Anisotropic local law and its application in random matrix theory

By

Haokai Xi

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
(Mathematics)

at the

UNIVERSITY OF WISCONSIN – MADISON

2017

Date of final oral examination: May 9, 2017

The dissertation is approved by the following members of the Final Oral Committee:

Jun Yin, Associate Professor, Mathematics (Advisor)
Timo Seppäläinen, Professor, Mathematics
Benedek Valkó, Professor, Mathematics
Philip Matchett Wood, Assistant Professor, Mathematics
Serguei Denissov, Professor, Mathematics

Abstract

This dissertation is broken into two chapters. The first chapter considers the local circular law for the product TX where T is a deterministic $N \times M$ matrix and X is a random $M \times N$ matrix with independent entries having zero mean and variance $(N \wedge M)^{-1}$. We prove that if z is away from the unit circle, the empirical spectral distribution(ESD) of TX converges to $\tilde{\chi}_{\mathbb{D}}(z)dA(z)$, where $\tilde{\chi}_{\mathbb{D}}$ is a rotation-invariant function determined by the singular values of T and A denotes the Lebesgue measure on $\mathbb{C}$. The local circular law is valid around z up to scale $(N \wedge M)^{-1/4+\epsilon}$ for any $\epsilon > 0$. Moreover, if $|z| > 1$ or the matrix entries of X have vanishing third moments, the local circular law is valid around z up to scale $(N \wedge M)^{-1/2+\epsilon}$ for any $\epsilon > 0$. The second chapter considers the convergence rate of the eigenvector empirical spectral distribution(VESD) of sample covariance matrices to the Marchenko-Pastur (MP) distribution. We consider sample covariance matrices of the form XX^*, where $X = (x_{ij})$ is an $M \times N$ random matrix, whose entries are independent random variables with mean zero and variance N^{-1}. We show that, under the finite 6th moment condition, the Kolmogorov distance between the *expected VESD* and the MP distribution is of order at most $N^{-1+\epsilon}$ for any fixed $\epsilon > 0$; under the finite 8th moment condition, the convergence rate of the VESD is $O(N^{-1/2+\epsilon})$ almost surely for any fixed $\epsilon > 0$. Both of the works are joint with Fan Yang and Jun Yin.

Acknowledgements

First and foremost, I want to thank my advisor, Prof. Jun Yin, for his guidance, support and encouragement. Without his advice, this thesis would not exist. I thank all my committee members for reading this work and providing their invaluable comments and suggestions. I would like to thank my collaborator Fan Yang for our fruitful discussions, and for all the help he has provided for my thesis. I want to thank Prof. Ignacio Uriarte-Tuero for encouraging me to pursue Ph.D when I was an undergraduate student. I am grateful to all my friends who shaped my life these years, especially for their company during the good and hard times. Last but not the least, I want to thank my family: my parents and my brother for supporting me spiritually throughout my Ph.D study and my life in general.

List of Figures

Contents

Chapter 1

Introduction

The study of random matrices dates back to Wishart's work [67] in the area of multivariate statistics in the 1920's. It became influential in the 1950's by the work of Wigner in the area of nuclear physics, where he produced the idea of using large random matrices to model statistical properties of the energy levels [66]. Since then random matrix problems attracted considerable attention in theoretical physics as well as mathematics community, and many essential tools helping to analyze properties of random matrices were developed. Among the interesting random matrix properties, one focus is the study of the empirical spectral distribution(ESD) of large random matrices. In the remainder of the introduction, we will briefly overview the types of matrix models people have studied on this problem. In Chapter 2 and 3, there are more detailed introduction to the questions considered there.

1.1 Classical matrix models

1.1.1 Wigner semicircle law

For a $N \times N$ Hermitian random matrix X_N, the ESD of X_N is defined as

$$\mu_{X_N} := \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_N$ are the eigenvalues of X_N. In other words, $\mu_{X_N}(I)$ counts the number of eigenvalues of X_N in I. If the upper-triangular entries X_{ij} $i > j$ are iid random variables with mean 0 and variance 1, and the diagonal entries X_{ii} are iid real random variables that are independent of the upper-triangular entries, and has bounded mean and variance, then we call X_N a Wigner matrix. Some Wigner matrix ensembles of special interest are Gaussian orthogonal ensembles(GOE), Gaussian unitary ensembles(GUE). Their names are derived from Dyson's categorization of random matrix ensembles by symmetry class [17].

Let X_N be an ensemble of Wigner matrices, after scaling by $1/\sqrt{N}$ (so the operator norm has typical size 1), it is well known that The ESD of $\frac{1}{\sqrt{N}}X_N$ converges weakly to the Wigner semicircle distribution

$$d\mu_{sc} := \frac{1}{2\pi}\sqrt{(4-|x|^2)_+}dx.$$

One approach to derive this result is the *Stieltjes transform method*, which starts with the identity

$$\int_{\mathbf{R}} \frac{1}{x-z}\, d\mu_{\frac{1}{\sqrt{N}}X_N}(x) = \frac{1}{N}\mathrm{Tr}\left(\frac{1}{\sqrt{N}}X_N - zI\right)^{-1}. \tag{1.1}$$

The left hand side of the above equation is the Stieltjes transform of $\mu_{\frac{1}{\sqrt{N}}X_N}$, and $(X_N/\sqrt{N} - zI)^{-1}$ on the right hand side is called the resolvent of $\frac{1}{\sqrt{N}}X_N$. Using the Schur's complement, we can get the self-consistent equation

$$m_c(z) = -\frac{1}{m_c(z)+z},$$

where m_c is the Stieltjes transform of the semicircle law μ_{sc}. One can solve the above equation explicitly and use inverse Stieltjes transform to obtain the exact form of μ_{sc}.

Moreover, we can get a local version of the semicircle law from the Stieltjes transform method. Note that the imaginary part of the left side of (1.1) (let $z = E + i\eta$) is

$$\int_{\mathbf{R}} \frac{\eta}{(x-E)^2+\eta^2}\, d\mu_{\frac{1}{\sqrt{N}}X_N}(x).$$

If we choose η to be sufficiently small, then the above quantity estimates the density of eigenvalues in a scale of order η around E. In [24], Erdos, Schlein, and Yau proved such a local law up to scale $N^{-1+\epsilon}$ for any $\epsilon > 0$. Later together with Knowles and Yin, they also produced a series of papers [25, 26, 29, 27, 28, 31, 30, 22] where the local law was applied to solve various problems.

1.1.2 Marchenko-Pastur (MP) law

The counterpart of the semicircle law for covariance matrix is called Marchenko-Pastur (MP) law, which is proved by Marchenko and Pastur[48]. Given a $M \times N$ i.i.d. matrix X and $\mathbb{E}X_{ij} = 0$, $\mathbb{E}|X_{ij}|^2 = 1/N$. Assume $M/N \to k$, the limiting ESD of XX^* is the Marchenko and Pastur law:

$$d\mu_{MP} = \frac{1}{2\pi}\frac{\sqrt{(\lambda_+ - x)(x-\lambda_-)}}{kx}\mathbf{1}_{[\lambda_-,\lambda_+]}dx + (1-k^{-1})_+\delta_0 dx$$

where $\lambda_- = (1-\sqrt{k})^2$, $\lambda_+ = (1+\sqrt{k})^2$. One can obtain the self consistence equation for the Stieltjes transform m_c of the limiting distribution μ_{MP}

$$m_c(d,z) + \frac{1}{z - (1-d^{-1}) + zd^{-1}m_c(d,z)} = 0.$$

Then by applying the inverse Stieltjes transform we can get the exact form of the MP law. The local law for the covariance matrix is also proved in [24] and simultaneously in [12]. In [44], Knowles and Yin proved the local law for the covariance matrix of the form $TXX^\dagger T^\dagger$, where T is some deterministic matrix.

1.1.3 Circular law

The study of the eigenvalue spectral of non-Hermitian random matrices goes back to the celebrated paper [33] by Ginibre, where he calculated the joint probability density for the eigenvalues of non-Hermitian random matrix with independent complex Gaussian entries. The joint density distribution is integrable with an explicit kernel (see [33, 49]), which allowed him to derive the circular law for the eigenvalues. For the Gaussian random matrix with real entries, the joint distribution of the eigenvalues is more complicated but still integrable, which leads to a proof of the circular law as well [8, 18, 32, 61].

For the random matrix with non-Gaussian entries, there is no explicit formula for the joint distribution of the eigenvalues. However, in many cases the eigenvalue spectrum of the non-Gaussian random matrices behaves similarly to the Gaussian case as $N \to \infty$, known as the universality phenomena. More precisely, consider an $N \times N$ iid matrix X_N whose entries have mean zero and variance one, the circular law states that the ESD of $\frac{1}{\sqrt{N}} X_N$ converges to the circular law

$$d\mu_{circ} := \frac{1}{\pi} \mathbf{1}_{|z| \leq 1} \, dA(z),$$

where A is the Lebesgue measure on $\mathbf{C}$.

A key step in this direction is made by Girko in [34], where he partially proved the circular law for non-Hermitian matrices with independent entries. The crucial insight of the paper is the *Hermitization technique*, which allowed Girko to translate the convergence of complex empirical measures of a non-Hermitian matrix into the convergence of logarithmic transforms for a family of Hermitian matrices, or, to be more precise,

$$\mathrm{Tr}\log[(X-z)^{\dagger}(X-z)] = \log\left[\det((X-z)^{\dagger}(X-z))\right], \tag{1.2}$$

with X being the random matrix and $z \in \mathbb{C}$. Due to the singularity of the log function at 0, the small eigenvalues of $(X-z)^{\dagger}(X-z)$ play a special role. The estimate on the smallest singular value of $X-z$ was not obtained in [34], but the gap was remedied later in a series of paper. Bai [2, 4] analyzed the ESD of $(X-z)^{\dagger}(X-z)$ through its Stieltjes transform and handled the logarithmic singularity by assuming bounded density and bounded high moments for the entries of X. Lower bounds on the smallest singular values were given by Rudelson and Vershynin [55, 56], and subsequently by Tao and Vu [62], Pan and Zhou [52] and Gőtze and Tikhomirov [36] under weakened moments and smoothness assumptions. The final result was presented in [65], where the circular law is proved under the optimal L^2 assumption.

1.2 Anisotropic local law

We denote the resolvent of a $N \times N$ Hermitian matrix H by $R(w) := (H-w)^{-1}$, where $w = E + i\eta$ is a spectral parameter with positive imaginary part η. Then the Stieltjes transform of the ESD of H is equal to $N^{-1}\mathrm{Tr}\, R(w)$, and we have the convergence estimate

$$N^{-1}\mathrm{Tr}\, R(w) \approx m_c(w) \tag{1.3}$$

with high probability for large N. Here m_c is the Stieltjes transform of the asymptotic eigenvalue distribution. The convergence in (1.3) is referred to as the *averaged law*. By taking the imaginary part of (1.3), we have seen that a control of the Stieltjes transform yields a control of the eigenvalue density on a small scale of order η around E (which contains an order ηN eigenvalues). A *local law* is an estimate of the form (1.3) for all $\eta \gg N^{-1}$. Such local laws have been a cornerstone of the modern random matrix theory (see [24] for local semicircle law and [10] for local circular law).

For the questions in Chapter 2 and 3, it turns out the averaged local law from (1.3) is not sufficient. We have to control not only the trace of $R(w)$, but also the matrix $R(w)$ itself by showing that $R(w)$ is close to some deterministic matrix $\Pi(w)$, provided that $\eta \gg N^{-1}$. This closeness can be established in the sense of individual matrix entries $R_{ij}(w) \approx \Pi_{ij}(z)$ (see e.g. [10, 30]). We call such an estimate an *entrywise local law.* More generally, we can establish it for *generalized matrix entries* (see [6, 43]):

$$\langle \mathbf{v}, R(w)\mathbf{u}\rangle \approx \langle \mathbf{v}, \Pi(w)\mathbf{u}\rangle, \quad \eta \gg N^{-1}, \quad \forall \|\mathbf{v}\|_2, \|\mathbf{u}\|_2 = 1. \tag{1.4}$$

We call the estimate in (1.4) an *anisotropic local law.* (If Π is a scalar matrix, (1.4) is also referred to as an *isotropic local law*, in the sense that $R(w)$ is approximately isotropic for large N.) Such an *anisotropic local law* is the key ingredient in proving the problems in Chapter 2 and 3.

Chapter 2

Local circular law for TX

2.1 Introduction

2.1.1 Local circular law

The history of circular law was briefly introduced in the first chapter, those works studied the circular law in the global regime, i.e. the convergence of ESD on subsets containing ηN eigenvalues for some small constant $\eta > 0$. Later in a series of papers [10, 11, 72], Bourgade, Yau and Yin proved the *local* version of the circular law up to the optimal scale $N^{-1/2+\epsilon}$ under the assumption that the distributions of the matrix entries satisfy a uniform sub-exponential decay condition. In [64], the local universality was proved by Tao and Vu under the assumption of first four moments matching the moments of a Gaussian random variable.

2.1.2 The matrix model in this chapter

In this chapter, we study the ESD of the product of a deterministic $N \times M$ matrix T with a random $M \times N$ matrix X, where we assume $N \sim M$. In Figure 1, we plot the eigenvalue distribution of TX when T have two distinct singular values (except the trivial zero singular values). The goal of this chapter is to prove a local circular law for

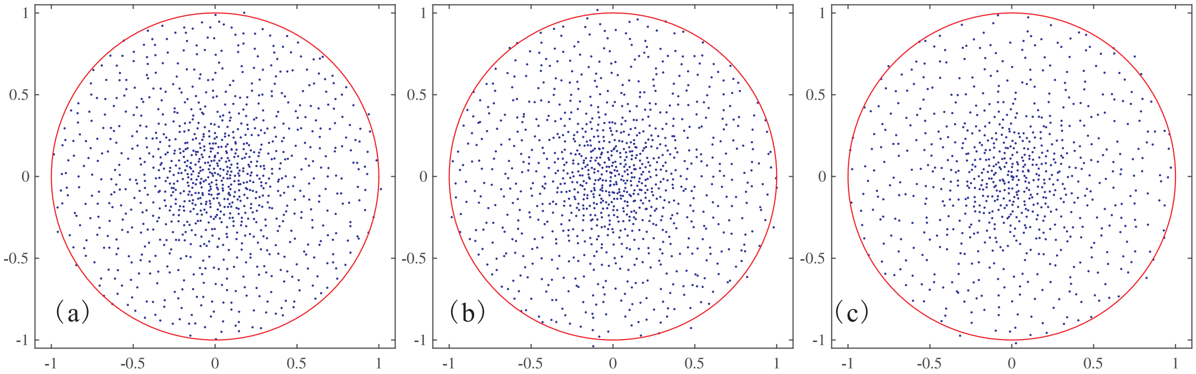


Figure 1: The eigenvalue distribution of the product TX of a deterministic $N \times M$ matrix T with a Gaussian random $M \times N$ matrix X. The entries of X have zero mean and variance $(N \wedge M)^{-1}$, and $TT^\dagger$ has $0.5(N \wedge M)$ eigenvalues as $2/17$ and $0.5(N \wedge M)$ eigenvalues as $32/17$. (a) $N = M = 1000$. (b) $N = 1000$, $M = 2000$. (c) $N = 1500$, $M = 750$.

the ESD of TX at any point z away from the unit circle. Following the idea in [10], the key ingredients for the proof are (a) the upper bound for the largest singular value of $TX - z$, (b) the lower bound for the least singular value of $TX - z$, and (c) rigidity of the singular values of $TX - z$. The upper bound for the largest singular value can be obtained by controlling the norm of $TX - z$ through a standard large deviation estimate (see e.g. [13, 47, 57] and (2.65)). The lower bound for the least singular value of $TX - z$ follows from the results in e.g. [56] and [62] (see also Lemma 2.2.23). Thus the bulk of this work is devoted to establish (c).

2.1.3 Basic ideas

To obtain the rigidity of the singular values of $Tx - z$, we study the ESD of $Q := (TX - z)^\dagger (TX - z)$ using Stieltjes transform as in [10]. We normalize X so that its entries have variance $(N \wedge M)^{-1}$. Then Q is an $N \times N$ Hermitian matrix with eigenvalues being typically of order 1. We denote its resolvent by $R(w) := (Q - w)^{-1}$, where $w = E + i\eta$ is a spectral parameter with positive imaginary part η. Then the Stieltjes transform of the ESD of Q is equal to $N^{-1}\mathrm{Tr}\, R(w)$. Let m_c be the Stieltjes transform of the asymptotic eigenvalue distribution. We have seen that the estimate $N^{-1}\mathrm{Tr}\, R(w) \approx m_c(w)$ is referred as the *averaged law*. In [10], a local law for the resolvent of $(X-z)^\dagger(X-z)$ was established to prove the local circular law.

In generalizing the proof in [10] to our setting, a main difficulty is that the entries of TX are not independent. We will use a new comparison method proposed in [44], which roughly states that if the local laws hold for $R(w)$ with Gaussian X, then they also hold in the case of a general X. For definiteness, we assume $N = M$ for now, and T is a square matrix with singular decomposition $T = UDV$. For a Gaussian $X \equiv X^{Gauss}$, we have $VX^{Gauss}U \overset{d}{=} \tilde{X}^{Gauss}$, where $\tilde{X}$ is another Gaussian random matrix. Then for the determinant in (1.2),

$$\det(TX^{Gauss} - z) = \det(DVX^{Gauss}U - z) \overset{d}{=} \det(D\tilde{X}^{Gauss} - z). \tag{2.1}$$

The problem is now reduced to the study of the singular values of $D\tilde{X}^{Gauss} - z$, which has independent entries. Notice the entries of $D\tilde{X}^{Gauss}$ are not identically distributed, which will make our proof much more complicated. However, this issue can be handled, e.g. as in [22], where a local law was obtained for generalized Wigner matrices with non-identically distributed entries.

To use the comparison method invented in [44], we will need the *anisotropic local law* of the form (1.4). Here we outline the three steps to establish the anisotropic local law for $Q = (TX - z)^\dagger(TX - z)$: (A) the entrywise local law and averaged local law when T is diagonal (Theorem 2.2.18); (B) the anisotropic local law when T is diagonal (Theorem 2.2.18); (C) the anisotropic local law and averaged local law when T is a general (rectangular) matrix (Theorem 2.2.19).

In performing Step (A), our proof is basically based on the methods in [10]. However, our multi-variable self-consistent equations and their solutions are much more complicated here. Thus a key part of the proof is to establish some basic properties of the asymptotic eigenvalue density and prove the stability of the self-consistent equations under small perturbations. These work need some new ideas and analytic techniques (see Appendix 2.7). In performing Step (B), we applied and extended the polynomialization method developed in [6, section 5]. Finally, as remarked around (2.1), (B) implies the anisotropic local law for a Gaussian X and a general T. Based on this fact we perform Step (C) using a self-consistent comparison argument in [44]. With the averaged local law proved in Step (C), we can prove the local circular law for TX. In general, the averaged local law we get is up to the non-optimal scale $\eta \gg (N \wedge M)^{-1/2}$. As a result, we can only prove the local circular law for TX up to the scale $(N \wedge M)^{-1/4+\epsilon}$. A new observation is that the non-optimal averaged local law can lead to the optimal local circular law for TX outside the unit circle (i.e. $|z| > 1$) (see Section 2.2.4). To prove the optimal local circular law inside the unit circle (i.e. $|z| < 1$), we need the optimal averaged local law up to the scale $\eta \gg (N \wedge M)^{-1}$, which can be obtained under the extra assumption that the entries of X have vanishing third moments.

2.1.4 Conventions

The fundamental large parameter is N and we assume that M is comparable to N (see (2.2)). All quantities that are not explicitly constant may depend on N, and we usually omit N from our notation. We use C to denote a generic large positive constant, which may depend on fixed parameters and whose value may change from one line to the next. Similarly, we use c or ϵ to denote a generic small positive constant. If a constant depend on a quantity a, we use $C(a)$ or C_a to indicate this dependence. We use $\tau > 0$ in various assumptions to denote a small positive constant, and use ζ, τ' to denote constants that depend on τ and may be chosen arbitrarily small. All constants C, c and ϵ may depend on τ; we neither indicate nor track this dependence.

For any (complex) matrix A, we use $A^\dagger$ to denote its conjugate transpose, A^T the transpose, $\|A\|$ the operator norm and $\|A\|_{HS}$ the Hilbert-Schmidt norm. We use the notation $\mathbf{v} = (v_i)_{i=1}^n$ for a vector in $\mathbb{C}^n$, and denote its Euclidean norm by $|\mathbf{v}| \equiv \|\mathbf{v}\|_2$. We usually write the $n \times n$ identity matrix I_n as 1 without causing any confusions.

For two quantities A_N and $B_N > 0$ depending on N, we use the notations $A_N = O(B_N)$ and $A_N \sim B_N$ to mean $|A_N| \le CB_N$ and $C^{-1}B_N \le |A_N| \le CB_N$, respectively, for some positive constant $C > 0$. We use $A_N = o(B_N)$ to mean $|A_N| \le c_N B_N$ for some positive constant $c_N \to 0$ as $N \to \infty$. If A_N is a matrix, we use the notations $A_N = O(B_N)$ and $A_N = o(B_N)$ to mean $\|A_N\| = O(B_N)$ and $\|A_N\| = o(B_N)$, respectively.

2.1.5 Outline of this chapter

The majority of this chapter is devoted to the proof of Theorem 2.2.18 and Theorem 2.2.19. In Section 2.3, we collect the basic tools that we shall use in the proof. In Section

2.4, we perform step (A) of the proof by proving the entrywise local law and averaged local law in Theorem 2.2.18 under the assumption that T is diagonal. We first prove a weak version of the entrywise local law in Sections 2.4.1-2.4.3, and then improve the weak law to the strong entrywise local law and averaged local law in Sections 2.4.4-2.4.5. In Section 2.5, we perform step (B) of the proof by proving the anisotropic local law in Theorem 2.2.18 using the entrywise local law proved in Section 2.4. Finally in Section 2.6 we finish the step (C) of the proof, where using Theorem 2.2.18, we prove Theorem 2.2.19 with a self-consistent comparison method.

2.2 Main results

In this section, we state and prove the main result of this chapter. In Section 2.2.1, we define our model and list our main assumptions. In Section 2.2.2, we first define the asymptotic eigenvalue density ρ_{2c} of $Q = (TX - z)^{\dagger}(TX - z)$, and then state the main theorem—Theorem 2.2.6—of this chapter. Its proof depends crucially on local estimates of the resolvent of Q, which are presented in Section 2.2.3. In Section 2.2.4, we prove Theorem 2.2.6 based on the local estimates stated in Section 2.2.3.

2.2.1 Definition of the model

In this chapter, we want to understand the local statistics of the eigenvalues of $TX - zI$, where T is a deterministic $N \times M$ matrix, X is a random $M \times N$ matrix, $z \in \mathbb{C}$ and I is the $N \times N$ identity matrix. We assume $M \sim N$, i.e.

$$\tau \leq \frac{M}{N} \leq \tau^{-1} \tag{2.2}$$

for some small constant $\tau > 0$. We assume the entries $X_{i\mu}$ of X are independent (not necessarily identically distributed) random variables satisfying

$$\mathbb{E}\, X_{i\mu} = 0, \quad \mathbb{E}\, |X_{i\mu}|^2 = \frac{1}{N \wedge M} \tag{2.3}$$

for all $1 \leq i \leq M, 1 \leq \mu \leq N$. For definiteness, in this chapter we only focus on the case where all the X entries are real. However, our results and proofs also hold, after minor changes, in the complex case if we assume in addition $\mathbb{E}X_{i\mu}^2 = 0$ for $X_{i\mu} \in \mathbb{C}$. We assume that for all $p \in \mathbb{N}$, there is an N-independent constant C_p such that

$$\mathbb{E}\, |\sqrt{N \wedge M} X_{i\mu}|^p \leq C_p \tag{2.4}$$

for all $1 \leq i \leq M, 1 \leq \mu \leq N$. We define $\Sigma := TT^\dagger$, and assume the eigenvalues of Σ satisfy that

$$\tau^{-1} \geq \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{N \wedge M} \geq \tau \tag{2.5}$$

and all other eigenvalues are 0. Furthermore, we can normalize T by multiplying a scalar such that

$$\frac{1}{N \wedge M} \sum_{i=1}^{N \wedge M} \sigma_i = 1. \tag{2.6}$$

We summarize our basic assumptions here for future reference.

Assumption 2.2.1. *We suppose that (2.2), (2.3), (2.4), (2.5) and (2.6) hold.*

2.2.2 Main theorem

To state the main theorem, we need to define the asymptotic eigenvalue density function for Q. We first introduce the self-consistent equations, then the asymptotic eigenvalue density will be closely related to their solutions. Let

$$\rho_\Sigma := \frac{1}{N \wedge M} \sum_{i=1}^{N \wedge M} \delta_{\sigma_i} \tag{2.7}$$

denote the empirical spectral density of Σ. Let $n := |\operatorname{supp}\rho_\Sigma|$ be the number of distinct nonzero eigenvalues of Σ, which are denoted as

$$\tau^{-1} \geq s_1 > s_2 > \cdots > s_n \geq \tau. \tag{2.8}$$

Let l_i be the multiplicity of s_i. By (2.6), l_i and s_i satisfy the normalization conditions

$$\frac{1}{N \wedge M}\sum_{i=1}^{n} l_i = 1, \quad \frac{1}{N \wedge M}\sum_{i=1}^{n} l_i s_i = 1. \tag{2.9}$$

For each $w \in \mathbb{C}_+ := \{w \in \mathbb{C} : \operatorname{Im} w > 0\}$, we define the self-consistent equations of (m_1, m_2) as

$$\frac{1}{m_2} = -w(1+m_1) + \frac{|z|^2}{1+m_1}, \tag{2.10}$$

$$m_1 = \frac{1}{N}\sum_{i=1}^{n} l_i s_i \left[-w\left(1+s_i m_2\right) + \frac{|z|^2}{1+m_1}\right]^{-1}. \tag{2.11}$$

If we plug (2.10) into (2.11), we get the self-consistent equation for m_1 only:

$$m_1 = \frac{1}{N}\sum_{i=1}^{n} l_i s_i \left[-w\left(1+\frac{s_i}{-w(1+m_1)+\frac{|z|^2}{1+m_1}}\right) + \frac{|z|^2}{1+m_1}\right]^{-1}. \tag{2.12}$$

The next lemma states that the solution to the functional equation (2.12) in $\mathbb{C}_+$ is unique if z is away from the unit circle. It will be proved in Appendix 2.7.3.

Lemma 2.2.2. *Fix $z \in \mathbb{C}$ such that $|z| \neq 1$. For $w \in \mathbb{C}_+$, there exists at most one analytic function $m_{1c,z,\Sigma}(w) : \mathbb{C}_+ \to \mathbb{C}_+$ such that (2.12) holds and $w m_{1c,z,\Sigma}(w) \in \mathbb{C}_+$. Moreover, $m_{1c,z,\Sigma,N}(w)$ is the Stieltjes transform of a positive integrable function ρ_{1c} with compact support in $[0,\infty)$.*

We shall abbreviate $m_{1c}(w) := m_{1c,z,\Sigma}(w)$. We also define $m_{2c}(w) := m_{2c,z,\Sigma}(w)$ by taking $m_1 = m_{1c}(w)$ in (2.10). Obviously, m_{2c} is also an analytic function of w.

Moreover, for any $w \in \mathbb{C}_+$ we can verify that $m_{2c}(w), wm_{2c}(w) \in \mathbb{C}_+$ by using (2.10) and that $m_{1c}, wm_{1c} \in \mathbb{C}_+$. We define two functions on $\mathbb{R}$ as

$$\rho_{1,2c}(x) = \frac{1}{\pi} \lim_{\eta \searrow 0} \operatorname{Im} m_{1,2c}(x + i\eta), \quad x \in \mathbb{R}. \tag{2.13}$$

It is easy to see that $\rho_{1,2c} \geq 0$ and $\operatorname{supp}(\rho_{1,2c}) \subseteq [0, \infty)$. Moreover, $\operatorname{supp} \rho_{2c} = \operatorname{supp} \rho_{1c}$ by (2.10). We shall call ρ_{2c} the asymptotic eigenvalue density of $Q = (TX - z)^\dagger (TX - z)$ (for a reason that will be made clear during the proof in Section 2.4). Since $\operatorname{Im}(wm_{2c}) \geq 0$, we have

$$\operatorname{Im} \left[-w \left(1 + s_i m_{2c}\right) + \frac{|z|^2}{1 + m_{1c}} \right] \leq -\operatorname{Im} w,$$

and (2.11) gives $|m_{1c}| \leq 1/\operatorname{Im} w \to 0$ as $\operatorname{Im} w \to \infty$. Similarly, $|m_{2c}| \leq 1/\operatorname{Im} w \to 0$ as $\operatorname{Im} w \to \infty$. Thus $m_{1,2c}(w)$ is indeed the Stieltjes transform of $\rho_{1,2c}$:

$$m_{1,2c}(w) = \int_{\mathbb{R}} \frac{\rho_{1,2c}(x)}{x - w} dx. \tag{2.14}$$

We now state the basic properties of ρ_{1c} and ρ_{2c}, which can be obtained by studying the solutions $m_{1,2c}(w)$ to the self-consistent equations (2.10) and (2.12) when $w \in (0, \infty)$. Here we extend the definition of $m_{1,2c}$ continuously down to the real axis by setting

$$m_{1,2c}(x) = \lim_{\eta \searrow 0} m_{1,2c}(x + i\eta), \quad x \in \mathbb{R}.$$

As a convention, for $w \in \overline{\mathbb{C}_+}$, we take $\sqrt{w}$ to be the branch with positive imaginary part. Denote $m := \sqrt{w}(1 + m_1)$ and $m_c := \sqrt{w}(1 + m_{1c})$. Equation (2.12) then becomes

$$f(\sqrt{w}, m) = 0, \tag{2.15}$$

where

$$f(\sqrt{w}, m) = -\sqrt{w} + m + \frac{1}{N} \sum_{i=1}^{n} l_i s_i \frac{m(m^2 - |z|^2)}{\sqrt{w} m^3 - (s_i + |z|^2) m^2 - \sqrt{w} |z|^2 m + |z|^4}. \tag{2.16}$$

The following lemma gives the basic structure of $\operatorname{supp}\rho_{1,2c}$. Its proof will be given in Appendix 2.7.1.

Lemma 2.2.3. *Fix $\tau \leq ||z|^2-1| \leq \tau^{-1}$. The support of $\rho_{1,2c}$ is a union of connected components:*

$$\operatorname{supp}\rho_{1,2c} \cap (0,+\infty) = \left(\bigcup_{1\leq k\leq L} [e_{2k}, e_{2k-1}]\right) \cap (0,\infty), \tag{2.17}$$

where $L \equiv L(n) \in \mathbb{N}$ and $C_1\tau^{-1} \geq e_1 > e_2 > \ldots > e_{2L} \geq 0$ for some constant $C_1 > 0$ that does not depend on τ. If $|z|^2 \leq 1-\tau$, we have $e_{2L}=0$; if $1+\tau \leq |z|^2 \leq 1+\tau^{-1}$, we have $e_{2L} \geq \epsilon(\tau)$ for some constant $\epsilon(\tau)>0$. Moreover, for every $e_i > 0$, there exists a unique $m_c(e_i)$ such that

$$\partial_m f(\sqrt{e_i}, m_c(e_i)) = 0. \tag{2.18}$$

We shall call e_i the edges of $\rho_{1,2c}$. For any $w \in (0,\infty)$ and $1 \leq i \leq n$, the cubic polynomial $\sqrt{w}m^3 - (s_i+|z|^2)m^2 - \sqrt{w}|z|^2 m + |z|^4$ in (2.16) has three distinct roots $a_i(w)>0$, $b_i(w)>0$ and $-c_i(w)<0$ (see Lemma 2.7.1). Our next assumption on ρ_Σ and $|z|$ takes the form of the following regularity conditions.

Definition 2.2.4. *(Regularity) Fix $\tau \leq ||z|^2-1| \leq \tau^{-1}$.*

(i) We say that the edge $e_k \neq 0$, $k=1,\ldots,2L$, is regular if

$$\min_{1\leq i\leq n}\{|m_c(e_k)-a_i(e_k)|, |m_c(e_k)-b_i(e_k)|, |m_c(e_k)+c_i(e_k)|\} \geq \epsilon, \tag{2.19}$$

and

$$\left|\partial_m^2 f(\sqrt{e_k}, m_c(e_k))\right| \geq \epsilon, \tag{2.20}$$

for some small constant $\epsilon>0$. In the case $|z|^2 \leq 1-\tau$, we always call $e_{2L}=0$ a regular edge.

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

Remark 1: The edge regularity conditions (i) has previously appeared (may be in slightly different forms) in several works on sample covariance matrices and Wigner matrices [5, 19, 38, 44, 45, 50]. The conditions (2.19) and (2.20) guarantees a regular square-root behavior of ρ_{1c} near e_k and ensures that the gap in the spectrum of ρ_{1c} adjacent to e_k does not close for large N (Lemma 2.7.5):

$$\min_{l \neq k} |e_l - e_k| \geq \epsilon \tag{2.21}$$

for some constant $\epsilon > 0$. The bulk regularity condition (ii) was introduced in [44]. It imposes a lower bound on the density of eigenvalues away from the edges. Without it, one can have points in the interior of $\operatorname{supp} \rho_{1c}$ with an arbitrarily small density and our arguments would fail.

Remark 2: The regularity conditions in Definition 2.2.4 are stable under perturbations of $|z|$ and ρ_Σ. In particular, fix ρ_Σ, suppose the regularity conditions are satisfied at $z = z_0$ with $\tau \leq ||z_0|^2 - 1| \leq \tau^{-1}$. Then for sufficiently small $c > 0$, the regularity conditions hold uniformly in $z \in \{z : ||z| - |z_0|| \leq c\}$. For a detailed discussion, see the remark at the end of Section 2.7.3.

We will use the following notion of stochastic domination, which was first introduced in [20] and subsequently used in many works on random matrix theory, such as [6, 7, 10, 21, 22, 44]. It simplifies the presentation of the results and their proofs by systematizing statements of the form "ξ is bounded by ζ with high probability up to a small power of N".

Definition 2.2.5 (Stochastic domination)**.** *(i) Let*

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

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

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

for large enough $N \geq N_0(\epsilon, D)$, and we use the notation $\xi \prec \zeta$. Throughout this chapter the stochastic domination will always be uniform in all parameters that are not explicitly fixed (such as matrix indices, and w and z that take values in some compact sets). Note that $N_0(\epsilon, D)$ may depend on quantities that are explicitly constant, such as τ and C_p in (2.2), (2.4) and (2.5).

(ii) If for some complex family ξ we have $|\xi| \prec \zeta$, we also write $\xi \prec \zeta$ or $\xi = O_{\prec}(\zeta)$. We also extend the definition of $O_{\prec}(\cdot)$ to matrices in the weak operator sense as follows. Let A be a family of complex square random matrices and ζ a family of nonnegative random variables. Then we use $A = O_{\prec}(\zeta)$ to mean $|\langle \mathbf{v}, A\mathbf{w}\rangle| \prec \zeta \|\mathbf{v}\|_2 \|\mathbf{w}\|_2$ uniformly for all deterministic vectors $\mathbf{v}$ and $\mathbf{w}$.

(iii) We say that an event Ξ holds with high probability if $1 - 1(\Xi) \prec 0$.

In the following, we denote the eigenvalues of TX by μ_j, $1 \leq j \leq N$. We are now ready to state our main theorem, i.e. the generalized local circular law for TX.

Theorem 2.2.6 (Local circular law for TX)**.** *Suppose Assumption 2.2.1 holds, and $\tau \leq \left||z_0|^2 - 1\right| \leq \tau^{-1}$ for any N (z_0 can depend on N). Suppose ρ_Σ (defined in (2.7)) and $|z_0|$ are such that all the edges and bulk components of ρ_{1c} are regular in the sense of*

Definition 2.2.4. We assume in addition that each entry of X *has a density bounded by* N^{C_2} *for some* $C_2 > 0$*. Let* F *be a smooth non-negative function which may depend on* N*, such that* $\|F\|_\infty \le C_1$*,* $\|F'\|_\infty \le N^{C_1}$ *and* $F(z) = 0$ *for* $|z| \ge C_1$*, for some constant* $C_1 > 0$ *independent of* N*. Let* $F_{z_0,a}(z) = K^{2a} F(K^a(z - z_0))$*, where* $K := N \wedge M$*. Then* TX *has* $(N - K)$ *trivial zero eigenvalues, and for the other eigenvalues* μ_j*,* $1 \le j \le K$*, we have*

$$\frac{1}{K} \sum_{j=1}^{K} F_{z_0,a}(\mu_j) - \frac{1}{\pi} \int F_{z_0,a}(z) \tilde{\chi}_{\mathbb{D}}(z) dA(z) \prec K^{-1/2+2a} \|\Delta F\|_{L^1}, \tag{2.22}$$

for any $a \in (0, 1/4]$*. Here*

$$\tilde{\chi}_{\mathbb{D}}(z) := \frac{1}{4} \int_0^\infty (\log x) \Delta_z \rho_{2c}(x, z) dx, \tag{2.23}$$

where $\rho_{2c} \equiv \rho_{2c,z,\Sigma}$ *is defined in (2.13). If* $1 + \tau \le |z_0|^2 \le 1 + \tau^{-1}$ *or the entries of* X *have vanishing third moments,*

$$\mathbb{E} X_{i\mu}^3 = 0, \quad 1 \le i \le M, 1 \le \mu \le N, \tag{2.24}$$

then we have the improved result

$$\frac{1}{K} \sum_{j=1}^{K} F_{z_0,a}(\mu_j) - \frac{1}{\pi} \int F_{z_0,a}(z) \tilde{\chi}_{\mathbb{D}}(z) dA(z) \prec K^{-1+2a} \|\Delta F\|_{L^1}, \tag{2.25}$$

for any $a \in (0, 1/2]$*. If the entries of* X *are identically distributed, then the bounded density condition is not necessary.*

Remark 1: Note that $F_{z_0,a}(z) = K^{2a} F(K^a(z - z_0))$ is an approximate delta function obtained from rescaling F to the size of order K^{-a} around z_0. Thus (2.22) gives a generalized circular law up to scale $K^{-1/4+\epsilon}$, while (2.25) gives a generalized circular law up to scale $K^{-1/2+\epsilon}$. The $\tilde{\chi}_{\mathbb{D}}$ in (2.23) gives the distribution of the eigenvalues of TX. It is rotationally symmetric, because $\rho_{2c}(x,z)$ depends only on $|z|$ (see (2.10) and (2.11)). If $TT^* = 1$ or $T^*T = 1$ (i.e. all the nontrivial singular values of T are equal to 1), then $\tilde{\chi}_{\mathbb{D}}$ becomes the indicator function $\chi_{\mathbb{D}}$ on the unit disk $\mathbb{D}$, and we get the well-known local circular law for X (see [10] for the $T = I$ case). For a general T, we do not have much understanding of $\tilde{\chi}_{\mathbb{D}}$ so far. This will be one of the topics of our future study. Also, we have assumed that z is strictly away from the unit circle. Our proof may be extended to the $|z-1| = o(1)$ case if we have a better understanding of the solutions $m_{1,2c}$.

Remark 2: As explained in the Introduction, our strategy is first to prove the anisotropic local law for the resolvent of Q when X is Gaussian, and then to get the anisotropic local law for a general X through a comparison with the Gaussian case. Without (2.24), our comparison arguments cannot give the anisotropic local law up to the optimal scale, so we can only prove the weaker bound (2.22). We will try to remove this assumption in the future work.

Remark 3: If the entries of X are identically distributed, then it was proved in [71] that the smallest singular value of $TX - z$ is larger than $N^{-1-\epsilon}$ with high probability for any $\epsilon > 0$. Otherwise, we need the extra bounded density condition, which is only used in Lemma 2.2.23 to get a lower bound for the smallest singular value of $TX - z$.

We conclude this section with two examples verifying the regularity conditions of Definition 2.2.4.

Example 2.2.7 (Bounded number of distinct eigenvalues)**.** *We suppose that* n *is fixed, and that* $s_1, \ldots, s_n$ *and* $\rho_\Sigma(\{s_1\}), \ldots, \rho_\Sigma(\{s_n\})$ *all converge as* $N \to \infty$*. We suppose that* $\lim_N e_k > \lim_N e_{k+1}$ *for all* k*, and furthermore for all* e_k *we have* $\partial_m^2 f(\sqrt{e_k}, m_c(e_k)) \neq 0$.

Then it is easy to check that all the edges and bulk components are regular in the sense of Definition 2.2.4 for small enough ϵ*.*

Example 2.2.8 (Continuous limit)**.** *We suppose* ρ_Σ *is supported in some interval* $[a,b]\subset(0,\infty)$*, and that* ρ_Σ *converges in distribution to some measure* ρ_∞ *that is absolutely continuous and whose density satisfies* $\tau \le d\rho_\infty(E)/dE \le \tau^{-1}$ *for* $E\in[a,b]$*. Then there are only a small number (which is independent of* n*) of connected components for* $\operatorname{supp}\rho_{1c}$*, and all the edges and bulk components are regular. See the remark at the end of Section 2.7.1.*

2.2.3 Hermitization and local laws for resolvents

In the following, we use the notation

$$Y \equiv Y_z := TX - zI, \tag{2.26}$$

where I is the identity matrix. Following Girko's Hermitization technique [34], the first step in proving the local circular law is to understand the local statistics of singular values of Y. In this subsection, we present the main local estimates concerning the resolvents $\left(YY^\dagger - w\right)^{-1}$ and $\left(Y^\dagger Y - w\right)^{-1}$. These results will be used later to prove Theorem 2.2.6.

Our local laws can be formulated in a simple, unified fashion using a $2N\times 2N$ block matrix, which is a linear function of X.

Definition 2.2.9 (Index sets)**.** *We define the index sets*

$$\mathcal{I}_1 := \{1,...,N\},\quad \mathcal{I}_1^M := \{1,\ldots,M\},\quad \mathcal{I}_2 := \{N+1,...,2N\},\quad \mathcal{I} := \mathcal{I}_1\cup\mathcal{I}_2,\quad \mathcal{I}^M := \mathcal{I}_1^M\cup\mathcal{I}_2\,.$$

We will consistently use the latin letters $i, j \in \mathcal{I}_1$ or $\mathcal{I}_1^M$, greek letters $\mu, \nu \in \mathcal{I}_2$, and $s, t \in \mathcal{I}$. We label the indices of the matrices according to

$$X = (X_{i\mu} : i \in \mathcal{I}_1^M, \mu \in \mathcal{I}_2), \quad T = (T_{ij} : i \in \mathcal{I}_1, j \in \mathcal{I}_1^M).$$

When $M = N$, we always identify $\mathcal{I}_1^M$ with $\mathcal{I}_1$. For $i \in \mathcal{I}_1$ and $\mu \in \mathcal{I}_2$, we introduce the notations $\bar{i} := i + N \in \mathcal{I}_2$ and $\bar{\mu} := \mu - N \in \mathcal{I}_1$.

Definition 2.2.10 (Groups)**.** *For an $\mathcal{I} \times \mathcal{I}$ matrix A, we define the 2×2 matrices $A_{[ij]}$ as*

$$A_{[ij]} = \begin{pmatrix} A_{ij} & A_{i\bar{j}} \\ A_{\bar{i}j} & A_{\bar{i}\bar{j}} \end{pmatrix}. \tag{2.27}$$

We shall call $A_{[ij]}$ a diagonal group if $i = j$, and an off-diagonal group otherwise .

Definition 2.2.11 (Linearizing block matrix)**.** *For $w := E + i\eta \in \mathbb{C}_+$, we define the $\mathcal{I} \times \mathcal{I}$ matrix*

$$H(w) \equiv H(T, X, z, w) := \begin{pmatrix} -wI & w^{1/2}Y \\ w^{1/2}Y^\dagger & -wI \end{pmatrix}, \tag{2.28}$$

where we take the branch of $\sqrt{w}$ with positive imaginary part. Define the $\mathcal{I} \times \mathcal{I}$ matrix

$$G(w) \equiv G(T, X, z, w) := H(w)^{-1}, \tag{2.29}$$

as well as the $\mathcal{I}_1 \times \mathcal{I}_1$ and $\mathcal{I}_2 \times \mathcal{I}_2$ matrices

$$G_L(w) = \left(YY^\dagger - w\right)^{-1}, \quad G_R(w) = \left(Y^\dagger Y - w\right)^{-1}. \tag{2.30}$$

Throughout the rest of this chapter, we frequently omit the argument w from our notations.

By Schur's complement formula, it is easy to see that

$$G(w) = \begin{pmatrix} G_L & w^{-1/2} G_L Y \\ w^{-1/2} Y^\dagger G_L & w^{-1} Y^\dagger G_L Y - w^{-1} I \end{pmatrix} = \begin{pmatrix} w^{-1} Y G_R Y^\dagger - w^{-1} I & w^{-1/2} Y G_R \\ w^{-1/2} G_R Y^\dagger & G_R \end{pmatrix}. \tag{2.31}$$

Therefore a control of G immediately yields controls of the resolvents G_L and G_R.

In the following, we only consider the $N \le M$ case. The $N > M$ case, as we will see, will be built easily upon $N \le M$ case. We introduce a deterministic matrix Π, which will turn out to be close to G with high probability.

Definition 2.2.12 (Deterministic limit of G)**.** *Suppose $N \le M$ and T has a singular decomposition*

$$T = U\bar{D}V, \quad \bar{D} = (D, 0), \tag{2.32}$$

where $D = diag(d_1, d_2, \ldots, d_N)$ is a diagonal matrix. Define $\pi_{[i]c}$ to be the 2×2 matrix such that

$$\left(\pi_{[i]c}\right)^{-1} = \begin{pmatrix} -w(1 + |d_i|^2 m_{2c}) & -w^{1/2} z \\ -w^{1/2} \bar{z} & -w(1 + m_{1c}) \end{pmatrix}. \tag{2.33}$$

Let Π_d be the $2N \times 2N$ matrix with $(\Pi_d)_{[ii]} = \pi_{[i]c}$ and all other entries being zero. Define

$$\Pi \equiv \Pi(\Sigma, z, w) := \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix} \Pi_d \begin{pmatrix} U^\dagger & 0 \\ 0 & U^\dagger \end{pmatrix} = \begin{pmatrix} -(1 + m_{1c}) A(\Sigma) & w^{-1/2} z A(\Sigma) \\ w^{-1/2} \bar{z} A(\Sigma) & -(1 + m_{2c}\Sigma) A(\Sigma) \end{pmatrix}, \tag{2.34}$$

where $\Sigma = TT^\dagger$ and $A(\Sigma) = \left[w(1 + m_{2c}\Sigma)(1 + m_{1c}) - |z|^2\right]^{-1}$.

Definition 2.2.13 (Averaged variables)**.** *Suppose $N \le M$. Define the averaged random variables*

$$m_1 := \frac{1}{N} \sum_{i \in \mathcal{I}_1} \left(\bar{\Sigma} G\right)_{ii}, \quad m_2 := \frac{1}{N} \sum_{\mu \in \mathcal{I}_2} \left(\bar{\Sigma} G\right)_{\mu\mu}, \tag{2.35}$$

where

$$\bar{\Sigma} := \begin{pmatrix} \Sigma & 0 \\ 0 & I \end{pmatrix}. \tag{2.36}$$

Define $\pi_{[i]}$ *to be the* 2×2 *matrix such that*

$$\left(\pi_{[i]}\right)^{-1} = \begin{pmatrix} -w(1+|d_i|^2 m_2) & -w^{1/2}z \\ -w^{1/2}\bar{z} & -w(1+m_1) \end{pmatrix}. \tag{2.37}$$

Remark: Note that under the above definition we have

$$m_2 = \frac{1}{N}\operatorname{Tr} G_R = \frac{1}{N}\operatorname{Tr} G_L,$$

which is the Stieltjes transform of the empirical eigenvalue density of $YY^\dagger$ and $Y^\dagger Y$. Moreover, we will see from the proof that $m_{1,2c}$ are the almost sure limits of $m_{1,2}$ as $N \to \infty$ with

$$m_{1c} = \frac{1}{N}\sum_{i\in\mathcal{I}_1} \left(\bar{\Sigma}\Pi\right)_{ii}, \quad m_{2c} = \frac{1}{N}\sum_{\mu\in\mathcal{I}_2} \left(\bar{\Sigma}\Pi\right)_{\mu\mu}. \tag{2.38}$$

The following two propositions summarize the properties of $\rho_{1,2c}$ and $m_{1,2c}$ that are needed to understand the main results in this section. They will be proved in Appendix 2.7. In Fig. 2, we plot ρ_{2c} for the example from Fig. 1 for different values of z.

Proposition 2.2.14 (Basic properties of $\rho_{1,2c}$)**.** *The density* ρ_{1c} *is compactly supported in* $[0,\infty)$ *and the following properties regarding* ρ_{1c} *hold.*

(i) The support of ρ_{1c} *is* $\bigcup_{1\le k\le L(n)}[e_{2k}, e_{2k-1}]$ *where* $e_1 > e_2 > \ldots > e_{2L} \ge 0$*. If* $1+\tau \le |z|^2 \le 1+\tau^{-1}$*, then* $e_{2L} \ge \epsilon$ *for some constant* $\epsilon > 0$*; if* $|z|^2 \le 1-\tau$*, then* $e_{2L} = 0$*.*

(ii) Suppose $[e_{2k}, e_{2k-1}]$ *is a regular bulk component. For any* $\tau' > 0$*, if* $x \in [e_{2k}+\tau', e_{2k-1}-\tau']$*, then* $\rho_{1c}(x) \sim 1$*.*

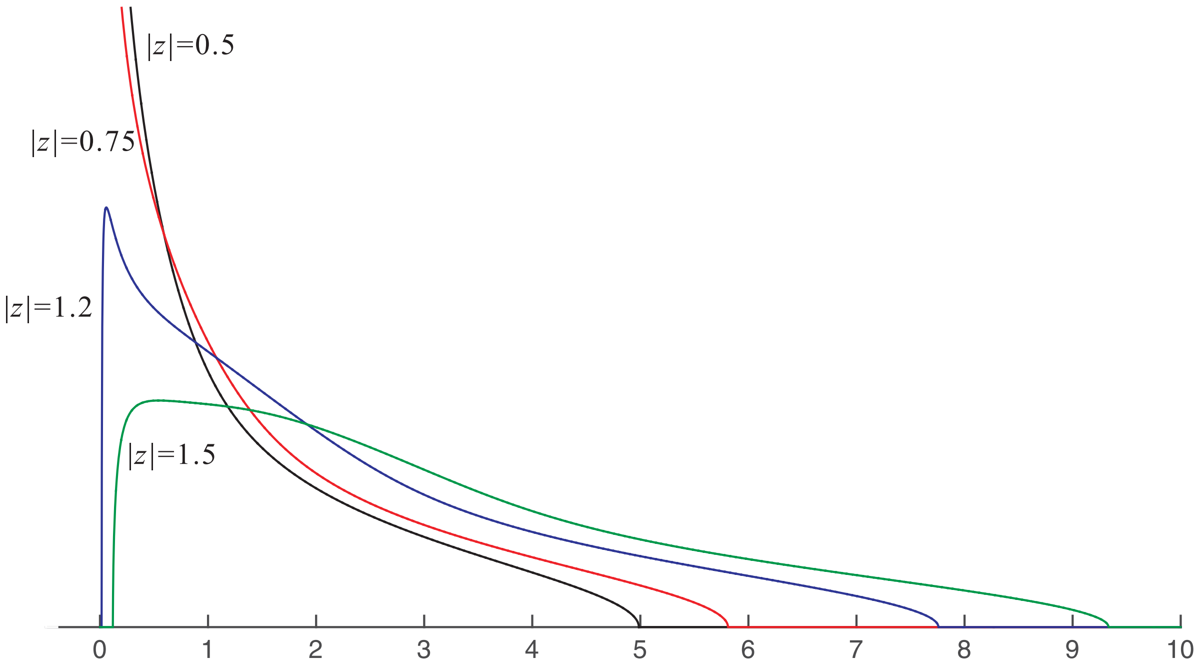


Figure 2: The densities $\rho_{2c}(x,z)$ when $|z| = 0.5,\ 0.75,\ 1.2,\ 1.5$. Here $\rho_\Sigma = 0.5\delta_{\sqrt{2/17}} + 0.5\delta_{4\sqrt{2/17}}$.

(iii) Suppose e_j *is a nonzero regular edge. If* j *is even, then* $\rho_{1c}(x) \sim \sqrt{x-e_j}$ *as* $x \to e_j$ *from above. Otherwise if* j *is odd, then* $\rho_{1c}(x) \sim \sqrt{e_j - x}$ *as* $x \to e_j$ *from below.*

(iv) If $|z|^2 \le 1-\tau$, *then* $\rho_{1c}(x) \sim x^{-1/2}$ *as* $x \searrow e_{2L} = 0$.

The same results also hold for ρ_{2c}. *In addition,* ρ_{2c} *is a probability density.*

Proposition 2.2.15. *The preceding proposition implies that, uniformly in* w *in any compact set of* $\mathbb{C}_+$,

$$|m_{1,2c}(w)| = O(|w|^{-1/2}). \tag{2.39}$$

Moreover, if $1+\tau \le |z|^2 \le 1+\tau^{-1}$, *then* $|m_{1,2c}(w)| \sim 1$ *for* w *in any compact set of* $\mathbb{C}_+$; *if* $|z|^2 \le 1-\tau$, *then* $|m_{1,2c}(w)| \sim |w|^{-1/2}$ *for* w *in any compact set of* $\mathbb{C}_+$.

We will consistently use the notation $E + i\eta$ for the spectral parameter w. In this chapter, we regard the quantities $E(w)$ and $\eta(w)$ as functions of w and usually omit the argument w. In the following we would like to define several spectral domains of w that will be used in the proof.

Definition 2.2.16 (Spectral domains)**.** *Fix a small constant* $\zeta > 0$ *which may depend on* τ*. The spectral parameter* w *is always assumed to be in the fundamental domain*

$$\mathbf{D} \equiv \mathbf{D}(\zeta, N) := \{w \in \mathbb{C}_+ : \zeta e_{2L} \leq E \leq \zeta^{-1}, N^{-1+\zeta}|m_{2c}|^{-1} \leq \eta \leq \zeta^{-1}\}, \tag{2.40}$$

unless otherwise indicated. Given a regular edge e_k*, we define the subdomain*

$$\mathbf{D}_k^e \equiv \mathbf{D}_k^e(\zeta, \tau', N) := \{w \in \mathbf{D}(\zeta, N) : |E - e_k| \leq \tau', E \geq 0\}. \tag{2.41}$$

Corresponding to a regular bulk component $[e_{2k}, e_{2k-1}]$*, we define the subdomain*

$$\mathbf{D}_k^b \equiv \mathbf{D}_k^b(\zeta, \tau', N) := \{w \in \mathbf{D}(\zeta, N) : E \in [e_{2k} + \tau', e_{2k-1} - \tau']\}. \tag{2.42}$$

For the component outside $\operatorname{supp} \rho_{1c}$*, we define the subdomain*

$$\mathbf{D}^o \equiv \mathbf{D}^o(\zeta, \tau', N) := \{w \in \mathbf{D}(\zeta, N) : \operatorname{dist}(E, \operatorname{supp} \rho_{1c}) \geq \tau'\}. \tag{2.43}$$

We also need the following domain with large η*,*

$$\mathbf{D}_L \equiv \mathbf{D}_L(\zeta) := \{w \in \mathbb{C}_+ : 0 \leq E \leq \zeta^{-1}, \eta \geq \zeta^{-1}\}, \tag{2.44}$$

and the subdomain of $\mathbf{D} \cup \mathbf{D}_L$*,*

$$\widehat{\mathbf{D}} \equiv \widehat{\mathbf{D}}(\zeta, N) := \{w \in \mathbf{D}(\zeta, N) : \eta \geq N^{-1/2+\zeta}|m_{2c}|^{-1}\} \cup \mathbf{D}_L(\zeta). \tag{2.45}$$

We call $\mathbf{S}$ *a regular domain if it is a regular* $\mathbf{D}_k^e$ *or* $\mathbf{D}_k^b$ *domain, a* $\mathbf{D}^o$ *domain or a* $\mathbf{D}_L$ *domain.*

Remark: In the definition of $\mathbf{D}$, we have suppressed the explicit w-dependence. Notice that when $|z|^2 < 1 - \tau$, since $|m_{2c}| \sim |w|^{-1/2}$ as $w \to 0$, we allow $\eta \sim |w| \sim N^{-2+2\zeta}$ in $\mathbf{D}$. In the definition of $\mathbf{D}_k^e$, the condition $E \geq 0$ is only useful for the edge at 0 when $|z|^2 \leq 1 - \tau$.

Now we are prepared to state the local laws satisfied by G defined in (3.27). Let

$$\Psi \equiv \Psi(w) := \sqrt{\frac{\operatorname{Im}(m_{1c} + m_{2c})}{N\eta}} + \frac{1}{N\eta} \tag{2.46}$$

be the deterministic control parameter.

Definition 2.2.17 (Local laws)**.** *Suppose $N \leq M$. Recall $G \equiv G(T, X, z, w)$ defined in (3.27) and $\Pi \equiv \Pi(\Sigma, z, w)$ defined in (2.34). Let $\mathbf{S}$ be a regular domain.*

(i) We say that the entrywise local law holds with parameters $(T, X, z, \mathbf{S})$ if

$$[G(T, X, z, w) - \Pi(\Sigma, z, w)]_{st} \prec \Psi(w) \tag{2.47}$$

uniformly in $w \in \mathbf{S}$ and $s, t \in \mathcal{I}$.

(ii) We say that the anisotropic local law holds with parameters $(T, X, z, \mathbf{S})$ if

$$G(T, X, z, w) - \Pi(\Sigma, z, w) = O_{\prec}(\Psi(w)) \tag{2.48}$$

uniformly in $w \in \mathbf{S}$ (recall Definition 3.2.4 (ii)).

(iii) We say that the averaged local law holds with parameters $(T, X, z, \mathbf{S})$ if

$$|m_2(T, X, z, w) - m_{2c}(\Sigma, z, w)| \prec \frac{1}{N\eta} \tag{2.49}$$

uniformly in $w \in \mathbf{S}$.

The local laws for G with a general T will be built upon the following result with a diagonal T.

Theorem 2.2.18 (Local laws when T is diagonal)**.** *Fix $\tau \le ||z|^2 - 1| \le \tau^{-1}$. Suppose Assumption 2.2.1 holds, $N = M$, and $T \equiv D := diag(d_1, ..., d_N)$ is a diagonal matrix. Let $\mathbf{S}$ be a regular domain. Then the entrywise local law, anisotropic local law and averaged local law hold with parameters $(D, X, z, \mathbf{S})$.*

Now suppose that $N \le M$ and T is an $N \times M$ matrix such that the eigenvalues of Σ satisfy (2.5) and (2.6). Consider the singular decomposition $T = U\bar{D}V$, where U is an $N \times N$ unitary matrix, V is an $M \times M$ unitary matrix and $\bar{D} = (D, 0)$ is an $N \times M$ matrix such that $D = \mathrm{diag}(d_1, d_2, \dots, d_N)$. Then we have

$$TX - z = UDV_1X - z, \tag{2.50}$$

where V_1 is an $N \times M$ matrix and V_2 is an $(M - N) \times M$ matrix defined through $V = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}$. If $X = X^{Gauss}$ is Gaussian, then $V_1 X^{Gauss} \overset{d}{=} \tilde{X}^{Gauss} U^\dagger$, where $\tilde{X}^{Gauss}$ is another $N \times N$ Gaussian random matrix. Then by the definition of G in (3.27),

$$G(T, X^{Gauss}, z, w) \overset{d}{=} \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix} G(D, \tilde{X}^{Gauss}, z, w) \begin{pmatrix} U^\dagger & 0 \\ 0 & U^\dagger \end{pmatrix}. \tag{2.51}$$

Since the anisotropic local law holds for $G(D, \tilde{X}^{Gauss}, z, w)$ by Theorem 2.2.18, we get immediately the anisotropic local law for $G(T, X^{Gauss}, z, w)$. The next theorem states that the anisotropic local law holds for general TX provided that the anisotropic local law holds for TX^{Gauss}.

Theorem 2.2.19 (Anisotropic local law when $N \le M$)**.** *Fix $\tau \le ||z|^2 - 1| \le \tau^{-1}$. Suppose Assumption 2.2.1 holds and $N \le M$. Let $T = U\bar{D}V$ be a singular decomposition of T, where $\bar{D} = (D, 0)$ with $D = diag(d_1, d_2, \dots, d_N)$. Let $\mathbf{S}$ be a regular domain. Then* *the anisotropic local law and averaged local law hold with parameters* $(T, X, z, \mathbf{S} \cap \widehat{\mathbf{D}})$. *If in addition (2.24) holds, then the anisotropic local law and averaged local law hold with parameters* $(T, X, z, \mathbf{S})$.

Finally we turn to the $N > M$ case. Suppose $T = U\bar{D}V$ is a singular decomposition of T, where U is an $N \times N$ unitary matrix, V is an $M \times M$ unitary matrix and $\bar{D} = \begin{pmatrix} D \\ 0 \end{pmatrix}$ is an $N \times M$ matrix such that $D = \operatorname{diag}(d_1, d_2, \ldots, d_M)$. Let $U = (U_1, U_2)$, where U_1 has size $N \times M$ and U_2 has size $N \times (N - M)$. Following Girko's idea of Hermitization [34], to prove the local circular law in Theorem 2.2.6 when $N > M$, it suffices to study $\det(TX - z)$ (see (2.53) below), for which we have

$$\det(TX - z) = \det \begin{pmatrix} DVXU_1 - z & DVXU_2 \\ 0 & -z \end{pmatrix} = \det(V^T D^T U_1^T X^T - z)(-z)^{N-M}. \tag{2.52}$$

Comparing with (2.50), we see that this case is reduced to the $N \le M$ case. The only difference is that the extra $(-z)^{N-M}$ term now corresponds to the $N - M$ zero eigenvalues of TX. Thus we make the following claim.

Claim 2.2.20. *The* $N < M$ *case of Theorem 2.2.6 implies the* $N > M$ *case of Theorem 2.2.6.*

2.2.4 Proof of Theorem 2.2.6

By Claim 2.2.20, it suffices to assume $N \le M$. Our main tool will be Theorem 2.2.19. A major part of the proof follows from [10, Section 5]. The following lemma collects basic properties of stochastic domination $\prec$, which will be used tacitly during the proof and throughout this chapter.

Lemma 2.2.21 (Lemma 3.2 in [6])**.** *(i) Suppose that* $\xi(u,v) \prec \zeta(u,v)$ *uniformly in* $u \in U$ *and* $v \in V$*. If* $|V| \le N^C$ *for some constant* C*, then*

$$\sum_{v\in V} \xi(u,v) \prec \sum_{v\in V} \zeta(u,w)$$

uniformly in u*.*

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

$$\xi_1(u)\xi_2(u) \prec \zeta_1(u)\zeta_2(u)$$

uniformly in $u \in U$*.*

(iii) Suppose that $\Psi(u) \ge N^{-C}$ *is deterministic and* $\xi(u)$ *is a nonnegative random variable such that* $E\xi(u)^2 \le N^C$ *for all* u*. Then if* $\xi(u) \prec \Psi(u)$ *uniformly in* u*, we have*

$$\mathbb{E}\xi(u) \prec \Psi(u)$$

uniformly in u*.*

The Girko's Hermitization technique [34] can be reformulated as the following (see e.g. [37]): for any smooth function g,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} g(\mu_j) &= \frac{1}{4\pi N}\int \Delta g(z) \sum_{j=1}^{N} \log(\mu_j - z)(\bar{\mu}_j - \bar{z}) dA(z) \\ &= \frac{1}{4\pi N}\int \Delta g(z) \log\left|\det(Y(z)Y^\dagger(z))\right| dA(z) = \frac{1}{4\pi N}\int \Delta g(z) \sum_{j=1}^{N} \log\lambda_j(z) dA(z),\end{aligned} \tag{2.53}$$

where $0 \le \lambda_1 \le \lambda_2 \le \ldots \le \lambda_N$ are the ordered eigenvalues of $Y(z)Y^\dagger(z)$. For $g = F_{z_0,a}$, we use the new variable $\xi = N^a(z - z_0)$ to write the above equation as

$$\frac{1}{N}\sum_{i=1}^{N} F_{z_0,a}(\mu_j) = \frac{N^{-1+2a}}{4\pi}\int (\Delta F)(\xi) \sum_{j=1}^{N} \log\lambda_j(z) dA(\xi). \tag{2.54}$$

Define the classical location $\gamma_j(z)$ of the j-th eigenvalue of $Y(z)Y^\dagger(z)$ by

$$\gamma_j(z) := \sup_x \left\{ \int_0^x \rho_{2c}(x)dx \le \frac{j}{N} \right\}, \quad 1 \le j \le N. \tag{2.55}$$

In fact, if γ_j lies in the bulk, then by the positivity of ρ_{2c} we can simply define γ_j through

$$\int_0^{\gamma_j} \rho_{2c}(x)dx = \frac{j}{N}.$$

By Proposition 2.2.14, we have that for any $\delta > 0$,

$$\left| \sum_{j=1}^N \log \gamma_j(z) - N \int_0^\infty (\log x)\rho_{2c}(x,z)dx \right| \le \sum_{j=1}^N \int_{\gamma_{j-1}(z)}^{\gamma_j(z)} \left|\log \gamma_j(z) - \log x\right| \rho_{2c}(x,z)dx \le N^\delta \tag{2.56}$$

for large enough N. Suppose we have the bound

$$\left| \sum_j \log \lambda_j - \sum_j \log \gamma_j \right| \prec N^b. \tag{2.57}$$

Plugging (2.56) and (2.57) into (2.54), we get

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^N F_{z_0}(\mu_j) &= \frac{N^{2a}}{4\pi} \int (\Delta F)(\xi) \int_0^\infty (\log x)\rho_{2c}(x,z)dxdA(\xi) + O_\prec(N^{-1+b+2a}\|\Delta F\|_{L_1}) \\ &= \frac{1}{4\pi} \int F(\xi) \int_0^\infty (\log x)\Delta_z \rho_{2c}(x,z)dxdA(\xi) + O_\prec(N^{-1+b+2a}\|\Delta F\|_{L_1}). \end{aligned}$$

Thus we obtain (2.22) if we can prove (2.57) for $b = 1/2$, and we obtain (2.25) if we can prove (2.57) for $b = 0$ when $1+\tau \le |z_0|^2 \le 1+\tau^{-1}$ or when the assumption (2.24) holds.

We need the following lemma which is a consequence of Theorem 2.2.19. Recall (2.17) and (2.21), the number L of the connected components is of order 1 and each component $[e_{2k}, e_{2k-1}]$ contains order N of γ_j's. We define the classical number of eigenvalues to the left of e_k, $1 \le k \le 2L$, as

$$N_k := \left\lceil N \int_0^{e_k} \rho_{2c}(x) \right\rceil. \tag{2.58}$$

Note that $N_{2L} = 0$, $N_1 = N$ and $N_{2k+1} = N_{2k}$, $1 \le k \le L-1$.

Lemma 2.2.22 (Singular value rigidity)**.** *Fix a small* $\epsilon > 0$*.*

(i) If the averaged local law holds with parameters $(T, X, z, \mathbf{D}(\zeta, N) \cap \widehat{\mathbf{D}}(\zeta, N))$ *for arbitrarily small* ζ*, then the following estimates hold. For any* $e_{2k} > 0$ *and* $N_{2k}+N^{1/2+\epsilon} \leq j \leq N_{2k-1} - N^{1/2+\epsilon}$*,*

$$\frac{|\lambda_j - \gamma_j|}{\gamma_j} \prec \left(\min \left\{ \frac{j - N_{2k}}{N}, \frac{N_{2k-1} - j}{N} \right\} \right)^{-1/3} N^{-1/2}. \tag{2.59}$$

In the case $|z|^2 \leq 1 - \tau$ *with* $e_{2L} = 0$*, we have for any* $N^{1/2+\epsilon} \leq j \leq N_{2L-1} - N^{1/2+\epsilon}$*,*

$$\frac{|\lambda_j - \gamma_j|}{\gamma_j} \prec j^{-1} \left(\frac{N_{2L-1} - j}{N} \right)^{-1/3} N^{1/2}. \tag{2.60}$$

Moreover, if $1 + \tau \leq |z|^2 \leq 1 + \tau^{-1}$*, then for any fixed* $0 < c < e_{2L}$*,*

$$\#\{j : 0 < \lambda_j < c\} \prec 1. \tag{2.61}$$

(ii) If the averaged local law holds with parameters $(T, X, z, \mathbf{D}(\zeta, N))$ *for arbitrarily small* ζ*, then the following estimates hold. For any* $e_{2k} > 0$ *and* $N_{2k} + N^{\epsilon} \leq j \leq N_{2k-1} - N^{\epsilon}$*,*

$$\frac{|\lambda_j - \gamma_j|}{\gamma_j} \prec \left(\min \left\{ \frac{j - N_{2k}}{N}, \frac{N_{2k-1} - j}{N} \right\} \right)^{-1/3} N^{-1}. \tag{2.62}$$

In the case $|z|^2 \leq 1 - \tau$ *with* $e_{2L} = 0$*, we have for any* $N^{\epsilon} \leq j \leq N_{2L-1} - N^{\epsilon}$*,*

$$\frac{|\lambda_j - \gamma_j|}{\gamma_j} \prec j^{-1} \left(\frac{N_{2L-1} - j}{N} \right)^{-1/3}. \tag{2.63}$$

Proof. The proof is similar to the proof of [10, Lemma 5.1]. See also [6, Theorem 2.10] or [22, Theorem 7.6] □

Using (2.59) and (2.60), we get that

$$\sum_{N_{2k}+N^{1/2+\epsilon} \leq j \leq N_{2k-1} - N^{1/2+\epsilon}} |\log \lambda_j - \log \gamma_j| \prec \sum_{N_{2k}+N^{1/2+\epsilon} \leq j \leq N_{2k-1} - N^{1/2+\epsilon}} \frac{|\lambda_j - \gamma_j|}{\gamma_j} \prec N^{1/2}. \tag{2.64}$$

By Theorem 2.10 of [6], there exists a constant $C > 0$ such that

$$\|X^*X\| \le C \ \text{ with high probability.} \tag{2.65}$$

Thus we have

$$\lambda_j \le \|Y\|^2 \le (\|T\|\|X\| + |z|)^2 \prec 1, \quad 1 \le j \le N. \tag{2.66}$$

Together with Lemma 2.2.23 concerning the smallest singular value of $TX - z$, we get

$$\sum_{k=1}^{2L} \sum_{|j-e_k|<N^{1/2+\epsilon}} |\log \lambda_j| \prec N^{1/2+\epsilon}. \tag{2.67}$$

Since $|\log \gamma_j| \prec 1$ by Proposition 2.2.14, we conclude

$$\sum_{k=1}^{2L} \sum_{|j-e_k|<N^{1/2+\epsilon}} |\log \lambda_j - \log \gamma_j| \prec N^{1/2+\epsilon}. \tag{2.68}$$

Combining (2.64) and (2.68), we get for any $\epsilon > 0$,

$$\sum_{1\le j\le N} |\log \lambda_j - \log \gamma_j| \prec N^{1/2+\epsilon} \tag{2.69}$$

for large enough N. This implies (2.57) for $b = 1/2$. If in addition the assumption (2.24) holds, the averaged local law holds with parameters $(T, X, z, \mathbf{D}(\zeta, N))$ for arbitrarily small ζ by Theorem 2.2.19. Then we can prove (2.57) for $b = 0$ using the better bounds (2.62) and (2.63).

Finally we show that when $|z_0|^2 \ge 1 + \tau$, with the bounds (2.59) we can still prove the estimate (2.57) for $b = 0$. By the averaged local law and the definition of γ_j in (2.55), we have

$$\left|\sum_{j=1}^{N} \frac{1}{\lambda_j - i\eta} - \sum_{j=1}^{N} \frac{1}{\gamma_j - i\eta}\right| \prec \frac{1}{\eta}, \tag{2.70}$$

uniformly in $N^{-1/2+\epsilon} \le \eta \le N^{1/2}$. Taking integral of (2.70) over η from $N^{-1/2+\epsilon}$ to $N^{1/2}$, we get

$$\left|\sum_{j=1}^{N} \log\left(\frac{\lambda_j - iN^{-1/2+\epsilon}}{\gamma_j - iN^{-1/2+\epsilon}}\right) - \sum_{j=1}^{N} \log\left(\frac{\lambda_j - iN^{1/2}}{\gamma_j - iN^{1/2}}\right)\right| \prec 1. \tag{2.71}$$

Then we use (2.59) and the bound (2.66) to estimate that

$$\left|\sum_{j=1}^{N} \log\left(\frac{\lambda_j - iN^{1/2}}{\gamma_j - iN^{1/2}}\right)\right| \prec \sum_{j=1}^{N} \left|(\lambda_j - \gamma_j)\, N^{-1/2}\right| \prec N^{\epsilon}.$$

Thus we conclude

$$\left|\sum_{j=1}^{N} \log\left(\frac{\lambda_j - iN^{-1/2+\epsilon}}{\gamma_j - iN^{-1/2+\epsilon}}\right)\right| \prec N^{\epsilon}. \tag{2.72}$$

Using $\gamma_j \sim 1$, (2.61) and (2.74), we get

$$\begin{aligned}\left|\sum_{j=1}^{N} \log\left(\frac{\lambda_j - iN^{-1/2+\epsilon}}{\gamma_j - iN^{-1/2+\epsilon}}\right) - \sum_{j=1}^{N} \log\frac{\lambda_j}{\gamma_j}\right| &\prec 1 + \left|\sum_{\lambda_j \geq c} \log\left(\frac{\lambda_j - iN^{-1/2+\epsilon}}{\gamma_j - iN^{-1/2+\epsilon}}\right) - \sum_{\lambda_j \geq c} \log\frac{\lambda_j}{\gamma_j}\right| \\ &\prec 1 + \sum_{\lambda_j \geq c} \left|(\lambda_j - \gamma_j)\, N^{-1/2+\epsilon}\right| \prec N^{2\epsilon}. \end{aligned} \tag{2.73}$$

Combing (2.72) and (2.73), we conclude (2.57) for $b = 0$.

If the entries of X are identically distributed, then instead of Lemma 2.2.23, we shall use the results in [71] to get a lower bound for the smallest singular value of $TX - z$ (see Remark 3 below Theorem 2.2.6). In particular, the bounded density condition for the entries of X is not needed anymore. This concludes the last statement of Theorem 2.2.6.

Lemma 2.2.23 (Lower bound on the smallest singular value)**.** *If $N \leq M$ and the entries of X have a density bounded by N^{C_3} for some $C_3 > 0$, then*

$$|\log \lambda_1(z)| \prec 1 \tag{2.74}$$

holds uniformly for z in any fixed compact set.

Proof. We already have an upper bound for λ_1; see (2.66). Hence to get (2.74), we still need to prove that

$$\mathbb{P}\left(\lambda_1(z) \leq e^{-N^{\epsilon}}\right) \leq N^{-C} \tag{2.75}$$

for any $\epsilon, C > 0$. By (2.50), we have that

$$TX - z = UD(V_1 X - D^{-1}U^{-1}z) =: UD\tilde{Y}(z).$$

Hence it suffices to control the smallest singular value of $\tilde{Y}(z)$, call it $\tilde{\lambda}_1(z)$. Notice the columns $\tilde{Y}_1, \ldots, \tilde{Y}_N$ of $\tilde{Y}(z)$ are independent vectors. From the variational characterization

$$\tilde{\lambda}_1(z) = \min_{|u|=1} \|\tilde{Y}(z)u\|^2,$$

we can easily get

$$\tilde{\lambda}_1(z)^{1/2} \geq N^{-1/2} \min_{1\leq k\leq N} \operatorname{dist}\left(\tilde{Y}_k, \operatorname{span}\{\tilde{Y}_l, l \neq k\}\right) = N^{-1/2} \min_{1\leq k\leq N} \left|\langle \tilde{Y}_k, u_k\rangle\right|, \tag{2.76}$$

where u_k is the unit normal vector of $\operatorname{span}\{\tilde{Y}_l, l \neq k\}$ and hence is independent of $\tilde{Y}_k$. By conditioning on u_k, we get immediately that

$$\mathbb{P}(\tilde{\lambda}_1(z) \leq N^{-C_0}) \leq CN^{-C_0/2+C_3+3/2}, \tag{2.77}$$

which is a much stronger result than (2.75). Here we have used Theorem 1.2 of [58] to conclude that $\langle \bar{Y}_k, u_k\rangle$ for fixed u_k has density bounded by CN^{C_3}. □

2.3 Basic tools

In this preliminary section, we collect various identities and estimates that we shall use throughout the following.

Definition 2.3.1 (Minors)**.** *For $J \subset \mathcal{I}$, we define the minor $H^{(J)} := \{H_{st} : s,t \in \mathcal{I} \setminus J\}$, and correspondingly $G^{(J)} := (H^{(J)})^{-1}$. Let $[J] := \{s \in \mathcal{I} : s \in J \text{ or } \bar{s} \in J\}$. We shall also denote $H^{[J]} := \{H_{st} : s,t \in \mathcal{I}\setminus[J]\}$ and $G^{[J]} := (H^{[J]})^{-1}$. We abbreviate $(\{s\}) \equiv (s)$, $(\{s,t\}) \equiv (st)$, $[\{s\}] \equiv [s]$ and $[\{s,t\}] = [st]$.*

Notice that by the definition, we have $H^{(J)}_{st} = 0$ and $G^{(J)}_{st} = 0$ if $s \in J$ or $t \in J$.

Lemma 2.3.2. *(Resolvent identities).*

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

$$\frac{1}{G_{ii}} = -w - w\left(YG^{(i)}Y^\dagger\right)_{ii}, \quad \frac{1}{G_{\mu\mu}} = -w - w\left(Y^\dagger G^{(\mu)}Y\right)_{\mu\mu}. \tag{2.78}$$

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

$$G_{ij} = wG_{ii}G^{(i)}_{jj}\left(YG^{(ij)}Y^\dagger\right)_{ij}, \quad G_{\mu\nu} = wG_{\mu\mu}G^{(\mu)}_{\nu\nu}\left(Y^\dagger G^{(\mu\nu)}Y\right)_{\mu\nu}. \tag{2.79}$$

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

$$G_{i\mu} = G_{ii}G^{(i)}_{\mu\mu}\left(-w^{1/2}Y_{i\mu} + w\big(YG^{(i\mu)}Y\big)_{i\mu}\right), \tag{2.80}$$

$$G_{\mu i} = G_{\mu\mu}G^{(\mu)}_{ii}\left(-w^{1/2}Y^\dagger_{\mu i} + w\left(Y^\dagger G^{(\mu i)}Y^\dagger\right)_{\mu i}\right). \tag{2.81}$$

(iii) For $r \in \mathcal{I}$ and $s, t \in \mathcal{I} \setminus \{r\}$,

$$G^{(r)}_{st} = G_{st} - \frac{G_{sr}G_{rt}}{G_{rr}}, \quad \frac{1}{G_{ss}} = \frac{1}{G^{(r)}_{ss}} - \frac{G_{sr}G_{rs}}{G_{ss}G^{(r)}_{ss}G_{rr}}. \tag{2.82}$$

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

Proof. All these identities can be proved using Schur's complement formula. They have been previously derived and summarized e.g. in [22, 23, 30]. □

Lemma 2.3.3. *(Resolvent identities for $G_{[ij]}$ groups).*

(i) For $i \in \mathcal{I}_1$, we have

$$G^{-1}_{[ii]} = H_{[ii]} - \sum_{k,l \neq i} H_{[ik]}G^{[i]}_{[kl]}H_{[li]}. \tag{2.83}$$

For $i \neq j \in \mathcal{I}_1$, we have

$$G_{[ij]} = -G_{[ii]} \sum_{k \neq i} H_{[ik]} G^{[i]}_{[kj]} = -\sum_{k \neq j} G^{[j]}_{[ik]} H_{[kj]} G_{[jj]} \tag{2.84}$$

$$= -G_{[ii]} H_{[ij]} G^{[i]}_{[jj]} + G_{[ii]} \sum_{k,l \notin \{i,j\}} H_{[ik]} G^{[ij]}_{[kl]} H_{[lj]} G^{[i]}_{[jj]}. \tag{2.85}$$

(ii) For $k \in \mathcal{I}_1$ and $i, j \in \mathcal{I}_1 \setminus \{k\}$,

$$G^{[k]}_{[ij]} = G_{[ij]} - G_{[ik]} G^{-1}_{[kk]} G_{[kj]}, \tag{2.86}$$

and

$$G^{-1}_{[ii]} = \left(G^{[k]}_{[ii]}\right)^{-1} - G^{-1}_{[ii]} G_{[ik]} G^{-1}_{[kk]} G_{[ki]} \left(G^{[k]}_{[ii]}\right)^{-1}. \tag{2.87}$$

(iii) All of the above identities hold for $G^{[J]}$ instead of G for $J \subset \mathcal{I}$.

Proof. These identities can be proved using Schur's complement formula. The details are left to the reader. □

Next we introduce the spectral decomposition of G. Let

$$Y = \sum_{k=1}^{N} \sqrt{\lambda_k} \xi_k \zeta^{\dagger}_{\bar{k}}$$

be the singular decomposition of Y, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$ and $\{\xi_k\}_{k=1}^{N}$ and $\{\zeta_{\bar{k}}\}_{k=1}^{N}$ are orthonormal bases of $\mathbb{C}^{\mathcal{I}_1}$ and $\mathbb{C}^{\mathcal{I}_2}$ respectively. Then by (2.31), we have

$$G(w) = \sum_{k=1}^{N} \frac{1}{\lambda_k - w} \begin{pmatrix} \xi_k \xi_k^{\dagger} & w^{-1/2} \sqrt{\lambda_k} \xi_k \zeta^{\dagger}_{\bar{k}} \\ w^{-1/2} \sqrt{\lambda_k} \zeta_{\bar{k}} \xi_k^{\dagger} & \zeta_{\bar{k}} \zeta^{\dagger}_{\bar{k}} \end{pmatrix}. \tag{2.88}$$

Definition 2.3.4 (Generalized entries)**.** *For $\mathbf{v}, \mathbf{w} \in \mathbb{C}^{\mathcal{I}}$, $s \in \mathcal{I}$ and an $\mathcal{I} \times \mathcal{I}$ matrix A, we shall denote*

$$A_{\mathbf{vw}} := \langle \mathbf{v}, A\mathbf{w} \rangle, \quad A_{\mathbf{v}s} := \langle \mathbf{v}, A\mathbf{e}_s \rangle, \quad A_{s\mathbf{w}} := \langle \mathbf{e}_s, A\mathbf{w} \rangle, \tag{2.89}$$

where $\mathbf{e}_s$ is the standard unit vector.

Given vectors $\mathbf{v} \in \mathbb{C}^{\mathcal{I}_1}$ and $\mathbf{w} \in \mathbb{C}^{\mathcal{I}_2}$, we always identify them with their natural embeddings $\begin{pmatrix} \mathbf{v} \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ \mathbf{w} \end{pmatrix}$ in $\mathbb{C}^{\mathcal{I}}$. The exact meanings will be clear from the context.

Lemma 2.3.5. *Fix $\tau > 0$. The following estimates hold uniformly for any $w \in \mathbf{D}(\zeta, N) \cup \mathbf{D}_L(\zeta)$. We have*

$$\|G\| \leq C\eta^{-1}, \ \|\partial_w G\| \leq C\eta^{-2}. \tag{2.90}$$

Let $\mathbf{v} \in \mathbb{C}^{\mathcal{I}_1}$ and $\mathbf{w} \in \mathbb{C}^{\mathcal{I}_2}$, we have the bounds

$$\sum_{\mu \in \mathcal{I}_2} |G_{\mathbf{w}\mu}|^2 = \sum_{\mu \in \mathcal{I}_2} |G_{\mu\mathbf{w}}|^2 = \frac{\operatorname{Im} G_{\mathbf{ww}}}{\eta}, \tag{2.91}$$

$$\sum_{i \in \mathcal{I}_1} |G_{\mathbf{v}i}|^2 = \sum_{i \in \mathcal{I}_1} |G_{i\mathbf{v}}|^2 = \frac{\operatorname{Im} G_{\mathbf{vv}}}{\eta}, \tag{2.92}$$

$$\sum_{i \in \mathcal{I}_1} |G_{\mathbf{w}i}|^2 = \sum_{i \in \mathcal{I}_1} |G_{i\mathbf{w}}|^2 = |w|^{-1} G_{\mathbf{ww}} + \bar{w} \, |w|^{-1} \frac{\operatorname{Im} G_{\mathbf{ww}}}{\eta}, \tag{2.93}$$

$$\sum_{\mu \in \mathcal{I}_2} |G_{\mathbf{v}\mu}|^2 = \sum_{\mu \in \mathcal{I}_2} |G_{\mu\mathbf{v}}|^2 = |w|^{-1} G_{\mathbf{vv}} + \bar{w} \, |w|^{-1} \frac{\operatorname{Im} G_{\mathbf{vv}}}{\eta}. \tag{2.94}$$

All of the above estimates remain true for $G^{(J)}$ instead of G for $J \subset \mathcal{I}$.

Proof. The estimates in (3.54) follow from (2.88). For any unit vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^{\mathcal{I}_1}$, we have

$$|\langle \mathbf{x}, G\mathbf{y} \rangle| \leq \sum_{k=1}^{N} \frac{|\langle \mathbf{x}, \xi_k \rangle| \left| \langle \xi_k^\dagger, \mathbf{y} \rangle \right|}{|\lambda_k - w|} \leq \frac{1}{\eta} \left[\sum_{k=1}^{N} |\langle \mathbf{x}, \xi_k \rangle|^2 \right]^{1/2} \left[\sum_{k=1}^{N} \left| \langle \xi_k^\dagger, \mathbf{y} \rangle \right|^2 \right]^{1/2} = \frac{1}{\eta}.$$

For any unit vectors $\mathbf{x} \in \mathbb{C}^{\mathcal{I}_1}$ and $\mathbf{y} \in \mathbb{C}^{I_2}$, we have

$$|\langle \mathbf{x}, G\mathbf{y} \rangle| \leq |w|^{-1/2} \sum_{k=1}^{N} \frac{\sqrt{\lambda_k} \, |\langle \mathbf{x}, \xi_k \rangle| \left| \langle \zeta_{\bar{k}}^\dagger, \mathbf{y} \rangle \right|}{|\lambda_k - w|} \leq \sum_{k=1}^{N} \frac{1}{2\eta} \left(|\langle \mathbf{x}, \xi_k \rangle|^2 + \left| \langle \zeta_{\bar{k}}^\dagger, \mathbf{y} \rangle \right|^2 \right) = \frac{1}{\eta},$$

where we have used that for $w = E + i\eta$, $|w|^{-1/2} \sqrt{\lambda_k} / |\lambda_k - w| \leq \eta^{-1}$. For the other two blocks of G, we can prove similar estimates. This gives the first bound in (3.54). It is trivial to generalize the proof to $\partial_w G$, where η^{-2} comes from the $(\lambda_k - w)^{-2}$ factor of $\partial_w G$. For (3.57), we observe that

$$\frac{\operatorname{Im} G_{\mathbf{w}\mathbf{w}}}{\eta} = \frac{1}{\eta} \operatorname{Im} \sum_{k=1}^{N} \frac{\langle \mathbf{w}, \zeta_{\bar{k}} \rangle \langle \zeta_{\bar{k}}^{\dagger}, \mathbf{w} \rangle}{\lambda_k - w} = \sum_{k=1}^{N} \frac{|\langle \mathbf{w}, \zeta_k \rangle|^2}{(\lambda_k - E)^2 + \eta^2},$$

and by (2.31),

$$\sum_{\mu \in \mathcal{I}_2} |G_{\mathbf{w}\mu}|^2 = \sum_{\mu \in \mathcal{I}_2} \langle \mathbf{w}, G_R e_\mu \rangle \langle e_\mu, G_R^{\dagger} \mathbf{w} \rangle = \left\langle \mathbf{w}, G_R G_R^{\dagger} \mathbf{w} \right\rangle = \sum_{k=1}^{N} \frac{|\langle \mathbf{w}, \zeta_k \rangle|^2}{(\lambda_k - E)^2 + \eta^2}. \quad (2.95)$$

Similarly, we can prove the identity for $\sum_{\mu \in \mathcal{I}_2} |G_{\mu\mathbf{w}}|^2$ and (3.55). For (3.56), first we can prove that $\sum_{i \in \mathcal{I}_1} |G_{\mathbf{w}i}|^2 = \sum_{i \in \mathcal{I}_1} |G_{i\mathbf{w}}|^2$ using (2.88). Then we use (2.31) and (2.95) to get that

$$\begin{aligned} \sum_{i \in \mathcal{I}_1} |G_{\mathbf{w}i}|^2 &= |w|^{-1} \left(G_R Y^{\dagger} Y G_R^{\dagger} \right)_{\mathbf{w}\mathbf{w}} = |w|^{-1} \left[G_R \left(Y^{\dagger} Y - \bar{w} \right) G_R^{\dagger} \right]_{\mathbf{w}\mathbf{w}} + \bar{w} \, |w|^{-1} \left(G_R G_R^{\dagger} \right)_{\mathbf{w}\mathbf{w}} \\ &= |w|^{-1} G_{\mathbf{w}\mathbf{w}} + \bar{w} \, |w|^{-1} \left(G_R G_R^{\dagger} \right)_{\mathbf{w}\mathbf{w}} = |w|^{-1} G_{\mathbf{w}\mathbf{w}} + \bar{w} \, |w|^{-1} \frac{\operatorname{Im} G_{\mathbf{w}\mathbf{w}}}{\eta}. \quad (2.96) \end{aligned}$$

Identity (3.58) can be proved in a similar way. □

The following Lemma gives useful large deviation bounds. See Theorem B.1 and Lemmas B.2-B.4 in [21] for the proof. See also Theorem C.1 of [22].

Lemma 2.3.6. *(Large deviation bounds) Let* $(X_i^{(N)})$, $(Y_i^{(N)})$ *be independent families of random variables and* $(a_{ij}^{(N)})$, $(b_i^{(N)})$ *be deterministic complex numbers. Suppose all entries* $X_i^{(N)}$ *and* $Y_i^{(N)}$ *are independent and satisfy (2.3) and (2.4). Then we have the following bounds:*

$$\sum_i b_i X_i \prec \frac{\left(\sum_i |b_i|^2 \right)^{1/2}}{\sqrt{N}}, \quad \sum_{i,j} a_{ij} X_i Y_j \prec \frac{\left(\sum_{i,j} |a_{ij}|^2 \right)^{1/2}}{N}, \quad \sum_{i \neq j} a_{ij} X_i X_j \prec \frac{\left(\sum_{i \neq j} |a_{ij}|^2 \right)^{1/2}}{N}, \quad (2.97)$$

where, for simplicity of notation, we omitted the superscript (N) in the above expressions. If the coefficients $(a_{ij}^{(N)})$ and $(b_i^{(N)})$ depend on some parameter u, then all of the above estimates are uniform in u.

We have stated some basic properties of $\rho_{1,2c}$ and $m_{1,2c}$ in Lemma 2.2.3 and Proposition 2.2.14. Now we collect more estimates for $m_{1,2c}$ that will be used in the proof. The next lemma is proved in Appendix 2.7.2. For $w = E + i\eta$, we define the distance to the spectral edge through

$$\kappa \equiv \kappa(E) := \min_{1 \le k \le 2L, e_k > 0} |E - e_k|. \tag{2.98}$$

Notice in the $|z| < 1$ case, we do not take into consideration the edge at $e_{2L} = 0$.

Lemma 2.3.7. *Fix $\tau > 0$ and suppose $\tau \le ||z|^2 - 1| \le \tau^{-1}$. We denote $w = E + i\eta$.*

Case 1 Fix $\tau' > 0$. Suppose the bulk component $[e_{2k}, e_{2k-1}]$ is regular in the sense of Definition 2.2.4. Then for $w \in \mathbf{D}_k^b(\zeta, \tau', N)$, we have

$$|1 + m_{1c}| \sim \operatorname{Im} m_{1c} \sim 1, \ \ |m_{2c}| \sim \operatorname{Im} m_{2c} \sim 1. \tag{2.99}$$

Case 2 Fix $\tau' > 0$. Then for $w \in \mathbf{D}^o(\zeta, \tau', N)$, we have

$$\operatorname{Im} m_{1,2c} \sim \eta, \ \ |1 + m_{1c}| \sim 1, \ \ |m_{2c}| \sim 1. \tag{2.100}$$

Case 3 Suppose $e_k \neq 0$ is a regular edge. Then for $w \in \mathbf{D}_k^e(\zeta, \tau', N)$, if $\tau' > 0$ is small enough, we have

$$\operatorname{Im} m_{1,2c} \sim \begin{cases} \sqrt{\kappa + \eta} & \text{if } E \in \operatorname{supp} \rho_{1,2c} \\ \eta / \sqrt{\kappa + \eta} & \text{if } E \notin \operatorname{supp} \rho_{1,2c} \end{cases}, \ \ |1 + m_{1c}| \sim 1, \ \ |m_{2c}| \sim 1. \tag{2.101}$$

Case 4 Suppose $|z|^2 \leq 1 - \tau$ *so that* $e_{2L} = 0$. *We take* $\tau' > 0$ *to be small enough. Then for* $w \in \mathbf{D}^e_{2L}(\zeta, \tau', N)$, *if* $\operatorname{Im} w \geq \tau'$, *we have*

$$|1 + m_{1c}| \sim \operatorname{Im} m_{1c} \sim 1, \ |m_{2c}| \sim \operatorname{Im} m_{2c} \sim 1; \tag{2.102}$$

if $|w| \leq 2\tau'$, *we have*

$$m_{1c} = i\frac{\sqrt{t}}{\sqrt{w}} + O(1), \ m_{2c} = \frac{i\sqrt{t}}{\sqrt{w}(t + |z|^2)} + O(1), \tag{2.103}$$

for some constant $t > 0$, *and*

$$\operatorname{Im} m_{1,2c} \sim |w|^{-1/2}. \tag{2.104}$$

Case 5 For $w \in \mathbf{D}_L(\zeta)$, *we have*

$$|m_{1c}| \sim \operatorname{Im} m_{1c} \sim \frac{1}{\eta}, \ |m_{2c}| \sim \operatorname{Im} m_{2c} \sim \frac{1}{\eta}. \tag{2.105}$$

In Cases 1-4, we have

$$\left|w\left(1 + s_i m_{2c}\right)\left(1 + m_{1c}\right) - |z|^2\right| \geq c, \tag{2.106}$$

where $c > 0$ *is some constant that may depend on* τ, τ' *and* ζ. *In Case 5, we have*

$$\left|w\left(1 + s_i m_{2c}\right)\left(1 + m_{1c}\right) - |z|^2\right| \geq \eta, \tag{2.107}$$

Note that the uniform bounds (2.106) and (2.107) guarantee that the matrix entries of $\Pi(w)$ remain bounded. We have the following Lemma, which will be proved in Appendix 2.7.2.

Lemma 2.3.8. *In Cases 1-4 of Lemma 2.3.7, we have*

$$\|\pi_{[i]c}\| \leq C|w|^{-1/2}, \quad \left\|\left(\pi_{[i]c}\right)^{-1}\right\| \leq C|w|^{1/2}, \tag{2.108}$$

and in Case 5 of Lemma 2.3.7, we have

$$\|\pi_{[i]c}\| \le C\eta^{-1}, \quad \left\| \left(\pi_{[i]c}\right)^{-1} \right\| \le C\eta. \tag{2.109}$$

For all the cases in Lemma 2.3.7,

$$\operatorname{Im}\Pi_{\mathbf{v}\mathbf{v}} \le C\operatorname{Im}(m_{1c}+m_{2c}), \tag{2.110}$$

uniformly in w *and any deterministic unit vector* $\mathbf{v} \in \mathbb{C}^{\mathcal{I}}$*.*

The self-consistent equation (2.12) can be written as

$$\Upsilon(w, m_1) = 0, \tag{2.111}$$

where

$$\Upsilon(w,m_1) = m_1 + \frac{1}{N}\sum_{i=1}^{n} l_i s_i (1+m_1)\left[w\left(1 + s_i \frac{1+m_1}{-w(1+m_1)^2 + |z|^2}\right)(1+m_1) - |z|^2 \right]^{-1}. \tag{2.112}$$

The stability of (2.111) roughly says that if $\Upsilon(w,m_1)$ is small and $m_1(w') - m_{1c}(w')$ is small for $w' := w + iN^{-10}$, then $m_1(w) - m_{1c}(w)$ is small. For an arbitrary $w \in \mathbf{D}$, we define the discrete set

$$L(w) := \{w\} \cup \{w' \in \mathbf{D} : \operatorname{Re} w' = \operatorname{Re} w, \operatorname{Im} w' \in [\operatorname{Im} w, 1] \cap (N^{-10}\mathbb{N})\}, \tag{2.113}$$

Thus, if $\operatorname{Im} w \ge 1$ then $L(w) = \{w\}$, and if $\operatorname{Im} w < 1$ then $L(w)$ is a 1-dimensional lattice with spacing N^{-10} plus the point w. Obviously, we have $|L(w)| \le N^{10}$.

Definition 2.3.9 (Stability of (2.111))**.** *We say that (2.111) is stable on* $\mathbf{D}$ *if the following holds. Suppose that* $N^{-2}|m_{1c}| \le \delta(w) \le (\log N)^{-1}|m_{1c}|$ *for* $w \in \mathbf{D}$ *and that* δ *is Lipschitz continuous with Lipschitz constant* $\le N^4$*. Suppose moreover that for each fixed*

E, the function $\eta \mapsto \delta(E+i\eta)$ is non-increasing for $\eta > 0$. Suppose that $u_1 : \mathbf{D} \to \mathbb{C}$ is the Stieltjes transform of a positive integrable function. Let $w \in \mathbf{D}$ and suppose that for all $w' \in L(w)$ we have

$$|\Upsilon(w, u_1)| \leq \delta(w). \tag{2.114}$$

Then

$$|u_1(w) - m_{1c}(w)| \leq \frac{C\delta}{\sqrt{\kappa+\eta+\delta}}, \tag{2.115}$$

for some constant $C > 0$ independent of w and N.

We say that (2.111) is stable on $\mathbf{D}_L$ if for $0 \leq \delta(w) \leq (\log N)^{-1}|m_{1c}|$, (2.114) implies

$$|u_1(w) - m_{1c}(w)| \leq C\delta, \tag{2.116}$$

for some constant $C > 0$ independent of w and N.

This stability condition has previously appeared in [6, 10, 44]. In [44], for example, the stability condition was established under various regularity assumptions. In the following lemma, we establish the stability on each regular domain. The proof is given in Appendix 2.7.3. This lemma leaves the case $|w|^{1/2} + |z|^2 = o(1)$ alone. We will handle this case in a different way in Section 2.4.5.

Lemma 2.3.10. *Fix $\tau > 0$ and let $\tau' > 0$ be sufficiently small depending on τ. Let $\tau \leq ||z|^2 - 1| \leq \tau^{-1}$.*

Case 1 Suppose the bulk component $[e_{2k}, e_{2k-1}]$ is regular in the sense of Definition 2.2.4. Then (2.111) is stable on $\mathbf{D}^b_k(\zeta, \tau', N)$ in the sense of Definition 2.3.9.

Case 2 (2.111) is stable on $\mathbf{D}^o(\zeta, \tau', N)$ in the sense of Definition 2.3.9.

Case 3 *Suppose* $e_k \neq 0$ *is a regular edge in the sense of Definition 2.2.4. Then (2.111) is stable on* $\mathbf{D}^e_k(\zeta, \tau', N)$ *in the sense of Definition 2.3.9.*

Case 4 *Suppose* $|z|^2 \leq 1 - \tau$ *so that* $e_{2L} = 0$*. If* $|w|^{1/2} + |z|^2 \geq \epsilon$ *for some constant* $\epsilon > 0$*, then (2.111) is stable on* $\mathbf{D}^e_{2L}(\zeta, \tau', N)$ *in the sense of Definition 2.3.9.*

Case 5 *(2.111) is stable on* $\mathbf{D}_L(\zeta)$ *in the sense of Definition 2.3.9.*

2.4 Entrywise local law when T is diagonal

In this section we prove the entrywise local law and averaged local law in Theorem 2.2.18 when T is diagonal. The proof is similar to the previous proofs of the entrywise local law in e.g. [6, 7, 10, 44]. We basically follow the idea in [10], and we will provide necessary details for the parts that are different from the previous proofs.

The main novel observation of this section is that the self-consistent equations (2.10) and (2.11) can be "derived" from the random matrix model by an application of Schur's complement formula. It is helpful to give a heuristic argument here. We introduce the conditional expectation

$$\mathbb{E}_{[i]}[\cdot] := \mathbb{E}[\cdot \mid H^{[i]}],$$

i.e. the partial expectation in the randomness of the i and $\bar{i}$-th rows and columns of H. For the diagonal $G_{[ii]}$ group, we ignore formally the random fluctuations in (2.83) to get that

$$G^{-1}_{[ii]} \approx \mathbb{E}_{[i]} H_{[ii]} - \sum_{k,l \neq i} \mathbb{E}_{[i]} \left(H_{[ik]} G^{[i]}_{[kl]} H_{[li]} \right) = \begin{pmatrix} -w & -w^{1/2} z \\ -w^{1/2} \bar{z} & -w \end{pmatrix} - \frac{w}{N} \sum_k \begin{pmatrix} |d_i|^2 G^{[i]}_{\bar{k}\bar{k}} & 0 \\ 0 & |d_k|^2 G^{[i]}_{kk} \end{pmatrix}$$

$$= \begin{pmatrix} -w & -w^{1/2} z \\ -w^{1/2}\bar{z} & -w \end{pmatrix} - w \begin{pmatrix} |d_i|^2 m_2 & 0 \\ 0 & m_1 \end{pmatrix}, \tag{2.117}$$

where we used the definitions of m_1 and m_2 in (2.35). The 11 entry of (2.117) gives the equation

$$G_{ii} \approx \frac{-1 - m_1}{w\left(1 + |d_i|^2 m_2\right)(1 + m_1) - |z|^2}, \tag{2.118}$$

from which we get that

$$G_{ii}\left[-w\left(1 + |d_i|^2 m_2\right) + \frac{|z|^2}{1 + m_1}\right] \approx 1.$$

Summing over i and using that $N^{-1}\sum_i G_{ii} = N^{-1}\sum_\mu G_{\mu\mu} = m_2$, the above equation becomes

$$-w\left(m_2 + m_1 m_2\right) + \frac{|z|^2 m_2}{1 + m_1} \approx 1,$$

which gives (2.10). Multiplying (2.118) with $|d_i|^2$ and summing over i, we get the self-consistent equation (2.11). In this section we give a justification of these approximations.

Before we start the proof, we make the following remark. In this section we mainly focus on the domain $\mathbf{D}$. On the domain $\mathbf{D}_L$, the proofs are much simpler and we only describe them briefly. The parameter z can be either inside or outside of the unit circle. Recall Lemma 2.3.7 and Lemma 2.3.10, the domain $\mathbf{D}$ of w can be divided roughly into four regions: w near a *nonzero* regular edge, $w \to 0$, w in the bulk, or w outside the spectrum. In this section we will only consider the case $|z|^2 \le 1 - \tau$ since it covers all four different behaviors of $m_{1,2c}$. Note that in this case $|m_{1,2c}(w)| \sim |w|^{-1/2}$ for w in any compact set of $\mathbb{C}_+$ by Proposition 2.2.15. Also due to the remark above Lemma 2.3.10, in Sections 2.4.1-2.4.4, we assume $|w|^{1/2} + |z|^2 \ge c$ for some $c > 0$. We will handle the $|w|^{1/2} + |z|^2 = o(1)$ case in Section 2.4.5.

2.4.1 The self-consistent equations

To begin with, we prove the following weak version of the entrywise local law.

Proposition 2.4.1 (Weak entrywise law). *Fix* $|z|^2 \le 1-\tau$ *and a small constant* $c>0$*. Suppose Assumption 2.2.1 holds,* $N=M$ *and* $T \equiv D := diag(d_1,...,d_N)$*. Then for any regular domain* $\mathbf{S} \subset \mathbf{D}$*,*

$$\max_{i,j\in\mathcal{I}_1} \left\| (G(w)-\Pi(w))_{[ij]} \right\| \prec \frac{1}{|w|^{1/2}} \left(\frac{|w|^{1/2}}{N\eta} \right)^{1/4} \tag{2.119}$$

for all $w\in\mathbf{S}$ *such that* $|w|^{1/2}+|z|^2 \ge c$*. For* $w \in \mathbf{D}_L$*, we have*

$$\max_{i,j\in\mathcal{I}_1} \left\| (G(w)-\Pi(w))_{[ij]} \right\| \prec \frac{1}{\eta}\sqrt{\frac{1}{N}}. \tag{2.120}$$

For the purpose of proof, we define the following random control parameters.

Definition 2.4.2 (Control parameters). *Suppose* $N=M$ *and* $T \equiv D := diag(d_1,...,d_N)$*. We define*

$$\Lambda := \max_{i,j\in\mathcal{I}_1} \left\| (G-\Pi)_{[ij]} \right\|, \quad \Lambda_o := \max_{i\ne j\in\mathcal{I}_1} \left\| (G-\Pi)_{[ij]} \right\|. \tag{2.121}$$

For $J\subseteq\mathcal{I}$*, define the averaged variables* $m_{1,2}^{(J)}$ $(m_{1,2}^{[J]})$ *by replacing* G *in (2.35) with* $G^{(J)}$ $(G^{[J]})$*, i.e.*

$$m_1^{(J)} := \frac{1}{N}\sum_{i\notin J} |d_i|^2 G_{ii}^{(J)}, \quad m_2^{(J)} := \frac{1}{N}\sum_{\mu\notin J} G_{\mu\mu}^{(J)}. \tag{2.122}$$

The averaged error and the random control parameter are defined as

$$\theta := |m_1-m_{1c}|+|m_2-m_{2c}| \quad \text{and} \quad \Psi_\theta := \sqrt{\frac{\operatorname{Im}(m_{1c}+m_{2c})+\theta}{N\eta}} + \frac{1}{N\eta}, \tag{2.123}$$

respectively.

Remark: By (2.5), we immediately get that

$$\tau \operatorname{Im} m_1^{(J)} \le \operatorname{Im} m_2^{(J)} \le \tau^{-1} \operatorname{Im} m_2^{(J)}, \tag{2.124}$$

and $\theta = O(\Lambda)$, since $|m_1 - m_{1c}| \le \tau^{-1}\Lambda$ and $|m_2 - m_{2c}| \le \Lambda$.

We introduce the Z variables:

$$Z_{[i]}^{[J]} := (1 - \mathbb{E}_{[i]}) \left(G_{[ii]}^{[J]} \right)^{-1}.$$

By the identity (2.83) we have

$$G_{[ii]}^{-1} = \mathbb{E}_{[i]} G_{[ii]}^{-1} + Z_{[i]} = \begin{pmatrix} -w - w\,|d_i|^2\, m_2^{[i]} & -w^{1/2} z \\ -w^{1/2} \bar{z} & -w - w m_1^{[i]} \end{pmatrix} + Z_{[i]}, \tag{2.125}$$

where

$$Z_{[i]} = w \begin{pmatrix} |d_i|^2 m_2^{[i]} - |d_i|^2 \left(X G^{[i]} X^\dagger \right)_{ii} & w^{-1/2} d_i X_{i\bar{i}} - \left(D X G^{[i]} D X \right)_{i\bar{i}} \\ w^{-1/2} \bar{d}_i X^\dagger_{\bar{i}i} - \left(X^\dagger D^\dagger G^{[i]} X^\dagger D^\dagger \right)_{\bar{i}i} & m_1^{[i]} - \left(X^\dagger D^\dagger G^{[i]} D X \right)_{\bar{i}\bar{i}} \end{pmatrix}. \tag{2.126}$$

Lemma 2.4.3. *For $J \subseteq \mathcal{I}_1$, the following crude bound on the difference between m_a and $m_a^{[J]}$ $(a = 1, 2)$ holds:*

$$\left| m_a - m_a^{[J]} \right| \le \frac{C\,|J|}{N\eta}, \quad a = 1, 2, \tag{2.127}$$

where $C = C(\tau)$ is a constant depending only on τ.

Proof. For $i \in \mathcal{I}_1$, we have

$$|m_1 - m_1^{(i)}| = \frac{1}{N} \left| \sum_{k \in \mathcal{I}_1} |d_k|^2 \frac{G_{ki} G_{ik}}{G_{ii}} \right| \le \frac{\tau^{-1}}{N |G_{ii}|} \sum_{k \in \mathcal{I}_1} |G_{ik}|^2 = \frac{\tau^{-1}}{N\eta} \frac{\operatorname{Im} G_{ii}}{|G_{ii}|} \le \frac{\tau^{-1}}{N\eta} \tag{2.128}$$

where in the first step we used (3.50), and in the second and third steps the equality (3.55). Similarly, using (3.50) and (3.56) we get

$$|m_1^{(i)} - m_1^{(i\bar{i})}| = \frac{1}{N} \left| \sum_{k \in \mathcal{I}_1} |d_k|^2 \frac{G^{(i)}_{k\bar{i}} G^{(i)}_{\bar{i}k}}{G^{(i)}_{\bar{i}\bar{i}}} \right| \le \frac{\tau^{-1}}{N |G^{(i)}_{\bar{i}\bar{i}}|} \left(\frac{G^{(i)}_{\bar{i}\bar{i}}}{|w|} + \frac{\bar{w}}{|w|} \frac{\operatorname{Im} G^{(i)}_{\bar{i}\bar{i}}}{\eta} \right) \le \frac{2\tau^{-1}}{N\eta}.$$

By induction on the indices in $[J]$, we can prove (3.52) for m_1. The proof for m_2 is similar. □

Lemma 2.4.4. *Suppose $|z|^2 \le 1-\tau$. For $i \in \mathcal{I}_1$, we have*

$$\left|\left(Z_{[i]}\right)_{11}\right| \prec |w|\sqrt{\frac{\operatorname{Im} m_2^{[i]}}{N\eta}}, \quad \left|\left(Z_{[i]}\right)_{22}\right| \prec |w|\sqrt{\frac{\operatorname{Im} m_1^{[i]}}{N\eta}}, \tag{2.129}$$

$$\left|\left(Z_{[i]}\right)_{st}\right| \prec |w|\left(\frac{|w|^{-1/2}}{\sqrt{N}}+\sqrt{\frac{|m_1^{[i]}|}{N\,|w|}}+\sqrt{\frac{\operatorname{Im} m_1^{[i]}}{N\eta}}\right) \quad \text{for } s \neq t \in \{1,2\}, \tag{2.130}$$

uniformly in $w \in \mathbf{D}\cup\mathbf{D}_L$. In particular, these imply that

$$Z_{[i]} \prec |w|\Psi_\theta, \tag{2.131}$$

uniformly in $w \in \mathbf{D}$, and

$$Z_{[i]} \prec |w|(N\eta)^{-1/2}, \tag{2.132}$$

uniformly in $w \in \mathbf{D}_L$.

Proof. Applying the large deviation Lemma 2.3.6 to $Z_{[i]}$ in (2.126), we get that

$$\begin{aligned}\left|\frac{\left(Z_{[i]}\right)_{11}}{w}\right| \prec & \frac{1}{N}\left[\left(\sum_{\mu}\left|G^{[i]}_{\mu\mu}\right|^2\right)^{1/2}+\left(\sum_{\mu\neq\nu}\left|G^{[i]}_{\mu\nu}\right|^2\right)^{1/2}\right] \le \frac{C}{N}\left(\sum_{\mu,\nu}\left|G^{[i]}_{\mu\nu}\right|^2\right)^{1/2}\\ = & \frac{C}{N}\left(\sum_{\mu}\frac{\operatorname{Im} G^{[i]}_{\mu\mu}}{\eta}\right)^{1/2} = C\sqrt{\frac{\operatorname{Im} m_2^{[i]}}{N\eta}}.\end{aligned}$$

where in the third step we used the equality (3.57). Similarly we can prove the bound for $\left(Z_{[i]}\right)_{22}$ using Lemma 2.3.6 and (3.55). Now we consider $\left(Z_{[i]}\right)_{12}$. First, we have $X_{i\bar{i}} \prec N^{-1/2}$ by (2.4). For the other part, we use Lemma 2.3.6 and (3.58) to get that

$$\left|\left(DXG^{[i]}DX\right)_{i\bar{i}}\right| \prec \frac{1}{N}\left(\sum_{j,\mu}|d_j|^2\left|G^{[i]}_{\mu j}\right|^2\right)^{1/2} = \frac{1}{N}\left[\sum_j |d_j|^2\left(|w|^{-1}G^{[i]}_{jj}+\frac{\bar{w}}{|w|}\frac{\operatorname{Im} G^{[i]}_{jj}}{\eta}\right)\right]^{1/2}$$

$$\leq \left[\frac{|m_1^{[i]}|}{N\,|w|} + \frac{\operatorname{Im} m_1^{[i]}}{N\eta}\right]^{1/2} \leq C\left(\sqrt{\frac{|m_1^{[i]}|}{N\,|w|}} + \sqrt{\frac{\operatorname{Im} m_1^{[i]}}{N\eta}}\right). \tag{2.133}$$

Similarly we can prove the estimate for $\left(Z_{[i]}\right)_{21}$.

Now we prove (2.131). By the definitions (2.123) and using (3.52), we get that

$$\left|\left(Z_{[i]}\right)_{11}\right| \prec |w|\sqrt{\frac{\operatorname{Im} m_2^{[i]}}{N\eta}} = |w|\sqrt{\frac{\operatorname{Im} m_{2c} + \operatorname{Im}\left(m_2^{[i]} - m_2\right) + \operatorname{Im}\left(m_2 - m_{2c}\right)}{N\eta}} \leq C|w|\Psi_\theta. \tag{2.134}$$

We can estimate $\left(Z_{[i]}\right)_{22}$ and the third term in (2.130) in a similar way. For the Cases 1-4 in Lemma 2.3.7, we have $|m_{1c}| \sim 1$ for $|w| \sim 1$, $\operatorname{Im} m_{1c} \sim |w|^{-1/2} \sim |m_{1c}|$ for $|w| \to 0$, and $\eta \leq C\operatorname{Im} m_{1c}$. Thus

$$\sqrt{\frac{|m_{1c}|}{N\,|w|}} \leq \frac{C}{\sqrt{N}} \leq C\Psi_\theta \text{ for } |w| \sim 1, \quad \text{and} \quad \sqrt{\frac{|m_{1c}|}{N\,|w|}} \leq C\sqrt{\frac{\operatorname{Im} m_{1c}}{N\eta}} \leq C\Psi_\theta \text{ for } |w| \to 0.$$

Then for the second term in (2.130), we have that

$$\sqrt{\frac{|m_1^{[i]}|}{N\,|w|}} \leq C\left(\frac{1}{N\eta} + \sqrt{\frac{\theta}{N\eta}} + \sqrt{\frac{|m_{1c}|}{N\,|w|}}\right) \leq C\Psi_\theta.$$

This concludes (2.131). Finally, the estimate (2.132) follows directly from (2.129), (2.130) and (3.54). □

Lemma 2.4.5. *Suppose* $|z|^2 \leq 1-\tau$. *Define the* w*-dependent event* $\Xi(w) := \{\theta \leq |w|^{-1/2}(\log N)^{-1}\}$. *Then we have that for* $w \in \mathbf{D}$,

$$\mathbf{1}(\Xi)m_2 = \mathbf{1}(\Xi)\left[\frac{1+m_1}{-w\left(1+m_1\right)^2 + |z|^2} + O_\prec(\Psi_\theta)\right], \quad \mathbf{1}(\Xi)\Upsilon(w,m_1) \prec \mathbf{1}(\Xi)\Psi_\theta, \tag{2.135}$$

where Υ *is defined in (2.112). For* $w \in \mathbf{D}_L$, *we have*

$$m_2 = \frac{1+m_1}{-w\left(1+m_1\right)^2 + |z|^2} + O_\prec\left(\eta^{-1}(N\eta)^{-1/2}\right), \quad \Upsilon(w,m_1) \prec \eta^{-1}\left(N\eta\right)^{-1/2}. \tag{2.136}$$

Proof. First, suppose that $w \in \mathbf{D}$. Using (2.125), we get

$$G_{[ii]}^{-1} = \pi_{[i]}^{-1} + \epsilon_{[i]}, \tag{2.137}$$

where $\pi_{[i]}$ is defined in (2.37) and

$$\epsilon_{[i]} = w \begin{pmatrix} |d_i|^2 \left(m_2 - m_2^{[i]}\right) & 0 \\ 0 & m_1 - m_1^{[i]} \end{pmatrix} + Z_{[i]}.$$

By (3.52) and (2.131), we have that $\epsilon_{[i]} \prec |w|\Psi_\theta$. Let $B_i = \pi_{[i]}^{-1} - \pi_{[i]c}^{-1}$, where $\pi_{[i]c}$ is defined in (2.33). By (2.108) and the definition of Ξ, we have $\mathbf{1}(\Xi)\|B_i\pi_{[i]c}\| \le C(\log N)^{-1}$. Thus we have the expansion

$$\mathbf{1}(\Xi)\pi_{[i]} = \mathbf{1}(\Xi)(\pi_{[i]c}^{-1} + B_i)^{-1} = \mathbf{1}(\Xi)\pi_{[i]c}\left(1 - B_i\pi_{[i]c} + (B_i\pi_{[i]c})^2 + \ldots\right) = \mathbf{1}(\Xi)(\pi_{[i]c} + \epsilon_a), \tag{2.138}$$

where ϵ_a can be estimated as $\mathbf{1}(\Xi)\|\epsilon_a\| \le \mathbf{1}(\Xi)C|w|^{-1/2}(\log N)^{-1}$. This shows that $\mathbf{1}(\Xi)\|\pi_{[i]}\| = \mathbf{1}(\Xi)O(|w|^{-1/2})$, and so $\mathbf{1}(\Xi)\left\|\epsilon_{[i]}\pi_{[i]}\right\| \prec \mathbf{1}(\Xi)|w|^{1/2}\Psi_\theta \le \mathbf{1}(\Xi)\, CN^{-\zeta/2}$ by the definition of $\mathbf{D}$ in (3.31). Again we do the expansion for (2.137):

$$\mathbf{1}(\Xi)G_{[ii]} = \mathbf{1}(\Xi)\left(\pi_{[i]}^{-1} + \epsilon_{[i]}\right)^{-1} = \mathbf{1}(\Xi)\pi_{[i]}\left(1 + \sum_{l=1}^{\infty}\left(-\epsilon_{[i]}\pi_{[i]}\right)^l\right) = \mathbf{1}(\Xi)\left(\pi_{[i]} + \epsilon_b\right), \tag{2.139}$$

where $\mathbf{1}(\Xi)\|\epsilon_b\| \prec \mathbf{1}(\Xi)\Psi_\theta$. Now the 11 entry of (2.139) gives that

$$\mathbf{1}(\Xi)G_{ii} = \mathbf{1}(\Xi)\frac{-1 - m_1}{w\left(1 + |d_i|^2 m_2\right)\left(1 + m_1\right) - |z|^2} + \mathbf{1}(\Xi)O_\prec\left(\Psi_\theta\right), \tag{2.140}$$

from which we get that

$$\mathbf{1}(\Xi)G_{ii}\left[-w\left(1 + |d_i|^2 m_2\right) + \frac{|z|^2}{1 + m_1}\right] = \mathbf{1}(\Xi)\left[1 + O_\prec\left(|w|^{1/2}\Psi_\theta\right)\right]. \tag{2.141}$$

Here we used that

$$\mathbf{1}(\Xi)\left[-w\left(1 + |d_i|^2 m_2\right) + \frac{|z|^2}{1 + m_1}\right] = O(|w|^{1/2}),$$

which follows from Lemma 2.3.7 and the definition of Ξ. Summing (2.141) over i, we get

$$\mathbf{1}(\Xi)\left[-w\left(m_2+m_1m_2\right)+\frac{|z|^2m_2}{1+m_1}\right]=\mathbf{1}(\Xi)\left[1+O_{\prec}\left(|w|^{1/2}\Psi_\theta\right)\right],$$

which gives

$$\mathbf{1}(\Xi)m_2=\mathbf{1}(\Xi)\frac{1+m_1}{-w\left(1+m_1\right)^2+|z|^2}+\mathbf{1}(\Xi)O_{\prec}\left(\Psi_\theta\right). \tag{2.142}$$

Now plugging (2.142) into (2.140), multiplying with $|d_i|^2$ and summing over i, we obtain that

$$\mathbf{1}(\Xi)m_1=\mathbf{1}(\Xi)\left[\frac{1}{N}\sum_{i=1}^{n}l_is_i\frac{-1-m_1}{w\left(1+s_i\frac{1+m_1}{-w(1+m_1)^2+|z|^2}\right)(1+m_1)-|z|^2}+O_{\prec}\left(\Psi_\theta\right)\right], \tag{2.143}$$

where we used (2.106) and $\mathbf{1}(\Xi)(1+m_1)=\mathbf{1}(\Xi)O(|w|^{-1/2})$. This concludes the proof.

Similarly, when $w\in\mathbf{D}_L$, it is easy to prove (2.136) using the estimates (2.132) and (3.54). Note that $|m_{1,2}|=O(\eta^{-1})$ by (3.54), which implies immediately the bounds $\|\pi_{[i]}\|=O(\eta^{-1})$ and $\|\left(\pi_{[i]}\right)^{-1}\|=O(\eta)$. Hence without introducing the event Ξ, we can obtain directly

$$G_{[ii]}=\pi_{[i]}+O_{\prec}(\eta^{-1}(N\eta)^{-1/2}). \tag{2.144}$$

The rest of the proof is essentially the same. $\square$

Notice that applying Lemma 2.3.10 to (2.136), we obtain that $|m_{1,2}-m_{1,2c}|\prec\eta^{-1}(N\eta)^{-1/2}$. Plugging it into (2.144), we immediately get (2.120) for $w\in\mathbf{D}_L$. This proves the entrywise law on $\mathbf{D}_L$, since $\eta^{-1}N^{-1/2}\le C\Psi$ by the definition (3.37) and the estimate (2.105).

2.4.2 The large η case

It remains to prove Proposition 2.4.1 on domain $\mathbf{D}$. We would like to fix E and then apply a continuity argument in η by first showing that the rough bound $\Lambda \leq |w|^{-1/2}(\log N)^{-1}$ in Lemma 2.4.5 holds for large η. To start the argument, we first need to establish the estimates on G when $\eta \sim 1$. The next lemma is a trivial consequence of (3.54).

Lemma 2.4.6. *For any $w \in \mathbf{D}$ and $\eta \geq c$ for fixed $c > 0$, we have the bound*

$$\max_{s,t} |G_{st}(w)| \leq C \tag{2.145}$$

for some $C > 0$. This estimate also holds if we replace G with $G^{(J)}$ for $J \subset \mathcal{I}$.

Lemma 2.4.7. *Fix $c > 0$ and $|z|^2 \leq 1 - \tau$. We have the following estimate*

$$\max_{w \in \mathbf{D}, \eta \geq c} \Lambda(w) \prec N^{-1/2}. \tag{2.146}$$

Proof. By the previous lemma, we have $|m_{1,2}^{[i]}| = O(1)$. So by Lemma 2.4.4, $\|Z_{[i]}\| \prec N^{-1/2}$ uniformly in $\eta \geq c$. Then as in (2.137), we have

$$G_{[ii]} = \left(\pi_{[i]}^{-1} + \epsilon_{[i]}\right)^{-1}, \tag{2.147}$$

where $\|\pi_{[i]}^{-1}\| = O(1)$ and $\|\epsilon_{[i]}\| \prec N^{-1/2}$. Notice since $G_{[ii]} = O(1)$, we have the estimate

$$\pi_i = \left(G_{[ii]}^{-1} - \epsilon_{[i]}\right)^{-1} = G_{[ii]}\left(1 - \epsilon_{[i]} G_{[ii]}\right)^{-1} = O_{\prec}(1).$$

Then we can expand (2.147) to get that

$$G_{[ii]} = \pi_i + O_{\prec}\left(N^{-1/2}\right). \tag{2.148}$$

The 11 and 22 entries of (2.148) lead to the equations

$$m_1 = \frac{1}{N} \sum_{i=1}^{N} |d_i|^2 \left[-w\left(1 + |d_i|^2 m_2\right) + \frac{|z|^2}{1 + m_1}\right]^{-1} + O_{\prec}\left(N^{-1/2}\right), \tag{2.149}$$

$$m_2 = \frac{1}{N}\sum_{i=1}^{N}\left[-w\left(1+m_1\right)+\frac{|z|^2}{1+|d_i|^2 m_2}\right]^{-1} + O_{\prec}\left(N^{-1/2}\right). \tag{2.150}$$

We claim that $\operatorname{Im} m_{1,2} \geq C(\log N)^{-1}$ with high probability for some $C > 0$.

Using the spectral decomposition (2.88), we note that for $l > 1$,

$$\begin{aligned}
&\frac{1}{N}\sum_{|\lambda_k - E|\geq l\eta}\frac{|E-\lambda_k|}{(\lambda_k-E)^2+\eta^2} \leq \frac{1}{l\eta},\\
&\frac{1}{N}\sum_{|\lambda_k - E|\leq l\eta}\frac{|E-\lambda_k|}{(\lambda_k-E)^2+\eta^2} \leq \frac{1}{N}\sum_{|\lambda_k - E|\leq l\eta}\frac{l\eta}{(\lambda_k-E)^2+\eta^2} \leq l\operatorname{Im} m_2.
\end{aligned}$$

Summing up these two inequalities and optimizing l, we get

$$|\operatorname{Re} m_2| \leq 2\sqrt{\frac{\operatorname{Im} m_2}{\eta}}. \tag{2.151}$$

Assume that $\operatorname{Im} m_2 \leq C(\log N)^{-1}$, then by (2.124) we also have $\operatorname{Im} m_1 \leq C\tau^{-1}(\log N)^{-1}$. From (2.151), we get $|m_2| \leq C(\log N)^{-1/2}$. Together with the estimate $m_1 = O(1)$, we get

$$\left|-w\left(1+m_1\right)+\frac{|z|^2}{1+|d_i|^2 m_2}\right| \leq C \text{ with high probability.} \tag{2.152}$$

On the other hand

$$\operatorname{Im}\left[-w\left(1+m_1\right)+\frac{|z|^2}{1+|d_i|^2 m_2}\right] \leq -\operatorname{Im} w = -\eta, \tag{2.153}$$

where we used $\operatorname{Im}[|z|^2/(1+|d_i|^2 m_2)] < 0$ and

$$\operatorname{Im}(w m_1) = \operatorname{Im}\left[\frac{1}{N}\sum_{k=1}^{N}|d_i|^2|\xi_k(i)|^2\left(-1+\frac{\lambda_k}{\lambda_k - w}\right)\right] \geq 0.$$

With (2.152) and (2.153), we get from (2.150) that $\operatorname{Im} m_2 \geq c'$ with high probability for some $c' > 0$. This contradicts $\operatorname{Im} m_2 \leq C(\log N)^{-1}$. Thus we must have $\operatorname{Im} m_2 \geq C(\log N)^{-1}$ with high probability, which also implies $\operatorname{Im} m_1 \geq C(\log N)^{-1}$ by (2.124).

Now we can proceed as in the proof of Lemma 2.4.5 and get that

$$m_2 = \frac{1+m_1}{-w\left(1+m_1\right)^2+|z|^2} + O_{\prec}\left(N^{-1/2}\right), \quad \Upsilon(w,m_1) \prec N^{1/2}. \tag{2.154}$$

We omit the details. Applying Lemma 2.3.10 to (2.154), we conclude $|m_{1,2}-m_{1,2c}| \prec N^{-1/2}$ uniformly in $\eta \geq c$. By (2.148), we get $\|(G-\Pi)_{[ii]}\| \prec N^{-1/2}$ uniformly in $\eta \geq c$ and $i \in \mathcal{I}_1$. Finally using (2.85), Lemma 3.2.14 and Lemma 2.3.6, we can prove the off-diagonal estimate; see (2.167) below. $\square$

2.4.3 Proof of the weak entrywise local law

In this subsection, we finish the proof of Proposition 2.4.1 on domain $\mathbf{D}$. We shall fix the real part E of $w = E + i\eta$ and decrease the imaginary part η. Recall that Lemma 2.4.5 is based on the condition $\theta \leq |w|^{-1/2}(\log N)^{-1}$. So far this is established only for large η in (2.146). We want to show that this condition also holds for small η by using a continuity argument.

It is convenient to introduce the random function

$$v(w) = \max_{w' \in L(w)} \theta(w')|w'|^{1/2}\left(\frac{N\operatorname{Im} w'}{|w'|^{1/2}}\right)^{1/4},$$

where $L(w)$ is defined in (2.113). Fix a regular domain $\mathbf{S}$, $\epsilon < \zeta/4$ and a large constant $D > 0$. Our goal is to prove that with high probability there is a gap in the range of v, i.e.

$$\mathbb{P}\left(v(w) \leq N^{\epsilon}, v(w) > N^{3\epsilon/4}\right) \leq N^{-D+21} \tag{2.155}$$

for all $w \in \mathbf{S}$ and large enough $N \geq N(\epsilon, D)$.

Suppose $v(w) \leq N^{\epsilon}$, then it is easy to verify that

$$\theta(w') \leq C|w'|^{-1/2}(\log N)^{-1} \tag{2.156}$$

for all $w' \in L(w)$. Hence $\{v(w) \le N^\epsilon\} \subset \Xi(w')$ for all $w' \in \mathbf{S} \cap L(w)$. Then by (2.135), for all $w' \in \mathbf{S} \cap L(w)$, there exists an $N_0 \equiv N_0(\epsilon, D)$ such that

$$P\left(v(w) \le N^\epsilon, \Upsilon(w') > \frac{N^\epsilon}{|w'|^{1/2}} \sqrt{\frac{|w'|^{1/2}}{N \operatorname{Im} w'}} \right) \le N^{-D}, \tag{2.157}$$

for all $N > N_0$. Taking the union bound we get

$$P\left(v(w) \le N^\epsilon, \max_{w' \in L(w)} \Upsilon(w') \sqrt{\frac{N \operatorname{Im} w'}{|w'|^{-1/2}}} > N^\epsilon \right) \le N^{-D+10}. \tag{2.158}$$

Now consider the event

$$\Xi_1 := \left\{ v(w) \le N^\epsilon, \max_{w' \in L(w)} \Upsilon(w') \sqrt{\frac{N \operatorname{Im} w'}{|w'|^{-1/2}}} \le N^\epsilon \right\}. \tag{2.159}$$

We have $1(\Xi_1)\Upsilon(w') \le \delta(w')$ for all $w' \in L(w)$ with $\delta(w') := \frac{N^\epsilon}{|w'|^{1/2}} \sqrt{\frac{|w'|^{1/2}}{N \operatorname{Im} w'}}$. We now apply Lemma 2.3.10. If $\kappa \ll 1$ (recall (2.98)), then $|w| \sim 1$ and we have

$$1(\Xi_1)|m_1(w') - m_{1c}(w')| \le C\sqrt{\delta(w')} \le C N^{\epsilon/2} \left(\frac{1}{N \operatorname{Im} w'} \right)^{1/4}$$

for all $w' \in L(w)$; if $\kappa \ge c > 0$ for some constant $c > 0$, then

$$1(\Xi_1)|m_1(w') - m_{1c}(w')| \le C\delta(w') \le C \frac{N^\epsilon}{|w'|^{1/2}} \left(\frac{|w'|^{1/2}}{N \operatorname{Im} w'} \right)^{1/2}$$

for all $w' \in L(w)$. Combining these two cases we get

$$1(\Xi_1)|m_1(w') - m_{1c}(w')| \le C \frac{N^{\epsilon/2}}{|w'|^{1/2}} \left(\frac{|w'|^{1/2}}{N \operatorname{Im} w'} \right)^{1/4} \tag{2.160}$$

for all $w' \in L(w)$. By (2.135), we have

$$1(\Xi_1)|m_2(w') - m_{2c}(w')| \prec 1(\Xi_1)|m_1(w') - m_{1c}(w')| + 1(\Xi_1)\Psi_\theta \prec \frac{N^{\epsilon/2}}{|w'|^{1/2}} \left(\frac{|w'|^{1/2}}{N \operatorname{Im} w'} \right)^{1/4},$$

for all $w' \in \mathbf{S} \cap L(w)$. Together with (2.160), this shows that there exists an $N_1 \equiv N_1(\epsilon, D)$ such that

$$\mathbb{P}\left(v(w) \le N^{\epsilon}, \max_{w' \in L(w)} \Upsilon(w') \sqrt{\frac{N \mathrm{Im}\, w'}{|w'|^{-1/2}}} \le N^{\epsilon}, \max_{w' \in L(w)} \theta(w') |w'|^{1/2} \left(\frac{N \mathrm{Im}\, w'}{|w'|^{1/2}} \right)^{1/4} > N^{3\epsilon/4} \right) \le N^{-D} \tag{2.161}$$

for $N \ge \max\{N_0, N_1\}$. Adding (2.158) and (2.161), we get

$$\mathbb{P}\left(v(w) \le N^{\epsilon}, \max_{w' \in L(w)} \theta(w') |w'|^{1/2} \left(\frac{N \mathrm{Im}\, w'}{|w'|^{1/2}} \right)^{1/4} > N^{3\epsilon/4} \right) \le N^{-D+11}.$$

Taking the union bound over $L(w)$ we get (2.155) for all $N \ge \max\{N_0, N_1\}$.

Now we conclude the proof of Proposition 2.4.1 by combining (2.155) with the large η estimate (2.146). We choose a lattice $\Delta \subset \mathbf{S}$ such that $|\Delta| \le N^{20}$ and for any $w \in \mathbf{S}$ there is a $w' \in \Delta$ with $|w' - w| \le N^{-9}$. Taking the union bound we get

$$\mathbb{P}\left(\exists w \in \Delta : v(w) \in (N^{3\epsilon/4}, N^{\epsilon}] \right) \le N^{-D+41}. \tag{2.162}$$

Since v has Lipshcitz constant bounded by, say, N^6, then we have

$$\mathbb{P}\left(\exists w \in \mathbf{S} : v(w) \in (2N^{3\epsilon/4}, N^{\epsilon}/2] \right) \le N^{-D+41}. \tag{2.163}$$

Combining with (2.146), we see that there exists $N_2 \equiv N_2(\epsilon, D)$ such that for all $N > N_2$,

$$\mathbb{P}\left(\forall w \in \mathbf{S} : v(w) \le 2N^{3\epsilon/4} \right) \ge 1 - 2N^{-D+41}.$$

Since ϵ and D are arbitrary, the above inequality shows that $v(w) \prec 1$ uniformly in $w \in \mathbf{S}$, or

$$\theta(w) \prec \frac{1}{|w|^{1/2}} \left(\frac{|w|^{1/2}}{N\eta} \right)^{1/4}. \tag{2.164}$$

In particular this shows that for all $w \in \mathbf{S}$, the event Ξ holds with high-probability.

Now using (2.139) and (2.164), we get

$$\left\|G_{[ii]} - \pi_{[i]c}\right\| \le \left\|G_{[ii]} - \pi_{[i]}\right\| + \left\|\pi_{[i]} - \pi_{[i]c}\right\| \prec \Psi_\theta + \theta \prec \frac{1}{|w|^{1/2}} \left(\frac{|w|^{1/2}}{N\eta}\right)^{1/4}. \tag{2.165}$$

To conclude Proposition 2.4.1, it remains to prove the estimate for the off-diagonal $G_{[ij]}$ groups. Using (3.52), it is not hard to get that

$$\left\|G^{[J]}_{[ii]} - \pi_{[i]c}\right\| \prec \frac{1}{|w|^{1/2}} \left(\frac{|w|^{1/2}}{N\eta}\right)^{1/4} \tag{2.166}$$

for any $|J| \le l$ with $l \in \mathbb{N}$ fixed. Thus we have $\left\|G^{[J]}_{[ii]}\right\| = O\left(|w|^{-1/2}\right)$ and $\left\|\left(G^{[J]}_{[ii]}\right)^{-1}\right\| = O\left(|w|^{1/2}\right)$ with high probability. Let $i \ne j \in \mathcal{I}_1$, using (2.85) and the above diagonal estimates, we get that

$$\left\|G_{[ij]}\right\| \prec |w|^{-1}\frac{|w|^{1/2}}{\sqrt{N}} + |w|^{-1}\left\|\sum_{k,l\notin\{i,j\}} H_{[ik]} G^{[ij]}_{[kl]} H_{[lj]}\right\| \prec \Psi_\theta \prec \frac{1}{|w|^{1/2}} \left(\frac{|w|^{1/2}}{N\eta}\right)^{1/4}, \tag{2.167}$$

where we used Lemma 3.2.14 and Lemma 2.3.6 to obtain that

$$|w|^{-1}\left\|\sum_{k,l\notin\{i,j\}} H_{[ik]} G^{[ij]}_{[kl]} H_{[lj]}\right\| = \left\|\begin{pmatrix} \sum_{k,l\notin\{i,j\}} X_{i\bar{k}} G^{[ij]}_{\bar{k}\bar{l}} X^\dagger_{\bar{l}j} & \sum_{k,l\notin\{i,j\}} X_{i\bar{k}} G^{[ij]}_{\bar{k}l} X_{l\bar{j}} \\ \sum_{k,l\notin\{i,j\}} X^\dagger_{\bar{i}k} G^{[ij]}_{k\bar{l}} X^\dagger_{\bar{l}j} & \sum_{k,l\notin\{i,j\}} X^\dagger_{\bar{i}k} G^{[ij]}_{kl} X_{l\bar{j}} \end{pmatrix}\right\| \prec \Psi_\theta. \tag{2.168}$$

Its proof is very similar to the proof of Lemma 2.4.4, so we omit the details.

2.4.4 Proof of the strong enterywise local law

In this section, we finish the proof of the (strong) entrywise local law and averaged local law in Theorem 2.2.18 on domain $\mathbf{D}$ and under the condition $|w|^{1/2} + |z|^2 \ge c$. In Lemma 2.4.5, we have proved an error estimate of the self-consistent equations of $m_{1,2}$ linearly in Ψ_θ. The core part of the proof is to improve this estimate to quadratic in Ψ_θ. For the sequence of random variables $Z_{[i]}$, we define the averaged quantities

$$[Z] = \frac{1}{N}\sum_{i=1}^{N} \pi_{[i]} Z_{[i]} \pi_{[i]}, \quad \langle Z \rangle = \frac{1}{N}\sum_{i=1}^{N} |d_i|^2 \pi_{[i]} Z_{[i]} \pi_{[i]}.$$

The following Lemma gives an improvement of Lemma 2.4.5.

Lemma 2.4.8. *Fix $|z|^2 \leq 1-\tau$. Then for $w \in \mathbf{D}$,*

$$m_2 = \frac{1+m_1}{-w\left(1+m_1\right)^2 + |z|^2} + O_\prec\left(|w|^{1/2}\Psi_\theta^2 + \|[Z]\| + \|\langle Z \rangle\|\right), \tag{2.169}$$

and

$$\Upsilon(w, m_1) \prec |w|^{1/2}\Psi_\theta^2 + \|[Z]\| + \|\langle Z \rangle\|. \tag{2.170}$$

For $w \in \mathbf{D}_L$,

$$m_2 = \frac{1+m_1}{-w\left(1+m_1\right)^2 + |z|^2} + O_\prec\left((N\eta)^{-1} + \|[Z]\| + \|\langle Z \rangle\|\right), \tag{2.171}$$

and

$$\Upsilon(w, m_1) \prec (N\eta)^{-1} + \|[Z]\| + \|\langle Z \rangle\|. \tag{2.172}$$

Proof. The proof is almost the same as the one for Lemma 2.4.5, we only lay out the difference. We first consider the case $w \in \mathbf{D}$. By Proposition 2.4.1, the event Ξ holds with high probability. Hence without loss of generality, we may assume Ξ holds throughout the following proof. Using (2.86), we get

$$\frac{1}{N}\sum_{k\in\mathcal{I}_1} \begin{pmatrix} |d_k|^2 & 0 \\ 0 & 1 \end{pmatrix} \left(G_{[kk]} - G_{[kk]}^{[i]}\right) = \begin{pmatrix} |d_i|^2 & 0 \\ 0 & 1 \end{pmatrix} \frac{G_{[ii]}}{N} + \frac{1}{N}\sum_{k\neq i} \begin{pmatrix} |d_k|^2 & 0 \\ 0 & 1 \end{pmatrix} G_{[ki]} G_{[ii]}^{-1} G_{[ik]}. \tag{2.173}$$

By Proposition 2.4.1, (2.108) and (2.167), we have

$$\left\| G_{[ki]} G_{[ii]}^{-1} G_{[ik]} \right\| \prec |w|^{1/2}\Psi_\theta^2.$$

By Lemma 2.3.7, it is easy to verify that $\left\|G_{[ii]}/N\right\| \le C|w|^{1/2}\Psi_\theta^2$. Plugging it into (2.173), we get

$$\left|m_{1,2}^{[i]} - m_{1,2}\right| \prec |w|^{1/2}\Psi_\theta^2. \tag{2.174}$$

By (2.131) and (2.174), the error ϵ_b in (2.139) is

$$\epsilon_b = O_\prec(|w|^{1/2}\Psi_\theta^2) - \pi_{[i]}Z_{[i]}\pi_{[i]}\left[1 + O_\prec(|w|^{1/2}\Psi_\theta)\right] = O_\prec(|w|^{1/2}\Psi_\theta^2) - \pi_{[i]}Z_{[i]}\pi_{[i]}.$$

Then following the arguments in Lemma 2.4.5, we can prove the desired result. For $w \in \mathbf{D}_L$, the proof is similar by using (2.120). □

In the following lemma, we shall prove stronger bounds on $[Z]$ and $\langle Z\rangle$ by keeping track of the cancellation effects due to the average over the index i. Its proof is given in Appendix 2.7.4.

Lemma 2.4.9. *(Fluctuation averaging) Fix* $|z|^2 \le 1-\tau$. *Suppose* Φ *and* Φ_o *are positive,* N*-dependent deterministic functions satisfying* $N^{-1/2} \le \Phi, \Phi_o \le N^{-c}$ *for some constant* $c > 0$. *Suppose moreover that* $\Lambda \prec |w|^{-1/2}\Phi$ *and* $\Lambda_o \prec |w|^{-1/2}\Phi_o$. *Then for* $w \in \mathbf{D}$,

$$\|[Z]\| + \|\langle Z\rangle\| \prec |w|^{-1/2}\,\Phi_o^2. \tag{2.175}$$

Now we finish the proof of the entrywise local law and averaged local law on the domain $\mathbf{D}$. By Proposition 2.4.1, we can take

$$\Phi_o = |w|^{1/2}\sqrt{\frac{\operatorname{Im}(m_{1c}+m_{2c}) + |w|^{-3/8}(N\eta)^{-1/4}}{N\eta}}, \quad \Phi = \left(\frac{|w|^{1/2}}{N\eta}\right)^{1/4},$$

in Lemma 2.4.9, with $\Lambda_o \prec \Psi_\theta \prec |w|^{-1/2}\Phi_o$ and $\Lambda \prec \Psi_\theta + \theta \prec |w|^{-1/2}\Phi$. Then (2.170) gives

$$\Upsilon(w, m_1) \prec \frac{|w|^{1/2}\operatorname{Im}(m_{1c}+m_{2c}) + |w|^{1/4}(N\eta)^{-1/4}}{N\eta}.$$

Using the stability Lemma 2.3.10, we get

$$|m_1-m_{1c}| \prec \frac{|w|^{1/2}\mathrm{Im}(m_{1c}+m_{2c})}{N\eta\sqrt{\kappa+\eta}}+\frac{|w|^{1/8}}{(N\eta)^{5/8}} \prec \frac{1}{N\eta}+\frac{|w|^{1/8}}{(N\eta)^{5/8}} \prec |w|^{-1/2}\left(\frac{|w|^{1/2}}{N\eta}\right)^{1/2+1/8}.$$

Here if $\sqrt{\kappa+\eta} \geq (\log N)^{-1}$, we use

$$\frac{|w|^{1/2}\mathrm{Im}(m_{1c}+m_{2c})}{N\eta\sqrt{\kappa+\eta}} \leq \frac{C\log N}{N\eta} \prec \frac{1}{N\eta};$$

if $\sqrt{\kappa+\eta} \leq (\log N)^{-1}$, we have $\mathrm{Im}(m_{1c}+m_{2c}) = O(\sqrt{\kappa+\eta})$, which also gives that

$$\frac{|w|^{1/2}\mathrm{Im}(m_{1c}+m_{2c})}{N\eta\sqrt{\kappa+\eta}} \prec \frac{1}{N\eta}.$$

We then use (2.169) to get that

$$\theta \prec |m_1-m_{1c}| + \frac{|w|^{1/2}\mathrm{Im}(m_{1c}+m_{2c})+|w|^{1/4}(N\eta)^{-1/4}}{N\eta} \prec |w|^{-1/2}\left(\frac{|w|^{1/2}}{N\eta}\right)^{1/2+1/8}. \tag{2.176}$$

Repeating the previous steps with the new estimate (2.176), we get the bound

$$\theta \prec |w|^{-1/2}\left(\frac{|w|^{1/2}}{N\eta}\right)^{\sum_{k=1}^{l}1/2^k+1/2^{l+2}}$$

after l iterations. This implies the averaged local law $\theta \prec (N\eta)^{-1}$ since l can be arbitrarily large. Finally as in (2.165) and (2.167), we have for $i \neq j \in \mathcal{I}_1$,

$$\left\|G_{[ii]}-\pi_{[i]c}\right\| + \left\|G_{[ij]}\right\| \prec \Psi_\theta + \theta \prec \sqrt{\frac{\mathrm{Im}(m_{1c}+m_{2c})}{N\eta}} + \frac{1}{N\eta}.$$

This concludes the proof of the entrywise local law and averaged local law on domain $\mathbf{D}$ when $|w|^{1/2}+|z|^2 \geq c$.

When $w \in \mathbf{D}_L$, we have proved the entrywise law (see the remark after (2.144)). Also we can prove a similar estimate as in Lemma 2.4.9, which implies

$$m_2 = \frac{1+m_1}{-w\left(1+m_1\right)^2+|z|^2} + O_\prec\left((N\eta)^{-1}\right), \quad \Upsilon(w,m_1) \prec (N\eta)^{-1}. \tag{2.177}$$

The averaged local law then follows from Lemma 2.3.10. We leave the details to the reader.

2.4.5 Proof of Theorem 2.2.18 when $|z|$ and $|w|$ are small

In the previous proof, we did not include the case where $|w|^{1/2} + |z|^2 \le \epsilon$ for some sufficiently small constant $\epsilon > 0$. The only reason is that Lemma 2.3.10 does not apply in this case. We deal with this problem in this subsection.

The main idea of this subsection is to use a different set of self-consistent equations, which has the desired stability when $|w|$ and $|z|$ are small. Multiplying (2.140) with $|d_i|^2$ and summing over i, we get

$$1(\Xi)m_1 = 1(\Xi)\left[\frac{1}{N}\sum_{i=1}^{n} l_i s_i \frac{-1-m_1}{w\left(1+s_i m_2\right)\left(1+m_1\right)-|z|^2} + O_{\prec}\left(\Psi_\theta\right)\right]. \tag{2.178}$$

Recall that $\Sigma := DD^\dagger = D^\dagger D$. We introduce a new matrix

$$\tilde{H}(w) := \begin{pmatrix} -w\Sigma^{-1} & w^{1/2}(X - D^{-1}z) \\ w^{1/2}(X - D^{-1}z)^\dagger & -wI \end{pmatrix}, \tag{2.179}$$

and define $\tilde{G} := \tilde{H}^{-1}$. By Schur's complement formula, the upper left block of $\tilde{G}$ is

$$\tilde{G}_L = \left[(X - D^{-1}z)(X - D^{-1}z)^\dagger - w\Sigma^{-1}\right]^{-1},$$

and the lower right block is

$$\tilde{G}_R = \left[(X - D^{-1}z)^\dagger \Sigma (X - D^{-1}z) - w\right]^{-1} = \left[(DX - z)^\dagger (DX - z) - w\right]^{-1} = G_R.$$

Now we write $m_{1,2}$ in another way as

$$m_1 = \frac{1}{N}\operatorname{Tr}\left[D^\dagger\left(YY^\dagger - w\right)^{-1} D\right] = \frac{1}{N}\operatorname{Tr}\tilde{G}_L, \tag{2.180}$$

$$\begin{aligned} m_2 &= \frac{1}{N}\operatorname{Tr}\tilde{G}_R = \frac{1}{N}\operatorname{Tr}\left[(X - D^{-1}z)^\dagger\Sigma(X - D^{-1}z) - w\right]^{-1} \\ &= \frac{1}{N}\operatorname{Tr}\left[(X - D^{-1}z)(X - D^{-1}z)^\dagger\Sigma - w\right]^{-1} = \frac{1}{N}\operatorname{Tr}\left(\Sigma^{-1}\tilde{G}_L\right). \end{aligned} \tag{2.181}$$

We apply the arguments in the proof of Lemma 2.4.5 to $\tilde{H}$, and obtain that

$$\tilde{G}_{[ii]}^{-1} = \begin{pmatrix} -w|d_i|^{-2} - wm_2 & -w^{1/2} z d_i^{-1} \\ -w^{1/2}\bar{z}\bar{d}_i^{-1} & -w - wm_1 \end{pmatrix} + O_{\prec}(|w|\Psi_\theta), \tag{2.182}$$

from which we get that

$$1(\Xi)\tilde{G}_{ii} = 1(\Xi)\left[\frac{-1-m_1}{w(|d_i|^{-2} + m_2)(1+m_1) - |z|^2|d_i|^{-2}} + O_{\prec}(\Psi_\theta)\right].$$

Plugging this into (2.181), we get

$$1(\Xi)m_2 = 1(\Xi)\left[\frac{1}{N}\sum_{i=1}^{n}\frac{l_i}{s_i}\frac{-1-m_1}{w(s_i^{-1} + m_2)(1+m_1) - |z|^2 s_i^{-1}} + O_{\prec}(\Psi_\theta)\right]. \tag{2.183}$$

We take the equations in (2.178) and (2.183) as our new self-consistent equations, namely,

$$1(\Xi)f_1(m_1,m_2) = 1(\Xi)O(\Psi_\theta), \quad 1(\Xi)f_2(m_1,m_2) = 1(\Xi)O(\Psi_\theta), \tag{2.184}$$

where

$$f_1(m_1,m_2) := m_1 + \frac{1}{N}\sum_i l_i s_i \frac{1+m_1}{w\left(1+s_i m_2\right)\left(1+m_1\right) - \left|z\right|^2}, \tag{2.185}$$

$$f_2(m_1,m_2) := m_2 + \frac{1}{N}\sum_i l_i \frac{1+m_1}{w(1+s_i m_2)(1+m_1) - |z|^2}. \tag{2.186}$$

According to the following lemma, this system of self-consistent equations are stable when $|w|$ and $|z|^2$ are small enough .

Lemma 2.4.10. *Suppose that* $N^{-2}|w|^{-1/2} \le \delta(w) \le (\log N)^{-1}|w|^{-1/2}$ *for* $w \in \mathbf{D}$. *Suppose* $u_{1,2} : \mathbf{D} \to \mathbb{C}$ *are Stieltjes transforms of positive integrable functions such that*

$$\max\left\{\left|f_1(u_1,u_2)(w)\right|, \left|f_2(u_1,u_2)(w)\right|\right\} \le \delta(w).$$

Then there exists an $\epsilon > 0$ *such that if* $\left|w\right|^{1/2} + |z|^2 \le \epsilon$, *we have*

$$|u_1(w) - m_{1c}(w)| + |u_2(w) - m_{2c}(w)| \le C\delta, \tag{2.187}$$

for some constant $C > 0$ *independent of* w, z *and* N.

Proof. The proof depends on the estimate of the Jacobian at (m_{1c}, m_{2c}). By (2.103) and (2.328), we have

$$m_{1c} = \frac{i\sqrt{t_0} + O(|w|^{1/2} + |z|^2)}{\sqrt{w}}, \quad m_{2c} = \frac{it_0^{-1/2} + O(|w|^{1/2} + |z|^2)}{\sqrt{w}},$$

where $t_0 = (N^{-1}\sum_{i=1}^n l_i/s_i)^{-1}$. Then we can calculate that

$$\det \begin{pmatrix} \partial_1 f_1 & \partial_2 f_1 \\ \partial_1 f_2 & \partial_2 f_2 \end{pmatrix}_{u_{1,2}=m_{1,2c}} = \det \begin{pmatrix} 1 + O(|z|^2) & t_0 + O(|w|^{1/2} + |z|^2) \\ O(|z|^2) & 2 + O(|w|^{1/2} + |z|^2) \end{pmatrix} = 2 + O(|w|^{1/2} + |z|^2).$$

We can conclude the stability by expanding $f_{1,2}(u_1, u_2)$ around (m_{1c}, m_{2c}) and using a fixed point argument as in the proof of Lemma 2.3.10 in Section 2.7.3. □

With this stability lemma, we can repeat all the arguments in the previous subsections to conclude the entrywise local law and averaged local law when $|w|^{1/2} + |z|^2 \le \epsilon$.

2.5 Anisotropic local law when T is diagonal

In this section we prove the anisotropic local law in Theorem 2.2.18 when T is diagonal. The basic idea of the proof follows from [6, section 5], and the core part of our proof is a novel way to perform the combinatorics. By the Definition 2.2.17 (ii) and Definition 3.2.4 (ii), it suffices to prove the following proposition for generalized entries of G.

Proposition 2.5.1. *Fix $|z|^2 \le 1 - \tau$ and suppose that the assumptions of Theorem 2.2.18 hold. Then for any regular domain $\mathbf{S} \subseteq \mathbf{D}$,*

$$|\langle \mathbf{u}, (G(w) - \Pi(w))\, \mathbf{v}\rangle| \prec \Psi \tag{2.188}$$

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

It is equivalent to prove that

$$\sum_{i,j\in\mathcal{I}_1} u_{[i]}^{\dagger}\left(G_{[ij]}-\Pi_{[ij]}\right)v_{[j]} \prec \Psi, \quad u_{[i]} := \begin{pmatrix} u_i \\ u_{\bar{i}} \end{pmatrix}, \quad v_{[j]} := \begin{pmatrix} v_j \\ v_{\bar{j}} \end{pmatrix}. \tag{2.189}$$

By the entrywise local law,

$$\left|\sum_{i,j} u_{[i]}^{\dagger}\left(G_{[ij]}-\Pi_{[ij]}\right)v_{[j]}\right| \le \sum_{i}\left\|G_{[ii]}-\Pi_{[ii]}\right\|\left|u_{[i]}\right|\left|v_{[i]}\right| + \left|\sum_{i\ne j} u_{[i]}^{\dagger}G_{[ij]}v_{[j]}\right| \prec \Psi + \left|\sum_{i\ne j} u_{[i]}^{\dagger}G_{[ij]}v_{[j]}\right|.$$

Thus to show (2.189), it suffices to prove

$$\left|\sum_{i\ne j} u_{[i]}^{\dagger}G_{[ij]}v_{[j]}\right| \prec \Psi. \tag{2.190}$$

Note that with the entrywise local law, one can only get that

$$\left|\sum_{i\ne j} u_{[i]}^{\dagger}G_{[ij]}v_{[j]}\right| \prec \Psi\|\mathbf{u}\|_1\|\mathbf{v}\|_1 \le N\Psi,$$

using $\|\mathbf{u}\|_1 \le N^{1/2}\|\mathbf{u}\|_2$ and $\|\mathbf{v}\|_1 \le N^{1/2}\|\mathbf{v}\|_2$. In particular, this estimate of the ℓ^1 norm is sharp when $\mathbf{u},\mathbf{v}$ are delocalized, i.e. their entries have size of order $N^{-1/2}$.

The estimate (2.190) follows from the Markov's inequality if we can prove the following lemma.

Lemma 2.5.2. *Suppose the assumptions in Proposition 2.5.1 hold. For any $p\in 2\mathbb{N}$, we have*

$$\mathbb{E}\left|\sum_{i\ne j} u_{[i]}^{\dagger}G_{[ij]}v_{[j]}\right|^p \prec \Psi^p.$$

The proof of Lemma 2.5.2 is based on the polynomialization method developed in [6, section 5]. For simplicity, we only consider the case with $w\in\mathbf{D}$ and $|z|^2\le 1-\tau$ in this section. If $w\in\mathbf{D}_L$ or $1+\tau\le|z|^2\le 1+\tau^{-1}$, the proof is almost the same.

2.5.1 Rescaling and partition of indices

For our purpose, it is convenient to define the rescaled matrix

$$R^{(J)} := w^{1/2} G^{(J)}, \tag{2.191}$$

for any $J \subset \mathcal{I}$ with $|J| \leq l$ for some fixed l. Consequently we define the control parameter Φ

$$\Phi = |w|^{1/2}\, \Psi. \tag{2.192}$$

By the entrywise law, for $w \in \mathbf{D}$,

$$R^{(J)}_{[ii]} = O_{\prec}(1), \quad \left(R^{(J)}_{[ii]}\right)^{-1} = O_{\prec}(1), \quad R^{(J)}_{[ij]} = O_{\prec}(\Phi) \text{ for } i \neq j, \tag{2.193}$$

under the above scaling. Now to prove Lemma 2.5.2, it is equivalent to prove

$$\mathbb{E}\left|\sum_{i\neq j} u^{\dagger}_{[i]} R_{[ij]} v_{[j]}\right|^{p} \prec \Phi^{p}. \tag{2.194}$$

We expand the product in (2.194) as

$$\left|\sum_{i\neq j} u^{\dagger}_{[i]} R_{[ij]} v_{[j]}\right|^{p} = \sum_{i_k \neq j_k \in \mathcal{I}_1} \prod_{k=1}^{p/2} u^{\dagger}_{[i_k]} R_{[i_k j_k]} v_{[j_k]} \cdot \prod_{k=p/2+1}^{p} \overline{u^{\dagger}_{[i_k]} R_{[i_k j_k]} v_{[j_k]}}.$$

Formally, we regard $\{i_1, ..., i_p, j_1, ..., j_p\}$ as the set of $2p$ (index) variables that take values in $\mathcal{I}_1$. Let $\mathcal{B}_p$ be the collection of all partitions of $\{i_1, ..., i_p, j_1, ..., j_p\}$ such that i_k, j_k are not in the same block for all $k = 1, ..., p$. For $\Gamma \in \mathcal{B}_p$, let $n(\Gamma)$ be the number of its blocks and define a set of $\mathcal{I}_1$-valued variables as

$$L(\Gamma) := \{b_1, ..., b_{n(\Gamma)}\}. \tag{2.195}$$

Now it is convenient to regard Γ as a symbol-to-symbol function,

$$\Gamma : \{i_1, ..., i_p, j_1, ..., j_p\} \to L(\Gamma), \tag{2.196}$$

such that each $\Gamma^{-1}(b_k)$ is a block of the partition. Then we can rewrite the sum as

$$\left|\sum_{i\neq j} u^{\dagger}_{[i]} R_{[ij]} v_{[j]}\right|^p = \sum_{\Gamma\in\mathcal{B}_p} \sum^{*}_{\substack{b_l\in\mathcal{I}_1,\\ l=1,\ldots,n(\Gamma)}} \prod_{k=1}^{p/2} u^{\dagger}_{[\Gamma(i_k)]} R_{[\Gamma(i_k)\Gamma(j_k)]} v_{[\Gamma(j_k)]} \cdot \prod_{k=p/2+1}^{p} \overline{u^{\dagger}_{[\Gamma(i_k)]} R_{[\Gamma(i_k)\Gamma(j_k)]} v_{[\Gamma(j_k)]}}, \tag{2.197}$$

where $\sum^*$ denotes the summation subject to the condition that the values of $b_1, ...b_n$ are ordered as $b_1 < b_2 < \ldots < b_n$. We pick one term from the above summation and denote

$$\Delta(\Gamma) := \prod_{k=1}^{p/2} u^{\dagger}_{[\Gamma(i_k)]} R_{[\Gamma(i_k)\Gamma(j_k)]} v_{[\Gamma(j_k)]} \cdot \prod_{k=p/2+1}^{p} \overline{u^{\dagger}_{[\Gamma(i_k)]} R_{[\Gamma(i_k)\Gamma(j_k)]} v_{[\Gamma(j_k)]}}. \tag{2.198}$$

Notations: For any $b_k \in L$, we can define a corresponding $\mathcal{I}_2$-valued variable $\bar{b}_k$ in the obvious way, and we denote

$$[L] := \{b_1, ..., b_n, \overline{b_1}, ..., \overline{b_n}\}. \tag{2.199}$$

For notational convenience, we will also use letters i, j, k, l to denote the symbols in L.

2.5.2 String and string operators

During the proof we will frequently use the following resolvent identities for rescaled matrix R. They follow immediately from Lemma 2.3.3.

Lemma 2.5.3 (Resolvent identities for $R_{[ij]}$ groups)**.** *For $k \notin J$ and $i, j \in \mathcal{I}_1 \setminus J \cup \{k\}$, we have*

$$R^{[J]}_{[ij]} = R^{[Jk]}_{[ij]} + R^{[J]}_{[ik]} \left(R^{[J]}_{[kk]}\right)^{-1} R^{[J]}_{[kj]}, \tag{2.200}$$

$$\left(R^{[J]}_{[ii]}\right)^{-1} = \left(R^{[Jk]}_{[ii]}\right)^{-1} - \left(R^{[J]}_{[ii]}\right)^{-1} R^{[J]}_{[ik]} \left(R^{[J]}_{[kk]}\right)^{-1} R^{[J]}_{[ki]} \left(R^{[Jk]}_{[ii]}\right)^{-1}, \tag{2.201}$$

$$\left(R^{[J]}_{[ii]}\right)^{-1} = w^{-1/2} H^{[J]}_{[ii]} - w^{-1} \sum_{l,l'\notin J\cup\{i\}} H^{[J]}_{[il]} R^{[Ji]}_{[ll']} H^{[J]}_{[l'i]}. \tag{2.202}$$

Furthermore, for $i \neq j$ and L defined in (2.195), we have

$$R^{[L\setminus\{ij\}]}_{[ij]} = R^{[L\setminus\{ij\}]}_{[ii]} S_{[ij]} R^{[L\setminus\{j\}]}_{[jj]}, \quad \textit{with} \quad S_{[ij]} = -w^{-1/2} H_{[ij]} + w^{-1} \sum_{k,l \notin L} H_{[ik]} R^{[L]}_{[kl]} H_{[lj]}. \tag{2.203}$$

In this section, we expand the R variables in $\Delta(\Gamma)$ using the identities in Lemma 2.5.3. During the expansion, we need to distinguish carefully between an algebraic expression and its value as a random variable.

Definition 2.5.4 (Strings)**.** *Let $\mathfrak{A}$ be an alphabet containing all symbols that may appear during the expansion, such as $R^{[J]}_{[ij]}$, $\left(R^{[J]}_{[ij]}\right)^{-1}$, $S_{[ij]}$, $u^{\dagger}_{[i]}$ and $v_{[j]}$ for $J \subset L(\Gamma)$. We define a string $\mathbf{s}$ to be a formal expression consisting of the symbols from $\mathfrak{A}$, and denote by $[\![\mathbf{s}]\!]$ the random variable represented by it. Let $\mathfrak{M}$ be the collection of all possible strings. We denote an empty string by $\emptyset$.*

Given a string $\mathbf{s}$, after an expansion of R's in it, we will get a different string $\mathbf{s}'$. However, they represent the same random variable $[\![\mathbf{s}]\!] = [\![\mathbf{s}']\!]$. During the proof, we will identify more elements of $\mathfrak{A}$ (see the symbols in (2.219)).

To perform the expansions in a systematical way, we define the following operators acting on strings. We call the symbols $R^{[J]}_{[ij]}$, $\left(R^{[J]}_{[ij]}\right)^{-1}$ to be *maximally expanded* if $J \cup \{i, j\} = L$. We call a string $\mathbf{s}$ to be *maximally expanded* if all the R symbols in $\mathbf{s}$ is maximally expanded.

Definition 2.5.5 (String operators)**.** *(i) Define an operator $\tau^{(k)}_0$ for $\Omega \in \mathfrak{M}$, in the following sense. Find the first $R^{[J]}_{[ij]}$ in Ω such that $k \notin J \cup \{i, j\}$, or the first $\left(R^{[J]}_{[ii]}\right)^{-1}$ such that $k \notin J \cup \{i\}$. If $R^{[J]}_{[ij]}$ is found, replace it with $R^{[Jk]}_{[ij]}$; if $\left(R^{[J]}_{[ii]}\right)^{-1}$ is found, replace it with $\left(R^{[Jk]}_{[ii]}\right)^{-1}$; if neither is found, $\tau^{(k)}_0(\Omega) = \Omega$ and we say that $\tau^{(k)}_0$ is trivial for Ω.*

(ii) Define an operator $\tau_1^{(k)}$ *for* $\Omega \in \mathfrak{M}$, *in the following sense. Find the first* $R_{[ij]}^{[J]}$ *in* Ω *such that* $k \notin J \cup \{i,j\}$, *or the first* $\left(R_{[ii]}^{[J]}\right)^{-1}$ *such that* $k \notin T \cup \{i\}$. *If* $R_{[ij]}^{[J]}$ *is found, replace it with* $R_{[ik]}^{[J]}\left(R_{[kk]}^{[J]}\right)^{-1}R_{[kj]}^{[J]}$; *if* $\left(R_{[ii]}^{[J]}\right)^{-1}$ *is found, replace it with* $-\left(R_{[ii]}^{[J]}\right)^{-1}R_{[ik]}^{[J]}\left(R_{[kk]}^{[J]}\right)^{-1}R_{[ki]}^{[J]}\left(R_{[ii]}^{[Jk]}\right)^{-1}$; *if neither is found,* $\tau_1^{(k)}(\Omega) = \emptyset$ *and we say that* $\tau_1^{(k)}$ *is null for* Ω.

(iii) Define an operator ρ *for* $\Omega \in \mathfrak{M}$, *in the following sense. Find each maximally expanded off-diagonal* $R_{[ij]}^{[L\setminus\{ij\}]}$ *in* Ω *and replace it with* $R_{[ii]}^{[L\setminus\{ij\}]}S_{[ij]}R_{[jj]}^{[L\setminus\{j\}]}$. *If nothing is found,* $\rho(\Omega) = \Omega$.

According to Lemma 2.5.3, for any $\Omega \in \mathfrak{M}$ we have

$$\left[\!\left[\left(\tau_0^{(k)} + \tau_1^{(k)}\right)(\Omega)\right]\!\right] = [\![\Omega]\!]\,, \quad [\![\rho(\Omega)]\!] = [\![\Omega]\!]\,. \tag{2.204}$$

Definition 2.5.6. *Define the function* $\mathcal{F}_{\text{d}-\max} : \mathfrak{M} \to \mathbb{N}$ *(where the subscript "d-max" stands for "distance to being maximally expanded") through*

$$\mathcal{F}_{\text{d}-\max}\left(R_{[ij]}^{[J]\,*}\right) = |L\setminus(J \cup \{i,j\})|\,,$$

where $*$ *could be* 1 *or* -1, *and*

$$\mathcal{F}_{\text{d}-\max}(\Omega) = \sum_{R \textit{ variables in } \Omega} \mathcal{F}_{\text{d}-\max}(R).$$

Define another function $\mathcal{F}_{\text{off}} : \mathfrak{M} \to \mathbb{N}$ *with* $\mathcal{F}_{\text{off}}(\Omega)$ *being the number of off-diagonal symbols in* Ω.

By off-diagonal symbols, we mean the terms of the form A_{st} with $s \notin \{t, \bar{t}\}$ or $A_{[ij]}$ with $i \neq j$, e.g. $R_{[ij]}^{[J]}$ and $S_{[ij]}$ with $i \neq j$. Later we will define other types of off-diagonal symbols (see (2.219)). Note that a R symbol is maximally expanded if and only if $\mathcal{F}_{\text{d}-\max}(R) = 0$ and a string Ω is maximally expanded if and only if $\mathcal{F}_{\text{d}-\max}(\Omega) = 0$. The next two lemmas are almost trivial by Definition 3.4.6.

Lemma 2.5.7. *Fix* $k \in L$. *If* $\tau_0^{(k)}(\Omega) = \Omega$ *and* $\tau_1^{(k)}(\Omega) = \emptyset$,

$$\mathcal{F}_{\mathrm{d-max}}\left(\tau_0^{(k)}(\Omega)\right) = \mathcal{F}_{\mathrm{d-max}}(\Omega), \quad \mathcal{F}_{\mathrm{d-max}}\left(\tau_1^{(k)}(\Omega)\right) = 0; \tag{2.205}$$

otherwise,

$$\mathcal{F}_{\mathrm{d-max}}\left(\tau_0^{(k)}(\Omega)\right) = \mathcal{F}_{\mathrm{d-max}}(\Omega) - 1, \quad \mathcal{F}_{\mathrm{d-max}}\left(\tau_1^{(k)}(\Omega)\right) \leq \mathcal{F}_{\mathrm{d-max}}(\Omega) + 4n(\Gamma). \tag{2.206}$$

For ρ, *we have*

$$\mathcal{F}_{\mathrm{d-max}}\left(\rho(\Omega)\right) = \mathcal{F}_{\mathrm{d-max}}(\Omega) + a, \tag{2.207}$$

where a *is the number of maximally expanded off-diagonal* R*'s in* Ω.

Lemma 2.5.8. *Fix* $k \in L$. *For any* $\Omega \in \mathfrak{M}$, *we have*

$$\mathcal{F}_{\mathrm{off}}\left(\tau_0^{(k)}(\Omega)\right) = \mathcal{F}_{\mathrm{off}}(\Omega), \quad \mathcal{F}_{\mathrm{off}}\left(\rho(\Omega)\right) = \mathcal{F}_{\mathrm{off}}(\Omega), \tag{2.208}$$

and

$$\mathcal{F}_{\mathrm{off}}(\Omega) + 1 \leq \mathcal{F}_{\mathrm{off}}\left(\tau_1^{(k)}(\Omega)\right) \leq \mathcal{F}_{\mathrm{off}}(\Omega) + 2 \quad \textit{if} \quad \tau_1^{(k)}(\Omega) \neq \emptyset. \tag{2.209}$$

2.5.3 Expansion of the strings

For simplicity of notations, throughout the rest of this section we omit the complex conjugates on the right hand side of (2.198) (if we keep the complex conjugates, the proof is the same but with slightly heavier notations). Suppose the right hand side of (2.198) is represented by a string Ω_Δ. Given a binary word $\mathbf{w} = a_1 a_2 ... a_m$ with $a_i \in \{0, 1\}$, we define the operation

$$(\Omega_\Delta)_{\mathbf{w}} := \rho\tau_{a_m}^{(b_m)} \cdots \rho\tau_{a_2}^{(b_2)}\rho\tau_{a_1}^{(b_1)}\left(\Omega_\Delta\right) \tag{2.210}$$

where $b_{qn+r} := b_r$ (recall (2.195)) for any $1 \leq r \leq n$ and $q \in \mathbb{N}$. So a binary word $\mathbf{w}$ uniquely determines an operator composition. By (2.204), $[\![(\Omega_\Delta)_{\mathbf{w}0}]\!] + [\![(\Omega_\Delta)_{\mathbf{w}1}]\!] = [\![(\Omega_\Delta)_{\mathbf{w}}]\!]$ and so we get

$$\sum_{|\mathbf{w}|=m} [\![(\Omega_\Delta)_{\mathbf{w}}]\!] = [\![\Omega_\Delta]\!]$$

for any $m \geq 1$, where $|\mathbf{w}|$ denotes the length of $\mathbf{w}$.

Lemma 2.5.9. *Given any* $\mathbf{w}$ *such that* $|\mathbf{w}| = (n^2+1)(p+6l_0)$ *and* $(\Omega_\Delta)_{\mathbf{w}} \neq \emptyset$*, then either* $\mathcal{F}_{\text{off}}((\Omega_\Delta)_{\mathbf{w}}) \geq l_0 := (8/\zeta + 2)\, p$*, or* $(\Omega_\Delta)_{\mathbf{w}}$ *is maximally expanded.*

Proof. We use m_0 to denote the number of 0's in $\mathbf{w}$, and m_1 to denote the number of 1's. Furthermore, we use $m_0^{(0)}$ to denote the number of 0's corresponding to the trivial τ_0's, and $m_0^{(1)}$ to denote the number of 0's corresponding to the non-trivial τ_0's. Assume $\mathcal{F}_{\text{off}}((\Omega_\Delta)_{\mathbf{w}}) < l_0$ and $(\Omega_\Delta)_{\mathbf{w}}$ is not maximally expanded. By (2.208) and (2.209), we have $m_1 \leq l_0 - p \leq l_0$. By (2.205)-(2.207), we have

$$\mathcal{F}_{\text{d}-\max}((\Omega_\Delta)_{\mathbf{w}}) \leq \mathcal{F}_{\text{d}-\max}(\Omega_\Delta) + l_0 + 4nm_1 - m_0^{(1)}.$$

Then with $\mathcal{F}_{\text{d}-\max}(\Omega_\Delta) = np$, we get a rough bound $m_0^{(1)} + m_1 < n(p+6l_0)$. By pigeonhole principle, there are at least n 0's in a row in $\mathbf{w}$ that correspond to trivial τ_0's. This indicates that $(\Omega_\Delta)_{\mathbf{w}}$ is maximally expanded, which gives a contradiction. □

Lemma 2.5.10. *There exists constants* $C_{p,l_0}, C_{p,\zeta} > 0$ *such that*

$$\sum_{\Gamma \in \mathcal{B}_p} \sideset{}{^*}\sum_{\substack{b_l \in \mathcal{I}_1,\\ l=1,\ldots,n(\Gamma)}} \left| \mathbb{E} \sum_{\substack{|\mathbf{w}|=(n^2+1)(p+6l_0),\\ \mathcal{F}_{\text{off}}((\Omega_{\Delta(\Gamma)})_{\mathbf{w}}) \geq l_0}} [\![(\Omega_{\Delta(\Gamma)})_{\mathbf{w}}]\!] \right| \prec C_{p,l_0} N^{2p} \Phi^{l_0} \leq C_{p,\zeta} \Phi^p. \tag{2.211}$$

Proof. The first bound is due to the fact that each summand is of the order $O_\prec(\Phi^{l_0})$ and there are at most N^{2p} of them. For the second bound, we used $\Phi \leq CN^{-\zeta/2}$. □

This lemma shows that all the strings with sufficiently many off-diagonal symbols contribute at most Φ^p. It remains to handle the maximally expanded strings. Define a diagonal symbol as

$$S_{[ii]} := -\begin{pmatrix} 0 & d_i X_{i\bar{i}} \\ \bar{d}_i X^\dagger_{\bar{i}i} & 0 \end{pmatrix} + w^{-1} \sum_{k,l \notin L} H_{[ik]} R^{[L]}_{[kl]} H_{[li]}, \tag{2.212}$$

such that

$$\left(R^{[L\setminus\{i\}]}_{[ii]} \right)^{-1} = \begin{pmatrix} -w^{1/2} & -z \\ -\bar{z} & -w^{1/2} \end{pmatrix} - S_{[ii]}. \tag{2.213}$$

Notice all the R symbols in a maximally expanded string are diagonal. We taylor expand $R^{[L\setminus\{i\}]}_{[ii]}$ as

$$R^{[L\setminus\{i\}]}_{[ii]} = \left[w^{-1/2} \pi^{-1}_{[i]c} + \left(S_{[ii]} - B_i \right) \right]^{-1} = \sum_{k=0}^{l_0 - 1} \tilde{\pi}_{ic} \left[\left(S_{[ii]} - B_i \right) \tilde{\pi}_{ic} \right]^k + O_\prec \left(\Phi^{l_0} \right), \tag{2.214}$$

where $\tilde{\pi}_{[i]c} := w^{1/2} \pi_{[i]c}$, $B_i := \begin{pmatrix} w^{1/2} |d_i|^2 m_{2c} & 0 \\ 0 & w^{1/2} m_{1c} \end{pmatrix}$, and for the error term,

$$S_{[ii]} - B_i = w^{-1/2} Z^{[L\setminus\{i\}]}_{[i]} + w^{1/2} \begin{pmatrix} |d_i|^2 (m_{2c} - m^{[L]}_2) & 0 \\ 0 & m_{1c} - m^{[L]}_1 \end{pmatrix} \prec \Phi$$

by (2.131) and the averaged local law. Now for all maximally expanded $(\Omega_\Delta)_{\mathbf{w}}$ with $|\mathbf{w}| = (n^2 + 1)(p + 6l_0)$, denote by $\sigma \left[\!\left[(\Omega_\Delta)_{\mathbf{w}} \right]\!\right]$ the expression after plugging in (2.213) and (2.214) without the tail terms. Similar to Lemma 2.5.10, we have

$$\sum_{\Gamma \in \mathcal{B}_p} \sum^{*}_{\substack{b_l \in \mathcal{I}_1, \\ l = 1, \ldots, n(\Gamma)}} \left| \mathbb{E} \sum_{\substack{|\mathbf{w}| = (n^2+1)(p+6l_0), \\ (\Omega_\Delta)_{\mathbf{w}} \text{ maximally expanded}}} \left(\left[\!\left[(\Omega_{\Delta(\Gamma)})_{\mathbf{w}} \right]\!\right] - \sigma \left[\!\left[(\Omega_{\Delta(\Gamma)})_{\mathbf{w}} \right]\!\right] \right) \right| \prec C_{p,\zeta} \Phi^p.$$

From the above bound and Lemmas 3.4.7, 2.5.10, we see that to prove (2.194), it suffices to show

$$\sum_{\Gamma\in\mathcal{B}_p}\ \sum^{*}_{\substack{b_l\in\mathcal{I}_1,\\ l=1,\dots,n(\Gamma)}}\left|\mathbb{E}\sum_{\substack{|\mathbf{w}|=(n^2+1)(p+6l_0),\\ (\Omega_\Delta)_{\mathbf{w}}\text{ maximally expanded}}}\sigma\left[\!\left[(\Omega_{\Delta(\Gamma)})_{\mathbf{w}}\right]\!\right]\right|\prec C_{p,\zeta}\Phi^p. \tag{2.215}$$

We write $\sigma\left[\!\left[(\Omega_\Delta)_{\mathbf{w}}\right]\!\right]$ as a sum of monomials in terms of $S_{[ij]}$:

$$\sigma\left[\!\left[(\Omega_\Delta)_{\mathbf{w}}\right]\!\right]=\sum_i M(\mathbf{w},\Delta(\Gamma),i), \tag{2.216}$$

where i is an index to label these monomials. Note that after plugging (2.216) into (2.215), the number of summands $M(\mathbf{w},\Delta(\Gamma),i)$ inside the expectation depends only on p and ζ. Thus to show (2.215), it suffices to prove the following lemma.

Lemma 2.5.11. *Fix any $\Gamma\in\mathcal{B}_p$ and binary word $\mathbf{w}$ with $|\mathbf{w}|=(n^2+1)(p+6l_0)$. Suppose $(\Omega_\Delta)_{\mathbf{w}}$ is maximally expanded. Let $M(\mathbf{w},\Delta(\Gamma))$ be a monomial in $\sigma\left[\!\left[(\Omega_{\Delta(\Gamma)})_{\mathbf{w}}\right]\!\right]$. Then we have*

$$\sum^{*}_{b_l\in\mathcal{I}_1,l=1,\dots,n(\Gamma)}|\mathbb{E}M(\mathbf{w},\Delta(\Gamma))|\prec C_{p,\zeta}\Phi^p \tag{2.217}$$

for some constant $C_{p,\zeta}$ that only depends on p and ζ.

For the rest of this section, we fix a $\Gamma\in\mathcal{B}_p$ and a maximally expanded $(\Omega_{\Delta(\Gamma)})_{\mathbf{w}}$ with $|\mathbf{w}|=(n^2+1)(p+6l_0)$. Then we fix a monomial $M(\mathbf{w},\Delta(\Gamma))$ in $\sigma\left[\!\left[(\Omega_{\Delta(\Gamma)})_{\mathbf{w}}\right]\!\right]$. Let Ω_M be the string form of $M(\mathbf{w},\Delta(\Gamma))$ in terms of $S_{[ij]}$. It is not hard to see that

$$\mathcal{F}_{\text{off}}(\Omega_M)=\mathcal{F}_{\text{off}}((\Omega_\Delta)_{\mathbf{w}}). \tag{2.218}$$

Now we decompose $S_{[ij]}$ as

$$S_{[ij]}=S^X_{i\bar j}+S^X_{\bar i j}+S^R_{i\bar j}+S^R_{ij}+S^R_{\bar i\bar j}+S^R_{\bar i j}, \tag{2.219}$$

where we define the following symbols in $\mathfrak{A}$:

$$S^X_{i\bar{j}} := d_i X_{i\bar{j}} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad S^X_{\bar{i}j} := \bar{d}_i X^\dagger_{\bar{i}j} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \tag{2.220}$$

$$S^R_{i\bar{j}} := \sum_{k,l\notin L} d_i d_l X_{i\bar{k}} X_{l\bar{j}} \begin{pmatrix} 0 & R^{[L]}_{\bar{k}l} \\ 0 & 0 \end{pmatrix}, \quad S^R_{ij} := \sum_{k,l\notin L} d_i \bar{d}_l X_{i\bar{k}} X^\dagger_{\bar{l}j} \begin{pmatrix} R^{[L]}_{\bar{k}\bar{l}} & 0 \\ 0 & 0 \end{pmatrix}, \tag{2.221}$$

$$S^R_{\bar{i}\bar{j}} := \sum_{k,l\notin L} \bar{d}_i d_l X^\dagger_{\bar{i}k} X_{l\bar{j}} \begin{pmatrix} 0 & 0 \\ 0 & R^{[L]}_{kl} \end{pmatrix}, \quad S^R_{\bar{i}j} := \sum_{k,l\notin L} \bar{d}_i \bar{d}_l X^\dagger_{\bar{i}k} X^\dagger_{\bar{l}j} \begin{pmatrix} 0 & 0 \\ R^{[L]}_{k\bar{l}} & 0 \end{pmatrix}. \tag{2.222}$$

We expand the $S_{[ij]}$'s in $M(\mathbf{w}, \Delta(\Gamma))$ using (2.219), and write $M(\mathbf{w}, \Delta(\Gamma))$ as a sum of monomials in terms of S^X_{st} and S^R_{st}:

$$M(\mathbf{w}, \Delta(\Gamma)) = \sum_i Q(\mathbf{w}, \Delta(\Gamma), i), \tag{2.223}$$

where i is an index to label these monomials. Again it is not hard to see that

$$\mathcal{F}_{\text{off}}(\Omega_Q) = \mathcal{F}_{\text{off}}(\Omega_M) = \mathcal{F}_{\text{off}}((\Omega_\Delta)_{\mathbf{w}}). \tag{2.224}$$

Since the number of summands in (2.223) is independent of N, to prove (2.217) it suffices to show

$$\sum^{*}_{b_l \in \mathcal{I}_1, l=1,\ldots,n(\Gamma)} |\mathbb{E} Q(\mathbf{w}, \Delta(\Gamma))| \prec C_{p,\zeta} \Phi^p \tag{2.225}$$

for any monomial $Q(\mathbf{w}, \Delta(\Gamma))$ in (2.223). Throughout the following, we fix a $Q(\mathbf{w}, \Delta(\Gamma))$ with nonzero expectation, and denote by Ω_Q the string form of $Q(\mathbf{w}, \Delta(\Gamma))$ in terms of S^X_{st} and S^R_{st}. Notice the R variables in S^R_{st} are maximally expanded. As a result, the S^X_{st} variables are independent of S^R_{st} variables in $Q(\mathbf{w}, \Delta(\Gamma))$. Therefore we make the following observation: if S^X_{st} appears as a symbol in Ω_Q, then Ω_Q contains at least two of them.

Definition 2.5.12. *Recall* Γ *defined in (2.196). Let* h *be the number of blocks of* Γ *whose size is* 1*, i.e.*

$$h := \sum_{l=1}^{n(\Gamma)} \mathbf{1}\left(\left|\Gamma^{-1}(b_l)\right| = 1\right). \tag{2.226}$$

For $l = 1, ..., n$*, define*

$$I_l := \left|\{i_1, \ldots, i_p\} \cap \Gamma^{-1}(b_l)\right|, \quad J_l := \left|\{j_1, \ldots, j_p\} \cap \Gamma^{-1}(b_l)\right|.$$

Lemma 2.5.13. *Suppose for any* $b_1, ..., b_n$ *taking distinct values in* $\mathcal{I}_1$*,*

$$|\,\mathbb{E}\, Q(\mathbf{w}, \Delta(\Gamma))| \prec C N^{-h/2} \Phi^p \prod_{l=1}^{n} \left|u_{[b_l]}\right|^{I_l} \left|v_{[b_l]}\right|^{J_l} \tag{2.227}$$

holds for some constant C *independent of* N*. Then the estimate (2.225) holds.*

Proof. By Cauchy-Schwarz inequality,

$$\sum_{k=1}^{N} \left|u_{[k]}\right|^{a} \left|v_{[k]}\right|^{b} \leq \begin{cases} N^{1/2} & \text{if } a+b=1 \\ 1 & \text{if } a+b \geq 2 \end{cases}.$$

Then using $h = \sum_{l=1}^{n} \mathbf{1}\left(I_l + J_l = 1\right)$, we get

$$\sum_{b_l \in \mathcal{I}_1, l=1,\ldots,n(\Gamma)}^{*} |\mathbb{E} Q(\mathbf{w}, \Delta(\Gamma))| \prec C \Phi^p N^{-h/2} \prod_{l=1}^{n} \sum_{b_l \in \mathcal{I}_1} \left|u_{[b_l]}\right|^{I_l} \left|v_{[b_l]}\right|^{J_l} \leq C \Phi^p.$$

□

Hence it suffices to prove (2.227). The key is to extract the $N^{-h/2}$ factor from $\mathbb{E}\, Q(\mathbf{w}, \Delta(\Gamma))$. For this purpose, we need to keep track of the indices in L during the expansion.

Definition 2.5.14. *Define a function* $\mathcal{F}_{\text{in}} : L \times \mathfrak{M} \to \mathbb{N}$ *with* $\mathcal{F}_{\text{in}}(l, \Omega)$ *giving the number of times* l *or* $\bar{l}$ *appears as an index of an off-diagonal* R *or* S *symbol in* Ω*.*

The following lemma follows immediately from Definition 3.4.6 and the expansions we have done to obtain Ω_Q from $(\Omega_\Delta)_{\mathbf{w}}$.

Lemma 2.5.15. *(1) For any string Ω, if $\tau_0^{(k)}$ is not trivial for Ω, then*

$$\mathcal{F}_{\text{in}}\left(l, \tau_0^{(k)}(\Omega)\right) = \mathcal{F}_{\text{in}}(l, \Omega), \quad \mathcal{F}_{\text{in}}\left(l, \tau_1^{(k)}(\Omega)\right) = \mathcal{F}_{\text{in}}(l, \Omega) + a, \ a \in \{0, 2\}. \tag{2.228}$$

(2) For any string Ω,

$$\mathcal{F}_{\text{in}}\left(l, \rho(\Omega)\right) = \mathcal{F}_{\text{in}}(l, \Omega). \tag{2.229}$$

(3) For any maximally expanded $(\Omega_\Delta)_{\mathbf{w}}$,

$$\mathcal{F}_{\text{in}}(l, \Omega_Q) = \mathcal{F}_{\text{in}}(l, (\Omega_\Delta)_{\mathbf{w}}). \tag{2.230}$$

Let Ω_Q^X be the substring of Ω_Q containing only S^X symbols, and Ω_Q^R be the substring of Ω_Q containing only S^R symbols. Define

$$\mathcal{V} := \{l \in L | \ \mathcal{F}_{\text{in}}(l, \Omega_\Delta) = 1\}, \tag{2.231}$$

and

$$\mathcal{V}_0 := \{l \in L | \ \mathcal{F}_{\text{in}}(l, \Omega_\Delta) = 1 \text{ and } \mathcal{F}_{\text{in}}(l, \Omega_Q^X) = 0\}, \tag{2.232}$$

$$\mathcal{V}_1 := \{l \in L | \ \mathcal{F}_{\text{in}}(l, \Omega_\Delta) = 1 \text{ and } \mathcal{F}_{\text{in}}(l, \Omega_Q^X) \geq 2\}. \tag{2.233}$$

Recall the observation above Definition 2.5.12, we have $\mathcal{V} = \mathcal{V}_0 \cup \mathcal{V}_1$ and

$$h = |\mathcal{V}| = |\mathcal{V}_0| + |\mathcal{V}_1|.$$

Let n_X be the number of off-diagonal S^X symbols in Ω_Q^X and n_R be the number of off-diagonal S^R symbols in Ω_Q^R. Note that $n_o := n_X + n_R$ is the total number of off-diagonal symbols in Ω_Q.

2.5.4 Introduction of graphs and conclusion of the proof

We introduce graphs to conclude the proof of (2.227). We use a connected graph to represent the string Ω_Q, call it by $\mathfrak{G}_{Q0}$. The indices in $[L]$ are represented by black nodes in $\mathfrak{G}_{Q0}$. The S^X_{st} or S^R_{st} symbols in Ω_Q are represented by edges connecting the nodes s and t. We also define colors for the nodes and edges, where the color set for nodes is $\{black, white\}$ and the color set for edges is $\{S^X, S^R, X, R\}$. In $\mathfrak{G}_{Q0}$, all the nodes are black, all S^X edges are assigned S^X color and all S^R edges are assigned S^R color. We show a possible graph in Fig. 3. In this subsection, we identify an index with its node representation, and a symbol with its edge representation.

Definition 2.5.16. *Define function* deg *on the nodes set* $[L]$ *such that* $\deg(l)$ *gives the number of* S^R *edges connecting to the node* l.

By Lemma 2.5.15, we see that for any $l \in \mathcal{V}_0$,

$$\mathcal{F}_{\rm in}(l, \Omega_Q) \equiv \deg(l) + \deg(\bar{l}) \equiv 1 \pmod 2. \tag{2.234}$$

Hence

$$|\mathcal{V}_0| = \sum_{l\in\mathcal{V}_0} \left[\mathcal{F}_{\rm in}(l, \Omega_Q) \mod 2\right] \le \sum_{l\in\mathcal{V}_0} \left[(\deg(l) \mod 2) + \big(\deg(\bar{l}) \mod 2\big)\right]. \tag{2.235}$$

Now we expand the S^R edges. Take the $S^R_{i\bar{j}}$ edge as an example (recall (2.221)). We replace the $S^R_{i\bar{j}}$ edge with an R-group, defined as following. We add two white colored nodes to represent the summation indices $\bar{k}, l \notin [L]$, two X-colored edges to represent $X_{i\bar{k}}$ and $X_{l\bar{j}}$, and an R-colored edge connecting $\bar{k}$ and l to represent $\begin{pmatrix} 0 & R^{[L]}_{\bar{k}l} \\ 0 & 0 \end{pmatrix}$. We call the subgraph consisting of the three new edges and their nodes an R-group. If $i = j$, we call it a diagonal R-group; otherwise, call it an off-diagonal R-group. We expand all the S^R edges in $\mathfrak{G}_{Q0}$ into R-groups and call the resulting graph $\mathfrak{G}_{Q1}$. For example, after expanding the S^R edges in Fig. 3, we get the graph in Fig. 8. In the graph $\mathfrak{G}_{Q1}$, the R edges, X edges and S^X edges are mutually independent, since the R symbols are maximally expanded, and the white nodes are different from the black nodes.

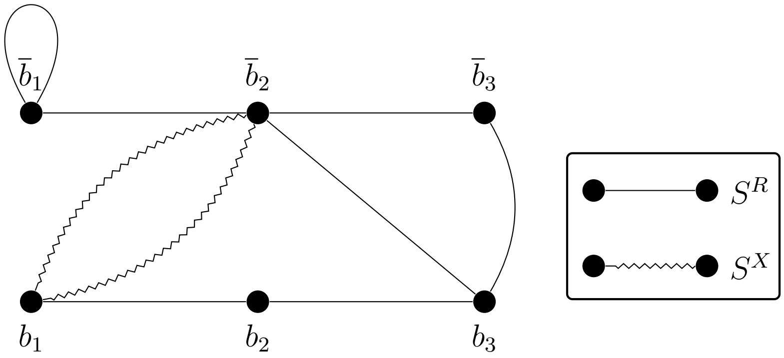


Figure 3: An example of the graph $\mathfrak{G}_{Q0}$.

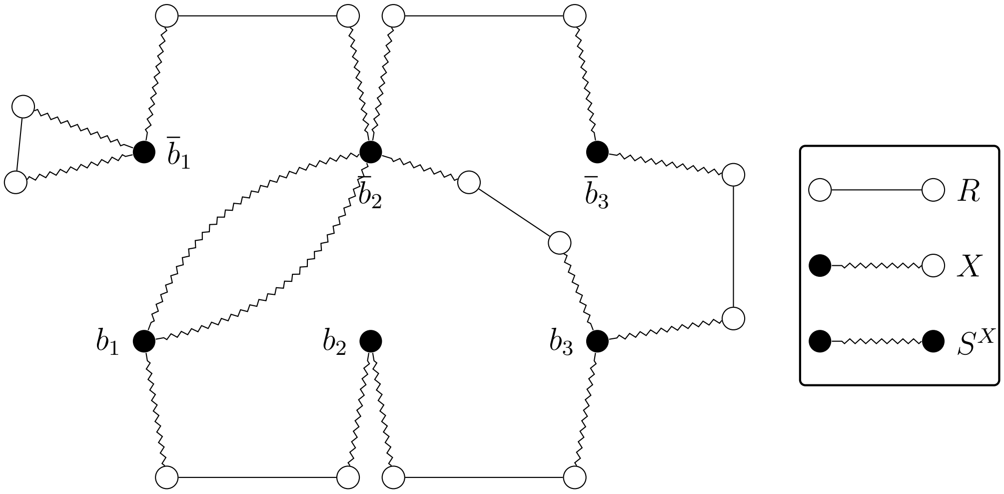
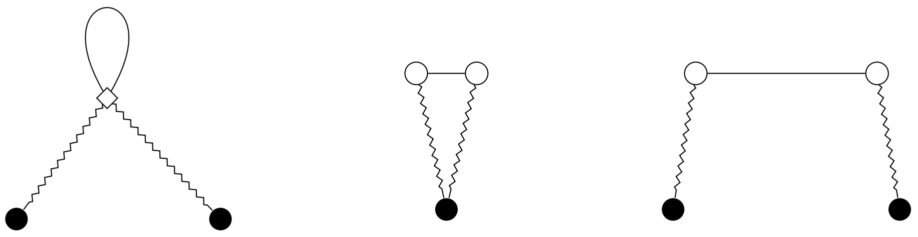


Figure 4: The resulting graph $\mathfrak{G}_{Q1}$ after expanding each S^R in Fig. 3 into R-groups.

Notice that each white node represents a summation index. As we have done for the black nodes, we first partition the white nodes into blocks and then assign values to the blocks when doing the summation. Let W be the set of all white nodes in $\mathfrak{G}_{Q1}$, and let $\mathcal{W}$ be the collection of all partitions of W. Fix a partition $\gamma \in \mathcal{W}$ and denote its blocks by $W_1, ..., W_{m(\gamma)}$. If two white nodes of some off-diagonal R-group happen to lie in the same block, then we merge the two nodes into one diamond white node (Fig. 5a). All the other white nodes are called normal (Fig. 5b). Let $n_R^{(d)}$ be the number of diamond nodes (which is $\leq$ the number of diagonal R-edges in $\mathfrak{G}_{Q1}$). Then we trivially have (recall Definition 2.5.16)

$$\# \text{ of white nodes} = -n_R^{(d)} + \sum_{k=1}^{n} \left[\deg(b_k) + \deg(\bar{b}_k) \right]. \tag{2.236}$$



(a) Diamond white node.

(b) Normal white nodes.

Figure 5: Two types of white nodes

By (2.235), there are at least $|\mathcal{V}_0|$ black nodes with odd deg in $[\mathcal{V}_0]$ (where $[\mathcal{V}_0]$ is defined in the obvious way). WLOG, we may assume these nodes are $b_1, ..., b_{|\mathcal{V}_0|}$. To have nonzero expectation, each white block must contain at least two white nodes. Therefore for each $k = 1, ..., |\mathcal{V}_0|$, there exists a block connecting to b_k which contains at least 3 white nodes. Call such a block $W(b_k)$, and denote by $A(b_k)$ the set of the adjacent white nodes to b_k in $W(b_k)$. Be careful that the $W(b_k)$'s or $A(b_k)$'s are not necessarily distinct.

WLOG, let $W_1, ..., W_d$ be the distinct blocks among all $W(b_k)$'s. Define

$$\mathcal{V}_{00} := \{b_k |\ A(b_k) \text{ has no normal white nodes, } 1 \le k \le |\mathcal{V}_0|\},$$

and

$$\mathcal{V}_{01} := \{b_k |\ A(b_k) \text{ has at least one normal white node, } 1 \le k \le |\mathcal{V}_0|\}.$$

The following lemma gives the key estimates we need.

Lemma 2.5.17. *For any partition* $\gamma \in \mathcal{W}$,

$$m(\gamma) \le \frac{-|\mathcal{V}_{00}| - |\mathcal{V}_{01}|/2 - n_R^{(d)} + \sum_{k=1}^{n} \left[\deg(b_k) + \deg(\bar{b}_k)\right]}{2}, \tag{2.237}$$

and

$$n_X + n_R \ge p + |\mathcal{V}_1| + |\mathcal{V}_{00}|, \quad n_X \ge |\mathcal{V}_1|, \quad n_R^{(d)} \ge |\mathcal{V}_{00}|. \tag{2.238}$$

Proof. The second inequality of (2.238) can be proved easily through

$$|\mathcal{V}_1| \le \left|\{k \in L | \mathcal{F}_{\text{in}}(k, \Omega_Q^X) \ge 2\}\right| \le n_X.$$

Notice for $b_k \in \mathcal{V}_0$, $A(b_k)$ contains at least three diamond white nodes, while each of the white node is shared by another b_l. Thus we trivially have $|\mathcal{V}_{00}| \le n_R^{(d)}$.

Now we prove (2.237). A diamond white node is connected to two black nodes and a normal white node is connected to one black node. Hence a diamond white node belongs to two sets $A(b_{k_1}), A(b_{k_2})$, and a normal white node belongs to exactly one set $A(b_k)$. Therefore for each $i = 1, ..., d$, if W_i contains exactly one $A(b_k)$, then

$$|W_i| \ge 3 \ge 2 + \mathbf{1}_{\mathcal{V}_{01}}(b_k) + \frac{\mathbf{1}_{\mathcal{V}_{00}}(b_k)}{2}.$$

Otherwise if W_i contains more than one $A(b_k)$, then

$$|W_i| \ge \sum_{b_k : A(b_k) \subseteq W_i} \left(2 \cdot \mathbf{1}_{\mathcal{V}_{01}}(b_k) + \frac{3}{2} \cdot \mathbf{1}_{\mathcal{V}_{00}}(b_k)\right) \ge 2 + \sum_{b_k : A(b_k) \subseteq W_i} \left(\mathbf{1}_{\mathcal{V}_{01}}(b_k) + \frac{\mathbf{1}_{\mathcal{V}_{00}}(b_k)}{2}\right).$$

Here the first inequality can be understood as following. For each black node b_k with $A(b_k) \subseteq W_i$, we count the number of white nodes in $A(b_k)$ and add them together. During the counting, we assign weight-1 to a normal white node and weight-1/2 to a diamond white node (since it is shared by two different black nodes). If $b_k \in \mathcal{V}_{00}$, there are at least three diamond white nodes in $A(b_k)$ with total weight $\geq 3/2$; if $b_k \in \mathcal{V}_{01}$, there are at least one normal white node and two other white nodes in $A(b_k)$ with total weight ≥ 2. Thus $\sum_{b_k: A(b_k) \subseteq W_i} (2 \cdot \mathbf{1}_{\mathcal{V}_{01}}(b_k) + 3/2 \cdot \mathbf{1}_{\mathcal{V}_{00}}(b_k))$ is smaller than the number of white nodes in W_i. Then summing $|W_i|$ over i, we get

$$\sum_{i=1}^{d} |W_i| \geq 2d + |\mathcal{V}_{01}| + \frac{|\mathcal{V}_{00}|}{2}.$$

For the other $m - d$ blocks, each of them contains at least two white nodes. Therefore

$$2m + |\mathcal{V}_{01}| + \frac{|\mathcal{V}_{00}|}{2} \leq \sum_{i=1}^{d} |W_i| + 2(m-d) \leq -n_R^{(d)} + \sum_{k=1}^{n} \left[\deg(b_k) + \deg(\bar{b}_k)\right],$$

where we used (2.236) in the last step. This proves (2.237).

For $b_k \in \mathcal{V}_{00}$, $A(b_k)$ contains at least three white nodes from off-diagonal R-groups,

$$\mathcal{V}_{00} \subseteq \{b_k \in L \mid \mathcal{F}_{\text{in}}(b_k, \Omega_\Delta) = 1 \text{ and } \mathcal{F}_{\text{in}}(b_k, \Omega_Q^R) \geq 3\} =: \mathcal{V}_2.$$

Recall Lemma 2.5.15, only $\tau_1^{(k)}$ can increase $\mathcal{F}_{\text{in}}$. Thus $\mathbf{w}$ contains $\tau_1^{(b_k)}$ for each $b_k \in \mathcal{V}_1 \cup \mathcal{V}_2$ (recall the definition of $\mathcal{V}_1$ in (2.233)). Therefore by (2.209), (2.224) and the fact that $\mathcal{V}_{00}$ and $\mathcal{V}_1$ are disjoint, we have

$$n_X + n_R = \mathcal{F}_{\text{off}}((\Omega_\Delta)_{\mathbf{w}}) \geq \mathcal{F}_{\text{off}}(\Omega_\Delta) + |\mathcal{V}_1 \cup \mathcal{V}_2| \geq p + |\mathcal{V}_1| + |\mathcal{V}_{00}|.$$

This proves the first inequality of (2.238). $\square$

Now we prove (2.227). By (2.4) and (2.193), a diagonal R edge contributes 1, an off-diagonal R edge contributes Φ, and an S^X or X edge contributes $N^{-1/2}$. Denote

$$\mathcal{U} = \prod_{l=1}^{n} \left|u_{[b_l]}\right|^{I_l} \left|v_{[b_l]}\right|^{J_l}.$$

Then using Lemma 3.2.5, we get

$$\begin{aligned}
|\mathbb{E}Q(\mathbf{w}, \Delta(\Gamma))| &\prec C\mathcal{U}\big(N^{-1/2}\big)^{n_X} \sum_{\gamma\in\mathcal{W}} \sideset{}{^*}\sum_{\gamma(W_1),\dots,\gamma(W_m)\in\mathcal{I}\setminus L} \Phi^{n_R-n_R^{(d)}} \prod_{k=1}^{n} \big(N^{-1/2}\big)^{\deg(b_k)+\deg\left(\bar{b}_k\right)} \\
&\le C\mathcal{U}N^{-n_X/2} \sum_{\gamma\in\mathcal{W}} N^{m-\frac{\sum_{k=1}^{n} \deg\left(b_k\right)+\deg\left(\bar{b}_k\right)}{2}} \Phi^{n_R-n_R^{(d)}} \\
&\le C\mathcal{U}N^{-n_X/2} \sum_{\gamma\in\mathcal{W}} N^{\frac{-|\mathcal{V}_{01}|-|\mathcal{V}_{00}|/2-n_R^{(d)}}{2}} \Phi^{n_R-n_R^{(d)}} \\
&\le C\mathcal{U}N^{-h/2} \sum_{\gamma\in\mathcal{W}} N^{-(n_X-|\mathcal{V}_1|)/2} N^{-(n_R^{(d)}-|\mathcal{V}_{00}|)/2} \Phi^{n_R-n_R^{(d)}} \\
&\le C\mathcal{U}N^{-h/2} \sum_{\gamma\in\mathcal{W}} \Phi^{n_X+n_R-|\mathcal{V}_1|-|\mathcal{V}_{00}|} \le C\mathcal{U}N^{-h/2}\Phi^{p},
\end{aligned}$$

where in the third step we used (2.237), in the fourth step $h = |\mathcal{V}| = |\mathcal{V}_1| + |\mathcal{V}_{00}| + |\mathcal{V}_{01}|$, in the fifth step $N^{-1/2} \le \Phi$ and (2.238), and in the last step (2.238). Thus we have proved (2.227), which concludes the proof of Proposition 2.5.1.

2.6 Anisotropic local law: self-consistent comparison

In this section we prove Theorem 2.2.19. We first prove the anisotropic and averaged local laws under the vanishing third moment assumption (2.24). When $\eta \ge N^{-1/2+\zeta}|m_{2c}|^{-1}$, the anisotropic and averaged local laws can be established without assuming (2.24). For convenience, we only consider the case with $w \in \mathbf{D}$ and $|z|^2 \le 1-\tau$ in this section. The proof for the other cases is very similar.

Following the notations in the arguments between Theorems 2.2.18 and 2.2.19, we have

$$H(TX-z,w)=\overline{T}\begin{pmatrix} -w(D^\dagger D)^{-1} & w^{1/2}(V_1X-(UD)^{-1}z) \\ w^{1/2}(V_1X-(UD)^{-1}z)^\dagger & -wI \end{pmatrix}\overline{T}^\dagger, \quad \overline{T}:=\begin{pmatrix} UD & 0 \\ 0 & I \end{pmatrix}. \tag{2.239}$$

Now we define

$$\mathcal{G}(w):=|w|^{1/2}\begin{pmatrix} -w(D^\dagger D)^{-1} & w^{1/2}\left(V_1X-(UD)^{-1}z\right) \\ w^{1/2}\left(V_1X-(UD)^{-1}z\right)^\dagger & -wI \end{pmatrix}^{-1}=|w|^{1/2}\overline{T}^\dagger G\overline{T}. \tag{2.240}$$

Since T is invertible and $\|T\|+\|T^{-1}\|\le\tau^{-1}$ by (2.5), to prove the anisotropic law in Theorem 2.2.19, it suffices to show that for all deterministic unit vectors $\mathbf{u},\mathbf{v}\in\mathbb{C}^{\mathcal{I}}$,

$$\left\langle \mathbf{u},\left(\mathcal{G}(w)-\widetilde{\Pi}(w)\right)\mathbf{v}\right\rangle \prec \Phi(w), \tag{2.241}$$

where

$$\widetilde{\Pi}(w):=|w|^{1/2}\overline{T}^\dagger\Pi(w)\overline{T}, \quad \Phi(w):=|w|^{1/2}\Psi(w). \tag{2.242}$$

Notice we have $\|\widetilde{\Pi}\|=O(1)$ by (2.108). By the anisotropic local law in Theorem 2.2.18 and the remark around (2.51), if $X=X^{Gauss}$ is Gaussian, then (2.241) holds. Hence for a general X, it suffices to prove that

$$\left\langle \mathbf{u},\left(\mathcal{G}(X,w)-\mathcal{G}(X^{Gauss},w)\right)\mathbf{v}\right\rangle \prec \Phi(w). \tag{2.243}$$

Similar to Lemma 3.2.14, we can prove the following estimates for $\mathcal{G}$.

Lemma 2.6.1. *For $i\in\mathcal{I}_1^M$, we define $\mathbf{v}_i=V_1\mathbf{e}_i\in\mathbb{C}^{\mathcal{I}_1}$, i.e. $\mathbf{v}_i$ is the i-th column vector of V_1. Let $\mathbf{u}\in\mathbb{C}^{\mathcal{I}_1}$ and $\mathbf{w}\in\mathbb{C}^{\mathcal{I}_2}$, then we have for some constant $C>0$,*

$$\sum_{\mu\in\mathcal{I}_2}\left|\mathcal{G}_{\mathbf{w}\mu}\right|^2=|w|^{1/2}\frac{\operatorname{Im}\mathcal{G}_{\mathbf{w}\mathbf{w}}}{\eta}, \tag{2.244}$$

$$\sum_{i\in\mathcal{I}_1^M}\left|\mathcal{G}_{\mathbf{u}\mathbf{v}_i}\right|^2\le C|w|^{1/2}\frac{\operatorname{Im}\mathcal{G}_{\mathbf{u}\mathbf{u}}}{\eta}, \tag{2.245}$$

$$\sum_{i\in\mathcal{I}_1^M}\left|\mathcal{G}_{\mathbf{w}\mathbf{v}_i}\right|^2\le C\left(|w|^{-1/2}\left|\mathcal{G}_{\mathbf{w}\mathbf{w}}\right|+|w|^{1/2}\frac{\operatorname{Im}\mathcal{G}_{\mathbf{w}\mathbf{w}}}{\eta}\right), \tag{2.246}$$

$$\sum_{\mu\in\mathcal{I}_2}\left|\mathcal{G}_{\mathbf{u}\mu}\right|^2\le C\left(|w|^{-1/2}\left|\mathcal{G}_{\mathbf{u}\mathbf{u}}\right|+|w|^{1/2}\frac{\operatorname{Im}\mathcal{G}_{\mathbf{u}\mathbf{u}}}{\eta}\right). \tag{2.247}$$

2.6.1 Self-consistent comparison

Our proof basically follows the arguments in [44, Section 7] with some minor modifications. Thus we will omit some details during the proof. By polarization, it suffices to prove the following proposition. In fact, we can obtain the more general bound (2.241) by applying (2.248) to the vectors $\mathbf{u}+\mathbf{v}$ and $\mathbf{u}+i\mathbf{v}$, respectively.

Proposition 2.6.2. *Fix $|z|^2\le 1-\tau$ and suppose that the assumptions of Theorem 2.2.19 hold. If (2.24) holds or $\eta\ge N^{-1/2+\zeta}|m_{2c}|^{-1}$, then for any regular domain $\mathbf{S}\subseteq\mathbf{D}$,*

$$\left\langle\mathbf{v},\left(\mathcal{G}(w)-\widetilde{\Pi}(w)\right)\mathbf{v}\right\rangle\prec\Phi(w) \tag{2.248}$$

uniformly in $w\in\mathbf{S}$ and any deterministic unit vectors $\mathbf{v}\in\mathbb{C}^{\mathcal{I}}$.

We first assume that (2.24) holds. Then we will show how to modify the arguments to prove the $\eta\ge N^{-1/2+\zeta}|m_{2c}|^{-1}$ case. The proof consists of a bootstrap argument from larger scales to smaller scales in multiplicative increments of $N^{-\delta}$, where

$$\delta\in\left(0,\frac{\zeta}{2C_0}\right), \tag{2.249}$$

with $C_0>0$ being a universal constant that will be chosen large enough in the proof. For any $\eta\ge\left|m_{1c}\right|^{-1}N^{-1+\zeta}$, we define

$$\eta_l:=\eta N^{\delta l}\ \text{for}\ \ l=0,...,L-1,\quad \eta_L:=1. \tag{2.250}$$

where $L \equiv L(\eta) := \max\left\{l \in \mathbb{N} \mid \eta N^{\delta(l-1)} < 1\right\}$. Note that $L \le 2\delta^{-1}$.

By (3.54), the function $w \mapsto \mathcal{G}(w) - \widetilde{\Pi}(w)$ is Lipschitz continuous in $\mathbf{S}$ with Lipschitz constant bounded by CN^3. Thus to prove (2.248) for all $w \in \mathbf{S}$, it suffices to show that (2.248) holds for all w in some discrete but sufficiently dense subset $\widehat{\mathbf{S}} \subset \mathbf{S}$. We will use the following discretized domain $\widehat{\mathbf{S}}$.

Definition 2.6.3. *Let* $\widehat{\mathbf{S}}$ *be an* N^{-10}*-net of* $\mathbf{S}$ *such that* $|\widehat{\mathbf{S}}| \le N^{20}$ *and*

$$E + i\eta \in \widehat{\mathbf{S}} \Rightarrow E + i\eta_l \in \widehat{\mathbf{S}} \text{ for } l = 1, ..., L(\eta).$$

The bootstrapping is formulated in terms of two scale-dependent properties $(\mathbf{A}_m)$ and $(\mathbf{C}_m)$ defined on the subsets

$$\widehat{\mathbf{S}}_m := \left\{w \in \widehat{\mathbf{S}} \mid \operatorname{Im} w \ge N^{-\delta m}\right\}.$$

$(\mathbf{A}_m)$ For all $w \in \widehat{\mathbf{S}}_m$, all deterministic unit vector $\mathbf{v}$, and all X satisfying (2.3)-(2.4), we have

$$\operatorname{Im} \mathcal{G}_{\mathbf{vv}}(w) \prec |w|^{1/2} \operatorname{Im}\left[m_{1c}(w) + m_{2c}(w)\right] + N^{C_0\delta}\Phi(w). \tag{2.251}$$

$(\mathbf{C}_m)$ For all $w \in \widehat{\mathbf{S}}_m$, all deterministic unit vector $\mathbf{v}$, and all X satisfying (2.3)-(2.4), we have

$$\left|\mathcal{G}_{\mathbf{vv}}(w) - \widetilde{\Pi}_{\mathbf{vv}}(w)\right| \prec N^{C_0\delta}\Phi(w). \tag{2.252}$$

It is trivial to see that property $(\mathbf{A}_0)$ holds. Moreover, it is easy to observe the following result.

Lemma 2.6.4. *For any* m*, property* $(\mathbf{C}_m)$ *implies property* $(\mathbf{A}_m)$*.*

Proof. This result follows from (2.110). $\square$

The key step is the following induction result.

Lemma 2.6.5. *For any* $1 \leq m \leq 2\delta^{-1}$*, property* $(\mathbf{A}_{m-1})$ *implies property* $(\mathbf{C}_m)$*.*

Combining Lemmas 2.6.4 and 2.6.5, we conclude that (2.252) holds for all $w \in \widehat{\mathbf{S}}$. Since δ can be chosen arbitrarily small under the condition (2.249), we conclude that (2.248) holds for all $w \in \widehat{\mathbf{S}}$, and Proposition 2.6.2 follows. What remains now is the proof of Lemma 2.6.5. Denote

$$F_{\mathbf{v}}(X, w) = \left| \mathcal{G}_{\mathbf{vv}}(X, w) - \widetilde{\Pi}_{\mathbf{vv}}(w) \right|. \tag{2.253}$$

By Markov's inequality, it suffices to prove the following lemma.

Lemma 2.6.6. *Fix* $p \in 2\mathbb{N}$ *and* $m \leq 2\delta^{-1}$*. Suppose that the assumptions of Proposition 2.6.2, (2.24) and property* $(\mathbf{A}_{m-1})$ *hold. Then we have*

$$\mathbb{E}F_{\mathbf{v}}^p(X, w) \leq \left(N^{C_0\delta}\Phi(w) \right)^p \tag{2.254}$$

for all $w \in \widehat{\mathbf{S}}_m$ *and all deterministic unit vectors* $\mathbf{v}$*.*

In the following, we focus on proving Lemma 2.6.6. First, in order to make use of the assumption $(\mathbf{A}_{m-1})$, which has spectral parameters in $\widehat{\mathbf{S}}_{m-1}$, to get some estimates for spectral parameters in $\widehat{\mathbf{S}}_m$, we shall use the following rough bounds for $\mathcal{G}_{\mathbf{xy}}$.

Lemma 2.6.7. *For any* $w = E + i\eta \in \mathbf{S}$ *and* $\mathbf{x}, \mathbf{y} \in \mathbb{C}^{\mathcal{I}}$*, we have*

$$\begin{aligned} \left| \mathcal{G}_{\mathbf{xy}}(w) - \widetilde{\Pi}_{\mathbf{xy}}(w) \right| \prec & N^{2\delta} \sum_{l=1}^{L(\eta)} [\operatorname{Im} \mathcal{G}_{\mathbf{x}_1\mathbf{x}_1}(E + i\eta_l) + \operatorname{Im} \mathcal{G}_{\mathbf{x}_2\mathbf{x}_2}(E + i\eta_l) \\ & + \operatorname{Im} \mathcal{G}_{\mathbf{y}_1\mathbf{y}_1}(E + i\eta_l) + \operatorname{Im} \mathcal{G}_{\mathbf{y}_2\mathbf{y}_2}(E + i\eta_l)] + \|\mathbf{x}\|_2 \|\mathbf{y}\|_2, \end{aligned}$$

where $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$ *and* $\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}$ *for* $\mathbf{x}_1, \mathbf{y}_1 \in \mathbb{C}^{\mathcal{I}_1}$ *and* $\mathbf{x}_2, \mathbf{y}_2 \in \mathbb{C}^{\mathcal{I}_2}$*, and* η_l *is defined in (2.250).*

Proof. The proof is similar to the one for [44, Lemma 7.12]. □

Recall that for a given family of complex square random matrices A, we use $A = O_\prec(\zeta)$ to mean $|\langle \mathbf{v}, A\mathbf{w}\rangle| \prec \zeta \|\mathbf{v}\|_2 \|\mathbf{w}\|_2$ uniformly for all deterministic vectors $\mathbf{v}$ and $\mathbf{w}$ (see Definition 3.2.4 (ii)).

Lemma 2.6.8. *Suppose* $(\mathbf{A}_{m-1})$ *holds, then*

$$\mathcal{G}(w) - \widetilde{\Pi}(w) = O_\prec(N^{2\delta}) \tag{2.255}$$

and

$$\operatorname{Im} \mathcal{G}_{\mathbf{v}\mathbf{v}} \le N^{2\delta}\left[|w|^{1/2}\operatorname{Im}\left(m_{1c}(w) + m_{2c}(w)\right) + N^{C_0\delta}\Phi(w)\right] \tag{2.256}$$

for all $w \in \widehat{\mathbf{S}}_m$ *and all deterministic unit vector* $\mathbf{v}$.

Proof. Let $w = E + i\eta \in \widehat{\mathbf{S}}_m$. Then $E + i\eta_l \in \widehat{\mathbf{S}}_{m-1}$ for $l = 1, \ldots, L(\eta)$, and (2.251) gives $\operatorname{Im} \mathcal{G}_{\mathbf{v}\mathbf{v}}(w) \prec 1$. The estimate (2.255) now follows immediately from Lemma 2.6.7. To prove (2.256), we remark that if $s(w)$ is the Stieltjes transform of any positive integrable function on $\mathbb{R}$, the map $\eta \mapsto \eta \operatorname{Im} s(E + i\eta)$ is nondecreasing and the map $\eta \mapsto \eta^{-1}\operatorname{Im} s(E + i\eta)$ is nonincreasing. We apply them to $|w|^{-1/2}\operatorname{Im} \mathcal{G}_{\mathbf{v}\mathbf{v}}(E + i\eta)$ and $\operatorname{Im} m_{1,2c}(E + i\eta)$ to get for $w_1 = E + i\eta_1 \in \widehat{\mathbf{S}}_{m-1}$,

$$\begin{aligned}\operatorname{Im} \mathcal{G}_{\mathbf{v}\mathbf{v}}(w) &\le N^{\delta}\frac{|w|^{1/2}}{|w_1|^{1/2}}\operatorname{Im} \mathcal{G}_{\mathbf{v}\mathbf{v}}(w_1) \prec N^{\delta}\left[|w|^{1/2}\operatorname{Im}\left(m_{1c}(w_1) + m_{2c}(w_1)\right) + N^{C_0\delta}\frac{|w|^{1/2}}{|w_1|^{1/2}}\Phi(w_1)\right] \\ &\le N^{2\delta}\left[|w|^{1/2}\operatorname{Im}\left(m_{1c}(w) + m_{2c}(w)\right) + N^{C_0\delta}\Phi(w)\right],\end{aligned}$$

where we used $\Phi(w) := |w|^{1/2}\Psi(w)$ and the fact that $\eta \mapsto \Psi(E + i\eta)$ is nonincreasing, which is clear from the definition (3.37). □

Now we apply the self-consistent comparison method introduced in [44, Section 7] to prove Lemma 2.6.6. To organize the proof, we divide it into two small subsections.

Interpolation and expansion

Definition 2.6.9 (Interpolating matrices)**.** *Introduce the notation* $X^0 := X^{Gauss}$ *and* $X^1 := X$*. Let* $\rho^0_{i\mu}$ *and* $\rho^1_{i\mu}$ *be the laws of* $X^0_{i\mu}$ *and* $X^1_{i\mu}$*, respectively, for* $i \in \mathcal{I}_1^M$ *and* $\mu \in \mathcal{I}_2$*. For* $\theta \in [0,1]$*, we define the interpolated law*

$$\rho^\theta_{i\mu} := (1-\theta)\rho^0_{i\mu} + \theta\rho^1_{i\mu}.$$

We shall work on the probability space consisting of triples (X^0, X^θ, X^1) *of independent* $\mathcal{I}_1^M \times \mathcal{I}_2$ *random matrices, where the matrix* $X^\theta = (X^\theta_{i\mu})$ *has law*

$$\prod_{i\in\mathcal{I}_1^M}\prod_{\mu\in\mathcal{I}_2} \rho^\theta_{i\mu}(dX^\theta_{i\mu}). \tag{2.257}$$

For $\lambda \in \mathbb{R}$*,* $i \in \mathcal{I}_1^M$ *and* $\mu \in \mathcal{I}_2$*, we define the matrix* $X^{\theta,\lambda}_{(i\mu)}$ *through*

$$\left(X^{\theta,\lambda}_{(i\mu)}\right)_{j\nu} := \begin{cases} X^\theta_{i\mu} & \text{if } (j,\nu) \neq (i,\mu) \\ \lambda & \text{if } (j,\nu) = (i,\mu) \end{cases}.$$

We also introduce the matrices

$$\mathcal{G}^\theta(w) := \mathcal{G}\left(X^\theta, w\right), \quad \mathcal{G}^{\theta,\lambda}_{(i\mu)}(w) := \mathcal{G}\left(X^{\theta,\lambda}_{(i\mu)}, w\right),$$

according to (2.240) and the Definition 3.2.1.

We shall prove Lemma 2.6.6 through interpolation matrices X^θ between X^0 and X^1. It holds for X^0 by the anisotropic law in Theorem 2.2.18 (see the remark above (2.243)).

Lemma 2.6.10. *Lemma 2.6.6 holds if* $X = X^0$*.*

Using (2.257) and fundamental calculus, we get the following basic interpolation formula.

Lemma 2.6.11. *For $F : \mathbb{R}^{\mathcal{I}_1^M \times \mathcal{I}_2} \to \mathbb{C}$ we have*

$$\frac{d}{d\theta}\mathbb{E}F(X^{\theta}) = \sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}\left[\mathbb{E}F\left(X_{(i\mu)}^{\theta,X_{i\mu}^1}\right) - \mathbb{E}F\left(X_{(i\mu)}^{\theta,X_{i\mu}^0}\right)\right] \tag{2.258}$$

provided all the expectations exist.

We shall apply Lemma 2.6.11 with $F(X) = F_{\mathbf{v}}^p(X,w)$ for $F_{\mathbf{v}}(X,w)$ defined in (2.253). The main work is devoted to proving the following self-consistent estimate for the right-hand side of (2.258).

Lemma 2.6.12. *Fix $p \in 2\mathbb{N}$ and $m \le 2\delta^{-1}$. Suppose (2.24) and $(\mathbf{A_{m-1}})$ holds, then we have*

$$\sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}\left[\mathbb{E}F_{\mathbf{v}}^p\left(X_{(i\mu)}^{\theta,X_{i\mu}^1}\right) - \mathbb{E}F_{\mathbf{v}}^p\left(X_{(i\mu)}^{\theta,X_{i\mu}^0}\right)\right] = O\left((N^{C_0\delta}\Phi)^p + \mathbb{E}F_{\mathbf{v}}^p(X^{\theta},w)\right) \tag{2.259}$$

for all $\theta \in [0,1]$, all $w \in \widehat{\mathbf{S}}_m$, and all deterministic unit vector $\mathbf{v}$.

Combining Lemmas 2.6.10, 2.6.11 and 2.6.12 with a Grönwall argument, we can conclude the proof of Lemma 2.6.6 and hence Proposition 2.6.2.

In order to prove Lemma 2.6.12, we compare $X_{(i\mu)}^{\theta,X_{i\mu}^0}$ and $X_{(i\mu)}^{\theta,X_{i\mu}^1}$ via a common $X_{(i\mu)}^{\theta,0}$, i.e. under the assumptions of Lemma 2.6.12, we will prove

$$\sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}\left[\mathbb{E}F_{\mathbf{v}}^p\left(X_{(i\mu)}^{\theta,X_{i\mu}^u}\right) - \mathbb{E}F_{\mathbf{v}}^p\left(X_{(i\mu)}^{\theta,0}\right)\right] = O\left((N^{C_0\delta}\Phi)^p + \mathbb{E}F_{\mathbf{v}}^p(X^{\theta},w)\right) \tag{2.260}$$

for all $u \in \{0,1\}$, all $\theta \in [0,1]$, all $w \in \widehat{\mathbf{S}}_m$, and all deterministic unit vector $\mathbf{v}$.

Underlying the proof of (2.260) is an expansion approach which we will describe below. Throughout the rest of the proof, we suppose that $(\mathbf{A}_{m-1})$ holds. Also the rest of the proof is performed at a single $w \in \widehat{\mathbf{S}}_m$. Define the $\mathcal{I}^M \times \mathcal{I}^M$ (recall Definition 2.2.9) matrix $\Delta_{(i\mu)}^{\lambda}$ through

$$\left(\Delta_{(i\mu)}^{\lambda}\right)_{st} := \lambda\delta_{is}\delta_{\mu t} + \lambda\delta_{it}\delta_{\mu s}, \quad i \in \mathcal{I}_1^M, \mu \in \mathcal{I}_2. \tag{2.261}$$

Then we have for any $\lambda, \lambda' \in \mathbb{R}$ and $K \in \mathbb{N}$,

$$\mathcal{G}^{\theta,\lambda'}_{(i\mu)} = \mathcal{G}^{\theta,\lambda}_{(i\mu)} + \sum_{k=1}^{K} \alpha^k \, \mathcal{G}^{\theta,\lambda}_{(i\mu)} \left(\overline{V} \Delta^{\lambda-\lambda'}_{(i\mu)} \overline{V}^{\dagger} \, \mathcal{G}^{\theta,\lambda}_{(i\mu)} \right)^k + \alpha^{K+1} \, \mathcal{G}^{\theta,\lambda'}_{(i\mu)} \left(\overline{V} \Delta^{\lambda-\lambda'}_{(i\mu)} \overline{V}^{\dagger} \, \mathcal{G}^{\theta,\lambda}_{(i\mu)} \right)^{K+1}, \tag{2.262}$$

where $\overline{V} := \begin{pmatrix} V_1 & 0 \\ 0 & I \end{pmatrix}$ and $\alpha := \frac{w^{1/2}}{|w|^{1/2}}$. The following result provides a priori bounds for the entries of $\mathcal{G}^{\theta,\lambda}_{(i\mu)}$.

Lemma 2.6.13. *Suppose that y is a random variable satisfying $|y| \prec N^{-1/2}$. Then*

$$\mathcal{G}^{\theta,y}_{(i\mu)} - \widetilde{\Pi} = O_{\prec}(N^{2\delta}) \tag{2.263}$$

for all $i \in \mathcal{I}_1^M$ and $\mu \in \mathcal{I}_2$.

Proof. See [44, Lemma 7.14]. □

In the following, for simplicity of notations, we introduce $f_{(i\mu)}(\lambda) := F^p_{\mathbf{v}}(X^{\theta,\lambda}_{(i\mu)})$. We use $f^{(n)}_{(i\mu)}$ to denote the n-th derivative of $f_{(i\mu)}$. By Lemma 2.6.13 and expansion (2.262) we get the following result.

Lemma 2.6.14. *Suppose that y is a random variable satisfying $|y| \prec N^{-1/2}$. Then for fixed $n \in \mathbb{N}$,*

$$\left| f^{(n)}_{(i\mu)}(y) \right| \prec N^{2\delta(n+p)}. \tag{2.264}$$

By this lemma, the Taylor expansion of $f_{(i\mu)}$ gives

$$f_{(i\mu)}(y) = \sum_{n=0}^{4p} \frac{y^n}{n!} f^{(n)}_{(i\mu)}(0) + O_{\prec}(\Phi^p), \tag{2.265}$$

provided C_0 is chosen large enough in (2.249). Therefore we have for $u \in \{0,1\}$,

$$\mathbb{E}F^p_{\mathbf{v}}\left(X^{\theta,X^u_{i\mu}}_{(i\mu)}\right) - \mathbb{E}F^p_{\mathbf{v}}\left(X^{\theta,0}_{(i\mu)}\right) = \mathbb{E}\left[f_{(i\mu)}\left(X^u_{i\mu}\right) - f_{(i\mu)}(0)\right]$$

$$= \mathbb{E}\, f_{(i\mu)}(0) + \frac{1}{2N}\,\mathbb{E}\, f^{(2)}_{(i\mu)}(0) + \sum_{n=4}^{4p} \frac{1}{n!}\,\mathbb{E}\, f^{(n)}_{(i\mu)}(0)\,\mathbb{E}\left(X^u_{i\mu}\right)^n + O_\prec(\Phi^p),$$

where we used that $X^u_{i\mu}$ has vanishing first and third moments and its variance is $1/N$. Thus to show (2.260), we only need to prove for $n = 4, 5, ..., 4p$,

$$N^{-n/2} \sum_{i\in\mathcal{I}_1^M} \sum_{\mu\in\mathcal{I}_2} \left|\mathbb{E}\, f^{(n)}_{(i\mu)}(0)\right| = O\left((N^{C_0\delta}\Phi)^p + \mathbb{E}F^p_{\mathbf{v}}(X^\theta, w)\right), \tag{2.266}$$

where we used (2.4). In order to get a self-consistent estimate in terms of the matrix X^θ on the right-hand side of (2.266), we want to replace $X^{\theta,0}_{(i\mu)}$ in $f_{(i\mu)}(0) := F^p_{\mathbf{v}}(X^{\theta,0}_{(i\mu)})$ with $X^\theta = X^{\theta, X^\theta_{(i\mu)}}_{(i\mu)}$.

Lemma 2.6.15. *Suppose that*

$$N^{-n/2} \sum_{i\in\mathcal{I}_1^M} \sum_{\mu\in\mathcal{I}_2} \left|\mathbb{E}\, f^{(n)}_{(i\mu)}(X^\theta_{i\mu})\right| = O\left((N^{C_0\delta}\Phi)^p + \mathbb{E}F^p_{\mathbf{v}}(X^\theta, w)\right) \tag{2.267}$$

holds for $n = 4, ..., 4p$, Then (2.266) holds for $n = 4, ..., 4p$.

Proof. From (2.265) we can get

$$f^{(l)}_{(i\mu)}(0) = f^{(l)}_{(i\mu)}(y) - \sum_{n=1}^{4p-l} \frac{y^n}{n!} f^{(l+n)}_{(i\mu)}(0) + O_\prec(N^{l/2}\Phi^p). \tag{2.268}$$

The result follows by repeatedly applying (2.268). The details can be found in [44, Lemma 7.16]. □

Conclusion of the proof with words

What remains now is to prove (2.267). For simplicity, we abbreviate $X^\theta \equiv X$ for the remainder of the proof. In order to exploit the detailed structure of the derivatives on the left-hand side of (2.267), we introduce the following algebraic objects.

Definition 2.6.16 (Words)**.** *Given* $i \in \mathcal{I}_1^M$ *and* $\mu \in \mathcal{I}_2$*. Let* $\mathcal{W}$ *be the set of words of even length in two letters* $\{\mathbf{i}, \boldsymbol{\mu}\}$*. We denote the length of a word* $w \in \mathcal{W}$ *by* $2n(w)$ *with* $n(w) \in \mathbb{N}$*. We use bold symbols to denote the letters of words. For instance,* $w = \mathbf{t}_1\mathbf{s}_2\mathbf{t}_2\mathbf{s}_3 \cdots \mathbf{t}_n\mathbf{s}_{n+1}$ *denotes a word of length* $2n$*. Define* $\mathcal{W}_n := \{w \in \mathcal{W} : n(w) = n\}$ *to be the set of words of length* $2n$*. We require that each word* $w \in \mathcal{W}_n$ *satisfies that* $\mathbf{t}_l\mathbf{s}_{l+1} \in \{\mathbf{i}\boldsymbol{\mu}, \boldsymbol{\mu}\mathbf{i}\}$ *for all* $1 \le l \le n$*.*

Next we assign each letter $*$ *its value* $[*]$ *through* $[\mathbf{i}] := \mathbf{v}_i$*,* $[\mu] := \mu$*, where* $\mathbf{v}_i \in \mathbb{C}^{\mathcal{I}_1}$ *is defined in Lemma 2.6.1 and is regarded as a summation index. Note that it is important to distinguish the abstract letter from its value, which is a summation index. Finally, to each word* w *we assign a random variable* $A_{\mathbf{v},i,\mu}(w)$ *as follows. If* $n(w) = 0$ *we define*

$$A_{\mathbf{v},i,\mu}(w) := \mathcal{G}_{\mathbf{v}\mathbf{v}} - \widetilde{\Pi}_{\mathbf{v}\mathbf{v}}.$$

If $n(w) \ge 1$*, say* $w = \mathbf{t}_1\mathbf{s}_2\mathbf{t}_2\mathbf{s}_3 \cdots \mathbf{t}_n\mathbf{s}_{n+1}$*, we define*

$$A_{\mathbf{v},i,\mu}(w) := \mathcal{G}_{\mathbf{v}[\mathbf{t}_1]}\, \mathcal{G}_{[\mathbf{s}_2][\mathbf{t}_2]} \cdots \mathcal{G}_{[\mathbf{s}_n][\mathbf{t}_n]}\, \mathcal{G}_{[\mathbf{s}_{n+1}]\,\mathbf{v}}\,. \tag{2.269}$$

Notice the words are constructed such that, by (2.262),

$$\left(\frac{\partial}{\partial X_{i\mu}}\right)^n \left(\mathcal{G}_{\mathbf{v}\mathbf{v}} - \widetilde{\Pi}_{\mathbf{v}\mathbf{v}}\right) = (-\alpha)^n n! \sum_{w \in \mathcal{W}_n} A_{\mathbf{v},i,\mu}(w)$$

for $n = 0, 1, 2, \ldots$, with which we get that

$$\left(\frac{\partial}{\partial X_{i\mu}}\right)^n F_{\mathbf{v}}^p(X) = (-\alpha)^n n! \sum_{n_1+\cdots+n_p=n} \prod_{r=1}^{p/2} \frac{1}{n_r! n_{r+p/2}!} \times \left(\sum_{w_r \in \mathcal{W}_{n_r}} \sum_{w_{r+p/2} \in \mathcal{W}_{n_{r+p/2}}} A_{\mathbf{v},i,\mu}(w_r) \overline{A_{\mathbf{v},i,\mu}(w_{r+p/2})} \right).$$

Then to prove (2.267), it suffices to show that

$$N^{-n/2} \sum_{i \in \mathcal{I}_1^M} \sum_{\mu \in \mathcal{I}_2} \left| \mathbb{E} \prod_{r=1}^{p/2} A_{\mathbf{v},i,\mu}(w_r) \overline{A_{\mathbf{v},i,\mu}(w_{r+p/2})} \right| = O\left((N^{C_0\delta}\Phi)^p + \mathbb{E}F_{\mathbf{v}}^p(X^\theta, w) \right) \tag{2.270}$$

for $4 \leq n \leq 4p$ and all words $w_1, ..., w_p \in \mathcal{W}$ satisfying $n(w_1)+\cdots+n(w_p) = n$. To avoid the unimportant notational complications associated with the complex conjugates, we in fact prove that

$$N^{-n/2} \sum_{i\in\mathcal{I}_1^M} \sum_{\mu\in\mathcal{I}_2} \left| \mathbb{E} \prod_{r=1}^{p} A_{\mathbf{v},i,\mu}(w_r) \right| = O\left((N^{C_0\delta}\Phi)^p + \mathbb{E}F_{\mathbf{v}}^p(X^\theta, w) \right). \tag{2.271}$$

The proof of (2.270) is essentially the same but with slightly heavier notations. Treating empty words separately, we find it suffices to prove

$$N^{-n/2} \sum_{i\in\mathcal{I}_1^M} \sum_{\mu\in\mathcal{I}_2} \mathbb{E} \left| A_{\mathbf{v},i,\mu}^{p-q}(w_0) \prod_{r=1}^{q} A_{\mathbf{v},i,\mu}(w_r) \right| = O\left((N^{C_0\delta}\Phi)^p + \mathbb{E}F_{\mathbf{v}}^p(X^\theta, w) \right) \tag{2.272}$$

for $4 \leq n \leq 4p$, $1 \leq q \leq p$, and w_r such that $n(w_0) = 0$, $\sum_r n(w_r) = n$ and $n(w_r) \geq 1$ for $r \geq 1$.

To estimate (2.272) we introduce the quantity

$$\mathcal{R}_s := |\mathcal{G}_{\mathbf{v}\mathbf{v}_s}| + |\mathcal{G}_{\mathbf{v}_s\mathbf{v}}|. \tag{2.273}$$

for $s \in \mathcal{I}$, where as a convention we let $\mathbf{v}_\mu = e_\mu$ for $\mu \in \mathcal{I}_2$.

Lemma 2.6.17. *For $w \in \mathcal{W}$ we have the rough bound*

$$|A_{\mathbf{v},i,\mu}(w)| \prec N^{2\delta(n(w)+1)}. \tag{2.274}$$

Furthermore, for $n(w) \geq 1$ we have

$$|A_{\mathbf{v},i,\mu}(w)| \prec (\mathcal{R}_i^2 + \mathcal{R}_\mu^2) N^{2\delta(n(w)-1)}. \tag{2.275}$$

For $n(w) = 1$ we have better bound

$$|A_{\mathbf{v},i,\mu}(w)| \prec \mathcal{R}_i \mathcal{R}_\mu. \tag{2.276}$$

Proof. (2.274) follows immediately from the rough bound (2.255) and definition (2.269). For (2.275), we break $A_{\mathbf{v},i,\mu}(w)$ into $\mathcal{G}_{\mathbf{v}[\mathbf{t}_1]}(\mathcal{G}_{[\mathbf{s}_2][\mathbf{t}_2]}\cdots\mathcal{G}_{[\mathbf{s}_n][\mathbf{t}_n]})^{1/2}$ times $(\mathcal{G}_{[\mathbf{s}_2][\mathbf{t}_2]}\cdots\mathcal{G}_{[\mathbf{s}_n][\mathbf{t}_n]})^{1/2}\,\mathcal{G}_{[\mathbf{s}_{n+1}]\,\mathbf{v}}$ and use Cauchy-Schwarz inequality. (2.276) follows from the constraint $\mathbf{t}_1 \neq \mathbf{s}_2$ in the definition (2.269). □

By pigeonhole principle, if $n \leq 2q-2$ there exists at least two words w_r with $n(w_r) = 1$. Therefore by Lemma 2.6.17 we have

$$\left|A^{p-q}_{\mathbf{v},i,\mu}(w_0)\prod_{r=1}^{q}A_{\mathbf{v},i,\mu}(w_r)\right| \prec N^{2\delta(n+q)}F_{\mathbf{v}}^{p-q}(X)\left(\mathbf{1}(n\geq 2q-1)(\mathcal{R}_i^2+\mathcal{R}_\mu^2)+\mathbf{1}(n\leq 2q-2)\mathcal{R}_i^2\mathcal{R}_\mu^2\right). \tag{2.277}$$

By Lemma 2.6.1, we have

$$\begin{aligned}\frac{1}{N}\sum_{i\in\mathcal{I}_1^M}\mathcal{R}_i^2+\frac{1}{N}\sum_{\mu\in\mathcal{I}_2}\mathcal{R}_\mu^2 &\prec \frac{|w|^{1/2}\mathrm{Im}\,\mathcal{G}_{\mathbf{vv}}+\eta|w|^{-1/2}\,|\mathcal{G}_{\mathbf{vv}}|}{N\eta}\\ &\prec N^{2\delta}\frac{|w|\mathrm{Im}(m_{1c}+m_{2c})+|w|^{1/2}N^{C_0\delta}\Phi}{N\eta}\prec N^{(C_0+2)\delta}\Phi^2,\end{aligned} \tag{2.278}$$

where in the second step we used the two bounds in Lemma 2.6.8, $|w|^{-1/2}\eta = O(|w|\mathrm{Im}\,m_{1c})$ by Lemma 2.3.7, and in the last step the definition of Φ. Using the same method we can get

$$\frac{1}{N^2}\sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}\mathcal{R}_i^2\mathcal{R}_\mu^2 \prec \left(N^{(C_0+2)\delta}\Phi^2\right)^2. \tag{2.279}$$

Plugging (2.278) and (2.279) into (2.277), we get that the left-hand side of (2.272) is bounded by

$$N^{-n/2+2}N^{2\delta(n+q+2)}\,\mathbb{E}\,F_{\mathbf{v}}^{p-q}(X)\left(\mathbf{1}(n\geq 2q-1)\left(N^{C_0\delta/2}\Phi\right)^2+\mathbf{1}(n\leq 2q-2)\left(N^{C_0\delta/2}\Phi\right)^4\right).$$

Using $\Phi \gtrsim N^{-1/2}$, we find that the left hand side of (2.272) is bounded by

$$N^{2\delta(n+q+2)}\,\mathbb{E}\,F_{\mathbf{v}}^{p-q}(X)\left(\mathbf{1}(n\geq 2q-1)\left(N^{C_0\delta/2}\Phi\right)^{n-2}+\mathbf{1}(n\leq 2q-2)\left(N^{C_0\delta/2}\Phi\right)^{n}\right)$$

$$\le \mathbb{E}\, F_{\mathbf{v}}^{p-q}(X)\left(\mathbf{1}(n \ge 2q-1)\left(N^{C_0\delta/2+12\delta}\Phi\right)^{n-2} + \mathbf{1}(n \le 2q-2)\left(N^{C_0\delta/2+12\delta}\Phi\right)^{n}\right)$$

where we used that $q \le n$ and $n \ge 4$. Choose $C_0 \ge 25$, then by (2.249) we have $N^{C_0\delta/2+12\delta} \le N^{\zeta/2}$ and hence $N^{C_0\delta/2+12\delta}\Phi \le 1$. Moreover, if $n \ge 4$ and $n \ge 2q-1$, then $n \ge q+2$. Therefore we conclude that the left-hand side of (2.272) is bounded by

$$\mathbb{E}\, F_{\mathbf{v}}^{p-q}(X)\left(N^{C_0\delta}\Phi\right)^{q}. \tag{2.280}$$

Now (2.272) follows from Holder's inequality. This concludes the proof of (2.267), and hence of (2.260), and hence of Lemma 2.6.5. This finishes the proof of Proposition 2.6.2 under the assumption (2.24).

In the rest of this section, we prove Proposition 2.6.2 when $\eta \ge N^{-1/2+\zeta}|m_{2c}|^{-1}$ without assuming (2.24). In this case, we can verify that

$$\Phi \le N^{-1/4-\zeta/2}. \tag{2.281}$$

Following the previous arguments, we see that it suffices to prove the estimate (2.267) for $n = 3$. In other words, we need to prove the following lemma.

Lemma 2.6.18. *Fix $1 \le m \le 2\delta^{-1}$ and $p \in 2\mathbb{N}$. Let $w \in \widehat{\mathbf{S}}_m \cap \widehat{\mathbf{D}}$ (recall (2.45)) and suppose $(\mathbf{A}_{m-1})$ holds. Then we have*

$$N^{-3/2}\sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}\left|\mathbb{E}\, f_{(i\mu)}^{(3)}(X_{i\mu}^{\theta})\right| = O\left((N^{C_0\delta}\Phi)^p + \mathbb{E}F_{\mathbf{v}}^{p}(X^{\theta}, w)\right). \tag{2.282}$$

Proof. The main new ingredient of the proof is a further iteration step at a fixed w. Suppose

$$\mathcal{G} - \tilde{\Pi} = O_{\prec}(N^{2\delta}\phi) \tag{2.283}$$

for some $\phi \le 1$. By the a priori bound (2.255), (2.283) holds for $\phi = 1$. Assuming (2.283), we shall prove a self-improving bound of the form

$$N^{-3/2}\sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}\left|\mathbb{E}\, f_{(i\mu)}^{(3)}(X_{i\mu}^{\theta})\right| = O\left((N^{C_0\delta}\Phi)^p + (N^{-\zeta/4}\phi)^p + \mathbb{E}F_{\mathbf{v}}^p(X^\theta, w)\right). \tag{2.284}$$

Once (2.284) is proved, we can use it iteratively to get an increasingly accurate bound for the left hand side of (2.252). After each step, we obtain a better a priori bound (2.283) where ϕ is reduced by $N^{-\zeta/4}$. Hence after $O(\zeta^{-1})$ iterations we can get (2.282).

As in Section 2.6.1, to prove (2.284) it suffices to show

$$N^{-3/2}\left|\sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}A_{\mathbf{v},i,\mu}^{p-q}(w_0)\prod_{r=1}^{q}A_{\mathbf{v},i,\mu}(w_r)\right| \prec F_{\mathbf{v}}^{p-q}(X)(N^{(C_0-1)\delta}\Phi + N^{-\zeta/2}\phi)^q, \tag{2.285}$$

which follows from the bound

$$N^{-3/2}\left|\sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}\prod_{r=1}^{q}A_{\mathbf{v},i,\mu}(w_r)\right| \prec (N^{(C_0-1)\delta}\Phi + N^{-\zeta/2}\phi)^q. \tag{2.286}$$

Each of the three cases $q = 1, 2, 3$ can be proved as in [44, Lemma 12.7], and we leave the details to the reader. This concludes Lemma 2.6.18. □

2.6.2 Averaged local law for TX

In this section we prove the averaged local law in Theorem 2.2.19. Again for convenience, we only consider the case with $w \in \mathbf{D}$ and $|z|^2 \le 1-\tau$. First we assume (2.24) holds. The anisotropic local law proved in the previous section gives a good a priori bound. In analogy to (2.253), we define

$$\widetilde{F}(X,w) := |w|^{1/2}|m_2(w) - m_{2c}(w)| = \left|\frac{1}{N}\sum_{\nu\in\mathcal{I}_2}\mathcal{G}_{\nu\nu}(w) - |w|^{1/2}m_{2c}(w)\right|.$$

Since $\Phi^2 = O(|w|^{1/2}/(N\eta))$, it suffices to prove that $\widetilde{F} \prec \Phi^2$. Following the argument in Section 2.6.1, analogous to (2.267), we only need to prove that

$$N^{-n/2} \sum_{i\in\mathcal{I}_1^M} \sum_{\mu\in\mathcal{I}_2} \left| \mathbb{E} \left(\frac{\partial}{\partial X_{i\mu}} \right)^n \widetilde{F}^p(X) \right| = O\left((N^{\delta}\Phi^2)^p + \mathbb{E}\widetilde{F}^p(X) \right) \tag{2.287}$$

for all $n = 4, ..., 4p$. Here $\delta > 0$ is an arbitrary positive constant. Analogously to (2.271), it suffices to prove that for $n = 4, ..., 4p$,

$$N^{-n/2} \sum_{i\in\mathcal{I}_1^M} \sum_{\mu\in\mathcal{I}_2} \left| \mathbb{E} \prod_{r=1}^{p} \left(\frac{1}{N} \sum_{\nu\in\mathcal{I}_2} A_{\mathbf{e}_\nu,i,\mu}(w_r) \right) \right| = O\left((N^{\delta}\Phi^2)^p + \mathbb{E}\widetilde{F}^p(X) \right) \tag{2.288}$$

for $\sum_r n(w_r) = n$. The only difference in the definition of $A_{\mathbf{v},i,\mu}(w)$ is that when $n(w) = 0$, we define

$$A_{\mathbf{v},i,\mu}(w) := \mathcal{G}_{\mathbf{v}\mathbf{v}} - |w|^{1/2} m_{2c}.$$

Similar to (2.273) we define

$$\mathcal{R}_{\nu,s} := |\mathcal{G}_{\nu\mathbf{v}_s}| + |\mathcal{G}_{\mathbf{v}_s\nu}|. \tag{2.289}$$

By the anisotropic local law, $\mathcal{G} - \widetilde{\Pi} = O_{\prec}(\Phi)$. Hence combining with Lemma 2.6.1 and (2.110), we get

$$\frac{1}{N} \sum_{\nu\in\mathcal{I}_2} \mathcal{R}_{\nu,s}^2 \prec \frac{|w|^{1/2}\operatorname{Im}\mathcal{G}_{\mathbf{v}_s\mathbf{v}_s} + \eta|w|^{-1/2}\,|\mathcal{G}_{\mathbf{v}_s\mathbf{v}_s}|}{N\eta} \prec \frac{|w|\operatorname{Im}(m_{1c}+m_{2c}) + |w|^{1/2}\Phi}{N\eta} = O(\Phi^2). \tag{2.290}$$

Since $\mathcal{G} = O_{\prec}(1)$ by the anisotropic local law, we have

$$\left| \frac{1}{N} \sum_{\nu\in\mathcal{I}_2} A_{\mathbf{e}_\nu,i,\mu}(w) \right| \prec \frac{1}{N} \sum_{\nu\in\mathcal{I}_2} \left(\mathcal{R}_{\nu,i}^2 + \mathcal{R}_{\nu,\mu}^2 \right) \prec \Phi^2 \text{ for } n(w) \ge 1. \tag{2.291}$$

Following (2.291), for $n \ge 4$, the left-hand side of (2.288) is bounded by

$$\mathbb{E}\,\widetilde{F}^{p-q}(X)(\Phi^2)^q.$$

Applying Holder's inequality, we conclude the proof.

Then we prove the averaged local law when $\eta \geq N^{-1/2+\zeta}|m_{2c}|^{-1}$. It suffices to prove

$$N^{-3/2}\left|\sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}\mathbb{E}\left(\frac{\partial}{\partial X_{i\mu}}\right)^3\widetilde{F}^p(X)\right| = O\left(\left(N^\delta\Phi^2\right)^p + \left(\frac{N^{-c_0/2}}{N\eta}\right)^p + \mathbb{E}\widetilde{F}^p(X)\right), \tag{2.292}$$

for some small constant $c_0 > 0$. Analogous to the above arguments, it reduces to show that

$$N^{-3/2}\left|\sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}\prod_{r=1}^{q}\left(\frac{1}{N}\sum_{\nu\in\mathcal{I}_2}A_{\mathbf{e}_\nu,i,\mu}(w_r)\right)\right| = O_\prec\left(\Phi^{2q} + \left(\frac{N^{-c_0}}{N\eta}\right)^q\right), \tag{2.293}$$

where q is the number of words with nonzero length. Again we can prove the three cases $q = 1, 2, 3$ as in [44, Lemma 12.8], and we leave the details to the reader. This concludes the averaged local law.

2.7 Appendix

2.7.1 Proof of Lemma 2.2.3 and Proposition 2.2.14

We now prove Lemma 2.2.3. First is a technical lemma for f defined in (2.16).

Lemma 2.7.1. *For $w > 0$ and $|z| > 0$, f can be written as*

$$f(\sqrt{w}, m) = -\sqrt{w} + m + w^{-1/2} + \frac{1}{N}\sum_{i=1}^{n} l_i s_i\left(\frac{A_i}{m-a_i} + \frac{B_i}{m-b_i} + \frac{C_i}{m+c_i}\right), \tag{2.294}$$

where we have the following estimates for the poles and the coefficients,

$$\max\left(|z|, \frac{s_i+|z|^2}{\sqrt{w}}\right) < a_i < \frac{s_i+|z|^2}{\sqrt{w}} + |z|, \quad a_n < a_{n-1} < \ldots < a_1, \tag{2.295}$$

$$0 < b_1 < b_2 < \ldots < b_n < \min\left(|z|, \frac{|z|^2}{\sqrt{w}}\right), \tag{2.296}$$

$$\frac{-(s_i+|z|^2)+\sqrt{(s_i+|z|^2)^2+4w|z|^2}}{2\sqrt{w}} < c_i < |z|, \quad c_1 < c_2 < \ldots < c_n, \tag{2.297}$$

and

$$0 < A_i \leq 2\frac{s_i+|z|^2+\sqrt{w}|z|}{w}, \; 0 < B_i \leq 2\frac{s_i+|z|^2+\sqrt{w}|z|}{w}, \; 0 < C_i \leq \frac{s_i+|z|^2+\sqrt{w}|z|}{w}. \tag{2.298}$$

Proof. The proof is based on basic algebraic arguments. Let

$$p_i = \sqrt{w}m^3 - (s_i+|z|^2)m^2 - \sqrt{w}|z|^2 m + |z|^4.$$

It is easy to verify that

$$\Delta = 18(s_i+|z|^2)w|z|^6 + 4(s_i+|z|^2)^3|z|^4 + (s_i+|z|^2)^2 w|z|^4 + 4w^2|z|^6 - 27w|z|^8 > 0.$$

Thus p_i has three distinct real roots. By the form of p_i, we see that there are two positive roots and one negative root, call them $a_i > b_i > 0 > -c_i$. Now we perform the partial fraction expansion for the rational functions in (2.16):

$$\frac{m^2-|z|^2}{\sqrt{w}m^3-(s_i+|z|^2)m^2-\sqrt{w}|z|^2m+|z|^4} = \frac{A_i'}{m-a_i} + \frac{B_i'}{m-b_i} - \frac{C_i'}{m+c_i}, \tag{2.299}$$

where

$$A_i' = \frac{a_i^2-|z|^2}{\sqrt{w}(a_i-b_i)(a_i+c_i)}, \; B_i' = \frac{b_i^2-|z|^2}{\sqrt{w}(b_i-a_i)(b_i+c_i)}, \; C_i' = \frac{-c_i^2+|z|^2}{\sqrt{w}(c_i+a_i)(c_i+b_i)}. \tag{2.300}$$

We take $s_i = 0$ in p_i and call the resulting polynomial as

$$p_0 = \sqrt{w}m^3 - |z|^2m^2 - \sqrt{w}|z|^2m + |z|^4 = \sqrt{w}\left(m - \frac{|z|^2}{\sqrt{w}}\right)\left(m^2-|z|^2\right),$$

which has roots $m = \pm|z|, |z|^2/\sqrt{w}$. By (2.8), we have $p_1 < p_2 < \ldots < p_n < p_0$ for all $m \neq 0$. Comparing the graphs of p_i's (as cubic functions of m) for $0 \leq i \leq n$, we get that

$$\max\left(|z|, \frac{|z|^2}{\sqrt{w}}\right) < a_n < a_{n-1} < \ldots < a_1,\ 0 < b_1 < b_2 < \ldots < b_n < \min\left(|z|, \frac{|z|^2}{\sqrt{w}}\right), \tag{2.301}$$

and

$$0 < c_1 < c_2 < \ldots < c_n < |z|. \tag{2.302}$$

Thus we get (2.296). By these bounds, we see that $a_i^2 - |z|^2 > 0$, $b_i^2 - |z|^2 < 0$ and $-c_i^2 + |z|^2 > 0$, which, by (2.300), give that $A_i' > 0$, $B_i' > 0$ and $C_i' > 0$. Plugging (2.299) into f, we get immediately (2.294) with $A_i = A_i' a_i$, $B_i = B_i' b_i$ and $C_i = C_i' c_i$. The $w^{-1/2}$ term can be obtained by comparing the coefficients of the m^3 terms in (2.16) and using the normalization condition (2.9).

Now we compare p_i with $p_i' := \sqrt{w}m^3 - (s_i + |z|^2)m^2 - \sqrt{w}|z|^2 m$, which has roots

$$m = 0,\ \frac{(s_i + |z|^2) \pm \sqrt{(s_i + |z|^2)^2 + 4w|z|^2}}{2\sqrt{w}}.$$

Since $p_i' < p_i$ for all m, we get

$$a_i < \frac{(s_i + |z|^2) + \sqrt{(s_i + |z|^2)^2 + 4w|z|^2}}{2\sqrt{w}} < \frac{s_i + |z|^2}{\sqrt{w}} + |z|, \tag{2.303}$$

and

$$c_i > \frac{-(s_i + |z|^2) + \sqrt{(s_i + |z|^2)^2 + 4w|z|^2}}{2\sqrt{w}}. \tag{2.304}$$

Combining (2.302) and (2.304), we get (2.297). Then we compare p_i with $p_i'' := \sqrt{w}m^3 - (s_i + |z|^2)m^2$, which has roots $w = 0$, $(s_i + |z|^2)/\sqrt{w}$. Note that $p_i'' > p_i$ for $m > |z|^2/\sqrt{w}$, which gives $a_i > (s_i + |z|^2)/\sqrt{w}$ since $a_i > |z|^2/\sqrt{w}$. Combining this bound with (2.301) and (2.303), we get (2.295).

Finally we estimate the coefficients A_i, B_i and C_i. Using (2.300) and (2.295)-(2.297), we first can estimate that

$$\begin{aligned}
A_i' &= \frac{(a_i - |z|)(a_i + |z|)}{\sqrt{w}(a_i - b_i)(a_i + c_i)} \le \frac{a_i + |z|}{\sqrt{w}(a_i + c_i)} \le \frac{2}{\sqrt{w}}, \\
B_i' &= \frac{(|z| + b_i)(|z| - b_i)}{\sqrt{w}(a_i - b_i)(b_i + c_i)} \le \frac{|z| + b_i}{\sqrt{w}(b_i + c_i)} \le 2\frac{s_i + |z|^2 + \sqrt{w}|z|}{w|z|}, \\
C_i' &= \frac{(|z| - c_i)(c_i + |z|)}{\sqrt{w}(c_i + a_i)(c_i + b_i)} \le \frac{|z| - c_i}{\sqrt{w}(c_i + b_i)} \le \frac{s_i + |z|^2 + \sqrt{w}|z|}{w|z|},
\end{aligned}$$

with which we can get that

$$A_i = A_i' a_i \le \frac{2}{\sqrt{w}}\left(\frac{s_i + |z|^2}{\sqrt{w}} + |z|\right) = 2\frac{s_i + |z|^2 + \sqrt{w}|z|}{w}, \tag{2.305}$$

$$B_i = B_i' b_i \le 2\frac{s_i + |z|^2 + \sqrt{w}|z|}{w|z|}|z| = 2\frac{s_i + |z|^2 + \sqrt{w}|z|}{w}, \tag{2.306}$$

$$C_i = C_i' c_i \le \frac{s_i + |z|^2 + \sqrt{w}|z|}{w|z|}|z| = \frac{s_i + |z|^2 + \sqrt{w}|z|}{w}. \tag{2.307}$$

□

In (2.294), it is sometimes convenient to reorder the terms and rename the constants to write f as

$$f(m) = -\sqrt{w} + m + w^{-1/2} + \frac{1}{N}\sum_{k=1}^{2n} \frac{C_k^+}{m - x_k} + \frac{1}{N}\sum_{l=1}^{n} \frac{C_l^-}{m + y_l}. \tag{2.308}$$

where all the constants C_k^+ and C_l^- are positive and chosen such that

$$0 < x_1 < x_2 < \ldots < x_{2n}, \; 0 < y_1 < y_2 < \ldots < y_n. \tag{2.309}$$

Clearly, f is smooth on the $3n+1$ open intervals of $\mathbb{R}$ defined by

$$I_{-n} := (-\infty, -y_n), \; I_{-k} := (-y_{k+1}, -y_k) \; (k = 1, \ldots, n-1), \; I_0 := (-y_1, x_1),$$

$$I_k := (x_k, x_{k+1}) \; (k = 1, \ldots, 2n-1), \; I_{2n} := (x_{2n}, +\infty).$$

Next, we introduce the multiset $\mathcal{C}$ of critical points of f (as a function of m), using the conventions that a nondegenerate critical point is counted once and a degenerated critical point twice. First we will prove the following elementary lemma about the structure of $\mathcal{C}$ (see Fig. 6 and Fig. 7).

Lemma 2.7.2. *(Critical points) We have* $|\mathcal{C} \cap I_{-n}| = |\mathcal{C} \cap I_{2n}| = 1$ *and* $|\mathcal{C} \cap I_k| \in \{0, 2\}$ *for* $k = -n+1, \ldots, 2n-1$.

Proof. We omit the dependence of f on w for now. By (2.308) we have

$$f'(m) = 1 - \frac{1}{N}\sum_{k=1}^{2n} \frac{C_k^+}{(m-x_k)^2} - \frac{1}{N}\sum_{l=1}^{n} \frac{C_l^-}{(m+y_l)^2}, \; f''(m) = \frac{1}{N}\sum_{k=1}^{2n} \frac{2C_k^+}{(m-x_k)^3} + \frac{1}{N}\sum_{l=1}^{n} \frac{2C_l^-}{(m+y_l)^3}.$$

We see that f'' is decreasing on all the intervals I_k for $k = -n+1, \ldots, 2n-1$. Thus there is at most one point $m \in I_k$ such that $f''(m) = 0$. We conclude that f has at most two critical points on I_k. By the boundary conditions of f' on ∂I_k, we get $|\mathcal{C} \cap I_k| \in \{0, 2\}$ for $k = -n+1, \ldots, 2n-1$. For $m < -y_n$, we have $f''(m) < 0$, while for $m > x_{2n}$, we have $f''(m) > 0$. By the boundary conditions of f' on ∂I_{-n} and ∂I_{2n}, we see that f' decreases from 1 to $-\infty$ when m increases from $-\infty$ to $-y_n$, while f' increases from $-\infty$ to 1 when m increases from x_{2n} to $+\infty$. Hence we conclude that each of the intervals $(-\infty, -y_n)$ and $(x_{2n}, +\infty)$ contains a unique critical point in it, i.e. $|\mathcal{C} \cap I_{-n}| = |\mathcal{C} \cap I_{2n}| = 1$. $\square$

From this lemma, we deduce that $|\mathcal{C}| = 2p$ is even. We denote by z_{2p} the critical point in I_{-n}, z_1 the critical point in I_{2n}, and $z_2 \geq \ldots \geq z_{2p-1}$ the $2p-2$ critical points in $I_{-n+1} \cup \ldots \cup I_{2n-1}$. For $k = 1, \ldots, 2p$, we define the critical values $h_k := f(z_k)$. The next lemma is crucial in establishing the basic properties of ρ_{1c} (see e.g. Fig. 6).

Lemma 2.7.3. *(Orderings of the critical values) The critical values are ordered as* $h_1 \geq h_2 \geq \ldots \geq h_{2p}$. *Furthermore, there is an absolute constant* $C_0 > 0$ *independent of* τ *such that* $h_k \in [-C_0(\tau^{-1}|w|^{-1/2} + |z|) - \sqrt{w}, C_0(\tau^{-1}|w|^{-1/2} + |z|) - \sqrt{w}]$ *for* $k = 1, \ldots, 2p$.

Proof. Notice for the equation (2.15), if we multiply both sides with the product of all denominators in f, we get a polynomial equation $P_w(m) = 0$ with P_w being a polynomial of degree $3n + 1$. An immediate consequence is that for any fixed $w > 0$ and $E \in \mathbb{R}$, $f(\sqrt{w}, m) = E$ can have at most $3n+1$ roots in m. This fact will be useful in the proof of this lemma and Lemma 2.2.3.

For $i = -n, \ldots, 2n$, define the subset $J_i(w) := \{m \in I_i : \partial_m f(\sqrt{w}, m) > 0\}$. From Lemma 2.7.2, we deduce that if $i = -n + 1, \ldots, 2n - 1$, then $J_i \neq \emptyset$ if and only if I_i contains two distinct critical points of f, in which case J_i is an interval. Moreover, we have $J_{-n} = (-\infty, z_{2p})$ and $J_{2n} = (z_1, +\infty)$. Next, we observe that for any $-n \leq i < j \leq 2n$, we have $f(J_i) \cap f(J_j) = \emptyset$. Otherwise if there were $E \in f(J_i) \cap f(J_j)$, we would have $|\{x : f(x) = E\}| > 3n + 1$. We hence conclude that the sets $f(J_i)$, $-n \leq i \leq 2n$ can be strictly ordered. The claim $h_1 \geq h_2 \geq \ldots \geq h_{2p}$ is now reformulated as

$$f(J_i) < f(J_j) \text{ whenever } i < j \text{ and } J_i, J_j \neq \emptyset. \tag{2.310}$$

To prove (2.310), we use a continuity argument. Let $t \in (0, 1]$ and introduce

$$f^t(m) = -\sqrt{w} + m + w^{-1/2} + \frac{t}{N} \sum_{k=1}^{2n} \frac{C_k^+}{m - x_k} + \frac{t}{N} \sum_{l=1}^{n} \frac{C_l^-}{m + y_l}.$$

It is easy to check (2.310) holds for small enough $t > 0$. We claim that

$$J_i \neq \emptyset \Rightarrow J_i^t \neq \emptyset \text{ for all } t \in (0, 1]. \tag{2.311}$$

This is trivial for $i = -n, 2n$. Recall that for $-n+1 \leq i \leq 2n-1$, $J_i^t \neq \emptyset$ is equivalent to I_i containing two distinct critical points. Moreover, $\partial_t \partial_m f^t(m) < 0$ in $I_{-n+1} \cup \ldots \cup I_{2n-1}$, from which we deduce that the number of distinct critical points in each I_i, $i = -n + 1, \ldots, 2n - 1$, does not decreases as t decreases. This proves (2.311).

Next, suppose that there exist $i < j$ such that $J_i, J_j \neq \emptyset$ and $f(J_i) > f(J_j)$. From (2.311), we deduce that $J_i^t, J_j^t \neq \emptyset$ for all $t \in (0,1]$. By a simple continuity argument, we get that $f^t(J_i^t) > f^t(J_j^t)$ for all $t \in (0,1]$. However, this is impossible for small enough t as explained before (2.311). This concludes the proof of (2.310).

To prove the second statement of Lemma 2.7.3, we only need to show that $h_1 \leq C_0(\tau^{-1}|w|^{-1/2} + |z|) - \sqrt{w}$ and $h_{2p} \geq -C_0(\tau^{-1}|w|^{-1/2} + |z|) - \sqrt{w}$ for some absolute constant C_0. We only give the proof for h_1; the proof for h_{2p} is similar. At z_1, we have

$$f(z_1)+\sqrt{w} \leq (z_1+y_n)\left[1 + \frac{1}{N}\sum_{k=1}^{2n} \frac{C_k^+}{(z_1 - x_k)^2} + \frac{1}{N}\sum_{l=1}^{n} \frac{C_l^-}{(z_1 + y_l)^2}\right] + w^{-1/2} = 2(z_1+y_n) + w^{-1/2},$$

where we used

$$0 = f'(z_1) = 1 - \frac{1}{N}\sum_{k=1}^{2n} \frac{C_k^+}{(z_1 - x_k)^2} - \frac{1}{N}\sum_{l=1}^{n} \frac{C_l^-}{(z_1 + y_l)^2}. \tag{2.312}$$

Now we would like to estimate $z_1 + y_n$. Again using (2.312), we have that

$$\frac{1}{N}\sum_{k=1}^{2n} \frac{C_k^+}{(z_1 - x_{2n})^2} + \frac{1}{N}\sum_{l=1}^{n} \frac{C_l^-}{(z_1 - x_{2n})^2} \geq 1.$$

Then by (2.298) we get

$$z_1 - x_{2n} \leq \sqrt{\frac{1}{N}\sum_{k=1}^{2n} C_k^+ + \frac{1}{N}\sum_{l=1}^{n} C_l^-} \leq \sqrt{5\frac{\tau^{-1} + |z|^2 + \sqrt{w}|z|}{w}}.$$

Using the above estimates and (2.295)-(2.297), we obtain that

$$f(z_1) \leq 2\left(\sqrt{5\frac{\tau^{-1} + |z|^2 + \sqrt{w}|z|}{w}} + \frac{s_1 + |z|^2}{\sqrt{w}} + 2|z|\right) + w^{-1/2} - \sqrt{w} \leq C_0(\tau^{-1}|w|^{-1/2} + |z|) - \sqrt{w}.$$

for some constant $C_0 > 0$ that does not depend on τ. □

Proof of Lemma 2.2.3. Let $J(w) := \bigcup_{i=-n}^{2n} J_i(w)$. Given $w > 0$ such that $0 \in f(J(w))$, then the set $\{m \in \mathbb{R} : f(\sqrt{w}, m) = 0\}$ has $3n + 1$ points. Since $f(\sqrt{w}, m) = 0$ has at most $3n+1$ solutions in m, we deduce that $m_c(w)$ is real and hence $m_{1c}(w)$ is also real. Since m_{1c} is the Stieltjes transform of ρ_{1c}, we conclude that $w \notin \operatorname{supp}\rho_{1c}$. On the other hand, suppose $w > 0$ and $0 \notin f(J(w))$. Then the set of preimages $\{m \in \mathbb{R} : f(\sqrt{w}, m) = 0\} = \{m \in \mathbb{R} : P_w(m) = 0\}$ has $3n-1$ points. Since $P_w(m)$ is a degree $3n+1$ polynomial with real coefficients, we conclude that P_w has a unique root with positive imaginary part. By the uniqueness of the solution of $P_{w+i\eta}$ in $\mathbb{C}_+$ (Lemma 2.2.2) and the continuity of the roots of $P_{w+i\eta}$ in η, we conclude that $\operatorname{Im} m_c(w) > 0$ and hence $\operatorname{Im} m_{1c}(w) > 0$ by taking $\eta \searrow 0$, i.e. $w \in \operatorname{supp}\rho_{1c}$. In sum, we get

$$\operatorname{supp}\rho_{1c} = \overline{\{w > 0 : 0 \notin f(J(w))\}}. \tag{2.313}$$

From Lemma 2.7.3, we see that there exists an absolute constant $C_1 > 0$ such that if $w \geq C_1\tau^{-1}$, then $h_1(\omega) \leq C_0(\tau^{-1}|w|^{-1/2} + |z|) - \sqrt{w} < 0$. Hence fix $w \geq C_1\tau^{-1}$, we have $0 \in f(J_{2n}(w))$ and $w \notin \operatorname{supp}\rho_{1c}$ (see the upper graphs in Fig. 6 and Fig. 7). This shows that ρ_{1c} is compactly supported in $[0, C_1\tau^{-1}]$. Now we decrease w so that $w < s_1 + |z|^2 + 1$. Then using (2.295), we have

$$h_1(w) > z_1 + w^{-1/2} - \sqrt{w} > \frac{s_1 + |z|^2 + 1 - w}{\sqrt{w}} > 0.$$

By continuity, there must be some $0 < w < C\tau^{-1}$ such that $0 \notin f(J(w))$. Thus $\operatorname{supp}\rho_{1c} \neq \emptyset$. By (2.313), it is not hard to see that $\operatorname{supp}\rho_{1c}$ is a disjoint union of (countably many) closed intervals,

$$\operatorname{supp}\rho_{1c} = \bigcup_k [e_{2k}, e_{2k-1}], \tag{2.314}$$

where $C_1\tau^{-1} \geq e_1 \geq e_2 \geq \ldots$. Furthermore, for e_i to be a boundary point, we must have that 0 is a critical value of $f(\sqrt{e_i}, m)$, i.e. there is a unique critical point $m = m_c(e_i)$

such that

$$f(\sqrt{e_i}, m_c(e_i)) = 0, \quad \partial_m f(\sqrt{e_i}, m_c(e_i)) = 0. \tag{2.315}$$

Notice the two equations in (2.315) are equivalent to two polynomial equations in $(\sqrt{w}, m)$ with order $3n+1$ and $6n$, respectively. By Bézout's theorem, there are at most finitely many solutions to the equations (2.315). Hence there are finitely many e_i's, call them $e_1 \geq \ldots \geq e_{2L}$, where $L \equiv L(n) \in \mathbb{N}$. The statement about e_{2L} follows from Lemma 2.7.4 below. This concludes Lemma 2.2.3. $\square$

Lemma 2.7.4. *If $1+\tau \leq |z|^2 \leq 1+\tau^{-1}$, there is a constant $\epsilon(\tau) > 0$ so that $e_{2L} \geq \epsilon(\tau)$. If $|z|^2 \leq 1-\tau$, $e_{2L} = 0$ and $\rho_{1c}(x) \sim x^{-1/2}$ when $x \searrow 0$.*

By this lemma, the behavior of the leftmost edge e_{2L} changes essentially when z crosses the unit circle. From the following proof, we will see that the singularity happens at $|z|^2 = N^{-1}\sum_{i=1}^n l_i s_i$. Thus the fact that the singular circle has radius 1 is due to our normalization (2.6) for T.

Proof of Lemma 2.7.4. We first study the equation (2.15) when $w \searrow 0$ in the case $1+\tau \leq |z|^2 \leq 1+\tau^{-1}$. We calculate the derivative of f as

$$\begin{aligned}\partial_m f(\sqrt{w}, m) = 1 &+ \frac{1}{N}\sum_{i=1}^n l_i s_i \frac{m^2 - |z|^2}{\sqrt{w}m^3 - (s_i + |z|^2)m^2 - \sqrt{w}|z|^2 m + |z|^4} \\ &- \frac{m}{N}\sum_{i=1}^n l_i s_i \frac{\sqrt{w}\left(m^2 - |z|^2\right)^2 + 2s_i|z|^2 m}{\left[\sqrt{w}m^3 - (s_i + |z|^2)m^2 - \sqrt{w}|z|^2 m + |z|^4\right]^2}.\end{aligned} \tag{2.316}$$

Recall the definition of J_i in the proof of Lemma 2.7.3. It is easy to see that $J_0 \neq \emptyset$ for all $w > 0$, since $\partial_m f(\sqrt{w}, 0) = 1 - |z|^{-2} > 0$ (see the lower graph in Fig. 6). Call the end points of J_0 as $z_k(w) > 0$ and $z_{k+1}(w) < 0$. By the definition of I_0, we have $z_k < b_1 < |z|$. Suppose $z_k = o(|z|)$ as $w \to 0$, then (2.316) gives that $0 = 1 - |z|^{-2} + o(1)$, which gives a contradiction. Thus $z_k \sim |z|$ as $w \to 0$. Now using $\partial_m f(\sqrt{w}, z_k) = 0$, we can estimate that

$$\begin{aligned} f(\sqrt{w}, z_k) &= -\sqrt{w} + \frac{z_k^2}{N}\sum_{i=1}^n l_i s_i \frac{\sqrt{w}\left(z_k^2 - |z|^2\right)^2 + 2s_i|z|^2 z_k}{\left[\sqrt{w}z_k^3 - (s_i + |z|^2)z_k^2 - \sqrt{w}|z|^2 z_k + |z|^4\right]^2} \\ &\geq -\sqrt{w} + \frac{1}{N}\sum_{i=1}^n l_i s_i \frac{2s_i|z|^2 z_k^3}{|z|^8} \geq c - \sqrt{w} \end{aligned} \tag{2.317}$$

for some constant $c > 0$ independent of w, where in the second step we used that

$$\sqrt{w}z_k^3 - (s_i + |z|^2)z_k^2 - \sqrt{w}|z|^2 z_k + |z|^4 > 0, \text{ and } \sqrt{w}z_k^3 - (s_i + |z|^2)z_k^2 - \sqrt{w}|z|^2 z_k < 0$$

which come from the fact that $0 < z_k < b_i < |z|$ for all $1 \leq i \leq n$. By (2.317), we can find ϵ small enough such that $f(\sqrt{w}, z_k) > 0$ for all $0 < w \leq \epsilon$. In this case, $0 \in f(J_0(w))$ and hence $w \notin \operatorname{supp} \rho_{1c}$. In fact, it is not hard to see that there is a solution $m_0 = \sqrt{w}|z|^2/(|z|^2 - 1) + o(\sqrt{w}) \in I_0$ such that $f(\sqrt{w}, m_0) = 0$ and $\partial_m f(\sqrt{w}, m_0) > 0$. This proves the first statement of Lemma 2.7.4.

Now we study equation (2.15) when $|z|^2 \leq 1 - \tau$ and $w \to 0$. For later purpose, we allow w to be complex and prove a more general result than what we need for this lemma. Let $w = 0$ in the equation (2.15), we get $m = 0$ or

$$0 = 1 + \frac{1}{N}\sum_{i=1}^n l_i s_i \frac{m^2 - |z|^2}{-(s_i + |z|^2)m^2 + |z|^4}. \tag{2.318}$$

We define

$$g(x) := 1 + \frac{1}{N}\sum_{i=1}^n l_i s_i \frac{x - |z|^2}{-(s_i + |z|^2)x + |z|^4} = \frac{|z|^2}{N}\sum_{i=1}^n l_i \frac{-x + |z|^2 - s_i}{-(s_i + |z|^2)x + |z|^4}. \tag{2.319}$$

It is easy to see that g is smooth and decreasing on the intervals defined through

$$K_1 := \left(-\infty, \frac{|z|^4}{s_1 + |z|^2}\right), \ K_i := \left(\frac{|z|^4}{s_{i-1} + |z|^2}, \frac{|z|^4}{s_i + |z|^2}\right) \ (i = 2, \ldots, n), \ K_{n+1} := \left(\frac{|z|^4}{s_n + |z|^2}, \infty\right).$$

By the boundary values of g on these intervals, we see that $g(x)$ has exactly one zero on intervals K_i for $i = 1, \dots, n$, and has no zero on K_{n+1}. Since $g(x) = 0$ is equivalent to a polynomial equation of order n, it has at most n solutions. We conclude that all of its solutions are real. Obviously, the zeros on the intervals K_i are positive for $i = 2, \dots, n$. Now we study the zero on K_1. Observe that $g(0) = 1 - |z|^{-2} < 0$ (as $|z|^2 \le 1 - \tau$), hence the zero on K_1 is negative, call it $-t$. Moreover, it is easy to verify that $g(-\tau^{-1}) > 0$ using (2.319), so $t < \tau^{-1}$. If $|z|^2 \ge \tau/2$, then by the concavity of g on the K_1, we get

$$t \ge \frac{g(0)}{g'(0)} \ge \frac{|z|^4(1 - |z|^2)}{s_1} \ge \frac{\tau^4}{4}. \tag{2.320}$$

In the case $|z|^2 \le \tau/2$, we have $|z|^2 - s_n \le -\tau/2$ and $g(|z|^2 - s_n) \le 0$ by (2.319). Hence we have

$$-t \le |z|^2 - s_n \le -\tau/2. \tag{2.321}$$

Combining (2.320) and (2.321), we get that $c\tau^4 \le t \le \tau^{-1}$ for some constant $c > 0$.

Now we return to the self-consistent equation (2.15). The previous discussion shows that

$$f(0, i\sqrt{t}) = 0, \quad \text{with } t \ge c\tau^4.$$

It is easy to see that there exist constants $c_1, \tau' > 0$ such that

$$\left| -(s_i + |z|^2)m^2 + |z|^4 + \sqrt{w}\left(m^3 - |z|^2 m\right) \right| \ge c_1 \text{ for } |m - i\sqrt{t}| \le \tau'. \tag{2.322}$$

First we consider the case $|z| \ge \epsilon > 0$. Expanding $f(\sqrt{w}, m)$ around $(0, i\sqrt{t})$ and using (2.322), we get

$$0 = \partial_{\sqrt{w}} f(0, i\sqrt{t})\sqrt{w} + \partial_m f(0, i\sqrt{t})(m - i\sqrt{t}) + o(\sqrt{w}) + o(m - i\sqrt{t}). \tag{2.323}$$

By (2.316), the partial derivative

$$\partial_{\sqrt{w}} f(\sqrt{w}, m) = -1 - \frac{m^2}{N}\sum_{i=1}^{n} l_i s_i \frac{\left(m^2 - |z|^2\right)^2}{\left[-(s_i + |z|^2)m^2 + |z|^4 + \sqrt{w}\left(m^3 - |z|^2 m\right)\right]^2}, \tag{2.324}$$

and (2.322), we obtain that $\left|\partial_{\sqrt{w}} f(0, i\sqrt{t})\right| \le C$ and

$$\partial_m f(0, i\sqrt{t}) = \frac{t}{N}\sum_{i=1}^{n} l_i s_i \frac{2s_i|z|^2}{\left[(s_i+|z|^2)t+|z|^4\right]^2} \ge c_2 \tag{2.325}$$

for some constant $c_2 > 0$. Using (2.325), we get from (2.323) that

$$m - i\sqrt{t} = O(\sqrt{w}), \quad \text{if } |z| \ge \epsilon. \tag{2.326}$$

Then assume that $|z|^2 < \epsilon$ for sufficiently small ϵ. From $g(-t) = 0$ and (2.319), we get that

$$\frac{1}{N}\sum_{i=1}^{n} l_i \frac{t+|z|^2-s_i}{(s_i+|z|^2)t+|z|^4} = 0. \tag{2.327}$$

From the leading order term, we get $t^{-1} = t_0^{-1} + O(|z|^2)$, where $t_0 := \left(N^{-1}\sum_i l_i/s_i\right)^{-1}$. Expanding (2.327) up to the first order of $|z|^2$, we get

$$t = t_0 + \left(\frac{t_0^2}{N}\sum_i \frac{l_i}{s_i^2} - 2\right)|z|^2 + O(|z|^4). \tag{2.328}$$

Now we write equation (2.15) as

$$F(\sqrt{w}, m) = 0, \tag{2.329}$$

where $F(\sqrt{w}, m) := f(\sqrt{w}, m)/m$. Expanding F around $(0, i\sqrt{t})$ and using (2.322), we get

$$\begin{aligned} 0 =& \partial_{\sqrt{w}} F(0, i\sqrt{t})\sqrt{w} + \partial_m F(0, i\sqrt{t})(m-i\sqrt{t}) + \partial_m\partial_{\sqrt{w}} F(0, i\sqrt{t})(m-i\sqrt{t})\sqrt{w} \\ &+ \frac{1}{2}\partial_{\sqrt{w}}^2 F(0, i\sqrt{t})w + \frac{1}{2}\partial_m^2 F(0, i\sqrt{t})(m-i\sqrt{t})^2 + o(w, |m-i\sqrt{t}|^2, |m-i\sqrt{t}|\sqrt{w}). \end{aligned} \tag{2.330}$$

We can calculate that (the partial derivatives of F can be obtained using (2.316) and (2.324))

$$\partial_m F(\sqrt{w}, i\sqrt{t}) = -\frac{2i|z|^2 + 2\sqrt{wt_0}}{t_0^{3/2}} + o(|z|^2, \sqrt{w}), \tag{2.331}$$

$$\partial_{\sqrt{w}}F(\sqrt{w},i\sqrt{t})=\left(i|z|^2+2\sqrt{wt_0}\right)\frac{\sqrt{t_0}}{N}\sum_{j=1}^{n}\frac{l_j}{s_j^2}+o(|z|^2,\sqrt{w}). \tag{2.332}$$

From (2.331) and (2.332), we get that

$$\partial_m F(0,i\sqrt{t})=-\frac{2i|z|^2}{t_0^{3/2}}+o(|z|^2),\ \partial_{\sqrt{w}}F(0,i\sqrt{t})=\frac{i|z|^2\sqrt{t_0}}{N}\sum_{j=1}^{n}\frac{l_j}{s_j^2}+o(|z|^2),$$

$$\partial_m\partial_{\sqrt{w}}F(0,i\sqrt{t})=-\frac{2}{t_0}+O(|z|^2),\ \partial^2_{\sqrt{w}}F(0,i\sqrt{t})=\frac{2t_0}{N}\sum_{j=1}^{n}\frac{l_j}{s_j^2}+O(|z|^2),\ \partial^2_m F(0,i\sqrt{t})=O(|z|^2).$$

Plugging the above results into (2.330), we get that

$$0=\left[\frac{i|z|^2\sqrt{t_0}+\sqrt{w}t_0}{N}\sum_{j=1}^{n}\frac{l_j}{s_j^2}+o(|z|^2)\right]\sqrt{w}+\left[-2\frac{i|z|^2+\sqrt{wt_0}}{t_0^{3/2}}+o(|z|^2)\right](m-i\sqrt{t})$$
$$+o(w,|m-i\sqrt{t}|^2,|m-i\sqrt{t}|\sqrt{w}). \tag{2.333}$$

Observing that $\left|i|z|^2\sqrt{t_0}+\sqrt{w}t_0\right|\sim|z|^2+\sqrt{|w|}$, we get

$$m-i\sqrt{t}=\left[\frac{t_0^2}{2N}\sum_{j=1}^{n}\frac{l_j}{s_j^2}+O(|w|^{1/2}+|z|^2)\right]\sqrt{w},\quad \text{if } |z|<\epsilon. \tag{2.334}$$

Combing (2.326) and (2.334), we get that if $|z|^2<1-\tau$, $m=i\sqrt{t}+O(\sqrt{w})$ when $w\to 0$. In particular, this shows that $|m|\approx \operatorname{Im} m\sim 1$ when $w\to 0$. Finally, we conclude the proof of Lemma 2.7.4 by using that $m_{1c}(w)=m_c(w)w^{-1/2}-1$. □

To prove Proposition 2.2.14, we need the following lemma, which is a consequence of the edge regularity conditions (2.19) and (2.20).

Lemma 2.7.5. *Suppose $e_k\neq 0$ is a regular edge. Then $|m_{1c}(w)-m_{1c}(e_k)|\sim|w-e_k|^{1/2}$ as $w\to e_k$ and $\min_{l\neq k}|e_l-e_k|\geq\delta$ for some constant $\delta>0$.*

Proof. Denote $m_k:=m_c(e_k)$ and let $w\to e_k$. Note that by Lemma 2.2.3 and Lemma 2.7.4, if $e_k\neq 0$, we have

$$\epsilon'\leq e_k\leq C\tau^{-1}, \tag{2.335}$$

for some constant $\epsilon' > 0$. Then we expand f around $(\sqrt{e_k}, m_k)$ to get that

$$\begin{aligned} 0 =& \partial_{\sqrt{w}} f(\sqrt{e_k}, m_k)(\sqrt{w} - \sqrt{e_k}) + \frac{1}{2}\partial_m^2 f(\sqrt{e_k}, m_k)(m_c(w) - m_k)^2 \\ &+ O\left[|\sqrt{w} - \sqrt{e_k}|^2 + |m_c(w) - m_k|^3 + |\sqrt{w} - \sqrt{e_k}||m_c(w) - m_k|\right], \end{aligned} \tag{2.336}$$

where by (2.324),

$$\partial_{\sqrt{w}} f(\sqrt{e_k}, m_k) = -1 - \frac{m_k^2}{N}\sum_{i=1}^{n} l_i s_i \frac{(m_k^2 - |z|^2)^2}{e_k(m_k - a_i)^2(m_k - b_i)^2(m_k + c_i)^2}, \tag{2.337}$$

and by (2.294),

$$\partial_m^2 f(\sqrt{e_k}, m_k) = \frac{2}{N}\sum_{i=1}^{n} l_i s_i \left[\frac{A_i}{(m_k - a_i)^3} + \frac{B_i}{(m_k - b_i)^3} + \frac{C_i}{(m_k + c_i)^3}\right]. \tag{2.338}$$

Applying (2.295)-(2.298), (2.335) and the conditions (2.19)-(2.20) to (2.337) and (2.338), we get that

$$1 \le \left|\partial_{\sqrt{w}} f(\sqrt{e_k}, m_k)\right| \le C_1, \quad \epsilon \le \left|\partial_m^2 f(\sqrt{e_k}, m_k)\right| \le C_2 \tag{2.339}$$

for some $C_1, C_2 > 0$. Similarly, if $|w - e_k| \le \tau'$ and $|m_c(w) - m_k| \le \tau'$ for some sufficiently small τ', using the condition (2.19) we can get that

$$\max\left\{\left|\partial_m^3 f(\sqrt{w}, m_c(w))\right|, \left|\partial_{\sqrt{w}}^2 f(\sqrt{w}, m_c(w))\right|, \left|\partial_m \partial_{\sqrt{w}} f(\sqrt{w}, m_c(w))\right|\right\} \le C_3. \tag{2.340}$$

Plugging them into equation (2.336), for $|w - e_k| \le \tau'$ and $|m_c(w) - m_k| \le \tau'$, we get $|m_c(w) - m_k| \sim |\sqrt{w} - \sqrt{e_k}|^{1/2}$ and

$$-\partial_{\sqrt{w}} f(\sqrt{e_k}, m_k)(\sqrt{w} - \sqrt{e_k}) + O(|\sqrt{w} - \sqrt{e_k}|^{3/2}) = \frac{1}{2}\partial_m^2 f(\sqrt{e_k}, m_k)(m_c(w) - m_k)^2. \tag{2.341}$$

By (2.335), we immediately get that $|\sqrt{w} - \sqrt{e_k}| \sim |w - e_k|$ and $|m_c(w) - m_k| \sim |m_{1c}(w) - m_{1c}(e_k)|$, which proves the first part of the lemma. By (2.341), if w is real and $|w - e_k| \le \tau'$, we have that

$$m_c(w) - m_k = \left[\frac{-2\partial_{\sqrt{w}} f(\sqrt{e_k}, m_k)}{\partial_m^2 f(\sqrt{e_k}, m_k)} + O(|\sqrt{w} - \sqrt{e_k}|^{1/2})\right]^{1/2} \left(\sqrt{w} - \sqrt{e_k}\right)^{1/2}. \tag{2.342}$$

Thus in a sufficiently small interval $U = [e_k - \delta, e_k + \delta]$, $m_c(w)$ has positive imaginary part for w on one side of e_k, while $m_c(w)$ is real for w on the other side. Hence U does not contain another edge. This shows that $\min_{l \neq k} |e_l - e_k| \geq \delta$. □

Proof of Proposition 2.2.14. The properties of ρ_{1c} have been proved in Lemmas 2.2.3, 2.7.4 and 2.7.5, and included in Definition 2.2.4. Since $\operatorname{supp} \rho_{2c} = \operatorname{supp} \rho_{1c}$ by the discussion after Lemma 2.2.2, we immediately get property (i) for ρ_{2c}. The conclusion ρ_{2c} being a probability measure is due to the definition of m_2 in (2.35) and the fact that m_{2c} is the almost sure limit of m_2.

The properties (ii) and (iv) for ρ_{2c} can be easily obtained by plugging m_{1c} into (2.10). To prove the property (iii) for ρ_{2c}, we need to know the behavior of $\operatorname{Im} m_{2c}(w)$ when $w \to e_j$ along the real line. By (2.10), it suffices to prove that if $|x - e_j| \leq \tau'$ for some small enough $\tau' > 0$, then

$$\left|-w(1 + m_{1c})^2 + |z|^2\right| = \left|m_c^2 - |z|^2\right| \geq \epsilon$$

for some constant $\epsilon > 0$. Suppose that $|m_c^2(w) - |z|^2| = o(1)$. Then plugging m_c into $\partial_m f(\sqrt{w}, m_c)$ in (2.316), and using condition (2.19) and Lemma 2.7.5, we get that

$$\partial_m f(\sqrt{w}, m_c(w)) = -1 + O(|m_c^2 - |z|^2|). \tag{2.343}$$

Again using condition (2.19) and Lemma 2.7.5, we can bound $\partial_{\sqrt{w}} \partial_m f(\sqrt{w}, m_c(w))$ and $\partial_m^2 f(\sqrt{w}, m_c(w))$ for w near e_j. Thus we shall have that

$$0 = \partial_m f(\sqrt{e_j}, m_c(e_j)) = \partial_m f(\sqrt{w}, m_c(w)) + O(|w - e_j|^{1/2}) = -1 + O(|m_c^2 - |z|^2| + |w - e_j|^{1/2}). \tag{2.344}$$

This gives a contradiction. Thus we must have a lower bound for $|m_c^2 - |z|^2|$. □

Remark: Here we add a small remark on Example 2.2.8. Given the assumptions in Example 2.2.8, it is easy to see that f can only take critical values on intervals I_{-n}, I_0, I_n and I_{2n}, since $\max\{|a_i - a_{i-1}|, |b_i - b_{i-1}|, |c_i - c_{i-1}|\} \to 0$ in this case. Thus the number of connected components of $\operatorname{supp} \rho_{1c}$ is independent of n, and all the edges and the bulk components are regular as in Example 2.2.7.

2.7.2 Proof of Lemmas 2.3.7 and 2.3.8

We first prove Lemma 2.3.7. We consider the five cases separately.

Case 1: For $w = E + i\eta \in \mathbf{D}_k^b(\zeta, \tau', N)$, we have

$$m_{1c}(w) = \int_{\mathbb{R}} \frac{\rho_{1c}(x)}{x - (E + i\eta)} dx, \quad \operatorname{Im} m_{1c}(w) = \int_{\mathbb{R}} \frac{\rho_{1c}(x, z)\eta}{(x - E)^2 + \eta^2} dx. \tag{2.345}$$

By the regularity condition of Definition 2.2.4 (ii), we get immediately $\operatorname{Im} m_{1c} \sim 1$. Since $\operatorname{Im} m_{1c} \le |1 + m_{1c}| \le C$ by Proposition 2.2.15, we get $|1 + m_{1c}| \sim 1$. Notice wm_{1c} can be expressed as

$$wm_{1c}(w) = \int_{\mathbb{R}} \frac{w\rho_{1c}(x, z)}{x - w} dx = -\int_{\mathbb{R}} \rho_{1c}(x, z) dx + \int_{\mathbb{R}} \frac{x\rho_c(x, z)}{x - w} dx.$$

By the same argument as above and using the fact that $x \ge \tau'$ for $x \in [e_{2k} + \tau', e_{2k-1} - \tau']$, we get

$$\operatorname{Im}(wm_{1c}) = \operatorname{Im} \int_{\mathbb{R}} \frac{x\rho_{1c}(x, z)}{x - w} dx \sim 1.$$

Since the imaginary parts of $-w$ and $|z|^2/(1 + m_{1c})$ are both negative, we get

$$\operatorname{Im} \left[-w(1 + m_{1c}) + \frac{|z|^2}{1 + m_{1c}} \right] \le -\operatorname{Im}(wm_{1c}). \tag{2.346}$$

Using the bounds for m_{1c} and $\operatorname{Im} m_{1c}$ proved above, it is easy to see that

$$\left| -w(1 + m_{1c}) + \frac{|z|^2}{1 + m_{1c}} \right| = O(1). \tag{2.347}$$

Equations (2.346) and (2.347) together give that $\operatorname{Im} m_{2c} \sim 1$ and $|m_{2c}| \sim 1$ by (2.10). Similarly, we can also prove that

$$wm_{2c} = \left[-(1+m_{1c}) + \frac{|z|^2}{w(1+m_{1c})}\right]^{-1} \in \mathbb{C}_+$$

and $\operatorname{Im}(wm_{2c}) \sim 1$. Then (2.106) follows from the bound

$$\operatorname{Im}\left(w + s_i w m_{2c} - \frac{|z|^2}{1+m_{1c}}\right) \geq s_i \operatorname{Im}(wm_{2c}).$$

Case 2: For $w = E + i\eta \in \mathbf{D}^o(\zeta, \tau', N)$, using (2.345) and $\operatorname{dist}(E, \operatorname{supp}\rho_{1,2c}) \geq \tau'$, we immediately get $\operatorname{Im} m_{1,2c} \sim \eta$. Now we prove the other estimates.

We first prove (2.106). If $\eta \sim 1$, the proof is the same as in Case 1. Hence we assume $\eta \leq c'$, where $c' \equiv c'(\tau, \tau') > 0$ is sufficiently small. Recall the definitions of $\mathbf{D}$ and $\mathbf{D}^o$ in (3.31) and (2.43), we always have $E \sim 1$ in this case.

We shall prove that

$$\min_i \{|m_c(w) - a_i(w)|, |m_c(w) - b_i(w)|, |m_c(w) + c_i(w)|\} \geq \epsilon', \tag{2.348}$$

for some constant ϵ'. This leads immediately to (2.106) since

$$\left| w\left(1 + s_i \frac{1+m_{1c}}{-w(1+m_{1c})^2 + |z|^2}\right)(1+m_{1c}) - |z|^2 \right| = \left| \frac{\sqrt{w}(m_c - a_i)(m_c - b_i)(m_c + c_i)}{-m_c^2 + |z|^2} \right|. \tag{2.349}$$

For $p_i = \sqrt{E} m^3 - (s_i + |z|^2) m^2 - \sqrt{E}|z|^2 m + |z|^4$, it is not hard to prove that the roots $a_i(E)$, $b_i(E)$ and $-c_i(E)$ decrease as E increase. Since $E \notin \operatorname{supp}\rho_{1c}$, we have $m_{1c}(E) \in \mathbb{R}$ and

$$\frac{dm_{1c}(E)}{dE} = \int_{\mathbb{R}} \frac{\rho_{1c}(x,z)}{(x-E)^2} dx \geq 0.$$

So $m_{1c}(E)$ (and hence $m_c(E)$) increases as E increases. Suppose e_k is the smallest edge that is bigger than E, then for $a_i(E)$ bigger than $m_c(E)$, we have that

$$a_i(E) - m_c(E) \geq a_i(e_k) - m_c(e_k) + \epsilon'(\tau') \geq \epsilon'(\tau'), \tag{2.350}$$

by using $|E - e_k| \geq \tau'$ (see (2.43)). On the other hand, If e_{k-1} is the largest edge value that is smaller than E, then for $a_i(E)$ smaller than $m_c(E)$, we have that

$$m_c(E) - a_i(E) \geq m_c(e_{k-1}) - a_i(e_{k-1}) + \epsilon'(\tau') \geq \epsilon'(\tau'). \tag{2.351}$$

Applying the same arguments to $b_i(E)$ and $-c_i(E)$, we get

$$\min_i \{|m_c(E) - a_i(E)|, |m_c(E) - b_i(E)|, |m_c(E) + c_i(E)|\} \geq \epsilon' \tag{2.352}$$

for $E \in (e_{2k+1}, e_{2k})$ for some k. Now we are only left with the case $E < e_{2L}$, the rightmost edge, when $|z|^2 \geq 1 + \tau$. In this case, we have seen that $0 < m_c(E) < b_i(E)$ for all i in the proof of Lemma 2.7.4. Thus we can use (2.350) to get lower bounds for $|m_c(E) - a_i(E)|$ and $|m_c(E) - b_i(E)|$. Since $c_i(E) \sim 1$ in this case (by (2.297) and using $E, |z| \sim 1$), $|m_c(E) + c_i(E)| \geq \epsilon$ is trivial. Again we get the estimate (2.352).

Then we consider $w = E + i\eta$ with $\eta \leq c'$. First, it is easy to check that $a_i(E + i\eta)$, $b_i(E + i\eta)$ and $c_i(E + i\eta)$ are continuous in η. On the other hand for $m_c(E + i\eta)$, we have

$$|\partial_w m_{1c}(w)| = \left| \int_{\mathbb{R}} \frac{\rho_{1c}(x, z)}{(x - w)^2} dx \right| \leq C \tag{2.353}$$

by the condition $\operatorname{dist}(E, \operatorname{supp} \rho_{1c}) \geq \tau'$. Thus we immediately get $|m_c(E+i\eta) - m_c(E)| = O(\eta)$. Hence as long as c' is small enough, (2.348) still holds true, which further gives (2.106).

Now we show that $|1 + m_{1c}| \sim 1$ for $w \in \mathbf{D}^o$ and $\eta \leq c'$. In fact, if $|m_c|$ can be arbitrarily small, then by (2.106) we get that

$$f(\sqrt{w}, m_c) = -\sqrt{w} + O(m_c) \neq 0,$$

which gives a contradiction. Finally we have $|m_{2c}| \sim 1$ for $w \in \mathbf{D}^o$ and $\eta \leq c'$ by Proposition 2.2.15.

Case 3: For a regular edge $e_k \neq 0$, we always have $e_k \geq \epsilon$ for some $\epsilon > 0$ by Lemma 2.7.4. Thus we always have $|w| \sim 1$ for $w = E + i\eta \in \mathbf{D}_k^e(\zeta, \tau', N)$ as long as τ' is sufficiently small. If $\eta \sim 1$, then $\sqrt{\kappa+\eta} \sim \eta/\sqrt{\kappa+\eta} \sim 1$ and the proof is the same as in Case 1. Now we pick τ' small and consider the case $\eta \leq \tau'$. By the regularity assumption (2.19) and Lemma 2.7.5, we have

$$\min_{1\leq i\leq n}\{|m_c(w) - a_i(w)|, |m_c(w) - b_i(w)|, |m_c(w) + c_i(w)|\} \geq \epsilon/2 \tag{2.354}$$

uniformly in $w \in \{w \in \mathbf{D}_k^e(\zeta, \tau', N) : \kappa(w) + \eta(w) \leq 2\tau'\}$, provided τ' is sufficiently small. The above bound implies (2.106). If $m_c(w) \to 0$, then using (2.106) we get from $f(\sqrt{w}, m_c) = 0$ that $-\sqrt{w} + O(m_c) = 0$, which gives a contradiction. Thus we must have $|1 + m_{1c}| \sim |m_c| \sim 1$. To show $|m_{2c}| \sim 1$, we can use Proposition 2.2.15.

We still need to prove the estimates for $\operatorname{Im} m_{1,2c}$ when $\eta \leq \tau'$. Recall the expansion (2.341) around e_k and equation (2.342), where both $\partial_{\sqrt{w}} f(\sqrt{e_k}, m_k)$ and $\partial_m^2 f(\sqrt{e_k}, m_k)$ are real (as e_k and m_k are real). Suppose k is odd, then $\operatorname{Im} m_c(E) = 0$ for $E \searrow e_k$ (i.e. $E \notin \operatorname{supp}\rho_c$) and $\operatorname{Im} m_c(E) > 0$ for $E \nearrow e_k$ (i.e. $E \in \operatorname{supp}\rho_c$). Thus (2.342) gives

$$m_c(w) - m_k = C_k(w)(w - e_k)^{1/2} + D_k(w),$$

with $C_k > 0$, $C_k \sim 1$, $|D_k| = O(|w - e_k|)$ and $\operatorname{Im} D_k = O(\eta)$. Then for $E \geq e_k$, we have

$$\operatorname{Im} m_c(E + i\eta) \sim \operatorname{Im}(\kappa + i\eta)^{1/2} + O(\eta) \sim \frac{\eta}{\sqrt{\kappa+\eta}},$$

and for $E \leq e_k$, we have

$$\operatorname{Im} m_c(E + i\eta) \sim \operatorname{Im}(-\kappa + i\eta)^{1/2} + O(\eta) \sim \sqrt{\kappa+\eta}.$$

If k is even, the proof is the same except that in this case, we have

$$m_c(w) - m_k = C_k(w)(e_k - w)^{1/2} + D_k(w).$$

For $m_{1c}(w)$ and $m_{2c}(w)$, we get the conclusion by noticing $w \approx e_k$ and

$$\operatorname{Im} m_{1c} = \operatorname{Im}\left(w^{-1/2} m_c\right) \sim \operatorname{Im} m_c(w), \quad \operatorname{Im} m_{2c} = \operatorname{Im}\left[\frac{m_c}{\sqrt{w}(-m_c^2 + |z|^2)}\right] \sim \operatorname{Im} m_c(w),$$

where we used that $|m_c^2 - |z|^2| \sim 1$ as observed in the proof of Proposition 2.2.14 in Section 2.7.1.

Case 4: Again if $\eta \sim 1$, the proof is the same as in Case 1. If $|w| \le 2\tau'$ for small enough τ', in the proof of Lemma 2.7.4, we have seen that $m_c = i\sqrt{t} + O(\sqrt{w})$, which gives the first equation in (2.103). Plugging it into (2.10), we get the second equation in (2.103). Taking the imaginary part, we obtain (2.104). Finally using (2.103), we can verify (2.106) easily.

Case 5: For $w = E + i\eta \in \mathbf{D}_L(\zeta, N)$, the bounds for $m_{1,2}$ and $\operatorname{Im} m_{1,2}$ in (2.105) follows from (2.345) directly.

Proof of Lemma 2.3.8. The estimates (2.108) and (2.109) follow immediately from (2.33), (2.106) and (2.107). For (2.110), we can write

$$\Pi_{\mathbf{vv}} = \left\langle \mathbf{v}, \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix} \Pi_d \begin{pmatrix} U^\dagger & 0 \\ 0 & U^\dagger \end{pmatrix} \mathbf{v} \right\rangle = (\Pi_d)_{\mathbf{uu}} = \sum_{i=1}^N \left\langle u_{[i]}, \pi_{[i]c} u_{[i]} \right\rangle,$$

where

$$\mathbf{u} := \begin{pmatrix} U^\dagger & 0 \\ 0 & U^\dagger \end{pmatrix} \mathbf{v}, \quad u_{[i]} := \begin{pmatrix} u_i \\ u_{\bar{i}} \end{pmatrix}.$$

To control $\operatorname{Im} \Pi_{\mathbf{vv}}$, it is enough to bound $\left\langle u_{[i]}, \pi_{[i]c} u_{[i]} \right\rangle$ for each i.

We first consider Cases 1-4 of Lemma 2.3.7. By the definition of $\pi_{[i]c}$ in (2.33), we get

$$\operatorname{Im} \pi_{ii,c} = \operatorname{Im}\left[-w(1 + |d_i|^2 m_{2c}) + \frac{|z|^2}{1 + m_{1c}}\right]^{-1} \le \frac{C}{|w|} \operatorname{Im}\left[w(1 + |d_i|^2 m_{2c}) - \frac{|z|^2}{1 + m_{1c}}\right]$$

$$= \frac{C}{|w|}\left[(1+|d_i|^2\mathrm{Re}\, m_{2c})\mathrm{Im}\, w + |d_i|^2(\mathrm{Re}\, w)\mathrm{Im}\, m_{2c} + \frac{|z|^2}{|1+m_{1c}|^2}\mathrm{Im}\, m_{1c}\right],$$

where in the second step we used (2.106) and $|1+m_{1c}| \sim |w|^{-1/2}$. In the first three cases of Lemma 2.3.7, we have $|w| \sim 1$ and $\mathrm{Im}\, w = O(\mathrm{Im}\, m_{1c})$, which give that $\mathrm{Im}\, \pi_{ii,c} \le C\mathrm{Im}(m_{1c}+m_{2c})$. In case 4 of Lemma 2.3.7, we use $|\mathrm{Im}\, w|+|\mathrm{Re}\, w|+|1+m_{1c}|^{-2} = O(|w|)$ and $\mathrm{Im}\, m_{1,2c} \sim |w|^{-1/2}$ to get that $\mathrm{Im}\, \pi_{ii,c} \le C\mathrm{Im}(m_{1c}+m_{2c})$. Similarly, we can get the bound $\mathrm{Im}\, \pi_{\bar{i}\bar{i},c} \le C\mathrm{Im}(m_{1c}+m_{2c})$. Finally we can estimate the following term using similar methods,

$$\begin{aligned}\mathrm{Im}\left(\bar{u}_{\bar{i}} u_i \pi_{\bar{i}i,c} + \bar{u}_i u_{\bar{i}} \pi_{i\bar{i},c}\right) &= 2\mathrm{Re}\,(\bar{u}_i u_{\bar{i}} z)\,\mathrm{Im}\left\{w^{-1/2}\left[w(1+|d_i|^2 m_{2c})(1+m_{1c}) - |z|^2\right]^{-1}\right\} \\ &\le C\mathrm{Re}\,(\bar{u}_i u_{\bar{i}} z)\,\mathrm{Im}(m_{1c}+m_{2c}) \le C\left(|u_i|^2+|u_{\bar{i}}|^2\right)\mathrm{Im}(m_{1c}+m_{2c}).\end{aligned}$$

Combining the above estimates we get $\mathrm{Im}\left\langle u_{[i]}, \pi_{[i]c} u_{[i]}\right\rangle \le C|u_{[i]}|^2\mathrm{Im}(m_{1c}+m_{2c})$, which implies (2.110). For the Case 5 of Lemma 2.3.7, we use (2.105) and (2.109) to get

$$\mathrm{Im}\left\langle u_{[i]}, \pi_{[i]c} u_{[i]}\right\rangle \le |u_{[i]}|^2\|\pi_{[i]c}\| \le C|u_{[i]}|^2\mathrm{Im}(m_{1c}+m_{2c}).$$

□

2.7.3 Proof of Lemma 2.3.10 and Lemma 2.2.2

We first prove Lemma 2.3.10. During the proof, we also use the following equivalent definition of the stability expressed in terms of $m = \sqrt{w}(1+m_1)$, $u = \sqrt{w}(1+u_1)$ and $f(\sqrt{w}, m)$. Suppose the assumptions in Definition 2.3.9 holds. Let $w \in \mathbf{D}$ and suppose that for all $w' \in L(w)$ we have $|f(\sqrt{w}, u)| \le |w|^{1/2}\delta(w)$. Then

$$|u(w) - m_c(w)| \le \frac{C|w|^{1/2}\delta}{\sqrt{\kappa+\eta+\delta}}. \tag{2.355}$$

Case 1: We take over the notations in Definition 2.3.9 and abbreviate $R := f(\sqrt{w}, u)$, so that $|R| \le |w|^{1/2}\delta$. Then we write the equation $f(\sqrt{w}, u) - f(\sqrt{w}, m_c) = R$ as

$$\alpha(u - m_c)^2 + \beta(u - m_c) = R, \tag{2.356}$$

where using (2.294), α and β can be expressed as

$$\alpha := \frac{1}{N}\sum_{i=1}^{n} l_i s_i \left[\frac{A_i}{(u - a_i)(m_c - a_i)^2} + \frac{B_i}{(u - b_i)(m_c - b_i)^2} + \frac{C_i}{(u + c_i)(m_c + c_i)^2}\right], \tag{2.357}$$

and

$$\beta := 1 - \frac{1}{N}\sum_{i=1}^{n} l_i s_i \left[\frac{A_i}{(m_c - a_i)^2} + \frac{B_i}{(m_c - b_i)^2} + \frac{C_i}{(m_c + c_i)^2}\right] = \partial_m f(\sqrt{w}, m_c). \tag{2.358}$$

We shall prove that

$$|\alpha| + |\partial_u \alpha| \le C, \ |\beta| \sim 1, \tag{2.359}$$

for $w \in \mathbf{D}_k^b$ and u satisfying $|u - m_c| \le (\log N)^{-1/3}$. If $|u - m_c| \le (\log N)^{-1/3}$, we also have $\operatorname{Im} u \sim 1$. By (2.106), we have

$$\min_i \{|m_c - a_i|, |m_c - b_i|, |m_c + c_i|\} \ge \epsilon \tag{2.360}$$

for some $\epsilon > 0$. Replacing the m_c in (2.106) with u, we also get that

$$\min_i \{|u - a_i|, |u - b_i|, |u + c_i|\} \ge \epsilon' \tag{2.361}$$

for some $\epsilon' > 0$. Using (2.360) and (2.361), we get immediately that $|\alpha| + |\partial_u \alpha| + |\beta| \le C$. What remains is the proof of the lower bound $|\beta| \ge c$. If $\operatorname{Im} w \ge \epsilon$ for some constant $\epsilon > 0$, the lower bound follows from Lemma 2.7.6 below. If $\operatorname{Im} w \le \epsilon$ for a sufficiently small $\epsilon > 0$, the lower bound follows from Lemma 2.7.7 below. Now given the estimate (2.359), it is easy to prove (2.355) with a fixed point argument. This proves the stability of (2.111).

Lemma 2.7.6. *Suppose that* $\operatorname{Im} w \sim 1$ *and* $|m_c| \sim \operatorname{Im} m_c \sim 1$*. Then* $|\partial_m f(\sqrt{w}, m_c)| \geq c$ *for some constant* $c > 0$*.*

Proof. Using (2.14), $m_c = \sqrt{w}(1 + m_{1c})$ and the conditions $\operatorname{Im} w \sim 1$, $\operatorname{Im} m_c \sim 1$, we can get that

$$\left|\frac{\partial_{\sqrt{w}} f(\sqrt{w}, m_c)}{\partial_m f(\sqrt{w}, m_c)}\right| = \left|\frac{\partial m_c}{\partial \sqrt{w}}\right| \leq C \Rightarrow \left|\partial_{\sqrt{w}} f(\sqrt{w}, m_c)\right| \leq C \left|\partial_m f(\sqrt{w}, m_c)\right|, \tag{2.362}$$

for some constant $C > 0$. Now we assume that $|\partial_m f(\sqrt{w}, m_c)|$ can be arbitrarily small. Then $\left|\partial_{\sqrt{w}} f(\sqrt{w}, m_c)\right|$ can also be arbitrarily small. Denote $a := \partial_m f(\sqrt{w}, m_c)$ and $b := \partial_{\sqrt{w}} f(\sqrt{w}, m_c)$. Using (2.316) and (2.324), we get that

$$a = \frac{\sqrt{w}}{m_c} - \frac{m_c}{N} \sum_{i=1}^{n} l_i s_i \frac{\sqrt{w}\left(m_c^2 - |z|^2\right)^2 + 2s_i |z|^2 m_c}{\left[-(s_i + |z|^2) m_c^2 + |z|^4 + \sqrt{w}\left(m_c^3 - |z|^2 m_c\right)\right]^2} \tag{2.363}$$

and

$$b = -1 - \frac{m_c^2}{N} \sum_{i=1}^{n} l_i s_i \frac{\left(m_c^2 - |z|^2\right)^2}{\left[-(s_i + |z|^2) m_c^2 + |z|^4 + \sqrt{w}\left(m_c^3 - |z|^2 m\right)\right]^2}. \tag{2.364}$$

Using (2.363) and (2.364), we can get that

$$\frac{(\sqrt{w} m_c - |z|^2)|z|^2}{m_c} b - \frac{1}{2}(m_c^2 - |z|^2)(m_c a - \sqrt{w} b) = \frac{(|z|^2 - \sqrt{w} m_c)(m_c^2 + |z|^2)}{m_c}, \tag{2.365}$$

where we used the equation $f(\sqrt{w}, m_c) = 0$ in the derivation. By our assumption, the left-hand side of (2.365) can be arbitrarily small. For the right-hand side of (2.365), we have $|m_c| \sim 1$ and $|\sqrt{w} m_c - |z|^2| \sim 1$ (since $\operatorname{Im}(\sqrt{w} m_c) = \operatorname{Im}(w + w m_{1c}) \sim 1$). Then if $|m_c - i|z|| \geq c'$ for some constant $c' > 0$, we have $|m^2 + |z|^2| \sim 1$, and hence

$$\left|\frac{(\sqrt{w} m_c - |z|^2)|z|^2}{m_c} b - \frac{1}{2}(m_c^2 - |z|^2)(m_c a - \sqrt{w} b)\right| \sim 1,$$

which gives a contradiction. Thus we must have a lower bound $|\partial_m f(\sqrt{w}, m_c)| \geq c$ if $|m - i|z|| \geq c'$.

We still need to deal with the case with $|m_c - i|z|| \leq c'$ for some sufficiently small c'. Notice $|z| \sim 1$ in this case. It is easy to calculate that

$$\frac{\partial f}{\partial \sqrt{w}}(\sqrt{w}, i|z|) = -1 + \frac{|z|^2}{N} \sum_{k=1}^{n} l_k s_k \frac{4|z|^4}{\left[(s_k + |z|^2)|z|^2 + |z|^4 - 2i\sqrt{w}|z|^3\right]^2}. \tag{2.366}$$

Denote $L_k := (s_k + |z|^2)|z|^2 + |z|^4 - 2i\sqrt{w}|z|^3$. Since $i\sqrt{w} = i(x+iy) = ix - y$ for some $x, y > 0$ and $x, y \sim 1$, we have $\operatorname{Re} L_k > 0$, $\operatorname{Im} L_k < 0$ and $|\operatorname{Re} L_k|, |\operatorname{Im} L_k| \sim 1$. In particular, this gives that $\operatorname{Im} L_k^2 < 0$ and $|\operatorname{Im} L_k^2| \sim 1$. Thus each fraction $4|z|^4/L_k^2$ in (2.366) has positive imaginary part of order 1. Therefore

$$\left|\frac{\partial f}{\partial \sqrt{w}}(\sqrt{w}, i|z|)\right| \geq \operatorname{Im}\left[\frac{\partial f}{\partial \sqrt{w}}(\sqrt{w}, i|z|)\right] \sim 1.$$

Then by (2.362), we get that $|\partial_m f(\sqrt{w}, i|z|)| \geq c$ for some $c > 0$. Using (2.106), it is easy to see that

$$\partial_m f(\sqrt{w}, m_c) = \partial_m f(\sqrt{w}, i|z|) + O(|m_c - i|z||).$$

Thus in the case $|m_c - i|z|| \leq c'$, we still have $|\partial_m f(\sqrt{w}, m_c)| \geq c/2$, provided that c' is sufficiently small. □

Lemma 2.7.7. *Suppose that $w \in \mathbf{D}_k^b$ and $\operatorname{Im} w \leq \epsilon$. Then for sufficiently small $\epsilon > 0$, we have $|\partial_m f(\sqrt{w}, m_c)| \sim 1$.*

Proof. By (2.99) and (2.106), we have $\partial_{\sqrt{w}}\partial_m f(w, m_c) = O(1)$ and $\partial_m^2 f(w, m_c) = O(1)$. Denote $w = E + i\eta$. Taking the imaginary part of the following equation

$$0 = f(\sqrt{E}, m_c(E)) = -\sqrt{E} + m_c + E^{-1/2} + \frac{1}{N}\sum_{i=1}^{n} l_i s_i \left(\frac{A_i}{m_c - a_i} + \frac{B_i}{m_c - b_i} + \frac{C_i}{m_c + c_i}\right), \tag{2.367}$$

and noticing that A_i, B_i, C_i and a_i, b_i, c_i are all positive real numbers for real E, we get

$$\frac{1}{N}\sum_{i=1}^{n} l_i s_i \left(\frac{A_i}{|m_c - a_i|^2} + \frac{B_i}{|m_c - b_i|^2} + \frac{C_i}{|m_c + c_i|^2}\right) = 1. \tag{2.368}$$

Using the above equation, we get

$$\partial_m f(\sqrt{E}, m_c(E)) = 1 - \frac{1}{N}\sum_{i=1}^{n} l_i s_i \left[\frac{A_i}{(m_c - a_i)^2} + \frac{B_i}{(m_c - b_i)^2} + \frac{C_i}{(m_c + c_i)^2}\right]$$
$$= \frac{1}{N}\sum_{i=1}^{n} l_i s_i \left[\frac{A_i}{|m_c - a_i|^2} - \frac{A_i}{(m_c - a_i)^2} + \frac{B_i}{|m_c - b_i|^2} - \frac{B_i}{(m_c - b_i)^2} + \frac{C_i}{|m_c + c_i|^2} - \frac{C_i}{(m_c + c_i)^2}\right]. \tag{2.369}$$

We look at, for example, the term

$$\frac{A_i}{|m_c - a_i|^2} - \frac{A_i}{(m_c - a_i)^2} = \frac{A_i}{|m_c - a_i|^2}(1 - e^{-2i\theta_i}),$$

where $m_c - a_i := |m_c - a_i|e^{i\theta_i}$. Using $\operatorname{Im} m_c \sim 1$, it is easy to see that $\operatorname{Re}(1 - e^{-2i\theta_i}) \ge c'$ for some constant $c' > 0$. Applying the same estimates to the B, C terms in (2.369), we get

$$\left|\partial_m f(\sqrt{E}, m_c(E))\right| \ge \operatorname{Re}\left[\partial_m f(\sqrt{E}, m_c(E))\right] \ge c \tag{2.370}$$

for some constant $c > 0$.

Now for $w = E + i\eta$ with $\eta \le \epsilon$, we can expand $\partial_m f(\sqrt{w}, m_c(w))$ around $\partial_m f(\sqrt{E}, m_c(E))$:

$$\partial_m f(\sqrt{w}, m_c(w)) = \partial_m f(E, m_c(E)) + O(\eta),$$

where we used (2.106). Combing with (2.370), we get $|\partial_m f(w, m_c(w))| \sim 1$ for small enough ϵ. □

Case 2: We mimic the argument in the proof of Case 1. We see that it suffices to prove $|\alpha| + |\partial_u \alpha| \le C$ and $|\beta| \sim 1$ for α, β defined in (2.357) and (2.358) and $|u - m_c| \le (\log N)^{-1/3}$. Using (2.106), it is not hard to prove that $|\alpha| + |\partial_u \alpha| + |\beta| \le C$. What remains is the proof of the lower bound $|\beta| \ge c$. For the $\operatorname{Im} w \sim 1$ case, the bound follows from Lemma 2.7.6. We are left with the case where $E = \operatorname{Re} w \sim 1$ and $\eta = \operatorname{Im} w \to 0$. Using (2.14), $m_c = \sqrt{w}(1 + m_{1c})$, $|w| \sim 1$ and $\operatorname{dist}(E, \operatorname{supp} \rho_{1c}) \geq \tau'$, we can get that

$$\left| \frac{\partial_{\sqrt{w}} f(\sqrt{w}, m_c)}{\partial_m f(\sqrt{w}, m_c)} \right| = \left| \frac{\partial m_c}{\partial \sqrt{w}} \right| \leq C$$

for some constant $C > 0$. Thus it suffices to prove that $\left|\partial_{\sqrt{w}} f(\sqrt{w}, m_c)\right|$ has a lower bound. Using (2.324) and noticing that $m_c(E) \in \mathbb{R}$, we get

$$\partial_{\sqrt{w}} f(\sqrt{E}, m_c(E)) = -1 - \frac{m_c^2}{N} \sum_{i=1}^{n} l_i s_i \frac{\left(m_c^2 - |z|^2\right)^2}{\left[-(s_i + |z|^2) m_c^2 + |z|^4 + \sqrt{E}\left(m_c^3 - |z|^2 m_c\right)\right]^2} \leq -1.$$

Expanding $\partial_{\sqrt{w}} f(\sqrt{w}, m_c(w))$ around $\partial_{\sqrt{w}} f(\sqrt{E}, m_c(E))$, using (2.106) and $|m_c(E+i\eta) - m_c(E)| \sim \eta$, we get for η small

$$\left|\partial_{\sqrt{w}} f(\sqrt{w}, m_c)\right| \geq 1 + O(\eta) \geq c.$$

This concludes the proof for *Case 2*.

Case 3: The case $\operatorname{Im} w \geq \tau'$ can be proved with the same method as in the proof of case 1. Hence we only consider the case $|w - e_k| \leq 2\tau'$ in the following. Note that $|w| \sim 1$ in this case. Suppose

$$|w - e_k| \leq 2\tau', \quad |u - m_c| \leq (\log N)^{-1/3}. \tag{2.371}$$

Then we claim that

$$|\alpha| \sim 1, \quad |\beta| \sim \sqrt{\kappa + \eta} \tag{2.372}$$

for small enough τ'. Using (2.371), (2.106), (2.20) and Lemma 2.7.5, we can get that

$$\alpha = \frac{1}{2} \partial_m^2 f(\sqrt{e_k}, m_c(e_k)) + O(|w - e_k|^{1/2} + (\log N)^{-1/3}) \sim 1.$$

To prove the estimate for β, we use (2.18), (2.106) and Lemma 2.7.5 to get that

$$\beta = \int_{e_k}^{w} \frac{d}{dw'} \partial_m f(\sqrt{w'}, m_c(w')) dw'$$

$$= \int_{e_k}^{w} \frac{\partial_{\sqrt{w'}}\partial_m f(\sqrt{w'}, m_c(w'))}{2\sqrt{w'}} dw' + \int_{e_k}^{w} \partial_m^2 f(\sqrt{w'}, m_c(w')) \frac{dm_c(w')}{dw'} dw'$$

$$= \int_{e_k}^{w} \frac{\partial_{\sqrt{w}}\partial_m f(\sqrt{e_k}, m_c(e_k)) + O(|w-e_k|^{1/2})}{2\sqrt{w'}} dw' + \int_{m_c(e_k)}^{m_c(w)} \left[\partial_m^2 f(\sqrt{e_k}, m_c(e_k)) + O(|w-e_k|^{1/2})\right] dm$$

$$= \partial_m^2 f(\sqrt{e_k}, m_k)(m_c(w) - m_c(e_k)) + O(|w-e_k|). \tag{2.373}$$

Thus we conclude for small enough τ' that

$$|\beta| \sim |w-e_k|^{1/2} \sim \sqrt{\kappa+\eta}.$$

With the estimate (2.372), we now proceed as in the proof of [6, Lemma 4.5], by solving the quadratic equation (2.356) for $u - m_c$ explicitly. We select the correct solution by a continuity argument using that (2.355) holds by assumption at $z + iN^{-10}$. The second assumption of (2.371) is obtained by continuity from the estimate on $|u - m_c|$ at the neighboring point $z + iN^{-10}$. We refer to [6, Lemma 4.5] for the full details. This concludes the proof for *Case 3*.

Case 4: The case when $\operatorname{Im} w \geq \tau'$ can be proved using the same method as in the proof of Case 1. Now we are left with the case $|w| \leq 2\tau'$ for some sufficiently small τ'. First we assume $|z| \geq c > 0$ for some small $c > 0$. Then mimicking the argument in the proof of Case 1, we see that it suffices to prove $|\alpha| + |\partial_u \alpha| \leq C$ and $|\beta| \sim 1$ when $|u - m_c| \leq (\log N)^{-1/3}$. Using (2.106), it is not hard to prove that $|\alpha| + |\partial_u \alpha| + |\beta| \leq C$. The lower bound $|\beta| \geq c$ can be obtained easily from (2.325).

Then suppose $|z|^2 < c$, but $|w|^{1/2} + |z|^2 \geq \epsilon$. According to (2.331) and using that $\left|i|z|^2 + \sqrt{wt_0}\right| \sim |w|^{1/2} + |z|^2$, we can verify that

$$\beta = \partial_m f(\sqrt{w}, m_c(w)) \sim |w|^{1/2} + |z|^2 \sim 1.$$

With (2.106), it is easy to check that

$$\partial_m^2 f(\sqrt{w},\xi) = O(1), \quad \partial_m^3 f(\sqrt{w},\xi) = O(1),$$

for $|\xi - m_c| \le (\log N)^{-1/3}$, from which we get that $|\alpha| + |\partial_u \alpha| = O(1)$. With a fixed point argument, we conclude (2.355).

Case 5: Again we following the arguments in the proof of Case 1. However, instead of $f(\sqrt{w}, m)$, we shall study $\Upsilon(w, m_1)$ in (2.112) directly. We take over the notations in Definition 2.3.9 and abbreviate $R := \Upsilon(w, u_1)$, so that $|R| \le \delta$. Then we write the equation $\Upsilon(w, u_1) - \Upsilon(w, m_{1c}) = R$ as

$$\alpha(u_1)(u_1 - m_{1c})^2 + \beta(u_1 - m_{1c}) = R, \tag{2.374}$$

where we used the same symbols as in (2.356) for notational convenience. As in *Case 1*, we have $\beta = \partial_{m_1}\Upsilon(w, m_{1c})$, and we can estimate that $|\alpha| + |\partial_{u_1}\alpha| \le C$ for $w \in \mathbf{D}_L$ and u_1 satisfying $|u_1 - m_{1c}| \ll |m_{1c}|$. Now to conclude (2.116), it suffices to prove $|\beta| \sim 1$ for $w \in \mathbf{D}_L$. In fact with (2.112), we can obtain that

$$\beta = 1 + O\left(\eta^{-1}\right) \sim 1,$$

for $\eta \ge \zeta^{-1}$. This concludes the proof.

Proof of Lemma 2.2.2. The fact that ρ_{1c} has compact support follows from Lemma 2.2.3; ρ_{1c} being integrable follows from Lemma 2.7.4. Note that in proving Lemmas 2.2.3 and 2.7.4, we do not make use of the regularity assumptions in Definition 2.2.4. It remains to show that for fixed $w \in \mathbb{C}_+$ and $|z| \neq 1$, there exists a unique $m_{1c}(w) \in \mathbb{C}_+$ satisfying equation (2.12). This follows from the proof of *Case 1* in this section under the extra condition $\eta \sim 1$. Again, we do not need the regularity assumptions for the proof, because η^{-1} provides a nice bound for the Stieltjes transforms in the global region with $\eta \sim 1$. $\square$

Remark: The estimate (2.106) has been used repeatedly during the proof of Lemma 2.3.10. Here we remark that it also gives the stability of the regularity conditions in Definition 2.2.4 under perturbations of $|z|$ and ρ_Σ. For example, we define the shifted empirical spectral density

$$\rho_{\Sigma,t} := \frac{1}{N \wedge M} \sum_{i=1}^{N \wedge M} \delta_{\sigma_i + t}, \tag{2.375}$$

and the associated $m_c(w,t)$ and function $f(\sqrt{w}, m, t)$. Given a regular edge e_k, we have that

$$f(\sqrt{e_k}, m_k, t=0) = 0, \quad \partial_m f(\sqrt{e_k}, m_k, t=0) = 0,$$

where we denote $m_k := m_c(e_k)$. We have the Jacobian

$$J := \det \begin{pmatrix} \partial_{\sqrt{w}} f & \partial_m f \\ \partial_{\sqrt{w}} \partial_m f & \partial_m^2 f \end{pmatrix}_{(\sqrt{w},m,t)=(\sqrt{e_k},m_k,0)} = \partial_{\sqrt{w}} f(\sqrt{e_k}, m_k, 0) \partial_m^2 f(\sqrt{e_k}, m_k, 0).$$

By (2.324), we have $\left|\partial_{\sqrt{w}} f(\sqrt{e_k}, m_k, 0)\right| \geq 1$. Combining with (2.20), we get $|J| \geq \epsilon$. Using (2.106), we can verify that $\partial_t f(\sqrt{e_k}, m_k, 0) = O(1)$ and $\partial_t \partial_m f(\sqrt{e_k}, m_k, 0) = O(1)$. Thus if we regard e_k and m_k as functions of t, then $\partial_t m_k(t=0) = O(1)$ and $\partial_t e_k(t = 0) = O(1)$ by the implicit function theorem. Then it is easy to verify

$$\partial_m^2 f\left(\sqrt{e_k(t)}, m_c(e_k, t)\right) = \partial_m^2 f\left(\sqrt{e_k}, m_c(e_k)\right) + O(t),$$

$$|m_c(e_k, t) - a_i(e_k, t)| = |m_c(e_k) - a_i(e_k)| + O(t),$$

and similar estimates for $|m_c - b_i|$ and $|m_c + c_i|$. Thus if Definition 2.2.4 (i) holds for some ρ_Σ, then it holds for all $\rho_{\Sigma,t}$ provided that t is small enough.

Now given a regular bulk component $[e_{2k}, e_{2k-1}]$ and $E \in [e_{2k} + \tau', e_{2k-1} - \tau']$. Differentiating the equation $f(\sqrt{E}, m_c(E,t), t) = 0$ in t yields

$$\partial_t m_c(E,t) = -\frac{\partial_t f(\sqrt{E}, m_c(E,t), t)}{\partial_m f(\sqrt{E}, m_c(E,t), t)}.$$

By (2.106), we find that $\partial_t f(\sqrt{E}, m_c(E), 0) = O(1)$, while by (2.359), $|\partial_m f(\sqrt{E}, m_c(E), 0)| = \beta \sim 1$. Thus $\partial_t m_c(E, 0) = O(1)$. A simple extension of this argument shows that $m_c(E, t) = m_c(E) + O(t)$ and hence $\operatorname{Im} m_c(E, t)$ is bounded from below by some $c' = c'(\tau, \tau')$. Thus we conclude that if Definition 2.2.4 (ii) holds for some ρ_Σ, then it holds for all $\rho_{\Sigma,t}$ with t in some fixed small interval around zero. Obviously, the above arguments also work for $|z|$ perturbation.

2.7.4 Proof of Lemma 2.4.9

Our proof of (2.175) is an extension of [6, Lemma 4.9], [10, Lemma 7.3] and [22, Theorem 4.7]. Here we only prove the bound for $\|[Z]\|$. The proof for $\|\langle Z\rangle\|$ is exactly the same. For $i \in \mathcal{I}_1$, we define $P_i := \mathbb{E}_{[i]}$ and $Q_i := 1 - P_i$. Recall that $Z_{[i]} = Q_i G_{[ii]}^{-1}$. Hence we need to prove

$$[Z] = \frac{1}{N}\sum_{i=1}^N \pi_{[i]}\left(Q_i G_{[ii]}^{-1}\right)\pi_{[i]} \prec |w|^{-1/2}\,\Phi_o^2,$$

for $w \in \mathbf{D}$. For $J \subset \mathcal{I}$, we define $\pi_{[i]}^{[J]}$ by replacing $m_{1,2}$ in (2.37) with $m_{1,2}^{[J]}$ defined in (2.122). As in (2.174), we can prove that $|m_{1,2}^{[i]} - m_{1,2}| \prec |w|^{-1/2}\,\Phi_o^2$, which further gives that

$$[Z] = \frac{1}{N}\sum_{i=1}^N \pi_{[i]}^{[i]}\left(Q_i G_{[ii]}^{-1}\right)\pi_{[i]}^{[i]} + O_\prec\left(|w|^{-1/2}\,\Phi_o^2\right) = \frac{1}{N}\sum_{i=1}^N Q_i\left(\pi_{[i]}^{[i]} G_{[ii]}^{-1}\pi_{[i]}^{[i]}\right) + O_\prec\left(|w|^{-1/2}\,\Phi_o^2\right).$$

Thus if we abbreviate $B_i := |w|^{1/2} Q_i\left(\pi_{[i]}^{[i]} G_{[ii]}^{-1}\pi_{[i]}^{[i]}\right)$, it suffices to prove that $B := N^{-1}\sum_i B_i \prec \Phi_o^2$. We will estimate B by bounding the p-th moment of its norm by Φ_o^{2p} for $p = 2n \in 2\mathbb{N}$, i.e. $\mathbb{E}\|B\|^p \prec \Phi_o^{2p}$. The lemma then follows from the Markov's inequality. Using $\|KK^\dagger\| = \|K\|^2$, we have that

$$\operatorname{Tr}(BB^\dagger)^n \geq \left\|BB^\dagger\right\|^n = \|B\|^{2n}.$$

Thus it suffices to prove that

$$\mathbb{E}\mathrm{Tr}(BB^{\dagger})^{p/2} \prec \Phi_o^{2p}, \quad \text{for } p = 2n. \tag{2.376}$$

This estimate can be proved with the same method as in [22, Appendix B], with the only complication being that $\pi_{[i]}$ is random and depends on i. In principle, this can be handle by using (2.86) and (2.87) to put any indices $j, k, ... \in \mathcal{I}_1$ (that we wish to include) into the superscripts of $\pi_{[i]}$. This leads to a minor modification of the proof in [22, Appendix B]. Here we describe the basic ideas of the proof, without writing down all the details.

The proof is based on a decomposition of the space of random variables using P_s and Q_s. It is evident that P_s and Q_s are projections, $P_s + Q_s = 1$ and all of these projections commute with each other. For a set $J \subset \mathcal{I}$, we denote $P_J := \prod_{s\in J} P_s$ and $Q_J := \prod_{s\in J} Q_s$. Let $p = 2n$ and introduce the shorthand notation $\tilde{B}_{k_s} := B_{k_s}$ for odd $s \le p$ and $\tilde{B}_{k_s} := B_{k_s}^{\dagger}$ for even $s \le p$. Then we get

$$\mathbb{E}\mathrm{Tr}(BB^{\dagger})^{p/2} = \frac{1}{N^p} \sum_{k_1,k_2,\ldots,k_p} \mathbb{E}\mathrm{Tr} \prod_{s=1}^{p} \tilde{B}_{k_s} = \frac{1}{N^p} \sum_{k_1,k_2,\ldots,k_p} \mathbb{E}\mathrm{Tr} \prod_{s=1}^{p} \left(\prod_{r=1}^{p} (P_{k_r} + Q_{k_r}) \tilde{B}_{k_s} \right). \tag{2.377}$$

Introducing the notations $\mathbf{k} = (k_1, k_2, \ldots, k_p)$ and $\{\mathbf{k}\} = \{k_1, k_2, \ldots, k_p\}$, we can write

$$\mathbb{E}\mathrm{Tr}(BB^{\dagger})^{p/2} = \frac{1}{N^p} \sum_{\mathbf{k}} \sum_{I_1,\ldots,I_p \subset \{\mathbf{k}\}} \mathbb{E}\mathrm{Tr} \prod_{s=1}^{p} \left(P_{I_s^c} Q_{I_s} \tilde{B}_{k_s} \right). \tag{2.378}$$

Following the arguments in [22, Appendix B], one can see that to conclude (2.376) it suffices to prove that for $k \in I$,

$$\|Q_I B_k\| \prec \Phi_o^{|I|}. \tag{2.379}$$

As in [22, Appendix B], it is not hard to prove that for $k \in I$,

$$|w|^{-1/2} \left\| Q_I G_{[kk]}^{-1} \right\| \prec \Phi_o^{|I|}, \quad \text{and} \quad |w|^{-1/2} \left\| Q_{I\setminus\{k\}} G_{[kk]}^{-1} \right\| \prec \Phi_o^{|I|} \text{ if } |I| \ge 2. \tag{2.380}$$

Now we extend the proof to obtain the estimate (2.379). For the case $|I| = 1$ (i.e. $I = \{k\}$),

$$\|B_k\| = |w|^{1/2}\|\pi^{[i]}_{[i]} Z_{[k]} \pi^{[i]}_{[i]}\| \le |w|^{-1/2}\|Z_{[k]}\| \prec \Phi_o,$$

where we used $\|Z_{[k]}\| \prec |w|^{1/2}\Phi_o$, which can be proved with the same arguments as in Lemma 2.4.4. For the case $|I| \ge 2$, WLOG, we may assume $k = 1$ and $I = \{1, \ldots, t\}$ with $t \ge 2$. It is enough to prove that

$$|w|^{1/2} \left\| Q_t \ldots Q_2 Q_1 \pi^{[1]}_{[1]} G^{-1}_{[11]} \pi^{[1]}_{[1]} \right\| \prec \Phi_o^t. \tag{2.381}$$

We take $t = 3$ as an example to describe the ideas for the proof of (2.381). Using (2.86), we get

$$\pi^{[1]}_{[1]} = \pi^{[12]}_{[1]} + |w|^{1/2}\epsilon^{[1]}_{11}\pi^{[12]}_{[1]} A_1 \pi^{[12]}_{[1]} + |w|^{1/2}\epsilon^{[1]}_{\bar{1}\bar{1}}\pi^{[12]}_{[1]} A_2 \pi^{[12]}_{[1]} + \mathrm{error}_{1,2}, \tag{2.382}$$

where $\epsilon^{[1]}_{11}$ and $\epsilon^{[1]}_{\bar{1}\bar{1}}$ are the upper left and lower right entries of

$$\epsilon^{[1]}_{[1]} := |w|^{1/2} \left(\frac{G^{[1]}_{[22]}}{N} + \frac{1}{N} \sum_{k \notin \{1,2\}} G^{[1]}_{[k2]} \left(G^{[1]}_{[22]} \right)^{-1} G^{[1]}_{[2k]} \right) \prec \Phi_o^2,$$

$A_{1,2}$ are deterministic matrices with operator norm $O(1)$, and $\|\mathrm{error}_{1,2}\| \prec |w|^{-1/2}\Phi_o^4$. Then we get

$$\begin{aligned} \pi^{[1]}_{[1]} G^{-1}_{[11]} \pi^{[1]}_{[1]} = {} & \pi^{[12]}_{[1]} G^{-1}_{[11]} \pi^{[12]}_{[1]} + |w|^{1/2}\epsilon^{[1]}_{11}\pi^{[12]}_{[1]} A_1 \pi^{[12]}_{[1]} G^{-1}_{[11]} \pi^{[12]}_{[1]} + |w|^{1/2}\epsilon^{[1]}_{\bar{1}\bar{1}}\pi^{[12]}_{[1]} A_2 \pi^{[12]}_{[1]} G^{-1}_{[11]} \pi^{[12]}_{[1]} \\ & + |w|^{1/2}\pi^{[12]}_{[1]} G^{-1}_{[11]} \epsilon^{[1]}_{11} \pi^{[12]}_{[1]} A_1 \pi^{[12]}_{[1]} + |w|^{1/2}\pi^{[12]}_{[1]} G^{-1}_{[11]} \epsilon^{[1]}_{\bar{1}\bar{1}} \pi^{[12]}_{[1]} A_2 \pi^{[12]}_{[1]} + O_{\prec}(|w|^{-1/2}\Phi_o^4). \end{aligned} \tag{2.383}$$

We first handle the $\pi^{[12]}_{[1]} G^{-1}_{[11]} \pi^{[12]}_{[1]}$ term. By (2.380), we have

$$Q_2 \pi^{[12]}_{[1]} G^{-1}_{[11]} \pi^{[12]}_{[1]} = \pi^{[12]}_{[1]} \left(Q_2 G^{-1}_{[11]} \right) \pi^{[12]}_{[1]} \prec |w|^{-1/2}\Phi_o^2.$$

For the remaining term, we first expand $\pi_{[1]}^{[12]} = \pi_{[1]}^{[123]} + O_\prec(|w|^{-1/2}\Phi_o^2)$ and use (2.380) to get

$$Q_3 Q_2 \pi_{[1]}^{[12]} G_{[11]}^{-1} \pi_{[1]}^{[12]} = \pi_{[1]}^{[123]} \left(Q_3 Q_2 G_{[11]}^{-1} \right) \pi_{[1]}^{[123]} + O_\prec \left(|w|^{-1/2}\Phi_o^4 \right) \prec |w|^{-1/2}\Phi_o^3.$$

Then we deal with the second terms in (2.383). We first expand $\epsilon_{[1]}^{[1]} = e_{[1]}^{[3]} + O_\prec(\Phi_o^3)$, where

$$e_{[1]}^{[3]} := |w|^{1/2} \left(\frac{G_{[22]}^{[13]}}{N} + \frac{1}{N} \sum_{k\notin\{1,2,3\}} G_{[k2]}^{[13]} \left(G_{[22]}^{[13]} \right)^{-1} G_{[2k]}^{[13]} \right).$$

Using the similar arguments as above, we get

$$Q_3 |w|^{1/2} e_{11}^{[3]} \pi_{[1]}^{[12]} A_1 \pi_{[1]}^{[12]} G_{[11]}^{-1} \pi_{[1]}^{[12]} = |w|^{1/2} e_{11}^{[3]} \pi_{[1]}^{[123]} A_1 \pi_{[1]}^{[123]} \left(Q_3 G_{[11]}^{-1} \right) \pi_{[1]}^{[123]} + O_\prec(|w|^{-1/2}\Phi_o^4)$$
$$\prec |w|^{-1/2}\Phi_o^4.$$

Thus we have

$$Q_2 Q_3 |w|^{1/2} \epsilon_{11}^{[1]} \pi_{[1]}^{[12]} A_1 \pi_{[1]}^{[12]} G_{[11]}^{-1} \pi_{[1]}^{[12]} \prec |w|^{-1/2}\Phi_o^3.$$

Obviously this kind of estimate works for the rest of the terms in (2.383). This proves (2.381) when $t = 3$.

We can continue in this manner for a general t. At the l-th step, we expand the leading order terms using (2.86) and (2.87), and after applying $Q_l \ldots Q_3 Q_2$ on them, the number of Φ_o factors increases by one at each step by (2.380). Trough induction we can prove (2.381). In fact the expansions can be performed in a systematic way using the method in [22, Appendix B], and we leave the details to the reader. Also we remark that similar techniques are used in the proof of anisotropic local law in Section 2.5, and we choose to present the details there (in fact the proof here is much easier than the one in Section 2.5).

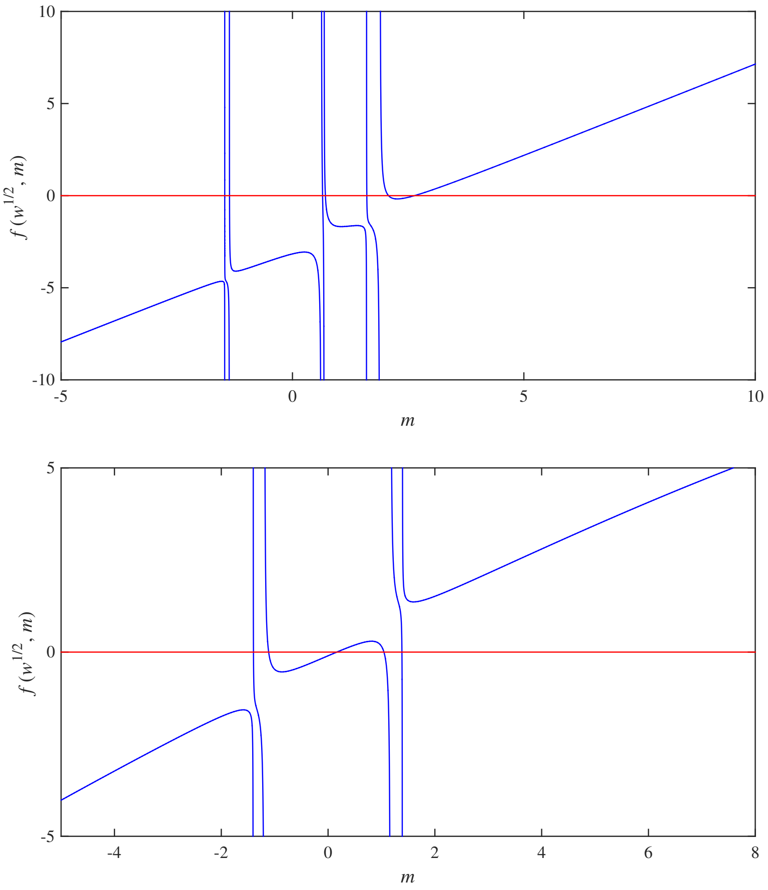


Figure 6: The graphs of $f(\sqrt{w}, m)$ for the example from Figure 1, i.e. $\rho_\Sigma = 0.5\delta_{\sqrt{2/17}} + 0.5\delta_{4\sqrt{2/17}}$. We take $|z| = 1.5$, and $w = 10$ and 0.01 in the upper and lower graphs, respectively. In the lower graph, we only plot the five branches near $m = 0$. The remaining two branches are far away.

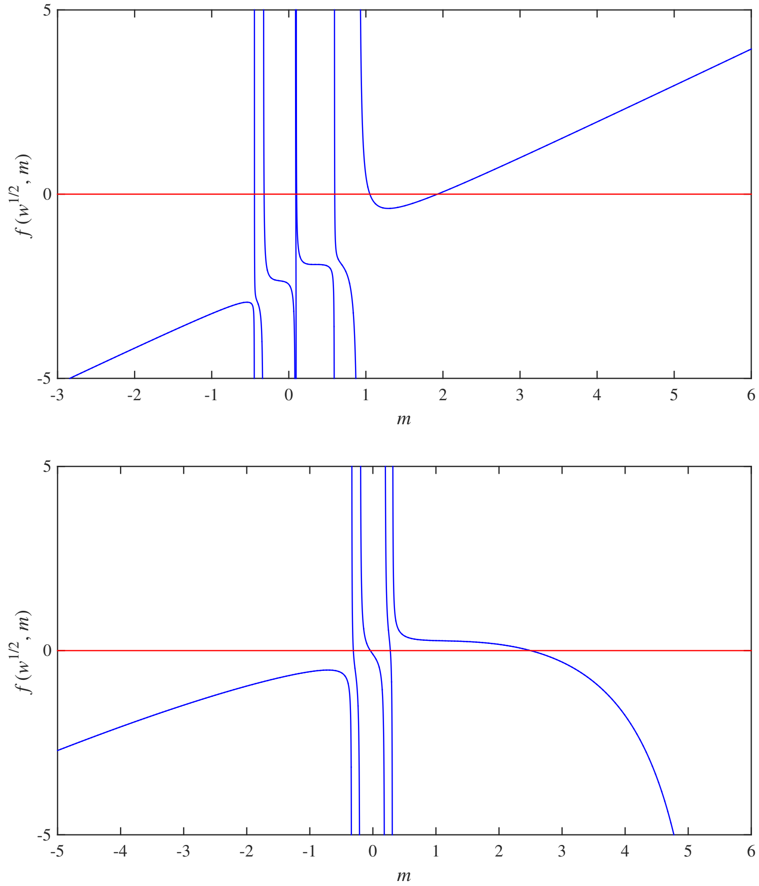


Figure 7: The graphs of $f(\sqrt{w}, m)$ for the example from Figure 1, i.e. $\rho_\Sigma = 0.5\delta_{\sqrt{2/17}} + 0.5\delta_{4\sqrt{2/17}}$. We take $|z| = 0.5$, and $w = 6$ and 0.01 in the upper and lower graphs, respectively. In the lower graph, we only plot the five branches near $m = 0$. The remaining two branches are far away.

Chapter 3

Convergence of eigenvector empirical spectral distribution of sample covariance matrices

3.1 Introduction

Let $X = (x_{ij})$ be an $M \times N$ real or complex data matrix whose entries are independent (but not necessarily identically distributed) random variables satisfying

$$\mathbb{E}x_{ij} = 0, \quad \mathbb{E}|x_{ij}|^2 = N^{-1}, \quad 1 \le i \le M, \ 1 \le j \le N, \tag{3.1}$$

and in addition,

$$\mathbb{E}x_{ij}^2 = 0, \quad \text{if } x_{ij} \text{ is complex.} \tag{3.2}$$

Then XX^* gives a class of $M \times M$ sample covariance matrices. Its empirical spectral distribution (ESD) is defined as

$$F_{XX^*}^{(M)}(x) := \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}_{\{\lambda_i(XX^*) \le x\}},$$

where $\lambda_1(XX^*) \geq \ldots \geq \lambda_M(XX^*)$ are the eigenvalues of XX^* and $\mathbf{1}_{\{\cdot\}}$ denotes the conventional indicator function. Similarly, one can also consider X^*X and its ESD

$$F_{X^*X}^{(N)}(x) := \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{\lambda_i(X^*X)\leq x\}}.$$

Sample covariance matrices are fundamental objects in modern multivariate statistics, where the advance of technology has led to high dimensional data such that M is comparable to or even larger than N. These large dimensional covariance matrices have many applications in various fields, such as statistics [15, 39, 40, 41], economics [51] and population genetics [53]. Define the aspect ratio

$$d_N := N/M.$$

We are interested in the regime where $\lim_{N\to\infty} d_N = d \in (0, \infty)$, i.e. M and N are proportional to each other. In this case, it is well-known that $F_{XX^*}^{(M)}$ converges weakly to the Marchenko-Pastur (MP) law $F_d(x)$ [48], which has a $(1-d)_+$ mass at $x = 0$ and has a density

$$\rho_{1c}(x) = (1-d)_+\delta_0 + \frac{d}{2\pi}\frac{\sqrt{[(\lambda_+ - x)(x-\lambda_-)]_+}}{x}, \quad \lambda_\pm = (1 \pm d^{-1/2})^2, \tag{3.3}$$

in the interval $[\lambda_-, \lambda_+]$ (where λ_+ and λ_- are often referred to as the soft edge and hard edge, respectively). For $z \in \mathbb{C}_+ := \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$, the Stieltjes transform $m_{1c}(d, z)$ of F_d satisfies the following self-consistent equation

$$m_{1c}(d,z) + \frac{1}{z - (1-d^{-1}) + zd^{-1}m_{1c}(d,z)} = 0, \tag{3.4}$$

and has the closed form expression

$$m_{1c}(d,z) = \frac{1 - d^{-1} - z + i\sqrt{(\lambda_+ - z)(z-\lambda_-)}}{2zd^{-1}}. \tag{3.5}$$

Since XX^* and X^*X share the same nonzero eigenvalues, it is easy to see that $F^{(N)}_{X^*X}$ also converges weakly to a deterministic law, whose Stieltjes transform $m_{2c}(z)$ satisfies

$$m_{2c}(d,z) = \frac{d^{-1}-1}{z} + d^{-1}m_{1c}(d,z). \tag{3.6}$$

In applications of spectral analysis of large dimensional random matrices, one of the important problems is the convergence rate of the ESD. Recall that the Kolmogorov distance between two distributions F_1 and F_2 is defined as

$$\|F_1 - F_2\| := \sup_x |F_1(x) - F_2(x)|.$$

Then we use $\|F^{(M)}_{XX^*} - F_{d_N}\|$ to measure the convergence rate of $F^{(M)}_{XX^*}$. The convergence rate of the ESD of sample covariance matrices was first studied in [1] using Berry-Esseen type inequalities for the difference of two distributions in terms of their Stieltjes transforms. The Berry-Esseen type inequalities were later improved in [35] to show that the convergence rate is $O(N^{-1/2})$ in probability under finite 8th moment condition. A sharper bound was obtained in [54], where the authors proved that $\|F^{(M)}_{XX^*} - F_{d_N}\| = O(N^{-1}(\log N)^{O(\log\log N)})$ in probability under the sub-exponential decay assumption.

The properties of eigenvectors of large dimensional random matrices are much harder to study. However, great progress has been made in this direction; see [24, 6] for the delocalization and isotropic delocalization of eigenvectors, [42, 63] for the universality of eigenvectors, [9] for the local quantum unique ergodicity of eigenvectors, [**?**] for the eigenvectors of principal components, [59, 60, 3, 69, 70] for the asymptotical Haar property of the eigenmatrix based on the VESD (see (3.8) below), to name a few. Note that some of these results are proved for Wigner matrices, but their generalizations to sample covariance matrices usually are straightforward.

3.1.1 Main results

This chapter is concerned with the *eigenvector empirical spectral distribution* (VESD) of sample covariance matrices, which we shall now define. Suppose XX^* has the spectral decomposition

$$XX^* = \sum_{k=1}^{M} \lambda_k(XX^*)\xi_k\xi_k^*, \tag{3.7}$$

where ξ_k are the eigenvectors. We define the VESD of XX^* as

$$F_{XX^*}^{(M)}(\mathbf{v}, x) = \sum_{i=1}^{M} |\langle \xi_k, \mathbf{v}\rangle|^2 \mathbf{1}_{\{\lambda_k(XX^*)\leq x\}}, \tag{3.8}$$

where $\mathbf{v}$ is a deterministic unit vector in $\mathbb{C}^M$. It was proved in [3, 6] that for any fixed $\mathbf{v}$, $F_{XX^*}^{(M)}(\mathbf{v}, x)$ converges weakly to the MP law as $N \to \infty$. Compared with ESD, much less has been done on the convergence rate of VESD. To the best of our knowledge, there is only one paper [70] studying on this topic. In that paper, the authors proved that if $d_N > 1$ and the entries of X are identically distributed, then $\|\mathbb{E}F_{XX^*}^{(M)}(\mathbf{v}, \cdot) - F_{d_N}(\cdot)\| = O(N^{-1/2})$ under the finite 10th moment assumption, and $\|F_{XX^*}^{(M)}(\mathbf{v}, \cdot) - F_{d_N}(\cdot)\| = O(N^{-1/4+\epsilon})$ almost surely under the finite 8th moment assumption. However, we find that both of these bounds are far away from being optimal, and can be improved using a different method. This is the main purpose of this work. As demonstrated in [59, 60, 70], the convergence of the VESD for any fixed unit vector $\mathbf{v}$ can be used to characterize the asymptotical Haar property of the eigenmatrix of XX^*. Thus we expect that a better bound for the convergence rate will lead to a better understanding of the Haar properties of the eigenvectors of large sample covariance matrices.

Now we are ready to state the main result of this chapter. For a reason that will be clear later (when we prove Corollary 3.1.2), we consider slightly more general random matrices $X = (x_{ij})$. More specifically, we define the following conditions: there exist constants $C_0, c_0 > 0$ such that for all $1 \leq i \leq M$ and $1 \leq j \leq N$,

$$|\mathbb{E}x_{ij}| \leq C_0 N^{-2-c_0}, \tag{3.9}$$

$$\left|\mathbb{E}|x_{ij}|^2 - N^{-1}\right| \leq C_0 N^{-2-c_0}, \tag{3.10}$$

$$\mathbb{E}|x_{ij}|^4 \leq C_0 N^{-2}, \tag{3.11}$$

$$\left|\mathbb{E}x_{ij}^2\right| \leq C_0 N^{-2-c_0}, \quad \text{if } x_{ij} \text{ is complex.} \tag{3.12}$$

Theorem 3.1.1. *Let $X = (x_{ij})$ be an $M \times N$ random matrix whose entries are independent random variables satisfying (3.9), (3.10), (3.11) and (3.12). Suppose there exist constants $C_1, \phi > 0$ such that*

$$\max_{1 \leq i \leq M, 1 \leq j \leq N} |x_{ij}| \leq C_1 N^{-\phi}. \tag{3.13}$$

Suppose $d_N \to d$ for some constant $d \neq 1$. Then for any fixed (small) $\epsilon > 0$ and (large) $D > 0$, we have

$$\|\mathbb{E}F_{XX^*}^{(M)}(\mathbf{v}, \cdot) - F_{d_N}(\cdot)\| \leq N^{-1+\epsilon}, \tag{3.14}$$

and

$$\mathbb{P}\left(\|F_{XX^*}^{(M)}(\mathbf{v}, \cdot) - F_{d_N}(\cdot)\| \geq N^{\epsilon}\left(N^{-2\phi} + N^{-1/2}\right)\right) \leq N^{-D}, \tag{3.15}$$

for all deterministic unit vectors $\mathbf{v} \in \mathbb{C}^M$, provided that N is large enough.

As an immediate corollary, we have the following result.

Corollary 3.1.2. *Let $X = (x_{ij})$ be an $M \times N$ random matrix whose entries are independent random variables satisfying (3.1) and (3.2). Suppose there exist constants $a, A > 0$*

such that

$$\max_{1\le i\le M,1\le j\le N}\mathbb{E}|\sqrt{N}x_{ij}|^a\le A \tag{3.16}$$

for all N*. Suppose* $d_N\to d$ *for some constant* $d\neq 1$*. Then if* $a\ge 6$*, we have for any fixed* $\epsilon>0$ *and deterministic unit vector* $\mathbf{v}\in\mathbb{C}^M$*,*

$$\|\mathbb{E}F^{(M)}_{XX^*}(\mathbf{v},\cdot)-F_{d_N}(\cdot)\|\le N^{-1+\epsilon} \tag{3.17}$$

for sufficiently large N*. Moreover, if* $a\ge 8$*, we have*

$$\mathbb{P}\left(\limsup_{N\to\infty}N^{1/2-\epsilon}\|F^{(M)}_{XX^*}(\mathbf{v},\cdot)-F_{d_N}(\cdot)\|\le 1\right)=1. \tag{3.18}$$

Proof. In the proof, we fix $a>4$ and choose a constant $\phi>0$ small enough such that $\left(N^{1/2-\phi}\right)^a\ge N^{2+\omega}$ for some constant $\omega>0$. Then we introduce the following truncation

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

By the moment condition (3.16), we have

$$\mathbb{P}(\tilde{X}\neq X)=O(N^{2-a/2+a\phi}). \tag{3.19}$$

Moreover, we have

$$\begin{aligned}\mathbb{P}(\tilde{X}\neq X\text{ i.o.})&=\lim_{k\to\infty}\mathbb{P}\left(\cup_{N=k}^\infty\cup_{i=1}^M\cup_{j=1}^N\left\{|x_{ij}|\ge N^{-\phi}\right\}\right)\\&=\lim_{k\to\infty}\mathbb{P}\left(\cup_{t=k}^\infty\cup_{N\in[2^t,2^{t+1})}\cup_{i=1}^M\cup_{j=1}^N\left\{|x_{ij}|\ge N^{-\phi}\right\}\right)\\&\le C\lim_{k\to\infty}\sum_{t=k}^\infty\left(2^{t+1}\right)^2\left(2^{t(1/2-\phi)}\right)^{-a}\le C\lim_{k\to\infty}\sum_{t=k}^\infty 2^{-\omega t}=0,\end{aligned} \tag{3.20}$$

i.e. $\tilde{X}=X$ almost surely as $N\to\infty$. Here in the above derivation, we regard $M\equiv N/d_N$ as a function depending on N, which, by the given condition on d_N, satisfies $M=O(N)$ for large enough N.

Using (3.16) and integration by parts, we can get that

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

which imply that

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

and

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

Moreover, we trivially have

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

Hence $\tilde{X}$ is a random matrix satisfying the assumptions in Theorem 3.1.1. Then using (3.14) and (3.19) with $a = 6$ and $\phi = \epsilon/6$, we conclude (3.17); using (3.15) and (3.20) with $\phi = (1-\epsilon)/4$ and $a = 8$, we conclude (3.18). □

Remark 3.1.3. *The estimates (3.17) and (3.18) improve the bounds obtained in [70] (and relax the moment assumptions as well). We believe that the convergence rates in (3.17) and (3.18) are close to optimal due to the following reasons. It was proved in [3] that for an analytic function f,*

$$\sqrt{N}\int f(x)d\left(F_{XX^*}^{(M)}(\mathbf{v},x) - F_{d_N}(x)\right) \to \mathcal{N}(0,\sigma_f), \tag{3.21}$$

where $\mathcal{N}(0,\sigma_f)$ denotes the Gaussian distribution with mean zero and variance σ_f. This shows that the fluctuation of $F_{XX^*}^{(M)}(x)$ is of order $N^{-1/2}$ and suggests the bound in (3.18). Taking expectation of (3.21), one can see that the order of $|\mathbb{E}F_{XX^*}^{(M)}(\mathbf{v},x) - F_{d_N}(x)|$ should be even smaller. On the other hand, the fluctuation of the eigenvalues of XX^* on the *microscopic scale* N^{-1} *will lead to an error of order at least* N^{-1}*. This shows that the bound (3.17) is very close to being optimal.*

Remark 3.1.4. *In [70], the authors can only handle the* $M < N$ *(i.e.* $d_N > 1$*) case, while our proof will work for both the* $d > 1$ *and* $d < 1$ *cases. However, in the case with* $d_N \to 1$*, we will encounter some difficulties near the hard edge* λ_-*, which converges to* 0 *as* $N \to \infty$ *by (3.3). However, we can still prove weaker versions of (3.14) and (3.15) by restricting ourself to the region away from* 0*. For instance, we have for any fixed* $\tau > 0$*,*

$$\sup_{x\geq\tau} |\mathbb{E}F^{(M)}_{XX^*}(\mathbf{v},\cdot) - F_{d_N}(\cdot)| \leq N^{-1+\epsilon}$$

under the assumptions in Theorem 3.1.1. Similarly, the bound in (3.15) also holds if we only take the $\sup$ *over* $x \geq \tau$*.*

3.1.2 Main ideas

A basic tool for the proof is the Stieltjes transform. For any $z = E + i\eta \in \mathbb{C}_+$, we define the resolvent of XX^* as

$$\mathcal{G}(X,z) := (XX^* - z)^{-1}.$$

Then the Stieltjes transform of $F^{(M)}_{XX^*}(\mathbf{v},\cdot)$ is equal to $\langle \mathbf{v}, \mathcal{G}(X,z)\mathbf{v}\rangle$, and we have the asymptotic estimate

$$\langle \mathbf{v}, \mathcal{G}(X,z)\mathbf{v}\rangle \approx m_{1c}(d_N, z) \tag{3.22}$$

for any fixed $\eta > 0$, when N is large. By taking the imaginary part, it is easy to see that a control of the Stieltjes transform $\langle \mathbf{v}, \mathcal{G}(X,z)\mathbf{v}\rangle$ yields a control of the VESD on a small scale of order η around E. An *isotropic local law* is an estimate of the form (3.22) for all $\eta \gg N^{-1}$. Such isotropic local law was first established in [6] for sample covariance matrices and generalized Wigner matrices, assuming the matrix entries have arbitrarily high moments.

Now we briefly describe the ideas for the proof of Theorem 3.1.1. Following the approach in [54] (which is used to prove the convergence rate of ESD), the main idea is that the estimates (3.14) and (3.15) follow from an appropriate isotropic local law for $\mathcal{G}(X, z)$ up to the optimal scale $\eta \gg N^{-1}$ (see Section 3.3). In fact, a generalization of the proof in [6] gives roughly the following estimate (see Theorem 3.2.8): for any fixed $\epsilon > 0$,

$$|\langle \mathbf{v}, \mathcal{G}(X, z)\mathbf{v}\rangle - m_{1c}(d_N, z)| \le N^{\epsilon}\left(N^{-2\phi} + (N\eta)^{-1/2}\right) \tag{3.23}$$

with extremely high probability for all $\operatorname{Im} z \ge N^{-1+\epsilon}$. Then this estimate will imply (3.15). However, to conclude (3.14), we need a much stronger bound for the expected resolvent, i.e., for any fixed $\epsilon > 0$,

$$|\mathbb{E}\langle \mathbf{v}, \mathcal{G}(X, z)\mathbf{v}\rangle - m_{1c}(d_N, z)| \le \frac{N^{\epsilon}}{N\eta}. \tag{3.24}$$

for all $\operatorname{Im} z \ge N^{-1+\epsilon}$. The improvement of the weak bound (3.23) to the almost optimal one in (3.24) constitutes the main novelty of this work.

A key observation is that (see Section 3.4.2), after taking the expectation the leading order term in $\Delta m(\mathbf{v}, X) := \langle \mathbf{v}, \mathcal{G}(X, z)\mathbf{v}\rangle - m_{1c}(d_N, z)$ vanishes, and hence make $\mathbb{E}\Delta m(\mathbf{v}, X)$ to be one order smaller than the bound in (3.23). In other words, we have

$$|\mathbb{E}\langle \mathbf{v}, \mathcal{G}(X, z)\mathbf{v}\rangle - m_{1c}(d_N, z)| \le N^{\epsilon}\left(N^{-4\phi} + (N\eta)^{-1}\right). \tag{3.25}$$

This already gives the estimate (3.24) if $\phi \ge 1/4$. For X satisfying (3.13) for some $\phi < 1/4$, we shall construct another random matrix $\tilde{X}$ which can well approximate X but has bounded entries, i.e. $\max_{i,j} \sqrt{N}|x_{ij}| = O(1)$ (see Lemma 3.5.1). Then the resolvent of $\tilde{X}\tilde{X}^*$ satisfies (3.24) by taking $\phi = 1/2$ in (3.25). On the other hand, with a resolvent comparison method developed in [46], we will show that the difference between $\mathbb{E}\langle \mathbf{v}, \mathcal{G}(X,z)\mathbf{v}\rangle$ and $\mathbb{E}\langle \mathbf{v}, \mathcal{G}(\tilde{X},z)\mathbf{v}\rangle$ is of order $(N\eta)^{-1}$; see Section 3.5. This concludes (3.24).

Remark 3.1.5. *It is possible to generalize our proof to more general random matrix models. For example, one may consider sample covariance matrices of the form* $Q := (TX)(TX)^*$ *(*T *is a general deterministic rectangular matrix), generalized Wigner matrices (i.e. Wigner ensembles whose entries have non-identical variances) and deformed Wigner matrices of the form* $H + A$ *(*H *is a Wigner matrix and* A *is a deterministic Hermitian matrix). The convergence of VESD of these models will be studied in future works. In particular, we expect that our proof applied to the Wigner matrices can improve the results obtained in [69].*

Remark 3.1.6. *For definiteness, we will focus on real sample covariance matrices during the proof. However, our proof also applies, after minor changes, to the complex case if we include the extra assumption (3.2) or (3.12). Also, we will only use* d_N *(instead of* d*) in the rest of this chapter. Correspondingly, we will use the quantities* $\rho_{1c}^{(N)}$, $m_{1,2c}^{(N)}$ *and* $\lambda_{\pm}^{(N)}$, *which are obtained by replacing* d *with* d_N *in (3.3)-(3.6). For simplicity, we shall always omit the superscript and still call them* ρ_{1c}, $m_{1,2c}$ *and* $\lambda_{\pm}$ *in the proof.*

The rest of this chapter is organized as follows. In Section 3.2, we introduce the notations and collect some tools that will be used in proving Theorem 3.1.1. The most important results in this section are Theorem 3.2.8 and Theorem 3.2.9, which give the isotropic local law for the resolvent $\mathcal{G}(X,z)$. In Section 3.3, we prove Theorem 3.1.1 using Theorem 3.2.8 and Theorem 3.2.9. Finally, the Theorem 3.2.8 and Theorem 3.2.9

are proved in Section 3.4 and Section 3.5, respectively.

3.1.3 Conventions

The fundamental large parameter is N, and we regard $M \equiv M_N$ as depending on N. All quantities that are not explicitly constant may depend on N, and we usually omit the argument N from our notations.

We use C to denote a generic large positive constant, which may depend on some fixed parameters and whose value may change from one line to the next. Similarly, we use c, ϵ, ϕ, τ, etc. to denote generic small positive constants. For two quantities a_N and b_N depending on N, the notation $a_N = O(b_N)$ means that $|a_N| \le C|b_N|$ for some constant $C > 0$, and $a_N = o(b_N)$ means that $|a_N| \le c_N|b_N|$ for some positive sequence $\{c_N\}$ with $c_N \to 0$ as $N \to \infty$. We also use the notation $a_N \sim b_N$ if $a_N = O(b_N)$ and $b_N = O(a_N)$. For a matrix A, we use $\|A\| := \|A\|_{l^2 \to l^2}$ to denote its operator norm. For a vector $\mathbf{v} = (v_i)_{i=1}^n \in \mathbb{C}^n$, $\|\mathbf{v}\| \equiv \|\mathbf{v}\|_2$ stands for the Euclidean norm of $\mathbf{v}$, while $|\mathbf{v}| \equiv \|\mathbf{v}\|_1$ stands for the l^1-norm. We denote the inner product in $\mathbb{C}^n$ by $\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{i=1}^n \bar{v}_i w_i$.

3.2 Main Tools

3.2.1 Resolvents and local Marchenko-Pastur law

Our study of sample covariance matrices can be performed in a simple and unified fashion using the following $(N+M) \times (N+M)$ self-adjoint matrix H, which is a linear function of X. It was used previously in [16, 44] to prove the local laws of sample covariance matrices.

Definition 3.2.1 (Resolvents)**.** *We define the* $(N+M)\times(N+M)$ *matrix*

$$H := \begin{pmatrix} 0 & X \\ X^* & 0 \end{pmatrix}. \tag{3.26}$$

For $z \in \mathbb{C}_+$, *we define the resolvent for* H*:*

$$G(X,z) := \begin{pmatrix} -I_{M\times M} & X \\ X^* & -zI_{N\times N} \end{pmatrix}^{-1}, \tag{3.27}$$

and the resolvents (or the Green functions) for XX^* *and* X^*X*:*

$$\mathcal{G}_1(X,z) := (XX^* - z)^{-1}, \quad \mathcal{G}_2(X,z) := (X^*X - z)^{-1}. \tag{3.28}$$

The Stieltjes transform of the ESD of XX^* *is given by*

$$m_1(X,z) := \int \frac{1}{x-z} dF^{(M)}_{XX^*}(x) = \frac{1}{M} \operatorname{Tr} \mathcal{G}_1(X,z).$$

Similarly, we also define $m_2(X,z) := N^{-1}\operatorname{Tr}\mathcal{G}_2(X,z)$. *During the proof, we often omit the arguments* X, z *from our notations.*

Remark 3.2.2. *Since the nonzero eigenvalues of* X^*X *and* XX^* *are identical and* XX^* *has* $M-N$ *more (or* $N-M$ *less) zero eigenvalues, we have*

$$F^{(M)}_{XX^*} = d_N F^{(N)}_{X^*X} + (1-d_N)\mathbf{1}_{[0,\infty)},$$

which implies that (see also (3.6))

$$m_2(z) = \frac{d_N^{-1}-1}{z} + d_N^{-1} m_1(z). \tag{3.29}$$

For simplicity of notations, we define the index sets

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

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

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

Using Schur complement formula, it is easy to check that

$$G = \begin{pmatrix} z\mathcal{G}_1 & \mathcal{G}_1 X \\ X^*\mathcal{G}_1 & \mathcal{G}_2 \end{pmatrix} = \begin{pmatrix} z\mathcal{G}_1 & X\mathcal{G}_2 \\ \mathcal{G}_2 X^* & \mathcal{G}_2 \end{pmatrix}. \tag{3.30}$$

Thus a control of G yields a control of the resolvents $\mathcal{G}_1$ and $\mathcal{G}_2$. Moreover, we have

$$m_1 = \frac{1}{Mz} \sum_{i \in \mathcal{I}_1} G_{ii}, \quad m_2 = \frac{1}{N} \sum_{\mu \in \mathcal{I}_2} G_{\mu\mu}.$$

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

$$\mathbf{D}(\zeta, N) := \{z \in \mathbb{C}_+ : \max(\zeta, \lambda_-/2) \le E \le 2\lambda_+, N^{-1+\zeta} \le \eta \le \zeta^{-1}\}, \tag{3.31}$$

for some small constant $\zeta > 0$, unless otherwise indicated. Note that if $d_N \to d$ for some constant $d \neq 1$, then by (3.3) we have $\lambda_- \sim 1$ when N is sufficiently large. Thus we can always take ζ to be sufficiently small such that $\zeta \le \lambda_-/2$. We define the distance to the spectral edges as

$$\kappa := \min\{|E - \lambda_+|, |E - \lambda_-|\}. \tag{3.32}$$

The next lemma gives some basic properties of $m_{1,2c}$, which can be proved through direct calculations using (3.5) and (3.6).

Lemma 3.2.3. *For $z \in \mathbf{D}$, we have*

$$|m_{1,2c}(z)| \sim 1, \quad \operatorname{Im} m_{1,2c}(z) \sim \begin{cases} \eta/\sqrt{\kappa+\eta}, & \text{if } E \notin [\lambda_-, \lambda_+] \\ \sqrt{\kappa+\eta}, & \text{if } E \in [\lambda_-, \lambda_+] \end{cases}. \tag{3.33}$$

We will use the following notion of stochastic domination, which was first introduced in [20] and subsequently used in many works on random matrix theory, such as [6, **?**, 21, 22, 44]. It simplifies the presentation of the results and their proofs by systematizing statements of the form "ξ is bounded by ζ with high probability up to a small power of N".

Definition 3.2.4 (Stochastic domination)**.** *(i) Let*

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

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

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

for large enough $N \ge N_0(\epsilon, D)$. Throughout this chapter the stochastic domination will always be uniform in all parameters that are not explicitly fixed (such as matrix indices, deterministic vectors, and spectral parameter $z \in \mathbf{D}$). Note that $N_0(\epsilon, D)$ may depend on quantities that are explicitly constant, such as d, C_1 and ϕ in Theorem 3.1.1.

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

(iii) We say that an event Ξ holds with high probability if $1 - \mathbf{1}(\Xi) \prec 0$.

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

Lemma 3.2.5 (Lemma 3.2 in [6]). *Let* ξ *and* ζ *be families of nonnegative random variables.*

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

$$\sum_{v \in V} \xi(u,v) \prec \sum_{v \in V} \zeta(u,v)$$

uniformly in u*.*

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

$$\xi_1(u)\xi_2(u) \prec \zeta_1(u)\zeta_2(u)$$

uniformly in $u \in U$*.*

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

$$\mathbb{E}\xi(u) \prec \Psi(u)$$

uniformly in u*.*

Definition 3.2.6 (Bounded support condition). *We say a family of random matrices* X *satisfy the bounded support condition with* q*, if*

$$\max_{i \in \mathcal{I}_1, \mu \in \mathcal{I}_2} |X_{i\mu}| \prec q. \tag{3.34}$$

Here $q \equiv q(N)$ *is deterministic and usually satisfies* $N^{-1/2} \leq q \leq N^{-\phi}$ *for some (small) constant* $\phi > 0$*. Whenever (3.34) holds, we say that* X *has support* q*.*

Remark 3.2.7. *If the entries of X satisfy (3.13), then X trivially satisfies the bounded support condition with $q = N^{-\phi}$. If we assume that $\sqrt{N}X_{i\mu}$ has arbitrarily high moments, i.e. for any $p \in \mathbb{N}$ there is a constant C_p such that*

$$\max_{i,\mu} \mathbb{E}|\sqrt{N}X_{i\mu}|^p \le C_p. \tag{3.35}$$

Then by Markov's inequality, X has support $N^{-1/2}$.

We define the deterministic limit

$$\Pi(z) := \begin{pmatrix} zm_{1c}(z)I_{M\times M} & 0 \\ 0 & m_{2c}(z)I_{N\times N} \end{pmatrix}, \tag{3.36}$$

and the control parameter

$$\Psi(z) := \sqrt{\frac{\operatorname{Im}(m_{1c} + m_{2c})}{N\eta}} + \frac{1}{N\eta}. \tag{3.37}$$

Note that by (3.33), we always have

$$\Psi \gtrsim N^{-1/2}, \quad \Psi^2 \lesssim (N\eta)^{-1}, \tag{3.38}$$

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

Theorem 3.2.8 (Local MP law)**.** *Let X be an $M \times N$ real random matrix whose entries are independent random variables satisfying (3.9), (3.10), (3.11) and the bounded support condition (3.34) with $q \le N^{-\phi}$ for some constant $\phi > 0$. Then the following estimates hold for all $z \in \mathbf{D}$:*

(1) the averaged local law:

$$|m_1(X, z) - m_{1c}(z)| \prec \frac{1}{N\eta}; \tag{3.39}$$

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

$$|\langle \mathbf{u}, G(X,z)\mathbf{v}\rangle - \langle \mathbf{u}, \Pi(z)\mathbf{v}\rangle| \prec q + \Psi(z); \tag{3.40}$$

(3) for all deterministic unit vector $\mathbf{v} \in \mathbb{C}^{\mathcal{I}_1}$*,*

$$|\langle \mathbf{v}, \mathcal{G}_1(X,z)\mathbf{v}\rangle - m_{1c}\rangle| \prec q^2 + \sqrt{\frac{1}{N\eta}}, \tag{3.41}$$

and

$$|\mathbb{E}\langle \mathbf{v}, \mathcal{G}_1(X,z)\mathbf{v}\rangle - m_{1c}(z)| \prec q^4 + \frac{1}{N\eta}. \tag{3.42}$$

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

The proof for Theorem 3.2.8 will be given in Section 3.4. Here we make some comments on the above estimates.

If we assume (3.1) (instead of (3.9) and (3.10)) and $q = N^{-1/2}$, then (3.39) and (3.40) have been proved in [6]. If we have (3.1) and $q \le N^{-\phi}$, then it was proved in Lemma 3.11 and Theorem 3.14 of [16] that the averaged local law (3.39) and the entrywise local law

$$\max_{a,b\in\mathcal{I}} |G_{ab}(X,z) - \Pi_{ab}(z)| \prec q + \Psi(z) \tag{3.43}$$

hold uniformly in $z \in \mathbf{D}$. With (3.43) and the moment assumption (3.11), one can repeat the arguments in [6, Section 5] or [68, Section 5] to get the isotropic local law (3.40). The main novelties of this Theorem are the bounds (3.41) and (3.42). The bound (3.41) is relatively easier to prove. In fact, if we only consider the upper left and lower right blocks of $G(X,z)$, we can get the following version of the entrywise law:

$$\max_{r=1,2}\max_{a,b\in\mathcal{I}_r} |G_{ab}(X,z) - \Pi_{ab}(z)| \prec q^2 + \sqrt{\frac{1}{N\eta}}, \tag{3.44}$$

which can be proved easily with (3.43) (see Appendix 3.6.1). Then with (3.44) and (3.11), we can apply the arguments in [6, Section 5] to conclude the isotropic local law (3.41) (see Appendix 3.6.2).

On the other hand, the improvement from (3.41) to (3.42) is more crucial, and is the main reason why we can improve the bound in [70] to the almost optimal one in (3.14). In fact, the leading order term of $\langle \mathbf{v}, \mathcal{G}_1 \mathbf{v}\rangle - m_{1c}$ vanishes after taking expectation, and hence leads to a bound that is one order smaller than the one in (3.41). The proof of (3.42) will be given in Sections 3.4.2-3.4.4, which constitutes the main novelty of this work.

Finally, if the variance assumption in (3.1) is relaxed to the one in (3.10), we can repeat the previous arguments to get the desired estimates (3.39)-(3.42). In fact, it is easy to check that the $O(N^{-2-c_0})$ term leads to a negligible error at each step, and the whole proof remains unchanged. The relaxation of the mean zero assumption in (3.1) to (3.9) is a little more involved, which will be handled with a centralization argument in Section 3.4.1.

If $q = N^{-1/4+\epsilon}$ for some sufficiently small constant $\epsilon > 0$, then (3.41) and (3.42) already give that

$$|\langle \mathbf{v}, \mathcal{G}_1(X,z)\mathbf{v}\rangle - m_{1c}\rangle| \prec \sqrt{\frac{1}{N\eta}}, \quad |\mathbb{E}\langle \mathbf{v}, \mathcal{G}_1(X,z)\mathbf{v}\rangle - m_{1c}(z)| \prec \frac{1}{N\eta},$$

which is sufficient to conclude Theorem 3.1.1. However, we observe that the above bound on $|\mathbb{E}\langle \mathbf{v}, \mathcal{G}_1(X,z)\mathbf{v}\rangle - m_{1c}(z)|$ is still valid under a much weaker support assumption. More specifically, we have the following theorem. Its proof will be given in Section 3.5. The main strategy is a resolvent comparison method that was developed in [46].

Theorem 3.2.9. *Let X be an $M \times N$ real random matrix satisfying the assumptions in*

Theorem 3.2.8. Then we have

$$|\mathbb{E}\langle \mathbf{v}, \mathcal{G}_1(X,z)\mathbf{v}\rangle - m_{1c}(z)| \prec \frac{1}{N\eta}, \tag{3.45}$$

uniformly in $z \in \mathbf{D}$ *and any deterministic unit vector* $\mathbf{v} \in \mathbb{C}^{\mathcal{I}_1}$.

We define the classical location γ_j of the j-th eigenvalue of XX^* as

$$\int_{\gamma_j}^{+\infty} \rho_{1c}(x)dx = \frac{j}{M}, \quad 1 \le j \le K,$$

where ρ_{1c} is defined in (3.3) and $K := \min\{M, N\}$. As a corollary of (3.39), we have the following rigidity of eigenvalues of XX^*. For its proof, one can refer to the arguments in [31, Section 5], [22, Section 7] and [54, Section 8].

Theorem 3.2.10 (Rigidity of eigenvalues)**.** *Suppose (3.39) holds and* $\lambda_- \ge c$ *for some constant* $c > 0$. *Then we have*

$$|\lambda_j(XX^*) - \gamma_j| \prec (\min\{j, K+1-j\})^{-1/3} N^{-2/3}, \quad 1 \le j \le K. \tag{3.46}$$

3.2.2 Resolvent estimates

In this subsection, we collect some useful identities from linear algebra and some simple estimates that follow from Theorem 3.2.8.

Definition 3.2.11 (Minors)**.** *For* $\mathbb{T} \subseteq \mathcal{I}$, *we define the minor* $H^{(\mathbb{T})} := (H_{ab} : a, b \in \mathcal{I}\backslash\mathbb{T})$ *obtained by removing all rows and columns of* H *indexed by* $a \in \mathbb{T}$. *Note that we keep the names of indices when defining* $H^{(\mathbb{T})}$, *i.e.* $(H^{(\mathbb{T})})_{ab} = \mathbf{1}_{\{a,b\notin\mathbb{T}\}} H_{ab}$. *Correspondingly, we define the Green function*

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

and the partial traces

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

We will abbreviate $(\{a\}) \equiv (a)$ *and* $(\{a,b\}) \equiv (ab)$ *in the proof.*

Lemma 3.2.12 (Resolvent identities). *(i) For* $i \in \mathcal{I}_1$ *and* $\mu \in \mathcal{I}_2$*, we have*

$$\frac{1}{G_{ii}} = -1 - \left(XG^{(i)}X^*\right)_{ii}, \quad \frac{1}{G_{\mu\mu}} = -z - \left(X^*G^{(\mu)}X\right)_{\mu\mu}. \tag{3.47}$$

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

$$G_{ij} = G_{ii}G_{jj}^{(i)}\left(XG^{(ij)}X^*\right)_{ij}, \tag{3.48}$$

and

$$G_{\mu\nu} = G_{\mu\mu}G_{\nu\nu}^{(\mu)}\left(X^*G^{(\mu\nu)}X\right)_{\mu\nu}. \tag{3.49}$$

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

$$G_{bc} = G_{bc}^{(a)} + \frac{G_{ba}G_{ac}}{G_{aa}}, \quad \frac{1}{G_{bb}} = \frac{1}{G_{bb}^{(a)}} - \frac{G_{ba}G_{ab}}{G_{bb}G_{bb}^{(a)}G_{aa}}. \tag{3.50}$$

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

Proof. The above identities can be proved using Schur complement formula. The reader can refer to e.g. [6, Lemmas 3.6 and 3.8] or [44, Lemma 4.4]. □

Lemma 3.2.13. *Suppose* $\Phi(z)$ *is a deterministic function on* $\mathbf{D}$ *satisfying* $N^{-1/2} \leq \Phi(z) \leq N^{-c}$ *for some constant* $c > 0$*. Suppose* $|G_{ab}(z) - \Pi_{ab}(z)| \prec \Phi(z)$ *uniformly in* $z \in \mathbf{D}$*. Fix an* $l \in \mathbb{N}$*. Then for any* $\mathbb{T} \subseteq \mathcal{I}$ *with* $|\mathbb{T}| \leq l$*, we have*

$$\left|G_{ab}(z) - G_{ab}^{(\mathbb{T})}(z)\right| \prec \Phi^2(z), \quad a, b \in \mathcal{I}\setminus\mathbb{T}, \tag{3.51}$$

and

$$\left|m_1(z) - m_1^{(\mathbb{T})}(z)\right| + \left|m_2(z) - m_2^{(\mathbb{T})}(z)\right| \prec \Phi^2(z), \tag{3.52}$$

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

Proof. The bound (3.51) can be proved by repeatedly applying the first resolvent expansion in (3.50) with respect to the indices in $\mathbb{T}$ and using the entrywise local law. The bound (3.52) is a trivial consequence of (3.51). $\square$

For $\mathbf{v}, \mathbf{w} \in \mathbb{C}^{\mathcal{I}}$, $a \in \mathcal{I}$ and any $\mathcal{I} \times \mathcal{I}$ matrix A, we abbreviate

$$A_{\mathbf{vw}} := \langle \mathbf{v}, A\mathbf{w} \rangle, \quad A_{\mathbf{v}a} := \langle \mathbf{v}, A\mathbf{e}_a \rangle, \quad A_{a\mathbf{w}} := \langle \mathbf{e}_a, A\mathbf{w} \rangle, \tag{3.53}$$

where $\mathbf{e}_a$ denotes the standard unit vector in the coordinate direction a. We shall call them the generalized matrix entries. We sometimes identify vectors $\mathbf{v} \in \mathbb{C}^{\mathcal{I}_1}$ and $\mathbf{w} \in \mathbb{C}^{\mathcal{I}_2}$ with their natural embeddings $\begin{pmatrix} \mathbf{v} \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ \mathbf{w} \end{pmatrix}$ in $\mathbb{C}^{\mathcal{I}}$. The exact meanings will be clear from the context.

Lemma 3.2.14. *For any $M \times N$ matrix Y, the following estimates hold for $G \equiv G(Y, z)$ and any $z \in \mathbf{D}$. There exists a constant $C > 0$ such that*

$$\|G\| \le C\eta^{-1}, \quad \|\partial_z G\| \le C\eta^{-2}. \tag{3.54}$$

Moreover, for $\mathbf{v} \in \mathbb{C}^{\mathcal{I}_1}$ and $\mathbf{w} \in \mathbb{C}^{\mathcal{I}_2}$, we have the following identities

$$\sum_{i \in \mathcal{I}_1} |G_{\mathbf{v}i}|^2 = \sum_{i \in \mathcal{I}_1} |G_{i\mathbf{v}}|^2 = \frac{|z|^2}{\eta} \operatorname{Im}\left(\frac{G_{\mathbf{vv}}}{z}\right) \tag{3.55}$$

$$\sum_{i \in \mathcal{I}_1} |G_{\mathbf{w}i}|^2 = \sum_{i \in \mathcal{I}_1} |G_{i\mathbf{w}}|^2 = G_{\mathbf{ww}} + \frac{\bar{z}}{\eta} \operatorname{Im} G_{\mathbf{ww}}, \tag{3.56}$$

$$\sum_{\mu \in \mathcal{I}_2} |G_{\mathbf{w}\mu}|^2 = \sum_{\mu \in \mathcal{I}_2} |G_{\mu\mathbf{w}}|^2 = \frac{\operatorname{Im} G_{\mathbf{ww}}}{\eta}, \tag{3.57}$$

$$\sum_{\mu\in\mathcal{I}_2} |G_{\mathbf{v}\mu}|^2 = \sum_{\mu\in\mathcal{I}_2} |G_{\mu\mathbf{v}}|^2 = \frac{G_{\mathbf{v}\mathbf{v}}}{z} + \frac{\bar{z}}{\eta}\mathrm{Im}\left(\frac{G_{\mathbf{v}\mathbf{v}}}{z}\right). \tag{3.58}$$

These estimates remain true for $G^{(\mathbb{T})}$ instead of G for any $\mathbb{T} \subseteq \mathcal{I}$.

Proof. These estimates and identities can be proved through simple calculations using (3.30) and the spectral decomposition of G. The reader can also refer to, for example, [44, Lemma 4.6], [68, Lemma 3.5] and [16, Lemma A.3]. $\square$

Suppose (3.40) holds. Then using (3.55)-(3.58) and (3.51), it is easy to verify that

$$\max\left\{ \sum_i \left|G^{(\mathbb{T})}_{\mathbf{v}i}\right|^2, \sum_i \left|G^{(\mathbb{T})}_{i\mathbf{v}}\right|^2, \sum_\mu \left|G^{(\mathbb{T})}_{\mathbf{v}\mu}\right|^2, \sum_\mu \left|G^{(\mathbb{T})}_{\mu\mathbf{v}}\right|^2 \right\} \prec \eta^{-1}, \tag{3.59}$$

for any deterministic unit vector $\mathbf{v} \in \mathbb{C}^{\mathcal{I}}$ and $\mathbb{T} \subseteq \mathcal{I}$ with fixed length.

3.3 Proof of Theorem 3.1.1

In this section, we prove Theorem 3.1.1 using Theorems 3.2.8-3.2.10. The following arguments have been used in previous papers to control the Kolmogorov distance between the ESD of a random matrix and the limiting law. For example, the reader can refer to [29, Lemma 6.1] and [54, Lemma 8.1]. By the remark below (3.31), we can choose the constant $\zeta > 0$ such that $\lambda_-/2 > \zeta$ for all sufficiently large N.

Proof of (3.14). The key inputs of the proof are the bounds (3.45) and (3.46). Suppose $\langle \mathbf{v}, \mathcal{G}_1(X,z)\mathbf{v}\rangle$ is the Stieltjes transform of $\hat{\rho}_{\mathbf{v}}$. Then we define

$$\hat{n}_{\mathbf{v}}(E) := \int \mathbf{1}_{[0,E]}(x)\hat{\rho}_{\mathbf{v}}dx, \quad n_c(E) := \int \mathbf{1}_{[0,E]}(x)\rho_{1c}dx, \tag{3.60}$$

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

$$\|\mathbb{E}F^{(M)}_{XX^*}(\mathbf{v},\cdot) - F_{d_N}(\cdot)\| = \sup_E |n_{\mathbf{v}}(E) - n_c(E)|.$$

For simplicity, we denote $\Delta\rho := \rho_{\mathbf{v}} - \rho_{1c}$ and its Stieltjes transform by

$$\Delta m(z) := \mathbb{E}\langle \mathbf{v}, \mathcal{G}_1(X,z)\mathbf{v}\rangle - m_{1c}(z).$$

Let $\chi(y)$ be a smooth cutoff function with support in $[-1,1]$, with $\chi(y)=1$ for $|y| \le 1/2$ and with bounded derivatives. Fix $\eta_0 = N^{-1+\zeta}$ and $3\lambda_-/4 \le E_1 < E_2 \le 3\lambda_+/2$. Let $f \equiv f_{E_1,E_2,\eta_0}$ be a smooth function supported in $[E_1-\eta_0, E_2+\eta_0]$ such that $f(x)=1$ if $x \in [E_1+\eta_0, E_2-\eta_0]$, and $|f'| \le C\eta_0^{-1}$, $|f''| \le C\eta_0^{-2}$ if $|x-E_i| \le \eta_0$. Using the Helffer-Sjöstrand calculus (see e.g. [14]), we have

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

Then we obtain that

$$\begin{aligned}
&\left|\int f(E)\Delta\rho(E)dE\right| \\
&\le C\int_{\mathbb{R}^2} \left(|f(x)| + |y||f'(x)|\right)|\chi'(y)||\Delta m(x+iy)|dxdy && (3.61)\\
&+ C\sum_i \left|\int_{|y|\le\eta_0}\int_{|x-E_i|\le\eta_0} yf''(x)\chi(y)\operatorname{Im}\Delta m(x+iy)dxdy\right| && (3.62)\\
&+ C\sum_i \left|\int_{|y|\ge\eta_0}\int_{|x-E_i|\le\eta_0} yf''(x)\chi(y)\operatorname{Im}\Delta m(x+iy)dxdy\right|. && (3.63)
\end{aligned}$$

By (3.45) with $\eta=\eta_0$, we have

$$\eta_0 \operatorname{Im}\mathbb{E}\langle \mathbf{v}, \mathcal{G}_1(X, E+i\eta_0)\mathbf{v}\rangle \prec N^{-1+\zeta}. \tag{3.64}$$

Since $\eta\operatorname{Im}\mathbb{E}\langle \mathbf{v}, \mathcal{G}_1(X,E+i\eta)\mathbf{v}\rangle$ and $\eta\operatorname{Im} m_{1c}(E+i\eta)$ are increasing with η, we obtain that

$$\eta\left|\operatorname{Im}\Delta m(E+i\eta)\right| \prec N^{-1+\zeta} \quad \text{for all } 0 \le \eta \le \eta_0. \tag{3.65}$$

Moreover, since $G(X,z)^* = G(X,\bar{z})$, the estimates (3.45) and (3.65) also hold for $z \in \mathbb{C}_-$.

Now we bound the terms (3.61), (3.62) and (3.63). Using (3.45) and that the support of χ' is in $1 \geq |y| \geq 1/2$, the term (3.61) is estimated by

$$\int_{\mathbb{R}^2} (|f(x)| + |y||f'(x)|) |\chi'(y)||\Delta m(x+iy)| dx dy \prec N^{-1}. \tag{3.66}$$

Using $|f''| \leq C\eta_0^{-2}$ and (3.65), we can bound the terms in (3.62) by

$$\left| \int_{|y|\leq \eta_0} \int_{|x-E_i|\leq \eta_0} y f''(x)\chi(y) \operatorname{Im} \Delta m(x+iy) dx dy \right| \prec N^{-1+\zeta}. \tag{3.67}$$

Finally, we integrate the term (3.63) by parts first in x, and then in y (and use the Cauchy-Riemann equation $\partial \mathrm{Im}(\Delta m)/\partial x = -\partial \mathrm{Re}(\Delta m)/\partial y$) to get that

$$\begin{aligned}
&\int_{y\geq \eta_0} \int_{|x-E_i|\leq \eta_0} y f''(x)\chi(y) \operatorname{Im} \Delta m(x+iy) dx dy \\
&= \int_{y\geq \eta_0} \int_{|x-E_i|\leq \eta_0} y f'(x)\chi(y) \frac{\partial \operatorname{Re} \Delta m(x+iy)}{\partial y} dx dy \\
&= -\int_{|x-E_i|\leq \eta_0} \eta_0 \chi(\eta_0) f'(x) \operatorname{Re} \Delta m(x+i\eta_0) dx \qquad (3.68)\\
&\quad - \int_{y\geq \eta_0} \int_{|x-E_i|\leq \eta_0} (y\chi'(y) + \chi(y)) f'(x) \operatorname{Re} \Delta m(x+iy) dx dy. \qquad (3.69)
\end{aligned}$$

We bound the term in (3.68) by $O_\prec(N^{-1})$ using (3.45) and $|f'| \leq C\eta_0^{-1}$. The first term in (3.69) can be estimated by $O_\prec(N^{-1})$ as in (3.66). For the second term in (3.69), we again use (3.45) and $|f'| \leq C\eta_0^{-1}$ to get that

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

Combining the above estimates, we obtain that

$$\left| \int_{y\geq \eta_0} \int_{|x-E_i|\leq \eta_0} y f''(x)\chi(y) \operatorname{Im} \Delta m(x+iy) dx dy \right| \prec N^{-1}.$$

Obviously, the same estimate also holds for the $y \leq -\eta_0$ part. Together with (3.66) and (3.67), we conclude that

$$\left| \int f(E) \Delta\rho(E) dE \right| \prec N^{-1+\zeta}. \tag{3.70}$$

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

$$\begin{aligned}\hat{n}_{\mathbf{v}}(E+\eta_0) - \hat{n}_{\mathbf{v}}(E-\eta_0) &= \sum_{\lambda_k \in (E-\eta_0, E+\eta_0]} |\langle \xi_k, \mathbf{v}\rangle|^2 \\ &\le 2\eta_0 \sum_{k=1}^{M} \frac{|\langle \xi_k, \mathbf{v}\rangle|^2 \eta_0}{(\lambda_k - E)^2 + \eta_0^2} = 2\eta_0 \mathrm{Im}\, \langle \mathbf{v}, \mathcal{G}_1(X, E+i\eta_0)\mathbf{v}\rangle,\end{aligned} \tag{3.71}$$

where we used the spectral decomposition

$$\mathcal{G}_1(X, E+i\eta) = \sum_{k=1}^{M} \frac{\xi_k \xi_k^*}{\lambda_k - E - i\eta},$$

which follows from (3.7). Then by (3.64) and Lemma 3.2.5, we get that

$$n_{\mathbf{v}}(E+\eta_0) - n_{\mathbf{v}}(E-\eta_0) \prec N^{-1+\zeta}. \tag{3.72}$$

On the other hand, we trivially have

$$n_c(E+\eta_0) - n_c(E-\eta_0) \le C\eta_0 = CN^{-1+\zeta} \tag{3.73}$$

since $\rho_{1c}(x)$ is bounded for x away from 0.

Now we set $E_2 = 3\lambda_+/2$. With (3.70), (3.72) and (3.73), we get that for any $E \in [3\lambda_-/4, E_2]$,

$$|(n_{\mathbf{v}}(E_2) - n_{\mathbf{v}}(E)) - (n_c(E_2) - n_c(E))| \prec N^{-1+\zeta}. \tag{3.74}$$

Note that by (3.46), the eigenvalues of XX^* are inside $\{0\} \cup [3\lambda_-/4, E_2]$ with high probability. Hence we have with high probability,

$$\hat{n}_{\mathbf{v}}(E_2) = n_c(E_2) = 1, \quad \hat{n}_{\mathbf{v}}(3\lambda_-/4) = \hat{n}_{\mathbf{v}}(0). \tag{3.75}$$

Together with (3.74), we get that

$$\sup_{E \ge 0} |n_{\mathbf{v}}(E) - n_c(E)| \prec N^{-1+\zeta}. \tag{3.76}$$

This concludes (3.14) since ζ can be arbitrarily small. □

Proof of (3.15). The proof for (3.15) is similar except that we shall use the estimate (3.41) instead of (3.45). By (3.41), we have

$$|\langle \mathbf{v}, \mathcal{G}_1(X,z)\mathbf{v}\rangle - m_{1c}(z)| \prec N^{-2\phi} + (N\eta)^{-1/2} \tag{3.77}$$

uniformly in $z \in \mathbf{D}$. Then we would like to bound (recall (3.60))

$$\|F^{(M)}_{XX^*}(\mathbf{v},\cdot) - F_{d_N}(\cdot)\| = \sup_E |\hat{n}_{\mathbf{v}}(E) - n_c(E)|,$$

where $\hat{n}_{\mathbf{v}}$ is defined in (3.60). We denote

$$\Delta\hat{\rho} := \hat{\rho}_{\mathbf{v}} - \rho_{1c}, \quad \Delta\hat{m} := \langle \mathbf{v}, \mathcal{G}_1(X,z)\mathbf{v}\rangle - m_{1c}(z).$$

Then for f_{E_1,E_2,η_0} defined in the previous proof, we can repeat the Helffer-Sjöstrand argument with the estimate (3.77) to get that

$$\sup_{E_1,E_2}\left|\int f_{E_1,E_2,\eta_0}(E)\Delta\hat{\rho}(E)dE\right| \prec N^{-2\phi} + N^{-1/2}, \tag{3.78}$$

which, together with (3.71) and (3.75), implies

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

This concludes (3.15) by the Definition 3.2.4. □

3.4 Proof of Theorem 3.2.8

3.4.1 Centralization

For X satisfying the assumptions in Theorem 3.2.8, we write $X = X_1 + B$, where $X_1 := X - \mathbb{E}X$ is a random matrix satisfying (3.10), (3.11) and

$$\mathbb{E}(X_1)_{i\mu} = 0, \quad i \in \mathcal{I}_1, \mu \in \mathcal{I}_2, \tag{3.79}$$

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

$$\max_{i,\mu} |B_{i\mu}| \le C_0 N^{-2-c_0}. \tag{3.80}$$

Claim 3.4.1. *If Theorem 3.2.8 holds for X_1, then it also holds for X.*

Proof. For $z \in \mathbf{D}$, we have

$$G(X,z) := \begin{pmatrix} -I_{M\times M} & X_1 + B \\ X_1^* + B^* & -zI_{N\times N} \end{pmatrix}^{-1} = \left(G_1^{-1} + V\right)^{-1}, \tag{3.81}$$

where we abbreviate $G_1(z) := G(X_1, z)$ and $V := \begin{pmatrix} 0 & B \\ B^* & 0 \end{pmatrix}$. By our assumption, (3.40) holds for G_1. Then we expand G using the resolvent expansion

$$G = G_1 - G_1VG_1 + (G_1V)^2G_1 - (G_1V)^3G. \tag{3.82}$$

For any unit vectors $\mathbf{v}, \mathbf{u} \in \mathbb{C}^{\mathcal{I}}$, we have

$$\begin{aligned} |\langle \mathbf{v}, G_1VG_1\mathbf{u}\rangle| &\le \sum_{b\in\mathcal{I}} \Big| \sum_{a\in\mathcal{I}} (G_1)_{\mathbf{v}a} V_{ab} \Big| |(G_1)_{b\mathbf{u}}| \\ &\prec \max_b \Big(\sum_{a\in\mathcal{I}} |V_{ab}|^2 \Big)^{1/2} \sum_{b\in\mathcal{I}} |(G_1)_{b\mathbf{u}}| \\ &\prec N^{-1-c_0} \Big(\sum_{b\in\mathcal{I}} |(G_1)_{b\mathbf{u}}|^2 \Big)^{1/2} \prec N^{-1-c_0}\eta^{-1/2}, \end{aligned} \tag{3.83}$$

where in the second step we used (3.40) for G_1, in the third step the Cauchy-Schwarz inequality and (3.80), and in the last step (3.59). With a similar argument, we obtain that

$$\left|\langle \mathbf{v}, (G_1V)^2G_1\mathbf{u}\rangle\right| \prec N^{-2-2c_0}\eta^{-1}. \tag{3.84}$$

Combining this estimate with the rough bound (3.54) for G, we get that

$$\begin{aligned}\left|\langle\mathbf{v},(G_1V)^3G\mathbf{u}\rangle\right| &= \Big|\sum_{a,b}\left((G_1V)^2G_1\right)_{\mathbf{v}a}V_{ab}G_{b\mathbf{u}}\Big| \\ &\prec \left(N^{-2-2c_0}\eta^{-1}\right)\eta^{-1}\sum_a\Big(\sum_b |V_{ab}|^2\Big)^{1/2} \le CN^{-3/2-3c_0}\eta^{-1},\end{aligned} \tag{3.85}$$

where we used $\eta \ge N^{-1}$ for $z\in\mathbf{D}$. Plugging the estimates (3.83)-(3.85) into (3.82), we conclude that

$$|\langle\mathbf{v},G\mathbf{u}\rangle - \langle\mathbf{v},G_1\mathbf{u}\rangle| \prec N^{-1-\omega/4}\eta^{-1/2} \le (N\eta)^{-1} \tag{3.86}$$

for all deterministic unit vectors $\mathbf{v},\mathbf{u}\in\mathbb{C}^{\mathcal{I}}$. We can then easily conclude the claim with this estimate. □

Thus in the following proof, we can assume that the entries of X are centered without loss of generality. According to the comments below Theorem 3.2.8, we can repeat the proof in [16] to get (3.39) and the entrywise local law (3.43). Then combining (3.43), the moment assumption (3.11) and the arguments in [6, Section 5], we can obtain (3.40) (see also the proof for Lemma 3.4.3 in Appendix 3.6.2). The bound (3.41) follows from Claim 3.4.1 and the next two lemmas.

Lemma 3.4.2. *Let X be an $M\times N$ real random matrix whose entries are independent random variables satisfying (3.79), (3.10), (3.11) and the bounded support condition (3.34) with $q\le N^{-\phi}$ for some constant $\phi>0$. Then if (3.39) and (3.43) holds, the local law (3.44) also holds for all $z\in\mathbf{D}$.*

Lemma 3.4.3. *Suppose the assumptions in Lemma 3.4.2 hold for X. Suppose $\Phi(z)$ is a deterministic function on $\mathbf{D}$ satisfying $c_0(N^{-1/2}+q^2)\le\Phi(z)\le N^{-c_0}$ for some constant $c_0>0$. If we have*

$$\max_{a,b\in\mathcal{I}}|G_{ab}(z)-\Pi_{ab}(z)|^2 \prec \Phi, \quad \max_{r=1,2}\max_{a,b\in\mathcal{I}_r}|G_{ab}(z)-\Pi_{ab}(z)| \prec \Phi \tag{3.87}$$

for all $z \in \mathbf{D}$*, then*

$$|\langle \mathbf{v}, \mathcal{G}_1(X,z)\mathbf{v}\rangle - m_{1c}\rangle| \prec \Phi(z) \tag{3.88}$$

uniformly in all $z \in \mathbf{D}$ *and all deterministic unit vector* $\mathbf{v} \in \mathbb{C}^{\mathcal{I}_1}$*.*

We will give the proof of Lemma 3.4.2 and Lemma 3.4.3 in appendix. In the rest of this section, we focus on proving our main estimate (3.42). For simplicity, we denote $\Phi := q^2 + (N\eta)^{-1/2}$ in the proof below. Also, by Claim 3.4.1, we can assume that the entries of X are centered.

3.4.2 Sketch of the proof for (3.42)

We want to estimate $|\mathbb{E}\langle \mathbf{v}, \mathcal{G}_1\mathbf{v}\rangle - m_{1c}|$ for any deterministic unit vector $\mathbf{v} \in \mathbb{C}^{\mathcal{I}_1}$. Note that (3.41) gives the a priori bound

$$\Big|\sum_{i,j} \bar{v}_i v_j \mathbb{E}\left(\mathcal{G}_1\right)_{ij} - m_{1c}\Big| \prec \Phi.$$

We will show that after taking expectation, the leading order term in $(\mathcal{G}_1)_{ij} - m_{1c}\delta_{ij}$ vanishes and gives the improved estimate (3.42). We deal with the diagonal and off-diagonal parts separately:

$$\sum_i |v_i|^2 \left[\mathbb{E}(\mathcal{G}_1)_{ii} - m_{1c}(z)\right], \quad \sum_{i\neq j} \bar{v}_i v_j \mathbb{E}\left(\mathcal{G}_1\right)_{ij}.$$

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

$$Z_i^{(\mathbb{T})} := (1-\mathbb{E}_i)(G^{(\mathbb{T})})_{ii}^{-1} = m_2^{(\mathbb{T}i)} - (XG^{(\mathbb{T}i)}X^*)_{ii}, \quad i \notin \mathbb{T}, \tag{3.89}$$

where $\mathbb{E}_i[\cdot] := \mathbb{E}[\cdot|H^{(i)}]$, i.e. it is the partial expectation in the randomness of the i-th row and column of H, and we used (3.47) in the second step. If $\mathbb{T} = \emptyset$, we abbreviate

$Z_i \equiv Z_i^{(\emptyset)}$. By (3.148), we have $|Z_i| \prec \Phi$. Then using (3.47) we get that

$$\begin{aligned}\mathbb{E}G_{ii} - zm_{1c} &= \mathbb{E}\frac{1}{-1-m_{2c}-(m_2^{(i)}-m_{2c})+Z_i} - zm_{1c} \\ &= \frac{1}{-1-m_{2c}} - zm_{1c} - \frac{1}{(1+m_{2c})^2}\mathbb{E}Z_i + O_\prec\left(\Phi^2 + \frac{1}{N\eta}\right) = O_\prec\left(\Phi^2\right),\end{aligned}$$

where in the second step we used the bound for Z_i, (3.39) and (3.52), and in the third step we used (3.4), (3.6) and $\mathbb{E}Z_i = 0$. Thus we can bound the diagonal part by

$$\sum_i |v_i|^2 \left[\mathbb{E}(\mathcal{G}_1)_{ii} - m_{1c}(z)\right] = \frac{1}{z}\sum_i |v_i|^2[\mathbb{E}G_{ii} - zm_{1c}(z)] \prec q^4 + \frac{1}{N\eta} \tag{3.90}$$

for $z \in \mathbf{D}$ (recall that $|z| \geq E \sim 1$ by (3.31)).

For the off-diagonal part, we claim that for all $i \neq j \in \mathcal{I}_1$,

$$\left|\mathbb{E}\left(\mathcal{G}_1\right)_{ij}\right| \prec N^{-1}\Phi^2. \tag{3.91}$$

Then using (3.91) and $\|\mathbf{v}\|_1 \leq \sqrt{M}$, we obtain that

$$\Big|\sum_{i\neq j}\bar{v}_i v_j \mathbb{E}\left(\mathcal{G}_1\right)_{ij}\Big| \prec \|\mathbf{v}\|_1^2 N^{-1}\Phi^2 \leq C\left(q^4 + \frac{1}{N\eta}\right).$$

Together with (3.90), this concludes (3.42).

To prove (3.91), we follow the arguments in [6, Section 5] and [68, Section 5]. We illustrate the basic idea with some simplified calculations. Using the resolvent identities (3.48) and (3.50), we get

$$\begin{aligned}\mathbb{E}G_{ij} &= \mathbb{E}G_{ii}G_{jj}^{(i)}\left(XG^{(ij)}X^*\right)_{ij} \\ &= \mathbb{E}G_{ii}^{(j)}G_{jj}^{(i)}\left(XG^{(ij)}X^*\right)_{ij} + \mathbb{E}\frac{G_{ij}G_{ji}}{G_{jj}}G_{jj}^{(i)}\left(XG^{(ij)}X^*\right)_{ij}.\end{aligned} \tag{3.92}$$

We now focus on the first term. Applying (3.47) gives that

$$\mathbb{E}G_{ii}^{(j)}G_{jj}^{(i)}\left(XG^{(ij)}X^*\right)_{ij} = \mathbb{E}\frac{\left(XG^{(ij)}X^*\right)_{ij}}{[1+(XG^{(ij)}X^*)_{ii}]\,[1+(XG^{(ij)}X^*)_{jj}]}$$

$$= \mathbb{E}\frac{\left(XG^{(ij)}X^*\right)_{ij}}{[(1+m_{2c})-\epsilon_i]\,[(1+m_{2c})-\epsilon_j]}. \tag{3.93}$$

where we have $|(1+m_{2c})^{-1}| = |zm_{1c}| \sim 1$ and

$$\epsilon_i := m_{2c} - (XG^{(ij)}X^*)_{ii} = m_{2c} - m_2^{(ij)} + Z_i^{(j)} \prec \Phi \tag{3.94}$$

by (3.41), (3.52) (with $\Phi = q + \Psi$) and (3.148). We now expand the fractions in (3.93) in order to take the expectation. Note that the $G^{(ij)}$ entries are independent of the X entries in the i, j-th rows and columns. Thus to attain a nonzero expectation, each X entry must appear at least twice in the expression. Due to this reason, the leading and next-to-leading order terms in the expansion vanish. The "real" leading order term is

$$\begin{aligned}
\mathbb{E}\frac{\epsilon_i\epsilon_j\left(XG^{(ij)}X^*\right)_{ij}}{(1+m_{2c})^4} &= \frac{1}{(1+m_{2c})^4}\mathbb{E}(XG^{(ij)}X^*)_{ii}(XG^{(ij)}X^*)_{jj}(XG^{(ij)}X^*)_{ij} \\
&= \frac{1}{(1+m_{2c})^4}\sum_{\mu,\nu}\frac{C_{\mu,\nu}}{N^3}\mathbb{E}G_{\mu\mu}^{(ij)}G_{\nu\nu}^{(ij)}G_{\mu\nu}^{(ij)} \\
&= \frac{1}{(1+m_{2c})^4}\sum_{\mu\neq\nu}\frac{C_{\mu,\nu}}{N^3}m_{2c}^2\mathbb{E}G_{\mu\nu}^{(ij)} + O_\prec(N^{-1}\Phi^2),
\end{aligned} \tag{3.95}$$

where the constants $C_{\mu,\nu}$ depend on the moments of $X_{i\mu}$ and $X_{j\mu}$ (recall (3.11)). Here in the last step, we used $|G_{\mu\mu}^{(ij)} - m_{2c}| \prec \Phi$ (by (3.41) and (3.51)), and bounded the $\mu = \nu$ terms by $O_\prec(N^{-2}) = O_\prec(N^{-1}\Phi^2)$. Now applying (3.49) to $G_{\mu\nu}^{(ij)}$, we get that

$$\begin{aligned}
\mathbb{E}G_{\mu\nu}^{(ij)} &= \mathbb{E}G_{\mu\mu}^{(ij)}G_{\nu\nu}^{(ij\mu)}\left(X^*G^{(ij\mu\nu)}X\right)_{\mu\nu} \\
&= \mathbb{E}m_{2c}^2\left(X^*G^{(ij\mu\nu)}X\right)_{\mu\nu} + O_\prec(\Phi^2) = O_\prec(\Phi^2),
\end{aligned} \tag{3.96}$$

where in the second step we used $|G_{\mu\mu}^{(ij)} - m_{2c}| + |G_{\nu\nu}^{(ij\mu)} - m_{2c}| \prec \Phi$ and

$$\left(X^*G^{(ij\mu\nu)}X\right)_{\mu\nu} = G_{\mu\nu}^{(ij)}\left(G_{\mu\mu}^{(ij)}G_{\nu\nu}^{(ij\mu)}\right)^{-1} \prec \Phi,$$

which follow easily from (3.41) and (3.51), and in the last step the leading order term vanishes since the two X entries are independent for $\mu \neq \nu$. Then by (3.96), the terms in (3.95) are bounded by $O_\prec(N^{-1}\Phi^2)$.

In general, after the expansion of the two fractions in (3.93), we get a summation of terms of the form

$$A_{m,n} := \mathbb{E}\epsilon_i^m \epsilon_j^n (XG^{(ij)}X^*)_{ij}, \quad i \neq j,$$

up to some constant coefficients of order 1. Since $|\epsilon_{i,j}| \prec \Phi \lesssim N^{-\zeta/2}$ for $z \in \mathbf{D}$ (we can take ζ small enough such that $N^{-\zeta/2} \geq q^2$), we only need to include the terms with $m + n \leq 2 + 2/\zeta$ and the tail will be smaller than $N^{-1}\Phi^2$. Note that in $A_{m,n}$, the X_{i*} entries, X_{j*} entries and $G^{(ij)}$ entries are mutually independent. Moreover, both the number of X_{i*} entries and the number of X_{j*} entries are odd. Thus to attain a nonzero expectation, we must pair the X entries such that there are two products of the forms $X_{i\mu}^{n_1}$ and $X_{j\nu}^{n_2}$ for some $n_1, n_2 \geq 3$. As a result, we lose $(n_1 - 2)/2 + (n_2 - 2)/2 \geq 1$ free indices, which contributes an N^{-1} factor. On the other hand, for the product of G entries, we have three cases: (1) if there are at least 2 off-diagonal G entries, then we bound them with $O_\prec(\Phi^2)$; (2) if there is only 1 off-diagonal G entry, then we can use the trick in (3.95) and the bound (3.96); (3) if there is no off-diagonal G entry, then we lose one more free index and get an extra N^{-1} factor. This gives the estimate (3.91) for the term in (3.93).

For the second term in (3.92), we again use (3.47), (3.48) and (3.50) to expand the G_{ij}, G_{ji} and G_{jj}^{-1} entries. Our goal is to expand all the G entries into polynomials of the terms

$$S_{kl} := (XG^{(ij)}X^*)_{kl}, \quad k, l \in \{i, j\}, \tag{3.97}$$

so that the X entries and $G^{(ij)}$ entries are independent in the resulting expression. In particular, the *maximally expanded* terms (see (3.98)) can be expanded into S_{kl} variables directly through (3.47) and (3.48). However, *non-maximally expanded* terms are also created along the expansions in (3.48) and (3.50). Then we need to further expand these newly appeared terms. In general, this process will not terminate. However, we will show in Lemma 3.4.7 that after sufficiently many expansions, the resulting expression either has enough off-diagonal terms, or is maximally expanded. In the former case, it suffices to bound each off-diagonal term by $O_{\prec}(\Phi)$. In the latter case, the expression will only consist of S_{kl} variables. Following the argument in the previous paragraph, the expectation over the X entries produces an N^{-1} factor, while the expectation over the G entries produces a Φ^2 factor.

In the rest of this section, we will give a rigorous proof based on the above arguments.

3.4.3 Resolvent expansion

To perform the resolvent expansion in a systematic way, we introduce the following notions of *string* and *string operator*. Recall the definition of S_{kl} in (3.97).

Definition 3.4.4 (Strings)**.** *Let* $\mathfrak{A}$ *be the alphabet containing all symbols that will appear during the expansion:*

$$\mathfrak{A} = \left\{G_{kl}, G_{kk}^{-1}, S_{kl} \text{ with } k,l \in \{i,j\}\right\} \cup \left\{G_{ii}^{(j)}, G_{jj}^{(i)}, (G_{ii}^{(j)})^{-1}, (G_{jj}^{(i)})^{-1}\right\}.$$

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

Remark 3.4.5. *It is important to distinguish the difference between a string* $\mathbf{s}$ *and its value* $[\![\mathbf{s}]\!]$*. For example, "*G_{ij}*" and "*$G_{ii}G_{jj}^{(i)}S_{ij}$*" are different strings, but they represent the same random variable by (3.48).*

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

$$\mathfrak{A}_{\max} = \left\{ G_{ij}, G_{ji}, G_{ii}^{(j)}, G_{jj}^{(i)}, (G_{ii}^{(j)})^{-1}, (G_{jj}^{(i)})^{-1}, S_{ii}, S_{jj}, S_{ij}, S_{ji} \right\}. \tag{3.98}$$

A string $\mathbf{s}$ is said to be maximally expanded if all of its symbols are in $\mathfrak{A}_{\max}$. We shall call $G_{ij}, G_{ji}, S_{ij}, S_{ji}$ the *off-diagonal* symbols and all the other symbols in $\mathfrak{A}$ *diagonal*. By the local law (3.41) and (3.51), we have $[\![\mathbf{a}_o]\!] \prec \Phi$ if $\mathbf{a}_o$ is an off-diagonal symbol (note that $S_{ij} = G_{ij}/(G_{ii}G_{jj}^{(i)}) \prec \Phi$ by (3.48)) and $[\![\mathbf{a}_o]\!] \prec 1$ if $\mathbf{a}_d$ is a diagonal symbol. We use $\mathcal{F}_{n\text{-}max}(\mathbf{s})$ and $\mathcal{F}_{\text{off}}(\mathbf{s})$ to denote the number of non-maximally expanded symbols and the number of off-diagonal symbols, respectively.

Definition 3.4.6 (String operators)**.** *Let $k \neq l \in \{i, j\}$.*

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

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

(iii) Define an operator ρ acting on a string $\mathbf{s}$ in the following sense. Replace each G_{kl} in $\mathbf{s}$ with $G_{kk}G_{ll}^{(k)}S_{kl}$.

By Lemma 3.2.12, it is clear that for any string $\mathbf{s}$,

$$[\![\tau_0(\mathbf{s})]\!] + [\![\tau_1(\mathbf{s})]\!] = [\![\mathbf{s}]\!], \quad [\![\rho(\mathbf{s})]\!] = [\![\mathbf{s}]\!]. \tag{3.99}$$

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

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

Then by (3.99) we have

$$\sum_{|w|=m} [\![\mathbf{s}_w]\!] = [\![\mathbf{s}]\!], \tag{3.100}$$

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

Lemma 3.4.7. *Consider the string* $\mathbf{s} =$ "$G_{ii}G_{jj}^{(i)}S_{ij}$"*. Let* w *be any binary sequence with* $|w| = 4l_0$ *and such that* $\mathbf{s}_w \neq \emptyset$*. Then either* $\mathcal{F}_{\text{off}}(\mathbf{s}_w) \geq 2l_0$ *or* $\mathbf{s}_w$ *is maximally expanded.*

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

By Definition 3.4.6, a nontrivial τ_0 reduces the number of non-maximally expanded symbols by 1, and keeps the number of off-diagonal symbols the same; a $\rho\tau_1$ increases the number of non-maximally expanded symbols by 2 or 3, and increases the number of off-diagonal symbols by 2. Hence $\mathcal{F}_{\text{off}}(\mathbf{s}_w) < 2l_0$ implies that there are at most $(l_0 - 1)$ 1's in w. These $\rho\tau_1$ operators increase $\mathcal{F}_{\textit{n-max}}$ at most by $3(l_0 - 1)$ in total. On the other hand, there are at least $3l_0$ 0's in w, which is sufficient to eliminate all the non-maximally expanded symbols, whose number is at most $3(l_0 - 1) + 1 = 3l_0 - 2$ in total (note that $\mathcal{F}_{\textit{n-max}}(\mathbf{s}) = 1$ for the initial string). □

Now we choose $l_0 = 1 + 1/\zeta$. Then we have

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

using $\Phi = O(N^{-\zeta/2})$. By Lemma 3.4.7, we see that to prove (3.91), it suffices to show that

$$|\mathbb{E}[\![\mathbf{s}_w]\!]| \prec N^{-1}\Phi^2 \tag{3.101}$$

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

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

$$G_{kk}^{(l)},\ (G_{kk}^{(l)})^{-1},\ S_{kl}, \quad \text{with } k \neq l \in \{i, j\}.$$

By (3.47), we can replace $(G_{kk}^{(l)})^{-1}$ with

$$(G_{kk}^{(l)})^{-1} = -1 - S_{kk}. \tag{3.102}$$

Note that $|S_{kk} - m_{2c}| \prec \Phi$ by (3.94). Then we can expand $G_{kk}^{(l)}$ as

$$\begin{aligned} G_{kk}^{(l)} &= \frac{1}{-1 - m_{2c} + (m_{2c} - S_{kk})} \\ &= \frac{-1}{1 + m_{2c}} \sum_{k=0}^{2l_0} \left(\frac{m_{2c} - S_{kk}}{1 + m_{2c}} \right)^k + O_{\prec}(N^{-1}\Phi^2). \end{aligned} \tag{3.103}$$

We apply the expansions (3.102) and (3.103) to the G symbols in $\mathbf{s}_w$, disregard the sufficiently small tails, and denote the resulting polynomial (in terms of the symbols S_{kl}) by P_w. Then P_w can be written as a finite sum of maximally expanded strings (or monomials) consisting of the S_{kl} symbols. Moreover, the number of such monomials depends only on l_0. Hence it suffices to show that for any such monomial M_w, we have

$$|\mathbb{E}[\![M_w]\!]| \prec N^{-1}\Phi^2. \tag{3.104}$$

Let N_i (N_j) be the number of times that i (j) appears as a (lower) index of the S symbols in M_w. We have $N_i = N_j = 3$ for the initial string $\mathbf{s}$ = "$G_{ii}G_{jj}^{(i)}S_{ij}$". From

Definition 3.4.6, it is easy to see that the operators τ_0, τ_1 and ρ do not change the parity of N_i and N_j. The expansions (3.102) and (3.103) also do not change the parity of N_i and N_j. This leads to the following key observation:

$$\textit{both } N_i \textit{ and } N_j \textit{ are odd in } M_w. \tag{3.105}$$

3.4.4 A graphical proof

In this subsection, we finish the proof of (3.104). Suppose $M_w = C(z)(S_{ii})^{m_1}(S_{jj})^{m_2}(S_{ij})^{m_3}(S_{ji})^{m_4}$, where $C(z)$ denotes a deterministic function of order 1 for all $z \in \mathbf{D}$. Then we write

$$\begin{aligned} \llbracket M_w \rrbracket \sim \sum_{\mu_*^{(*)},\nu_*^{(*)} \in \mathcal{I}_2} & \prod_{a=1}^{m_1} X_{i\mu_a^{(1)}} G^{(ij)}_{\mu_a^{(1)}\nu_a^{(1)}} X^*_{\nu_a^{(1)}i} \prod_{b=1}^{m_2} X_{j\mu_b^{(2)}} G^{(ij)}_{\mu_b^{(2)}\nu_b^{(2)}} X^*_{\nu_b^{(2)}j} \\ & \prod_{c=1}^{m_3} X_{i\mu_c^{(3)}} G^{(ij)}_{\mu_c^{(3)}\nu_c^{(3)}} X^*_{\nu_c^{(3)}j} \prod_{d=1}^{m_4} X_{j\mu_d^{(4)}} G^{(ij)}_{\mu_d^{(4)}\nu_d^{(4)}} X^*_{\nu_d^{(4)}i}. \end{aligned} \tag{3.106}$$

To avoid the heavy expressions, we introduce the following graphical notations. We use a connected graph (V, E) to represent the string M_w, where the vertex set V consists of the indices in (3.106) and the edge set E consists of the X and G variables. The indices i, j are represented by the black vertices in the graph, while the μ, ν indices are represented by the white vertices. The X edges are represented by the zig-zag lines and the G edges are represented by the straight lines. One can refer to Fig. 8 for an example of such a graph.

We organize the summation in (3.106) in the following way. We first partition the white vertices into blocks by requiring that any pair of white vertices take the same value if they are in the same block, and take different values otherwise. Then we do the summation over the white blocks which take values in $\mathcal{I}_2$. Finally, we sum over all possible partitions. Note that the number of different partitions depends only on the

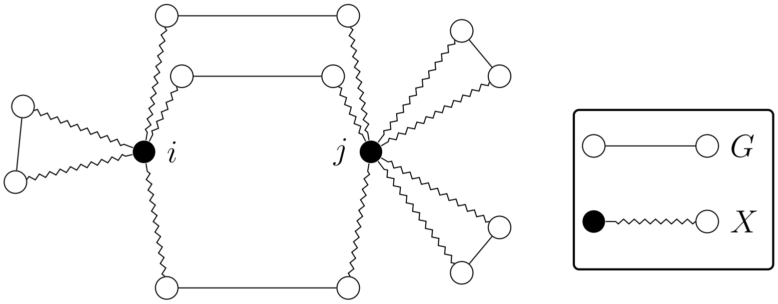


Figure 8: The resulting graph after expanding $S_{ii}(S_{ij})^3(S_{jj})^2$.

total number of S variables in M_w, which in turn depends only on l_0.

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

$$[\![M_w]\!] \sim \sum_{\Gamma} \sideset{}{^*}\sum_{b_1,...,b_k} G(\Gamma) \prod_{l=1}^{k} (X_{ib_l})^{n_l^i} (X_{jb_l})^{n_l^j}, \tag{3.107}$$

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

From (3.106), it is easy to see that the X edges are independent of $G(\Gamma)$. Thus taking expectation of (3.107) gives that

$$\begin{aligned} |\mathbb{E}[\![M_w]\!]| \leq & C \sum_{\Gamma} \sideset{}{^*}\sum_{b_1,...,b_k} |\mathbb{E}G(\Gamma)| \prod_{l=1}^{k} |\mathbb{E}(X_{ib_l})^{n_l^i}||\mathbb{E}(X_{jb_l})^{n_l^j}| \\ \leq & C \sum_{\Gamma} \sideset{}{^*}\sum_{b_1,...,b_k} |\mathbb{E}G(\Gamma)| \prod_{l=1}^{k} \mathbb{E}|X_{ib_l}|^{n_l^i} \mathbb{E}|X_{jb_l}|^{n_l^j} \mathbf{1}(n_l^i \neq 1, n_l^j \neq 1), \end{aligned} \tag{3.108}$$

Note that we must have $n_l^i + n_l^j \geq 2$ for $1 \leq l \leq k$, because we only consider nonempty blocks. On the other hand, if all n_l^i are even, then $N_i = \sum_{l=1}^{k} n_l^i$ must be even, which contradicts (3.105). Hence we can find some $1 \le l_1 \le k$ such that $n^i_{l_1}$ is odd and $n^i_{l_1} \ge 3$. Similarly, we can also find some $1 \le l_2 \le k$ such that $n^j_{l_2}$ is odd and $n^j_{l_2} \ge 3$. We abbreviate $\hat{n}^i_l := n^i_l \wedge 3$ and $\hat{n}^j_l := n^j_l \wedge 3$. From the above discussions, we see that

$$\frac{1}{2}\sum_{l=1}^{k}\left(\hat{n}^i_l + \hat{n}^j_l\right) \ge \frac{1}{2}\sum_{l \neq l_1, l_2}^{k}(\hat{n}^i_l + \hat{n}^j_l) + \frac{3}{2} + \frac{3}{2} \ge (k-2) + 3 = k + 1. \tag{3.109}$$

Now using the moment assumption (3.11), we can bound (3.108) by

$$|\mathbb{E}[\![M_w]\!]| \le C \sum_{\Gamma} \sum_{b_1,\dots,b_k}^{*} |\mathbb{E}G(\Gamma)| N^{-\sum_{l=1}^{k}(\hat{n}^i_l + \hat{n}^j_l)/2}. \tag{3.110}$$

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

(1) there are at least 2 off-diagonal G-edges in $G(\Gamma)$;

(2) there is only 1 off-diagonal G-edge in $G(\Gamma)$;

(3) there is no off-diagonal G-edge in $G(\Gamma)$.

In case (1), we trivially have $|\mathbb{E}G(\Gamma)| \prec \Phi^2$, because the diagonal edges are of order $O_\prec(1)$, while the off-diagonal edges are of order $O_\prec(\Phi)$.

In case (2), we use the same trick as in (3.95). Let the off-diagonal G-edge be $G^{(ij)}_{\mu\nu}$. For each diagonal $G^{(ij)}_{\alpha\alpha}$, we replace it with

$$(G^{(ij)}_{\alpha\alpha} - m_{2c}) + m_{2c} = m_{2c} + O_\prec(\Phi).$$

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

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

where we used (3.96) in the second step.

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

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

Otherwise, if $l_1 = l_2$, then

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

Now applying the above estimates and (3.109) to (3.110), we obtain that

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

This concludes the proof of (3.104), and hence finishes the proof of (3.91).

3.5 Proof of Theorem 3.2.9

3.5.1 Basic notations

Without loss of generality, by (3.86), we can assume

$$\mathbb{E}X_{i\mu} = 0, \quad i \in \mathcal{I}_1,\ \mu \in \mathcal{I}_2, \tag{3.111}$$

in the following proof. Then given X satisfying the assumptions in Theorem 3.2.9 and (3.111), we first construct another random matrix $\tilde{X}$ whose entries have the same first four moments as those of X but have size of order $N^{-1/2}$.

Lemma 3.5.1 (Lemma 5.1 of [46]). *Suppose X satisfies the assumptions in Theorem 3.2.8 and (3.111). Then there exists another matrix $\tilde{X} = (\tilde{X}_{i\mu})$ such that $\mathbb{P}(\max_{i,\mu} |\tilde{X}_{i\mu}| \leq CN^{-1/2}) = 1$ for some constant $C > 0$ and the first four moments of the entries of X and $\tilde{X}$ match, i.e.*

$$\mathbb{E}X_{i\mu}^k = \mathbb{E}\tilde{X}_{i\mu}^k, \quad k = 1, 2, 3, 4. \tag{3.112}$$

Taking $q = N^{-1/2}$ in (3.42), we see that (3.45) holds for $\mathcal{G}_1(\tilde{X}, z)$. Then due to (3.112), we expect that $G(X, z)$ has "similar" properties as $G(\tilde{X}, z)$, so that (3.45) also holds for $\mathcal{G}_1(X, z)$. This will be proved through a resolvent comparison approach that is developed in [46, Sections 6] and [16, Section 6]. More specifically, we will apply the Lindeberg replacement strategy, i.e., we change $\tilde{X}$ to X entry by entry and show that the error (due to the resolvent expansion) appeared at each step is negligible. In this subsection, we introduce some notations that will simplify the presentation of our proof.

Fix a bijective ordering map Φ on the index set of X,

$$\Phi : \{(i, \mu) : i \in \mathcal{I}_1, \mu \in \mathcal{I}_2\} \to \{1, \ldots, \gamma_{\max} = MN\}.$$

For any $1 \leq \gamma \leq \gamma_{\max}$, we define the matrix $X^\gamma = (X_{i\mu}^\gamma)$ such that $X_{i\mu}^\gamma = X_{i\mu}$ if $\Phi(i, \mu) \leq \gamma$, and $X_{i\mu}^\gamma = \tilde{X}_{i\mu}$ otherwise. Note that $X^0 = \tilde{X}$, $X^{\gamma_{\max}} = X$, and X^γ satisfies the bounded support condition with $q \leq N^{-\phi}$ for all $0 \leq \gamma \leq \gamma_{\max}$. Correspondingly, we define

$$H^\gamma := \begin{pmatrix} 0 & X^\gamma \\ (X^\gamma)^* & 0 \end{pmatrix}, \quad G^\gamma := \begin{pmatrix} -I_{M\times M} & X^\gamma \\ (X^\gamma)^* & -zI_{N\times N} \end{pmatrix}^{-1}. \tag{3.113}$$

Note that H^γ and $H^{\gamma-1}$ differ only at (i, μ) and (μ, i) elements, where $\Phi(i, \mu) = \gamma$. Then we define two $\mathcal{I} \times \mathcal{I}$ matrices V and W by

$$V_{ab} = (\delta_{ai}\delta_{b\mu} + \delta_{a\mu}\delta_{bi})X_{i\mu}, \quad W_{ab} = (\delta_{ai}\delta_{b\mu} + \delta_{a\mu}\delta_{bi})\tilde{X}_{i\mu},$$

such that H^γ and $H^{\gamma-1}$ can be written as

$$H^\gamma = Q + V, \quad H^{\gamma-1} = Q + W, \tag{3.114}$$

for some $\mathcal{I} \times \mathcal{I}$ matrix Q satisfying $Q_{i\mu} = Q_{\mu i} = 0$.

For simplicity, for any $1 \le \gamma \le \gamma_{\max}$, we denote the resolvents by

$$S^\gamma := G^\gamma, \quad T^\gamma := G^{\gamma-1}, \quad R^\gamma := \left(Q - \begin{pmatrix} I_{M\times M} & 0 \\ 0 & zI_{N\times N} \end{pmatrix} \right)^{-1}. \tag{3.115}$$

We often omit the superscript if γ is fixed. By (3.114), we can write

$$S = \left(Q - \begin{pmatrix} I_{M\times M} & 0 \\ 0 & zI_{N\times N} \end{pmatrix} + V \right)^{-1} = (1 + RV)^{-1} R. \tag{3.116}$$

Thus we can expand S using the resolvent expansion

$$S = R - RVR + (RV)^2 R + \ldots + (-1)^m (RV)^m R + (-1)^{m+1} (RV)^{m+1} S. \tag{3.117}$$

On the other hand, we can also expand R in terms of S:

$$R = (1 - SV)^{-1} S = S + SVS + (SV)^2 S + \ldots + (SV)^m S + (SV)^{m+1} R. \tag{3.118}$$

We can get similar expansions for T and R by replacing V, S with W, T in (3.117) and (3.118).

By the bounded support conditions for X and $\tilde{X}$, we have

$$\max_{a,b\in\mathcal{I}} |V_{ab}| = |X_{i\mu}| \prec N^{-\phi}, \quad \max_{a,b\in\mathcal{I}} |W_{ab}| = |\tilde{X}_{i\mu}| \le CN^{-1/2}. \tag{3.119}$$

Also, note that S, R, T satisfy the following deterministic bound by (3.54):

$$\sup_{z\in\mathbf{D}} \max_\gamma \max\left\{ \|S^\gamma\|, \|T^\gamma\|, \|R^\gamma\| \right\} \le \sup_{z\in\mathbf{D}} (C\eta^{-1}) \le N. \tag{3.120}$$

Then using expansion (3.118) in terms of T, W with $m=3$, the isotropic local law (3.40) for T, and the bound (3.120) for R, we can get that for any fixed unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}}$, $|R_{\mathbf{uv}}| = O(1)$ with high probability. Thus there exists a uniform constant $C_1 > 0$ such that with high probability,

$$\sup_{z\in\mathbf{D}} \max_{\gamma} \sup_{\text{deterministic unit } \mathbf{u},\mathbf{v}} \max\left\{|S^{\gamma}_{\mathbf{uv}}|, |T^{\gamma}_{\mathbf{uv}}|, |R^{\gamma}_{\mathbf{uv}}|\right\} \le C_1. \tag{3.121}$$

From the definitions of V and W, one can see that it is helpful to introduce the following notations to simplify the expressions.

Definition 3.5.2 (Matrix operators $*_\gamma$)**.** *For $\mathcal{I}\times\mathcal{I}$ matrices A and B, we define $A *_\gamma B$ as*

$$(A *_\gamma B)_{ab} = A_{ai}B_{\mu b} + A_{a\mu}B_{ib}, \quad \Phi(i,\mu) = \gamma. \tag{3.122}$$

We denote the m-th power of A under $*_\gamma$-product by $A^{*_\gamma m}$, i.e.,

$$A^{*_\gamma m} := A *_\gamma A *_\gamma A *_\gamma \ldots *_\gamma A. \tag{3.123}$$

Definition 3.5.3 ($\mathcal{P}_{\gamma,\mathbf{k}}$ and $\mathcal{P}_{\gamma,k}$)**.** *For $k\in\mathbb{N}$, $\mathbf{k}=(k_1,\cdots,k_s)\in\mathbb{N}^s$ and $\gamma=\Phi(i,\mu)$, we define*

$$\mathcal{P}_{\gamma,k}G_{\mathbf{uv}} := G^{*_\gamma(k+1)}_{\mathbf{uv}}, \quad \mathcal{P}_{\gamma,\mathbf{k}}\left(\prod_{t=1}^{s} G_{\mathbf{u}_t\mathbf{v}_t}\right) := \prod_{t=1}^{s}\mathcal{P}_{\gamma,k_t}G_{\mathbf{u}_t\mathbf{v}_t}. \tag{3.124}$$

If $\mathfrak{G}_1$ and $\mathfrak{G}_2$ are products of resolvent entries as above, then we define

$$\mathcal{P}_{\gamma,\mathbf{k}}(\mathfrak{G}_1+\mathfrak{G}_2) := \mathcal{P}_{\gamma,\mathbf{k}}\mathfrak{G}_1 + \mathcal{P}_{\gamma,\mathbf{k}}\mathfrak{G}_2. \tag{3.125}$$

Note that $\mathcal{P}_{\gamma,k}$ and $\mathcal{P}_{\gamma,\mathbf{k}}$ are not linear operators, but just notations we use for simplification.

Using Definition 3.5.3, we may write, for example,

$$\mathcal{P}_{\gamma,\mathbf{k}}\left(\prod_{t=1}^{s} G^{\gamma}_{\mathbf{u}_t\mathbf{v}_t}\right) := \prod_{t=1}^{s} S^{*_\gamma(k_t+1)}_{\mathbf{u}_t\mathbf{v}_t}, \quad \mathcal{P}_{\gamma,\mathbf{k}}\left(\prod_{t=1}^{s} G^{\gamma-1}_{\mathbf{u}_t\mathbf{v}_t}\right) := \prod_{t=1}^{s} T^{*_\gamma(k_t+1)}_{\mathbf{u}_t\mathbf{v}_t}.$$

For $k, s \in \mathbb{N}$ and $\mathbf{k} \in \mathbb{N}^{s+1}$, it is easy to verify that

$$G^{*_\gamma s} *_\gamma G^{*_\gamma k} = G^{*_\gamma (s+k)}, \quad \mathcal{P}_{\gamma,\mathbf{k}}(\mathcal{P}_{\gamma,s} G_{\mathbf{uv}}) = \mathcal{P}_{\gamma,s+|\mathbf{k}|} G_{\mathbf{uv}}, \tag{3.126}$$

where $|\mathbf{k}| = \sum_{t=1}^{s} k_t$. For the second equality, note that $\mathcal{P}_{\gamma,s} G_{\mathbf{uv}}$ is a sum of the products of the G entries, where each product contains $s+1$ entries.

3.5.2 Proof of (3.45)

As mentioned in the last subsection, we will prove (3.45) with the resolvent comparison method. The basic idea is that we expand S and T in terms of R by repeatedly applying the expansions (3.117) and (3.118), and then compare the resulting expressions. The main terms will cancel since $X_{i\mu}$ and $\tilde{X}_{i\mu}$ have the same first four moments, and the error terms are small since $X_{i\mu}$ and $\tilde{X}_{i\mu}$ have support bounded by $N^{-\phi}$.

The proof of the following Lemma 3.5.4 is almost the same as the one for [46, Lemma 6.5]. In fact, we can copy their arguments almost verbatim, except for some notational differences. Hence we omit the details. In the following expressions, for any $\mathbf{k} = (k_1, \ldots, k_p) \in \mathbb{N}^p$, we use $|\mathbf{k}| = \sum k_i$ to denote its l^1-norm.

Lemma 3.5.4. *Assume $z \in \mathbf{D}$ and $\gamma = \Phi(i,\mu)$. Fix any $p \in \mathbb{N}$ and $r > 0$. Then for S, R in (3.115), we have*

$$\begin{aligned} \mathbb{E} \prod_{t=1}^{p} S_{\mathbf{u}_t \mathbf{v}_t} = & \sum_{0 \le k \le 4} A_k \mathbb{E}\left[(-X_{i\mu})^k\right] \\ & + \sum_{5 \le |\mathbf{k}| \le r/\phi, \mathbf{k} \in \mathbb{N}^p} \mathcal{A}_{\mathbf{k}} \mathbb{E}\, \mathcal{P}_{\gamma,\mathbf{k}} \prod_{t=1}^{p} S_{\mathbf{u}_t \mathbf{v}_t} + O_\prec(N^{-r}), \end{aligned} \tag{3.127}$$

where A_k, $0 \le k \le 4$, depend only on R, $\mathcal{A}_{\mathbf{k}}$'s are independent of $(\mathbf{u}_t, \mathbf{v}_t)$, $1 \le t \le s$, and we have the bound

$$|\mathcal{A}_{\mathbf{k}}| \prec N^{-|\mathbf{k}|\phi/10-2}. \tag{3.128}$$

It is obvious that a result similar to Lemma 3.5.4 also holds for the product of T entries. As in (3.127), we define the notation $\mathcal{A}^{\gamma,a}$, $a = 0, 1$ as follows:

$$\begin{aligned} \mathbb{E}\prod_{t=1}^{p} S_{\mathbf{u}_t\mathbf{v}_t} = & \sum_{0\le k\le 4} A_k \mathbb{E}\left[(-X_{i\mu})^k\right] \\ & + \sum_{5\le|\mathbf{k}|\le r/\phi, \mathbf{k}\in\mathbb{N}^p} \mathcal{A}^{\gamma,0}_{\mathbf{k}} \mathbb{E}\,\mathcal{P}_{\gamma,\mathbf{k}} \prod_{t=1}^{p} S_{\mathbf{u}_t\mathbf{v}_t} + O_\prec(N^{-r}), \end{aligned} \tag{3.129}$$

$$\begin{aligned} \mathbb{E}\prod_{t=1}^{p} T_{\mathbf{u}_t\mathbf{v}_t} = & \sum_{0\le k\le 4} A_k \mathbb{E}\left[(-\tilde{X}_{i\mu})^k\right] \\ & + \sum_{5\le|\mathbf{k}|\le r/\phi, \mathbf{k}\in\mathbb{N}^p} \mathcal{A}^{\gamma,1}_{\mathbf{k}} \mathbb{E}\,\mathcal{P}_{\gamma,\mathbf{k}} \prod_{t=1}^{p} T_{\mathbf{u}_t\mathbf{v}_t} + O_\prec(N^{-r}). \end{aligned} \tag{3.130}$$

Since A_k, $0 \le k \le 4$, depend only on R and $X_{i\mu}$, $\tilde{X}_{i\mu}$ have the same first four moments, we get from (3.129) and (3.130) that

$$\begin{aligned} & \mathbb{E}\prod_{t=1}^{p} G_{\mathbf{u}_t\mathbf{v}_t} - \mathbb{E}\prod_{t=1}^{p} \tilde{G}_{\mathbf{u}_t\mathbf{v}_t} = \sum_{\gamma=1}^{\gamma_{\max}} \left(\mathbb{E}\prod_{t=1}^{p} G^{\gamma}_{\mathbf{u}_t\mathbf{v}_t} - \mathbb{E}\prod_{t=1}^{p} G^{\gamma-1}_{\mathbf{u}_t\mathbf{v}_t}\right) \\ & = \sum_{\gamma=1}^{\gamma_{\max}} \sum_{\mathbf{k}\in\mathbb{N}^p}^{5\le|\mathbf{k}|\le r/\phi} \left(\mathcal{A}^{\gamma,0}_{\mathbf{k}} \mathbb{E}\,\mathcal{P}_{\gamma,\mathbf{k}} \prod_{t=1}^{p} G^{\gamma}_{\mathbf{u}_t\mathbf{v}_t} - \mathcal{A}^{\gamma,1}_{\mathbf{k}} \mathbb{E}\,\mathcal{P}_{\gamma,\mathbf{k}} \prod_{t=1}^{p} G^{\gamma-1}_{\mathbf{u}_t\mathbf{v}_t}\right) \\ & \quad + O_\prec(N^{-r+2}). \end{aligned} \tag{3.131}$$

where we abbreviate $G := G(X, z)$ and $\tilde{G} := G(\tilde{X}, z)$.

Applying (3.131) with $p = 1$, $r = 3$ and fixed unit vector $\mathbf{u}_t = \mathbf{v}_t = \mathbf{v} \in \mathbb{C}^{\mathcal{I}_1}$, we obtain that

$$\left|\mathbb{E}(G - \tilde{G})_{\mathbf{v}\mathbf{v}}\right| \le \sum_{\gamma=1}^{\gamma_{\max}} \sum_{a=0,1} \sum_{5\le k\le 3/\phi} |\mathcal{A}^{\gamma,a}_k| \left|\mathbb{E}\mathcal{P}_{\gamma,k} G^{\gamma-a}_{\mathbf{v}\mathbf{v}}\right| + O_\prec(N^{-1}). \tag{3.132}$$

Using (3.121), (3.128) and Lemma 3.2.5, we can bound the sum in (3.132) by

$$\sum_{\gamma=1}^{\gamma_{\max}} \sum_{a=0,1} \sum_{5\le k\le 3/\phi} |\mathcal{A}^{\gamma,a}_k| \left|\mathbb{E}\mathcal{P}_{\gamma,k} G^{\gamma-a}_{\mathbf{v}\mathbf{v}}\right| \prec \sum_{5\le k\le 3/\phi} N^{-k\phi/10} \prec N^{-\phi/2}. \tag{3.133}$$

To apply Lemma 3.2.5 (iii), we need a second moment bound for $|\mathcal{P}_{\gamma,k}G^{\gamma-a}_{\mathbf{vv}}|$, which follows easily from (3.120). Recall that $\mathcal{P}_{\gamma,k}G^{\gamma-a}_{\mathbf{vv}}$ is also a sum of the products of G entries. Then applying (3.131) to $|\mathbb{E}\mathcal{P}_{\gamma,k}G^{\gamma-a}_{\mathbf{vv}}|$ and replacing $\gamma_{\max}$ with $\gamma - a$, we obtain that

$$\begin{aligned}\left|\mathbb{E}\mathcal{P}_{\gamma,k}G^{\gamma-a}_{\mathbf{vv}}\right| &\le \left|\mathbb{E}\mathcal{P}_{\gamma,k}G^{0}_{\mathbf{vv}}\right| \\ &+\sum_{\gamma'=1}^{\gamma-a}\sum_{a'=0,1}\sum_{\mathbf{k}'\in\mathbb{N}^{1+k}}^{5\le|\mathbf{k}'|\le 3/\phi}\left|\mathcal{A}^{\gamma',a'}_{\mathbf{k}'}\right|\left|\mathbb{E}\mathcal{P}_{\gamma',\mathbf{k}'}\mathcal{P}_{\gamma,k}G^{\gamma'-a'}_{\mathbf{vv}}\right| + O_{\prec}(N^{-1}).\end{aligned} \tag{3.134}$$

Together with (3.132) and (3.128), we get that

$$\begin{aligned}\left|\mathbb{E}(G-\tilde{G})_{\mathbf{vv}}\right| &\le \sum_{\gamma,a}\sum_{k}|\mathcal{A}^{\gamma,a}_{k}|\left|\mathbb{E}\mathcal{P}_{\gamma,k}G^{0}_{\mathbf{vv}}\right| \\ &+\sum_{\gamma,\gamma'}\sum_{a,a'}\sum_{k,\mathbf{k}'}\left|\mathcal{A}^{\gamma,a}_{k}\mathcal{A}^{\gamma',a'}_{\mathbf{k}'}\right|\left|\mathbb{E}\mathcal{P}_{\gamma',\mathbf{k}'}\mathcal{P}_{\gamma,k}G^{\gamma'-a'}_{\mathbf{vv}}\right| + O_{\prec}(N^{-1}).\end{aligned}$$

Again using (3.121), (3.128) and Lemma 3.2.5, we obtain that

$$\sum_{\gamma,\gamma'}\sum_{a,a'}\sum_{k,\mathbf{k}'}\left|\mathcal{A}^{\gamma,a}_{k}\mathcal{A}^{\gamma',a'}_{\mathbf{k}'}\right|\left|\mathbb{E}\mathcal{P}_{\gamma',\mathbf{k}'}\mathcal{P}_{\gamma,k}G^{\gamma'-a'}_{\mathbf{vv}}\right| \prec N^{-\phi}, \tag{3.135}$$

where we used that $k + |\mathbf{k}'| \ge 10$. Repeating this process, we can make the remainder term smaller and smaller. At the end, we obtain that

$$\begin{aligned}\left|\mathbb{E}(G-\tilde{G})_{\mathbf{vv}}\right| \le &\sum_{n=0}^{2/\phi}\sum_{\gamma_1,\cdots,\gamma_n}\sum_{a_1,\cdots,a_n}\sum_{\mathbf{k}_1,\cdots,\mathbf{k}_n}\left|\prod_j \mathcal{A}^{\gamma_j,a_j}_{\mathbf{k}_j}\right|\left|\mathbb{E}\mathcal{P}_{\gamma_n,\mathbf{k}_n}\cdots\mathcal{P}_{\gamma_1,\mathbf{k}_1}G^{0}_{\mathbf{vv}}\right| \\ &+ O_{\prec}(N^{-1}),\end{aligned}$$

where

$$\mathbf{k}_1\in\mathbb{N}^1,\ \mathbf{k}_2\in\mathbb{N}^{1+|\mathbf{k}_1|},\ \mathbf{k}_3\in\mathbb{N}^{1+|\mathbf{k}_1|+|\mathbf{k}_2|},\ \text{etc.,}\quad\text{and}\quad 5\le|\mathbf{k}_i|\le\frac{3}{\phi}. \tag{3.136}$$

Using (3.128) and Lemma 3.2.5, we obtain that

$$\begin{aligned}&\left|\mathbb{E}(G-\tilde{G})_{\mathbf{vv}}\right| \\ &\prec \max_{\mathbf{k},n}(N^{-2})^n(N^{-\frac{\phi}{10}})^{\sum_i|\mathbf{k}_i|}\sum_{\gamma_1,\cdots,\gamma_n}\left|\mathbb{E}\mathcal{P}_{\gamma_n,\mathbf{k}_n}\cdots\mathcal{P}_{\gamma_1,\mathbf{k}_1}G^{0}_{\mathbf{vv}}\right| + N^{-1}.\end{aligned} \tag{3.137}$$

Now we finish the proof of (3.45) using the estimate (3.137) and the bound (3.45) for $G^0 = G(\tilde{X}, z)$. We see that it suffices to control the term

$$\mathcal{P}_{\gamma_n,\mathbf{k}_n} \cdots \mathcal{P}_{\gamma_1,\mathbf{k}_1} \tilde{G}_{\mathbf{v}\mathbf{v}} \tag{3.138}$$

for $\mathbf{k}_1, \ldots, \mathbf{k}_n$ satisfying (3.136). By definition of $\mathcal{P}$, (3.138) is a sum of at most $C^{\sum|\mathbf{k}_i|}$ products of $G_{\mathbf{v}b}$, $G_{b\mathbf{v}}$ and G_{ab} entries, where the total number of G entries in each product is at most $\sum|\mathbf{k}_i|+1 = O(\phi^{-2})$. Due to the deterministic bound (3.120), (3.138) is always bounded by $N^{O(\phi^{-2})}$, and hence Lemma 3.2.5 (iii) can be applied.

For each product in (3.138), there are two $\mathbf{v}$'s in the indices of G. These two $\mathbf{v}$'s appear as $G_{\mathbf{v}a}G_{b\mathbf{v}}$ in the product, where a, b come from some γ_k and γ_l ($1 \le k, l \le n$) via $\mathcal{P}$. Thus after taking the average $N^{-2}\sum_{\gamma_k}$ and $N^{-2}\sum_{\gamma_l}$, the term $G_{\mathbf{v}a}G_{b\mathbf{v}}$ contributes a factor $O_{\prec}((N\eta)^{-1})$ by (3.59) and Cauchy-Schwarz inequality. For all other G factors in the product with no $\mathbf{v}$'s, we control them by $O_{\prec}(1)$ using (3.121). Thus for any fixed $\gamma_1, \ldots, \gamma_n$, $\mathbf{k}_1, \ldots, \mathbf{k}_n$, we have proved that

$$N^{-2n} \sum_{\gamma_1,\cdots,\gamma_n} \left| \mathbb{E}\mathcal{P}_{\gamma_n,\mathbf{k}_n} \cdots \mathcal{P}_{\gamma_1,\mathbf{k}_1} \tilde{G}_{\mathbf{v}\mathbf{v}} \right| \prec \frac{1}{N\eta}.$$

Then using (3.137) and (3.45) for $\tilde{G}$, we obtain that

$$|\mathbb{E}(\mathcal{G}_1)_{\mathbf{v}\mathbf{v}} - m_{1c}(z)| \prec |\mathbb{E}(\tilde{\mathcal{G}}_1)_{\mathbf{v}\mathbf{v}} - m_{1c}(z)| + \frac{1}{N\eta} \prec \frac{1}{N\eta},$$

where we abbreviated $\mathcal{G}_1 = z^{-1}G$ and $\tilde{\mathcal{G}}_1 = z^{-1}\tilde{G}$. This concludes (3.45).

3.6 Appendix

3.6.1 Proof of Lemma 3.4.2

We only prove

$$\max_{i,j\in\mathcal{I}_1} |G_{ij}(X,z) - \Pi_{ij}(z)| \prec q^2 + (N\eta)^{-1/2}. \tag{3.139}$$

The proof for (3.44) with $a, b \in \mathcal{I}_2$ is exactly the same. First, we recall the following large deviation bounds proved in [23].

Lemma 3.6.1 (Lemma 3.8 of [23])**.** *Let (x_i), (y_i) be independent families of centered and independent random variables, and (A_i), (B_{ij}) be families of deterministic complex numbers. Suppose the entries x_i and y_j have variance $O(N^{-1})$ and satisfy (3.34) with $N^{-1/2} \le q \le N^{-\phi}$ for some fixed $\phi > 0$. Then for $K = O(N)$, we have the following bounds:*

$$\Big| \sum_{1\le i,j\le K} x_i B_{ij} y_j \Big| \prec q^2 B_d + q B_o + \frac{1}{N}\Big(\sum_{i\ne j} |B_{ij}|^2 \Big)^{1/2}, \tag{3.140}$$

$$\Big| \sum_{1\le i\ne j\le K} \bar{x}_i B_{ij} x_j \Big| \prec q B_o + \frac{1}{N}\Big(\sum_{i\ne j} |B_{ij}|^2 \Big)^{1/2}, \tag{3.141}$$

$$\Big| \sum_{1\le i\le K} \big(|x_i|^2 - \mathbb{E}|x_i|^2 \big) B_{ii} \Big| \prec q B_d, \tag{3.142}$$

where $B_d := \max_i |B_{ii}|$ and $B_o := \max_{i\ne j} |B_{ij}|$.

In fact, these bounds are stated in slightly stronger forms in [23] with a different notion for high probability events. Here we choose to present (3.140)-(3.142) in terms of the stochastic domination, which will be more convenient for our use. Moreover, if we assume the fourth moment of x_i is bounded for all i as in (3.11), then we have a better bound for the LHS of (3.142).

Lemma 3.6.2. *Suppose the assumptions in Lemma 3.6.1 hold and* x_i*,* $1 \le i \le K$*, satisfy (3.11). Then we have*

$$\Big| \sum_i \left(|x_i|^2 - \mathbb{E}|x_i|^2 \right) B_{ii} \Big| \prec \left(q^2 + N^{-1/2} \right) B_d. \tag{3.143}$$

Proof. We abbreviate $z_i := \left(|x_i|^2 - \mathbb{E}|x_i|^2 \right) B_{ii}/B_d$. By Markov's inequality, it suffices to prove that for any fixed $p \in \mathbb{N}$,

$$\mathbb{E}\Big| \sum_i z_i \Big|^{2p} \prec \left(q^2 + N^{-1/2} \right)^{2p}. \tag{3.144}$$

Note that by the assumption, we have

$$\mathbb{E} z_i = 0, \quad \mathbb{E}|z_i|^n \prec q^{2n-4} N^{-2} \text{ for fixed } n \ge 2. \tag{3.145}$$

Now we expand the LHS of (3.144) to get

$$\mathbb{E}\Big| \sum_i z_i \Big|^{2p} = \sum_{i_1, \dots, i_{2p}} \mathbb{E} y_{i_1} \cdots y_{i_{2p}},$$

where we denote $y_{i_l} := z_{i_l}$ for $1 \le l \le p$ and $y_{i_l} := \bar{z}_{i_l}$ for $p+1 \le l \le 2p$. To organize the summation over the indices $i_1, \dots, i_{2p}$, we look at the partitions Γ of the set of the labels $\{1, ..., 2p\}$ according to the equivalence relation that k, l are in the same class if and only if $i_k = i_l$. We use b_l, $1 \le l \le k$, to denote the equivalence classes of Γ and n_l to denote the size of b_l. Obviously, k, b_l and n_l all depend on Γ, but we will omit this dependence in the following expressions. Moreover, since the random variables are centered, we must have $n_l \ge 2$ for all l to attain a nonzero expectation. Hence we have

$$\mathbb{E}\Big| \sum_i z_i \Big|^{2p} \le \sum_{\Gamma} \sum_{b_1, \dots, b_k}^{*} \mathbb{E}|y_{\mu_{b_1}}|^{n_1} \dots \mathbb{E}|y_{\mu_{b_k}}|^{n_k}, \tag{3.146}$$

where $\sum^*$ denotes the summation subject to the conditions that $b_1, \dots, b_k$ are all distinct, $n_l \ge 2$ for all l, and $\sum_{l=1}^k n_l = 2p$. Note that under these conditions, we trivially have $k \le p$.

Using (3.145), we get

$$\sum_{b_1,\dots,b_k}^{*}\mathbb{E}|y_{\mu_{b_1}}|^{n_1}\dots\mathbb{E}|y_{\mu_{b_k}}|^{n_k}\prec\sum_{b_1,\dots,b_k}^{*}(q^{2n_1-4}N^{-2})\dots(q^{2n_k-4}N^{-2})$$
$$=\sum_{b_1,\dots,b_k}^{*}N^{-2k}q^{4p-4k}\le CN^{-k}q^{4p-4k}.$$

Since the number of partitions of $\{1,...,2p\}$ is finite and depends only on p, (3.146) can be bounded by

$$\mathbb{E}\Big|\sum_i z_i\Big|^{2p}\prec\max_{1\le k\le p}N^{-k}q^{4p-4k}\le q^{4p}+N^{-p},$$

where in the last step, q^{4p} and N^{-p} can be obtained from the extreme cases $k=0$ and $k=p$, respectively. This concludes (3.144). □

Now using (3.48) and (3.140), we get that

$$|G_{ij}|\prec\left|\sum_{\mu,\nu}X_{i\mu}G^{(ij)}_{\mu\nu}X^*_{\nu j}\right|\prec q^2\max_{\mu}|G^{(ij)}_{\mu\mu}|+q\max_{\mu\neq\nu}|G^{(ij)}_{\mu\nu}|+\Big(\frac{1}{N^2}\sum_{\mu\neq\nu}|G^{(ij)}_{\mu\nu}|^2\Big)^{1/2}$$
$$\prec q^2+q(q+\Psi)+\left(\frac{1}{N\eta}\right)^{1/2}\prec q^2+(N\eta)^{-1/2},\qquad(3.147)$$

where we used (3.43), (3.51) and the bound (3.59). For the diagonal estimate, we need to control the Z variables defined in (3.89). Using (3.141) and (3.143), we get that for any $\mathbb{T}\subset\mathcal{I}$ with fixed length,

$$\begin{aligned}|Z_i^{(\mathbb{T})}|&=\Big|\sum_{\mu}G^{(\mathbb{T}i)}_{\mu\mu}\left(|X_{i\mu}|^2-\mathbb{E}|X_{i\mu}|^2\right)+\sum_{\mu\neq\nu}X_{i\mu}G^{(\mathbb{T}i)}_{\mu\nu}X^*_{\nu i}\Big|\\&\prec\left(q^2+N^{-1/2}\right)+q\max_{\mu\neq\nu}|G^{(\mathbb{T}i)}_{\mu\nu}|+\Big(\frac{1}{N^2}\sum_{\mu\neq\nu}|G^{(\mathbb{T}i)}_{\mu\nu}|^2\Big)^{1/2}\\&\prec q^2+(N\eta)^{-1/2},\end{aligned}\qquad(3.148)$$

where we used (3.43), (3.51) and (3.59) again. Then using (3.47), we get

$$G_{ii}-zm_{1c}=\frac{1}{-1-m_{2c}-(m_2^{(i)}-m_{2c})+Z_i}-zm_{1c}$$

$$
\begin{aligned}
&= \frac{1}{-1-m_{2c}} - zm_{1c} + O_\prec\left(q^2 + (N\eta)^{-1/2}\right) \\
&= O_\prec\left(q^2 + (N\eta)^{-1/2}\right)
\end{aligned}
$$

where we used (3.148), (3.39), (3.52) (with $\Phi = q + \Psi$) in the second step, and (3.4), (3.6) in the third step. Together with (3.147), this concludes (3.139).

3.6.2 Proof of Lemma 3.4.3

Note that by (3.87), we immediately get $\sum_i |v_i|^2 \left((\mathcal{G}_1)_{ii} - m_{1c}\right) \prec \Phi$. Hence it remains to show that

$$\sum_{i \neq j} \bar{v}_i v_j \left(\mathcal{G}_1\right)_{ij} \prec \Phi.$$

By Markov's inequality and (3.30), it suffices to prove that

$$\mathbb{E}\Big|\sum_{i\neq j} \bar{v}_i v_j G_{ij}\Big|^{2p} \prec \Phi^{2p} \tag{3.149}$$

for any fixed $p \in \mathbb{N}$. The proof of (3.149) is similar to the ones in [6, Section 5] and [68, Section 5]. The main difference is that in [6, 68], the matrix entries are assumed to have arbitrarily high moments, while we assume that the X entries have finite third moment and support bounded by q in our proof. In particular, for any fixed $n \geq 3$, we have

$$\mathbb{E}|X_{i\mu}|^n \prec q^{n-3} N^{-3/2}, \quad i \in \mathcal{I}_1,\ \mu \in \mathcal{I}_2. \tag{3.150}$$

Note that we have a stronger moment assumption in (3.11). However, the finite fourth moment condition will not be used in the proof below. We only need the weaker bound (3.150). Also we remark that some of the basic ideas have been illustrated in the proof for (3.42) in Section 3.4.

We first rewrite the product in (3.149) as

$$\begin{aligned}\Big|\sum_{i\neq j}\bar v_i G_{ij} v_j\Big|^{2p} &= \sum_{i_k\neq j_k\in\mathcal{I}_1}\prod_{k=1}^{p}\bar v_{i_k}G_{i_kj_k}v_{j_k}\cdot\prod_{k=p+1}^{2p}\overline{\bar v_{i_k}G_{i_kj_k}v_{j_k}}\\ &=\sum_{\Gamma}\sum_{b_1,\ldots,b_n}^{*}\prod_{k=1}^{p}\bar v_{\Gamma(i_k)}G_{\Gamma(i_k)\Gamma(j_k)}v_{\Gamma(j_k)}\cdot\prod_{k=p+1}^{2p}\overline{\bar v_{\Gamma(i_k)}G_{\Gamma(i_k)\Gamma(j_k)}v_{\Gamma(j_k)}},\end{aligned}$$

where (recall the notations in the proof for Lemma 3.6.2) Γ ranges over all partitions of the set of the labels $\{i_1,...,i_{2p},j_1,...,j_{2p}\}$ with the restriction that i_k, j_k cannot be in the same equivalence class for all k, $\{b_1,...,b_n\}$ is the set of equivalence classes for a fixed Γ, $\Gamma(\cdot)$ is regarded as a symbolic mapping from the set of labels to the set of equivalence classes, and $\sum^*$ denotes the summation subject to the condition that $b_1,\ldots,b_n$ all take distinct values and $\Gamma(i_k)\neq\Gamma(j_k)$ for all k.

Since the number of such partitions Γ is finite and depends only on p, it suffices to show that for any fixed Γ,

$$\mathbb{E}\sum_{b_1,\ldots,b_n}^{*}\prod_{k=1}^{p}\bar v_{\Gamma(i_k)}G_{\Gamma(i_k)\Gamma(j_k)}v_{\Gamma(j_k)}\cdot\prod_{k=p+1}^{2p}\overline{\bar v_{\Gamma(i_k)}G_{\Gamma(i_k)\Gamma(j_k)}v_{\Gamma(j_k)}}\prec\Phi^{2p}. \tag{3.151}$$

We abbreviate

$$P(b_1,...,b_n):=\prod_{k=1}^{p}G_{\Gamma(i_k)\Gamma(j_k)}\cdot\prod_{k=p+1}^{2p}\overline{G_{\Gamma(i_k)\Gamma(j_k)}}.$$

For simplicity, we shall omit the overline for complex conjugate in the following proof. In this way, we can avoid a lot of immaterial notational complexities that do not affect the proof.

For $k=1,...,n$, we denote by $\deg(b_k,P)$ the number of times that b_k appears as an index of the G entries in P, i.e. $\deg(b_k,P):=|\Gamma^{-1}(b_k)|$. We define $h:=\#\{1\leq k\leq n:\deg(b_k,P)=1\}$, i.e. h is the number of b_k's that only appear once in the indices of P. Without loss of generality, we assume these b_k's are $b_1,...,b_h$. Then we have the following properties:

$$\sum_{k=1}^{n} \deg(b_k, P) = 4p, \quad \text{and} \quad \deg(b_k, P) = 1, \quad \text{for } k = 1, ..., h. \tag{3.152}$$

Now we claim that

$$|\mathbb{E}P| \prec N^{-h/2}\Phi^{2p}. \tag{3.153}$$

Note that by $\|\mathbf{v}\|_2 = 1$ and Cauchy-Schwarz inequality, we have $\sum_i |v_i| \le \sqrt{M}$ and $\sum_i |v_i|^n \le 1$ for $n \ge 2$. Then if (3.153) holds, we can bound the left hand side of (3.151) by

$$N^{-h/2}\Phi^{2p} \prod_{k=1}^{n} \sum_{b_k} |v_{b_k}|^{\deg(b_k,P)} \le N^{-h/2}\Phi^{2p}(\sqrt{M})^{h/2} \le C\Phi^{2p}.$$

Hence it suffices to prove (3.153).

We define the S variables as (one can compare them with (3.97))

$$S_{ij} := (XG^{(L)}X^*)_{ij}, \tag{3.154}$$

for $i, j \in \mathcal{I}_1$ and $L := \{b_1, ..., b_n\}$. As in (3.147) and (3.148), we can verify that $|S_{ij} - m_{2c}\delta_{ij}| \prec \Phi$ for $i, j \in \mathcal{I}_1$ using (3.87), (3.51) and Lemmas 3.6.1-3.6.2. Then as in Section 3.4.3, we keep expanding the G entries in P using the resolvent expansions in Lemma 3.2.12, until each monomial in the expression either consists of S variables only or has sufficiently many off-diagonal terms. The following lemma corresponds to the previous Lemma 3.4.7 and has been proved in [6, Lemma 5.9] and [68, Lemma 5.9].

Lemma 3.6.3. *After finitely many expansions, we can write P as*

$$P = \sum_{\alpha=1}^{A} c_\alpha Q_\alpha + O_\prec(N^{-h/2}\Phi^{2p}), \tag{3.155}$$

where $A \in \mathbb{N}$ depends only on p and c_0 (recall that $\Phi(z) \le N^{-c_0}$ by assumption), c_α's are constants of order $O(1)$, and Q_α are products of S variables only and the number of

S variables in each product again depends only on p and c_0. Moreover, for $k = 1, ..., n$ and $\alpha = 1, ..., A$, we have that

$$\deg_o(b_k, Q_\alpha) \geq \deg_o(b_k, P), \quad \deg_o(b_k, Q_\alpha) = \deg_o(b_k, P) \mod 2, \tag{3.156}$$

and the number of off-diagonal S variables in Q is at least $2p$. Here $\deg_o(b_k, Q_\alpha)$ denotes the number of times that b_k appears as an index of the off-diagonal S variables in Q_α and $\deg_o(b_k, P) := \deg(b_k, P)$ (which is consistent with the previous definition since P only contains off-diagonal entries).

Now given the expansion in (3.155), we see that to conclude (3.153), it suffices to show that for any Q_α,

$$|\mathbb{E}Q_\alpha| \prec N^{-h/2}\Phi^{2p}. \tag{3.157}$$

In the following proof, we fixe one such $Q \equiv Q_\alpha$ and write

$$\begin{aligned} Q = \prod_{j=1}^{J} S_{b_{k_j} b_{l_j}} &= \sum_{\mu_j, \nu_j \in \mathcal{I}_2} \prod_{j=1}^{J} X_{b_{k_j}\mu_j} G_{\mu_j \nu_j} X^*_{\nu_j b_{l_j}} \\ &= \sum_{W} \sum_{w_1, ..., w_m}^{*} \prod_{j=1}^{J} X_{b_{k_j} W(\mu_j)} G_{W(\mu_j)W(\nu_j)} X_{b_{l_j} W(\nu_j)} \end{aligned}$$

where J is the number of S-variables in Q, W ranges over all partitions of the set of the labels $\{\mu_1, ..., \mu_J, \nu_1, ..., \nu_J\}$, $\{w_1, ..., w_m\}$ denotes the set of distinct equivalence classes for a particular W, $W(\cdot)$ is regarded as a symbolic mapping from the set of labels to the set of equivalence classes, and $\sum^*$ denotes the summation subject to the condition that $w_1, \ldots, w_m$ all take distinct values. Note that the number of partitions depends only on J. For a fixed partition W, we denote

$$R(w_1, ..., w_m; W) := \prod_{j=1}^{J} X_{b_{k_j} W(\mu_j)} G_{W(\mu_j)W(\nu_j)} X_{b_{l_j} W(\nu_j)}.$$

Then to prove (3.157), it suffices to show that

$$|\mathbb{E}R(w_1,...,w_m;W)| \prec N^{-m-h/2}\Phi^{2p}. \tag{3.158}$$

for any partition W.

To facilitate the proof, we introduce the graphical notations as in Section 3.4.4. We use a connected graph (V,E) to represent R, where the vertex set V consists of black vertices $b_1,\dots,b_n$ and white vertices $w_1,\dots,w_m$, and the edge set E consists of (k,α) edges representing $X_{b_k w_\alpha}$ and (α,β) edges representing $G_{w_\alpha w_\beta}$. We denote

$$e_{k\alpha} := \text{number of } (k,\alpha) \text{ edges in } R, \quad d_\alpha := \text{number of } (\alpha,\alpha) \text{ edges in } R.$$

Note that to attain a nonzero expectation, we must have

$$e_{k\alpha} = 0 \;\text{ or }\; e_{k\alpha} \ge 2 \quad \text{for all } k,\alpha. \tag{3.159}$$

We also define

$$e^{(o)}_{k\alpha} := \text{number of } (k,\alpha) \text{ edges that are from off-diagonal } S \text{ in } Q.$$

Then we have

$$\sum_\alpha e^{(o)}_{k\alpha} = \deg_o(b_k,Q) \tag{3.160}$$

By (3.152), (3.159) and the parity conservation due to (3.156), there exist edges $(1,\alpha_1),...,(h,\alpha_h)$ such that $e_{k\alpha_k}$ is odd and $e_{k\alpha_k}\ge 3$, $1\le k\le h$. Let $H:=\{(1,\alpha_1),...,(h,\alpha_h)\}$ be the set of these edges. Denote by F the set of (k,α) edge such that $e_{k\alpha}\ge 2$ and $(k,\alpha)\notin H$. Denote

$$s_\alpha := \sum_{k=1}^n e_{k\alpha}, \quad h_{k\alpha} := \mathbf{1}_{(k,\alpha)\in H}, \quad h_\alpha := \sum_{k=1}^n h_{k\alpha}, \quad f_\alpha := \sum_{k=1}^n \mathbf{1}_{(k,\alpha)\in F}$$

for all $k = 1, ..., n$ and $\alpha = 1, ..., m$. By the above definitions, we have $s_\alpha \geq 2$ and $h_\alpha + f_\alpha > 0$ (since the classes w_α are nontrivial), $s_\alpha \geq 2d_\alpha$, and

$$\sum_\alpha h_{k\alpha} = \mathbf{1}(1 \leq k \leq h), \quad \sum_\alpha h_\alpha = h. \tag{3.161}$$

Note that there are $\frac{1}{2}\sum_{k,\alpha} e_{k\alpha} - d_\alpha$ off-diagonal G edges in R. Hence by (3.87) and (3.150), we have

$$\begin{aligned} |\mathbb{E}R| &\prec \prod_{\alpha=1}^{m} \left(\Phi^{-d_\alpha} \prod_{k=1}^{n} \Phi^{\frac{1}{2}e_{k\alpha}} \mathbb{E}|X_{b_k w_\alpha}|^{e_{k\alpha}} \right) \\ &\prec \prod_{\alpha=1}^{m} \Phi^{s_\alpha/2-d_\alpha} \Big(\prod_{(k,\alpha)\in H} q^{e_{k\alpha}-3} N^{-3/2} \Big) \Big(\prod_{(k,\alpha)\in F} q^{e_{k\alpha}-2} N^{-1} \Big) =: \prod_{\alpha=1}^{m} R_\alpha. \end{aligned}$$

Now we consider the following four cases for R_α.

(i) $d_\alpha = 0$. In this case we have

$$\begin{aligned} R_\alpha &\prec \Phi^{s_\alpha/2} \prod_{(k,\alpha)\in H} N^{-3/2} \prod_{(k,\alpha)\in F} N^{-1} = \Phi^{s_\alpha/2} (N^{-1})^{h_\alpha + f_\alpha} N^{-h_\alpha/2} \\ &\prec \Phi^{s_\alpha/2} N^{-1} N^{-h_\alpha/2} \prec \Phi^{\sum_{k=1}^{h} h_{k\alpha}/2 + \sum_{k=h+1}^{n} e_{k\alpha}^{(o)}/2} N^{-1} N^{-h_\alpha/2} \end{aligned}$$

where in the third step we used $h_l + f_l > 0$, and in the fourth step we used

$$s_\alpha \geq \sum_k e_{k\alpha}^{(o)} \geq \sum_{k=1}^{h} h_{k\alpha} + \sum_{k=h+1}^{n} e_{k\alpha}^{(o)},$$

where we used that $e_{k\alpha}^{(o)} \geq h_{k\alpha}$ for $1 \leq k \leq h$ (recall that if $(k, \alpha) \in H$, then $e_{k\alpha_k}$ is odd and hence one of the edges must come from the off-diagonal S).

(ii) $d_\alpha \neq 0$, $h_\alpha = 1$ and $f_\alpha = 0$. Then there is only one k such that $e_{k\alpha} > 0$ and $s_\alpha = e_{k\alpha}$ is odd. Hence we have $s_l/2 \geq d_l + 1/2$ and we can bound R_α as

$$R_\alpha \prec \Phi^{\frac{1}{2}s_\alpha - d_\alpha} (N^{-1})^{h_\alpha + f_\alpha} N^{-h_\alpha/2} \prec \Phi^{1/2} N^{-1} N^{-h_\alpha/2}$$

$$= \Phi^{\sum_{k=1}^{h} h_{k\alpha}/2+\sum_{k=h+1}^{n} e^{(o)}_{k\alpha}/2} N^{-1} N^{-h_\alpha/2},$$

where in the last step we used

$$1 = \sum_{k=1}^{h} h_{k\alpha} + \sum_{k=h+1}^{n} e^{(o)}_{k\alpha},$$

since all the summands except one $h_{k\alpha}$ are 0.

(iii) $d_\alpha \neq 0$, $h_\alpha = 0$ and $f_\alpha = 1$. Then there is only one k such that $e_{k\alpha} > 0$ and $s_\alpha = e_{k\alpha}$. Thus the (α, α) edges are expanded from the diagonal S variables (otherwise α must connect to at least two different k's), which implies $\frac{1}{2}s_\alpha - d_\alpha = \frac{1}{2}e^{(o)}_{k\alpha}$. Then we can bound R_α by

$$\begin{aligned} R_\alpha &\prec \Phi^{\frac{1}{2}s_\alpha - d_\alpha} \big(N^{-1}\big)^{h_\alpha + f_\alpha} N^{-h_\alpha/2} = \Phi^{\sum_k e^{(o)}_{k\alpha}/2} N^{-1} N^{-h_\alpha/2} \\ &\prec \Phi^{\sum_{k=1}^{h} h_{k\alpha}/2+\sum_{k=h+1}^{n} e^{(o)}_{k\alpha}/2} N^{-1} N^{-h_\alpha/2} \end{aligned}$$

where, as in Case (i), we used $e^{(o)}_{k\alpha} \geq h_{k\alpha}$ for $1 \leq k \leq h$.

(iv) $d_\alpha \neq 0$ and $h_\alpha + f_\alpha \geq 2$. Then using $s_\alpha \geq 2d_\alpha$, $q \prec \Phi^{1/2}$ and $N^{-1/2} \prec \Phi$, we get that

$$\begin{aligned} R_\alpha &\prec \prod_{(k,\alpha)\in H} \Phi^{e_{k\alpha}/2-3/2} N^{-3/2} \prod_{(k,\alpha)\in F} \Phi^{e_{k\alpha}/2-1} N^{-1} \\ &\prec \prod_{(k,\alpha)\in H} \Phi^{e_{k\alpha}/2-1/2} N^{-1} \prod_{(k,\alpha)\in F} \Phi^{e_{k\alpha}/2} N^{-1/2} \\ &= \Phi^{(s_\alpha - h_\alpha)/2} N^{-(h_\alpha+f_\alpha)/2} N^{-h_\alpha/2} \leq \Phi^{(s_\alpha - h_\alpha)/2} N^{-1} N^{-h_\alpha/2} \\ &\leq \Phi^{\sum_{k=1}^{h} h_{k\alpha}/2+\sum_{k=h+1}^{n} e^{(o)}_{k\alpha}/2} N^{-1} N^{-h_\alpha/2} \end{aligned}$$

where in the last step we used the definitions of s_α and h_α, $e_{k\alpha} \geq 2h_{k\alpha}$ for $1 \leq k \leq h$ (since $e_{k\alpha} \geq 3$ whenever $h_{k\alpha} = 1$), and $h_{k\alpha} = 0$ for $k \geq h + 1$.

Combining the above four cases, we obtain that

$$|\mathbb{E}R| = \prod_{\alpha=1}^{m} R_\alpha \prec N^{-m} N^{-\frac{1}{2}\sum_\alpha h_\alpha} \Phi^{\sum_\alpha \left(\sum_{k=1}^{h} h_{k\alpha}/2 + \sum_{k=h+1}^{n} e_{k\alpha}^{(o)}/2\right)}.$$

Recall that $\sum_\alpha h_\alpha = h$. Then to prove (3.158), it remains to show that

$$\sum_\alpha \left(\sum_{k=1}^{h} h_{k\alpha} + \sum_{k=h+1}^{n} e_{k\alpha}^{(o)} \right) \ge 4p. \tag{3.162}$$

For $k = 1, ..., h$, using (3.161) and (3.152) we get that

$$\sum_{\alpha=1}^{m} h_{k\alpha} = 1 = \deg(b_k, P).$$

For $k = h+1, ..., n$, using (3.160) and (3.156) we get that

$$\sum_{\alpha=1}^{m} e_{k\alpha}^{(o)} = \deg_o(b_k, Q) \ge \deg(b_k, P).$$

With (3.152), we then conclude (3.162), which finishes our proof.

Bibliography

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

[2] ——, *Circular law*, Ann. Probab., 25 (1997), pp. 494–529.

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

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

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

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

[7] A. Bloemendal, A. Knowles, H.-T. Yau, and J. Yin, *On the principal components of sample covariance matrices*, Prob. Theor. Rel. Fields, 164 (2016), pp. 459–552.

[8] A. Borodin and C. D. Sinclair, *The Ginibre ensemble of real random matrices and its scaling limits*, Commun. Math. Phys., 291 (2009), pp. 177–224.

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

[10] P. Bourgade, H.-T. Yau, and J. Yin, *Local circular law for random matrices*, Probab. Theory Relat. Fields, 159 (2014), pp. 545–595.

[11] ——, *The local circular law II: the edge case*, Probab. Theory Relat. Fields, 159 (2014), pp. 619–660.

[12] C. Cacciapuoti, A. Maltsev, and B. Schlein, *Local marchenko-pastur law at the hard edge of sample covariance matrices*, Journal of Mathematical Physics, 54 (2013).

[13] K. R. Davidson and S. J. Szarek, *Local operator theory, random matrices and banach spaces*, vol. 1 of Handbook of the Geometry of Banach Spaces, North-Holland, Amsterdam, 2001, pp. 317 – 366.

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

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

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

[17] F. Dyson, *The threefold way. algebraic structure of symmetry groups and ensembles in quantum mechanics*, J. Math. Phys., 3 (1962), pp. 1199–1215.

[18] A. Edelman, *The probability that a random real gaussian matrix has k real eigenvalues, related distributions, and the circular law*, J. Multivar. Anal., 60 (1997), pp. 203 – 232.

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

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

[21] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *Delocalization and diffusion profile for random band matrices*, Commun. Math. Phys., 323 (2013), pp. 367–416.

[22] ——, *The local semicircle law for a general class of random matrices*, Electron. J. Probab., 18 (2013), pp. 1–58.

[23] ——, *Spectral statistics of Erdős-Rényi graphs I: Local semicircle law*, Ann. Probab., 41 (2013), pp. 2279–2375.

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

[25] ——, *Semicircle law on short scales and delocalization of eigenvectors for wigner random matrices*, Ann. Probab., 37 (2009), pp. 815–852.

[26] ——, *Wegner estimate and level repulsion for wigner random matrices*, Int. Math. Res. Notices, (2010), pp. 436–479.

[27] ——, *Universality of random matrices and local relaxation flow*, Invent. Math, 185 (2011), pp. 75–119.

[28] L. Erdős, B. Schlein, H.-T. Yau, and J. Yin, *The local relaxation flow approach to universality of the local statistics for random matrices*, Ann. Inst. H. Poincaré Probab. Statist., 48 (2012), pp. 1–46.

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

[30] ——, *Bulk universality for generalized Wigner matrices*, Probab. Theory Relat. Fields, 154 (2012), pp. 341–407.

[31] ——, *Rigidity of eigenvalues of generalized Wigner matrices*, Advances in Mathematics, 229 (2012), pp. 1435 – 1515.

[32] P. J. Forrester and T. Nagao, *Eigenvalue statistics of the real Ginibre ensemble*, Phys. Rev. Lett., 99 (2007), p. 050603.

[33] J. Ginibre, *Statistical ensembles of complex, quaternion, and real matrices*, J. Math. Phys., 6 (1965), pp. 440–449.

[34] V. Girko, *The circular law*, Russ. Teor. Veroyatnost. i Primenen., 29 (1984), pp. 669–679.

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

[36] ——, *The circular law for random matrices*, Ann. Probab., 38 (2010), pp. 1444–1491.

[37] A. Guionnet, M. Krishnapur, and O. Zeitouni, *The single ring theorem*, Ann. Math., 174 (2011), pp. 1189–1217.

[38] W. Hachem, A. Hardy, and J. Najim, *Large complex correlated wishart matrices: Fluctuations and asymptotic independence at the edges*, arXiv:1409.7548.

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

[40] ——, *On the distribution of the largest eigenvalue in principal components analysis*, Ann. Statist., 29 (2001), pp. 295–327.

[41] ——, *Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy-Widom limits and rates of convergence*, Ann. Statist., 36 (2008), pp. 2638–2716.

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

[43] ——, *The isotropic semicircle law and deformation of wigner matrices*, Comm. Pure Appl. Math., 66 (2013), pp. 1663–1749.

[44] ——, *Anisotropic local laws for random matrices*, Probability Theory and Related Fields, (2016), pp. 1–96.

[45] J. O. Lee and K. Schnelli, *Tracy-widom distribution for the largest eigenvalue of real sample covariance matrices with general population*, arXiv:1409.4979.

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

[47] A. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann, *Smallest singular value of random matrices and geometry of random polytopes*, Adv. Math., 195 (2005), pp. 491 – 523.

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

[49] M. L. Mehta, *Random matrices*, vol. 142 of Pure and Applied Mathematics, Elsevier, Amsterdam, 3 ed., 2004.

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

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

[52] G. Pan and W. Zhou, *Circular law, extreme singular values and potential theory*, J. Multivar. Anal., 101 (2010), pp. 645–656.

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

[54] N. S. Pillai and J. Yin, *Universality of covariance matrices*, Ann. Appl. Probab., 24 (2014), pp. 935–1001.

[55] M. Rudelson, *Invertibility of random matrices: norm of the inverse*, Ann. Math., 168 (2008), pp. 575–600.

[56] M. Rudelson and R. Vershynin, *The littlewood-offord problem and invertibility of random matrices*, Adv. Math., 218 (2008), pp. 600–633.

[57] ——, *The smallest singular value of a random rectangular matrix*, Comm. Pure Appl. Math., 62 (2009), pp. 1707–1739.

[58] ——, *Small ball probabilities for linear images of high-dimensional distributions*, Int. Math. Res. Notices, 2015 (2015), pp. 9594–9617.

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

[60] ——, *Weak convergence of random functions defined by the eigenvectors of sample covariance matrices*, Ann. Probab., 18 (1990), pp. 1174–1194.

[61] C. D. Sinclair, *Averages over Ginibre's ensemble of random real matrices*, Int. Math. Res. Not., 2007 (2007), pp. 1–15.

[62] T. Tao and V. Vu, *Random matrices: the circular law*, Commun. Contemp. Math., 10 (2008), pp. 261–307.

[63] ——, *Random matrices: Universal properties of eigenvectors*, Random Matrices: Theory and Applications, 01 (2012), p. 1150001.

[64] ——, *Random matrices: Universality of local spectral statistics of non-Hermitian matrices*, Ann. Probab., 43 (2015), pp. 782–874.

[65] T. Tao, V. Vu, and M. Krishnapur, *Random matrices: Universality of ESDs and the circular law*, Ann. Probab., 38 (2010), pp. 2023–2065.

[66] E. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, Ann. Math., 62 (1955), pp. 548–564.

[67] J. Wishart, *The generalised product moment distribution in samples from a normal multivariate population*, Biometrika, 20A (1928), pp. 32–52.

[68] H. Xi, F. Yang, and J. Yin, *Local circular law for the product of a deterministic matrix with a random matrix*, arXiv:1603.04066.

[69] N. Xia and Z. Bai, *Convergence rate of eigenvector empirical spectral distribution of large Wigner matrices*, Statistical Papers, (2016), pp. 1–33.

[70] N. Xia, Y. Qin, and Z. Bai, *Convergence rates of eigenvector empirical spectral distribution of large dimensional sample covariance matrix*, Ann. Statist., 41 (2013), pp. 2572–2607.

[71] F. Yang, *The smallest singular value of deformed random rectangular matrices*, arXiv:1702.04050.

[72] J. Yin, *The local circular law III: general case*, Probab. Theory Relat. Fields, 160 (2014), pp. 679–732.