(0) = O((n^{24\delta}\Psi)^{2q} + \|\mathbb{E}|F(\mathbf{X}^t)|^{2q}\|_{\infty}), \tag{7.58}$$

for k=4,...,8q. At this moment we would like to point out that $\mathbb{E}|g_{ab}(\mathbf{X}^{t,0}_{(i\mu)})|^{2q}$ in (7.44) equals

$$\mathbb{E}|F_{ab}(\mathbf{X}^{t,0}_{(i\mu)})|^{2q}+\frac{1}{2(m+n-p)}\mathbb{E}f^{(2)}_{(i\mu)}(0).$$

We will not prove (7.58) directly. Instead we will prove the following claim in order to obtain a self-consistent estimation of $\mathbf{X}^t$. We claim that if

$$n^{-k/2}\sum_{i=1}^{p}\sum_{\mu=1}^{n}\mathbb{E}f^{(k)}_{(i\mu)}(X^t_{i\mu})=O((n^{24\delta}\Psi)^{2q}+\|\mathbb{E}|F(\mathbf{X}^t)|^{2q}\|_\infty), \tag{7.59}$$

is true for k=4,...,16q, then (7.58) holds for k=4,...,8q. Indeed, in order to apply (7.59) to prove (7.58) we denote $f_{(i\mu)}$ and $X^t_{i\mu}$ by f and X respectively for simplicity. Similar to (7.57), by (7.55) we have

$$\mathbb{E}f^{(l)}(0)=\mathbb{E}f^{(l)}(X)-\sum_{k=1}^{16q-l}\mathbb{E}f^{(l+k)}(0)\frac{\mathbb{E}X^k}{k!}+O_\prec(n^{l/2-1/2-8q+40\delta q}). \tag{7.60}$$

It follows from (7.60) that

$$\begin{aligned}
\mathbb{E}f^{(k)}(0) &= \mathbb{E}f^{(k)}(X)-\sum_{\substack{k_1\geq 1\\ k+k_1\leq 16q}}\mathbb{E}f^{(k+k_1)}(0)\frac{\mathbb{E}X^{k_1}}{k_1!}+O_\prec(n^{k/2-1/2-8q+40\delta q})\\
&= \mathbb{E}f^{(k)}(X)-\sum_{\substack{k_1\geq 1\\ k+k_1\leq 16q}}\mathbb{E}f^{(k+k_1)}(X)\frac{\mathbb{E}X^{k_1}}{k_1!}\\
&\quad+\sum_{\substack{k_1,k_2\geq 1\\ k+k_1+k_2\leq 16q}}\mathbb{E}f^{(k+k_1+k_2)}(0)\frac{\mathbb{E}X^{k_1}}{k_1!}\frac{\mathbb{E}X^{k_2}}{k_2!}+O_\prec(n^{k/2-1/2-8q+40\delta q})\\
&= \ldots=\sum_{r=0}^{16q-k}(-1)^r\sum_{\substack{k_1,k_2,\ldots,k_r\geq 1\\ k+\sum k_i\leq 16q}}\mathbb{E}f^{(k+\sum k_i)}(X)\prod_i\frac{\mathbb{E}X^{k_i}}{k_i!}+O_\prec(n^{k/2-1/2-8q+40\delta q}).
\end{aligned}$$

This, together with (7.22) and the definition of Ψ in (7.8), implies (7.58) immediately, as claimed.

It then suffices to prove (7.59). Recall that

$$f^{(k)}_{(i\mu)}(X^t_{i\mu})=\frac{\partial^k\left(|F_{ab}(\mathbf{X}^{t,X^t_{i\mu}}_{(i\mu)})|^{2q}\right)}{\partial(X^t_{i\mu})^k}, \tag{7.61}$$

where $F_{st}(\cdot)$ is given in (7.37). Since $\mathbf{X}^t=\mathbf{X}^{t,X^t_{i\mu}}_{(i\mu)}$ is the only matrix we focus on we below use $\mathbf{X}=(X_{i\mu})$ instead of $\mathbf{X}^t=(X^t_{i\mu})$ to simplify notation because the entries of both of them have bounded higher moments. To prove (7.59) we need to study (7.61).

7.1.3 Estimate of higher order derivatives (7.61) in (7.59)

We first look at the higher order derivatives of $(\mathbf{J}_1\mathbf{F}(z)\mathbf{J}_2^T)_{ab}$ with respect to $\mathbf{X}_{i\mu}$. Noting that $\mathbf{F}(z)$ is a 3×3 block matrix we need to analyze the derivatives of $(\mathbf{J}_1\mathbf{F}(z)\mathbf{J}_2^T)_{ab}$ block by block. It turns

out that the higher order derivatives of $(\mathbf{J}_1\mathbf{F}(z)\mathbf{J}_2^T)_{ab}$ are quite complicated even if we analyze them block by block. Fortunately, as will be seen, the exact expressions of the higher order derivatives of $(\mathbf{J}_1\mathbf{F}(z)\mathbf{J}_2^T)_{ab}$ are not important. Moreover we claim an important fact that the higher order derivatives of $\mathbf{F}(z)$ with respect to $\mathbf{X}_{i\mu}$ can be generated by some sum or products of (part of) common matrices $\mathbf{U}_1, \mathbf{U}_2, \mathbf{\Sigma}, \mathbf{e}_i\mathbf{e}_\mu^T, \mathbf{e}_\mu\mathbf{e}_i^T, \mathbf{X}, \Gamma(\mathbf{X})$ (we call these common matrices atoms). Indeed, recalling $\Gamma(\mathbf{X}) = (U_2\mathbf{X}\mathbf{X}^TU_2^T)^{-1}$ simple calculations indicate that

$$\frac{\partial \mathbf{X}\mathbf{X}^T}{\partial \mathbf{X}_{i\mu}} = \mathbf{X}\mathbf{e}_\mu\mathbf{e}_i^T + \mathbf{e}_i\mathbf{e}_\mu^T\mathbf{X}^T, \ \frac{\partial \Gamma(\mathbf{X})}{\partial \mathbf{X}_{i\mu}} = -\Gamma(\mathbf{X})(U_2\mathbf{X}\mathbf{e}_\mu\mathbf{e}_i^TU_2^T + U_2\mathbf{e}_i\mathbf{e}_\mu\mathbf{X}^TU_2^T)\Gamma(\mathbf{X}). \tag{7.62}$$

It's easy to see that the first derivative of each block of $\mathbf{F}(z)$ with respect to $\mathbf{X}_{i\mu}$ can be constructed by sum or products of these atoms. Assuming that the kth derivative of each block of $\mathbf{F}(z)$ is constructed by these atoms we find that the $(k+1)$th derivative of each block of $\mathbf{F}(z)$ is also constructed by these atoms by (7.62). Based on the above fact we can describe the higher order derivatives of $(\mathbf{J}_1\mathbf{F}(z)\mathbf{J}_2^T)_{ab}$ easier. By dropping $\mathbf{e}_i\mathbf{e}_\mu^T$ and $\mathbf{e}_\mu\mathbf{e}_i^T$ from the atoms we define the set

$$\mathcal{Q}(k) = \{\text{The matrices constructed from sum or product of (part of) } \ \mathbf{U}_1, \mathbf{U}_2, \mathbf{X}, \mathbf{\Sigma}, \ \Gamma(\mathbf{X})\}. \tag{7.63}$$

Any kth order derivative of each block of $\mathbf{F}(z)$ with respect to $\mathbf{X}_{i\mu}$ belongs to some product(s) between some matrices in $\mathcal{Q}(k)$ and $\mathbf{e}_i\mathbf{e}_\mu^T$ or $\mathbf{e}_\mu\mathbf{e}_i^T$.

Lemma 3 and (7.3) imply that $\|\Gamma(\mathbf{X})\| \leq M$ and $\|\mathbf{X}\mathbf{X}^*\| \leq M$ with high probability. Recalling (7.26), in view of the arguments above we conclude that for any $\mathbf{Q} \in \mathcal{Q}(k)$,

$$\|\mathbf{Q}\| \prec 1 \tag{7.64}$$

and the cardinality of $\mathcal{Q}(k)$ satisfies $|\mathcal{Q}(k)| \leq M(k)$, where M(k) is a constant depending on k. Moreover, for the function $\mathcal{T}_n(\mathbf{X})$, if $\mathcal{T}_n(\mathbf{X})$ is differentiated, then by simple and tedious calculations, from the definition of the smooth cutoff function, (7.15) and (7.16) we have

$$\left|D_{i\mu}^j\mathcal{T}_n(\mathbf{X})\right| \prec 0 \tag{7.65}$$

and

$$\left|\mathbb{E}D_{i\mu}^j\mathcal{T}_n(\mathbf{X})\right| \leq n^{-l} \tag{7.66}$$

for any positive integer l and sufficient large n. The above properties about $\mathcal{T}_n(\mathbf{X})$ and the matrices belonging to $\mathcal{Q}(k)$ are enough for our proof below and we don't need to investigate the precise expression.

We next look at the higher order derivatives of $(\mathbf{J}_1\mathbf{G}(z)\mathbf{J}_2^T)_{ab}$ with respect to $\mathbf{X}_{i\mu}$. To characterize its higher order derivative conveniently we define group g of size k to be the set of paired indices:

$$g = \{a_1b_1, a_2b_2, \cdots, a_{k+1}b_{k+1}\},$$

where each of $\{a_j, b_j, j = 1, \cdots, k+1\}$ equals one of four letters $s, t, i, (\mu+p)$. Here we would remind readers that the size of group g is defined to be k instead of (k+1) in order to simplify the argument below. Denote the size of the group g by $k = k(g)$ and introduce the set $\mathfrak{G}_k = \{g : \ k(g) = k\}$ consisting of groups of size k. Moreover, we require each group in $\mathfrak{G}_k$ to satisfy three conditions specified below:

(i) $a_1 = a$ and $b_{k+1} = b$.

(ii) For $l \in [2, k+1]$ we have $a_l \in \{i, \mu + p\}$ and $b_{l-1} \in \{i, \mu + p\}$.

(iii) For $k \in [1, k]$ we have $b_{l-1}a_l \in \{i(\mu + p), (\mu + p)i\}$.

As will be seen, groups g are connected with the high order derivatives of $(\mathbf{J}_1\mathbf{G}(z)\mathbf{J}_2^T)_{ab}$.

Moreover write $\mathbf{F}(z) = \sum_{j=1}^{7} \mathbf{F}_j(z)$ where each $\mathbf{F}_j(z)$ corresponds to a non-zero block of $\mathbf{F}(z)$. As before, to characterize the higher order derivative of each block conveniently we define groups $g^{(j)}$ of size k to be the set of paired indices:

$$g^{(j)} = \{a_{j1}b_{j1}, a_{j2}b_{j2}, \cdots, a_{j(k+1)}b_{j(k+1)}\},$$

where each s_{jm} and t_{jm} equals s, t, i, μ. Moreover introduce the set $\mathfrak{G}_{jk} = \{g^{(j)} : \ k(g^{(j)}) = k\}$ consisting of groups of size k. We require each group in $\mathfrak{G}_{jk}$ to satisfy conditions:

(i) $a_{j1} = a$ and $b_{j(k+1)} = b$.

(ii) For $l \in [2, k+1]$ we have $a_{jl} \in \{i, \mu\}$ and $b_{j(l-1)} \in \{i, \mu\}$.

(iii) For $k \in [1, k]$ we have $b_{j(l-1)}a_{jl} \in \{i\mu, \mu i\}$.

As will be seen groups $g^{(j)}$ are linked to the high order derivatives of $(\mathbf{J}_1\mathbf{F}(z)\mathbf{J}_2^T)_{ab}$.

We below associate a random variable $B_{a,b,i,\mu}(g, g^{(1)}, \cdots, g^{(7)})$ with each group $g, g^{(j)}, j = 1, \cdots, 7$. When $k(g) = k(^{(j)}) = 0$ we define

$$B_{a,b,i,\mu}(g, g^{(1)}, \cdots, g^{(7)})) = (\mathbf{J}_1\mathbf{G}(z)\mathbf{J}_2^T)_{ab} - (\mathbf{J}_1\mathbf{F}(z)\mathbf{J}_2^T)_{ab}.$$

When $k(g) \geq 1$ and $k(g^{(j)}) \geq 1$, define

$$B_{a,b,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,7k+1}}(g, g^{(1)}, ..., g^{(7)}) = C_{a,b,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,7k+1}}(g, g^{(1)}, \cdots, g^{(7)})) \tag{7.67}$$

$$- \sum_{j=1}^{7} (\mathbf{J}_1\mathcal{R}_{j1})_{(a_{j1}b_{j1})}(\mathcal{R}_{j2})_{(a_{j2}b_{j2})}...(\mathcal{R}_{jk})_{(a_{jk}b_{jk})}(\mathcal{R}_{jk+1}\mathbf{J}_2^T)_{(a_{jk+1}b_{jk+1})},$$

with

$$C_{a,b,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,7k+1}}(g, g^{(1)}, \cdots, g^{(7)})) = (\mathbf{J}_1 G\mathbf{A}_5)_{(a_1b_1)}(\mathbf{R}_2)_{(a_2b_2)}...(\mathbf{R}_k)_{(a_kb_k)}(\mathbf{A}_4\mathbf{G}\mathbf{J}_2^T)_{(a_{k+1}b_{k+1})}, \tag{7.68}$$

where $\mathbf{R}_j (2 \le j \le n)$ has the expression of $\mathbf{R}_j = \mathbf{A}_4 \mathbf{G} \mathbf{A}_5$ with $\mathbf{A}_4 \in \{1, \mathbf{\Delta}\}$, $\mathbf{A}_5 \in \{1, \mathbf{\Delta}^T\}$ and the non-zero block $\mathcal{R}_{jl}$ belongs to $\mathcal{Q}(k)$ in (7.63). Moreover the selection of 1 and $\mathbf{\Delta}$ in $\mathbf{A}_4$ and $\mathbf{A}_5$ is subject to the constraint that the total number of $\mathbf{\Delta}$ and $\mathbf{\Delta}^T$ contained in $B_{a,b,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,7k+1}}(g, g^{(1)}, ..., g^{(7)})$ is k. One should also notice that if $k(g) = 1$, the terms R_j will disappear. It follows from (7.64) that

$$\|\mathcal{R}_{jl}\| \prec 1. \tag{7.69}$$

It is easy to see that

$$\frac{\partial \mathbf{G}}{\partial X_{i\mu}} = -\mathbf{G}(\mathbf{e}_{\mu+p}\mathbf{e}_i^T \mathbf{\Delta} + \mathbf{\Delta}^T \mathbf{e}_i \mathbf{e}_{\mu+p}^T)\mathbf{G}, \tag{7.70}$$

(one may see (7.50) for the derivative). We first demonstrate how to apply the above definitions about groups $g^{(j)}$ and $B_{a,b,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,7k+1}}(g, g^{(1)}, ..., g^{(7)})$ and hence write

$$\begin{aligned} &\frac{\partial^k}{\partial (X_{i\mu})^k}\Big([(\mathbf{J}_1 \mathbf{G}(z) \mathbf{J}_2^T)_{ab} - (\mathbf{J}_1 \mathbf{F}(z) \mathbf{J}_2^T)_{ab}]\mathcal{T}_n(\mathbf{X})\Big) \\ &= (-1)^k \sum_{\substack{g \in \mathfrak{G}_k, g^{(j)} \in \mathfrak{G}_{jk} \\ \mathbf{R}_i, i=2,\ldots,k \\ \mathcal{R}_{jl}, j=1,..7, l=1,\ldots,k+1}} B_{a,b,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,7k+1}}(g, g^{(1)}, ..., g^{(7)})\mathcal{T}_n(\mathbf{X}) + O_{\prec}(0), \end{aligned} \tag{7.71}$$

where the term $O_{\prec}(0)$ comes from the derivative on $\mathcal{T}_n(\mathbf{X})$ by (7.65), (7.29) and (7.27). To simplify the notations, we furthermore omit $\mathbf{R}_{2\cdots,k}, \mathcal{R}_{11,\ldots,7k+1}, g^{(1)}, ..., g^{(7)}$ in the sequel and write

$$B_{a,b,i,\mu}(g) = B_{a,b,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,7k+1}}(g, g^{(1)}, ..., g^{(7)}), \tag{7.72}$$

$$C_{a,b,i,\mu}(g) = C_{a,b,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,7k+1}}(g, g^{(1)}, ..., g^{(7)}), \tag{7.73}$$

(here one should notice that the sizes of g and $g^{(j)}$ are the same according to definition (7.67)). More generally we furthermore have

$$\begin{aligned} &\frac{\partial^k}{\partial (X_{i\mu})^k}\Big(|F_{ab}(\mathbf{X})|^{2q}\Big) = (-1)^k \sum_{\substack{k_1,\ldots,k_q,\tilde{k}_1,\ldots,\tilde{k}_q \in \mathbb{N} \\ \sum_r (k_r + \tilde{k}_r) = k}} \frac{k!}{\prod_r k_r! \tilde{k}_r!} \\ &\times \prod_{r=1}^{q} \Big(\sum_{\substack{g_r \in \mathfrak{G}_{k_r} \cup \mathfrak{G}_{jk_r} \\ \mathbf{R}_i, i=2,\ldots,k \\ \mathcal{R}_{jl}, j=1,..7, l=1,\ldots,k+1}} \sum_{\substack{\tilde{g}_r \in \mathfrak{G}_{\tilde{k}_r} \cup \mathfrak{G}_{j\tilde{k}_r} \\ \bar{\mathbf{R}}_i, i=2,\ldots,k \\ \bar{\mathcal{R}}_{jl}, j=1,..7, l=1,\ldots,k+1}} B_{a,b,i,\mu}(g_r)\overline{B_{a,b,i,\mu}(\tilde{g}_r)}\mathcal{T}_n^2(\mathbf{X})\Big) + O_{\prec}(0), \end{aligned} \tag{7.74}$$

where $g_r \in \mathfrak{G}_{k_r} \cup \mathfrak{G}_{jk_r}$ means that the groups associated with the derivatives of $\mathbf{G}(z)$ belong to $\mathfrak{G}_{k_r}$ and the groups associated with the derivatives of $\mathbf{F}(z)$ belong to $\mathfrak{G}_{jk_r}$. In view of (7.74) and (7.61) to prove (7.59) it then suffices to show that

$$n^{-k/2} \sum_{i=1}^{p} \sum_{\mu=1}^{n} \mathbb{E}\left[\prod_{r=1}^{q} B_{a,b,i,\mu}(g_r)\overline{B_{a,b,i,\mu}(\tilde{g}_r)}\mathcal{T}_n^{2q}(\mathbf{X})\right] = O((n^{24\delta}\Psi)^{2q} + \|\mathbb{E}|F(\mathbf{X})|^{2q}\|_\infty), \tag{7.75}$$

for $4 \le k \le 16q$ and groups $g_r \in \mathfrak{G}_{k_r} \cup \mathfrak{G}_{jk_r}$, $\tilde{g}_r \in \mathfrak{G}_{\tilde{k}_r} \cup \mathfrak{G}_{j\tilde{k}_r}$ satisfying $\sum_r (k(g_r) + \tilde{k}(g_r)) = k$. To simplify notations, we drop complex conjugates (which will complicate the notations but the proof

is the same) from the left hand side of (7.75). Without loss of generality, suppose there are (2q-l) terms such that $k(g_r)=0$ and denote each of them by g_0. (7.75) reduces to

$$n^{-k/2}\sum_{i=1}^{p}\sum_{\mu=1}^{n}\mathbb{E}\left[B_{a,b,i,\mu}(g_0)^{2q-l}\prod_{r=1}^{l}B_{a,b,i,\mu}(g_r)\mathcal{T}_n^{2q}(\mathbf{X})\right]=O((n^{24\delta}\Psi)^{2q}+\|\mathbb{E}|F(\mathbf{X})|^{2q}\|_\infty), \quad (7.76)$$

for $4\le k\le 16q$ and groups $g_r\in\mathfrak{G}_{k_r}\cup\mathfrak{G}_{jk_r}$ satisfying $\sum_r k(g_r)=k$ and $k(g_0)=0$.

To estimate the left hand of (7.76), we introduce the notations

$$\mathcal{H}_i=\mathcal{H}_{1i}+\mathcal{H}_{abi},\quad \mathcal{H}_{1i}=|(\mathbf{J}_1\mathbf{G}\mathbf{\Delta})_{ai}|+|(\mathbf{\Delta}^T\mathbf{G}\mathbf{J}_2^T)_{ib}|,\quad \mathcal{H}_{abi}=\sum_{\mathcal{R}\in\mathcal{Q}(k)}(|(\mathbf{J}_1\mathcal{R})_{ai}|+|(\mathcal{R}\mathbf{J}_2^T)_{ib}|),$$

$$\mathcal{H}_\mu=\mathcal{H}_{1\mu}+\mathcal{H}_{a\mu},\quad \mathcal{H}_{1\mu}=|(\mathbf{J}_1\mathbf{G})_{a(\mu+p)}|+|(\mathbf{G}\mathbf{J}_2^T)_{(\mu+p)b}|,\quad \mathcal{H}_{a\mu}=\sum_{\mathcal{R}\in\mathcal{Q}(k)}(|(\mathbf{J}_1\mathcal{R})_{a\mu}|+|(\mathcal{R}\mathbf{J}_2^T)_{\mu a}|,$$

where the lower indices i and μ at $\mathbf{J}_1\mathcal{R}$ and $\mathcal{R}\mathbf{J}_2^T$ respectively represent the index i, $i+p-m$, $i+p$, and μ, $\mu+p-n$ or $\mu+p$ depending on which block we consider (or differentiate). By (7.53), (7.27) and (7.64) we have

$$\mathcal{H}_i+\mathcal{H}_\mu\prec n^{2\delta}. \quad (7.77)$$

Moreover for $g_r\in\mathfrak{G}_{k_r}\cup\mathfrak{G}_{jk_r}$, we similarly obtain from (7.53), (7.27), (7.64) and definition (7.67) that

$$|B_{a,b,i,\mu}(g_r)|\prec n^{2\delta(k(g)+1)}, \quad (7.78)$$

(recall $k(g)=k(g^{(j)})$ from definition (7.67)). Likewise, for $k(g)\ge 1$, we have

$$|B_{a,b,i,\mu}(g_r)|\prec(\mathcal{H}_i^2+\mathcal{H}_\mu^2)n^{2\delta(k(g_r)-1)}, \quad (7.79)$$

while k(g)=1,

$$|B_{a,b,i,\mu}(g_r)|\prec\mathcal{H}_i\mathcal{H}_\mu. \quad (7.80)$$

When $k\le 2l-2$ there must exist at least 2 g_r's satisfying $k(g_r)=1$ because $\sum_{r=1}^{l}k(g_r)=k\le 2l-2$. It follows from (7.78) and (7.80) that

$$\begin{aligned}|B_{a,b,i,\mu}(g_0)^{2q-l}\prod_{r=1}^{l}B_{a,b,i,\mu}(g_r)| \;&\prec\; n^{2\delta(k+l)}F_{ab}^{2q-l}(\mathbf{X})\Big(I(k\ge 2l-1)(\mathcal{H}_i^2+\mathcal{H}_\mu^2)\\ &+\; I(k\le 2l-2)\mathcal{H}_i^2\mathcal{H}_\mu^2\Big). \quad (7.81)\end{aligned}$$

Recalling the notation $\mathbf{\Delta}$ in (7.51) we have $\|\mathbf{\Delta}\mathbf{\Delta}^T\|\le M$. In view of (7.64) it is easy to see that

$$\sum_{i=1}^{p}\mathcal{H}_{1i}^2+\sum_{\mu=1}^{n}\mathcal{H}_{1\mu}^2\prec n\phi_a^2+n\phi_b^2, \quad (7.82)$$

$$\sum_{i \text{ or } a \text{ or } b}^{p} \mathcal{H}_{abi}^2 + \sum_{a \text{ or } \mu}^{n} \mathcal{H}_{a\mu}^2 \prec 1, \tag{7.83}$$

where i or a or b means the summation over either i or a or b and

$$\phi_a^2 = \frac{\Im(\mathbf{JGJ}^*)_{aa} + \eta}{n\eta},$$

with $\mathbf{J} \in \mathcal{L}$ defined in (7.37). This implies that

$$\sum_{i=1}^{p} \mathcal{H}_i^2 + \sum_{\mu=1}^{n} \mathcal{H}_\mu^2 \prec n\phi_a^2 + n\phi_b^2. \tag{7.84}$$

From (7.22) and (7.25)

$$\phi_a^2 = \frac{\Im(\mathbf{JFJ}^*)_{aa} + \Im(\mathbf{J}(\mathbf{G}-\mathbf{F})\mathbf{J}^*)_{aa} + \eta}{n\eta} \prec \frac{\Im m + \Im(\mathbf{J}(\mathbf{G}-\mathbf{F})\mathbf{J}^*)_{aa}}{n\eta}.$$

Recalling the definition of Ψ in (7.8) we conclude that

$$\phi_a^2 \prec \Psi(\Psi + F_{aa}(\mathbf{X})). \tag{7.85}$$

By (7.13), (7.26), (7.28), (7.30), the definition of $\mathcal{T}_n(\mathbf{X})$ and definition (7.67) we have

$$|B_{a,b,i,\mu}(g_0)B_{a,b,i,\mu}(g_r)\mathcal{T}_n(\mathbf{X})| \le n^{M_0}. \tag{7.86}$$

From (7.81), (7.84), (7.86) and (7.39) the left hand side of (7.76) is bounded in absolute value by

$$n^{-k/2+2}n^{3\delta(k+l)}\mathbb{E}F_{ab}^{2q-l}(\mathbf{X})\Big[I(k \ge 2l-1)(\phi_a^2+\phi_b^2) + I(k \le 2l-2)(\phi_a^4+\phi_b^4)\Big] + n^{-D}. \tag{7.87}$$

Set

$$F_1^{2q} = F_{aa}^{2q} + F_{ba}^{2q} + F_{ab}^{2q}.$$

We conclude from (7.85)-(7.86) and (7.39) that the left hand side of (7.76) is bounded in absolute value by

$$\begin{cases} n^{3\delta(k+l)}\left(\Psi^{k-2}\mathbb{E}F_1^{2q-l}(\mathbf{X}) + \Psi^{k-3}\mathbb{E}F_1^{2q-l+1}(\mathbf{X})\right) + n^{-D}, & \text{if } \ k \ge 2l-1, \\ n^{3\delta(k+l)}\left(\Psi^{k}\mathbb{E}F_1^{2q-l}(\mathbf{X}) + \Psi^{k-2}\mathbb{E}F_1^{2q-l+2}(\mathbf{X})\right) + n^{-D}, & \text{if } \ k \le 2l-2. \end{cases} \tag{7.88}$$

Since $l \le k$ (7.88) is further bounded by

$$\begin{cases} (n^{24\delta}\Psi)^{k-2}\mathbb{E}F_1^{2q-l}(\mathbf{X}) + (n^{24\delta}\Psi)^{k-3}\mathbb{E}F_1^{2q-l+1}(\mathbf{X}) + n^{-D}, & \text{if } \ k \ge 2l-1, \\ (n^{24\delta}\Psi)^{k}\mathbb{E}F_1^{2q-l}(\mathbf{X}) + (n^{24\delta}\Psi)^{k-2}\mathbb{E}F_1^{2q-l+2}(\mathbf{X}) + n^{-D}, & \text{if } \ k \le 2l-2. \end{cases} \tag{7.89}$$

This ensures that the left hand side of (7.88) is bounded in absolute value by

$$(n^{24\delta}\Psi)^{l}\mathbb{E}F_1^{2q-l}(\mathbf{X}) + (n^{24\delta}\Psi)^{l-1}\mathbb{E}F_1^{2q-l+1}(\mathbf{X}) + (n^{24\delta}\Psi)^{l-2}I(l \ge 3)\mathbb{E}F_1^{2q-l+2}(\mathbf{X}) + n^{-D}, \tag{7.90}$$

where we use the facts that $k \geq l+2$ when $k \geq 4$ and $k \geq 2l-1$ and that $k \geq l$ and $l \geq 3$ when $k \leq 2l-2$ and $k \geq 4$. When $l \geq 2$, (7.76) follows from (7.90), the facts that $(E|X|^r)^{1/r}$ is a nondecreasing function of r and that $n^{-D} \leq (n^{24\delta}\Psi)^{2q}$ for sufficiently large D. For example

$$(n^{24\delta}\Psi)^{l-2}\mathbb{E}F_1^{2q-l+2}(\mathbf{X}) \leq \Big(\mathbb{E}\Big(F_1^{2q-l+2}(\mathbf{X})\Big)^{\frac{2q}{2q-l+2}}\Big)^{\frac{2q-l+2}{2q}}\Big((n^{24\delta}\Psi)^{2q}\Big)^{\frac{l-2}{2q}} \tag{7.91}$$

$$\leq \Big(\mathbb{E}\Big(F_1^{2q}(\mathbf{X})\Big) + (n^{24\delta}\Psi)^{2q}\Big)^{\frac{2q-l+2+l-2}{2q}}.$$

When $l=1$, the first term can be handled similarly and the second term directly implies (7.76). Thus we have proved (7.19) in Theorem 7.1.

7.2 Local law (7.20)

This subsection is to prove (7.20) in Theorem 7.1, i.e.

$$|\underline{m}_n(z) - \underline{m}(z)| \prec \frac{1}{n\eta}. \tag{7.92}$$

As pointed out in the paragraph containing (7.32), (7.92) holds when the underlying distribution of X_{ij} of $\mathbf{X}$ is the standard Gaussian distribution. Moreover, we need to use the interpolation method to prove (7.20) for the general distributions as in proving (7.19). However we do not need induction on the imaginary part of z unlike before due to existence of (7.19).

In order to prove (7.92) it suffices to show that

$$|\underline{m}_n(z) - \underline{m}(z)|\mathcal{T}_n(\mathbf{X}) \prec \frac{1}{n\eta}. \tag{7.93}$$

As in (7.37) we introduce the notation $\hat{F}^{2q}(\mathbf{X}, z)$ as follows

$$\hat{F}^{2q}(\mathbf{X}, z) = |\underline{m}_n(z) - \underline{m}(z)|^{2q}\mathcal{T}_n^{2q}(\mathbf{X}) = |\frac{1}{m}\sum_k^m G_{kk}(z) - \underline{m}(z)|^{2q}\mathcal{T}_n^{2q}(\mathbf{X}).$$

Checking on Lemmas 4, 6, 7, (7.45) and (7.59) in the last section we only need to show

$$n^{-k/2}\sum_{i=1}^p\sum_{\mu=1}^n \mathbb{E}\Big[(\frac{\partial}{\partial \mathbf{X}_{i\mu}})^k \hat{F}^{2q}(\mathbf{X}, z)\Big] = O((n^{\delta}\Psi^2)^{2q} + \|\hat{F}^{2q}(\mathbf{X}, z)\|_\infty),\ k \geq 4 \tag{7.94}$$

where δ is sufficiently small so that n^δ is smaller than n^ε before (7.93) due to the definition of the partial order. Applying the definition of $B_{a,b,i,\mu}$ in the previous section with $\mathbf{J}_1 = \mathbf{J}_2 = 1$ and $a = b = k$, it suffices to show that

$$n^{-k/2}\sum_{i=1}^p\sum_{\mu=1}^n \mathbb{E}\prod_{h=1}^{2q}\Big[\frac{1}{m}\sum_{k=1}^m B_{k,k,i,\mu}(g_h)\Big] = O((n^{\delta}\Psi^2)^{2q} + \|\mathbb{E}\hat{F}^{2q}(\mathbf{X}, z)\|_\infty). \tag{7.95}$$

Notice that (7.19) holds uniformly for any unit determinant vectors $\mathbf{v}$, $\mathbf{w}$ and $z \in S$. This, together with (7.85), implies that

$$\phi_s^2 \prec \Psi^2.$$

We then conclude from (7.81) and (7.84) that

$$\frac{1}{m}\sum_{k=1}^{m} B_{k,k,i,\mu}(g_h) \prec \Psi^2, \quad \text{for} \quad g(w) \geq 1. \tag{7.96}$$

For future use, recalling (7.68) and (7.73) we also obtain from (7.83) and (7.96)

$$\frac{1}{m}\sum_{k=1}^{m} C_{k,k,i,\mu}(g_h) \prec \Psi^2, \quad \text{for} \quad g(w) \geq 1. \tag{7.97}$$

As in (7.81) we then have

$$|\frac{1}{m}\sum_{k=1}^{m} B_{k,k,i,\mu}(g_0)^{2q-l} \prod_{r=1}^{l} \frac{1}{m}\sum_{k=1}^{m} B_{k,k,i,\mu}(g_r)| \prec \hat{F}^{2q-l}(\mathbf{X}, z)\Psi^{2l}.$$

(7.95) and hence (7.20) then follow via (7.39) and an argument similar to (7.91).

7.3 Convergence rate on the right edge and universality

7.3.1 Convergence rate on the right edge

The aim of this subsection is to prove the following Lemma.

Lemma 13. *Denote by λ_1 the largest eigenvalue of $\mathbf{A}$ in (7.1). Under conditions of Theorem 2.1,*

$$\lambda_1 - \hat{\mu}_m = O_{\prec}(n^{-\frac{2}{3}}).$$

Proof. The approach is similar to that in [8], ([19]) and [4]. Checking on the proof of Theorem 4.1 in [4] carefully, we find that (ii) in Theorem 4.1 in [4] and hence the lower bound of λ_1 of Lemma 13 still hold in our case because of (7.23) and (7.32). It then suffices to prove that for any small positive constant τ

$$\lambda_1 \leq \hat{\mu}_m + n^{-2/3+\tau} \tag{7.98}$$

holds with high probability. Note that by (7.3) and Lemma 3

$$\|\mathbf{A}\| \leq M \tag{7.99}$$

with high probability for sufficient large positive constant M (here one should notice that $\|\mathbf{D}^{-1}\| \leq M$ with high probability due to (6.2) and (6.3)). For a suitably small τ, set $z = E + i\eta$ and $\kappa = |E - \hat{\mu}_m|$ where $E \in [\hat{\mu}_m + n^{-2/3+\tau}, \hat{\mu}_m + \tau^{-1}]$ and $\eta = n^{-1/2-\tau/4}\kappa^{1/4}$. By Lemma 2.3 of [4], we have

$$\Im \underline{m} \asymp \frac{\eta}{\sqrt{\kappa+\eta}} \ll \frac{1}{n\eta}, \tag{7.100}$$

where $\ll$ means much less than.

We furthermore claim that with high probability

$$|\underline{m}_n - \underline{m}| \ll \frac{1}{n\eta} \tag{7.101}$$

Indeed, (7.101) holds when $\mathbf{X}$ reduces to $\mathbf{X}^0$ due to (4.6) in [4], (7.100) and (7.32). For the general distributions, (7.101) follows from (7.94) and (7.47). It follows from (7.100) and (7.101) that with high probability

$$\Im(\underline{m}_n) \ll \frac{1}{n\eta}.$$

Moreover note that with high probability

$$\sum_i I(E-\eta \le \lambda_i \le E+\eta) \le Mn\eta\Im(\underline{m}_n) \ll 1.$$

As a consequence there is no eigenvalue in $[E-\eta, E+\eta]$ with high probability. This, together with (7.99), ensures (7.98).

□

7.3.2 Universality

The aim of this subsection is to prove (ii) of Theorem 2.1. By (6.12) and (6.13), it suffices to prove edge universality at the rightmost edge of the support $\hat{\mu}_m$. In other words, the asymptotic distribution of λ_1 is not affected by the distribution of the entries of $\mathbf{X}$ under the 3rd moment matching condition. Similar to theorem 6.4 of [8], we first show the following green function comparison theorem.

Theorem 7.2. *There exists $\varepsilon_0 > 0$. For any $\epsilon < \varepsilon_0$, set $\eta = n^{-2/3-\epsilon}$, E_1, $E_2 \in \mathbb{R}$ with $E_1 < E_2$ and*

$$|E_1 - \hat{\mu}_m|, |E_2 - \hat{\mu}_m| \le n^{-2/3+\epsilon}.$$

Suppose that $K : \mathbb{R} \to \mathbb{R}$ is a smooth function with bounded derivatives up to fifth order. Then there exists a constant $\phi > 0$ such that for large enough n

$$|\mathbb{E}K(n\int_{E_1}^{E_2} \Im m_{\mathbf{X}^1}(x+i\eta)dx) - \mathbb{E}K(n\int_{E_1}^{E_2} \Im m_{\mathbf{X}^0}(x+i\eta)dx)| \le n^{-\phi}, \tag{7.102}$$

(see Definition 1 or (2.8) for $\mathbf{X}^1$ and $\mathbf{X}^0$).

Proof. Unlike [15], [8] and [3] we use the interpolation method (7.43), which is succinct and powerful when proving green function comparison theorem. In view of (7.15) and (7.16) we have

$$\begin{aligned}
&|\mathbb{E}K(n\int_{E_1}^{E_2} \Im m_{\mathbf{X}^1}(x+i\eta)dx) - \mathbb{E}K(n\int_{E_1}^{E_2} \Im m_{\mathbf{X}^0}(x+i\eta)dx)| = \\
&\left|\mathbb{E}K(n\int_{E_1}^{E_2} \Im m_{\mathbf{X}^1}(x+i\eta)\mathcal{T}_n(\mathbf{X}^1)dx) - \mathbb{E}K(n\int_{E_1}^{E_2} \Im m_{\mathbf{X}^0}(x+i\eta)\mathcal{T}_n(\mathbf{X}^0)dx)\right| + O(n^{-1}).
\end{aligned} \tag{7.103}$$

Applying (7.43) with $F(\mathbf{X}) = K(n\int_{E_1}^{E_2} \Im m_{\mathbf{X}}(x+i\eta)\mathcal{T}_n(\mathbf{X}))$ we only need to bound the following

$$\sum_{i=1}^{m}\sum_{\mu=1}^{p}\left|\mathbb{E}g(X_{i\mu}^1) - \mathbb{E}g(X_{i\mu}^0))\right|, \tag{7.104}$$

where

$$g(X_{i\mu}^u) = K(n\int_{E_1}^{E_2} \Im m_{\mathbf{X}_{(i\mu)}^{t,X_{i\mu}^u}}(x+i\eta)\mathcal{T}_n(\mathbf{X}_{(i\mu)}^{t,X_{i\mu}^u})dx), \quad u = 0, 1.$$

As in (7.56) and (7.57), we use Taylor's expansion up to order five to expand two functions $g(X_{i\mu}^u), u = 0, 1$ at the point 0. Then take the difference of the Taylor's expansions of $g(X_{i\mu}^u), u = 0, 1$. By the 3rd moments matching condition it then suffices to bound the fourth derivative

$$\sum_{r=1}^{4}\sum_{\substack{k_1,..,k_r\in\mathbb{N}_+\\ k_1+..+k_r=4}} M_r \max_x |K^{(r)}(x)|\mathbb{E}\prod_{i=1}^{r}\left(n\int_{E_1}^{E_2}\left|m_{\mathbf{X}_{(i\mu)}^{t,0}}^{(k_i)}(x+i\eta)\mathcal{T}_n(\mathbf{X}_{(i\mu)}^{t,0})\right|dx\right), \tag{7.105}$$

and the fifth derivative corresponding to the remainder of integral form

$$\frac{1}{\sqrt{n}}\sum_{r=1}^{5}\sum_{\substack{k_1,..,k_r\in\mathbb{N}_+\\ k_1+..+k_r=4}} M_r \max_x |K^{(r)}(x)|\mathbb{E}\prod_{i=1}^{r}\left(n\int_{E_1}^{E_2}\left|m_{\mathbf{X}_{(i\mu)}^{t,\theta X_{i\mu}^u}}^{(k_i)}(x+i\eta)\mathcal{T}_n(\mathbf{X}_{(i\mu)}^{t,\theta X_{i\mu}^u})\right|dx\right), \tag{7.106}$$

where M_r is a constant depending on r only, $m_{\mathbf{X}_{(i\mu)}^{t,0}}^{(k_i)}(\cdot)$ denotes the k_ith derivative with respect to $X_{i\mu}^u$ and $0 \le \theta \le 1$. Here we ignore the terms involving the derivatives of $\mathcal{T}_n(\mathbf{X}_{(i\mu)}^{t,\theta X_{i\mu}^u})$ due to (7.15), (7.16) and (7.28).

To investigate (7.105) and (7.106) we claim that it suffices to prove that

$$\left(n\int_{E_1}^{E_2}\left|m_{\mathbf{X}_{(i\mu)}^{u,X_{i\mu}^1}}^{(k)}(x+i\eta)\mathcal{T}_n(\mathbf{X}_{(i\mu)}^{u,X_{i\mu}^1})\right|dx\right) \prec (n^{\frac{1}{3}+\epsilon}\Psi^2), \tag{7.107}$$

where $k \ge 1$. Indeed, if (7.107) holds then (7.107) still holds if $X_{i\mu}^1$ is replaced by $\theta X_{i\mu}^1$ by checking on the argument of (7.107). We then conclude that the facts that (7.105) $\prec (n^{\frac{1}{3}+\epsilon}\Psi^2)$ and that (7.106) $\prec (n^{-\frac{1}{2}+\frac{1}{3}+\epsilon}\Psi^2)$ follow from Lemma 5, (7.28) and an application of (7.56).

By (7.68) and (7.97) we have for $k \ge 1$

$$\left|m_{\mathbf{X}_{(i\mu)}^{u,X_{i\mu}^1}}^{(k)}(x+i\eta)\mathcal{T}_n(\mathbf{X}_{(i\mu)}^{u,X_{i\mu}^1})\right| \prec \Psi^2,$$

which implies that (7.107) $\prec (n^{\frac{1}{3}+\epsilon}\Psi^2)$. Here we would point out that the derivatives $m_{\mathbf{X}_{(i\mu)}^{u,X_{i\mu}^1}}^{(k)}(\cdot)$ are of the form $\frac{1}{m}\sum_{k=1}^{m} C_{k,k,i,\mu}(g_h)$ from (7.67), (7.68), (7.71), (7.73), (7.94) and (7.95). By Lemma 2.3 of [4] we have

$$\Psi^2 \asymp \frac{1}{n\sqrt{\eta}} = O(n^{-\frac{2}{3}+\epsilon/2}).$$

Summarizing the above we have shown that

$$|\mathbb{E}K(n\int_{E_1}^{E_2}\Im m_{\mathbf{X}^1}(x+i\eta)dx)-\mathbb{E}K(n\int_{E_1}^{E_2}\Im m_{\mathbf{X}^0}(x+i\eta)dx)|\prec n^{-\frac{1}{3}+2\epsilon}.$$

The proof is complete by choosing an appropriate ϵ. □

In order to prove the Tracy-Widom law, we need to connect the probability $\mathbb{P}(\lambda_1\leq E)$ with Theorem 7.2.

By Lemma 13 we can fix $E^*\prec n^{-\frac{2}{3}}$ such that it suffices to consider $\lambda_1\leq\hat{\mu}_m+E^*$. Choosing $|E-\hat{\mu}_m|\prec n^{-\frac{2}{3}}$, $\eta=n^{-\frac{2}{3}-9\epsilon}$ and $l=\frac{1}{2}n^{-\frac{2}{3}-\epsilon}$, then for some sufficiently small constant $\epsilon>0$ and sufficiently large constant D, there exists a constant $n_0(\epsilon,D)$ such that

$$\mathbb{E}K(\frac{n}{\pi}\int_{E-l}^{\hat{\mu}_m+E^*}\Im m_{\mathbf{X}^1}(x+i\eta)dx)\leq\mathbb{P}(\lambda_1\leq E)\leq\mathbb{E}K(\frac{n}{\pi}\int_{E+l}^{\hat{\mu}_m+E^*}\Im m_{\mathbf{X}^1}(x+i\eta)dx)+n^{-D},\tag{7.108}$$

where $n\geq n_0(\epsilon,D)$ and K is a smooth cutoff function satisfying the condition of K in Theorem 7.2. We omit the proof of (7.108) because it is a standard procedure and one can refer to [8] or Corollary 5.1 of [4] for instance. Combining (7.108) with Theorem 7.2 one can prove Tracy-Widom's law directly (see the proof of Theorem 1.3 of [3]).

Acknowledgment. G. M. Pan was partially supported by a MOE Tier 2 grant 2014-T2-2-060 and by a MOE Tier 1 Grant RG25/14 at the Nanyang Technological University, Singapore.

References

[1] BAI, Z. D. and SILVERSTEIN, J. W. (2006). *Spectral analysis of large dimensional random matrices*, 1st ed. Springer, New York.

[2] BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of Large Sample Covariance Matrices of Spiked Population Models. *J. Multivariate Anal.* **97**, 1382–1408.

[3] BAO, Z. G., PAN, G. M. and ZHOU, W. (2015). Universality for the largest eigenvalue of sample covariance matrices with general population. *Ann. Statist.* **43**(1), 382–421.

[4] BAO, Z. G., PAN, G. M. and ZHOU, W. Local density of the spectrum on the edge for sample covariance matrices with general population. *Preprint.* Available at http://www. ntu. edu. sg/home/gmpan/publications. html.

[5] DHARMAWANSA, P., JOHNSTONE, I. M. and ONATSKI, A. (2014). Local Asymptotic Normality of the spectrum of high-dimensional spiked F-ratios. http://arxiv.org/pdf/1411.3875.pdf.

[6] EL KAROUI, N. (2007). Tracy-Widom Limit for the Largest Eigenvalue of a Large Class of Complex Sample Covariance Matrices, *Ann. Probab.* **35**,663-714.

[7] ERDÖS, L., SCHLEIN, B. and YAU, H.-T. (2009). Local Semicircle Law and Complete Delocalization for Wigner Random Matrices. Communications in Mathematical Physics, **287**(2), 641-655.

[8] ERDÖS, L., YAU, H.-T., and YIN, J.(2011). Rigidity of Eigenvalues of Generalized Wigner Matrices , *Advances in Mathematics*, **229(3)**, 1435-1515.

[9] ERDÖS, L., KNOWLES, A. and YAU, H.-T.(2013). Averaging fluctuations in resolvents of random band matrices, *Ann. H. Poincaré*, **14**, 1837C1926.

[10] FÉRAL, D., PÉCHÉ, S.(2009). The largest eigenvalues of sample covariance matrices for a spiked population: Diagonal case. J. Math. Phys. **50**, 073302.

[11] JOHANSSON, K. (2000). Shape fluctuations and random matrices. Comm. Math. Phys. **209**, No. 2, 437-476.

[12] JOHNSTONE, I.M. (2001). On the Distribution of the Largest Eigenvalue in Principal Component Analysis, *Ann. Statist.* **29**, 295-327.

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

[14] JOHNSTONE, I. M. (2009). Approximation null distribution of the largest root in multivariate analysis. *Ann. Appl. Statist.* **3** No.4 1616–1633.

[15] KNOWLES, A. and YIN, J. (2015). Anisotropic local laws for random matrices. *arXiv:1410.3516v3.*

[16] LEE, J. O. and SCHNELLI, K. (2014). Tracy-Widom Distribution for the Largest Eigenvalue of Real Sample Covariance Matrices with General Population. *arXiv:1409.4979v1.*

[17] MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sb. Math.* **4** 457–483.

[18] MUIRHEAD, R. J. (1982). Aspects of Multivariate Statistical Theory. *Wiley, New York.* MR0652932.

[19] PILLALI, N. S. and YIN, J. (2011). Universality of covariance matrices. *Ann. Appl. Prob.* **24** No.3,935–1001.

[20] SILVERSTEIN, J. W. and CHOI,S.-I (1995). Analysis of the Limiting Spectral Distribution of Large Dimensional Random Matrices. *Journal of Multivariate Analysis*, 54(2), 295C309.

[21] SOSHNIKOV, A. (2002). A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. Jour. Stat. Phys. **108**(5), 1033-1056.

[22] TAO, T. and VU,V. (2011). Random matrices: Universality of local eigenvalue statistics. Acta Mathematica, **206**(1), 127-204.

[23] TAO, T. and VU, V. (2012). Random covariance matrices: Universality of local statistics of eigenvalues. Ann. Probab. **40**(3), 1285-1315.

[24] TRACY, C. A. and WIDOM,H. (1994). Level-spacing distributions and the Airy kernel. Comm. Math. Phys. **159**, No. 1, 151-174.

[25] TRACY, C. A. and WIDOM,H. (1996). On orthogonal and symplectic matrix ensembles. Comm. Math. Phys. **177**, No. 3, 727-754.

[26] WACHTER, K. (1980) The limiting empirical measure of multiple discriminant ratios, The Annals of Statistics 8, 937-957.

[27] WANG, K. (2012). Random covariance matrices: Universality of local statistics of eigenvalues up to the edge. Random matrices: Theory and Applications, **1**(1), 1150005.

[28] WANG, Q. W. and YAO, J. F.(2015). Extreme eigenvalues of large-dimensional spiked Fisher matrices with application. http://arxiv.org/pdf/1504.05087.pdf.

[29] ZHENG, S. R. (2012). Central Limit Theorem for Linear Spectral Statistics of Large Dimensional F Matrix. Ann. Institut Henri Poincare Probab. Statist. 48, 444-476.