Electronic Journal of Probability

Electron. J. Probab. **22** (2017), no. 60, 1–77.
ISSN: 1083-6489 DOI: 10.1214/17-EJP76

Local circular law for the product of a deterministic matrix with a random matrix*

Haokai Xi † Fan Yang † Jun Yin †

Abstract

It is well known that the spectral measure of eigenvalues of a rescaled square non-Hermitian random matrix with independent entries satisfies the circular law. In this paper, we consider 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 a general local circular law for the empirical spectral distribution (ESD) of TX at any point z away from the unit circle under the assumptions that $N \sim M$, and the matrix entries X_{ij} have sufficiently high moments. More precisely, if z satisfies $||z|-1| \ge \tau$ for arbitrarily small $\tau > 0$, the 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 dA 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 $\epsilon > 0$.

Keywords: local circular law; anisotropic local law; deformation; universality.
AMS MSC 2010: Primary 15B52, Secondary 60B20; 82B44.
Submitted to EJP on June 5, 2016, final version accepted on June 15, 2017.
Supersedes arXiv:1603.04066.

1 Introduction

Circular law for non-Hermitian random matrices. The study of the eigenvalue spectral of non-Hermitian random matrices goes back to the celebrated paper [19] 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 [19, 28]), 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 [6, 10, 18, 35].

*Partially supported by NSF Career Grant DMS-1552192 and Sloan fellowship.
†University of Wisconsin-Madison, USA. E-mail: haokai@math.wisc.edu, fyang75@math.wisc.edu, jyin@math.wisc.edu

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. A key step in this direction is made by Girko in [20], where he partially proved the circular law for non-Hermitian matrices with independent entries. The crucial insight of this 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,

$$\operatorname{Tr}\log[(X - z)^\dagger (X - z)] = \log \det((X - z)^\dagger (X - z)) , \tag{1.1}$$

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 [20], but the gap was remedied later in a series of paper. Bai [1, 2] 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 [31, 32], and subsequently by Tao and Vu [36], Pan and Zhou [30] and Gőtze and Tikhomirov [21] under weakened moments and smoothness assumptions. The final result was presented in [38], where the circular law is proved under the optimal L^2 assumption. These papers studied the circular law in the global regime, i.e. the convergence of ESD on subsets containing ϵN eigenvalues for some small constant $\epsilon > 0$. Later in a series of papers [7, 8, 40], 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 [37], the local universality was proved by Tao and Vu under the assumption of first four moments matching the moments of a Gaussian random variable.

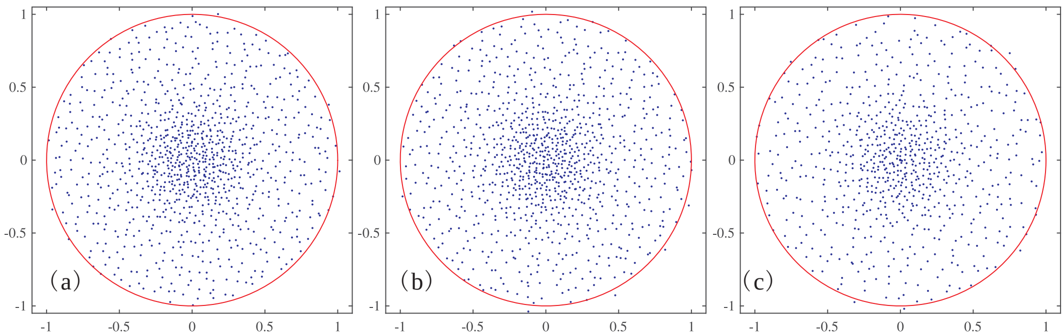


Figure 1: The eigenvalue distribution of the product TX of a deterministic $N \times M$ matrix T with a Gaussian random $M \times N$ matrix X. The entries of X have zero mean and variance $(N \wedge M)^{-1}$, and $TT^\dagger$ has $0.5(N \wedge M)$ eigenvalues as $2/17$ and $0.5(N \wedge M)$ eigenvalues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

In this paper, 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 $N \le M$. In Figure 1, we plot the eigenvalue distribution of TX when T has two distinct singular values (except the trivial zero singular values). The goal of this paper is to prove a local circular law for the ESD of TX at any point z away from the unit circle. Following the idea in [7], 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 X through a standard large deviation estimate (see e.g. [9, 27, 33]) or by studying the eigenvalue rigidity of $X^* X$ (see e.g. [4] and (2.64)). The lower bound for the least singular value of $TX - z$ follows from the results in e.g. [32, 36, 39] (see also Lemma 2.23). Thus the bulk of this paper is devoted to establishing (c).

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 [7]. 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 $N^{-1}\mathrm{Tr}\, R(w)$, and we have the convergence estimate

$$N^{-1}\mathrm{Tr}\, R(w) \approx m_c(w) \tag{1.2}$$

with high probability for large N. Here m_c is the Stieltjes transform of the asymptotic eigenvalue density, and the convergence in (1.2) is referred to as the *averaged law*. By taking the imaginary part of (1.2), it is easy to see 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 (1.2) for all $\eta \gg N^{-1}$. Such local laws have been a cornerstone of the modern random matrix theory. In [16], a local law was first derived for Wigner matrices. Subsequently in [7], 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 [7] to our setting, a main difficulty is that the entries of TX are not independent. We will use a new comparison method proposed in [24], which roughly states that if the local law holds for $R(w)$ with a Gaussian X, then it also holds in the case with a general X. For definiteness, we assume $N = M$ for now, and let T be a square matrix with singular decomposition $T = UDV$. For a Gaussian $X \equiv X^{Gauss}$, we have $VX^{Gauss}U \stackrel{d}{=} \tilde{X}^{Gauss}$, where $\tilde{X}$ is another Gaussian random matrix. Then for the determinant in (1.1), we have

$$\det(TX^{Gauss} - z) = \det(DVX^{Gauss}U - z) \stackrel{d}{=} \det(D\tilde{X}^{Gauss} - z). \tag{1.3}$$

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 [14], where a local law was obtained for generalized Wigner matrices with non-identically distributed entries.

To use the comparison method invented in [24], it turns out the averaged local law from (1.2) 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 $R_{ij}(w) \approx \Pi_{ij}(w)$ (see e.g. [7, 17]). We call such an estimate an *entrywise local law*. More generally, in [4, 25] the following closeness was established for *generalized matrix entries*:

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

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.) This kind of anisotropic local law is needed in applying the method in [24]. 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 2.18); (B) the anisotropic local law when T is diagonal (Theorem 2.18); (C) the anisotropic local law and averaged local law when T is a general (rectangular) matrix (Theorem 2.19).

In performing Step (A), our proof is basically based on the methods in [7]. 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 A). In performing Step (B), we applied and extended the polynomialization method developed in [4, section 5]. Finally, as remarked around (1.3), (B) implies the anisotropic local law and averaged local law for Gaussian X and general T. Based on this fact we perform Step (C) using a self-consistent comparison argument in [24]. With the averaged local law proved in Step (C), we can obtain a generalized (inhomogeneous) 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 2.4). 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.

Conventions. The fundamental large parameter is N and we assume that M is comparable to N (see (2.1)). 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 ϵ, ϵ' 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, $\|A\| := \|A\|_{l^2 \to l^2}$ 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 sequence $\{c_N\}$ with $c_N \to 0$ as $N \to \infty$. If A_N is a deterministic 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 The main results

In this section, we state and prove the main result of this paper. In Section 2.1, we define our model and list our main assumptions. In Section 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 2.6—of this paper. Its proof depends crucially on local estimates of the resolvent of Q, which are presented in Section 2.3. In Section 2.4, we prove Theorem 2.6 based on the local estimates stated in Section 2.3.

2.1 Definition of the model

In this paper, 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.

$$\tau \le \frac{M}{N} \le \tau^{-1} \tag{2.1}$$

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

$$\mathbb{E} X_{i\mu} = 0, \quad \mathbb{E} |X_{i\mu}|^2 = \frac{1}{N \wedge M} \tag{2.2}$$

for all $1 \le i \le M, 1 \le \mu \le N$. For definiteness, in this paper 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

$$\mathbb{E} |\sqrt{N \wedge M} X_{i\mu}|^p \le C_p \tag{2.3}$$

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

$$\tau^{-1} \ge \sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_{N \wedge M} \ge \tau \tag{2.4}$$

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

$$\frac{1}{N \wedge M} \sum_{i=1}^{N \wedge M} \sigma_i = 1. \tag{2.5}$$

We summarize our basic assumptions here for future reference.

Assumption 2.1. *We suppose that (2.1), (2.2), (2.3), (2.4) and (2.5) hold.*

2.2 The 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

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

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

$$\tau^{-1} \ge s_1 > s_2 > \ldots > s_n \ge \tau. \tag{2.7}$$

Let l_i be the multiplicity of s_i. By (2.5), l_i and s_i satisfy the normalization conditions

$$\frac{1}{N \wedge M} \sum_{i=1}^{n} l_i = 1, \quad \frac{1}{N \wedge M} \sum_{i=1}^{n} l_i s_i = 1. \tag{2.8}$$

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

$$\frac{1}{m_2} = -w(1 + m_1) + \frac{|z|^2}{1 + m_1}, \tag{2.9}$$

$$m_1 = \frac{1}{N}\sum_{i=1}^{N} l_i s_i \left[w\left(1+s_i m_2\right) + \frac{|z|^2}{1+m_1}\right]^{-1}. \tag{2.10}$$

If we plug (2.9) into (2.10), we get the self-consistent equation for m_1 only:

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

The next lemma states that the solution to the functional equation (2.11) in $\mathbb{C}_+$ is unique if z is away from the unit circle. It will be proved in Appendix A.3.

Lemma 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 (2.11) 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 (2.9). Obviously, m_{2c} is also an analytic function of w. Moreover, for any $w \in \mathbb{C}_+$ we can verify that $m_{2c}(w), w m_{2c}(w) \in \mathbb{C}_+$ by using (2.9) and that $m_{1c}, w m_{1c} \in \mathbb{C}_+$. We define two functions on $\mathbb{R}$ as

$$\rho_{1,2c}(x) = \frac{1}{\pi}\lim_{\eta \searrow 0} \operatorname{Im} m_{1,2c}(x+i\eta), \quad x \in \mathbb{R}. \tag{2.12}$$

It is easy to see that $\rho_{1,2c} \geq 0$ and $\operatorname{supp}(\rho_{1,2c}) \subset [0,\infty)$. Moreover, $\operatorname{supp}\rho_{2c} = \operatorname{supp}\rho_{1c}$ by (2.9). 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 4). Since $\operatorname{Im}(w m_{2c}) \geq 0$, we have

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

and (2.10) 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 $\rho_{1,2c}$:

$$m_{1,2c}(w) = \int_{\mathbb{R}} \frac{\rho_{1,2c}(x)}{x-w}dx. \tag{2.13}$$

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 (2.9) and (2.11) 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 (2.11) then becomes

$$f(\sqrt{w}, m) = 0, \tag{2.14}$$

where

$$f(\sqrt{w}, m) = -\sqrt{w} + m + \frac{1}{N}\sum_{i=1}^{N} l_i s_i \frac{m(m^2-|z|^2)}{\sqrt{w}m^3 - (s_i+|z|^2)m^2 - \sqrt{w}|z|^2 m + |z|^4}. \tag{2.15}$$

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

Lemma 2.3. *Fix $\tau \le ||z|^2 - 1| \le \tau^{-1}$. The support of $\rho_{1,2c}$ is a union of connected components:*

$$\operatorname{supp}\rho_{1,2c} \cap (0, +\infty) = \left(\bigcup_{1 \le k \le L} [e_{2k}, e_{2k-1}] \right) \cap (0, \infty), \tag{2.16}$$

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

$$\partial_m f(\sqrt{e_i}, m_c(e_i)) = 0. \tag{2.17}$$

We shall call e_i the edges of $\rho_{1,2c}$. For any $w \in (0, \infty)$ and $1 \le i \le n$, the cubic polynomial $\sqrt{w}m^3 - (s_i + |z|^2)m^2 - \sqrt{w}|z|^2 m + |z|^4$ in (2.15) has three distinct roots $a_i(w) > 0$, $b_i(w) > 0$ and $-c_i(w) < 0$ (see Lemma A.1). Our next assumption on Σ and $|z|$ takes the form of the following regularity conditions.

Definition 2.4. *(Regularity) Fix $\tau \le ||z|^2 - 1| \le \tau^{-1}$.*

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

$$\min_{1 \le i \le n} \{ |m_c(e_k) - a_i(e_k)|, |m_c(e_k) - b_i(e_k)|, |m_c(e_k) + c_i(e_k)| \} \ge \epsilon, \tag{2.18}$$

and

$$|\partial_m^2 f(\sqrt{e_k}, m_c(e_k))| \ge \epsilon, \tag{2.19}$$

for some small constant $\epsilon > 0$. In the case $|z|^2 \le 1 - \tau$, 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 [3, 11, 23, 24, 26, 29]. The conditions (2.18) and (2.19) 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 A.5):

$$\min_{l \ne k} |e_l - e_k| \ge \epsilon \tag{2.20}$$

for some constant $\epsilon > 0$. The bulk regularity condition (ii) was introduced in [24]. 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 2.4 are stable under perturbations of $|z|$ and Σ. In particular, fix τ, suppose the regularity conditions are satisfied at $z = z_0$ with $\tau \le ||z_0|^2 - 1| \le \tau^{-1}$. Then for sufficiently small $c > 0$, the regularity conditions hold uniformly in $z \in \{z : ||z| - |z_0|| \le c\}$. For a detailed discussion, see the remark at the end of Section A.3.

We will use the following notion of stochastic domination, which was first introduced in [12] and subsequently used in many works on random matrix theory, such as [4, 5, 7, 13, 14, 24]. 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 2.5 (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) $D > 0$,

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

for large enough $N \ge N_0(\epsilon, D)$, and we use the notation $\xi \prec \zeta$. Throughout this paper 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 (2.1), (2.3) and (2.4).

(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 be 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 $1 - \mathbf{1}(\Xi) \prec 0$.

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

Theorem 2.6 (Local circular law for TX)**.** *Suppose Assumption 2.1 holds, and $||z_0|^2 - 1| \le \tau^{-1}$ for any N (z_0 can depend on N). Suppose Σ (defined in (2.6)) and $|z_0|$ are such that all the edges and bulk components of ρ_{1c} are regular in the sense of Definition 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*

$$\left|\frac{1}{K}\sum_{j=1}^{K} F_{z_0,a}(\mu_j) - \frac{1}{\pi}\int F_{z_0,a}(z)\tilde{\rho}_{\mathbb{D}}(z)dA(z)\right| \prec K^{-1/2+2a}\|\Delta F\|_{L^1}, \tag{2.21}$$

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

$$\tilde{\rho}_{\mathbb{D}}(z) := \frac{1}{4\pi}\int_0^\infty (\log x)\Delta_z \rho_{2c}(x, z)dx, \tag{2.22}$$

where $\rho_{2c} \equiv \rho_{2c,z,\Sigma}$ is defined in (2.12). If $1 + \tau \le |z_0|^2 \le 1 + \tau^{-1}$ or the entries of X have vanishing third moments,

$$\mathbb{E}X_{i\mu}^3 = 0, \quad 1 \le i \le M, 1 \le \mu \le N, \tag{2.23}$$

then we have the improved result

$$\left|\frac{1}{K}\sum_{j=1}^{K} F_{z_0,a}(\mu_j) - \frac{1}{\pi}\int F_{z_0,a}(z)\tilde{\rho}_{\mathbb{D}}(z)dA(z)\right| \prec K^{-1+2a}\|\Delta F\|_{L^1}, \tag{2.24}$$

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 (2.21) gives a generalized circular law up to scale $K^{-1/4+\epsilon}$, while (2.24) gives a generalized circular law up to scale $K^{-1/2+\epsilon}$. The $\tilde{\rho}_{\mathbb{D}}$ in (2.22) gives the distribution of the eigenvalues of TX. It is rotationally symmetric, because $\rho_{2c}(x, z)$ depends only on $|z|$ (see (2.9) and (2.10)). If $TT^* = 1$ or $T^*T = 1$ (i.e. all the nontrivial singular values of T are equal to 1), then $\tilde{\rho}_{\mathbb{D}}$ 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 [7] for the $T = I$ case). For a general T, we do not have much understanding of $\tilde{\rho}_{\mathbb{D}}$ 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, the basic strategy of this paper 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 (2.23), our comparison arguments cannot give the anisotropic local law up to the optimal scale, so we can only prove the weaker bound (2.21). 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 [39] 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 2.23 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 2.4.

Example 2.7 (Bounded number of distinct eigenvalues)**.** We suppose that n is fixed, and that $s_1, \ldots, s_n$ and $\pi(\{s_1\}), \ldots, \pi(\{s_n\})$ 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 2.4 for small enough τ.

Example 2.8 (Continuous limit)**.** We suppose π is supported in some interval $[a, b] \subset (0, \infty)$, and that π converges in distribution to some measure π_∞ that is absolutely continuous and whose density satisfies $\tau \le d\pi_\infty(E)/dE \le \tau^{-1}$ for $E \in [a, b]$. Then there are only a small number (which is independent of n) of connected components for supp ρ_{1c}, and all the edges and bulk components are regular. See the remark at the end of Section A.1.

2.3 Hermitization and local laws for resolvents

In the following, we use the notation

$$Y \equiv Y_z := TX - zI, \tag{2.25}$$

where I is the identity matrix. Following Girko's Hermitization technique [20], 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 $(YY^\dagger - w)^{-1}$ and $(Y^\dagger Y - w)^{-1}$. These results will be used later to prove Theorem 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 2.9 (Index sets)**.** *We define the index sets*

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

and

$$I := I_1 \cup I_2, \quad I^M := I_1^M \cup I_2.$$

We will consistently use the latin letters $i, j \in I_1$ or I_1^M, greek letters $\mu, \nu \in I_2$, and $s, t \in I$. We label the indices of the matrices according to

$$X = (X_{i\mu} : i \in I_1^M, \mu \in I_2), \quad T = (T_{ij} : i \in I_1, j \in I_1^M).$$

When $M = N$, we always identify I_1^M with I_1. For $i \in I_1$ and $\mu \in I_2$, we introduce the notations $\bar{i} := i + N \in I_2$ and $\bar{\mu} := \mu - N \in I_1$.

Definition 2.10 (Groups)**.** *For an $I \times I$ matrix A, we define the 2×2 matrices $A_{[ij]}$ as*

$$A_{[ij]} = \begin{pmatrix} A_{ij} & A_{i\bar{j}} \\ A_{\bar{i}j} & A_{\bar{i}\bar{j}} \end{pmatrix}. \tag{2.26}$$

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

Definition 2.11 (Linearizing block matrix)**.** *For $w := E + i\eta \in \mathbb{C}_+$, we define the $I \times I$ matrix*

$$H(w) \equiv H(T, X, z, w) := \begin{pmatrix} -wI & w^{1/2} Y \\ w^{1/2} Y^{\dagger} & -wI \end{pmatrix}, \tag{2.27}$$

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

$$G(w) \equiv G(T, X, z, w) := H(w)^{-1}, \tag{2.28}$$

as well as the $I_1 \times I_1$ and $I_2 \times I_2$ matrices

$$G_L(w) = \left(YY^{\dagger} - w \right)^{-1}, \quad G_R(w) = \left(Y^{\dagger}Y - w \right)^{-1}. \tag{2.29}$$

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

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

$$G(w) = \begin{pmatrix} G_L & w^{-1/2} G_L Y \\ w^{-1/2} Y^{\dagger} G_L & w^{-1} Y^{\dagger} G_L Y - w^{-1} I \end{pmatrix} = \begin{pmatrix} w^{-1} Y G_R Y^{\dagger} - w^{-1} I & w^{-1/2} Y G_R \\ w^{-1/2} G_R Y^{\dagger} & G_R \end{pmatrix}. \tag{2.30}$$

Therefore a control of G immediately yields a control 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 2.12 (Deterministic limit of G)**.** *Suppose $N \le M$ and T has a singular decomposition*

$$T = U \bar{D} V, \quad \bar{D} = (D, 0), \tag{2.31}$$

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

$$\left(\pi_{[i]c}\right)^{-1} = \begin{pmatrix} -w(1 + |d_i|^2 m_{2c}) & -w^{1/2} z \\ -w^{1/2} \bar{z} & -w(1 + m_{1c}) \end{pmatrix}. \tag{2.32}$$

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

$$\Pi \equiv \Pi(\Sigma, z, w) := \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix} \Pi_d \begin{pmatrix} U^{\dagger} & 0 \\ 0 & U^{\dagger} \end{pmatrix} = \begin{pmatrix} (1 + m_{1c}) A(\Sigma) & -w^{-1/2} z A(\Sigma) \\ -w^{-1/2} \bar{z} A(\Sigma) & (1 + m_{2c}\Sigma) A(\Sigma) \end{pmatrix}, \tag{2.33}$$

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

Definition 2.13 (Averaged variables)**.** *Suppose $N \;\; M$. Define the averaged random variables*

$$m_1 := \frac{1}{N}\sum_{i \in I_1} \left(\bar{\Sigma} G\right)_{ii}; \quad m_2 := \frac{1}{N}\sum_{\in I_2} \left(\bar{\Sigma} G\right) ; \tag{2.34}$$

where

$$\bar{\Sigma} := \begin{pmatrix} \Sigma & 0 \\ 0 & I \end{pmatrix} : \tag{2.35}$$

Define $_{[i]}$ to be the $2 \;\; 2$ matrix such that

$$_{[i]}{}^{1} = \begin{pmatrix} w(1 + |d_i|^2 m_2) & w^{1/2} z \\ w^{1/2}\bar{z} & w(1 + m_1) \end{pmatrix} : \tag{2.36}$$

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

$$m_{1c} = \frac{1}{N}\sum_{i \in I_1} \left(\bar{\Sigma}\Pi\right)_{ii}; \quad m_{2c} = \frac{1}{N}\sum_{\in I_2} \left(\bar{\Sigma}\Pi\right) : \tag{2.37}$$

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

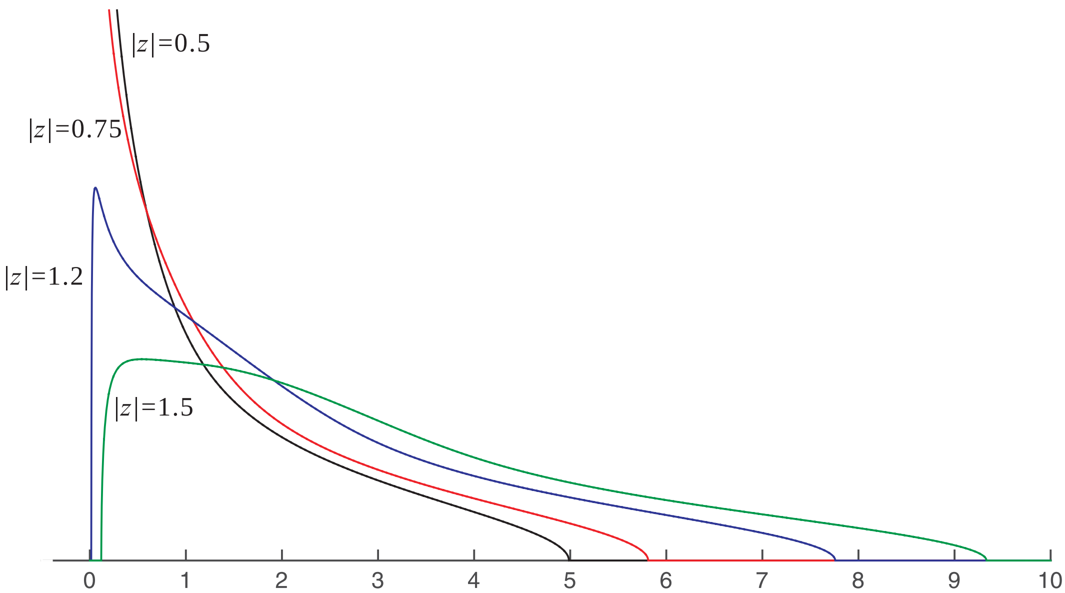


Figure 2: The densities $_{2c}(x; z)$ when $|z| = 0.5, 0.75, 1.2, 1.5$. Here $= 0.5 \; \sqrt{2/17} + 0.5 \; _4\sqrt{2/17}$.

Proposition 2.14 (Basic properties of $_{1;2c}$)**.** *The density $_{1c}$ is compactly supported in $[0, \infty)$ and the following properties regarding $_{1c}$ hold.*

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

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

$$|m_{1,2c}(w)| = O(|w|^{-1/2}). \tag{2.38}$$

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 paper, 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 2.16 (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*

$$\mathbf{D} \equiv \mathbf{D}(\zeta, N) := \{w \in \mathbb{C}_+ : e_{2L} \le E \le \zeta^{-1}, N^{-1+\zeta}|m_{2c}|^{-1} \le \eta \le \zeta^{-1}\}, \tag{2.39}$$

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

$$\mathbf{D}_k^e \equiv \mathbf{D}_k^e(\zeta, \tau', N) := \{w \in \mathbf{D}(\zeta, N) : |E - e_k| \le \tau', E \ge 0\}. \tag{2.40}$$

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

$$\mathbf{D}_k^b \equiv \mathbf{D}_k^b(\zeta, \tau', N) := \{w \in \mathbf{D}(\zeta, N) : E \in [e_{2k} + \tau', e_{2k-1} - \tau']\}. \tag{2.41}$$

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

$$\mathbf{D}^o \equiv \mathbf{D}^o(\zeta, \tau', N) := \{w \in \mathbf{D}(\zeta, N) : \operatorname{dist}(E, \operatorname{supp} \rho_{1c}) \ge \tau'\}. \tag{2.42}$$

We also need the following domain with large η,

$$\mathbf{D}_L \equiv \mathbf{D}_L(\zeta) := \{w \in \mathbb{C}_+ : 0 \le E \le \zeta^{-1}, \eta \ge \zeta^{-1}\}, \tag{2.43}$$

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

$$\widetilde{\mathbf{D}} \equiv \widetilde{\mathbf{D}}(\zeta, N) := \{w \in \mathbf{D}(\zeta, N) : \eta \ge N^{-1/2+\zeta}|m_{2c}|^{-1}\} \cup \mathbf{D}_L(\zeta). \tag{2.44}$$

We call $\mathbf{S}$ *a regular domain if it is a regular* $\mathbf{D}_k^e$ *domain, a regular* $\mathbf{D}_k^b$ *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 $E \ge 0$ is only useful for the edge at 0 when $|z|^2 \le 1 - \tau$.

Now we are prepared to state the local laws satisfied by G defined in (2.28). Let

$$\Psi \equiv \Psi(w) := \sqrt{\frac{\operatorname{Im}(m_{1c} + m_{2c})}{N\eta}} + \frac{1}{N\eta} \tag{2.45}$$

be the deterministic control parameter.

Definition 2.17 (Local laws). *Suppose $N \le M$. Recall $G \equiv G(T,X,z,w)$ defined in (2.28) and $\Pi \equiv \Pi(\Sigma,z,w)$ defined in (2.33). Let $\mathbf{S}$ be a regular domain.*

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

$$\left|[G(T,X,z,w) - \Pi(\Sigma,z,w)]_{st}\right| \prec \Psi(w) \tag{2.46}$$

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

$$G(T,X,z,w) - \Pi(\Sigma,z,w) = O_\prec(\Psi(w)) \tag{2.47}$$

uniformly in $w \in \mathbf{S}$ (recall Definition 2.5 (ii)).

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

$$|m_2(T,X,z,w) - m_{2c}(\Sigma,z,w)| \prec \frac{1}{N\eta} \tag{2.48}$$

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 2.18 (Local laws when T is diagonal). *Fix $\tau \le ||z|^2 - 1| \le \tau^{-1}$. Suppose Assumption 2.1 holds, $N = M$, and $T \equiv D := diag(d_1,\ldots,d_N)$ 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 (2.4) and (2.5). 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 $D = \mathrm{diag}(d_1,d_2,\ldots,d_N)$. Then we have

$$TX - z = UDV_1X - z, \tag{2.49}$$

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 $V_1X^{Gauss} \stackrel{d}{=} \tilde{X}^{Gauss}U^\dagger$, where $\tilde{X}^{Gauss}$ is another $N \times N$ Gaussian random matrix. Then by the definition of G in (2.28),

$$G(T,X^{Gauss},z,w) \stackrel{d}{=} \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix} G(D,\tilde{X}^{Gauss},z,w) \begin{pmatrix} U^\dagger & 0 \\ 0 & U^\dagger \end{pmatrix}. \tag{2.50}$$

Since the anisotropic local law holds for $G(D,\tilde{X}^{Gauss},z,w)$ by Theorem 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 2.19 (Anisotropic local law when $N \le M$). *Fix $\tau \le ||z|^2 - 1| \le \tau^{-1}$. Suppose Assumption 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 $D = diag(d_1,d_2,\ldots,d_N)$. 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 (2.23) 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 $D = \mathrm{diag}(d_1,d_2,\ldots,d_M)$. 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 [20], to prove the local circular law in Theorem 2.6 when $N > M$, it suffices to study $\det(TX - z)$ (see (2.52) below), for which we have

$$\det(TX - z) = \det\begin{pmatrix} DVXU_1 - z & DVXU_2 \\ 0 & -z \end{pmatrix} = \det(V^T D^T U_1^T X^T - z)(-z)^{N-M}. \tag{2.51}$$

Comparing with (2.49), 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 2.20. *The $N < M$ case of Theorem 2.6 implies the $N > M$ case of Theorem 2.6.*

2.4 Proof of Theorem 2.6

By Claim 2.20, it suffices to assume $N \le M$. Our main tool will be Theorem 2.19. A major part of the proof follows from [7, Section 5]. The following lemma collects basic properties of stochastic domination $\prec$, which will be used tacitly during the proof and throughout this paper.

Lemma 2.21 (Lemma 3.2 in [4]). *(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 $\mathbb{E}\xi(u)^2 \le N^C$ 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 [20] can be reformulated as the following (see e.g. [22]): for any smooth function g,

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} g(\mu_j) &= \frac{1}{4\pi N}\int \Delta g(z) \sum_{j=1}^{N} \log(\mu_j - z)(\bar{\mu}_j - \bar{z})\, dA(z) \\
&= \frac{1}{4\pi N}\int \Delta g(z) \log\left|\det(Y(z)Y^{\dagger}(z))\right| dA(z) \\
&= \frac{1}{4\pi N}\int \Delta g(z) \sum_{j=1}^{N} \log \lambda_j(z)\, dA(z),
\end{aligned} \tag{2.52}$$

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

$$\frac{1}{N}\sum_{i=1}^{N} F_{z_0,a}(\mu_j) = \frac{N^{-1+2a}}{4\pi}\int (\Delta F)(\xi) \sum_{j=1}^{N} \log \lambda_j(z)\, dA(\xi). \tag{2.53}$$

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

$$\gamma_j(z) := \begin{cases} \sup_x \left\{\int_0^x \rho_{2c}(x)dx \le \frac{j}{N}\right\}, & \text{if } 1 \le j \le N-1 \\ e_1, & \text{if } j = N \end{cases}. \tag{2.54}$$

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 2.14, we have that for any $\epsilon > 0$,

$$\begin{aligned} &\left|\sum_{j=1}^N \log \lambda_j(z) - N\int_0^\infty (\log x)\rho_{2c}(x;z)dx\right| \\ &\le \sum_{j=1}^N N\int_{\gamma_{j-1}(z)}^{\gamma_j(z)} |\log \lambda_j(z) - \log x|\rho_{2c}(x;z)dx \le N^{\epsilon} \end{aligned} \tag{2.55}$$

for large enough N. Suppose we have the bound

$$\left|\sum_j \log \lambda_j - \sum_j \log \gamma_j\right| \prec N^b. \tag{2.56}$$

Plugging (2.55) and (2.56) into (2.53), we get

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

Thus we obtain (2.21) if we can prove (2.56) for $b = 1/2$, and we obtain (2.24) if we can prove (2.56) for $b = 0$ when $1+\tau \le |z_0|^2 \le 1+\tau^{-1}$ or when the assumption (2.23) holds.

We need the following lemma which is a consequence of Theorem 2.19. Recall (2.16) and (2.20), the number L of the connected components is of order 1 and the number of γ_j's in each component $[e_{2k}, e_{2k-1}]$ is of order N. We define the classical number of eigenvalues to the left of e_k, $1 \le k \le 2L$, as

$$N_k := \left\lceil N\int_0^{e_k}\rho_{2c}(x)\right\rceil. \tag{2.57}$$

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

Lemma 2.22 (Singular value rigidity)**.** *Fix a small $\epsilon > 0$.*

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

$$\frac{|\lambda_j - \gamma_j|}{\gamma_j} \prec \left[\min\left(\frac{j-N_{2k}}{N}, \frac{N_{2k-1}-j}{N}\right)\right]^{-1/3} N^{-1/2}. \tag{2.58}$$

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

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

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

$$\#\{j : 0 < \lambda_j < c\} \prec 1. \tag{2.60}$$

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

$$\frac{|\lambda_j - \gamma_j|}{\gamma_j} \prec \left[\min\left(\frac{j-N_{2k}}{N}, \frac{N_{2k-1}-j}{N}\right)\right]^{-1/3} N^{-1}. \tag{2.61}$$

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

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

Proof. The proof is similar to the proof of [7, Lemma 5.1]. See also [4, Theorem 2.10] or [14, Theorem 7.6] □

Using (2.58) and (2.59), we get that

$$\sum_{N_{2k}+N^{1/2+\epsilon} \le j \le N_{2k-1} - N^{1/2+\epsilon}} |\log \lambda_j - \log \gamma_j| \prec \sum_{N_{2k}+N^{1/2+\epsilon} \le j \le N_{2k-1} - N^{1/2+\epsilon}} \frac{|\lambda_j - \gamma_j|}{\gamma_j} \prec N^{1/2}. \tag{2.63}$$

By Theorem 2.10 of [4], there exists a constant $C > 0$ such that

$$\|X^* X\| \le C \text{ with high probability}. \tag{2.64}$$

Thus we have

$$\lambda_j \le \|Y\|^2 \le (\|T\|\|X\| + |z|)^2 \prec 1, \quad 1 \le j \le N. \tag{2.65}$$

Together with Lemma 2.23 concerning the smallest singular value of $TX - z$, we get

$$\sum_{k=1}^{2L} \sum_{|j - e_k| < N^{1/2+\epsilon}} |\log \lambda_j| \prec N^{1/2+\epsilon}. \tag{2.66}$$

Since $|\log \gamma_j| \prec 1$ by Proposition 2.14, we conclude

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

Combining (2.63) and (2.67), we get for any $\epsilon > 0$,

$$\sum_{1 \le j \le N} |\log \lambda_j - \log \gamma_j| \le N^{1/2+\epsilon} \tag{2.68}$$

for large enough N. This implies (2.56) for $b = 1/2$. If in addition the assumption (2.23) holds, the averaged local law holds with parameters $(T, X, z, \mathbf{D}(\epsilon, N))$ for arbitrarily small ϵ by Theorem 2.19. Then we can prove (2.56) for $b = 0$ using the better bounds (2.61) and (2.62).

Finally we show that when $|z_0|^2 \ge 1 + \epsilon$, with the bounds (2.58) we can still prove the estimate (2.56) for $b = 0$. By the averaged local law and the definition of γ_j in (2.54), we have

$$\sum_{j=1}^{N} \frac{1}{\lambda_j - i\eta} - \sum_{j=1}^{N} \frac{1}{\gamma_j - i\eta} \prec \frac{1}{\eta}, \tag{2.69}$$

uniformly in $N^{-1/2+\epsilon} \le \eta \le N^{1/2}$. Taking integral of (2.69) over η from $N^{-1/2+\epsilon}$ to $N^{1/2}$, we get

$$\sum_{j=1}^{N} \log \frac{\lambda_j - iN^{-1/2+\epsilon}}{\gamma_j - iN^{-1/2+\epsilon}} - \sum_{j=1}^{N} \log \frac{\lambda_j - iN^{1/2}}{\gamma_j - iN^{1/2}} \prec 1. \tag{2.70}$$

Then we use (2.58) and the bound (2.65) to estimate that

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

Thus we conclude

$$\sum_{j=1}^{N} \log \frac{_j \quad iN^{-1/2+}}{_j \quad iN^{-1/2+}} \quad N . \tag{2.71}$$

Using $_j \quad 1$, (2.60) and (2.73), we get

$$\sum_{j=1}^{N} \log \frac{_j \quad iN^{-1/2+}}{_j \quad iN^{-1/2+}} \quad \sum_{j=1}^{N} \log \frac{_j}{_j} \quad 1 + \sum_{_j \quad c} \log \frac{_j \quad iN^{-1/2+}}{_j \quad iN^{-1/2+}} \quad \sum_{_j \quad c} \log \frac{_j}{_j}$$
$$1 + \sum_{_j \quad c} (_j \quad _j) N^{1/2+} \quad N^2 . \tag{2.72}$$

Combing (2.71) and (2.72), we conclude (2.56) for $b = 0$.

If the entries of X are identically distributed, then instead of Lemma 2.23, we shall use the results in [39] to get a lower bound for the smallest singular value of $TX \quad z$ (see Remark 3 below Theorem 2.6). In particular, the bounded density condition for the entries of X is not needed anymore. This concludes the last statement of Theorem 2.6.

Lemma 2.23 (Lower bound on the smallest singular value)**.** *If $N \quad M$ and the entries of X have a density bounded by N^{C_3} for some $C_3 > 0$, then*

$$|\log \quad _1(z)| \quad 1 \tag{2.73}$$

holds uniformly for z in any fixed compact set.

Proof. We already have an upper bound for $_1$; see (2.65). Hence to get (2.73), we still need to prove that

$$\mathbb{P}\left(_1(z) \quad e^{\quad N} \right) \quad N^{\quad C} \tag{2.74}$$

for any $\, ; C > 0$. By (2.49), we have that

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

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

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

we can easily get

$$\tilde{\ }_1(z)^{1/2} \quad N^{\quad 1/2} \min_{1 \quad k \quad N} \operatorname{dist}\left(\tilde{Y}_k, \operatorname{span}\{\tilde{Y}_l, l \neq k\} \right) = N^{\quad 1/2} \min_{1 \quad k \quad N} \left| \langle \tilde{Y}_k, u_k \rangle \right| , \tag{2.75}$$

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

$$\mathbb{P}(\tilde{\ }_1(z) \quad N^{\quad C_0}) \quad CN^{\quad C_0/2 + C_3 + 3/2}, \tag{2.76}$$

which is a much stronger result than (2.74). Here we have used Theorem 1.2 of [34] to conclude that $\langle \tilde{Y}_k, u_k \rangle$ for fixed u_k has density bounded by CN^{C_3}. □

2.5 Outline of the paper

The rest of this paper is devoted to the proof of Theorem 2.18 and Theorem 2.19. In Section 3, we collect the basic tools that we shall use in the proof. In Section 4, we perform step (A) of the proof by proving the entrywise local law and averaged local law in Theorem 2.18 under the assumption that T is diagonal. We first prove a weak version of the entrywise local law in Sections 4.1-4.3, and then improve the weak law to the strong entrywise local law and averaged local law in Sections 4.4-4.5. In Section 5, we perform step (B) of the proof by proving the anisotropic local law in Theorem 2.18 using the entrywise local law proved in Section 4. Finally in Section 6 we finish the step (C) of the proof, where using Theorem 2.18, we prove Theorem 2.19 with a self-consistent comparison method.

The first part of Appendix A establishes the basic properties of $\rho_{1,2c}$ stated in Lemma 2.3 and Proposition 2.14. In Sections A.2 and A.3, we prove some key estimates about $m_{1,2c}$ and the stability of the self-consistent equation (2.11) on regular domains. In Appendix B, we prove a fluctuation averaging lemma that will be used in the proof of the strong entrywise local law.

3 Basic tools

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

Definition 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 $G^{(J)} := \{(H^{(J)})^{-1}_{st} : s,t \in \mathcal{I} \setminus J\}$. 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 $G^{[J]} := \{(H^{[J]})^{-1}_{st} : s,t \in \mathcal{I} \setminus [J]\}$. We will 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 3.2. *(Resolvent identities).*

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

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

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

$$G_{ij} = wG_{ii}G^{(i)}_{jj}\left(YG^{(ij)}Y^\dagger\right)_{ij}, \quad G_{\mu\nu} = wG_{\mu\mu}G^{(\mu)}_{\nu\nu}\left(Y^\dagger G^{(\mu\nu)}Y\right)_{\mu\nu}. \tag{3.2}$$

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

$$G_{i\mu} = G_{ii}G^{(i)}_{\mu\mu}\left(-w^{1/2}Y_{i\mu} + w\left(YG^{(i\mu)}Y\right)_{i\mu}\right), \tag{3.3}$$

$$G_{\mu i} = G_{\mu\mu}G^{(\mu)}_{ii}\left(-w^{1/2}Y^\dagger_{\mu i} + w\left(Y^\dagger G^{(\mu i)}Y^\dagger\right)_{\mu i}\right). \tag{3.4}$$

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

$$G^{(r)}_{st} = G_{st} - \frac{G_{sr}G_{rt}}{G_{rr}}, \quad \frac{1}{G_{ss}} = \frac{1}{G^{(r)}_{ss}} - \frac{G_{sr}G_{rs}}{G_{ss}G^{(r)}_{ss}G_{rr}}. \tag{3.5}$$

(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 [14, 15, 17]. ☐

Lemma 3.3. *(Resolvent identities for $G_{[ij]}$ groups).*

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

$$G^{-1}_{[ii]} = H_{[ii]} - \sum_{k,l \ne i} H_{[ik]}G^{[i]}_{[kl]}H_{[li]}. \tag{3.6}$$

For $i \neq j \in I_1$, we have

$$G_{[ij]} = -G_{[ii]} \sum_{k \neq i} H_{[ik]} G^{[i]}_{[kj]} = -\sum_{k \neq j} G^{[j]}_{[ik]} H_{[kj]} G_{[jj]} \tag{3.7}$$

$$= -G_{[ii]} H_{[ij]} G^{[i]}_{[jj]} + G_{[ii]} \sum_{k,l \notin \{i,j\}} H_{[ik]} G^{[ij]}_{[kl]} H_{[lj]} G^{[i]}_{[jj]}. \tag{3.8}$$

(ii) For $k \in I_1$ and $i, j \in I_1 \setminus \{k\}$,

$$G^{[k]}_{[ij]} = G_{[ij]} - G_{[ik]} G^{-1}_{[kk]} G_{[kj]}, \tag{3.9}$$

and

$$G^{-1}_{[ii]} = \left(G^{[k]}_{[ii]}\right)^{-1} - G^{-1}_{[ii]} G_{[ik]} G^{-1}_{[kk]} G_{[ki]} \left(G^{[k]}_{[ii]}\right)^{-1}. \tag{3.10}$$

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

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

$$Y = \sum_{k=1}^{N} \sqrt{\sigma_k}\, \xi_k \zeta_k^\dagger$$

be the singular decomposition of Y, where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_N \geq 0$ and $\{\xi_k\}_{k=1}^N$ and $\{\zeta_k\}_{k=1}^N$ are orthonormal bases of $\mathbb{C}^{I_1}$ and $\mathbb{C}^{I_2}$ respectively. Then by (2.30), we have

$$G(w) = \sum_{k=1}^{N} \frac{1}{\sigma_k - w} \begin{pmatrix} \xi_k \xi_k^\dagger & w^{-1/2} \sqrt{\sigma_k}\, \xi_k \zeta_k^\dagger \\ w^{-1/2} \sqrt{\sigma_k}\, \zeta_k \xi_k^\dagger & \zeta_k \zeta_k^\dagger \end{pmatrix}. \tag{3.11}$$

Definition 3.4 (Generalized entries)**.** *For $\mathbf{v}, \mathbf{w} \in \mathbb{C}^I$, $s \in I$ and an $I \times I$ matrix A, we shall denote*

$$A_{\mathbf{vw}} := \langle \mathbf{v}, A\mathbf{w} \rangle, \quad A_{\mathbf{v}s} := \langle \mathbf{v}, A\mathbf{e}_s \rangle, \quad A_{s\mathbf{w}} := \langle \mathbf{e}_s, A\mathbf{w} \rangle, \tag{3.12}$$

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

Given vectors $\mathbf{v} \in \mathbb{C}^{I_1}$ and $\mathbf{w} \in \mathbb{C}^{I_2}$, we always identify them with their natural embeddings $\begin{pmatrix} \mathbf{v} \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ \mathbf{w} \end{pmatrix}$ in $\mathbb{C}^I$. The exact meanings will be clear from the context.

Lemma 3.5. *Fix $\tau > 0$. The following estimates hold uniformly for any $w \in \mathbf{D}(\tau, N) \cup \mathbf{D}_L(\tau)$. We have*

$$\|G\| \leq C\eta^{-1}, \quad \|\partial_w G\| \leq C\eta^{-2}. \tag{3.13}$$

Let $\mathbf{v} \in \mathbb{C}^{I_1}$ and $\mathbf{w} \in \mathbb{C}^{I_2}$, we have the bounds

$$\sum_{\mu \in I_2} |G_{\mathbf{w}\mu}|^2 = \sum_{\mu \in I_2} |G_{\mu\mathbf{w}}|^2 = \frac{\operatorname{Im} G_{\mathbf{ww}}}{\eta}, \tag{3.14}$$

$$\sum_{i \in I_1} |G_{\mathbf{v}i}|^2 = \sum_{i \in I_1} |G_{i\mathbf{v}}|^2 = \frac{\operatorname{Im} G_{\mathbf{vv}}}{\eta}, \tag{3.15}$$

$$\sum_{i \in I_1} |G_{\mathbf{w}i}|^2 = \sum_{i \in I_1} |G_{i\mathbf{w}}|^2 = |w|^{-1} G_{\mathbf{ww}} + \bar{w}|w|^{-1} \frac{\operatorname{Im} G_{\mathbf{ww}}}{\eta}, \tag{3.16}$$

$$\sum_{\mu \in I_2} |G_{\mathbf{v}\mu}|^2 = \sum_{\mu \in I_2} |G_{\mu\mathbf{v}}|^2 = |w|^{-1} G_{\mathbf{vv}} + \bar{w}|w|^{-1} \frac{\operatorname{Im} G_{\mathbf{vv}}}{\eta}. \tag{3.17}$$

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

Proof. The estimates in (3.13) follow from (3.11). For any unit vectors $\mathbf{x},\mathbf{y}\in\mathbb{C}^{\mathcal{I}_1}$, we have

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

For any unit vectors $\mathbf{x}\in\mathbb{C}^{\mathcal{I}_1}$ and $\mathbf{y}\in\mathbb{C}^{\mathcal{I}_2}$, we have

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

where we have used that for $w = E + i\eta$, $|w|^{-1/2}\sqrt{\lambda_k}/|\lambda_k - w| \le \eta^{-1}$. For the other two blocks of G, we can prove similar estimates. This gives the first bound in (3.13). 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 (3.14), we observe that

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

and by (2.30),

$$\sum_{\mu\in\mathcal{I}_2}|G_{\mathbf{w}\mu}|^2 = \sum_{\mu\in\mathcal{I}_2}\langle \mathbf{w}, G_R e_\mu\rangle\langle e_\mu, G_R^{\dagger}\mathbf{w}\rangle = \left\langle \mathbf{w}, G_R G_R^{\dagger}\mathbf{w}\right\rangle = \sum_{k=1}^{N}\frac{|\langle \mathbf{w},\xi_k\rangle|^2}{(\lambda_k - E)^2+\eta^2}. \tag{3.18}$$

Similarly, we can prove the identity for $\sum_{\mu\in\mathcal{I}_2}|G_{\mu\mathbf{w}}|^2$ and (3.15). For (3.16), 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 (3.11). Then we use (2.30) and (3.18) to get that

$$\begin{aligned}\sum_{i\in\mathcal{I}_1}|G_{\mathbf{w}i}|^2 &= |w|^{-1}\left(G_R Y^{\dagger}Y G_R^{\dagger}\right)_{\mathbf{ww}}\\ &= |w|^{-1}\left[G_R\left(Y^{\dagger}Y - \bar{w}\right)G_R^{\dagger}\right]_{\mathbf{ww}} + \bar{w}|w|^{-1}\left(G_R G_R^{\dagger}\right)_{\mathbf{ww}}\\ &= |w|^{-1}G_{\mathbf{ww}} + \bar{w}|w|^{-1}\left(G_R G_R^{\dagger}\right)_{\mathbf{ww}} = |w|^{-1}G_{\mathbf{ww}} + \bar{w}|w|^{-1}\frac{\operatorname{Im} G_{\mathbf{ww}}}{\eta}.\end{aligned} \tag{3.19}$$

Identity (3.17) 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 [13] for the proof. See also Theorem C.1 of [14].

Lemma 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 (2.2) and (2.3). Then we have the following bounds:*

$$\left|\sum_i b_iX_i\right| \prec \frac{\left(\sum_i |b_i|^2\right)^{1/2}}{\sqrt{N}}, \quad \left|\sum_{i,j} a_{ij}X_iY_j\right| \prec \left(\frac{\sum_{i,j}|a_{ij}|^2}{N}\right)^{1/2}, \quad \left|\sum_{i\neq j} a_{ij}X_iX_j\right| \prec \left(\frac{\sum_{i\neq j}|a_{ij}|^2}{N}\right)^{1/2}, \tag{3.20}$$

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 2.3 and Proposition 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 A.2. For $w = E + i\eta$, we define the distance to the spectral edge through

$$\kappa(E) := \min_{1 \le k \le 2L, e_k > 0} |E - e_k|. \tag{3.21}$$

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

Lemma 3.7. *Fix $\tau > 0$ and suppose $||z|^2 - 1| \ge \tau^{-1}$. 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 2.4. Then for $w \in \mathbf{D}_k^b(\zeta, \tau', N)$, we have*

$$|1 + m_{1c}| \sim \operatorname{Im} m_{1c} \sim 1, \ |m_{2c}| \sim \operatorname{Im} m_{2c} \sim 1. \tag{3.22}$$

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

$$\operatorname{Im} m_{1,2c} \sim \eta, \ |1 + m_{1c}| \sim 1, \ |m_{2c}| \sim 1. \tag{3.23}$$

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

$$\operatorname{Im} m_{1,2c} \sim \begin{cases} \sqrt{\kappa + \eta} & \text{if } E \in \operatorname{supp} \rho_{1,2c} \\ \eta/\sqrt{\kappa + \eta} & \text{if } E \notin \operatorname{supp} \rho_{1,2c} \end{cases}, \ |1 + m_{1c}| \sim 1, \ |m_{2c}| \sim 1. \tag{3.24}$$

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

$$|1 + m_{1c}| \sim \operatorname{Im} m_{1c} \sim 1, \ |m_{2c}| \sim \operatorname{Im} m_{2c} \sim 1; \tag{3.25}$$

if $|w| \le 2\tau'$, we have

$$m_{1c} = i\frac{\sqrt{t}}{\sqrt{w}} + O(1), \ m_{2c} = \frac{i\sqrt{t}}{\sqrt{w}(t + |z|^2)} + O(1), \tag{3.26}$$

for some constant $t > 0$, and

$$\operatorname{Im} m_{1,2c} \sim |w|^{-1/2}. \tag{3.27}$$

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

$$|m_{1c}| \sim \operatorname{Im} m_{1c} \sim \frac{1}{\eta}, \ |m_{2c}| \sim \operatorname{Im} m_{2c} \sim \frac{1}{\eta}. \tag{3.28}$$

In Cases 1-4, we have

$$|w(1 + s_i m_{2c})(1 + m_{1c}) - |z|^2| \ge c, \tag{3.29}$$

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

$$|w(1 + s_i m_{2c})(1 + m_{1c}) - |z|^2| \ge \eta, \tag{3.30}$$

Note that the uniform bounds (3.29) and (3.30) guarantee that the matrix entries of $\Pi(w)$ remain bounded. We have the following Lemma, which will be proved in Appendix A.2.

Lemma 3.8. *In Cases 1-4 of Lemma 3.7, we have*

$$\| {}_{[i]c}\| \quad C|w|^{1/2}; \qquad {}_{[i]c}{}^{1} \quad C|w|^{1/2}; \tag{3.31}$$

and in Case 5 of Lemma 3.7, we have

$$\| {}_{[i]c}\| \quad C \quad ^{1}; \qquad {}_{[i]c}{}^{1} \quad C \ : \tag{3.32}$$

For all the cases in Lemma 3.7,

$$\operatorname{Im}\Pi_{\mathbf{vv}} \quad C\operatorname{Im}(m_{1c}+m_{2c}); \tag{3.33}$$

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

The self-consistent equation (2.11) can be written as

$$\Upsilon(w; m_1) = 0; \tag{3.34}$$

where

$$\Upsilon(w; m_1) = m_1 + \frac{1}{N}\sum_{i=1} l_i s_i (1+m_1)\left[w \quad 1 + s_i \frac{1+m_1}{w(1+m_1)^2+|z|^2} \right) (1+m_1) \quad |z|^2 \Bigg]^{1} : \tag{3.35}$$

The stability of (3.34) roughly says that if $\Upsilon(w; m_1)$ is small and $m_1(w') \quad m_{1c}(w')$ is small for $w' := w + iN^{10}$, then $m_1(w) \quad m_{1c}(w)$ is small. For an arbitrary $w \in \mathbf{D}$, we define the discrete set

$$L(w) := \{w\} \cup \{w' \in \mathbf{D} : \operatorname{Re} w' = \operatorname{Re} w; \operatorname{Im} w' \in [\operatorname{Im} w; 1] \cap (N^{10}\mathbb{N})\}; \tag{3.36}$$

Thus, if $\operatorname{Im} w \quad 1$ 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 $|L(w)| \quad N^{10}$.

Definition 3.9 (Stability of (3.34))**.** *We say that (3.34) is stable on $\mathbf{D}$ if the following holds. Suppose that $N^{2}|m_{1c}| \quad (w) \quad (\log N)^{1}|m_{1c}|$ for $w \in \mathbf{D}$ and that is Lipschitz continuous with Lipschitz constant N^{4}. Suppose moreover that for each fixed E, the function $\mapsto (E + i\ )$ is non-increasing for > 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*

$$|\Upsilon(w; u_1)| \quad (w): \tag{3.37}$$

Then

$$|u_1(w) \quad m_{1c}(w)| \quad \frac{C}{\sqrt{\ + \ }}; \tag{3.38}$$

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

We say that (3.34) is stable on $\mathbf{D}_L$ if for $0 \quad (w) \quad (\log N)^{1}|m_{1c}|$, (3.37) implies

$$|u_1(w) \quad m_{1c}(w)| \quad C\ ; \tag{3.39}$$

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

This stability condition has previously appeared in [4, 7, 24]. In [24], 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 A.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 4.5.

Lemma 3.10. *Fix > 0 and let $' > 0$ be sufficiently small depending on . Let $||z|^2 \quad 1| \quad ^{1}$.*

Case 1 *Suppose the bulk component* $[e_{2k}; e_{2k\ \ 1}]$ *is regular in the sense of Definition 2.4. Then (3.34) is stable on* $\mathsf{D}_k^b(\ ;\ {}^0; N)$ *in the sense of Definition 3.9.*

Case 2 *(3.34) is stable on* $\mathsf{D}^o(\ ;\ {}^0; N)$ *in the sense of Definition 3.9.*

Case 3 *Suppose* $e_k \neq 0$ *is a regular edge in the sense of Definition 2.4. Then (3.34) is stable on* $\mathsf{D}_k^e(\ ;\ {}^0; N)$ *in the sense of Definition 3.9.*

Case 4 *Suppose* $|z|^2 \quad 1 \qquad$ *so that* $e_{2L} = 0$*. If* $|w|^{1/2} + |z|^2 \qquad$ *for some constant* $\quad > 0$*, then (3.34) is stable on* $\mathsf{D}_{2L}^e(\ ;\ {}^0; N)$ *in the sense of Definition 3.9.*

Case 5 *(3.34) is stable on* $\mathsf{D}_L(\ )$ *in the sense of Definition 3.9.*

4 Entrywise local law when T is diagonal

In this section we prove the entrywise local law and averaged local law in Theorem 2.18 when T is diagonal. The proof is similar to the previous proofs of the entrywise local law in e.g. [4, 5, 7, 24]. We basically follow the idea in [7], 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 (2.9) and (2.10) 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

$$\mathsf{E}_{[i]}[\ ] := \mathsf{E}[\ | H^{[i]}];$$

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 (3.6) to get that

$$\begin{aligned} G_{[ii]}^{\ 1} \quad & \mathsf{E}_{[i]} H_{[ii]} \quad \sum_{k,l \neq i} \mathsf{E}_{[i]} \quad H_{[ik]} G_{[kl]}^{[i]} H_{[li]} \\ &= \begin{pmatrix} w & w^{1/2} z \\ w^{1/2} \bar{z} & w \end{pmatrix} \quad \frac{w}{N} \sum_k \begin{pmatrix} |d_i|^2 G_{kk}^{[i]} & 0 \\ 0 & |d_k|^2 G_{\bar{k}\bar{k}}^{[i]} \end{pmatrix} \\ &= \begin{pmatrix} w & w^{1/2} z \\ w^{1/2} \bar{z} & w \end{pmatrix} \quad w \begin{pmatrix} |d_i|^2 m_2 & 0 \\ 0 & m_1 \end{pmatrix}; \end{aligned} \tag{4.1}$$

where we used the definitions of m_1 and m_2 in (2.34). The 11 entry of (4.1) gives the equation

$$G_{ii} \qquad \frac{1 \quad m_1}{w\,(1 + |d_i|^2 m_2)\,(1 + m_1) \quad |z|^2}; \tag{4.2}$$

from which we get that

$$G_{ii} \quad w \ \ 1 + |d_i|^2 m_2 \ \ + \frac{|z|^2}{1 + m_1} \qquad 1.$$

Summing over i and using that $N^{\ 1} \sum_i G_{ii} = N^{\ 1} \sum \ \ G \quad = m_2$, the above equation becomes

$$w\,(m_2 + m_1 m_2) + \frac{|z|^2 m_2}{1 + m_1} \qquad 1;$$

which gives (2.9). Multiplying (4.2) with $|d_i|^2$ and summing over i, we get the self-consistent equation (2.10). 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 D. On the domain 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 3.7 and Lemma 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 2.15. Also due to the remark above Lemma 3.10, in Sections 4.1-4.4, 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 4.5.

4.1 The self-consistent equations

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

Proposition 4.1 (Weak entrywise law)**.** *Fix $|z|^2 \le 1 - \tau$ and a small constant $c > 0$. Suppose Assumption 2.1 holds, $N = M$ and $T \equiv D := diag(d_1, \ldots, d_N)$. Then for any regular domain $\mathbf{S} \subseteq \mathbf{D}$,*

$$\max_{i,j \in \mathcal{I}_1} \left\| (G(w) - \Pi(w))_{[ij]} \right\| \prec \frac{1}{|w|^{1/2}} \left(\frac{|w|^{1/2}}{N\eta} \right)^{1/4} \tag{4.3}$$

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

$$\max_{i,j \in \mathcal{I}_1} \left\| (G(w) - \Pi(w))_{[ij]} \right\| \prec \frac{1}{\eta} \sqrt{\frac{1}{N}}. \tag{4.4}$$

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

Definition 4.2 (Control parameters)**.** *Suppose $N = M$ and $T \equiv D := diag(d_1, \ldots, d_N)$. We define*

$$\Lambda := \max_{i,j \in \mathcal{I}_1} \left\| (G - \Pi)_{[ij]} \right\|, \quad \Lambda_o := \max_{i \ne j \in \mathcal{I}_1} \left\| (G - \Pi)_{[ij]} \right\|. \tag{4.5}$$

For $J \subseteq \mathcal{I}$, define the averaged variables $m_{1,2}^{(J)}$ ($m_{1,2}^{[J]}$) by replacing G in (2.34) with $G^{(J)}$ ($G^{[J]}$), i.e.

$$m_1^{(J)} := \frac{1}{N} \sum_{i \notin J} |d_i|^2 G_{ii}^{(J)}, \quad m_2^{(J)} := \frac{1}{N} \sum_{\mu \notin J} G_{\mu\mu}^{(J)}. \tag{4.6}$$

The averaged error and the random control parameter are defined as

$$\theta := |m_1 - m_{1c}| + |m_2 - m_{2c}| \quad \text{and} \quad \Psi_\theta := \sqrt{\frac{\mathrm{Im}\,(m_{1c} + m_{2c}) + \theta}{N\eta}} + \frac{1}{N\eta}, \tag{4.7}$$

respectively.

Remark: By (2.4), we immediately get that

$$\mathrm{Im}\, m_1^{(J)} \lesssim \mathrm{Im}\, m_2^{(J)} \lesssim \tau^{-1} \mathrm{Im}\, m_2^{(J)}, \tag{4.8}$$

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

We introduce the Z variables:

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

By the identity (3.6) we have

$$G_{[ii]}^{-1} = \mathbb{E}_{[i]} G_{[ii]}^{-1} + Z_{[i]} = \begin{pmatrix} -w - w|d_i|^2 m_2^{[i]} & -w^{1/2} z \\ -w^{1/2} \bar{z} & -w - w m_1^{[i]} \end{pmatrix} + Z_{[i]}, \tag{4.9}$$

where

$$Z_{[i]} = w \begin{pmatrix} |d_i|^2 m_2^{[i]} - |d_i|^2 \left(X G^{[i]} X^\dagger \right)_{ii} & -w^{-1/2} d_i X_{ii} - \left(D X G^{[i]} D X \right)_{ii} \\ -w^{-1/2} \bar{d}_i X_{ii}^\dagger - \left(X^\dagger D^\dagger G^{[i]} X^\dagger D^\dagger \right)_{ii} & m_1^{[i]} - \left(X^\dagger D^\dagger G^{[i]} D X \right)_{ii} \end{pmatrix}. \tag{4.10}$$

Lemma 4.3. *For $J \subset I_1$, the following crude bound on the difference between m_a and $m_a^{[J]}$ ($a=1,2$) holds:*

$$\left|m_a - m_a^{[J]}\right| \le \frac{C|J|}{N\eta}, \quad a = 1,2, \tag{4.11}$$

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

Proof. For $i \in I_1$, we have

$$|m_1 - m_1^{(i)}| = \frac{1}{N}\left|\sum_{k\in I_1} |d_k|^2 \frac{G_{ki}G_{ik}}{G_{ii}}\right| \le \frac{1}{N|G_{ii}|}\sum_{k\in I_1}|G_{ik}|^2 = \frac{\eta^{-1}}{N}\frac{\operatorname{Im} G_{ii}}{|G_{ii}|} \le \frac{\eta^{-1}}{N} \tag{4.12}$$

where in the first step we used (3.5), and in the second and third steps the equality (3.15). Similarly, using (3.5) and (3.16) we get

$$|m_1^{(i)} - m_1^{(ii)}| = \frac{1}{N}\left|\sum_{k\in I_1} |d_k|^2 \frac{G^{(i)}_{ki}G^{(i)}_{ik}}{G^{(i)}_{ii}}\right| \le \frac{1}{N|G^{(i)}_{ii}|}\left(\frac{G^{(i)}_{ii}}{|w|} + \frac{\bar w}{|w|}\frac{\operatorname{Im} G^{(i)}_{ii}}{\eta}\right) \le \frac{2\eta^{-1}}{N}.$$

By induction on the indices in $[J]$, we can prove (4.11) for m_1. The proof for m_2 is similar. □

Lemma 4.4. *Suppose $|z|^2 \le 1 - \tau$. For $i \in I_1$, we have*

$$|(Z_{[i]})_{11}| \prec |w|\sqrt{\frac{\operatorname{Im} m_2^{[i]}}{N\eta}}; \quad |(Z_{[i]})_{22}| \prec |w|\sqrt{\frac{\operatorname{Im} m_1^{[i]}}{N\eta}}; \tag{4.13}$$

$$|(Z_{[i]})_{st}| \prec |w|^{1/2}\left(\frac{|w|^{1/2}}{\sqrt{N}} + \sqrt{\frac{|m_1^{[i]}|}{N|w|}} + \sqrt{\frac{\operatorname{Im} m_1^{[i]}}{N\eta}}\right) \quad \text{for } s \ne t \in \{1,2\}; \tag{4.14}$$

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

$$Z_{[i]} \prec |w|\Psi, \tag{4.15}$$

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

$$Z_{[i]} \prec |w|(N\eta)^{-1/2}, \tag{4.16}$$

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

Proof. Applying the large deviation Lemma 3.6 to $Z_{[i]}$ in (4.10), we get that

$$\left|\frac{(Z_{[i]})_{11}}{w}\right| \prec \frac{1}{N}\left[\left(\sum_{k} \left|G^{[i]}_{kk}\right|^2\right)^{1/2} + \left(\sum_{k\ne l}\left|G^{[i]}_{kl}\right|^2\right)^{1/2}\right] \le \frac{C}{N}\left(\sum_{k,l}\left|G^{[i]}_{kl}\right|^2\right)^{1/2}$$
$$= \frac{C}{N}\left(\sum_{k}\frac{\operatorname{Im} G^{[i]}_{kk}}{\eta}\right)^{1/2} = C\sqrt{\frac{\operatorname{Im} m_2^{[i]}}{N\eta}}.$$

where in the third step we used the equality (3.14). Similarly we can prove the bound for $(Z_{[i]})_{22}$ using Lemma 3.6 and (3.15). Now we consider $(Z_{[i]})_{12}$. First, we have $X_{ii} \prec N^{-1/2}$ by (2.3). For the other part, we use Lemma 3.6 and (3.17) to get that

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

$$\left[\frac{|m_1^{[i]}|}{N|w|}+\frac{\operatorname{Im} m_1^{[i]}}{N}\right]^{1/2} \quad C\left(\sqrt{\frac{|m_1^{[i]}|}{N|w|}}+\sqrt{\frac{\operatorname{Im} m_1^{[i]}}{N}}\right). \tag{4.17}$$

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

Now we prove (4.15). By the definitions (4.7) and using (4.11), we get that

$$Z_{[i]}\ {}_{11} \quad |w| \sqrt{\frac{\operatorname{Im} m_2^{[i]}}{N}} = |w| \sqrt{\frac{\operatorname{Im} m_{2c}+\operatorname{Im}\left(m_2^{[i]} \quad m_2\right)+\operatorname{Im}(m_2 \quad m_{2c})}{N}} \quad C|w|\Psi. \tag{4.18}$$

We can estimate $Z_{[i]}\ {}_{22}$ and the third term in (4.14) in a similar way. For the Cases 1-4 in Lemma 3.7, we have $|m_{1c}| \quad 1$ for $|w| \quad 1$, $\operatorname{Im} m_{1c} \quad |w|^{1/2} \quad |m_{1c}|$ for $|w| \to 0$, and $C\operatorname{Im} m_{1c}$. Thus

$$\sqrt{\frac{|m_{1c}|}{N|w|}} \quad \frac{C}{\sqrt{N}} \quad C\Psi \ \text{ for } |w| \quad 1; \ \text{ and } \ \sqrt{\frac{|m_{1c}|}{N|w|}} \quad C\sqrt{\frac{\operatorname{Im} m_{1c}}{N}} \quad C\Psi \ \text{ for } |w| \to 0.$$

Then for the second term in (4.14), we have that

$$\sqrt{\frac{|m_1^{[i]}|}{N|w|}} \quad C\left(\frac{1}{N}+\sqrt{\frac{}{N}}+\sqrt{\frac{|m_{1c}|}{N|w|}}\right) \quad C\Psi.$$

This concludes (4.15). Finally, the estimate (4.16) follows directly from (4.13), (4.14) and (3.13). □

Lemma 4.5. *Suppose $|z|^2 \quad 1$. Define the w-dependent event $\Xi(w) := \{ |w|^{1/2}(\log N)^{1}\}$. Then we have that for $w \in \mathbf{D}$,*

$$\mathbf{1}(\Xi)m_2=\mathbf{1}(\Xi)\left[\frac{1+m_1}{w(1+m_1)^2+|z|^2}+O(\Psi)\right]; \quad \mathbf{1}(\Xi)\Upsilon(w;m_1) \quad \mathbf{1}(\Xi)\Psi; \tag{4.19}$$

where Υ is defined in (3.35). For $w \in \mathbf{D}_L$, we have

$$m_2=\frac{1+m_1}{w(1+m_1)^2+|z|^2}+O\left(\ ^{1}(N\ )^{1/2}\right); \quad \Upsilon(w;m_1) \quad {}^{1}(N\ )^{1/2}. \tag{4.20}$$

Proof. First, suppose that $w \in \mathbf{D}$. Using (4.9), we get

$$G_{[ii]}^{-1} = {}_{[i]}^{-1} + {}_{[i]}, \tag{4.21}$$

where ${}_{[i]}$ is defined in (2.36) and

$${}_{[i]} = w \begin{pmatrix} |d_i|^2 \ m_2 \quad m_2^{[i]} & 0 \\ 0 & m_1 \quad m_1^{[i]} \end{pmatrix} + Z_{[i]}.$$

By (4.11) and (4.15), we have that ${}_{[i]} \quad |w|\Psi$. Let $B_i = {}_{[i]}^{-1} \quad {}_{[i]c}^{-1}$, where ${}_{[i]c}$ is defined in (2.32). By (3.31) and the definition of Ξ, we have $\mathbf{1}(\Xi)\|B_i \ {}_{[i]c}\| \quad C(\log N)^{1}.$ Thus we have the expansion

$$\mathbf{1}(\Xi)\ {}_{[i]} = \mathbf{1}(\Xi)\left({}_{[i]c}^{-1}+B_i\right)^{-1} = \mathbf{1}(\Xi)\ {}_{[i]c}\left(1 \quad B_i\ {}_{[i]c}+(B_i\ {}_{[i]c})^2+\cdots\right) = \mathbf{1}(\Xi)\left({}_{[i]c}+ \ {}_{a}\right); \tag{4.22}$$

where ${}_a$ can be estimated as $\mathbf{1}(\Xi)\|{}_a\| \quad \mathbf{1}(\Xi)C|w|^{1/2}(\log N)^{1}$. This shows that $\mathbf{1}(\Xi)\|{}_{[i]}\| = \mathbf{1}(\Xi)O(|w|^{1/2})$, and so $\mathbf{1}(\Xi)\;{}_{[i]}\;{}_{[i]} \quad \mathbf{1}(\Xi)|w|^{1/2}\Psi \quad \mathbf{1}(\Xi)\,CN^{=2}$ by the definition of D in (2.39). Again we do the expansion for (4.21):

$$\mathbf{1}(\Xi)G_{[ii]} = \mathbf{1}(\Xi)\;{}_{[i]}^{1} + {}_{[i]}\;{}^{1} = \mathbf{1}(\Xi)\;{}_{[i]}\left(1 + \sum_{l=1}\;{}_{[i]}\;{}_{[i]}\;{}^{l}\right) = \mathbf{1}(\Xi)\;{}_{[i]} + {}_{b}\;; \tag{4.23}$$

where $\mathbf{1}(\Xi)\|{}_b\| \quad \mathbf{1}(\Xi)\Psi$. Now the 11 entry of (4.23) gives that

$$\mathbf{1}(\Xi)G_{ii} = \mathbf{1}(\Xi)\frac{1 \quad m_1}{w\,(1+|d_i|^2 m_2)\,(1+m_1) \quad |z|^2} + \mathbf{1}(\Xi)O\;(\Psi\;)\,; \tag{4.24}$$

from which we get that

$$\mathbf{1}(\Xi)G_{ii}\left[w\;1+|d_i|^2 m_2 + \frac{|z|^2}{1+m_1}\right] = \mathbf{1}(\Xi)\left[1+O\;|w|^{1/2}\Psi\right]. \tag{4.25}$$

Here we used that

$$\mathbf{1}(\Xi)\quad w\;1+|d_i|^2 m_2 + \frac{|z|^2}{1+m_1} = O(|w|^{1/2});$$

which follows from Lemma 3.7 and the definition of Ξ. Summing (4.25) over i, we get

$$\mathbf{1}(\Xi)\quad w\,(m_2+m_1m_2) + \frac{|z|^2 m_2}{1+m_1} = \mathbf{1}(\Xi)\left[1+O\;|w|^{1/2}\Psi\right];$$

which gives

$$\mathbf{1}(\Xi)m_2 = \mathbf{1}(\Xi)\frac{1+m_1}{w\,(1+m_1)^2+|z|^2} + \mathbf{1}(\Xi)O\;(\Psi\;). \tag{4.26}$$

Now plugging (4.26) into (4.24), multiplying with $|d_i|^2$ and summing over i, we obtain that

$$\mathbf{1}(\Xi)m_1 = \mathbf{1}(\Xi)\left[\frac{1}{N}\sum_{i=1} l_i s_i \frac{1 \quad m_1}{w\;1+s_i\frac{1+m_1}{w(1+m_1)^2+|z|^2}\;(1+m_1) \quad |z|^2} + O\;(\Psi\;)\right]; \tag{4.27}$$

where we used (3.29) and $\mathbf{1}(\Xi)(1+m_1) = \mathbf{1}(\Xi)O(|w|^{1/2})$. This concludes the proof.

Similarly, when $w \in \mathrm{D}_L$, it is easy to prove (4.20) using the estimates (4.16) and (3.13). Note that $|m_{1;2}| = O(\;{}^{1})$ by (3.13), which implies immediately the bounds $\|{}_{[i]}\| = O(\;{}^{1})$ and $\|{}_{[i]}\;{}^{1}\| = O(\;)$. Hence without introducing the event Ξ, we can obtain directly

$$G_{[ii]} = {}_{[i]} + O\;(\;{}^{1}(N\;)^{1/2}). \tag{4.28}$$

The rest of the proof is essentially the same. □

Notice that applying Lemma 3.10 to (4.20), we obtain that $|m_{1;2} \quad m_{1;2c}| \quad {}^{1}(N\;)^{1/2}$. Plugging it into (4.28), we immediately get (4.4) for $w \in \mathrm{D}_L$. This proves the entrywise law on D_L, since ${}^{1}N^{1/2} \quad C\Psi$ by the definition (2.45) and the estimate (3.28).

4.2 The large case

It remains to prove Proposition 4.1 on domain D. We would like to fix E and then apply a continuity argument in by first showing that the rough bound $\Lambda \quad |w|^{1/2}(\log N)^{1}$ in Lemma 4.5 holds for large . To start the argument, we first need to establish the estimates on G when $\quad 1$. The next lemma is a trivial consequence of (3.13).

Lemma 4.6. *For any $w \in \mathbf{D}$ and $\ \ge c$ for fixed $c > 0$, we have the bound*

$$\max_{s,t} |G_{st}(w)| \le C \tag{4.29}$$

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

Lemma 4.7. *Fix $c > 0$ and $|z|^2 \le 1 \ $. We have the following estimate*

$$\max_{w \in \mathbf{D};\ \ge c} \Lambda(w) \le N^{-1/2}. \tag{4.30}$$

Proof. By the previous lemma, we have $|m^{[i]}_{1,2}| = O(1)$. So by Lemma 4.4, $\|Z_{[i]}\| \le N^{-1/2}$ uniformly in $\ \ge c$. Then as in (4.21), we have

$$G_{[ii]} = \left({}_{[i]}^{-1} + {}_{[i]} \right)^{-1}; \tag{4.31}$$

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

$$ {}_i = \left(G_{[ii]}^{-1} - {}_{[i]} \right)^{-1} = G_{[ii]} \left(1 - {}_{[i]} G_{[ii]} \right)^{-1} = O(1).$$

Then we can expand (4.31) to get that

$$G_{[ii]} = {}_i + O\left(N^{-1/2} \right). \tag{4.32}$$

The 11 and 22 entries of (4.32) lead to the equations

$$m_1 = \frac{1}{N} \sum_{i=1}^{N} |d_i|^2 \left[-w\left(1 + |d_i|^2 m_2\right) + \frac{|z|^2}{1 + m_1} \right]^{-1} + O\left(N^{-1/2} \right); \tag{4.33}$$

$$m_2 = \frac{1}{N} \sum_{i=1}^{N} \left[-w(1 + m_1) + \frac{|z|^2}{1 + |d_i|^2 m_2} \right]^{-1} + O\left(N^{-1/2} \right). \tag{4.34}$$

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

Using the spectral decomposition (3.11), we note that for $l > 1$,

$$\frac{1}{N} \sum_{|\ _k - E| \ge l} \frac{|E - \ _k|}{(\ _k - E)^2 + \ ^2} \le \frac{1}{l};$$

$$\frac{1}{N} \sum_{|\ _k - E| \le l} \frac{|E - \ _k|}{(\ _k - E)^2 + \ ^2} \le \frac{1}{N} \sum_{|\ _k - E| \le l} \frac{l}{(\ _k - E)^2 + \ ^2} \le l \operatorname{Im} m_2.$$

Summing up these two inequalities and optimizing l, we get

$$|\operatorname{Re} m_2| \le 2 \sqrt{\frac{\operatorname{Im} m_2}{\ }}. \tag{4.35}$$

Assume that $\operatorname{Im} m_2 \le C(\log N)^{-1}$, then by (4.8) we also have $\operatorname{Im} m_1 \le C^{-1}(\log N)^{-1}$. From (4.35), we get $|m_2| \le C(\log N)^{-1/2}$. Together with the estimate $m_1 = O(1)$, we get

$$\left| w(1 + m_1) + \frac{|z|^2}{1 + |d_i|^2 m_2} \right| \le C \text{ with high probability.} \tag{4.36}$$

On the other hand

$$\operatorname{Im} \left[-w(1 + m_1) + \frac{|z|^2}{1 + |d_i|^2 m_2} \right] \le -\operatorname{Im} w = -\ ; \tag{4.37}$$

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

$$\operatorname{Im}(wm_1) = \operatorname{Im}\left[\frac{1}{N}\sum_{k=1}^{N} |d_i|^2 |\xi_k(i)|^2 \left(1+\frac{\lambda_k}{\lambda_k - w}\right)\right] \ge 0.$$

With (4.36) and (4.37), we get from (4.34) that $\operatorname{Im} m_2 \ge c'$ with high probability for some $c' > 0$. This contradicts $\operatorname{Im} m_2 \le C(\log N)^{-1}$. Thus we must have $\operatorname{Im} m_2 \ge C(\log N)^{-1}$ with high probability, which also implies $\operatorname{Im} m_1 \ge C(\log N)^{-1}$ by (4.8).

Now we can proceed as in the proof of Lemma 4.5 and get that

$$m_2 = \frac{1+m_1}{w(1+m_1)^2 + |z|^2} + O\left(N^{-\epsilon/2}\right), \quad \Upsilon(w;m_1) \le N^{\epsilon/2}. \tag{4.38}$$

We omit the details. Applying Lemma 3.10 to (4.38), we conclude $|m_{1,2} - m_{1,2c}| \le N^{-\epsilon/2}$ uniformly in $\mathcal{D}_c$. By (4.32), we get $\|(G-\Pi)_{[ii]}\| \le N^{-\epsilon/2}$ uniformly in $\mathcal{D}_c$ and $i \in \mathcal{I}_1$. Finally using (3.8), Lemma 3.5 and Lemma 3.6, we can prove the off-diagonal estimate; see (4.51) below. ☐

4.3 Proof of the weak entrywise local law

In this subsection, we finish the proof of Proposition 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 4.5 is based on the condition $\eta \ge |w|^{1/2}(\log N)^{-1}$. So far this is established only for large η in (4.30). 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

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

where $L(w)$ is defined in (3.36). Fix a regular domain $\mathbf{S}$, $\epsilon < \delta/4$ 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.

$$\mathbb{P}\left(v(w) \le N^{\epsilon}, v(w) > N^{3\epsilon/4}\right) \le N^{-D+21} \tag{4.39}$$

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

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

$$\eta(w') \ge C|w'|^{1/2}(\log N)^{-1} \tag{4.40}$$

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 (4.19), for all $w' \in \mathbf{S} \cap L(w)$, there exists an $N_0 \equiv N_0(\epsilon, D)$ such that

$$P\left(v(w) \le N^{\epsilon}, \Upsilon(w') > \frac{N^{\epsilon}}{|w'|^{1/2}}\sqrt{\frac{|w'|^{1/2}}{N \operatorname{Im} w'}}\right) \le N^{-D}, \tag{4.41}$$

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

$$P\left(v(w) \le N^{\epsilon}, \max_{w' \in L(w)} \Upsilon(w')\sqrt{\frac{N \operatorname{Im} w'}{|w'|^{1/2}}} > N^{\epsilon}\right) \le N^{-D+10}. \tag{4.42}$$

Now consider the event

$$\Xi_1 := \left\{v(w) \le N^{\epsilon}, \max_{w' \in L(w)} \Upsilon(w')\sqrt{\frac{N \operatorname{Im} w'}{|w'|^{1/2}}} \le N^{\epsilon}\right\}. \tag{4.43}$$

We have $1(\Xi_1)\Upsilon(w') \le \gamma(w')$ for all $w' \in L(w)$ with $\gamma(w') := \frac{N^{\epsilon}}{|w'|^{1/2}}\sqrt{\frac{|w'|^{1/2}}{N\operatorname{Im} w'}}$. We now apply Lemma 3.10. If $\gamma \le 1$ (recall (3.21)), then $|w| \le 1$ and we have

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

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

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

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

$$1(\Xi_1)|m_1(w') - m_{1c}(w')| \le C\frac{N^{\epsilon/2}}{|w'|^{1/2}}\left(\frac{|w'|^{1/2}}{N\operatorname{Im} w'}\right)^{1/4} \tag{4.44}$$

for all $w' \in L(w)$. By (4.19), we have

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

for all $w' \in \mathbf{S} \cap L(w)$. Together with (4.44), this shows that there exists an $N_1 \equiv N_1(\epsilon; D)$ such that

$$\mathbb{P}\Bigg(v(w) \ge N^{-\epsilon};\ \max_{w' \in L(w)} \Upsilon(w')\sqrt{\frac{N\operatorname{Im} w'}{|w'|^{1/2}}} \le N^{\epsilon};$$
$$\max_{w' \in L(w)} \gamma(w')|w'|^{1/2}\left(\frac{N\operatorname{Im} w'}{|w'|^{1/2}}\right)^{1/4} > N^{3\epsilon/4}\Bigg) \le N^{-D} \tag{4.45}$$

for $N \ge \max\{N_0, N_1\}$. Adding (4.42) and (4.45), we get

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

Taking the union bound over $L(w)$ we get (4.39) for all $N \ge \max\{N_0, N_1\}$.

Now we conclude the proof of Proposition 4.1 by combining (4.39) with the large estimate (4.30). 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

$$\mathbb{P}\left(\exists w \in \Delta : v(w) \in (N^{-3\epsilon/4}, N^{-\epsilon}]\right) \le N^{-D+41}. \tag{4.46}$$

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

$$\mathbb{P}\left(\exists w \in \mathbf{S} : v(w) \in (2N^{-3\epsilon/4}, N^{-\epsilon}/2]\right) \le N^{-D+41}. \tag{4.47}$$

Combining with (4.30), 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

$$\Lambda(w) \prec \frac{1}{|w|^{1/2}}\left(\frac{|w|^{1/2}}{N}\right)^{1/4}. \tag{4.48}$$

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

Now using (4.23) and (4.48), we get

$$\left\| G_{[ii]} - \Pi_{[i]c} \right\| \le \left\| G_{[ii]} - \Pi_{[i]} \right\| + \left\| \Pi_{[i]} - \Pi_{[i]c} \right\| \prec \Psi + \frac{1}{|w|^{1/2}} \left(\frac{|w|^{1/2}}{N} \right)^{1/4}. \quad (4.49)$$

To conclude Proposition 4.1, it remains to prove the estimate for the off-diagonal $G_{[ij]}$ groups. Using (4.11), it is not hard to get that

$$\left\| G^{[J]}_{[ii]} - \Pi_{[i]c} \right\| \prec \frac{1}{|w|^{1/2}} \left(\frac{|w|^{1/2}}{N} \right)^{1/4} \quad (4.50)$$

for any $|J| \le l$ with $l \in \mathbb{N}$ fixed. Thus we have $\| G^{[J]}_{[ii]} \| = O\left(|w|^{-1/2}\right)$ 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 (3.8) and the above diagonal estimates, we get that

$$\left\| G_{[ij]} \right\| \prec |w|^{-1} \frac{|w|^{1/2}}{\sqrt{N}} + |w|^{-1} \left\| \sum_{k,l \notin \{i,j\}} H_{[ik]} G^{[ij]}_{[kl]} H_{[lj]} \right\| \prec \Psi + \frac{1}{|w|^{1/2}} \left(\frac{|w|^{1/2}}{N} \right)^{1/4}, \quad (4.51)$$

where we used Lemma 3.5 and Lemma 3.6 to obtain that

$$|w|^{-1} \sum_{k,l \notin \{i,j\}} H_{[ik]} G^{[ij]}_{[kl]} H_{[lj]} = \begin{pmatrix} \sum_{k,l \notin \{i,j\}} X_{ik} G^{[ij]}_{kl} X^{\dagger}_{lj} & \sum_{k,l \notin \{i,j\}} X_{ik} G^{[ij]}_{kl} X_{lj} \\ \sum_{k,l \notin \{i,j\}} X^{\dagger}_{ik} G^{[ij]}_{kl} X^{\dagger}_{lj} & \sum_{k,l \notin \{i,j\}} X^{\dagger}_{ik} G^{[ij]}_{kl} X_{lj} \end{pmatrix} \prec \Psi. \quad (4.52)$$

Its proof is very similar to the proof of Lemma 4.4, so we omit the details.

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 2.18 on domain $\mathbf{D}$ and under the condition $|w|^{1/2} + |z|^2 \ge c$. In Lemma 4.5, we have proved an error estimate of the self-consistent equations of $m_{1,2}$ linearly in Ψ. The core part of the proof is to improve this estimate to quadratic in Ψ. 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 4.5.

Lemma 4.8. *Fix $|z|^2 \le 1 - \tau$. Then for $w \in \mathbf{D}$,*

$$m_2 = \frac{1 + m_1}{w(1 + m_1)^2 + |z|^2} + O_{\prec}\left(|w|^{1/2} \Psi^2 + \|[Z]\| + \|\langle Z \rangle\| \right), \quad (4.53)$$

and

$$\Upsilon(w, m_1) \prec |w|^{1/2} \Psi^2 + \|[Z]\| + \|\langle Z \rangle\|. \quad (4.54)$$

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

$$m_2 = \frac{1 + m_1}{w(1 + m_1)^2 + |z|^2} + O_{\prec}\left((N\eta)^{-1} + \|[Z]\| + \|\langle Z \rangle\| \right), \quad (4.55)$$

and

$$\Upsilon(w, m_1) \prec (N\eta)^{-1} + \|[Z]\| + \|\langle Z \rangle\|. \quad (4.56)$$

Proof. The proof is almost the same as the one for Lemma 4.5, we only lay out the difference. We first consider the case $w \in \mathbf{D}$. By Proposition 4.1, the event Ξ holds with high probability. Hence without loss of generality, we may assume Ξ holds throughout the following proof. Using (3.9), we get

$$\frac{1}{N}\sum_{k\in I_1}\begin{pmatrix}|d_k|^2 & 0\\ 0 & 1\end{pmatrix}\left(G_{[kk]} - G^{[i]}_{[kk]}\right) = \begin{pmatrix}|d_i|^2 & 0\\ 0 & 1\end{pmatrix}\frac{G_{[ii]}}{N} + \frac{1}{N}\sum_{k\neq i}\begin{pmatrix}|d_k|^2 & 0\\ 0 & 1\end{pmatrix} G_{[ki]}G_{[ii]}^{-1}G_{[ik]}. \tag{4.57}$$

By Proposition 4.1, (3.31) and (4.51), we have

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

By Lemma 3.7, it is easy to verify that $\|G_{[ii]}/N\| \le C|w|^{1/2}\Psi^2.$ Plugging it into (4.57), we get

$$\left\|m^{[i]}_{1,2} - m_{1,2}\right\| \prec |w|^{1/2}\Psi^2. \tag{4.58}$$

By (4.15) and (4.58), the error ϵ_b in (4.23) is

$$\epsilon_b = O_\prec(|w|^{1/2}\Psi^2) - \mathcal{G}_{[i]}Z_{[i]}\mathcal{G}_{[i]}\left[1 + O_\prec(|w|^{1/2}\Psi)\right] = O_\prec(|w|^{1/2}\Psi^2) - \mathcal{G}_{[i]}Z_{[i]}\mathcal{G}_{[i]}.$$

Then following the arguments in Lemma 4.5, we can prove the desired result. For $w \in \mathbf{D}_L$, the proof is similar by using (4.4). □

In the following lemma, we shall give 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 B.

Lemma 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 $w \in \mathbf{D}$,*

$$\|[Z]\| + \|\langle Z\rangle\| \prec |w|^{-1/2}\Phi_o^2. \tag{4.59}$$

Now we finish the proof of the entrywise local law and averaged local law on the domain $\mathbf{D}$. By Proposition 4.1, we can take

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

in Lemma 4.9, with $\Lambda_o \prec \Psi \le |w|^{-1/2}\Phi_o$ and $\Lambda \prec \Psi + \theta \le |w|^{-1/2}\Phi$. Then (4.54) gives

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

Using the stability Lemma 3.10, we get

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

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

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

if $\sqrt{\kappa+\eta} \le (\log N)^{-1}$, 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 (4.53) to get that

$$|m_1 - m_{1c}| + \frac{|w|^{1/2}\mathrm{Im}(m_{1c}+m_{2c}) + |w|^{1/4}(N\eta)^{-1/4}}{N\eta} \prec |w|^{-1/2}\left(\frac{|w|^{1/2}}{N\eta}\right)^{1/2+1/8}. \tag{4.60}$$

Repeating the previous steps with the new estimate (4.60), we get the bound

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

after l iterations. This implies the averaged local law $\prec (N\eta)^{-1}$ since l can be arbitrarily large. Finally as in (4.49) and (4.51), we have for $i \ne j \in \mathcal{I}_1$,

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

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

When $w \in \mathbf{D}_L$, we have proved the entrywise law (see the remark after (4.28)). Also we can prove a similar estimate as in Lemma 4.9, which implies

$$m_2 = -\frac{1+m_1}{w(1+m_1)^2 + |z|^2} + O_\prec\left((N\eta)^{-1}\right), \quad \Upsilon(w, m_1) \prec (N\eta)^{-1}. \tag{4.61}$$

The averaged local law then follows from Lemma 3.10. We leave the details to the reader.

4.5 Proof of Theorem 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 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 (4.24) with $|d_i|^2$ and summing over i, we get

$$1(\Xi)m_1 = 1(\Xi)\left[\frac{1}{N}\sum_{i=1}^{N} l_i s_i \frac{1-m_1}{-w(1+s_i m_2)(1+m_1) - |z|^2} + O_\prec(\Psi_\theta)\right]. \tag{4.62}$$

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

$$\tilde{H}(w) := \begin{pmatrix} -w\Sigma^{-1} & w^{1/2}(X - D^{-1}z) \\ w^{1/2}(X - D^{-1}z)^\dagger & -wI \end{pmatrix}, \tag{4.63}$$

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

$$m_1 = \frac{1}{N}\mathrm{Tr}\left[D^\dagger\left(YY^\dagger - w\right)^{-1}D\right] = \frac{1}{N}\mathrm{Tr}\,\tilde{G}_L, \tag{4.64}$$

$$m_2 = \frac{1}{N}\mathrm{Tr}\,\tilde G_R = \frac{1}{N}\mathrm{Tr}\left[(X - D^{-1}z)^\dagger \Sigma (X - D^{-1}z) - w\right]^{-1}$$
$$= \frac{1}{N}\mathrm{Tr}\left[(X - D^{-1}z)(X - D^{-1}z)^\dagger \Sigma - w\right]^{-1} = \frac{1}{N}\mathrm{Tr}\,\Sigma^{-1}\tilde G_L. \qquad (4.65)$$

We apply the arguments in the proof of Lemma 4.5 to $\tilde H$, and obtain that

$$\tilde G_{[ii]}^{-1} = \begin{pmatrix} -w|d_i|^{-2} - wm_2 & -w^{1/2} z d_i^{-1} \\ -w^{1/2}\bar z \bar d_i^{-1} & -w - wm_1 \end{pmatrix} + O_\prec(|w|\Psi), \qquad (4.66)$$

from which we get that

$$1(\Xi)\tilde G_{ii} = 1(\Xi)\frac{-1-m_1}{w(|d_i|^{-2}+m_2)(1+m_1) - |z|^2|d_i|^{-2}} + O_\prec(\Psi).$$

Plugging this into (4.65), we get

$$1(\Xi)m_2 = 1(\Xi)\left[\frac{1}{N}\sum_{i=1}^{n}\frac{l_i}{s_i}\frac{-1-m_1}{w(s_i^{-1}+m_2)(1+m_1) - |z|^2 s_i^{-1}}\right] + O_\prec(\Psi). \qquad (4.67)$$

We take the equations in (4.62) and (4.67) as our new self-consistent equations, namely,

$$1(\Xi)f_1(m_1,m_2) = 1(\Xi)O_\prec(\Psi), \quad 1(\Xi)f_2(m_1,m_2) = 1(\Xi)O_\prec(\Psi), \qquad (4.68)$$

where

$$f_1(m_1,m_2) := m_1 + \frac{1}{N}\sum_i l_i s_i \frac{1+m_1}{w\left(1+s_i m_2\right)\left(1+m_1\right) - |z|^2}, \qquad (4.69)$$

$$f_2(m_1,m_2) := m_2 + \frac{1}{N}\sum_i l_i \frac{1+m_1}{w(1+s_i m_2)(1+m_1) - |z|^2}. \qquad (4.70)$$

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

Lemma 4.10. *Suppose that $N^{-2}|w|^{-1/2} \le \eta(w) \le (\log N)^{-1}|w|^{-1/2}$ for $w \in \mathbf{D}$. Suppose $u_{1,2} : \mathbf{D} \to \mathbb{C}$ are Stieltjes transforms of positive integrable functions such that*

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

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

$$|u_1(w) - m_{1c}(w)| + |u_2(w) - m_{2c}(w)| \le C\delta, \qquad (4.71)$$

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 (3.26) and (A.35), 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 3.10 in Section A.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 $|w|^{1/2} + |z|^2 \le \epsilon$.

5 Anisotropic local law when T is diagonal

In this section we prove the anisotropic local law in Theorem 2.18 when T is diagonal. The basic idea of the proof follows from [4, section 5], and the core part of our proof is a novel way to perform the combinatorics. By the Definition 2.17 (ii) and Definition 2.5 (ii), it suffices to prove the following proposition for generalized entries of G.

Proposition 5.1. *Fix $|z|^2 \le 1-\tau$ and suppose that the assumptions of Theorem 2.18 hold. Then for any regular domain $\mathbf{S} \subset \mathbf{D}$,*

$$|\langle \mathbf{u}, (G(w)-\Pi(w))\mathbf{v}\rangle| \prec \Psi \tag{5.1}$$

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

It is equivalent to prove that

$$\left|\sum_{i,j\in I_1} u_{[i]}^{\dagger}\left(G_{[ij]}-\Pi_{[ij]}\right)v_{[j]}\right| \prec \Psi; \quad u_{[i]} := \begin{pmatrix} u_i \\ u_{\bar i}\end{pmatrix}; \quad v_{[j]} := \begin{pmatrix} v_j \\ v_{\bar j}\end{pmatrix}. \tag{5.2}$$

By the entrywise local law,

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

Thus to show (5.2), it suffices to prove

$$\left|\sum_{i\ne j} u_{[i]}^{\dagger}G_{[ij]}v_{[j]}\right| \prec \Psi. \tag{5.3}$$

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

Lemma 5.2. *Suppose the assumptions in Proposition 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 5.2 is based on the polynomialization method developed in [4, 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.

5.1 Rescaling and partition of indices

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

$$R^{(J)} := w^{1/2}G^{(J)}, \tag{5.4}$$

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

$$\Phi = |w|^{1/2}\,\Psi. \tag{5.5}$$

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

$$R^{(J)}_{[ii]} = O_\prec(1); \quad \left(R^{(J)}_{[ii]}\right)^{-1} = O_\prec(1); \quad R^{(J)}_{[ij]} = O_\prec(\Phi) \text{ for } i \neq j; \tag{5.6}$$

under the above scaling. Now to prove Lemma 5.2, it is equivalent to prove

$$\mathbb{E}\left|\sum_{i\neq j} u^\dagger_{[i]} R_{[ij]} v_{[j]}\right|^p \prec \Phi^p. \tag{5.7}$$

We expand the product in (5.7) as

$$\left|\sum_{i\neq j} u^\dagger_{[i]} R_{[ij]} v_{[j]}\right|^p = \sum_{i_k\neq j_k \in I_1} \prod_{k=1}^{p/2} u^\dagger_{[i_k]} R_{[i_k j_k]} v_{[j_k]} \prod_{k=p/2+1}^{p} \overline{u^\dagger_{[i_k]} R_{[i_k j_k]} v_{[j_k]}}.$$

Formally, we regard $\{i_1, \ldots, i_p, j_1, \ldots, j_p\}$ as the set of $2p$ (index) variables that take values in I_1. Let B_p be the collection of all partitions of $\{i_1, \ldots, i_p, j_1, \ldots, j_p\}$ such that i_k, j_k are not in the same block for all $k = 1, \ldots, p$. For $\Gamma \in B_p$, let $n(\Gamma)$ be the number of its blocks and define a set of I_1-valued variables as

$$L(\Gamma) := \{b_1, \ldots, b_{n(\Gamma)}\}. \tag{5.8}$$

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

$$\Gamma : \{i_1, \ldots, i_p, j_1, \ldots, j_p\} \to L(\Gamma), \tag{5.9}$$

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

$$\begin{aligned}&\left|\sum_{i\neq j} u^\dagger_{[i]} R_{[ij]} v_{[j]}\right|^p \\ &= \sum_{\Gamma\in B_p} \sum^{*}_{\substack{b_l \in I_1,\\ l=1,\ldots,n(\Gamma)}} \prod_{k=1}^{p/2} u^\dagger_{[\Gamma(i_k)]} R_{[\Gamma(i_k)\Gamma(j_k)]} v_{[\Gamma(j_k)]} \prod_{k=p/2+1}^{p} \overline{u^\dagger_{[\Gamma(i_k)]} R_{[\Gamma(i_k)\Gamma(j_k)]} v_{[\Gamma(j_k)]}},\end{aligned} \tag{5.10}$$

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

$$\Delta(\Gamma) := \prod_{k=1}^{p/2} u^\dagger_{[\Gamma(i_k)]} R_{[\Gamma(i_k)\Gamma(j_k)]} v_{[\Gamma(j_k)]} \prod_{k=p/2+1}^{p} \overline{u^\dagger_{[\Gamma(i_k)]} R_{[\Gamma(i_k)\Gamma(j_k)]} v_{[\Gamma(j_k)]}}. \tag{5.11}$$

Notations: For any $b_k \in L$, we can define a corresponding I_2-valued variable $\bar{b}_k$ in the obvious way, and we denote

$$[L] := \{b_1, \ldots, b_n, \overline{b_1}, \ldots, \overline{b_n}\}. \tag{5.12}$$

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

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 3.3.

Lemma 5.3 (Resolvent identities for $R_{[ij]}$ groups). *For $k \notin J$ and $i, j \in I_1 \setminus J \cup \{k\}$, we have*

$$R^{[J]}_{[ij]} = R^{[Jk]}_{[ij]} + R^{[J]}_{[ik]} \left(R^{[J]}_{[kk]}\right)^{-1} R^{[J]}_{[kj]}, \tag{5.13}$$

$$\left(R^{[J]}_{[ii]}\right)^{-1} = \left(R^{[Jk]}_{[ii]}\right)^{-1} - \left(R^{[J]}_{[ii]}\right)^{-1} R^{[J]}_{[ik]} \left(R^{[J]}_{[kk]}\right)^{-1} R^{[J]}_{[ki]} \left(R^{[Jk]}_{[ii]}\right)^{-1}, \tag{5.14}$$

$$\left(R^{[J]}_{[ii]}\right)^{-1} = -w^{-1/2} H^{[J]}_{[ii]} - w^{-1} \sum_{l,l' \notin J \cup \{i\}} H^{[J]}_{[il]} R^{[Ji]}_{[ll']} H^{[J]}_{[l'i]}. \tag{5.15}$$

Furthermore, for $i \neq j$ and L defined in (5.8), we have

$$R^{[L \setminus \{i,j\}]}_{[ij]} = R^{[L \setminus \{i,j\}]}_{[ii]} S_{[ij]} R^{[L \setminus \{j\}]}_{[jj]}, \quad \text{with } S_{[ij]} = -w^{-1/2} H_{[ij]} + w^{-1} \sum_{k,l \notin L} H_{[ik]} R^{[L]}_{[kl]} H_{[lj]}. \tag{5.16}$$

In this section, we expand the R variables in $\Delta(\Gamma)$ using the identities in Lemma 5.3. During the expansion, we need to distinguish carefully between an algebraic expression and its value as a random variable.

Definition 5.4 (Strings)**.** *Let $\mathcal{A}$ 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 s to be a formal expression consisting of the symbols from $\mathcal{A}$, and denote by $[\![\mathsf{s}]\!]$ the random variable represented by it. Let $\mathcal{M}$ be the collection of all possible strings. We denote an empty string by $\emptyset$.*

Given a string s, after an expansion of R's in it, we will get a different string s'. However, they represent the same random variable $[\![\mathsf{s}]\!] = [\![\mathsf{s}']\!]$. During the proof, we will identify more elements of $\mathcal{A}$ (see the symbols in (5.32)).

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 s to be *maximally expanded* if all the R symbols in s is maximally expanded.

Definition 5.5 (String operators)**.** *(i) Define an operator $\mathcal{U}^{(k)}_0$ for $\Omega \in \mathcal{M}$, 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, $\mathcal{U}^{(k)}_0(\Omega) = \Omega$ and we say that $\mathcal{U}^{(k)}_0$ is trivial for Ω.*

(ii) Define an operator $\mathcal{U}^{(k)}_1$ for $\Omega \in \mathcal{M}$, 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 $k \notin T \cup \{i\}$. If $R^{[J]}_{[ij]}$ is found, replace it with $R^{[J]}_{[ik]} \left(R^{[J]}_{[kk]}\right)^{-1} R^{[J]}_{[kj]}$; if $\left(R^{[J]}_{[ii]}\right)^{-1}$ is found, replace it with $-\left(R^{[J]}_{[ii]}\right)^{-1} R^{[J]}_{[ik]} \left(R^{[J]}_{[kk]}\right)^{-1} R^{[J]}_{[ki]} \left(R^{[Jk]}_{[ii]}\right)^{-1}$; if neither is found, $\mathcal{U}^{(k)}_1(\Omega) = \emptyset$ and we say that $\mathcal{U}^{(k)}_1$ is null for Ω.

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

According to Lemma 5.3, for any $\Omega \in \mathcal{M}$ we have

$$\left[\!\!\left[\left(\mathcal{U}^{(k)}_0 + \mathcal{U}^{(k)}_1\right)(\Omega) \right]\!\!\right] = [\![\Omega]\!], \quad [\![\mathcal{O}(\Omega)]\!] = [\![\Omega]\!]. \tag{5.17}$$

Definition 5.6. *Define the function $F_{\text{d-max}} : \mathcal{M} \to \mathbb{N}$ (where the subscript "d-max" stands for "distance to being maximally expanded") through*

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

where τ could be 1 or -1, and

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

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

By off-diagonal symbols, we mean the terms of the form A_{st} with $s \in \{t, \bar t\}$ or $A_{[ij]}$ with $i \ne j$, e.g. $R^{[J]}_{[ij]}$ and $S_{[ij]}$ with $i \ne j$. Later we will define other types of off-diagonal symbols (see (5.32)). Note that a R symbol is maximally expanded if and only if $\mathcal F_{d-max}(R) = 0$ and a string Ω is maximally expanded if and only if $\mathcal F_{d-max}(\Omega) = 0$. The next two lemmas are almost trivial by Definition 5.5.

Lemma 5.7. *Fix $k \in \mathcal L$. If $\rho_0^{(k)}(\Omega) = \Omega$ and $\rho_1^{(k)}(\Omega) = \emptyset$,*

$$\mathcal F_{d-max}\left(\rho_0^{(k)}(\Omega)\right) = \mathcal F_{d-max}(\Omega); \quad \mathcal F_{d-max}\left(\rho_1^{(k)}(\Omega)\right) = 0; \tag{5.18}$$

otherwise,

$$\mathcal F_{d-max}\left(\rho_0^{(k)}(\Omega)\right) = \mathcal F_{d-max}(\Omega) - 1; \quad \mathcal F_{d-max}\left(\rho_1^{(k)}(\Omega)\right) \le \mathcal F_{d-max}(\Omega) + 4n(\Gamma). \tag{5.19}$$

For τ, we have

$$\mathcal F_{d-max}(\tau(\Omega)) = \mathcal F_{d-max}(\Omega) + a; \tag{5.20}$$

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

Lemma 5.8. *Fix $k \in \mathcal L$. For any $\Omega \in \mathcal M$, we have*

$$\mathcal F_{off}\left(\rho_0^{(k)}(\Omega)\right) = \mathcal F_{off}(\Omega); \quad \mathcal F_{off}(\tau(\Omega)) = \mathcal F_{off}(\Omega); \tag{5.21}$$

and

$$\mathcal F_{off}(\Omega) + 1 \le \mathcal F_{off}\left(\rho_1^{(k)}(\Omega)\right) \le \mathcal F_{off}(\Omega) + 2 \quad \text{if} \quad \rho_1^{(k)}(\Omega) \ne \emptyset. \tag{5.22}$$

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 (5.11) (if we keep the complex conjugates, the proof is the same but with slightly heavier notations). Suppose the right hand side of (5.11) is represented by a string Ω. Given a binary word $\mathbf w = a_1 a_2 \ldots a_m$ with $a_i \in \{0, 1\}$, we define the operation

$$(\Omega)_{\mathbf w} := \rho_{a_m}^{(b_m)} \circ \cdots \circ \rho_{a_2}^{(b_2)} \circ \rho_{a_1}^{(b_1)}(\Omega) \tag{5.23}$$

where $b_{qn+r} := b_r$ (recall (5.8)) for any $1 \le r \le n$ and $q \in \mathbb N$. So a binary word $\mathbf w$ uniquely determines an operator composition. By (5.17), $[\![(\Omega)_{\mathbf w 0}]\!] + [\![(\Omega)_{\mathbf w 1}]\!] = [\![(\Omega)_{\mathbf w}]\!]$ and so we get

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

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

Lemma 5.9. *Given any $\mathbf w$ such that $|\mathbf w| = (n^2+1)(p + 6l_0)$ and $(\Omega)_{\mathbf w} \ne \emptyset$, then either $\mathcal F_{off}((\Omega)_{\mathbf w}) \ge l_0 := (8/\delta + 2)p$, or $(\Omega)_{\mathbf w}$ 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 ρ_0's, and $m_0^{(1)}$ to denote the number of 0's corresponding to the non-trivial ρ_0's. Assume

$F_o((\Omega)_{\mathbf{w}}) < l_0$ and $(\Omega)_{\mathbf{w}}$ is not maximally expanded. By (5.21) and (5.22), we have $m_1 \quad l_0 \quad p \quad l_0$. By (5.18)-(5.20), we have

$$F_{d\ max}((\Omega)_{\mathbf{w}}) \quad F_{d\ max}(\Omega) + l_0 + 4nm_1 \quad m_0^{(1)}.$$

Then with $F_{d\ max}(\Omega) = np$, 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 w that correspond to trivial $_0$'s. This indicates that $(\Omega)_{\mathbf{w}}$ is maximally expanded, which gives a contradiction. □

Lemma 5.10. *There exists constants $C_{p;l_0}; C_{p;} > 0$ such that*

$$\sum_{2 B_p} \sum_{\substack{b_l \in I_1; \\ l=1;\dots;n()}} \mathbb{E} \sum_{\substack{|\mathbf{w}| = (n^2+1)(p+6l_0); \\ F_{\mathrm{off}}((\Delta(\Gamma))_{\mathbf{w}}) \quad l_0}} q_{(\Omega())_{\mathbf{w}}} y \quad C_{p;l_0} N^{2p} \Phi^{l_0} \quad C_{p;} \Phi^p. \tag{5.24}$$

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

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

$$S_{[ii]} := \begin{pmatrix} 0 & d_i X_{ii} \\ \bar{d}_i X_{ii}^y & 0 \end{pmatrix} + w^{\ 1} \sum_{k;l \notin L} H_{[ik]} R_{[kl]}^{[L]} H_{[li]}, \tag{5.25}$$

such that

$$R_{[ii]}^{[L n fig]\ \ 1} = \begin{pmatrix} w^{1=2} & z \\ \bar{z} & w^{1=2} \end{pmatrix} \quad S_{[ii]}. \tag{5.26}$$

Notice all the R symbols in a maximally expanded string are diagonal. We taylor expand $R_{[ii]}^{[L n fig]}$ as

$$R_{[ii]}^{[L n fig]} = \left[w^{\ 1=2}\ {}_{[i]c}^{\ 1} + S_{[ii]} \quad B_i \right]^{\ 1} = \sum_{k=0}^{l_0 \ 1} \tilde{}_{ic} \quad S_{[ii]} \quad B_i \ \tilde{}_{ic}^{\ k} + O\ \Phi^{l_0}; \tag{5.27}$$

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

$$S_{[ii]} \quad B_i = w^{\ 1=2} Z_{[i]}^{[L n fig]} + w^{1=2} \begin{pmatrix} |d_i|^2 (m_{2c} \quad m_2^{[L]}) & 0 \\ 0 & m_{1c} \quad m_1^{[L]} \end{pmatrix}! \quad \Phi$$

by (4.15) and the averaged local law. Now for all maximally expanded $(\Omega)_{\mathbf{w}}$ with $|\mathbf{w}| = (n^2+1)(p+6l_0)$, denote by $⟦(\Omega)_{\mathbf{w}}⟧$ the expression after plugging in (5.26) and (5.27) without the tail terms. Similar to Lemma 5.10, we have

$$\sum_{2 B_p} \sum_{\substack{b_l \in I_1; \\ l=1;\dots;n()}} \mathbb{E} \sum_{\substack{|\mathbf{w}| = (n^2+1)(p+6l_0); \\ (\Delta)_{\mathbf{w}} \text{ maximally expanded}}} q_{(\Omega())_{\mathbf{w}}} y \quad q_{(\Omega())_{\mathbf{w}}} y \quad C_{p;} \Phi^p.$$

From the above bound and Lemmas 5.9, 5.10, we see that to prove (5.7), it suffices to show

$$\sum_{2 B_p} \sum_{\substack{b_l \in I_1; \\ l=1;\dots;n()}} \mathbb{E} \sum_{\substack{|\mathbf{w}| = (n^2+1)(p+6l_0); \\ (\Delta)_{\mathbf{w}} \text{ maximally expanded}}} q_{(\Omega())_{\mathbf{w}}} y \quad C_{p;} \Phi^p. \tag{5.28}$$

We write $\langle(\Omega\ )_{\mathbf{w}}\rangle$ as a sum of monomials in terms of $S_{[ij]}$:

$$\langle(\Omega\ )_{\mathbf{w}}\rangle = \sum_i M(\mathbf{w};\Delta(\Gamma);i), \tag{5.29}$$

where i is an index to label these monomials. Note that after plugging (5.29) into (5.28), the number of summands $M(\mathbf{w};\Delta(\Gamma);i)$ inside the expectation depends only on p and . Thus to show (5.28), it suffices to prove the following lemma.

Lemma 5.11. *Fix any $\Gamma \in B_p$ and binary word $\mathbf{w}$ with $|\mathbf{w}| = (n^2+1)(p+6l_0)$. Suppose $(\Omega\ )_{\mathbf{w}}$ is maximally expanded. Let $M(\mathbf{w};\Delta(\Gamma))$ be a monomial in $\sqrt{(\Omega_{(\ )})_{\mathbf{w}}^{\dagger}}$. Then we have*

$$\sum_{b_l \in I_1;\, l=1,\dots,n(\ )} |\mathbb{E} M(\mathbf{w};\Delta(\Gamma))| \le C_{p,}\ \Phi^p \tag{5.30}$$

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

For the rest of this section, we fix a $\Gamma \in B_p$ and a maximally expanded $\sqrt{(\Omega_{(\ )})_{\mathbf{w}}^{\dagger}}$ with $|\mathbf{w}| = (n^2+1)(p+6l_0)$. Then we fix a monomial $M(\mathbf{w};\Delta(\Gamma))$ in $\sqrt{(\Omega_{(\ )})_{\mathbf{w}}^{\dagger}}$. 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

$$F_o\ (\Omega_M) = F_o\ ((\Omega\ )_{\mathbf{w}}). \tag{5.31}$$

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

$$S_{[ij]} = S^X_{ij} + S^X_{ij} + S^R_{ij} + S^R_{ij} + S^R_{ij} + S^R_{ij}, \tag{5.32}$$

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

$$S^X_{ij} := d_i X_{ij}\begin{pmatrix} 0 & 1 \\ 0 & 0\end{pmatrix}, \quad S^X_{ij} := \bar d_i X^{\dagger}_{ij}\begin{pmatrix} 0 & 0 \\ 1 & 0\end{pmatrix}, \tag{5.33}$$

$$S^R_{ij} := \sum_{k,l \notin L} d_i d_l X_{ik} X_{lj}\begin{pmatrix} 0 & R^{[L]}_{kl} \\ 0 & 0\end{pmatrix}, \quad S^R_{ij} := \sum_{k,l \notin L} d_i \bar d_l X_{ik} X^{\dagger}_{lj}\begin{pmatrix} R^{[L]}_{kl} & 0 \\ 0 & 0\end{pmatrix}, \tag{5.34}$$

$$S^R_{ij} := \sum_{k,l \notin L} \bar d_i d_l X^{\dagger}_{ik} X_{lj}\begin{pmatrix} 0 & 0 \\ 0 & R^{[L]}_{kl}\end{pmatrix}, \quad S^R_{ij} := \sum_{k,l \notin L} \bar d_i \bar d_l X^{\dagger}_{ik} X^{\dagger}_{lj}\begin{pmatrix} 0 & 0 \\ R^{[L]}_{kl} & 0\end{pmatrix}. \tag{5.35}$$

We expand the $S_{[ij]}$'s in $M(\mathbf{w};\Delta(\Gamma))$ using (5.32), and write $M(\mathbf{w};\Delta(\Gamma))$ as a sum of monomials in terms of S^X_{st} and S^R_{st}:

$$M(\mathbf{w};\Delta(\Gamma)) = \sum_i Q(\mathbf{w};\Delta(\Gamma);i), \tag{5.36}$$

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

$$F_o\ (\Omega_Q) = F_o\ (\Omega_M) = F_o\ ((\Omega\ )_{\mathbf{w}}). \tag{5.37}$$

Since the number of summands in (5.36) is independent of N, to prove (5.30) it suffices to show

$$\sum_{b_l \in I_1;\, l=1,\dots,n(\ )} |\mathbb{E} Q(\mathbf{w};\Delta(\Gamma))| \le C_{p,}\ \Phi^p \tag{5.38}$$

for any monomial $Q(\mathbf{w};\Delta(\Gamma))$ in (5.36). 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 5.12. *Recall Γ defined in (5.9). Let h be the number of blocks of Γ whose size is 1, i.e.*

$$h := \sum_{l=1}^{n(\Gamma)} \mathbf{1}\left(\left|\Gamma^{-1}(b_l)\right| = 1 \right). \tag{5.39}$$

For $l = 1, \ldots, n$, 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 5.13. *Suppose for any $b_1, \ldots, b_n$ taking distinct values in $\mathcal{I}_1$,*

$$\left| \mathbb{E}\, Q(\mathrm{w}, \Delta(\Gamma)) \right| \le C N^{-h/2} \Phi^p \prod_{l=1}^{n} \left| u_{[b_l]} \right|^{I_l} \left| v_{[b_l]} \right|^{J_l} \tag{5.40}$$

holds for some constant C independent of N. Then the estimate (5.38) holds.

Proof. By Cauchy-Schwarz inequality,

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

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

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

Hence it suffices to prove (5.40). The key is to extract the $N^{-h/2}$ factor from $\mathbb{E}\, Q(\mathrm{w}, \Delta(\Gamma))$. For this purpose, we need to keep track of the indices in $\mathcal{L}$ during the expansion.

Definition 5.14. *Define a function $F_{\rm in} : \mathcal{L} \times \mathcal{M} \to \mathbb{N}$ with $F_{\rm in}(l, \Omega)$ 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 5.5 and the expansions we have done to obtain Ω_Q from $(\Omega\ )_{\mathbf{w}}$.

Lemma 5.15. *(1) For any string Ω, if ${}_0^{(k)}$ is not trivial for Ω, then*

$$F_{\rm in}\left(l, {}_0^{(k)}(\Omega)\right) = F_{\rm in}(l, \Omega); \quad F_{\rm in}\left(l, {}_1^{(k)}(\Omega)\right) = F_{\rm in}(l, \Omega) + a, \; a \in \{0, 2\}. \tag{5.41}$$

(2) For any string Ω,

$$F_{\rm in}\left(l, (\Omega)\right) = F_{\rm in}(l, \Omega). \tag{5.42}$$

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

$$F_{\rm in}(l, \Omega_Q) = F_{\rm in}(l, (\Omega\ )_{\mathbf{w}}). \tag{5.43}$$

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

$$V := \{ l \in \mathcal{L} \mid F_{\rm in}(l, \Omega\ ) = 1 \}, \tag{5.44}$$

and

$$V_0 := \{ l \in \mathcal{L} \mid F_{\rm in}(l, \Omega\ ) = 1 \text{ and } F_{\rm in}(l, \Omega_Q^X) = 0 \}, \tag{5.45}$$

$$V_1 := \{ l \in \mathcal{L} \mid F_{\rm in}(l, \Omega\ ) = 1 \text{ and } F_{\rm in}(l, \Omega_Q^X) \ge 2 \}. \tag{5.46}$$

Recall the observation above Definition 5.12, we have $V = V_0 \cup V_1$ and

$$h = |V| = |V_0| + |V_1|.$$

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.

5.4 Introduction of graphs and conclusion of the proof

We introduce graphs to conclude the proof of (5.40). We use a connected graph to represent the string Ω_Q, call it by G_{Q0}. The indices in $[L]$ are represented by black nodes in 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 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 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 5.15, we see that for any $l \in V_0$,

$$F_{\text{in}}(l;\Omega_Q) \equiv \deg(l) + \deg(\bar{l}) \equiv 1 \pmod 2). \tag{5.47}$$

Hence

$$|V_0| = \sum_{l \in V_0} [F_{\text{in}}(l;\Omega_Q) \mod 2] \le \sum_{l \in V_0} \left((\deg(l) \mod 2) + (\deg(\bar{l}) \mod 2) \right). \tag{5.48}$$

Now we expand the S^R edges. Take the S^R_{ij} edge as an example (recall (5.34)). We replace the S^R_{ij} 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_{ik} and X_{lj}, and an R-colored edge connecting $\bar{k}$ and l to represent $\begin{pmatrix} 0 & R^{[L]}_{kl} \\ 0 & 0 \end{pmatrix}$. 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 G_{Q0} into R-groups and call the resulting graph G_{Q1}. For example, after expanding the S^R edges in Fig. 3, we get the graph in Fig. 4. In the graph 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.

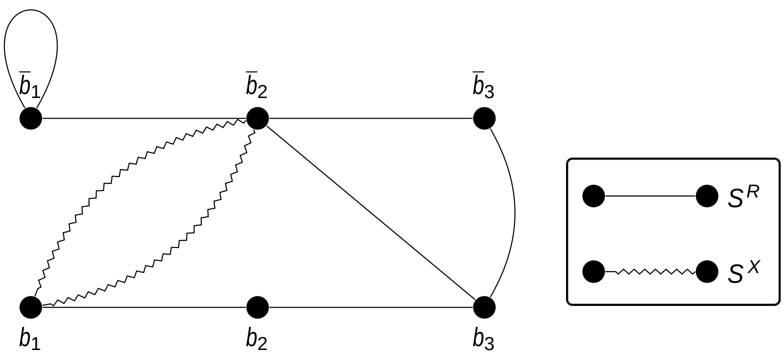


Figure 3: An example of the graph G_{Q0}.

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

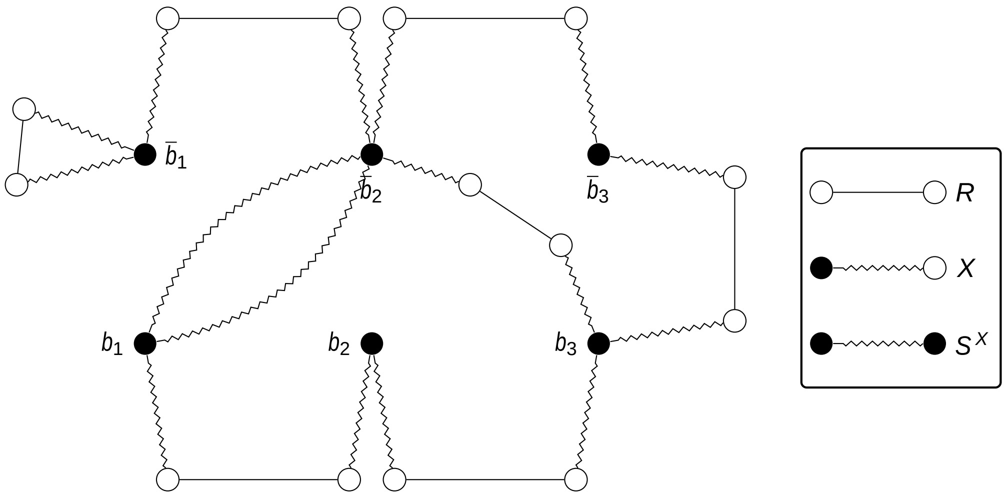
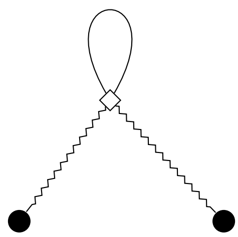


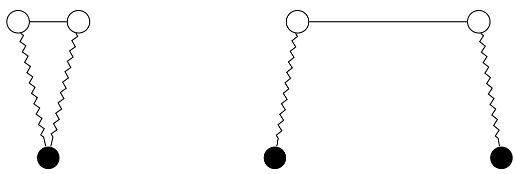
Figure 4: The resulting graph G_{Q1} after expanding each S^R in Fig. 3 into R-groups.

blocks when doing the summation. Let W be the set of all white nodes in G_{Q1}, and let $\mathcal{W}$ be the collection of all partitions of W. Fix a partition $\pi \in \mathcal{W}$ and denote its blocks by $W_1, \ldots, W_{m(\pi)}$. 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 $\ge$ the number of diagonal R-edges in G_{Q1}). Then we trivially have (recall Definition 5.16)

$$\# \text{ of white nodes} = n_R^{(d)} + \sum_{k=1}^{} \deg(b_k) + \deg(\bar b_k) \,. \tag{5.49}$$



(a) Diamond white node.



(b) Normal white nodes.

Figure 5: Two types of white nodes

By (5.48), there are at least $|V_0|$ black nodes with odd deg in $[V_0]$ (where $[V_0]$ is defined in the obvious way). WLOG, we may assume these nodes are $b_1, \ldots, b_{|V_0|}$. To have nonzero expectation, each white block must contain at least two white nodes. Therefore for each $k = 1, \ldots, |V_0|$, 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 $W_1, \ldots, W_d$ be the distinct blocks among all $W(b_k)$'s. Define

$$V_{00} := \{b_k \mid A(b_k) \text{ has no normal white nodes, } 1 \le k \le |V_0|\},$$

and

$$V_{01} := \{b_k \mid A(b_k) \text{ has at least one normal white node, } 1 \le k \le |V_0|\}.$$

The following lemma gives the key estimates we need.

Lemma 5.17. *For any partition $\gamma \in W$,*

$$m(\gamma) \le \frac{|V_{00}| - |V_{01}|/2 - n_R^{(d)} + \sum_{k=1}^{n} \left(\deg(b_k) + \deg(\bar{b}_k)\right)}{2}, \tag{5.50}$$

and

$$n_X + n_R \ge p + |V_1| + |V_{00}|; \quad n_X \ge |V_1|; \quad n_R^{(d)} \ge |V_{00}|. \tag{5.51}$$

Proof. The second inequality of (5.51) can be proved easily through

$$|V_1| \le |\{k \in L \mid F_{\rm in}(k; \Omega_Q^X) \ge 2\}| \le n_X.$$

Notice for $b_k \in V_0$, $A(b_k)$ contains at least three diamond white nodes, while each of the white node is shared by another b_l. Thus we trivially have $|V_{00}| \le n_R^{(d)}$.

Now we prove (5.50). 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 $i = 1, \ldots, d$, if W_i contains exactly one $A(b_k)$, then

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

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

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

Here the first inequality can be understood as following. For each black node b_k with $A(b_k) \subset W_i$, 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 $b_k \in V_{00}$, there are at least three diamond white nodes in $A(b_k)$ with total weight $\ge 3/2$; if $b_k \in V_{01}$, there are at least one normal white node and two other white nodes in $A(b_k)$ with total weight ≥ 2. Thus $\sum_{b_k : A(b_k) \subset W_i} (2 \cdot \mathbf{1}_{V_{01}}(b_k) + 3/2 \cdot \mathbf{1}_{V_{00}}(b_k))$ is smaller than the number of white nodes in W_i. Then summing $|W_i|$ over i, we get

$$\sum_{i=1}^{d} |W_i| \ge 2d + |V_{01}| + \frac{|V_{00}|}{2}.$$

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

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

where we used (5.49) in the last step. This proves (5.50).

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

$$V_{00} \subset \{b_k \in L \,|\, F_{\rm in}(b_k;\Omega\ ) = 1 \text{ and } F_{\rm in}(b_k;\Omega_Q^R) \ge 3\} =: V_2.$$

Recall Lemma 5.15, only ${}_1^{(k)}$ can increase $F_{\rm in}$. Thus w contains ${}_1^{(b_k)}$ for each $b_k \in V_1 \cup V_2$ (recall the definition of V_1 in (5.46)). Therefore by (5.22), (5.37) and the fact that V_{00} and V_1 are disjoint, we have

$$n_X + n_R = F_{\rm o}\,((\Omega\ )_{\mathbf{w}}) \le F_{\rm o}\,(\Omega\ ) + |V_1 \cup V_2| \le p + |V_1| + |V_{00}|.$$

This proves the first inequality of (5.51). □

Now we prove (5.40). By (2.3) and (5.6), 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

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

Then using Lemma 2.21, we get

$$\begin{aligned}
&|\mathbb{E}Q(\mathrm{w};\Delta(\Gamma))| \\
&\quad \prec CU\, N^{-1/2\, n_X} \sum_{W} \sum_{(W_1),\dots,(W_m)\in I \setminus L} \Phi^{n_R - n_R^{(d)}} \prod_{k=1} N^{-1/2\,(\deg(b_k)+\deg(\bar b_k))} \\
&\quad \prec CUN^{-n_X/2} \sum_{W} N^{m - \frac{\sum_{k=1}^{p} \deg(b_k)+\deg(\bar b_k)}{2}} \Phi^{n_R - n_R^{(d)}} \\
&\quad \le CUN^{-n_X/2} \sum_{W} N^{-\frac{|V_{01}| - |V_{00}|/2 - n_R^{(d)}}{2}} \Phi^{n_R - n_R^{(d)}} \\
&\quad \le CUN^{-h/2} \sum_{W} N^{-(n_X - |V_1|)/2} N^{-(n_R^{(d)} - |V_{00}|)/2} \Phi^{n_R - n_R^{(d)}} \\
&\quad \le CUN^{-h/2} \sum_{W} \Phi^{n_X + n_R - |V_1| - |V_{00}|} \le CUN^{-h/2} \Phi^{p},
\end{aligned}$$

where in the third step we used (5.50), in the fourth step $h = |V| = |V_1| + |V_{00}| + |V_{01}|$, in the fifth step $N^{-1/2} \le \Phi$ and (5.51), and in the last step (5.51). Thus we have proved (5.40), which concludes the proof of Proposition 5.1.

6 Anisotropic local law: self-consistent comparison

In this section we prove Theorem 2.19. We first prove the anisotropic and averaged local laws under the vanishing third moment assumption (2.23). When $N^{-1/2+\epsilon} \le |m_{2c}|^{-1}$, the anisotropic and averaged local laws can be established without assuming (2.23). 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 2.18 and 2.19, we have

$$H(TX - z;w) = \overline{T} \begin{pmatrix} -w(D^{\dagger}D)^{-1} & w^{1/2}(V_1X - (UD)^{-1}z) \\ w^{1/2}(V_1X - (UD)^{-1}z)^{\dagger} & -wI \end{pmatrix} \overline{T}^{\dagger},$$
$$\overline{T} := \begin{pmatrix} UD & 0 \\ 0 & I \end{pmatrix}. \tag{6.1}$$

Now we define

$$G(w) := |w|^{1/2} \begin{pmatrix} -w(D^{\dagger}D)^{-1} & w^{1/2}\left(V_1X - (UD)^{-1}z\right) \\ \left(w^{1/2}\left(V_1X - (UD)^{-1}z\right)\right)^{\dagger} & -wI \end{pmatrix}^{-1} = |w|^{1/2}\overline{T}^{\dagger} G \overline{T}. \tag{6.2}$$

Since T is invertible and $\|T\| + \|T^{-1}\| \lesssim 1$ by (2.4), to prove the anisotropic law in Theorem 2.19, it suffices to show that for all deterministic unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^{\mathcal{I}}$,

$$\left| \left\langle \mathbf{u}, \left(G(w) - \widehat{\Pi}(w) \right) \mathbf{v} \right\rangle \right| \prec \Phi(w), \tag{6.3}$$

where

$$\widehat{\Pi}(w) := |w|^{1/2}\overline{T}^{\dagger}\Pi(w)\overline{T}, \quad \Phi(w) := |w|^{1/2}\Psi(w). \tag{6.4}$$

Notice we have $\|\widehat{\Pi}\| = O(1)$ by (3.31). By the anisotropic local law in Theorem 2.18 and the remark around (2.50), if $X = X^{Gauss}$ is Gaussian, then (6.3) holds. Hence for a general X, it suffices to prove that

$$\left| \left\langle \mathbf{u}, \left(G(X;w) - G(X^{Gauss};w) \right) \mathbf{v} \right\rangle \right| \prec \Phi(w). \tag{6.5}$$

Similar to Lemma 3.5, we can prove the following estimates for G.

Lemma 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 $C > 0$,*

$$\sum_{\mu \in \mathcal{I}_2} |G_{\mathbf{w}\mu}|^2 = |w|^{1/2}\frac{\operatorname{Im} G_{\mathbf{ww}}}{\eta}; \tag{6.6}$$

$$\sum_{i \in \mathcal{I}_1^M} |G_{\mathbf{u}\mathbf{v}_i}|^2 \le C|w|^{1/2}\frac{\operatorname{Im} G_{\mathbf{uu}}}{\eta}; \tag{6.7}$$

$$\sum_{i \in \mathcal{I}_1^M} |G_{\mathbf{w}\mathbf{v}_i}|^2 \le C\left(|w|^{-1/2}|G_{\mathbf{ww}}| + |w|^{1/2}\frac{\operatorname{Im} G_{\mathbf{ww}}}{\eta} \right); \tag{6.8}$$

$$\sum_{\mu \in \mathcal{I}_2} |G_{\mathbf{u}\mu}|^2 \le C\left(|w|^{-1/2}|G_{\mathbf{uu}}| + |w|^{1/2}\frac{\operatorname{Im} G_{\mathbf{uu}}}{\eta} \right). \tag{6.9}$$

6.1 Self-consistent comparison

Our proof basically follows the arguments in [24, 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 (6.3) by applying (6.10) to the vectors $\mathbf{u} + \mathbf{v}$ and $\mathbf{u} + i\mathbf{v}$, respectively.

Proposition 6.2. *Fix $|z|^2 \le 1 - \tau$ and suppose that the assumptions of Theorem 2.19 hold. If (2.23) holds or $N^{-1/2+\epsilon} \le |m_{2c}|^{-1}$, then for any regular domain $\mathbf{S} \subseteq \mathbf{D}$,*

$$\left| \left\langle \mathbf{v}, \left(G(w) - \widehat{\Pi}(w) \right) \mathbf{v} \right\rangle \right| \prec \Phi(w) \tag{6.10}$$

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

We first assume that (2.23) holds. Then we will show how to modify the arguments to prove the $N^{-1/2+\epsilon} \le |m_{2c}|^{-1}$ case. The proof consists of a bootstrap argument from larger scales to smaller scales in multiplicative increments of $N^{-\delta}$, where

$$\delta \in \left(0, \frac{\epsilon}{2C_0}\right), \tag{6.11}$$

with $C_0 > 0$ being a universal constant that will be chosen large enough in the proof. For any $\eta \geq |m_{1c}|^{-1} N^{-1+\epsilon}$, we define

$$\eta_l := \eta N^{\delta l} \text{ for } l = 0, \ldots, L-1, \quad \eta_L := 1, \tag{6.12}$$

where $L \equiv L(\eta) := \max\left\{ l \in \mathbb{N} \,|\, \eta N^{\delta(l-1)} < 1 \right\}$. Note that $L \leq 2\delta^{-1}$.

By (3.13), the function $w \mapsto G(w) - \widehat{\Pi}(w)$ is Lipschitz continuous in $\mathbf{S}$ with Lipschitz constant bounded by CN^3. Thus to prove (6.10) for all $w \in \mathbf{S}$, it suffices to show that (6.10) 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 6.3. *Let $\widehat{\mathbf{S}}$ be an N^{-10}-net of $\mathbf{S}$ such that $|\widehat{\mathbf{S}}| \leq N^{20}$ and*

$$E + i\eta \in \widehat{\mathbf{S}} \Rightarrow E + i\eta_l \in \widehat{\mathbf{S}} \text{ for } l = 1, \ldots, L(\eta).$$

The bootstrapping is formulated in terms of two scale-dependent properties $(\mathbf{A}_m)$ and $(\mathbf{C}_m)$ defined on the subsets

$$\widehat{\mathbf{S}}_m := \left\{ w \in \widehat{\mathbf{S}} \,|\, \operatorname{Im} w \geq N^{-\delta m} \right\}.$$

$(\mathbf{A}_m)$ For all $w \in \widehat{\mathbf{S}}_m$, all deterministic unit vector $\mathbf{v}$, and all X satisfying (2.2)-(2.3), we have

$$\operatorname{Im} G_{\mathbf{vv}}(w) \leq |w|^{1/2} \operatorname{Im}\left[m_{1c}(w) + m_{2c}(w)\right] + N^{C_0\delta}\Phi(w). \tag{6.13}$$

$(\mathbf{C}_m)$ For all $w \in \widehat{\mathbf{S}}_m$, all deterministic unit vector $\mathbf{v}$, and all X satisfying (2.2)-(2.3), we have

$$\left| G_{\mathbf{vv}}(w) - \widehat{\Pi}_{\mathbf{vv}}(w) \right| \leq N^{C_0\delta}\Phi(w). \tag{6.14}$$

It is trivial to see that property $(\mathbf{A}_0)$ holds. Moreover, it is easy to observe the following result.

Lemma 6.4. *For any m, property $(\mathbf{C}_m)$ implies property $(\mathbf{A}_m)$.*

Proof. This result follows from (3.33). □

The key step is the following induction result.

Lemma 6.5. *For any $1 \leq m \leq 2\delta^{-1}$, property $(\mathbf{A}_{m-1})$ implies property $(\mathbf{C}_m)$.*

Combining Lemmas 6.4 and 6.5, we conclude that (6.14) holds for all $w \in \widehat{\mathbf{S}}$. Since δ can be chosen arbitrarily small under the condition (6.11), we conclude that (6.10) holds for all $w \in \widehat{\mathbf{S}}$, and Proposition 6.2 follows. What remains now is the proof of Lemma 6.5. Denote

$$F_{\mathbf{v}}(X, w) = \left| G_{\mathbf{vv}}(X, w) - \widehat{\Pi}_{\mathbf{vv}}(w) \right|. \tag{6.15}$$

By Markov's inequality, it suffices to prove the following lemma.

Lemma 6.6. *Fix $p \in 2\mathbb{N}$ and $m \leq 2\delta^{-1}$. Suppose that the assumptions of Proposition 6.2, (2.23) and property $(\mathbf{A}_{m-1})$ hold. Then we have*

$$\mathbb{E} F_{\mathbf{v}}^p(X, w) \leq \left[N^{C_0\delta}\Phi(w) \right]^p \tag{6.16}$$

for all $w \in \widehat{\mathbf{S}}_m$ and all deterministic unit vectors $\mathbf{v}$.

In the following, we focus on proving Lemma 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 $G_{\mathbf{xy}}$.

Lemma 6.7. *For any $w = E + i\eta \in \mathbf{S}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{C}^{\mathcal{I}}$, we have*

$$\left|G_{\mathbf{x}\mathbf{y}}(w) - \Pi_{\mathbf{x}\mathbf{y}}(w)\right| \prec N^{2\delta} \sum_{l=1}^{L(\eta)} \Big[\operatorname{Im} G_{\mathbf{x}_1\mathbf{x}_1}(E + i\eta_l) + \operatorname{Im} G_{\mathbf{x}_2\mathbf{x}_2}(E + i\eta_l)$$
$$+\operatorname{Im} G_{\mathbf{y}_1\mathbf{y}_1}(E + i\eta_l) + \operatorname{Im} G_{\mathbf{y}_2\mathbf{y}_2}(E + i\eta_l)\Big] + \|\mathbf{x}\|_2\|\mathbf{y}\|_2,$$

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 (6.12).

Proof. The proof is similar to the one for [24, 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 2.5 (ii)).

Lemma 6.8. *Suppose $(\mathbf{A}_{m-1})$ holds, then*

$$G(w) - \Pi(w) = O_\prec(N^{2\delta}\phi) \tag{6.17}$$

and

$$\operatorname{Im} G_{\mathbf{v}\mathbf{v}} \prec N^{2\delta}\left[|w|^{1/2}\operatorname{Im}\left(m_{1c}(w) + m_{2c}(w)\right) + N^{C_0\delta}\,\Phi(w)\right] \tag{6.18}$$

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 (6.13) gives $\operatorname{Im} G_{\mathbf{v}\mathbf{v}}(w) \prec 1.$ The estimate (6.17) now follows immediately from Lemma 6.7. To prove (6.18), 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 $|w|^{-1/2}\operatorname{Im}\, G_{\mathbf{v}\mathbf{v}}(E + i\eta)$ and $\operatorname{Im} m_{1,2c}(E + i\eta)$ to get for $w_1 = E + i\eta_1 \in \widehat{\mathbf{S}}_{m-1}$,

$$\begin{aligned}
\operatorname{Im} G_{\mathbf{v}\mathbf{v}}(w) &\le N^{\delta}\frac{|w|^{1/2}}{|w_1|^{1/2}}\operatorname{Im} G_{\mathbf{v}\mathbf{v}}(w_1) \\
&\prec N^{\delta}\left[|w|^{1/2}\operatorname{Im}\left(m_{1c}(w_1) + m_{2c}(w_1)\right) + N^{C_0\delta}\frac{|w|^{1/2}}{|w_1|^{1/2}}\Phi(w_1)\right] \\
&\le N^{2\delta}\left[|w|^{1/2}\operatorname{Im}\left(m_{1c}(w) + m_{2c}(w)\right) + N^{C_0\delta}\,\Phi(w)\right],
\end{aligned}$$

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 (2.45). □

Now we apply the self-consistent comparison method introduced in [24, Section 7] to prove Lemma 6.6. To organize the proof, we divide it into two small subsections.

6.1.1 Interpolation and expansion

Definition 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 $I_1^M \times I_2$ random matrices, where the matrix $X^\theta = (X^\theta_{i\mu})$ has law

$$\prod_{i \in I_1^M} \prod_{\mu \in I_2} \rho^\theta_{i\mu}(dX_{i\mu}). \tag{6.19}$$

For $\lambda \in \mathbb{R}$, $i \in I_1^M$ and $\mu \in I_2$, we define the matrix $X^{\theta,\lambda}_{(i\mu)}$ through

$$\left(X^{\theta,\lambda}_{(i\mu)}\right)_{j\nu} := \begin{cases} X^\theta_{j\nu} & \text{if } (j,\nu) \neq (i,\mu) \\ \lambda & \text{if } (j,\nu) = (i,\mu) \end{cases}.$$

We also introduce the matrices

$$G^\theta(w) := G\left(X^\theta, w\right), \quad G^{\theta,\lambda}_{(i\mu)}(w) := G\left(X^{\theta,\lambda}_{(i\mu)}, w\right),$$

according to (6.2) and the Definition 2.11.

We shall prove Lemma 6.6 through interpolation matrices X^θ between X^0 and X^1. It holds for X^0 by the anisotropic law in Theorem 2.18 (see the remark above (6.5)).

Lemma 6.10. *Lemma 6.6 holds if $X = X^0$.*

Using (6.19) and fundamental calculus, we get the following basic interpolation formula.

Lemma 6.11. *For $F : \mathbb{R}^{I_1^M \times I_2} \to \mathbb{C}$ we have*

$$\frac{d}{d\theta}\mathbb{E}F(X^\theta) = \sum_{i \in I_1^M} \sum_{\mu \in I_2} \left[\mathbb{E}F\left(X^{\theta,X^1_{i\mu}}_{(i\mu)}\right) - \mathbb{E}F\left(X^{\theta,X^0_{i\mu}}_{(i\mu)}\right)\right] \tag{6.20}$$

provided all the expectations exist.

We shall apply Lemma 6.11 with $F(X) = F^p_{\mathbf{v}}(X, w)$ for $F_{\mathbf{v}}(X, w)$ defined in (6.15). The main work is devoted to proving the following self-consistent estimate for the right-hand side of (6.20).

Lemma 6.12. *Fix $p \in 2\mathbb{N}$ and $m \leq 2\delta^{-1}$. Suppose (2.23) and $(\mathbf{A}_{m-1})$ holds, then we have*

$$\sum_{i \in I_1^M} \sum_{\mu \in I_2} \left[\mathbb{E}F^p_{\mathbf{v}}\left(X^{\theta,X^1_{i\mu}}_{(i\mu)}\right) - \mathbb{E}F^p_{\mathbf{v}}\left(X^{\theta,X^0_{i\mu}}_{(i\mu)}\right)\right] = O\left((N^{C_0\delta}\Phi)^p + \mathbb{E}F^p_{\mathbf{v}}(X^\theta, w)\right) \tag{6.21}$$

for all $\theta \in [0,1]$, all $w \in \mathbf{S}_m$, and all deterministic unit vector $\mathbf{v}$.

Combining Lemmas 6.10, 6.11 and 6.12 with a Grönwall argument, we can conclude the proof of Lemma 6.6 and hence Proposition 6.2.

In order to prove Lemma 6.12, we compare $X^{\theta,X^0_{i\mu}}_{(i\mu)}$ and $X^{\theta,X^1_{i\mu}}_{(i\mu)}$ via a common $X^{\theta,0}_{(i\mu)}$, i.e. under the assumptions of Lemma 6.12, we will prove

$$\sum_{i \in I_1^M} \sum_{\mu \in I_2} \left[\mathbb{E}F^p_{\mathbf{v}}\left(X^{\theta,X^u_{i\mu}}_{(i\mu)}\right) - \mathbb{E}F^p_{\mathbf{v}}\left(X^{\theta,0}_{(i\mu)}\right)\right] = O\left((N^{C_0\delta}\Phi)^p + \mathbb{E}F^p_{\mathbf{v}}(X^\theta, w)\right) \tag{6.22}$$

for all $u \in \{0,1\}$, all $\theta \in [0,1]$, all $w \in \mathbf{S}_m$, and all deterministic unit vector $\mathbf{v}$.

Underlying the proof of (6.22) 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 $w \in \mathbf{S}_m$. Define the $\mathcal{I}^M \times \mathcal{I}^M$ (recall Definition 2.9) matrix $\Delta_{(i\mu)}$ through

$$\left(\Delta_{(i\mu)}\right)_{st} := \delta_{is}\delta_{\mu t} + \delta_{it}\delta_{\mu s}, \quad i \in I_1^M, \ \mu \in I_2. \tag{6.23}$$

Then we have for any $;\ ^0 \in \mathbb{R}$ and $K \in \mathbb{N}$,

$$G_{(i\ )}^{;\ 0} = G_{(i\ )}^{;} + \sum_{k=1}^{K} {}^{k} G_{(i\ )}^{;} \left(\overline{V} \Delta_{(i\ )}^{0} \overline{V}^{y} G_{(i\ )}^{;} \right)^{k} + {}^{K+1} G_{(i\ )}^{;\ 0} \left(\overline{V} \Delta_{(i\ )}^{0} \overline{V}^{y} G_{(i\ )}^{;} \right)^{K+1}; \tag{6.24}$$

where $\overline{V} := \begin{pmatrix} V_1 & 0 \\ 0 & I \end{pmatrix}$ and $:= \frac{w^{1/2}}{|w|^{1/2}}$. The following result provides a priori bounds for the entries of $G_{(i\ )}^{;}$.

Lemma 6.13. *Suppose that y is a random variable satisfying $|y| \le N^{-1/2}$. Then*

$$\left| G_{(i\ )}^{;y} \right|_{\max} = O\ (N^{2}\ ) \tag{6.25}$$

for all $i \in I_1^M$ and $\in I_2$.

Proof. See [24, Lemma 7.14]. □

In the following, for simplicity of notations, we introduce $f_{(i\ )}(\ ) := F_{\mathbf{v}}^{p}(X_{(i\ )}^{;})$. We use $f_{(i\ )}^{(n)}$ to denote the n-th derivative of $f_{(i\ )}$. By Lemma 6.13 and expansion (6.24) we get the following result.

Lemma 6.14. *Suppose that y is a random variable satisfying $|y| \le N^{-1/2}$. Then for fixed $n \in \mathbb{N}$,*

$$\left| f_{(i\ )}^{(n)}(y) \right| \le N^{2\ (n+p)}. \tag{6.26}$$

By this lemma, the Taylor expansion of $f_{(i\ )}$ gives

$$f_{(i\ )}(y) = \sum_{n=0}^{4p} \frac{y^n}{n!} f_{(i\ )}^{(n)}(0) + O\ (\Phi^p); \tag{6.27}$$

provided C_0 is chosen large enough in (6.11). Therefore we have for $u \in \{0, 1\}$,

$$\begin{aligned} &\mathbb{E} F_{\mathbf{v}}^{p}\left(X_{(i\ )}^{;X_i^u} \right) - \mathbb{E} F_{\mathbf{v}}^{p}\left(X_{(i\ )}^{;0} \right) \\ &\quad = \mathbb{E}\left[f_{(i\ )}\left(X_i^u \right) - f_{(i\ )}(0) \right] \\ &\quad = \mathbb{E}\, f_{(i\ )}(0) + \frac{1}{2N} \mathbb{E}\, f_{(i\ )}^{(2)}(0) + \sum_{n=4}^{4p} \frac{1}{n!} \mathbb{E}\, f_{(i\ )}^{(n)}(0)\, \mathbb{E}\left(X_i^u \right)^n + O\ (\Phi^p); \end{aligned}$$

where we used that X_i^u has vanishing first and third moments and its variance is $1/N$. Thus to show (6.22), we only need to prove for $n = 4, 5, \ldots, 4p$,

$$N^{-n/2} \sum_{i \in I_1^M} \sum_{\in I_2} \left| \mathbb{E}\, f_{(i\ )}^{(n)}(0) \right| = O\ \left((N^{C_0}\ \Phi)^p + \mathbb{E} F_{\mathbf{v}}^{p}(X\ ; w) \right); \tag{6.28}$$

where we used (2.3). In order to get a self-consistent estimate in terms of the matrix X on the right-hand side of (6.28), we want to replace $X_{(i\ )}^{;0}$ in $f_{(i\ )}(0) := F_{\mathbf{v}}^{p}(X_{(i\ )}^{;0})$ with $X\ = X_{(i\ )}^{;X_{(i\ )}}$.

Lemma 6.15. *Suppose that*

$$N^{-n/2} \sum_{i \in I_1^M} \sum_{\in I_2} \left| \mathbb{E}\, f_{(i\ )}^{(n)}(X_{i\ }) \right| = O\ \left((N^{C_0}\ \Phi)^p + \mathbb{E} F_{\mathbf{v}}^{p}(X\ ; w) \right) \tag{6.29}$$

holds for $n = 4, \ldots, 4p$, Then (6.28) holds for $n = 4, \ldots, 4p$.

Proof. From (6.27) we can get

$$f_{(i\mu)}^{(l)}(0) = f_{(i\mu)}^{(l)}(y) - \sum_{n=1}^{4p-l} \frac{y^n}{n!} f_{(i\mu)}^{(l+n)}(0) + O_\prec(N^{-l/2}\Phi^p). \tag{6.30}$$

The result follows by repeatedly applying (6.30). The details can be found in [24, Lemma 7.16]. □

6.1.2 Conclusion of the proof with words

What remains now is to prove (6.29). For simplicity, we abbreviate $X \quad X$ for the remainder of the proof. In order to exploit the detailed structure of the derivatives on the left-hand side of (6.29), we introduce the following algebraic objects.

Definition 6.16 (Words)**.** *Given $i \in I_1^M$ and $\mu \in I_2$. 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 $[\mathbf{i}] := \mathbf{v}_i$, $[\boldsymbol{\mu}] := \mu$, where $\mathbf{v}_i \in \mathbb{C}^{I_1}$ is defined in Lemma 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

$$A_{\mathbf{v},i,\mu}(w) := G_{\mathbf{v}\mathbf{v}} - \Pi_{\mathbf{v}\mathbf{v}}.$$

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

$$A_{\mathbf{v},i,\mu}(w) := G_{\mathbf{v}[\mathbf{t}_1]} G_{[\mathbf{s}_2][\mathbf{t}_2]} \cdots G_{[\mathbf{s}_n][\mathbf{t}_n]} G_{[\mathbf{s}_{n+1}]\mathbf{v}}. \tag{6.31}$$

Notice the words are constructed such that, by (6.24),

$$\left(\frac{\partial}{\partial X_{i\mu}}\right)^n \left(G_{\mathbf{v}\mathbf{v}} - \Pi_{\mathbf{v}\mathbf{v}}\right) = (-1)^n n! \sum_{w \in \mathcal{W}_n} A_{\mathbf{v},i,\mu}(w)$$

for $n = 0, 1, 2, \ldots$, with which we get that

$$\left(\frac{\partial}{\partial X_{i\mu}}\right)^n F_{\mathbf{v}}^p(X) = (-1)^n n! \sum_{n_1+\cdots+n_p=n} \prod_{r=1}^{p/2} \frac{1}{n_r! n_{r+p/2}!} \left(\sum_{w_r \in \mathcal{W}_{n_r}} \sum_{w_{r+p/2} \in \mathcal{W}_{n_{r+p/2}}} A_{\mathbf{v},i,\mu}(w_r) \overline{A_{\mathbf{v},i,\mu}(w_{r+p/2})} \right).$$

Then to prove (6.29), it suffices to show that

$$N^{-n/2} \sum_{i \in I_1^M} \sum_{\mu \in I_2} \left| \mathbb{E} \prod_{r=1}^{p/2} A_{\mathbf{v},i,\mu}(w_r) \overline{A_{\mathbf{v},i,\mu}(w_{r+p/2})} \right| = O_\prec\left((N^{C_0\delta}\Phi)^p + \mathbb{E}F_{\mathbf{v}}^p(X, w)\right) \tag{6.32}$$

for $4 \le n \le 4p$ and all words $w_1, \ldots, w_p \in \mathcal{W}$ 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

$$N^{-n/2} \sum_{i \in I_1^M} \sum_{\mu \in I_2} \left| \mathbb{E} \prod_{r=1}^{p} A_{\mathbf{v},i,\mu}(w_r) \right| = O_\prec\left((N^{C_0\delta}\Phi)^p + \mathbb{E}F_{\mathbf{v}}^p(X, w)\right). \tag{6.33}$$

The proof of (6.32) is essentially the same but with slightly heavier notations. Treating empty words separately, we find it suffices to prove

$$N^{-n/2}\sum_{i\in I_1^M}\sum_{\mu\in I_2}\mathbb{E}\, A^{p-q}_{\mathbf{v};i;\mu}(w_0)\prod_{r=1}^{q}A_{\mathbf{v};i;\mu}(w_r) = O\left((N^{C_0}\Phi)^p+\mathbb{E}F^p_{\mathbf{v}}(X;w)\right) \tag{6.34}$$

for $4\le n\le 4p$, $1\le q\le p$, and w_r such that $n(w_0)=0$, $\sum_r n(w_r)=n$ and $n(w_r)\ge 1$ for $r\ge 1$.

To estimate (6.34) we introduce the quantity

$$R_s := |G_{\mathbf{v}\mathbf{v}_s}|+|G_{\mathbf{v}_s\mathbf{v}}|. \tag{6.35}$$

for $s\in I$, where as a convention we let $\mathbf{v}_\mu = e_\mu$ for $\mu\in I_2$.

Lemma 6.17. *For $w\in W$ we have the rough bound*

$$|A_{\mathbf{v};i;\mu}(w)|\prec N^{2\delta(n(w)+1)}. \tag{6.36}$$

Furthermore, for $n(w)\ge 1$ we have

$$|A_{\mathbf{v};i;\mu}(w)|\prec (R_i^2+R_\mu^2)N^{2\delta(n(w)-1)}. \tag{6.37}$$

For $n(w)=1$ we have better bound

$$|A_{\mathbf{v};i;\mu}(w)|\prec R_iR_\mu. \tag{6.38}$$

Proof. (6.36) follows immediately from the rough bound (6.17) and definition (6.31). For (6.37), we break $A_{\mathbf{v};i;\mu}(w)$ into $G_{\mathbf{v}[\mathbf{t}_1]}(G_{[\mathbf{s}_2][\mathbf{t}_2]}\cdots G_{[\mathbf{s}_n][\mathbf{t}_n]})^{1/2}$ times $(G_{[\mathbf{s}_2][\mathbf{t}_2]}\cdots G_{[\mathbf{s}_n][\mathbf{t}_n]})^{1/2}G_{[\mathbf{s}_{n+1}]\mathbf{v}}$ and use Cauchy-Schwarz inequality. (6.38) follows from the constraint $\mathbf{t}_1\ne\mathbf{s}_2$ in the definition (6.31). □

By pigeonhole principle, if $n\le 2q-2$ there exists at least two words w_r with $n(w_r)=1$. Therefore by Lemma 6.17 we have

$$\left|A^{p-q}_{\mathbf{v};i;\mu}(w_0)\prod_{r=1}^{q}A_{\mathbf{v};i;\mu}(w_r)\right| \prec N^{2\delta(n+q)}F^{p-q}_{\mathbf{v}}(X)\left(\mathbf{1}(n\ge 2q-1)(R_i^2+R_\mu^2)+\mathbf{1}(n\le 2q-2)R_i^2R_\mu^2\right). \tag{6.39}$$

By Lemma 6.1, we have

$$\frac{1}{N}\sum_{i\in I_1^M}R_i^2+\frac{1}{N}\sum_{\mu\in I_2}R_\mu^2 \prec \frac{|w|^{1/2}\operatorname{Im}G_{\mathbf{v}\mathbf{v}}+|w|^{-1/2}|G_{\mathbf{v}\mathbf{v}}|}{N} \prec N^{2\delta}\frac{|w|\operatorname{Im}(m_{1c}+m_{2c})+|w|^{1/2}N^{C_0}\Phi}{N}\le N^{(C_0+2)\delta}\Phi^2, \tag{6.40}$$

where in the second step we used the two bounds in Lemma 6.8, $|w|^{-1/2}=O(|w|\operatorname{Im}m_{1c})$ by Lemma 3.7, and in the last step the definition of Φ. Using the same method we can get

$$\frac{1}{N^2}\sum_{i\in I_1^M}\sum_{\mu\in I_2}R_i^2R_\mu^2 \prec \left(N^{(C_0+2)\delta}\Phi^2\right)^2. \tag{6.41}$$

Plugging (6.40) and (6.41) into (6.39), we get that the left-hand side of (6.34) is bounded by

$$N^{-n/2+2}N^{2\delta(n+q+2)}\mathbb{E}F^{p-q}_{\mathbf{v}}(X)\left(\mathbf{1}(n\ge 2q-1)\left(N^{C_0/2}\Phi\right)^2+\mathbf{1}(n\le 2q-2)\left(N^{C_0/2}\Phi\right)^4\right).$$

Using $\Phi \gtrsim N^{-1/2}$, we find that the left hand side of (6.34) is bounded by

$$N^{2-(n+q+2)} \mathbb{E} F_{\mathbf{v}}^{p-q}(X) \left[\mathbb{1}(n \geq 2q-1)\left(N^{C_0 \epsilon/2}\Phi\right)^{n-2} + \mathbb{1}(n \leq 2q-2)\left(N^{C_0 \epsilon/2}\Phi\right)^{n} \right]$$

$$\leq \mathbb{E} F_{\mathbf{v}}^{p-q}(X) \left[\mathbb{1}(n \geq 2q-1)\left(N^{C_0 \epsilon/2+1/2}\,\Phi\right)^{n-2} + \mathbb{1}(n \leq 2q-2)\left(N^{C_0 \epsilon/2+1/2}\,\Phi\right)^{n} \right]$$

where we used that $q \leq n$ and $n \geq 4$. Choose $C_0 \geq 25$, then by (6.11) we have $N^{C_0\epsilon/2+1/2} \leq N^{-\epsilon/2}$ and hence $N^{C_0\epsilon/2+1/2}\,\Phi \leq 1$. Moreover, if $n \geq 4$ and $n \geq 2q-1$, then $n \geq q+2$. Therefore we conclude that the left-hand side of (6.34) is bounded by

$$\mathbb{E} F_{\mathbf{v}}^{p-q}(X) \left(N^{C_0 \epsilon}\,\Phi\right)^{q}. \tag{6.42}$$

Now (6.34) follows from Holder's inequality. This concludes the proof of (6.29), and hence of (6.22), and hence of Lemma 6.5. This finishes the proof of Proposition 6.2 under the assumption (2.23).

In the rest of this section, we prove Proposition 6.2 when $\eta \leq N^{-1/2+\epsilon}|m_{2c}|^{-1}$ without assuming (2.23). In this case, we can verify that

$$\Phi \leq N^{-1/4-\epsilon/2}. \tag{6.43}$$

Following the previous arguments, we see that it suffices to prove the estimate (6.29) for $n = 3$. In other words, we need to prove the following lemma.

Lemma 6.18. *Fix $1 \leq m \leq 2\epsilon^{-1}$ and $p \in 2\mathbb{N}$. Let $w \in \mathbf{S}_m \setminus \mathbf{D}$ (recall (2.44)) and suppose $(\mathbf{A}_{m-1})$ holds. Then we have*

$$N^{-3/2} \sum_{i \in \mathcal{I}_1^M} \sum_{\mu \in \mathcal{I}_2} \mathbb{E}\, f^{(3)}_{(i\mu)}(X_{i\mu}) = O\left((N^{C_0\epsilon}\,\Phi)^p + \mathbb{E}F^p_{\mathbf{v}}(X, w) \right). \tag{6.44}$$

Proof. The main new ingredient of the proof is a further iteration step at a fixed w. Suppose

$$G - \tilde{\Pi} = O_\prec(N^{2\delta}\Phi) \tag{6.45}$$

for some $\delta \leq 1$. By the a priori bound (6.17), (6.45) holds for $\delta = 1$. Assuming (6.45), we shall prove a self-improving bound of the form

$$N^{-3/2} \sum_{i \in \mathcal{I}_1^M} \sum_{\mu \in \mathcal{I}_2} \mathbb{E}\, f^{(3)}_{(i\mu)}(X_{i\mu}) = O\left((N^{C_0\epsilon}\,\Phi)^p + (N^{\delta-\epsilon/4}\Phi)^p + \mathbb{E}F^p_{\mathbf{v}}(X, w) \right). \tag{6.46}$$

Once (6.46) is proved, we can use it iteratively to get an increasingly accurate bound for the left hand side of (6.14). After each step, we obtain a better a priori bound (6.45) where δ is reduced by $N^{-\epsilon/4}$. Hence after $O(\epsilon^{-1})$ iterations we can get (6.44).

As in Section 6.1.2, to prove (6.46) it suffices to show

$$N^{-3/2} \sum_{i \in \mathcal{I}_1^M} \sum_{\mu \in \mathcal{I}_2} \left| A^{p-q}_{\mathbf{v},i,\mu}(w_0) \prod_{r=1}^{q} A_{\mathbf{v},i,\mu}(w_r) \right| \prec F^{p-q}_{\mathbf{v}}(X)\left(N^{(C_0-1)\epsilon}\,\Phi + N^{\delta-\epsilon/2}\Phi\right)^q, \tag{6.47}$$

which follows from the bound

$$N^{-3/2} \sum_{i \in \mathcal{I}_1^M} \sum_{\mu \in \mathcal{I}_2} \left| \prod_{r=1}^{q} A_{\mathbf{v},i,\mu}(w_r) \right| \prec \left(N^{(C_0-1)\epsilon}\,\Phi + N^{\delta-\epsilon/2}\Phi\right)^q. \tag{6.48}$$

Each of the three cases $q = 1, 2, 3$ can be proved as in [24, Lemma 12.7], and we leave the details to the reader. This concludes Lemma 6.18. □

6.2 Averaged local law for TX

In this section we prove the averaged local law in Theorem 2.19. Again for convenience, we only consider the case with $w \in \mathbf{D}$ and $|z|^2 \le 1 - \tau$. First we assume (2.23) holds. The anisotropic local law proved in the previous section gives a good a priori bound. In analogy to (6.15), we define

$$\tilde{P}(X;w) := |w|^{1/2}|m_2(w) - m_{2c}(w)| = \left|\frac{1}{N}\sum_{\mu\in\mathcal{I}_2}\left(G_{\mu\mu}(w) - |w|^{1/2}m_{2c}(w)\right)\right|.$$

Since $\Phi^2 = O(|w|^{1/2}/(N\eta))$, it suffices to prove that $\tilde{P} \prec \Phi^2$. Following the argument in Section 6.1, analogous to (6.29), we only need to prove that

$$N^{-n/2}\sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}\left|\mathbb{E}\left(\frac{\partial}{\partial X_{i\mu}}\right)^n \tilde{P}^p(X)\right| = O\left((N^{\delta}\Phi^2)^p + \mathbb{E}\tilde{P}^p(X)\right) \tag{6.49}$$

for all $n = 4, \ldots, 4p$. Here $\delta > 0$ is an arbitrary positive constant. Analogously to (6.33), it suffices to prove that for $n = 4, \ldots, 4p$,

$$N^{-n/2}\sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}\left|\mathbb{E}\prod_{r=1}^{p}\left(\frac{1}{N}\sum_{\nu\in\mathcal{I}_2}A_{\mathbf{e}_\nu;i\mu}(w_r)\right)\right| = O\left((N^{\delta}\Phi^2)^p + \mathbb{E}\tilde{P}^p(X)\right) \tag{6.50}$$

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

$$A_{\mathbf{v};i\mu}(w) := G_{\mathbf{vv}} - |w|^{1/2}m_{2c}.$$

Similar to (6.35) we define

$$R_{\alpha;s} := |G_{\alpha\mathbf{v}_s}| + |G_{\mathbf{v}_s\alpha}|. \tag{6.51}$$

By the anisotropic local law, $G - \Pi = O_\prec(\Phi)$. Hence combining with Lemma 6.1 and (3.33), we get

$$\frac{1}{N}\sum_{\mu\in\mathcal{I}_2}R^2_{\mu;s} \prec \frac{|w|^{1/2}\operatorname{Im}G_{\mathbf{v}_s\mathbf{v}_s} + \eta|w|^{-1/2}|G_{\mathbf{v}_s\mathbf{v}_s}|}{N\eta} \prec \frac{|w|\operatorname{Im}(m_{1c}+m_{2c}) + |w|^{1/2}\Phi}{N\eta} = O(\Phi^2). \tag{6.52}$$

Since $G = O_\prec(1)$ by the anisotropic local law, we have

$$\left|\frac{1}{N}\sum_{\nu\in\mathcal{I}_2}A_{\mathbf{e}_\nu;i\mu}(w)\right| \prec \frac{1}{N}\sum_{\nu\in\mathcal{I}_2}\left(R^2_{\nu;i} + R^2_{\nu;\mu}\right) \prec \Phi^2 \text{ for } n(w) \ge 1. \tag{6.53}$$

Following (6.53), for $n \ge 4$, the left-hand side of (6.50) is bounded by

$$\mathbb{E}\tilde{P}^{p-q}(X)(\Phi^2)^q.$$

Applying Holder's inequality, we conclude the proof.

Then we prove the averaged local law when $\eta \le N^{-1/2+\epsilon}|m_{2c}|^{-1}$. It suffices to prove

$$N^{-3/2}\sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}\left|\mathbb{E}\left(\frac{\partial}{\partial X_{i\mu}}\right)^3\tilde{P}^p(X)\right| = O\left(\left(N^{\delta}\Phi^2\right)^p + \left(\frac{N^{-c_0/2}}{N\eta}\right)^p + \mathbb{E}\tilde{P}^p(X)\right), \tag{6.54}$$

for some small constant $c_0 > 0$. Analogous to the above arguments, it reduces to show that

$$N^{-3/2}\sum_{i\in\mathcal{I}_1^M}\sum_{\mu\in\mathcal{I}_2}\left|\mathbb{E}\prod_{r=1}^{q}\left(\frac{1}{N}\sum_{\nu\in\mathcal{I}_2}A_{\mathbf{e}_\nu;i\mu}(w_r)\right)\right| = O\left(\Phi^{2q} + \left(\frac{N^{-c_0}}{N\eta}\right)^q\right), \tag{6.55}$$

where q is the number of words with nonzero length. Again we can prove the three cases $q = 1, 2, 3$ as in [24, Lemma 12.8], and we leave the details to the reader. This concludes the averaged local law.

A Properties of $\rho_{1,2c}$ and Stability of (2.11)

A.1 Proof of Lemma 2.3 and Proposition 2.14

We now prove Lemma 2.3. First is a technical lemma for f defined in (2.15).

Lemma A.1. *For $w > 0$ and $|z| > 0$, f can be written as*

$$f(\sqrt{w}, m) = -\left(\sqrt{w} + m + w^{-1/2} + \frac{1}{N}\sum_{i=1}^{n} l_i s_i \left(\frac{A_i}{m-a_i} + \frac{B_i}{m-b_i} + \frac{C_i}{m+c_i}\right)\right), \tag{A.1}$$

where we have the following estimates for the poles and the coefficients,

$$\max\left(|z|, \frac{s_i+|z|^2}{\sqrt{w}}\right) < a_i < \frac{s_i+|z|^2}{\sqrt{w}} + |z|, \quad a_n < a_{n-1} < \dots < a_1, \tag{A.2}$$

$$0 < b_1 < b_2 < \dots < b_n < \min\left(|z|, \frac{|z|^2}{\sqrt{w}}\right), \tag{A.3}$$

$$\frac{-(s_i+|z|^2) + \sqrt{(s_i+|z|^2)^2 + 4w|z|^2}}{2\sqrt{w}} < c_i < |z|, \quad c_1 < c_2 < \dots < c_n, \tag{A.4}$$

and

$$0 < A_i \le 2\frac{s_i+|z|^2+\sqrt{w}|z|}{w}, \quad 0 < B_i \le 2\frac{s_i+|z|^2+\sqrt{w}|z|}{w}, \quad 0 < C_i \le \frac{s_i+|z|^2+\sqrt{w}|z|}{w}. \tag{A.5}$$

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 (2.15):

$$\frac{m^2 - |z|^2}{\sqrt{w}m^3 - (s_i+|z|^2)m^2 - \sqrt{w}|z|^2 m + |z|^4} = \frac{A_i'}{m-a_i} + \frac{B_i'}{m-b_i} - \frac{C_i'}{m+c_i}, \tag{A.6}$$

where

$$A_i' = \frac{a_i^2 - |z|^2}{\sqrt{w}(a_i-b_i)(a_i+c_i)}, \quad B_i' = \frac{b_i^2 - |z|^2}{\sqrt{w}(b_i-a_i)(b_i+c_i)}, \quad C_i' = \frac{c_i^2 + |z|^2}{\sqrt{w}(c_i+a_i)(c_i+b_i)}. \tag{A.7}$$

We take $s_i = 0$ in p_i and call the resulting polynomial as

$$p_0 = \sqrt{w}m^3 - |z|^2m^2 - \sqrt{w}|z|^2 m + |z|^4 = \sqrt{w}\left(m - \frac{|z|^2}{\sqrt{w}}\right)\left(m^2 - |z|^2\right),$$

which has roots $m = \pm|z|, |z|^2/\sqrt{w}$. By (2.7), we have $p_1 < p_2 < \dots < p_n < p_0$ for all $m \ne 0$. Comparing the graphs of p_i's (as cubic functions of m) for $0 \le i \le n$, we get that

$$\max\left(|z|, \frac{|z|^2}{\sqrt{w}}\right) < a_n < a_{n-1} < \dots < a_1, \quad 0 < b_1 < b_2 < \dots < b_n < \min\left(|z|, \frac{|z|^2}{\sqrt{w}}\right), \tag{A.8}$$

and

$$0 < c_1 < c_2 < \cdots < c_n < |z|. \tag{A.9}$$

Thus we get (A.3). 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 (A.7), give that $A_i' > 0$, $B_i' > 0$ and $C_i' > 0$. Plugging (A.6) into f, we get immediately (A.1) 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 (2.15) and using the normalization condition (2.8).

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

$$a_i < \frac{(s_i+|z|^2) + \sqrt{(s_i+|z|^2)^2 + 4w|z|^2}}{2\sqrt{w}} < \frac{s_i+|z|^2}{\sqrt{w}} + |z|, \tag{A.10}$$

and

$$c_i > \frac{-(s_i+|z|^2) + \sqrt{(s_i+|z|^2)^2 + 4w|z|^2}}{2\sqrt{w}}. \tag{A.11}$$

Combining (A.9) and (A.11), we get (A.4). 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 (A.8) and (A.10), we get (A.2).

Finally we estimate the coefficients A_i, B_i and C_i. Using (A.7) and (A.2)-(A.4), 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

$$A_i = A_i' a_i \le \frac{2}{\sqrt{w}}\left(\frac{s_i+|z|^2}{\sqrt{w}} + |z|\right) = 2\frac{s_i+|z|^2+\sqrt{w}|z|}{w}, \tag{A.12}$$

$$B_i = B_i' b_i \le 2\frac{s_i+|z|^2+\sqrt{w}|z|}{w|z|}|z| = 2\frac{s_i+|z|^2+\sqrt{w}|z|}{w}, \tag{A.13}$$

$$C_i = C_i' c_i \le \frac{s_i+|z|^2+\sqrt{w}|z|}{w|z|}|z| = \frac{s_i+|z|^2+\sqrt{w}|z|}{w}. \tag{A.14}$$

This completes the proof. □

In (A.1), it is sometimes convenient to reorder the terms and rename the constants to write f as

$$f(m) = -\sqrt{w} + m + w^{-1/2} + \frac{1}{N}\sum_{k=1}^{2n} \frac{C_k^+}{m - x_k} + \frac{1}{N}\sum_{l=1}^{n} \frac{C_l^-}{m + y_l}, \tag{A.15}$$

where all the constants C_k^+ and C_l^- are positive and chosen such that

$$0 < x_1 < x_2 < \cdots < x_{2n}, \ 0 < y_1 < y_2 < \cdots < y_n. \tag{A.16}$$

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 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 C (see Fig. 6 and Fig. 7).

Lemma A.2. *(Critical points) We have $|C \cap I_{-n}| = |C \cap I_{2n}| = 1$ and $|C \cap I_k| \in \{0, 2\}$ for $k = -n+1, \ldots, 2n-1$.*

Proof. We omit the dependence of f on w for now. By (A.15) we have

$$f'(m) = 1 - \frac{1}{N}\sum_{k=1}^{2n} \frac{C_k^+}{(m-x_k)^2} - \frac{1}{N}\sum_{l=1}^{n} \frac{C_l^-}{(m+y_l)^2},$$

and

$$f''(m) = \frac{1}{N}\sum_{k=1}^{2n} \frac{2C_k^+}{(m-x_k)^3} + \frac{1}{N}\sum_{l=1}^{n} \frac{2C_l^-}{(m+y_l)^3}.$$

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 $|C \cap I_k| \in \{0, 2\}$ 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. $|C \cap I_{-n}| = |C \cap I_{2n}| = 1$. □

From this lemma, we deduce that $|C| = 2p$ is even. We denote by z_{2p} the critical point in I_{-n}, z_1 the critical point in I_{2n}, and $z_2 \geq \cdots \geq z_{2p-1}$ the $2p-2$ critical points in $I_{-n+1} \cup \cdots \cup I_{2n-1}$. 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 ${}_{1c}$ (see e.g. Fig. 6).

Lemma A.3. *(Orderings of the critical values) The critical values are ordered as $h_1 \geq h_2 \geq \cdots \geq h_{2p}$. 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 (2.14), 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 2.3.

For $i = -n, \ldots, 2n$, define the subset $J_i(w) := \{m \in I_i : \partial_m f(\sqrt{w}, m) > 0\}$. From Lemma A.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 $h_1 \geq h_2 \geq \cdots \geq h_{2p}$ is now reformulated as

$$f(J_i) < f(J_j) \text{ whenever } i < j \text{ and } J_i, J_j \neq \emptyset. \tag{A.17}$$

To prove (A.17), we use a continuity argument. Let $t \in (0,1]$ and introduce

$$f^t(m) = -\left(\sqrt{w} + m + w^{-1/2} + \frac{t}{N}\sum_{k=1}^{2n}\frac{C_k^+}{m - x_k} + \frac{t}{N}\sum_{l=1}^{n}\frac{C_l^-}{m+y_l}\right).$$

It is easy to check (A.17) holds for small enough $t > 0$. We claim that

$$J_i \neq \emptyset \implies J_i^t \neq \emptyset \text{ for all } t \in (0,1]. \tag{A.18}$$

This is trivial for $i = -n, 2n$. Recall that for $-n+1 \le i \le 2n-1$, $J_i^t \neq \emptyset$ is equivalent to I_i containing two distinct critical points. Moreover, $\partial_t \partial_m f^t(m) < 0$ in $I_{-n+1} \cup \dots \cup I_{2n-1}$, from which we deduce that the number of distinct critical points in each I_i, $i = -n+1, \dots, 2n-1$, does not decreases as t decreases. This proves (A.18).

Next, suppose that there exist $i < j$ such that $J_i, J_j \neq \emptyset$ and $f(J_i) > f(J_j)$. From (A.18), 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 (A.18). This concludes the proof of (A.17).

To prove the second statement of Lemma A.3, we only need to show that $h_1 \ge C_0(\tau^{-1}|w|^{-1/2} + |z|)\sqrt{w}$ and $h_{2p} \ge C_0(\tau^{-1}|w|^{-1/2} + |z|)\sqrt{w}$ 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

$$\begin{aligned} f(z_1) + \sqrt{w} &\ge (z_1 + y_n)\left[1 + \frac{1}{N}\sum_{k=1}^{2n}\frac{C_k^+}{(z_1 - x_k)^2} + \frac{1}{N}\sum_{l=1}^{n}\frac{C_l^-}{(z_1+y_l)^2}\right] + w^{-1/2} \\ &= 2(z_1 + y_n) + w^{-1/2}, \end{aligned}$$

where we used

$$0 = f'(z_1) = 1 - \frac{1}{N}\sum_{k=1}^{2n}\frac{C_k^+}{(z_1 - x_k)^2} - \frac{1}{N}\sum_{l=1}^{n}\frac{C_l^-}{(z_1+y_l)^2}. \tag{A.19}$$

Now we would like to estimate $z_1 + y_n$. Again using (A.19), we have that

$$\frac{1}{N}\sum_{k=1}^{2n}\frac{C_k^+}{(z_1 - x_{2n})^2} + \frac{1}{N}\sum_{l=1}^{n}\frac{C_l^-}{(z_1 - x_{2n})^2} \ge 1.$$

Then by (A.5) we get

$$z_1 - x_{2n} \le \sqrt{\frac{1}{N}\sum_{k=1}^{2n} C_k^+ + \frac{1}{N}\sum_{l=1}^{n} C_l^-} \le 5\sqrt{\frac{\tau^{-1} + |z|^2 + \sqrt{w}|z|}{w}}.$$

Using the above estimates and (A.2)-(A.4), we obtain that

$$\begin{aligned} f(z_1) &\le 2\left(5\sqrt{\frac{\tau^{-1} + |z|^2 + \sqrt{w}|z|}{w}} + \sqrt{\frac{1+|z|^2}{\sqrt{w}}} + 2|z|\right) + w^{-1/2} - \sqrt{w} \\ &\le C_0(\tau^{-1}|w|^{-1/2} + |z|)\sqrt{w}. \end{aligned}$$

for some constant $C_0 > 0$ that does not depend on τ. □

Proof of Lemma 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

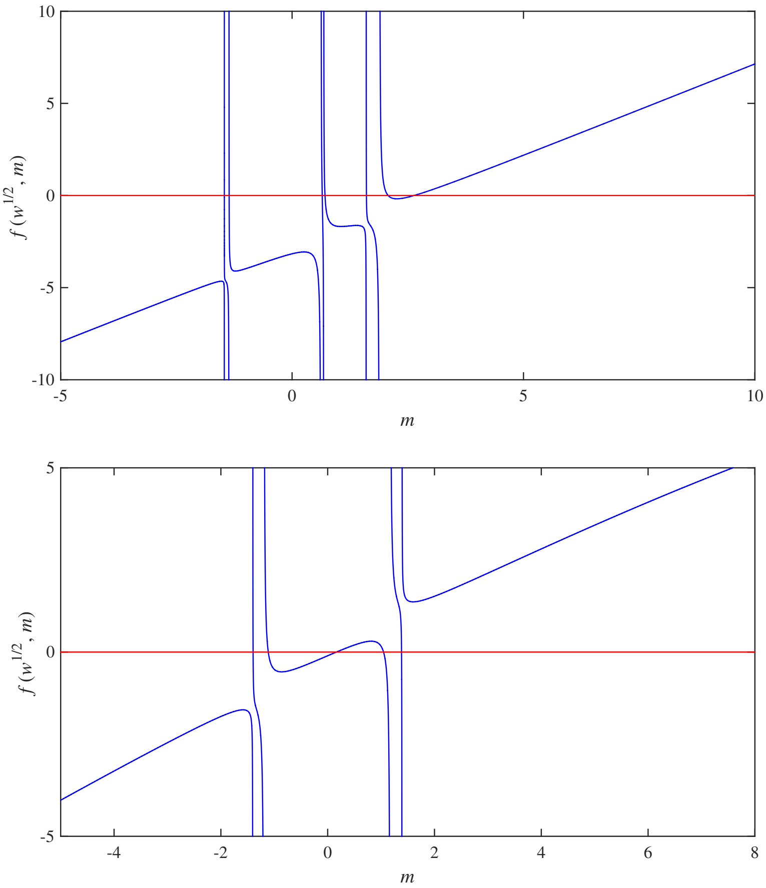


Figure 6: The graphs of $f(\sqrt{w}, m)$ for the example from Figure 1, i.e. $\rho = 0.5\,\delta_{\sqrt{2/17}} + 0.5\,\delta_{4\sqrt{2/17}}$. We take $|z| = 1.5$, and $w = 10$ and 0.01 in the upper and lower graphs, respectively. In the lower graph, we only plot the five branches near $m = 0$. The remaining two branches are far away.

uniqueness of the solution of $P_{w+i\eta}$ in $\mathbb{C}_+$ (Lemma 2.2) and the continuity of the roots of $P_{w+i\eta}$ in η, we conclude that $\mathrm{Im}\, m_c(w) > 0$ and hence $\mathrm{Im}\, m_{1c}(w) > 0$ by taking $\eta \searrow 0$, i.e. $w \in \mathrm{supp}\, \rho_{1c}$. In sum, we get

$$\mathrm{supp}\, \rho_{1c} = \overline{\{w > 0 : 0 \notin f(J(w))\}}. \tag{A.20}$$

From Lemma A.3, we see that there exists an absolute constant $C_1 > 0$ such that if $w \ge C_1^{-1}$, then $h_1(\omega) \le C_0(\tau^{-1}|w|^{-1/2} + |z|) - \sqrt{w} < 0$. Hence fix $w \ge C_1^{-1}$, we have $0 \in f(J_{2n}(w))$ and $w \notin \mathrm{supp}\, \rho_{1c}$ (see the upper graphs in Fig. 6 and Fig. 7). This shows that ρ_{1c} is compactly supported in $[0, C_1^{-1}]$. Now we decrease w so that $w < s_1 + |z|^2 + 1$. Then using (A.2), 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 $0 < w < C_1^{-1}$ such that $0 \notin f(J(w))$. Thus $\mathrm{supp}\, \rho_{1c} \ne \emptyset$. By (A.20), it is not hard to see that $\mathrm{supp}\, \rho_{1c}$ is a disjoint union of (countably many) closed intervals,

$$\mathrm{supp}\, \rho_{1c} = \bigcup_k [e_{2k}, e_{2k-1}], \tag{A.21}$$

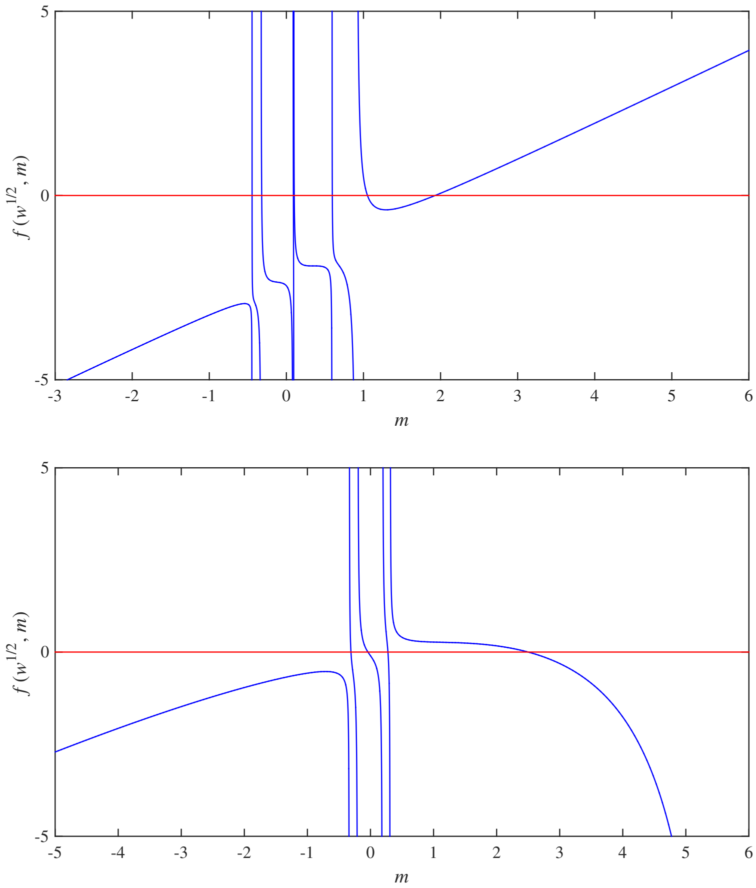


Figure 7: The graphs of $f(\sqrt{w}; m)$ for the example from Figure 1, i.e. $= 0.5 \; \sqrt{2/17} + 0.5 \; {}_4\sqrt{2/17}$. We take $|z| = 0.5$, and $w = 6$ and 0.01 in the upper and lower graphs, respectively. In the lower graph, we only plot the five branches near $m = 0$. The remaining two branches are far away.

where $C_1 \;\; ^{1} \;\; e_1 \;\; e_2 \;\; \ldots$ 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

$$f(\sqrt{e_i}; m_c(e_i)) = 0; \quad \partial_m f(\sqrt{e_i}; m_c(e_i)) = 0. \tag{A.22}$$

Notice the two equations in (A.22) 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 (A.22). Hence there are finitely many e_i's, call them $e_1 \;\; \ldots \;\; e_{2L}$, where $L \;\; L(n) \in \mathbb{N}$. The statement about e_{2L} follows from Lemma A.4 below. This concludes Lemma 2.3. □

Lemma A.4. *If $1 + \;\; |z|^2 \;\; 1 + \;\; ^{1}$, there is a constant $(\;) > 0$ so that $e_{2L} \;\; (\;)$. If $|z|^2 \;\; 1 \;\;$, $e_{2L} = 0$ and ${}_{1c}(x) \;\; x^{1/2}$ 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 (2.5) for T.

Proof of Lemma A.4. We first study the equation (2.14) when $w \searrow 0$ in the case $1 + \tau \le |z|^2 \le 1 + \tau^{-1}$. We calculate the derivative of f as

$$\partial_m f(\sqrt{w}, m) = 1 + \frac{1}{N}\sum_{i=1}^{n} l_i s_i \frac{m^2 - |z|^2}{\sqrt{w}m^3 - (s_i + |z|^2)m^2 - \sqrt{w}|z|^2 m + |z|^4}$$
$$- \frac{m}{N}\sum_{i=1}^{n} l_i s_i \frac{\sqrt{w}\left(m^2 - |z|^2\right)^2 + 2s_i|z|^2 m}{\left[\sqrt{w}m^3 - (s_i + |z|^2)m^2 - \sqrt{w}|z|^2 m + |z|^4\right]^2}. \tag{A.23}$$

Recall the definition of J_i in the proof of Lemma A.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 (A.23) 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

$$f(\sqrt{w}, z_k) = -\sqrt{w} + \frac{z_k^2}{N}\sum_{i=1}^{n} l_i s_i \frac{\sqrt{w}\left(z_k^2 - |z|^2\right)^2 + 2s_i|z|^2 z_k}{\left[\sqrt{w}z_k^3 - (s_i + |z|^2)z_k^2 - \sqrt{w}|z|^2 z_k + |z|^4\right]^2}$$
$$\geq -\sqrt{w} + \frac{1}{N}\sum_{i=1}^{n} l_i s_i \frac{2s_i|z|^2 z_k^3}{|z|^8} \geq c - \sqrt{w} \tag{A.24}$$

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 $1 \le i \le n$. By (A.24), we can find η small enough such that $f(\sqrt{w}, z_k) > 0$ for all $0 < w \le \eta$. 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 A.4.

Now we study equation (2.14) when $|z|^2 \le 1 - \tau$ 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 (2.14), we get $m = 0$ or

$$0 = 1 + \frac{1}{N}\sum_{i=1}^{n} l_i s_i \frac{m^2 - |z|^2}{(s_i + |z|^2)m^2 + |z|^4}. \tag{A.25}$$

We define

$$g(x) := 1 + \frac{1}{N}\sum_{i=1}^{n} l_i s_i \frac{x - |z|^2}{(s_i + |z|^2)x + |z|^4} = \frac{|z|^2}{N}\sum_{i=1}^{n} l_i \frac{x + |z|^2 - s_i}{(s_i + |z|^2)x + |z|^4}. \tag{A.26}$$

It is easy to see that g is smooth and decreasing on the intervals defined through

$$K_1 := \left(-\infty, -\frac{|z|^4}{s_1 + |z|^2}\right), \quad K_{n+1} := \left(-\frac{|z|^4}{s_n + |z|^2}, \infty\right),$$

and

$$K_i := \left(-\frac{|z|^4}{s_{i-1} + |z|^2}, -\frac{|z|^4}{s_i + |z|^2}\right), \quad i = 2, \ldots, n.$$

By the boundary values of g on these intervals, we see that $g(x)$ has exactly one zero on intervals K_i for $i = 1, \ldots, n$, 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 $i = 2, \ldots, n$. 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 (A.26), so $t < \tau^{-1}$. If $|z|^2 \le \tau/2$, then by the concavity of g on the K_1, we get

$$t \ge -\frac{g(0)}{g'(0)} \ge \frac{|z|^4(1-|z|^2)}{s_1} \ge \frac{\epsilon^4}{4}. \tag{A.27}$$

In the case $|z|^2 \ge \epsilon/2$, we have $|z|^2 - s_n \ge \epsilon/2$ and $g(|z|^2 - s_n) \ge 0$ by (A.26). Hence we have

$$t \ge |z|^2 - s_n \ge \epsilon/2. \tag{A.28}$$

Combining (A.27) and (A.28), we get that $c\epsilon^4 \le t \le \epsilon^{-1}$ for some constant $c > 0$.

Now we return to the self-consistent equation (2.14). The previous discussion shows that

$$f(0, i\sqrt{t}) = 0, \quad \text{with } t \ge c\epsilon^4.$$

It is easy to see that there exist constants $c_1, \delta_0 > 0$ such that

$$\left|(s_i+|z|^2)m^2 + |z|^4 + \sqrt{w}\left(m^3 - |z|^2 m\right)\right| \ge c_1 \text{ for } |m - i\sqrt{t}| \le \delta_0. \tag{A.29}$$

First we consider the case $|z| \ge \epsilon > 0$. Expanding $f(\sqrt{w}, m)$ around $(0, i\sqrt{t})$ and using (A.29), we get

$$0 = \partial_{\sqrt{w}} f(0, i\sqrt{t})\sqrt{w} + \partial_m f(0, i\sqrt{t})(m - i\sqrt{t}) + o(\sqrt{w}) + o(m - i\sqrt{t}). \tag{A.30}$$

By (A.23), the partial derivative

$$\partial_{\sqrt{w}} f(\sqrt{w}, m) = -1 - \frac{m^2}{N}\sum_{i=1}^{N} l_i s_i \frac{\left(m^2 - |z|^2\right)^2}{\left[-(s_i+|z|^2)m^2 + |z|^4 + \sqrt{w}(m^3 - |z|^2 m)\right]^2}, \tag{A.31}$$

and (A.29), we obtain that $|\partial_{\sqrt{w}} f(0, i\sqrt{t})| \le C$ and

$$\partial_m f(0, i\sqrt{t}) = \frac{t}{N}\sum_{i=1}^{N} l_i s_i \frac{2 s_i |z|^2}{[(s_i+|z|^2)t + |z|^4]^2} \ge c_2 \tag{A.32}$$

for some constant $c_2 > 0$. Using (A.32), we get from (A.30) that

$$m - i\sqrt{t} = O(\sqrt{w}), \quad \text{if } |z| \ge \epsilon. \tag{A.33}$$

Then assume that $|z|^2 < \epsilon$ for sufficiently small ϵ. From $g(-t) = 0$ and (A.26), we get that

$$\frac{1}{N}\sum_{i=1}^{N} l_i \frac{t + |z|^2 - s_i}{(s_i+|z|^2)t + |z|^4} = 0. \tag{A.34}$$

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 (A.34) up to the first order of $|z|^2$, we get

$$t = t_0 + \left(\frac{t_0^2}{N}\sum_i \frac{l_i}{s_i^2} - 2\right)|z|^2 + O(|z|^4). \tag{A.35}$$

Now we write equation (2.14) as

$$F(\sqrt{w}, m) = 0, \tag{A.36}$$

where $F(\sqrt{w}, m) := f(\sqrt{w}, m)/m$. Expanding F around $(0, i\sqrt{t})$ and using (A.29), we get

$$0 = \partial_{\sqrt{w}} F(0, i\sqrt{t})\sqrt{w} + \partial_m F(0, i\sqrt{t})(m - i\sqrt{t}) + \partial_m \partial_{\sqrt{w}} F(0, i\sqrt{t})(m - i\sqrt{t})\sqrt{w}$$

$$+\frac{1}{2}\partial^2_{\sqrt{w}}F(0,i\sqrt{t})w+\frac{1}{2}\partial^2_m F(0,i\sqrt{t})(m-i\sqrt{t})^2+o(w,|m-i\sqrt{t}|^2,|m-i\sqrt{t}|\sqrt{w}). \tag{A.37}$$

We can calculate that (the partial derivatives of F can be obtained using (A.23) and (A.31))

$$\partial_m F(\sqrt{w},i\sqrt{t})=-\frac{2i|z|^2+2\sqrt{wt_0}}{t_0^{3/2}}+o(|z|^2,\sqrt{w}), \tag{A.38}$$

$$\partial_{\sqrt{w}}F(\sqrt{w},i\sqrt{t})=\left(i|z|^2+2\sqrt{wt_0}\right)\frac{\sqrt{t_0}}{N}\sum_{j=1}\frac{l_j}{s_j^2}+o(|z|^2,\sqrt{w}). \tag{A.39}$$

From (A.38) and (A.39), we get that

$$\partial_m F(0,i\sqrt{t})=-\frac{2i|z|^2}{t_0^{3/2}}+o(|z|^2),\quad \partial_{\sqrt{w}}F(0,i\sqrt{t})=\frac{i|z|^2\sqrt{t_0}}{N}\sum_{j=1}\frac{l_j}{s_j^2}+o(|z|^2),$$

$$\partial_m\partial_{\sqrt{w}}F(0,i\sqrt{t})=-\frac{2}{t_0}+O(|z|^2),\quad \partial^2_{\sqrt{w}}F(0,i\sqrt{t})=\frac{2t_0}{N}\sum_{j=1}\frac{l_j}{s_j^2}+O(|z|^2),$$

$$\partial^2_m F(0,i\sqrt{t})=O(|z|^2).$$

Plugging the above results into (A.37), we get that

$$0=\left[\frac{i|z|^2\sqrt{t_0}+\sqrt{w}t_0}{N}\sum_{j=1}\frac{l_j}{s_j^2}+o(|z|^2)\right]\sqrt{w}+\left[-2\frac{i|z|^2+\sqrt{wt_0}}{t_0^{3/2}}+o(|z|^2)\right](m-i\sqrt{t})$$
$$+o(w,|m-i\sqrt{t}|^2,|m-i\sqrt{t}|\sqrt{w}). \tag{A.40}$$

Observing that $|i|z|^2\sqrt{t_0}+\sqrt{w}t_0| \sim |z|^2+\sqrt{|w|}$, we get

$$m-i\sqrt{t}=\left[\frac{t_0^2}{2N}\sum_{j=1}\frac{l_j}{s_j^2}+O(|w|^{1/2}+|z|^2)\right]\sqrt{w},\quad \text{if } |z|<\epsilon. \tag{A.41}$$

Combing (A.33) and (A.41), we get that if $|z|^2<1-\epsilon$, $m=i\sqrt{t}+O(\sqrt{w})$ when $w\to 0$. In particular, this shows that $|m|\sim \operatorname{Im} m\sim 1$ when $w\to 0$. Finally, we conclude the proof of Lemma A.4 by using that $m_{1c}(w)=m_c(w)w^{-1/2}\sim 1$. □

To prove Proposition 2.14, we need the following lemma, which is a consequence of the edge regularity conditions (2.18) and (2.19).

Lemma A.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 $\min_{l\neq k}|e_l-e_k|\geq\epsilon$ for some constant $\epsilon>0$.*

Proof. Denote $m_k:=m_c(e_k)$ and let $w\to e_k$. Note that by Lemma 2.3 and Lemma A.4, if $e_k\neq 0$, we have

$$\epsilon'\leq e_k\leq C\epsilon^{-1}, \tag{A.42}$$

for some constant $\epsilon'>0$. Then we expand f around $(\sqrt{e_k},m_k)$ to get that

$$0=\partial_{\sqrt{w}}f(\sqrt{e_k},m_k)(\sqrt{w}-\sqrt{e_k})+\frac{1}{2}\partial^2_m f(\sqrt{e_k},m_k)(m_c(w)-m_k)^2$$
$$+O\left(|\sqrt{w}-\sqrt{e_k}|^2+|m_c(w)-m_k|^3+|\sqrt{w}-\sqrt{e_k}||m_c(w)-m_k|\right), \tag{A.43}$$

where by (A.31),

$$\partial_{\sqrt{w}}f(\sqrt{e_k},m_k)=1-\frac{m_k^2}{N}\sum_{i=1}l_is_i\frac{\left(m_k^2-|z|^2\right)^2}{e_k(m_k-a_i)^2(m_k-b_i)^2(m_k+c_i)^2}, \tag{A.44}$$

and by (A.1),

$$\partial_m^2 f(\sqrt{e_k}, m_k) = \frac{2}{N}\sum_{i=1} l_i s_i \left(\frac{A_i}{(m_k-a_i)^3} + \frac{B_i}{(m_k-b_i)^3} + \frac{C_i}{(m_k+c_i)^3}\right). \tag{A.45}$$

Applying (A.2)-(A.5), (A.42) and the conditions (2.18)-(2.19) to (A.44) and (A.45), we get that

$$1 \lesssim \partial_{\sqrt{w}} f(\sqrt{e_k}, m_k) \le C_1, \quad |\partial_m^2 f(\sqrt{e_k}, m_k)| \ge C_2 \tag{A.46}$$

for some $C_1, C_2 > 0$. Similarly, if $|w-e_k| \le \epsilon'$ and $|m_c(w) - m_k| \le \epsilon'$ for some sufficiently small ϵ', using the condition (2.18) we can get that

$$\max\left\{ |\partial_m^3 f(\sqrt{w}, m_c(w))|, |\partial^2_{\sqrt{w}} f(\sqrt{w}, m_c(w))|, |\partial_m \partial_{\sqrt{w}} f(\sqrt{w}, m_c(w))| \right\} \le C_3. \tag{A.47}$$

Plugging them into equation (A.43), for $|w-e_k| \le \epsilon'$ and $|m_c(w) - m_k| \le \epsilon'$, we get $|m_c(w) - m_k| \sim |\sqrt{w} - \sqrt{e_k}|^{1/2}$ and

$$\partial_{\sqrt{w}} f(\sqrt{e_k}, m_k)(\sqrt{w} - \sqrt{e_k}) + O(|\sqrt{w} - \sqrt{e_k}|^{3/2}) = -\frac{1}{2}\partial_m^2 f(\sqrt{e_k}, m_k)(m_c(w) - m_k)^2. \tag{A.48}$$

By (A.42), 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 (A.48), if w is real and $|w-e_k| \le \epsilon'$, we have that

$$m_c(w) - m_k = \left(-\frac{2\partial_{\sqrt{w}} f(\sqrt{e_k}, m_k)}{\partial_m^2 f(\sqrt{e_k}, m_k)} + O(|\sqrt{w} - \sqrt{e_k}|^{1/2})\right)^{1/2} \left(\sqrt{w} - \sqrt{e_k}\right)^{1/2}. \tag{A.49}$$

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 $\min_{l \ne k} |e_l - e_k| \ge \delta$. ◻

Proof of Proposition 2.14. The properties of ρ_{1c} have been proved in Lemmas 2.3, A.4 and A.5, and included in Definition 2.4. Since supp ρ_{2c} = supp ρ_{1c} by the discussion after Lemma 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 (2.34) 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 (2.9). To prove the property (iii) for ρ_{2c}, we need to know the behavior of $\mathrm{Im}\, m_{2c}(w)$ when $w \to e_j$ along the real line. By (2.9), it suffices to prove that if $|x - e_j| \le \epsilon'$ for some small enough $\epsilon' > 0$, then

$$\left|w(1+m_{1c})^2 + |z|^2\right| = \left|m_c^2 - |z|^2\right| \ge \tau$$

for some constant $\tau > 0$. Suppose that $|m_c^2(w) - |z|^2| = o(1)$. Then plugging m_c into $\partial_m f(\sqrt{w}, m_c)$ in (A.23), and using condition (2.18) and Lemma A.5, we get that

$$\partial_m f(\sqrt{w}, m_c(w)) = -1 + O(|m_c^2 - |z|^2|). \tag{A.50}$$

Again using condition (2.18) and Lemma A.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

$$0 = \partial_m f(\sqrt{e_j}, m_c(e_j)) = \partial_m f(\sqrt{w}, m_c(w)) + O(|w-e_j|^{1/2}) = -1 + O(|m_c^2 - |z|^2| + |w-e_j|^{1/2}). \tag{A.51}$$

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 2.8. Given the assumptions in Example 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 supp ρ_{1c} is independent of n, and all the edges and the bulk components are regular as in Example 2.7.

A.2 Proof of Lemmas 3.7 and 3.8

We first prove Lemma 3.7. We consider the five cases separately.

Case 1: For $w = E + i\eta \in \mathbf{D}^b_k(\tau, \tau'; N)$, we have

$$m_{1c}(w) = \int_{\mathbb{R}} \frac{\rho_{1c}(x)}{x - (E + i\eta)} dx, \quad \operatorname{Im} m_{1c}(w) = \int_{\mathbb{R}} \frac{\eta\rho_{1c}(x, z)}{(x - E)^2 + \eta^2} dx. \tag{A.52}$$

By the regularity condition of Definition 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 2.15, 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

$$\operatorname{Im}\left[-w(1 + m_{1c}) + \frac{|z|^2}{1 + m_{1c}}\right] \le -\operatorname{Im}(wm_{1c}). \tag{A.53}$$

Using the bounds for m_{1c} and $\operatorname{Im} m_{1c}$ proved above, it is easy to see that

$$\left|-w(1 + m_{1c}) + \frac{|z|^2}{1 + m_{1c}}\right| = O(1). \tag{A.54}$$

Equations (A.53) and (A.54) together give that $\operatorname{Im} m_{2c} \sim 1$ and $|m_{2c}| \sim 1$ by (2.9). 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 (3.29) follows from the bound

$$\operatorname{Im}\left(-w + s_i wm_{2c} - \frac{|z|^2}{1 + m_{1c}}\right) \ge s_i \operatorname{Im}(wm_{2c}).$$

Case 2: For $w = E + i\eta \in \mathbf{D}^o(\tau, \tau'; N)$, using (A.52) and $\operatorname{dist}(E, \operatorname{supp} \rho_{1,2c}) \ge \tau'$, we immediately get $\operatorname{Im} m_{1,2c} \sim \eta$. Now we prove the other estimates.

We first prove (3.29). If $\eta \gtrsim 1$, the proof is the same as in Case 1. Hence we assume $\eta \le c'$, where $c' \equiv c'(\tau, \tau') > 0$ is sufficiently small. Recall the definitions of $\mathbf{D}$ and $\mathbf{D}^o$ in (2.39) and (2.42), we always have $E \sim 1$ in this case.

We shall prove that

$$\min_i \{|m_c(w) - a_i(w)|, |m_c(w) - b_i(w)|, |m_c(w) + c_i(w)|\} \ge \tau'', \tag{A.55}$$

for some constant τ''. This leads immediately to (3.29) since

$$\left| w\left(1 + s_i \frac{1 + m_{1c}}{-w(1 + m_{1c})^2 + |z|^2}\right)(1 + m_{1c}) - |z|^2 \right| = \left| \frac{\sqrt{w}(m_c - a_i)(m_c - b_i)(m_c + c_i)}{m_c^2 + |z|^2} \right|. \tag{A.56}$$

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

$$\frac{dm_{1c}(E)}{dE} = \int_{\mathbb{R}} \frac{\rho_{1c}(x;z)}{(x-E)^2}dx \ge 0.$$

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

$$a_i(E) - m_c(E) \ge a_i(e_k) - m_c(e_k) + \tau'(\epsilon') \gtrsim \tau'(\epsilon'); \tag{A.57}$$

by using $|E - e_k| \ge \epsilon'$ (see (2.42)). 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

$$m_c(E) - a_i(E) \ge m_c(e_{k-1}) - a_i(e_{k-1}) + \tau'(\epsilon') \gtrsim \tau'(\epsilon'). \tag{A.58}$$

Applying the same arguments to $b_i(E)$ and $-c_i(E)$, we get

$$\min_i \{|m_c(E) - a_i(E)|, |m_c(E) - b_i(E)|, |m_c(E) + c_i(E)|\} \gtrsim \tau' \tag{A.59}$$

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 $|z|^2 \le 1 + \tau$. In this case, we have seen that $0 < m_c(E) < b_i(E)$ for all i in the proof of Lemma A.4. Thus we can use (A.57) to get lower bounds for $|m_c(E) - a_i(E)|$ and $|m_c(E) - b_i(E)|$. Since $c_i(E) \gtrsim 1$ in this case (by (A.4) and using $E, |z| \lesssim 1$), $|m_c(E) + c_i(E)| \gtrsim \tau$ is trivial. Again we get the estimate (A.59).

Then we consider $w = E + i\eta$ with $\eta \le c'$. 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

$$|\partial_w m_{1c}(w)| = \left|\int_{\mathbb{R}} \frac{\rho_{1c}(x;z)}{(x-w)^2}dx\right| \le C \tag{A.60}$$

by the condition $\operatorname{dist}(E, \operatorname{supp}\rho_{1c}) \ge \epsilon'$. Thus we immediately get $|m_c(E + i\eta) - m_c(E)| = O(\eta)$. Hence as long as c' is small enough, (A.55) still holds true, which further gives (3.29).

Now we show that $|1 + m_{1c}| \sim 1$ for $w \in \mathbf{D}^o$ and $\eta \le c'$. In fact, if $|m_c|$ can be arbitrarily small, then by (3.29) we get that

$$f(\sqrt{w}, m_c) = -\sqrt{w} + O(m_c) \ne 0,$$

which gives a contradiction. Finally we have $|m_{2c}| \sim 1$ for $w \in \mathbf{D}^o$ and $\eta \le c'$ by Proposition 2.15.

Case 3: For a regular edge $e_k \ne 0$, we always have $e_k \ge \tau$ for some $\tau > 0$ by Lemma A.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 $\kappa \gtrsim 1$, 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 $\kappa \le \tau'$. By the regularity assumption (2.18) and Lemma A.5, we have

$$\min_{1 \le i \le n} \{|m_c(w) - a_i(w)|, |m_c(w) - b_i(w)|, |m_c(w) + c_i(w)|\} \ge \tau/2 \tag{A.61}$$

uniformly in $w \in \{w \in \mathbf{D}_k^e(\zeta, \tau', N) : \kappa(w) + \eta(w) \le 2\tau'\}$, provided τ' is sufficiently small. The above bound implies (3.29). If $m_c(w) \to 0$, then using (3.29) 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 2.15.

We still need to prove the estimates for $\operatorname{Im} m_{1,2c}$ when $\kappa \le \epsilon'$. Recall the expansion (A.48) around e_k and equation (A.49), 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 (A.49) 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 $E \ge e_k$, 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 $E \le e_k$, 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 $w \sim e_k$ and

$$\operatorname{Im} m_{1c} = \operatorname{Im}\left(w^{-1/2} m_c\right) \sim \operatorname{Im} m_c(w), \quad \operatorname{Im} m_{2c} = \operatorname{Im} \frac{m_c}{\sqrt{w}(-m_c^2 + |z|^2)} \sim \operatorname{Im} m_c(w),$$

where we used that $|m_c^2 - |z|^2| \sim 1$ as observed in the proof of Proposition 2.14 in Section A.1.

Case 4: Again if $\kappa \gtrsim 1$, the proof is the same as in Case 1. If $|w| \le 2\epsilon'$ for small enough ϵ', in the proof of Lemma A.4, we have seen that $m_c = i\sqrt{t} + O(\sqrt{w})$, which gives the first equation in (3.26). Plugging it into (2.9), we get the second equation in (3.26). Taking the imaginary part, we obtain (3.27). Finally using (3.26), we can verify (3.29) easily.

Case 5: For $w = E + i\eta \in \mathbf{D}_L(\varsigma; N)$, the bounds for $m_{1,2}$ and $\operatorname{Im} m_{1,2}$ in (3.28) follows from (A.52) directly.

Proof of Lemma 3.8. The estimates (3.31) and (3.32) follow immediately from (2.32), (3.29) and (3.30). For (3.33), 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

$$\mathbf{u} := \begin{pmatrix} U^\dagger & 0 \\ 0 & U^\dagger \end{pmatrix} \mathbf{v}, \quad u_{[i]} := \begin{pmatrix} u_i \\ u_i \end{pmatrix}.$$

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 3.7. By the definition of $\pi_{[i]c}$ in (2.32), we get

$$\begin{aligned}
\operatorname{Im} \pi_{ii,c} &= \operatorname{Im}\left(-w(1 + |d_i|^2 m_{2c}) + \frac{|z|^2}{1 + m_{1c}}\right)^{-1} \le \frac{C}{|w|} \operatorname{Im}\left(w(1 + |d_i|^2 m_{2c}) - \frac{|z|^2}{1 + m_{1c}}\right) \\
&= \frac{C}{|w|}\left((1 + |d_i|^2 \operatorname{Re} m_{2c}) \operatorname{Im} w + |d_i|^2 (\operatorname{Re} w) \operatorname{Im} m_{2c} + \frac{|z|^2}{|1 + m_{1c}|^2} \operatorname{Im} m_{1c}\right),
\end{aligned}$$

where in the second step we used (3.29) and $|1 + m_{1c}| \gtrsim |w|^{1/2}$. In the first three cases of Lemma 3.7, we have $|w| \sim 1$ and $\operatorname{Im} w = O(\operatorname{Im} m_{1c})$, which give that $\operatorname{Im} \pi_{ii,c}$

$C\mathrm{Im}(m_{1c}+m_{2c})$. In case 4 of Lemma 3.7, we use $|\mathrm{Im}\, w|+|\mathrm{Re}\, w|+|1+m_{1c}|^{-2}=O(|w|)$ and $\mathrm{Im}\, m_{1,2c}\gtrsim |w|^{-1/2}$ to get that $\mathrm{Im}\,\mathcal{G}_{ii,c}\le C\mathrm{Im}(m_{1c}+m_{2c})$. Similarly, we can get the bound $\mathrm{Im}\,\mathcal{G}_{\bar{i}\bar{i},c}\le C\mathrm{Im}(m_{1c}+m_{2c})$. Finally we can estimate the following term using similar methods,

$$\mathrm{Im}\left(\bar{u}_i u_{\bar{i}}\,\mathcal{G}_{i\bar{i},c}+\bar{u}_{\bar{i}} u_i\,\mathcal{G}_{\bar{i}i,c}\right)=2\mathrm{Re}\,(\bar{u}_i u_{\bar{i}} z)\,\mathrm{Im}\left\{w^{-1/2}\left[w(1+|d_i|^2 m_{2c})(1+m_{1c})-|z|^2\right]^{-1}\right\}$$
$$\le C\mathrm{Re}\,(\bar{u}_i u_{\bar{i}} z)\,\mathrm{Im}(m_{1c}+m_{2c})\le C\left(|u_i|^2+|u_{\bar{i}}|^2\right)\mathrm{Im}(m_{1c}+m_{2c}).$$

Combining the above estimates we get $\mathrm{Im}\,\langle u_{[i]},\mathcal{G}_{[i]c}u_{[i]}\rangle\le C|u_{[i]}|^2\mathrm{Im}(m_{1c}+m_{2c})$, which implies (3.33). For the Case 5 of Lemma 3.7, we use (3.28) and (3.32) to get

$$\mathrm{Im}\,\langle u_{[i]},\mathcal{G}_{[i]c}u_{[i]}\rangle\le|u_{[i]}|^2\|\mathcal{G}_{[i]c}\|\le C|u_{[i]}|^2\mathrm{Im}(m_{1c}+m_{2c}).$$

This completes the proof. □

A.3 Proof of Lemma 3.10 and Lemma 2.2

We first prove Lemma 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 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

$$|u(w)-m_c(w)|\le\frac{C|w|^{1/2}\delta}{\sqrt{\kappa+\eta+\delta}}. \tag{A.62}$$

Case 1: We take over the notations in Definition 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

$$\alpha(u-m_c)^2+\beta(u-m_c)=R, \tag{A.63}$$

where using (A.1), α and β can be expressed as

$$\alpha:=\frac{1}{N}\sum_{i=1}^{}l_i s_i\left(\frac{A_i}{(u-a_i)(m_c-a_i)^2}+\frac{B_i}{(u-b_i)(m_c-b_i)^2}+\frac{C_i}{(u+c_i)(m_c+c_i)^2}\right), \tag{A.64}$$

and

$$\beta:=1-\frac{1}{N}\sum_{i=1}^{}l_i s_i\left(\frac{A_i}{(m_c-a_i)^2}+\frac{B_i}{(m_c-b_i)^2}+\frac{C_i}{(m_c+c_i)^2}\right)=\partial_m f(\sqrt{w},m_c). \tag{A.65}$$

We shall prove that

$$|\alpha|+|\partial_u\alpha|\le C,\quad|\beta|\sim 1, \tag{A.66}$$

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 $\mathrm{Im}\, u\sim 1$. By (3.29), we have

$$\min_i\{|m_c-a_i|,|m_c-b_i|,|m_c+c_i|\}\ge\tau \tag{A.67}$$

for some $\tau>0$. Replacing the m_c in (3.29) with u, we also get that

$$\min_i\{|u-a_i|,|u-b_i|,|u+c_i|\}\ge\tau' \tag{A.68}$$

for some $\tau'>0$. Using (A.67) and (A.68), we get immediately that $|\alpha|+|\partial_u\alpha|+|\beta|\le C$. What remains is the proof of the lower bound $|\beta|\ge c$. If $\mathrm{Im}\, w\ge\epsilon$ for some constant $\epsilon>0$, the lower bound follows from Lemma A.6 below. If $\mathrm{Im}\, w\le\epsilon$ for a sufficiently small $\epsilon>0$, the lower bound follows from Lemma A.7 below. Now given the estimate (A.66), it is easy to prove (A.62) with a fixed point argument. This proves the stability of (3.34).

Lemma A.6. *Suppose that* $\operatorname{Im} w \ll 1$ *and* $|m_c| \sim \operatorname{Im} m_c \sim 1$*. Then* $|\partial_m f(\sqrt{w}, m_c)| \ge c$ *for some constant* $c > 0$*.*

Proof. Using (2.13), $m_c = \sqrt{w}(1 + m_{1c})$ and the conditions $\operatorname{Im} w \ll 1$, $\operatorname{Im} m_c \sim 1$, we can get that

$$\frac{\partial_{\sqrt{w}} f(\sqrt{w}, m_c)}{\partial_m f(\sqrt{w}, m_c)} = -\frac{\partial m_c}{\partial \sqrt{w}} \le C \Rightarrow \left|\partial_{\sqrt{w}} f(\sqrt{w}, m_c)\right| \le C \left|\partial_m f(\sqrt{w}, m_c)\right|, \tag{A.69}$$

for some constant $C > 0$. Now we assume that $|\partial_m f(\sqrt{w}, m_c)|$ can be arbitrarily small. Then $\partial_{\sqrt{w}} f(\sqrt{w}, m_c)$ 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 (A.23) and (A.31), we get that

$$a = \frac{\sqrt{w}}{m_c} - \frac{m_c}{N} \sum_{i=1} l_i s_i \frac{\sqrt{w}\left(m_c^2 - |z|^2\right)^2 + 2 s_i |z|^2 m_c}{\left[-(s_i + |z|^2) m_c^2 + |z|^4 + \sqrt{w}(m_c^3 - |z|^2 m_c)\right]^2} \tag{A.70}$$

and

$$b = -1 - \frac{m_c^2}{N} \sum_{i=1} l_i s_i \frac{\left(m_c^2 - |z|^2\right)^2}{\left[-(s_i + |z|^2) m_c^2 + |z|^4 + \sqrt{w}(m_c^3 - |z|^2 m)\right]^2}. \tag{A.71}$$

Using (A.70) and (A.71), we can get that

$$\frac{(\sqrt{w} m_c - |z|^2)|z|^2}{m_c} b - \frac{1}{2}(m_c^2 - |z|^2)(m_c a - \sqrt{w} b) = \frac{(|z|^2 - \sqrt{w} m_c)(m_c^2 + |z|^2)}{m_c}, \tag{A.72}$$

where we used the equation $f(\sqrt{w}, m_c) = 0$ in the derivation. By our assumption, the left-hand side of (A.72) can be arbitrarily small. For the right-hand side of (A.72), we have $|m_c| \sim 1$ and $|\sqrt{w} m_c - |z|^2| \gtrsim 1$ (since $\operatorname{Im}(\sqrt{w} m_c) = \operatorname{Im}(w + w m_{1c}) \sim 1$). Then if $|m_c - i|z|| \ge c'$ for some constant $c' > 0$, we have $|m^2 + |z|^2| \gtrsim 1$, and hence

$$\left|\frac{(\sqrt{w} m_c - |z|^2)|z|^2}{m_c} b - \frac{1}{2}(m_c^2 - |z|^2)(m_c a - \sqrt{w} b)\right| \gtrsim 1,$$

which gives a contradiction. Thus we must have a lower bound $|\partial_m f(\sqrt{w}, m_c)| \ge c$ if $|m - i|z|| \ge c'$.

We still need to deal with the case with $|m_c - i|z|| \le c'$ for some sufficiently small c'. Notice $|z| \sim 1$ in this case. It is easy to calculate that

$$\frac{\partial f}{\partial \sqrt{w}}(\sqrt{w}, i|z|) = -1 + \frac{|z|^2}{N} \sum_{k=1} l_k s_k \frac{4|z|^4}{\left[(s_k + |z|^2)|z|^2 + |z|^4 - 2i\sqrt{w}|z|^3\right]^2}. \tag{A.73}$$

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 $x, y \ll 1$, 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 (A.73) has positive imaginary part of order 1. Therefore

$$\left|\frac{\partial f}{\partial \sqrt{w}}(\sqrt{w}, i|z|)\right| \ge \operatorname{Im}\, \frac{\partial f}{\partial \sqrt{w}}(\sqrt{w}, i|z|) \gtrsim 1.$$

Then by (A.69), we get that $|\partial_m f(\sqrt{w}, i|z|)| \ge c$ for some $c > 0$. Using (3.29), 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 $|m_c - i|z|| \le c'$, we still have $|\partial_m f(\sqrt{w}, m_c)| \ge c/2$, provided that c' is sufficiently small. □

Lemma A.7. *Suppose that $w \in \mathrm{D}_k^b$ and $\mathrm{Im}\, w \le \eta$. Then for sufficiently small $\eta > 0$, we have $|\partial_m f(\sqrt{w}; m_c)| \gtrsim 1$.*

Proof. By (3.22) and (3.29), 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

$$0 = f(\sqrt{E}; m_c(E)) = -\sqrt{E} + m_c + E^{1/2} + \frac{1}{N}\sum_{i=1} l_i s_i \left(\frac{A_i}{m_c - a_i} + \frac{B_i}{m_c - b_i} + \frac{C_i}{m_c + c_i}\right), \tag{A.74}$$

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

$$\frac{1}{N}\sum_{i=1} l_i s_i \left(\frac{A_i}{|m_c - a_i|^2} + \frac{B_i}{|m_c - b_i|^2} + \frac{C_i}{|m_c + c_i|^2}\right) = 1. \tag{A.75}$$

Using the above equation, we get

$$\begin{aligned}\partial_m f(\sqrt{E}; m_c(E)) &= 1 - \frac{1}{N}\sum_{i=1} l_i s_i \left(\frac{A_i}{(m_c - a_i)^2} + \frac{B_i}{(m_c - b_i)^2} + \frac{C_i}{(m_c + c_i)^2}\right)\\ &= \frac{1}{N}\sum_{i=1} l_i s_i \left(\frac{A_i}{|m_c - a_i|^2} - \frac{A_i}{(m_c - a_i)^2} + \frac{B_i}{|m_c - b_i|^2} - \frac{B_i}{(m_c - b_i)^2} + \frac{C_i}{|m_c + c_i|^2} - \frac{C_i}{(m_c + c_i)^2}\right).\end{aligned} \tag{A.76}$$

We look at, for example, the term

$$\frac{A_i}{|m_c - a_i|^2} - \frac{A_i}{(m_c - a_i)^2} = \frac{A_i}{|m_c - a_i|^2}\left(1 - e^{-2i\theta_i}\right),$$

where $m_c - a_i := |m_c - a_i| e^{i\theta_i}$. Using $\mathrm{Im}\, m_c \gtrsim 1$, it is easy to see that $\mathrm{Re}(1 - e^{-2i\theta_i}) \ge c'$ for some constant $c' > 0$. Applying the same estimates to the B, C terms in (A.76), we get

$$|\partial_m f(\sqrt{E}; m_c(E))| \ge \mathrm{Re}\left[\partial_m f(\sqrt{E}; m_c(E))\right] \ge c \tag{A.77}$$

for some constant $c > 0$.

Now for $w = E + i\eta$ with $\eta \le \ldots$, 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 (3.29). Combing with (A.77), we get $|\partial_m f(w; m_c(w))| \gtrsim 1$ for small enough η. □

Case 2: We mimic the argument in the proof of Case 1. We see that it suffices to prove $|\alpha| + |\partial_u \beta| \le C$ and $|\beta| \gtrsim 1$ for α, β defined in (A.64) and (A.65) and $|u - m_c| \le (\log N)^{-1/3}$. Using (3.29), 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 $\mathrm{Im}\, w \gtrsim 1$ case, the bound follows from Lemma A.6. We are left with the case where $E = \mathrm{Re}\, w \gtrsim 1$ and $\eta = \mathrm{Im}\, w \to 0$. Using (2.13), $m_c = \sqrt{w}(1 + m_{1c})$, $|w| \gtrsim 1$ and $\mathrm{dist}(E, \mathrm{supp}\, \rho_{1c}) \ge \kappa'$, we can get that

$$\left|\frac{\partial_{\sqrt{w}} f(\sqrt{w}; m_c)}{\partial_m f(\sqrt{w}; m_c)}\right| = \left|\frac{\partial m_c}{\partial \sqrt{w}}\right| \le C$$

for some constant $C > 0$. Thus it suffices to prove that $\partial_{\sqrt{w}} f(\sqrt{w}; m_c)$ has a lower bound. Using (A.31) and noticing that $m_c(E) \in \mathbb{R}$, we get

$$\partial_{\sqrt{w}} f(\sqrt{E}; m_c(E)) = -1 - \frac{m_c^2}{N}\sum_{i=1} l_i s_i \frac{\left(m_c^2 - |z|^2\right)^2}{\left[(s_i + |z|^2)m_c^2 + |z|^4 + \sqrt{E}\,(m_c^3 - |z|^2 m_c)\right]^2} \le -1.$$

Expanding $\partial_{\sqrt{w}} f(\sqrt{w}; m_c(w))$ around $\partial_{\sqrt{w}} f(\sqrt{E}; m_c(E))$, using (3.29) and $|m_c(E+i\ ) \ m_c(E)|$, we get for small

$$\partial_{\sqrt{w}} f(\sqrt{w}; m_c) \quad 1 + O(\ ) \quad c.$$

This concludes the proof for *Case 2*.

Case 3: The case $\operatorname{Im} w \quad {}'$ can be proved with the same method as in the proof of case 1. Hence we only consider the case $|w \quad e_k| \quad 2\ '$ in the following. Note that $|w| \quad 1$ in this case. Suppose

$$|w \quad e_k| \quad 2\ '; \quad |u \quad m_c| \quad (\log N)^{\ 1/3}. \tag{A.78}$$

Then we claim that

$$|\ | \quad 1; \quad |\ | \quad \sqrt{\ +\ } \tag{A.79}$$

for small enough $'$. Using (A.78), (3.29), (2.19) and Lemma A.5, we can get that

$$= \frac{1}{2}\partial_m^2 f(\sqrt{e_k}; m_c(e_k)) + O(|w \quad e_k|^{1/2} + (\log N)^{\ 1/3}) \quad 1.$$

To prove the estimate for , we use (2.17), (3.29) and Lemma A.5 to get that

$$\begin{aligned}
&= \int_{e_k}^{w} \frac{d}{dw'} \partial_m f(\sqrt{w'}; m_c(w'))dw' \\
&= \int_{e_k}^{w} \frac{\partial_{\sqrt{w'}}\partial_m f(\sqrt{w'}; m_c(w'))}{2\sqrt{w'}} dw' + \int_{e_k}^{w} \partial_m^2 f(\sqrt{w'}; m_c(w')) \frac{dm_c(w')}{dw'} dw' \\
&= \int_{e_k}^{w} \frac{\partial_{\sqrt{w}}\partial_m f(\sqrt{e_k}; m_c(e_k)) + O(|w \quad e_k|^{1/2})}{2\sqrt{w'}} dw' \\
&\quad + \int_{m_c(e_k)}^{m_c(w)} \Big[\partial_m^2 f(\sqrt{e_k}; m_c(e_k)) + O(|w \quad e_k|^{1/2}) \Big] dm \\
&= \partial_m^2 f(\sqrt{e_k}; m_k)(m_c(w) \quad m_c(e_k)) + O(|w \quad e_k|).
\end{aligned} \tag{A.80}$$

Thus we conclude for small enough $'$ that

$$|\ | \quad |w \quad e_k|^{1/2} \quad \sqrt{\ +\ }.$$

With the estimate (A.79), we now proceed as in the proof of [4, Lemma 4.5], by solving the quadratic equation (A.63) for $u \quad m_c$ explicitly. We select the correct solution by a continuity argument using that (A.62) holds by assumption at $z + iN^{\ 10}$. The second assumption of (A.78) is obtained by continuity from the estimate on $|u \quad m_c|$ at the neighboring point $z + iN^{\ 10}$. We refer to [4, Lemma 4.5] for the full details. This concludes the proof for *Case 3*.

Case 4: The case when $\operatorname{Im} w \quad {}'$ can be proved using the same method as in the proof of Case 1. Now we are left with the case $|w| \quad 2\ '$ for some sufficiently small $'$. First we assume $|z| \quad c > 0$ for some small $c > 0$. Then mimicking the argument in the proof of Case 1, we see that it suffices to prove $|\ | + |\partial_u \ | \quad C$ and $|\ | \quad 1$ when $|u \quad m_c| \quad (\log N)^{\ 1/3}$. Using (3.29), it is not hard to prove that $|\ | + |\partial_u \ | + |\ | \quad C$. The lower bound $|\ | \quad c$ can be obtained easily from (A.32).

Then suppose $|z|^2 < c$, but $|w|^{1/2} + |z|^2$. According to (A.38) and using that $i|z|^2 + \sqrt{wt_0} \quad |w|^{1/2} + |z|^2$, we can verify that

$$= \partial_m f(\sqrt{w}; m_c(w)) \quad |w|^{1/2} + |z|^2 \quad 1.$$

With (3.29), it is easy to check that

$$\partial_m^2 f(\sqrt{w};\ ) = O(1); \quad \partial_m^3 f(\sqrt{w};\ ) = O(1);$$

for $| \quad m_c| \quad (\log N)^{\ 1/3}$, from which we get that $| \ | + |\partial_u \ | = O(1)$. With a fixed point argument, we conclude (A.62).

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 (3.35) directly. We take over the notations in Definition 3.9 and abbreviate $R := \Upsilon(w; u_1)$, so that $|R| \quad$. Then we write the equation $\Upsilon(w; u_1) \quad \Upsilon(w; m_{1c}) = R$ as

$$(u_1)(u_1 \quad m_{1c})^2 + \quad (u_1 \quad m_{1c}) = R; \tag{A.81}$$

where we used the same symbols as in (A.63) for notational convenience. As in *Case 1*, we have $\ = \partial_{m_1}\Upsilon(w; m_{1c})$; and we can estimate that $| \ | + |\partial_{u_1} \ | \quad C$ for $w \in \mathbf{D}_L$ and u_1 satisfying $|u_1 \quad m_{1c}| \quad |m_{1c}|$. Now to conclude (3.39), it suffices to prove $| \ | \quad 1$ for $w \in \mathbf{D}_L$. In fact with (3.35), we can obtain that

$$= 1 + O \quad ^{1} \quad 1;$$

for $\quad ^{1}$. This concludes the proof.

Proof of Lemma 2.2. The fact that $\ _{1c}$ has compact support follows from Lemma 2.3; $\ _{1c}$ being integrable follows from Lemma A.4. Note that in proving Lemmas 2.3 and A.4, we do not make use of the regularity assumptions in Definition 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 (2.11). This follows from the proof of *Case 1* in this section under the extra condition $\quad 1$. Again, we do not need the regularity assumptions for the proof, because $\ ^{1}$ provides a nice bound for the Stieltjes transforms in the global region with $\quad 1$. □

Remark: The estimate (3.29) has been used repeatedly during the proof of Lemma 3.10. Here we remark that it also gives the stability of the regularity conditions in Definition 2.4 under perturbations of $|z|$ and $\quad$. For example, we define the shifted empirical spectral density

$$_{;t} := \frac{1}{N \wedge M} \sum_{i=1}^{N \wedge M} \ _{i + t}; \tag{A.82}$$

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 (A.31), we have $\partial_{\sqrt{w}} f(\sqrt{e_k}; m_k; 0) \quad 1$. Combining with (2.19), we get $|J| \quad$. Using (3.29), 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

$$\partial_m^2 f \left(\sqrt{e_k(t)}; m_c(e_k; t)\right) = \partial_m^2 f(\sqrt{e_k}; m_c(e_k)) + O(t);$$

$$|m_c(e_k; t) \quad a_i(e_k; t)| = |m_c(e_k) \quad a_i(e_k)| + O(t);$$

and similar estimates for $|m_c \quad b_i|$ and $|m_c + c_i|$. Thus if Definition 2.4 (i) holds for some $\ $, then it holds for all $\ _{;t}$ 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 (3.29), we find that $\partial_t f(\sqrt{E}, m_c(E), 0) = O(1)$, while by (A.66), $|\partial_m f(\sqrt{E}, m_c(E), 0)| \gtrsim 1$. 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 $\mathrm{Im}\, m_c(E,t)$ is bounded from below by some $c' = c'(\tau, \tau')$. Thus we conclude that if Definition 2.4 (ii) holds for some Σ, then it holds for all $\Sigma_{\cdot,t}$ with t in some fixed small interval around zero. Obviously, the above arguments also work for $|z|$ perturbation.

B Proof of Lemma 4.9

Our proof of (4.59) is an extension of [4, Lemma 4.9], [7, Lemma 7.3] and [14, 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 $Z_{[i]} = Q_i G^{-1}_{[ii]}$. Hence we need to prove

$$\|[Z]\| = \left\| \frac{1}{N} \sum_{i=1}^{N} \Pi_{[i]} \left(Q_i G^{-1}_{[ii]}\right) \Pi_{[i]} \right\| \prec |w|^{-1/2} \Phi_o^2,$$

for $w \in \mathbf{D}$. For $J \subset \mathcal{I}$, we define $\Pi^{[J]}_{[i]}$ by replacing $m_{1,2}$ in (2.36) with $m^{[J]}_{1,2}$ defined in (4.6). As in (4.58), we can prove that $|m^{[i]}_{1,2} - m_{1,2}| \prec |w|^{-1/2} \Phi_o^2$, which further gives that

$$\begin{aligned} \|[Z]\| &= \left\| \frac{1}{N} \sum_{i=1}^{N} \Pi^{[i]}_{[i]} \left(Q_i G^{-1}_{[ii]}\right) \Pi^{[i]}_{[i]} \right\| + O_\prec\left(|w|^{-1/2} \Phi_o^2\right) \\ &= \left\| \frac{1}{N} \sum_{i=1}^{N} Q_i \left(\Pi^{[i]}_{[i]} G^{-1}_{[ii]} \Pi^{[i]}_{[i]}\right) \right\| + O_\prec\left(|w|^{-1/2} \Phi_o^2\right). \end{aligned}$$

Thus if we abbreviate $B_i := |w|^{1/2} Q_i \left(\Pi^{[i]}_{[i]} G^{-1}_{[ii]} \Pi^{[i]}_{[i]}\right)$, 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

$$\mathrm{Tr}(BB^\dagger)^n \ge \left\|\left(BB^\dagger\right)^n\right\| = \|B\|^{2n}.$$

Thus it suffices to prove that

$$\mathbb{E}\,\mathrm{Tr}(BB^\dagger)^{p/2} \prec \Phi_o^{2p}, \quad \text{for } p = 2n. \tag{B.1}$$

This estimate can be proved with the same method as in [14, Appendix B], with the only complication being that $\Pi_{[i]}$ is random and depends on i. In principle, this can be handle by using (3.9) and (3.10) to put any indices $j, k, \ldots \in \mathcal{I}_1$ (that we wish to include) into the superscripts of $\Pi_{[i]}$. This leads to a minor modification of the proof in [14, 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 $\tilde{B}_{k_s} := B^\dagger_{k_s}$

for even $s \quad p$. Then we get

$$\mathrm{E}\mathrm{Tr}(BB^{y})^{p/2} = \frac{1}{N^p} \sum_{k_1,k_2,\dots,k_p} \mathrm{E}\mathrm{Tr} \prod_{s=1} \tilde{B}_{k_s} = \frac{1}{N^p} \sum_{k_1,k_2,\dots,k_p} \mathrm{E}\mathrm{Tr} \prod_{s=1} \left(\prod_{r=1} (P_{k_r} + Q_{k_r}) \tilde{B}_{k_s} \right) . \tag{B.2}$$

Introducing the notations $\mathbf{k} = (k_1, k_2, \dots, k_p)$ and $f\mathbf{k}g = fk_1, k_2, \dots, k_pg$, we can write

$$\mathrm{E}\mathrm{Tr}(BB^{y})^{p/2} = \frac{1}{N^p} \sum_{\mathbf{k}} \sum_{I_1,\dots,I_p \quad f\mathbf{k}g} \mathrm{E}\mathrm{Tr} \prod_{s=1} P_{I_s^c} Q_{I_s} \tilde{B}_{k_s} . \tag{B.3}$$

Following the arguments in [14, Appendix B], one can see that to conclude (B.1) it suffices to prove that for $k \in I$,

$$\| Q_I B_k \| \quad \Phi_o^{|I|}. \tag{B.4}$$

As in [14, Appendix B], it is not hard to prove that for $k \in I$,

$$|w|^{1/2} \quad Q_I G_{[kk]}^{1} \quad \Phi_o^{|I|}; \quad \text{and} \quad |w|^{1/2} \quad Q_{I \setminus fkg} G_{[kk]}^{1} \quad \Phi_o^{|I|} \text{ if } |I| \quad 2. \tag{B.5}$$

Now we extend the proof to obtain the estimate (B.4). For the case $|I| = 1$ (i.e. $I = fkg$),

$$\| B_k \| = |w|^{1/2} \| \,{}^{[i]}_{[i]} Z_{[k]} \,{}^{[i]}_{[i]} \| \quad |w|^{1/2} \| Z_{[k]} \| \quad \Phi_o,$$

where we used $\| Z_{[k]} \| \quad |w|^{1/2} \Phi_o$, which can be proved with the same arguments as in Lemma 4.4. For the case $|I| \quad 2$, WLOG, we may assume $k = 1$ and $I = f1, \dots, tg$ with $t \quad 2$. It is enough to prove that

$$|w|^{1/2} \quad Q_t \cdots Q_2 Q_1 \,{}^{[1]}_{[1]} G_{[11]}^{1} \,{}^{[1]}_{[1]} \quad \Phi_o^{t}. \tag{B.6}$$

We take $t = 3$ as an example to describe the ideas for the proof of (B.6). Using (3.9), we get

$$ {}^{[1]}_{[1]} = {}^{[12]}_{[1]} + |w|^{1/2} \,{}^{[1]}_{11} \,{}^{[12]}_{[1]} A_1 \,{}^{[12]}_{[1]} + |w|^{1/2} \,{}^{[1]}_{11} \,{}^{[12]}_{[1]} A_2 \,{}^{[12]}_{[1]} + \mathrm{error}_{1,2}, \tag{B.7}$$

where ${}^{[1]}_{11}$ and ${}^{[1]}_{11}$ are the upper left and lower right entries of

$$ {}^{[1]}_{[1]} := |w|^{1/2} \left(\frac{G_{[22]}^{[1]}}{N} + \frac{1}{N} \sum_{k \notin f1,2g} G_{[k2]}^{[1]} \quad G_{[22]}^{[1]} \quad ^{1} G_{[2k]}^{[1]} \right) \quad \Phi_o^2,$$

$A_{1,2}$ are deterministic matrices with operator norm $O(1)$, and $\| \mathrm{error}_{1,2} \| \quad |w| \quad ^{1/2} \Phi_o^4$. Then we get

$$\begin{aligned} {}^{[1]}_{[1]} G_{[11]}^{1} \,{}^{[1]}_{[1]} = {} & {}^{[12]}_{[1]} G_{[11]}^{1} \,{}^{[12]}_{[1]} + |w|^{1/2} \,{}^{[1]}_{11} \,{}^{[12]}_{[1]} A_1 \,{}^{[12]}_{[1]} G_{[11]}^{1} \,{}^{[12]}_{[1]} + |w|^{1/2} \,{}^{[1]}_{11} \,{}^{[12]}_{[1]} A_2 \,{}^{[12]}_{[1]} G_{[11]}^{1} \,{}^{[12]}_{[1]} \\ & + |w|^{1/2} \,{}^{[12]}_{[1]} G_{[11]}^{1} \,{}^{[1]}_{11} \,{}^{[12]}_{[1]} A_1 \,{}^{[12]}_{[1]} + |w|^{1/2} \,{}^{[12]}_{[1]} G_{[11]}^{1} \,{}^{[1]}_{11} \,{}^{[12]}_{[1]} A_2 \,{}^{[12]}_{[1]} + O \quad (|w| \quad ^{1/2} \Phi_o^4). \end{aligned} \tag{B.8}$$

We first handle the ${}^{[12]}_{[1]} G_{[11]}^{1} \,{}^{[12]}_{[1]}$ term. By (B.5), we have

$$Q_2 \,{}^{[12]}_{[1]} G_{[11]}^{1} \,{}^{[12]}_{[1]} = {}^{[12]}_{[1]} \quad Q_2 G_{[11]}^{1} \quad {}^{[12]}_{[1]} \quad |w| \quad ^{1/2} \Phi_o^2.$$

For the remaining term, we first expand ${}^{[12]}_{[1]} = {}^{[123]}_{[1]} + O \quad (|w| \quad ^{1/2} \Phi_o^2)$ and use (B.5) to get

$$Q_3 Q_2 \,{}^{[12]}_{[1]} G_{[11]}^{1} \,{}^{[12]}_{[1]} = {}^{[123]}_{[1]} \quad Q_3 Q_2 G_{[11]}^{1} \quad {}^{[123]}_{[1]} + O \quad |w| \quad ^{1/2} \Phi_o^4 \quad |w| \quad ^{1/2} \Phi_o^3.$$

Then we deal with the second terms in (B.8). We first expand $\mathcal{E}^{[1]}_{[1]} = \mathcal{E}^{[3]}_{[1]} + O_\prec(\Phi_o^3)$, where

$$\mathcal{E}^{[3]}_{[1]} := |w|^{1/2}\left(\frac{G^{[13]}_{[22]}}{N} + \frac{1}{N}\sum_{k\in\{1,2,3\}} G^{[13]}_{[k2]}\left(G^{[13]}_{[22]}\right)^{-1} G^{[13]}_{[2k]}\right).$$

Using the similar arguments as above, we get

$$Q_3|w|^{1/2}\mathcal{E}^{[3]}_{11}\,{}^{[12]}_{[1]}A_1\,{}^{[12]}_{[1]}G^{-1}_{[11]}\,{}^{[12]}_{[1]} = |w|^{1/2}\mathcal{E}^{[3]}_{11}\,{}^{[123]}_{[1]}A_1\,{}^{[123]}_{[1]}\,Q_3G^{-1}_{[11]}\,{}^{[123]}_{[1]} + O_\prec(|w|^{-1/2}\Phi_o^4)$$
$$\prec |w|^{-1/2}\Phi_o^4.$$

Thus we have

$$Q_2Q_3|w|^{1/2}\,{}^{[1]}_{11}\,{}^{[12]}_{[1]}A_1\,{}^{[12]}_{[1]}G^{-1}_{[11]}\,{}^{[12]}_{[1]} \prec |w|^{-1/2}\Phi_o^3.$$

Obviously this kind of estimate works for the rest of the terms in (B.8). This proves (B.6) 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 (3.9) and (3.10), and after applying $Q_l\cdots Q_3Q_2$ on them, the number of Φ_o factors increases by one at each step by (B.5). Trough induction we can prove (B.6). In fact the expansions can be performed in a systematic way using the method in [14, 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 5, and we choose to present the details there (in fact the proof here is much easier than the one in Section 5).

References

[1] Z. D. Bai. Circular law. *Ann. Probab.*, 25(1):494–529, 1997. MR-1428519

[2] Z. D. Bai and J. W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*, volume 2 of *Mathematics Monograph Series*. Science Press, Beijing, 2006. MR-2567175

[3] Z. Bao, G. Pan, and W. Zhou. Universality for the largest eigenvalue of sample covariance matrices with general population. *Ann. Statist.*, 43(1):382–421, 2015. MR-3311864

[4] A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.*, 19(33):1–53, 2014. MR-3183577

[5] A. Bloemendal, A. Knowles, H.-T. Yau, and J. Yin. On the principal components of sample covariance matrices. *Prob. Theor. Rel. Fields*, 164(1):459–552, 2016. MR-3449395

[6] A. Borodin and C. D. Sinclair. The Ginibre ensemble of real random matrices and its scaling limits. *Commun. Math. Phys.*, 291(1):177–224, 2009. MR-2530159

[7] P. Bourgade, H.-T. Yau, and J. Yin. Local circular law for random matrices. *Probab. Theory Relat. Fields*, 159:545–595, 2014. MR-3230002

[8] P. Bourgade, H.-T. Yau, and J. Yin. The local circular law II: the edge case. *Probab. Theory Relat. Fields*, 159(3):619–660, 2014. MR-3230004

[9] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and banach spaces. volume 1 of *Handbook of the Geometry of Banach Spaces*, pages 317–366. North-Holland, Amsterdam, 2001. MR-1863696

[10] A. Edelman. The probability that a random real gaussian matrix has k real eigenvalues, related distributions, and the circular law. *J. Multivar. Anal.*, 60(2):203–232, 1997. MR-1437734

[11] N. El Karoui. Tracy-widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *Ann. Probab.*, 35(2):663–714, 2007. MR-2308592

[12] L. Erdős, A. Knowles, and H.-T. Yau. Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré*, 14:1837–1926, 2013. MR-3119922

[13] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Delocalization and diffusion profile for random band matrices. *Commun. Math. Phys.*, 323:367–416, 2013. MR-3085669

[14] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. The local semicircle law for a general class of random matrices. *Electron. J. Probab.*, 18:1–58, 2013. MR-3068390

[15] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Spectral statistics of Erdős-Rényi graphs I: Local semicircle law. *Ann. Probab.*, 41(3B):2279–2375, 2013. MR-3098073

[16] L. Erdős, B. Schlein, and H.-T. Yau. Local semicircle law and complete delocalization for Wigner random matrices. *Commun. Math. Phys.*, 287(2):641–655, 2008. MR-2481753

[17] L. Erdős, H.-T. Yau, and J. Yin. Bulk universality for generalized Wigner matrices. *Probab. Theory Relat. Fields*, 154(1):341–407, 2012. MR-2981427

[18] P. J. Forrester and T. Nagao. Eigenvalue statistics of the real Ginibre ensemble. *Phys. Rev. Lett.*, 99:050603, 2007.

[19] J. Ginibre. Statistical ensembles of complex, quaternion, and real matrices. *J. Math. Phys.*, 6(3):440–449, 1965. MR-0173726

[20] V. Girko. The circular law. *Russ. Teor. Veroyatnost. i Primenen.*, 29(4):669–679, 1984. MR-0773436

[21] F. Gőtze and A. Tikhomirov. The circular law for random matrices. *Ann. Probab.*, 38(4):1444–1491, 2010. MR-2663633

[22] A. Guionnet, M. Krishnapur, and O. Zeitouni. The single ring theorem. *Ann. Math.*, 174(2):1189–1217, 2011. MR-2831116

[23] W. Hachem, A. Hardy, and J. Najim. Large complex correlated Wishart matrices: Fluctuations and asymptotic independence at the edges. arXiv:1409.7548.

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

[25] A. Knowles and J. Yin. The isotropic semicircle law and deformation of Wigner matrices. *Comm. Pure Appl. Math.*, 66(11):1663–1749, 2013. MR-3103909

[26] J. O. Lee and K. Schnelli. Tracy-widom distribution for the largest eigenvalue of real sample covariance matrices with general population. arXiv:1409.4979.

[27] A. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann. Smallest singular value of random matrices and geometry of random polytopes. *Adv. Math.*, 195(2):491–523, 2005. MR-2146352

[28] M. L. Mehta. *Random matrices*, volume 142 of *Pure and Applied Mathematics*. Elsevier, Amsterdam, 3 edition, 2004. MR-2129906

[29] A. Onatski. The tracy-widom limit for the largest eigenvalues of singular complex wishart matrices. *Ann. Appl. Probab.*, 18(2):470–490, 2008. MR-2398763

[30] G. Pan and W. Zhou. Circular law, extreme singular values and potential theory. *J. Multivar. Anal.*, 101(3):645–656, 2010. MR-2575411

[31] M. Rudelson. Invertibility of random matrices: norm of the inverse. *Ann. Math.*, 168(2):575–600, 2008. MR-2434885

[32] M. Rudelson and R. Vershynin. The littlewood-offord problem and invertibility of random matrices. *Adv. Math.*, 218:600–633, 2008. MR-2407948

[33] M. Rudelson and R. Vershynin. The smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math.*, 62:1707–1739, 2009. MR-2569075

[34] M. Rudelson and R. Vershynin. Small ball probabilities for linear images of high-dimensional distributions. *Int. Math. Res. Notices*, 2015(19):9594–9617, 2015. MR-3431603

[35] C. D. Sinclair. Averages over Ginibre's ensemble of random real matrices. *Int. Math. Res. Not.*, 2007:1–15. MR-2341601

[36] T. Tao and V. Vu. Random matrices: the circular law. *Commun. Contemp. Math.*, 10(2):261–307, 2008. MR-2409368

[37] T. Tao and V. Vu. Random matrices: Universality of local spectral statistics of non-Hermitian matrices. *Ann. Probab.*, 43(2):782–874, 2015. MR-3306005

[38] T. Tao, V. Vu, and M. Krishnapur. Random matrices: Universality of ESDs and the circular law. *Ann. Probab.*, 38(5):2023–2065, 2010. MR-2722794

[39] F. Yang. The smallest singular value of deformed random rectangular matrices. arXiv:1702.04050.

[40] J. Yin. The local circular law III: general case. *Probab. Theory Relat. Fields*, 160(3):679–732, 2014. MR-3278919

Acknowledgments. The third author would like to thank Terence Tao, Mark Rudelson and Roman Vershynin for fruitful discussions and valuable suggestions.