Rigidity of Eigenvalues of Generalized Wigner Matrices

Citation
Erdos, László, Horng-Tzer Yau, and Jun Yin. 2012. "Rigidity of Eigenvalues of Generalized Wigner Matrices." Advances in Mathematics 229 (3) (February): 1435–1515. doi:10.1016/j.aim.2011.12.010.

Published Version
doi://10.1016/j.aim.2011.12.010

Rigidity of Eigenvalues of Generalized Wigner Matrices

László Erdős[1]*, Horng-Tzer Yau[2†] and Jun Yin[2‡]

Institute of Mathematics, University of Munich,
Theresienstr. 39, D-80333 Munich, Germany
lerdos@math.lmu.de [1]

Department of Mathematics, Harvard University
Cambridge MA 02138, USA
htyau@math.harvard.edu, jyin@math.harvard.edu [2]

Oct 25, 2011

Abstract

Consider $N \times N$ hermitian or symmetric random matrices H with independent entries, where the distribution of the (i, j) matrix element is given by the probability measure ν_{ij} with zero expectation and with variance σ_{ij}^2. We assume that the variances satisfy the normalization condition $\sum_i \sigma_{ij}^2 = 1$ for all j and that there is a positive constant c such that $c \le N\sigma_{ij}^2 \le c^{-1}$. We further assume that the probability distributions ν_{ij} have a uniform subexponential decay. We prove that the Stieltjes transform of the empirical eigenvalue distribution of H is given by the Wigner semicircle law uniformly up to the edges of the spectrum with an error of order $(N\eta)^{-1}$ where η is the imaginary part of the spectral parameter in the Stieltjes transform. There are three corollaries to this strong local semicircle law: (1) Rigidity of eigenvalues: If $\gamma_j = \gamma_{j,N}$ denotes the *classical location* of the j-th eigenvalue under the semicircle law ordered in increasing order, then the j-th eigenvalue λ_j is close to γ_j in the sense that for some positive constants C, c

$$\mathbb{P}\Big(\exists\, j:\ |\lambda_j - \gamma_j| \ge (\log N)^{C\log\log N}\Big[\min\big(j, N-j+1\big)\Big]^{-1/3} N^{-2/3}\Big) \le C\exp\big[-(\log N)^{c\log\log N}\big]$$

for N large enough. (2) The proof of *Dyson's conjecture* [15] which states that the time scale of the Dyson Brownian motion to reach local equilibrium is of order N^{-1} up to logarithmic corrections. (3) The edge universality holds in the sense that the probability distributions of the largest (and the smallest) eigenvalues of two generalized Wigner ensembles are the same in the large N limit provided that the second moments of the two ensembles are identical.

AMS Subject Classification (2010): 15B52, 82B44

Keywords: Random matrix, Local semicircle law, Tracy-Widom distribution, Dyson Brownian motion.

*Partially supported by SFB-TR 12 Grant of the German Research Council
†Partially supported by NSF grants DMS-0757425, 0804279
‡Partially supported by NSF grants DMS-1001655

1 Introduction

Random matrices were introduced by E. Wigner to model the excitation spectrum of large nuclei. The central idea is based on the observation that the eigenvalue gap distribution for a large complicated system is universal in the sense that it depends only on the symmetry class of the physical system but not on other detailed structures. As a special case of this general belief, the eigenvalue gap distribution of random matrices should be independent of the probability distributions of the ensembles and thus is given by the classical Gaussian ensembles. Besides the eigenvalue gap distribution, similar predictions hold also for short distance correlation functions of the eigenvalues. Since the gap distribution can be expressed in terms of correlation functions, mathematical analysis is usually performed on correlation functions. From now on, we refer to *universality* for the fact that the short distance behavior of the eigenvalue correlation functions of a random matrix ensemble are the same as those of the Gaussian ensemble of the same symmetry class (Gaussian unitary, orthogonal or symplectic ensemble, i.e., GUE, GOE, GSE).

The universality question can be roughly divided into the bulk universality in the interior of the spectrum and the edge universality near the spectral edges. Over the past two decades, spectacular progress on bulk and edge universality was made for invariant ensembles, see, e.g., [8, 12, 13, 30] and [2, 10, 11] for a review. For non-invariant ensembles with i.i.d. matrix elements (*Standard Wigner ensembles*) edge universality can be proved via the moment method and its various generalizations, see, e.g., [34, 37, 35]. In a striking contrast, the only rigorous results for the bulk universality of non-invariant Wigner ensembles were the work by Johansson [26] and subsequent improvements [6, 27] on *Gaussian divisible Hermitian ensembles*, i.e., Hermitian ensembles of the form

$$H_s = H_0 + sV, \tag{1.1}$$

where H_0 is a Wigner matrix, V is an independent standard GUE matrix and s is a fixed positive constant independent of N. The Hermitian assumption is essential since the key formula used in [26] and the earlier work [9] is valid only for *Hermitian ensembles*.

The bulk universality, however, was expected to hold for general classes of Wigner matrices, see Mehta's book [28], Conjectures 1.2.1 and 1.2.2 on page 7. We will refer to these two conjectures as the Wigner-Dyson-Gaudin-Mehta conjecture due to their pioneering work. Until a few years ago this conjecture remained unsolved, mainly due to the fact that all existing methods on local eigenvalue statistics depended on explicit formulas which were not available for Wigner matrices. In a series of papers [16, 17, 18, 20, 19, 21, 22, 23], we developed a new approach to understand local eigenvalue statistics. This approach, in particular, led to the first proof [20] of the Wigner-Dyson-Gaudin-Mehta conjecture for Hermitian Wigner matrices with smooth distributions for the matrix elements. We now give a brief summary of this approach which motivates the current paper.

The first step was to derive a *local semicircle law*, a precise estimate of the local eigenvalue density, down to energy scales containing around N^ε eigenvalues. In fact, we also obtain precise bounds on the matrix elements of the Green function. The second step is a general approach for the universality of Gaussian divisible ensembles by embedding the matrix (1.1) into a stochastic flow of matrices and use that the eigenvalues evolve according to a distinguished coupled system of stochastic differential equations, called the Dyson Brownian motion [15]. The central idea is to estimate the time to local equilibrium for the Dyson Brownian motion with the introduction of a new stochastic flow, the *local relaxation flow*, which locally behaves like a Dyson Brownian motion but has a faster decay to global equilibrium. This approach [19, 21] entirely eliminates the usage of explicit formulas and it provides a unified proof for the universality of Gaussian divisible ensembles for all symmetry classes. Furthermore, it also gives a conceptual interpretation that the origin of the universality is due to the local ergodicity of Dyson Brownian motion.

More precisely, we will use a slightly different version of (1.1), namely

$$H_t = e^{-t/2} H_0 + (1 - e^{-t})^{1/2} V, \tag{1.2}$$

to ensure that the variance of H_t remains independent of t. Denote by λ_j the j-th eigenvalue of the random matrix H_t, labelled in increasing order, $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$, and γ_j the *classical location* of the j-th eigenvalue, i.e., γ_j is defined by

$$N \int_{-\infty}^{\gamma_j} \varrho_{sc}(x) \mathrm{d}x = j, \qquad 1 \leq j \leq N, \tag{1.3}$$

where $\varrho_{sc}(x) = \frac{1}{2\pi}\sqrt{(4-x^2)_+}$ is the semicircle law. Our main result on the universality for the Dyson Brownian motion states that, roughly speaking, the short distance correlation functions for H_t at the time $t \sim N^{-2\mathfrak{a}}$ and $H_{t=\infty}$ are identical in weak sense provided that the following main condition holds:

Assumption III. There exists an $\mathfrak{a} > 0$ such that

$$\sup_{t \geq N^{-2\mathfrak{a}}} \frac{1}{N} \mathbb{E}_t \sum_{j=1}^{N} (\lambda_j - \gamma_j)^2 \leq C N^{-1-2\mathfrak{a}} \tag{1.4}$$

with a constant C uniformly in N. Here $\mathbb{E}_t$ is the expectation w.r.t. Dyson Brownian motion at the time t. The condition (1.4) has been derived from a sufficiently strong version of the local semicircle law.

Once the universality for the Gaussian divisible ensemble is established, the last step is to approximate all matrix ensembles by Gaussian divisible ones. This step can be done via a *reverse heat flow* argument [20, 21] for ensembles with smooth probability distributions or more generally via the *Green function comparison theorem* [22] which compares the distributions of eigenvalues of two ensembles around a fixed energy. The key input for the latter approach was to prove a-priori estimates on the matrix elements of the Green function. These estimates have been obtained together with the local semicircle law.

To summarize, our approach to universality consists of the following three main steps: *Step 1. Local semicircle law. Step 2. Universality for Gaussian divisible ensembles. Step 3. Approximation by Gaussian divisible ensembles.* Both Step 2 and 3 rely on a strong local semicircle law from the Step 1.

Shortly after the preprint [20] appeared, another method for the universality was posted by Tao and Vu [40]. This method contains similar three ingredients as in [20]; their key result, prior to the Green function comparison theorem appeared in [22], states that the probability distributions of the j-th eigenvalue of two ensembles for a fixed label j in the bulk are identical as $N \to \infty$ provided that the first four moments of the matrix elements of the two ensembles are identical. This result also implies the universality of the correlation functions for Hermitian Wigner ensembles [40] by combining it with the Gaussian divisible results of [26, 6] for the Step 2. For symmetric ensembles, it requires the first four moments matching those of GOE. As in our approach, a key analytic input for [40] is the local semicircle law established in [17]. The bulk universality in the case of symmetric matrices in the generality as stated in Mehta's book [28] (in particular, without the assumption to match four moments), was proved in [19, 23]. The key input is to link universality to local ergodicity of Dyson Brownian motion, reviewed in the previous paragraphs.

Due to the fundamental role of the local semicircle law, its error estimates were improved many times since its first proof in [17]. Furthermore, it was extended to sample covariance ensembles [21] and *generalized Wigner ensembles* [22] whose matrix elements are allowed to have different but comparable variances. The best existing error estimates for local semicircle law of generalized Wigner ensembles, given in [23], are already almost optimal in the bulk of the spectrum, but not near the edges. In this paper, we will prove a *strong local semicircle law*, Theorem 2.1, which, up to $\log N$ factors, gives optimal error estimates everywhere in the spectrum. There are four important consequences of this result:

1. It implies that Assumption III holds with the right hand side of (1.4) given by $N^{-2}(\log N)^{C \log\log N}$ for some constant C, i.e., $\mathfrak{a}$ can be chosen arbitrary close to $1/2$. Thus the Dyson Brownian motion reaches local equilibrium at $t \sim N^{-1+\delta}$ for arbitrary small δ. Up to the factor N^{δ}, this is optimal. Since the time to the global equilibrium for the Dyson Brownian motion is order one, we have thus established *Dyson's conjecture* [15] that the Dyson Brownian motion reaches equilibrium in two well-separated stages with time scales of order one and N^{-1}. As a historical note, we mention that Dyson had obtained the two time scales via heuristic physical argument and commented that a rigorous proof of his prediction is lacking. Furthermore, the notion of local equilibrium was used by Dyson in a very vague sense, see [19] for a more detailed discussion.

2. It implies certain explicit error estimates for the universality of correlation functions in short scales.

3. It implies the *rigidity of eigenvalues* in the sense that

$$\mathbb{P}\left\{ \exists j \ : \ |\lambda_j - \gamma_j| \geq (\log N)^{C \log\log N} \Big[\min\big(j, N-j+1 \big) \Big]^{-1/3} N^{-2/3} \right\} \leq C \exp\big[-(\log N)^{c \log\log N} \big] \tag{1.5}$$

   for some positive constants C and c. In other words, the eigenvalue is near its classical location with an error of at most $N^{-1}(\log N)^{C \log\log N}$ for generalized Wigner matrices in the bulk and the estimate deteriorates by a factor $\left(\frac{N}{j}\right)^{1/3}$ near the edge $j \ll N$.

4. It implies the *edge universality* in the sense that the probability distributions of the largest (and the smallest) eigenvalues of two generalized Wigner ensembles are equal in the large N limit provided that the second moments of the two ensembles are identical. We recall the standard assumption that the first moments of the matrix elements are always zero for all generalized Wigner ensembles. The comparison between our edge universality theorem and the previous results will be given at the end of Section 2 after the statement of Theorem 2.4.

It is well-known that the gaps between extremal eigenvalues and their fluctuations are of order $N^{-2/3}$. Thus the edge deterioration factor in (1.5) is the natural interpolation between N^{-1} in the bulk and $N^{-2/3}$ on the edges. The surprising feature of the rigidity estimate is that even if one eigenvalue is at a slightly wrong location, the probability is already extremely small. We remark that, without the $(\log N)^{C \log\log N}$ factor, the rigidity estimate (1.5) would be wrong since, at least for the classical GUE or GOE ensembles, the eigenvalues are known to fluctuate on a scale $\sqrt{\log N}/N$, see [25, 29]. For these ensembles, the distribution of $\lambda_j - \gamma_j$ is Gaussian in the bulk. However, the rigidity estimate (1.5) in this strong probabilistic form was not available even for the classical Gaussian ensembles.

2 Main results

Let $H = (h_{ij})_{i,j=1}^N$ be an $N \times N$ hermitian or symmetric matrix where the matrix elements $h_{ij} = \overline{h}_{ji}$, $i \leq j$, are independent random variables given by a probability measure ν_{ij} with mean zero and variance $\sigma_{ij}^2 \geq 0$:

$$\mathbb{E}\, h_{ij} = 0, \qquad \sigma_{ij}^2 := \mathbb{E}|h_{ij}|^2. \tag{2.1}$$

The distribution ν_{ij} and its variance σ_{ij}^2 may depend on N, but we omit this fact in the notation. Denote by $B := \{\sigma_{ij}^2\}_{i,j=1}^N$ the matrix of variances. The following assumptions on B are made throughout the paper:

(A) For any j fixed

$$\sum_{i=1}^N \sigma_{ij}^2 = 1\,. \tag{2.2}$$

Thus B is symmetric and double stochastic and, in particular, it satisfies $-1 \leq B \leq 1$.

(B) We assume that there exists two positive constants, δ_- and δ_+, independent of N, such that

$$1 \text{ is a simple eigenvalue of } B \text{ and } \mathrm{Spec}(B) \subset [-1+\delta_-, 1-\delta_+] \cup \{1\}. \tag{2.3}$$

(C) There is a constant C_0, independent of N, such that

$$\max_{ij}\{\sigma_{ij}^2\} \leq \frac{C_0}{N}. \tag{2.4}$$

For the orientation of the reader, we mention two special cases that provided the main motivation for our work.

Example 1. Generalized Wigner matrix. Define $C_{inf}(N)$ and $C_{sup}(N)$ by

$$C_{inf}(N) := \inf_{i,j}\{N\sigma_{ij}^2\} \leq \sup_{i,j}\{N\sigma_{ij}^2\} =: C_{sup}(N). \tag{2.5}$$

The ensemble is called *generalized Wigner ensemble* provided that

$$0 < C_- \leq C_{inf}(N) \leq C_{sup}(N) \leq C_+ < \infty, \tag{2.6}$$

for some $C_\pm$ independent of N. In this case, one can easily prove that 1 is a simple eigenvalue of B and (2.3) holds with some

$$\delta_\pm \geq C_-, \tag{2.7}$$

i.e., apart from the trivial eigenvalue, the spectrum of B is separated away ± 1 by positive constants that are independent of N. The special case $C_{inf} = C_{sup} = 1$ reduces to the standard Wigner matrices.

Example 2. Certain band matrices with bandwidth of order N. Band matrices are characterized by the property that σ_{ij}^2 is a function of $|i-j|$ on scale W, which is called the bandwidth. More precisely, the variances of a band matrix with bandwidth $1 \leq W \leq N/2$ are given by

$$\sigma_{ij}^2 = W^{-1} f\Big(\frac{[i-j]_N}{W}\Big), \tag{2.8}$$

where $f : \mathbb{R} \to \mathbb{R}_+$ is a bounded nonnegative symmetric function with $\int f = 1$ and we defined $[i-j]_N \in \mathbb{Z}$ by the property that $[i-j]_N \equiv i-j \mod N$ and $-\frac{1}{2}N < [i-j]_N \leq \frac{1}{2}N$. We often consider the case when $W = W(N)$, i.e. the bandwidth is a function of N. The condition **(A)** holds only asymptotically as $W(N) \to \infty$ but it can be remedied by an irrelevant rescaling. If the bandwidth is comparable with N, then we also have to assume that $f(x)$ is supported in $|x| \leq N/(2W)$.

It is easy to see that many band matrices satisfy the spectral assumption (2.3). The lower spectral bound, $-1 + \delta_- \le B$ with some $\delta_- > 0$ depending only on f, holds for any sufficiently large W, see Appendix A of [21]. The parameter δ_+ in the upper spectral bound typically behaves as of order $(W/N)^2$. Thus, for the condition **(B)** to hold, we need to assume that the bandwidth is comparable with N, i.e., it satisfies $W \ge cN$ with some positive constant c. The same assumption also guarantees that condition **(C)** holds.

We remark that the special case $W = N/2$ and $f(x) \ge c > 0$ for $|x| \le 1$ was already covered by Example 1, but Example 2 allows more general band matrices that may have vanishing variances. For example, with the choice of $f(x) = \frac{1}{2} \cdot \mathbf{1}(|x| \le 1)$, the ensemble with variances

$$\sigma_{ij}^2 = (N/2)^{-1} \mathbf{1}\left([i-j]_N \le N/4\right) \tag{2.9}$$

is a band matrix with bandwidth $W = N/4$.

Define the Green function of H by

$$G_{ij}(z) = \left(\frac{1}{H-z}\right)_{ij}, \qquad z = E + i\eta, \qquad E \in \mathbb{R}, \quad \eta > 0. \tag{2.10}$$

The Stieltjes transform of the empirical eigenvalue distribution of H is given by

$$m(z) = m_N(z) := \frac{1}{N} \sum_j G_{jj}(z) = \frac{1}{N} \mathrm{Tr}\, \frac{1}{H-z}\,. \tag{2.11}$$

Define $m_{sc}(z)$ as the unique solution of

$$m_{sc}(z) + \frac{1}{z + m_{sc}(z)} = 0, \tag{2.12}$$

with positive imaginary part for all z with $\mathfrak{Im}\, z > 0$, i.e.,

$$m_{sc}(z) = \frac{-z + \sqrt{z^2-4}}{2}, \tag{2.13}$$

where the square root function is chosen with a branch cut in the segment $[-2, 2]$ so that asymptotically $\sqrt{z^2-4} \sim z$ at infinity. This guarantees that the imaginary part of m_{sc} is non-negative for $\eta = \mathfrak{Im}\, z > 0$ and in the $\eta \to 0$ limit it is the Wigner semicircle distribution

$$\varrho_{sc}(E) := \lim_{\eta \to 0+0} \frac{1}{\pi} \mathfrak{Im}\; m_{sc}(E + i\eta) = \frac{1}{2\pi}\sqrt{(4-E^2)_+}. \tag{2.14}$$

The Wigner semicircle law [45] states that $m_N(z) \to m_{sc}(z)$ for any fixed z, i.e., provided that $\eta = \mathfrak{Im}\, z > 0$ is independent of N. Let $z = E + i\eta$ ($\eta > 0$) and denote $\kappa := ||E| - 2|$ the distance of E to the spectral edges ± 2. We have proved [23] a local version of this result for generalized Wigner matrices in the form of the following probability estimate:

$$\mathbb{P}\left(|m_N(z) - m_{sc}(z)| \ge \frac{N^\varepsilon}{N\eta\,\kappa}\right) \le \frac{C(\varepsilon, K)}{N^K} \tag{2.15}$$

that holds for any fixed positive constants ε and K and for any $z = E + i\eta$ such that $|E| \le 10$, $N\eta\kappa^{3/2} \ge N^\varepsilon$. Note that this estimate deteriorates near the spectral edges as $\kappa \ll 1$.

In this paper we prove the following local semicircle law that provides essentially the optimal estimate *uniformly* in $E = \mathfrak{Re}\, z$. We will estimate not only the deviation of $m(z)$ from $m_{sc}(z)$, but also the deviation of each diagonal matrix element of the resolvent, $G_{kk}(z)$, from $m_{sc}(z)$. Moreover, we show that the off-diagonal elements of the resolvent are small.

Let

$$v_k := G_{kk} - m_{sc}, \qquad m := \frac{1}{N}\sum_{k=1}^N G_{kk}, \qquad [v] := \frac{1}{N}\sum_{k=1}^N v_k = m - m_{sc}.$$

Our goal is to estimate the following quantities

$$\Lambda_d := \max_k |v_k| = \max_k |G_{kk} - m_{sc}|, \qquad \Lambda_o := \max_{k \neq \ell} |G_{k\ell}|, \qquad \Lambda := |m - m_{sc}|, \tag{2.16}$$

where the subscripts refer to "diagonal" and "off-diagonal" matrix elements. All these quantities depend on the spectral parameter z and on N but for simplicity we often omit this fact from the notation.

Theorem 2.1 (Strong local semicircle law) *Let $H = (h_{ij})$ be a hermitian or symmetric $N \times N$ random matrix, $N \geq 3$, with $\mathbb{E}\, h_{ij} = 0$, $1 \leq i,j \leq N$, and assume that the variances σ_{ij}^2 satisfy Assumptions* **(A)**, **(B)**, **(C)**, *i.e.* (2.2), (2.3) *and* (2.4). *Suppose that the distributions of the matrix elements have a uniformly subexponential decay in the sense that there exists a constant $\vartheta > 0$, independent of N, such that for any $x \geq 1$ and $1 \leq i,j \leq N$ we have*

$$\mathbb{P}(|h_{ij}| > x\sigma_{ij}) \leq \vartheta^{-1} \exp\big(-x^{\vartheta}\big). \tag{2.17}$$

Then there exist positive constants $A_0 > 1$, C, c and $\phi < 1$ depending only on ϑ, on $\delta_\pm$ from Assumption **(B)** *and on C_0 from Assumption* **(C)**, *such that for all L with*

$$A_0 \log\log N \leq L \leq \frac{\log(10N)}{10 \log\log N} \tag{2.18}$$

the following estimates hold for any sufficiently large $N \geq N_0(\vartheta, \delta_\pm, C_0)$:

(i) The Stieltjes transform of the empirical eigenvalue distribution of H satisfies

$$\mathbb{P}\Big(\bigcup_{z \in \mathbf{S}_L} \Big\{\Lambda(z) \geq \frac{(\log N)^{4L}}{N\eta}\Big\}\Big) \leq C \exp\big[-c(\log N)^{\phi L}\big], \tag{2.19}$$

where

$$\mathbf{S} := \mathbf{S}_L = \Big\{z = E + i\eta \;:\; |E| \leq 5, \quad N^{-1}(\log N)^{10L} < \eta \leq 10\Big\}. \tag{2.20}$$

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

$$\mathbb{P}\left(\bigcup_{z \in \mathbf{S}_L} \left\{\Lambda_d(z) + \Lambda_o(z) \geq (\log N)^{4L}\sqrt{\frac{\mathfrak{Im}\, m_{sc}(z)}{N\eta}} + \frac{(\log N)^{4L}}{N\eta}\right\}\right) \leq C \exp\big[-c(\log N)^{\phi L}\big]. \tag{2.21}$$

(iii) The largest eigenvalue of H is bounded by $2 + N^{-2/3}(\log N)^{9L}$ in the sense that

$$\mathbb{P}\Big(\max_{j=1,\ldots,N} |\lambda_j| \geq 2 + N^{-2/3}(\log N)^{9L}\Big) \leq C \exp\big[-c(\log N)^{\phi L}\big]. \tag{2.22}$$

The subexponential decay condition (2.17) can be weakened if we are not aiming at error estimates faster than any power law of N. This can be easily carried out and we will not pursue it in this paper. We also note that the upper bound on L originates from the natural requirement that $\mathbf{S}_L \neq \emptyset$.

Prior to our results in [22] and [23], a central limit theorem for the semicircle law on macroscopic scale for band matrices was established by Guionnet [24] and Anderson and Zeitouni [3]; a semicircle law for Gaussian band matrices was proved by Disertori, Pinson and Spencer [14]. For a review on band matrices, see the recent article [39] by Spencer.

The local semicircle estimates imply that the empirical counting function of the eigenvalues is close to the semicircle counting function and that the locations of the eigenvalues are close to their classical location in mean square deviation sense. Recall that $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_N$ are the ordered eigenvalues of H. We define the *normalized empirical counting function* by

$$\mathfrak{n}(E) := \frac{1}{N} \#\{\lambda_j \le E\}. \tag{2.23}$$

Let

$$n_{sc}(E) := \int_{-\infty}^{E} \varrho_{sc}(x) \mathrm{d}x \tag{2.24}$$

be the distribution function of the semicircle law and recall that $\gamma_j = \gamma_{j,N}$ denote the classical location of the j-th point under the semicircle law, see (1.3).

Theorem 2.2 (Rigidity of Eigenvalues) *Suppose that Assumptions* **(A)**, **(B)**, **(C)** *and the condition* (2.17) *hold. Then there exist positive constants $A_0 > 1$, C, c and $\phi < 1$ depending only on ϑ, on $\delta_\pm$ from Assumption* **(B)** *and on C_0 from Assumption* **(C)** *such that for any L with*

$$A_0 \log\log N \le L \le \frac{\log(10N)}{10 \log\log N}$$

we have

$$\mathbb{P}\left\{ \exists j \;:\; |\lambda_j - \gamma_j| \ge (\log N)^L \Big[\min\big(j, N-j+1 \big) \Big]^{-1/3} N^{-2/3} \right\} \le C \exp\big[-c(\log N)^{\phi L} \big] \tag{2.25}$$

and

$$\mathbb{P}\left\{ \sup_{|E|\le 5} \big| \mathfrak{n}(E) - n_{sc}(E) \big| \ge \frac{(\log N)^L}{N} \right\} \le C \exp\big[-c(\log N)^{\phi L} \big] \tag{2.26}$$

for any sufficiently large $N \ge N_0(\vartheta, \delta_\pm, C_0)$.

For standard Wigner matrices, (2.26) with the factor N^{-1} replaced by $N^{-2/5}$ (in a weaker sense with some modifications in the statement) was established in [5] and a stronger $N^{-1/2}$ control was proven for $\mathbb{E}\mathfrak{n}(E) - n_{sc}(E)$. If we replaced $(\log N)^L$ factor by N^δ for arbitrary $\delta > 0$, (2.26) was proved in [23] (Theorem 6.3) with some deterioration near the spectral edges and with a slightly weaker probability estimate. In Theorem 1.3 of a recent preprint [42], the following estimate (in our scaling)

$$\Big(\mathbb{E}\big[|\lambda_j - \gamma_j|^2 \big] \Big)^{1/2} \le \Big[\min\big(j, N-j+1 \big) \Big]^{-1/3} N^{-1/6-\varepsilon_0} \tag{2.27}$$

with some small positive ε_0 was proved under the assumption that the third moment of the matrix element vanishes and all variances of the matrix elements are identical, i.e., for the standard Wigner matrices with vanishing third moment. In the same paper, it was conjectured that the factor $N^{-1/6-\varepsilon_0}$ on the right hand side of (2.27) should be replaced by $N^{-2/3+\varepsilon}$. Prior to the work [42], the estimate (2.25) away from the edges with a slightly weaker probability estimate and with the $(\log N)^L$ factor replaced by N^δ for arbitrary $\delta > 0$ was proved in [23] (see the equation before (7.8) in [23]). For Wigner matrices whose matrix element distributions matching the standard Gaussian random variable up to the third moment, it was proved in [40] that $|\lambda_j - \gamma_j| \leq N^{-1+\varepsilon}$ holds in the bulk in probability (Theorem 32). More detailed behavior can be obtained if one assumes further that the fourth moment also matches the standard Gaussian random variable, see Corollary 21 of [40] for more details. Near the edge, (2.25) with $N^{-2/3}$ replaced by $N^{-1/2}$ and the probability estimate on the right side replaced by a Gaussian type estimate was proved in [1].

We remark that all results in this paper are stated for both the hermitian or symmetric case, but the statements hold for quaternion self-dual random matrices as well (see, e.g., Section 3.1 of [21]). The proofs will be presented for the hermitian case for definiteness but with obvious modifications they are valid for the other two cases as well.

We will frequently use the notation C and c for generic positive constants and N_0 for the lower threshold for N in this paper. We adopt the convention that, unless stated otherwise, these constants and also the implicit constants in the $O(\cdot)$ notation may depend on the basic parameters of our model, namely on ϑ, $\delta_\pm$ and C_0. The values of these generic constants may change from line to line.

2.1 Bulk Universality

We now use Theorem 2.2 to establish the speed of convergence for local statistics of Dyson Brownian motion. In fact, we will replace the Brownian motion in the definition of Dyson Brownian motion by an Ornstein-Uhlenbeck process. We thus consider a flow of random matrices H_t satisfying the following matrix valued stochastic differential equation

$$\mathrm{d}H_t = \frac{1}{\sqrt{N}}\mathrm{d}\beta_t - \frac{1}{2}H_t\mathrm{d}t, \tag{2.28}$$

where β_t is a hermitian matrix valued process whose diagonal matrix elements are standard real Brownian motions and the off-diagonal elements are independent standard complex Brownian motions; with all Brownian motions being independent. The initial condition H_0 is the original hermitian Wigner matrix. For any fixed $t \geq 0$, the distribution of H_t coincides with that of

$$e^{-t/2}H_0 + (1-e^{-t})^{1/2}\,V, \tag{2.29}$$

where V is an independent GUE matrix whose matrix elements are centered Gaussian random variables with variance $1/N$. For the symmetric case, the matrix elements of β_t in (2.28) are real Brownian motions and V in (2.29) is a GOE matrix. It is well-known that the eigenvalues of H_t follow a process that is also called the Dyson Brownian motion (in our case with a drift but we will still call it Dyson Brownian motion).

More precisely, let

$$\mu = \mu_N(\mathrm{d}\mathbf{x}) = \frac{e^{-\mathcal{H}(\mathbf{x})}}{Z_\beta}\mathrm{d}\mathbf{x}, \qquad \mathcal{H}(\mathbf{x}) = N\left[\beta\sum_{i=1}^N \frac{x_i^2}{4} - \frac{\beta}{N}\sum_{i<j}\log|x_j - x_i|\right] \tag{2.30}$$

be the probability measure of the eigenvalues of the general β ensemble, with $\beta \geq 1$ ($\beta = 2$ for GUE, $\beta = 1$ for GOE). Here Z_β is the normalization factor so that μ is probability measure. In this section, we often use the notation x_j instead of λ_j for the eigenvalues to follow the notations of [21]. Denote the distribution of the eigenvalues at time t by $f_t(\mathbf{x})\mu(\mathrm{d}\mathbf{x})$. Then f_t satisfies

$$\partial_t f_t = \mathscr{L} f_t. \tag{2.31}$$

where

$$\mathscr{L} = \sum_{i=1}^N \frac{1}{2N}\partial_i^2 + \sum_{i=1}^N \Big(-\frac{\beta}{4}x_i + \frac{\beta}{2N}\sum_{j\neq i}\frac{1}{x_i - x_j}\Big)\partial_i. \tag{2.32}$$

For any $n \geq 1$ we define the n-point correlation functions (marginals) of the probability measure $f_t \mathrm{d}\mu$ by

$$p_{t,N}^{(n)}(x_1, x_2, \dots, x_n) = \int_{\mathbb{R}^{N-n}} f_t(\mathbf{x})\mu(\mathbf{x})\mathrm{d}x_{n+1}\dots \mathrm{d}x_N. \tag{2.33}$$

With a slight abuse of notations, we will sometimes also use μ to denote the density of the measure μ with respect to the Lebesgue measure. The correlation functions of the equilibrium measure are denoted by

$$p_{\mu,N}^{(n)}(x_1, x_2, \dots, x_n) = \int_{\mathbb{R}^{N-n}} \mu(\mathbf{x})\mathrm{d}x_{n+1}\dots \mathrm{d}x_N. \tag{2.34}$$

The main result in [21] concerning Dyson Brownian motion, Theorem 2.1, states that the local ergodicity of Dyson Brownian motion holds for $t \geq N^{-2\mathfrak{a}+\delta}$ for any $\delta > 0$ provided that the Assumption III (1.4) holds. In fact, the estimate on the relaxation to the local equilibrium [21] is not restricted to Dyson Brownian motion; it applies to all flows satisfying four general assumptions, labelled as Assumption I-IV in [21]. Instead of repeating these assumptions in their general forms, we will give only simple sufficient conditions. Assumption I requires that the probability density of the global equilibrium measure is given by a Hamiltonian of the form

$$\mathcal{H} = \mathcal{H}_N(\mathbf{x}) = \beta\Big[\sum_{j=1}^N U(x_j) - \frac{1}{N}\sum_{i<j}\log|x_i - x_j|\Big], \tag{2.35}$$

where $\beta \geq 1$ and the function $U : \mathbb{R} \to \mathbb{R}$ is smooth with $U'' \geq \delta$ for some positive δ. This is clearly satisfied since the equilibrium measures are either GUE or GOE in the setting of this paper. Assumption II requires a limiting continuous density for the eigenvalue distribution. In our case, the density is given by the semicircle law. Assumption IV asserts that the local density of eigenvalues is bounded down to scale $\eta = N^{-1+\sigma}$ for any $\sigma > 0$. This assumption follows from the large deviation estimate (2.19) since a bound on $\Lambda(z)$, $z = E + i\eta$, can be easily used to prove an upper bound on the local density of eigenvalues in a window of size η about E. As usual, the additional condition in [21] on the entropy $S_\mu(f_{t_0}) \leq CN^m$ for $t_0 = N^{-2\mathfrak{a}}$ holds due to the regularization property of the Ornstein-Uhlenbeck process. Thus for a given $0 < \varepsilon' < 1$, choosing $\mathfrak{a} = 1/2 - \varepsilon'/2$, $A = \varepsilon'$ in the second part of Theorem 2.1 in [21] and using (2.25), we have the following theorem.

Theorem 2.3 (Strong local ergodicity of Dyson Brownian motion) *Let H be a hermitian or symmetric $N \times N$ random matrix with $\mathbb{E}\, h_{ij} = 0$ and suppose that Assumptions* **(A)***,* **(B)***,* **(C)** *and* (2.17) *hold with parameters $\delta_\pm, C_0$ and ϑ. Then for any $\varepsilon' > 0$, $\delta > 0$, $c > 0$ positive numbers, for any integer $n \geq 1$ and for any compactly supported continuous test function $O : \mathbb{R}^n \to \mathbb{R}$ there exists a constant C depending on all these parameters and on O such that*

$$\begin{aligned} &\sup_{t \geq N^{-1+\delta+\varepsilon'}} \Big| \int_{E-b}^{E+b} \frac{\mathrm{d}E'}{2b} \int_{\mathbb{R}^n} \mathrm{d}\alpha_1 \dots \mathrm{d}\alpha_n \, O(\alpha_1, \dots, \alpha_n) \frac{1}{\varrho(E)^n} \\ &\quad \times \Big(p_{t,N}^{(n)} - p_{\mu,N}^{(n)}\Big)\Big(E' + \frac{\alpha_1}{N\varrho(E)}, \dots, E' + \frac{\alpha_n}{N\varrho(E)}\Big)\Big| \leq CN^{2\varepsilon'}\Big[b^{-1}N^{-1+\varepsilon'} + b^{-1/2}N^{-\delta/2}\Big], \end{aligned} \tag{2.36}$$

holds for any fixed $E \in [2-c, 2+c]$ and for any $b = b_N \in (0, c/2)$ that may depend on N. Here $p_{t,N}^{(n)}$ and $p_{\mu,N}^{(n)}$, (2.33)–(2.34)*, are the correlation functions of the eigenvalues of the Dyson Brownian motion flow* (2.29) *and those of the equilibrium measure, respectively.*

Besides a weaker version of Theorem 2.3 was proved in [23], a similar result, with no error estimate, was obtained in [20] for the hermitian case by using an explicit formula related to Johansson's formula [26]. Theorem 2.3, however, contains explicit estimates and is valid for a time range much bigger than the previous results. In particular, we mention the following three special cases:

- If we choose $\delta = 1 - 2\varepsilon'$ and thus $t = N^{-\varepsilon'}$, then we can choose $b \sim N^{-1}$ and the universality is valid with essentially no averaging in E.
- If we choose the energy window of size $b \sim 1$ and the time $t = N^{-\varepsilon'}$, then the error estimate is of order $\sim N^{-1/2}$.
- If we choose $b \sim 1$, then the smallest time scale for which we can prove the universality is $t = N^{-1+\varepsilon'}$. This scale, up to the arbitrary small exponent ε', is optimal in accordance with the time scale to local equilibrium conjectured by Dyson [15].

For generalized Wigner matrices with a subexponential decay, i.e. assuming (2.6) in addition to the conditions of Theorem 2.3, the universality result with no explicit error estimate holds for any time $t \geq 0$. More precisely, for any fixed $b > 0$ we have

$$\begin{aligned} \lim_{N\to 0} \sup_{t\geq 0} \Big| \int_{E-b}^{E+b} \frac{\mathrm{d}E'}{2b} \int_{\mathbb{R}^n} \mathrm{d}\alpha_1 \ldots \mathrm{d}\alpha_n \, O(\alpha_1, \ldots, \alpha_n) \frac{1}{\varrho(E)^n} \\ \times \Big(p_{t,N}^{(n)} - p_{\mu,N}^{(n)} \Big) \Big(E' + \frac{\alpha_1}{N\varrho(E)}, \ldots, E' + \frac{\alpha_n}{N\varrho(E)} \Big) \Big| = 0. \end{aligned} \tag{2.37}$$

This result, with slightly stronger conditions on the distributions of the ensemble, was already proved in [23]. Similarly to [23], the extension of the universality from a small positive time to zero time requires a different method, the Green function comparison theorem [22] in our approach. The reasons of universality for zero time and time bigger than $1/N$ are very different: Theorem 2.3 shows that the local correlation functions have already reached their equilibrium under the Dyson Brownian motion flow for any time larger than $1/N$. For time smaller than $1/N$, in particular the important case $t = 0$, the universality is valid because we can compare the local correlation functions at time $t = 0$ with the ones generated by the flow at time $t = N^{-\varepsilon}$ with *specially adjusted initial data* (see, e.g., the Matching Lemma 3.4 [23]). The same argument as in Section 3 of [23] can be used to prove (2.37) from (2.36). In fact, since our new version of the strong local ergodicity of Dyson Brownian motion, Theorem 2.3, holds for very short times, the two ensembles to be compared are already very close to each other. Furthermore, effective error estimates instead of a limiting statement (2.37) can also be obtained and the parameter b may also be chosen N-dependent. For the case that b is N-independent, the time to local equilibrium as remarked above is $N^{-1+\varepsilon}$. Hence the condition (2.6) can be replaced by the following condition: there are constants $c, \varepsilon > 0$ such that

$$|\{(i,j) : N\sigma_{ij}^2 \leq c\}| \leq N^{2-\varepsilon} \, . \tag{2.38}$$

Since these extensions require only minor modifications of the current method, we will not pursue these directions in this paper.

2.2 Edge distribution

Recall that λ_N is the largest eigenvalue of the random matrix. The probability distribution functions of λ_N for the classical Gaussian ensembles are identified by Tracy and Widom [43, 44] to be

$$\lim_{N\to\infty} \mathbb{P}(N^{2/3}(\lambda_N - 2) \le s) = F_\beta(s), \tag{2.39}$$

where the function $F_\beta(s)$ can be computed in terms of Painlevé equations and $\beta = 1, 2, 4$ corresponds to the standard classical ensembles. The distribution of λ_N is believed to be universal and independent of the Gaussian structure. The strong local semicircle law, Theorem 2.1, combined with a modification of the Green function comparison theorem (Theorem 6.3) implies the following version of universality of the extreme eigenvalues.

Theorem 2.4 (Universality of extreme eigenvalues) *Suppose that we have two $N \times N$ matrices, $H^{(v)}$ and $H^{(w)}$, with matrix elements h_{ij} given by the random variables $N^{-1/2}v_{ij}$ and $N^{-1/2}w_{ij}$, respectively, with v_{ij} and w_{ij} satisfying the uniform subexponential decay condition* (2.17). *Let $\mathbb{P}^{\mathbf{v}}$ and $\mathbb{P}^{\mathbf{w}}$ denote the probability and $\mathbb{E}^{\mathbf{v}}$ and $\mathbb{E}^{\mathbf{w}}$ the expectation with respect to these collections of random variables. Suppose that Assumptions* **(A)**, **(B)**, **(C)** *hold for both ensembles. If the first two moments of v_{ij} and w_{ij} are the same, i.e.*

$$\mathbb{E}^{\mathbf{v}} \bar{v}_{ij}^l v_{ij}^u = \mathbb{E}^{\mathbf{w}} \bar{w}_{ij}^l w_{ij}^u, \qquad 0 \le l + u \le 2, \tag{2.40}$$

then there is an $\varepsilon > 0$ and $\delta > 0$ depending on ϑ in (2.17) *such that for any real parameter s (may depend on N) we have*

$$\mathbb{P}^{\mathbf{v}}(N^{2/3}(\lambda_N - 2) \le s - N^{-\varepsilon}) - N^{-\delta} \le \mathbb{P}^{\mathbf{w}}(N^{2/3}(\lambda_N - 2) \le s) \le \mathbb{P}^{\mathbf{v}}(N^{2/3}(\lambda_N - 2) \le s + N^{-\varepsilon}) + N^{-\delta} \tag{2.41}$$

for $N \ge N_0$ sufficiently large, where N_0 is independent of s. Analogous result holds for the smallest eigenvalue λ_1.

Theorem 2.4 can be extended to finite correlation functions of extreme eigenvalues. For example, we have the following extension to (2.41):

$$\begin{aligned}
&\mathbb{P}^{\mathbf{v}}\Big(N^{2/3}(\lambda_N - 2) \le s_1 - N^{-\varepsilon}, \ldots, N^{2/3}(\lambda_{N-k} - 2) \le s_{k+1} - N^{-\varepsilon}\Big) - N^{-\delta} \\
&\le \mathbb{P}^{\mathbf{w}}\Big(N^{2/3}(\lambda_N - 2) \le s_1, \ldots, N^{2/3}(\lambda_{N-k} - 2) \le s_{k+1}\Big) \\
&\le \mathbb{P}^{\mathbf{v}}\Big(N^{2/3}(\lambda_N - 2) \le s_1 + N^{-\varepsilon}, \ldots, N^{2/3}(\lambda_{N-k} - 2) \le s_{k+1} + N^{-\varepsilon}\Big) + N^{-\delta}
\end{aligned} \tag{2.42}$$

for all k fixed and N sufficiently large. The proof of (2.42) is similar to that of (2.41) and we will not provide details except stating the general form of the Green function comparison theorem (Theorem 6.4) needed in this case. We remark that edge universality is usually formulated in terms of joint distributions of edge eigenvalues in the form (2.42) with fixed parameters $s_1, s_2, \ldots$. Our result holds uniformly in these parameters, i.e., they may depend on N. However, the interesting regime is $|s_j| \le (\log N)^{C \log\log N}$, otherwise the rigidity estimate (2.25) gives a stronger control than (2.42).

The edge universality for Wigner matrices was first proved via the moment method by Soshnikov [37] (see also the earlier work [34]) for Hermitian and orthogonal ensembles with symmetric distributions to ensure that all odd moments vanish. By combining the moment method and Chebyshev polynomials, Sodin proved edge universality of band matrices and some special class of sparse matrices [35, 36].

The removal of the symmetry assumption was not straightforward. The approach of [35, 36] is restricted to ensembles with symmetric distributions. The symmetry assumption was partially removed in [31, 32] and significant progress was made in [41] which assumes only that the first three moments of two Wigner ensembles are identical. In other words, the symmetry assumption was replaced by the vanishing third moment condition for Wigner matrices. For a special class of ensembles, the Gaussian divisible Hermitian ensembles, edge universality was proved [27] under the sole condition that the fourth moment is finite, which in our scaling means that $\mathbb{E}|\sqrt{N}h_{ij}|^4$ is a positive constant. Using this result [27], one can remove the vanishing third moment condition in [41] for Hermitian Wigner ensembles.

In comparison with these results, Theorem 2.4 does not imply the edge universality of band matrices or sparse matrices [35, 36], but it implies in particular that, for the purpose to identify the distribution of the top eigenvalue for a generalized Wigner matrix with the subexponential decay condition, it suffices to consider generalized Wigner ensembles with Gaussian distribution. Since the distributions of the top eigenvalues of the Gaussian Wigner ensembles are given by F_β (2.39), Theorem 2.4 implies the edge universality of the standard Wigner matrices under the subexponential decay assumption alone. We remark that one can use Theorem 2.2 as an input in the approach of [27] to prove that the distributions of the top eigenvalues of the generalized hermitian Wigner ensembles with Gaussian distributions are given by F_2. Therefore the Tracy-Widom distribution also holds for any generalized hermitian Wigner ensemble with subexponential decay. But for ensembles in different symmetry classes (e.g., symmetric Wigner ensembles), there is no corresponding result to identify the distribution of the top eigenvalue with F_β if the variances are allowed to vary.

Finally, we comment that the subexponential decay assumption in our approach, though can be weakened, is far from optimal, see [4, 7, 33, 38] for discussions on optimal moment assumptions. Our approach based on the local semicircle law, however, gives both the bulk and edge universality and the symmetry of the distribution of matrix elements plays no role.

3 Apriori bound for the strong local semicircle law

We first prove a weaker form of Theorem 2.1, and in Section 4 we will use this apriori bound to obtain the stronger form as claimed in Theorem 2.1.

Theorem 3.1 *Let $H = (h_{ij})$ be a hermitian $N \times N$ random matrix, $N \geq 3$, with $\mathbb{E}\, h_{ij} = 0$, $1 \leq i, j \leq N$, and assume that the variances σ_{ij}^2 satisfy Assumptions* **(A)**, **(B)**, **(C)** *and assume the uniform subexponential decay* (2.17). *Then there exist constants $0 < \phi < 1$, $C \geq 1$ and $c > 0$, depending only on ϑ from* (2.17), *$\delta_\pm$ from Assumption* **(B)** *and on C_0 is from Assumption* **(C)** *such that for any ℓ with $4/\phi \leq \ell \leq C \log N / \log \log N$ and for any $z = E + i\eta \in \mathbf{S}_\ell$ we have*

$$\mathbb{P}\left\{ \max_i |G_{ii}(z) - m_{sc}(z)| \geq \frac{(\log N)^\ell}{(N\eta)^{1/3}} \right\} \leq C \exp\big[-c(\log N)^{\phi \ell} \big] \tag{3.1}$$

and

$$\mathbb{P}\left\{ \max_{i \neq j} |G_{ij}(z)| \geq \frac{(\log N)^\ell}{(N\eta)^{1/2}} \right\} \leq C \exp\big[-c(\log N)^{\phi \ell} \big] \tag{3.2}$$

for any sufficiently large $N \geq N_0(\theta, \delta_\pm, C_0)$.

We remark that the probabilistic estimates in Theorem 3.1 are stated for fix $z \in \mathbf{S}_\ell$, but it is easy to deduce from them probabilistic statements that hold simultaneously for all z, e.g.

$$\mathbb{P}\left(\bigcup_{z\in\mathbf{S}_\ell}\left\{\max_i |G_{ii}(z)-m_{sc}(z)| \geq \frac{(\log N)^\ell}{(N\eta)^{1/3}}\right\}\right) \leq C\exp\big[-c(\log N)^{\phi\ell}\big].$$

This holds true because in the set $\mathbf{S}_\ell$ the Green function and $m_{sc}(z)$ are Lipschitz continuous in z with a Lipschitz constant bounded by $\eta^{-2} \ll N^2$; for example $|\partial_z G_{ij}(z)| \leq |\mathfrak{Im}\, z|^{-2} \leq N^2$. Consider an N^{-10}-net in the compact set $\mathbf{S}_\ell$, i.e., a set of points $\{z_k\} \subset \mathbf{S}_\ell$ such that $\min_k |z-z_k| \leq N^{-10}$ for any $z \in \mathbf{S}_\ell$ and such that the cardinality of $\{z_k\}$ is at most CN^{20}. Using that the estimates (3.1)–(3.2) hold simultaneously for all points z_k (since these estimates decay faster than any polynomial in N by $\phi\ell > 1$), we see that similar estimates, with a smaller c, hold simultaneously for any $z \in \mathbf{S}_\ell$.

We will follow the self consistent perturbation ideas initiated in [22, 23]. We first introduce some notations.

Definition 3.1 *Let $\mathbb{T} = \{k_1, k_2, \ldots, k_t\} \subset \{1,2,\ldots,N\}$ be an unordered set of $|\mathbb{T}| = t$ elements and let $H^{(\mathbb{T})}$ be the $N-t$ by $N-t$ minor of H after removing the k_i-th $(1 \leq i \leq t)$ rows and columns. For $\mathbb{T} = \emptyset$, we define $H^{(\emptyset)} = H$. Similarly, we define $\mathbf{a}^{(j;\,\mathbb{T})}$ to be j-th column of H with the k_i-th $(1 \leq i \leq t)$ elements removed. Sometimes, we just use the short notation $\mathbf{a}^j = \mathbf{a}^{(j;\,\mathbb{T})}$. Note that the ℓ-th entry of $\mathbf{a}^j$ is $\mathbf{a}^j_\ell = h_{\ell j}$ for $\ell \notin \mathbb{T}$. For any $\mathbb{T} \subset \{1,2,\ldots,N\}$ we introduce the following notations:*

$$\begin{aligned}
G^{(\mathbb{T})}_{ij} &:= (H^{(\mathbb{T})}-z)^{-1}(i,j), \qquad i,j \notin \mathbb{T}\\
Z^{(\mathbb{T})}_{ij} &:= \mathbf{a}^i \cdot (H^{(\mathbb{T})}-z)^{-1}\mathbf{a}^j = \sum_{k,\ell\notin\mathbb{T}} \overline{\mathbf{a}^i_k} G^{(\mathbb{T})}_{k\ell}\mathbf{a}^j_\ell\\
K^{(\mathbb{T})}_{ij} &:= h_{ij} - z\delta_{ij} - Z^{(\mathbb{T})}_{ij}.
\end{aligned} \tag{3.3}$$

These quantities depend on z, but we mostly neglect this dependence in the notation.

The following formulas were proved in Lemma 4.2 of [22].

Lemma 3.2 (Self-consistent perturbation formulas) *Let $\mathbb{T} \subset \{1,2,\ldots,N\}$. For simplicity, we use the notation $(i\,\mathbb{T})$ for $(\{i\}\cup\mathbb{T})$ and $(ij\,\mathbb{T})$ for $(\{i,j\}\cup\mathbb{T})$. Then we have the following identities:*

1. *For any $i \notin \mathbb{T}$*
$$G^{(\mathbb{T})}_{ii} = (K^{(i\,\mathbb{T})}_{ii})^{-1}. \tag{3.4}$$

2. *For $i \neq j$ and $i,j \notin \mathbb{T}$*
$$G^{(\mathbb{T})}_{ij} = -G^{(\mathbb{T})}_{jj} G^{(j\,\mathbb{T})}_{ii} K^{(ij\,\mathbb{T})}_{ij} = -G^{(\mathbb{T})}_{ii} G^{(i\,\mathbb{T})}_{jj} K^{(ij\,\mathbb{T})}_{ij}. \tag{3.5}$$

3. *For $i \neq j$ and $i,j \notin \mathbb{T}$*
$$G^{(\mathbb{T})}_{ii} - G^{(j\,\mathbb{T})}_{ii} = G^{(\mathbb{T})}_{ij} G^{(\mathbb{T})}_{ji} (G^{(\mathbb{T})}_{jj})^{-1}. \tag{3.6}$$

4. *For any indices i, j and k that are different and $i,j,k \notin \mathbb{T}$*
$$G^{(\mathbb{T})}_{ij} - G^{(k\,\mathbb{T})}_{ij} = G^{(\mathbb{T})}_{ik} G^{(\mathbb{T})}_{kj} (G^{(\mathbb{T})}_{kk})^{-1}. \tag{3.7}$$

The following large deviation estimates concerning independent random variables were proved in Appendix B of [22].

Lemma 3.3 *Let a_i $(1 \le i \le N)$ be independent complex random variables with mean zero, variance σ^2 and having a uniform subexponential decay*

$$\mathbb{P}(|a_i| \ge x\sigma) \le \vartheta^{-1} \exp\big(-x^{\vartheta}\big), \qquad \forall\, x \ge 1, \tag{3.8}$$

with some $\vartheta > 0$. Let A_i, $B_{ij} \in \mathbb{C}$ $(1 \le i, j \le N)$. Then there exists a constant $0 < \phi < 1$, depending on ϑ, such that for any $\zeta > 1$ we have

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} a_i A_i\right| \ge (\log N)^{\zeta}\sigma\Big(\sum_i |A_i|^2\Big)^{1/2}\right\} \le \exp\big[-(\log N)^{\phi\zeta}\big], \tag{3.9}$$

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} \overline{a}_i B_{ii} a_i - \sum_{i=1}^{N} \sigma^2 B_{ii}\right| \ge (\log N)^{\zeta}\sigma^2\Big(\sum_{i=1}^{N} |B_{ii}|^2\Big)^{1/2}\right\} \le \exp\big[-(\log N)^{\phi\zeta}\big], \tag{3.10}$$

$$\mathbb{P}\left\{\left|\sum_{i\neq j} \overline{a}_i B_{ij} a_j\right| \ge (\log N)^{\zeta}\sigma^2\Big(\sum_{i\neq j} |B_{ij}|^2\Big)^{1/2}\right\} \le \exp\big[-(\log N)^{\phi\zeta}\big] \tag{3.11}$$

for any sufficiently large $N \ge N_0$, where $N_0 = N_0(\vartheta)$ depends on ϑ.

The following lemma (Lemma 4.2 from [23]) collects elementary properties of the Stieljes transform of the semicircle law. As a technical note, we use the notation $f \sim g$ for two positive functions in some domain D if there is a positive universal constant C such that $C^{-1} \le f(z)/g(z) \le C$ holds for all $z \in D$.

Lemma 3.4 *We have for all z with $\mathfrak{Im}\, z > 0$ that*

$$|m_{sc}(z)| = |m_{sc}(z) + z|^{-1} \le 1. \tag{3.12}$$

From now on, let $z = E + i\eta$ with $|E| \le 5$ and $0 < \eta \le 10$ and we set $\kappa = \big||E| - 2\big|$. Then we have

$$|m_{sc}(z)| \sim 1, \qquad |1 - m_{sc}^2(z)| \sim \sqrt{\kappa + \eta} \tag{3.13}$$

and the following two bounds:

$$\mathfrak{Im}\, m_{sc}(z) \sim \begin{cases} \frac{\eta}{\sqrt{\kappa+\eta}} & \text{if } \kappa \ge \eta \text{ and } |E| \ge 2 \\[2ex] \sqrt{\kappa + \eta} & \text{if } \kappa \le \eta \text{ or } |E| \le 2. \end{cases} \tag{3.14}$$

□

3.1 Self-consistent perturbation equations

Following [22, 23], we define the following quantities:

$$A_i := \sigma_{ii}^2 G_{ii} + \sum_{j\neq i} \sigma_{ij}^2 \frac{G_{ij}G_{ji}}{G_{ii}} \tag{3.15}$$

$$Z_i := \sum_{k,l\neq i} \Big[\overline{\mathbf{a}_k^i} G_{k\,l}^{(i)} \mathbf{a}_l^i - \mathbb{E}_{\mathbf{a}^i} \overline{\mathbf{a}_k^i} G_{k\,l}^{(i)} \mathbf{a}_l^i \Big] = Z_{ii}^{(i)} - \mathbb{E}_{\mathbf{a}^i} Z_{ii}^{(i)}, \tag{3.16}$$

$$\Upsilon_i := A_i + \Big(K_{ii}^{(i)} - \mathbb{E}_{\mathbf{a}^i} K_{ii}^{(i)} \Big) = A_i + h_{ii} - Z_i, \tag{3.17}$$

where $\mathbb{E}_{\mathbf{a}^i}$ indicates the expectation with respect to the matrix elements in the i-th column. Using (3.4) from Lemma 3.2, we obtain the following system of self-consistent equations for the deviation from m_{sc} of the diagonal matrix elements of the resolvent;

$$v_i = G_{ii} - m_{sc} = \frac{1}{-z - m_{sc} - \Big(\sum_j \sigma_{ij}^2 v_j - \Upsilon_i \Big)} - m_{sc}. \tag{3.18}$$

For the off-diagonal terms, we will use the equation (3.5). All the quantities defined so far depend on the spectral parameter $z = E + i\eta$, but we will mostly omit this fact from the notation.

The key quantities Λ, Λ_d and Λ_o (2.16) appearing in Theorem 3.1 will be typically small and we will prove in this section that their size is less than $(N\eta)^{-1/3}$, modulo logarithmic corrections. We thus define the exceptional (bad) event

$$\mathbf{B} = \mathbf{B}(z) := \Big\{ \Lambda_d(z) + \Lambda_o(z) \geq (\log N)^{-2} \Big\}. \tag{3.19}$$

We will always work in the complement set $\mathbf{B}^c$, i.e., we will have

$$\Lambda_d(z) + \Lambda_o(z) \leq (\log N)^{-2}. \tag{3.20}$$

We collect some basic properties of the Green function in the following elementary lemma.

Lemma 3.5 *Let $\mathbb{T}$ be a subset of $\{1, \ldots, N\}$. Then there exists a constant $C = C_{\mathbb{T}}$ depending on C_0 from (2.4) and on $|\mathbb{T}|$, the cardinality of $\mathbb{T}$, such that the following hold in $\mathbf{B}^c$*

$$|G_{kk}^{(\mathbb{T})} - m_{sc}| \leq \Lambda_d + C\Lambda_o^2 \qquad \text{for all } k \notin \mathbb{T}, \tag{3.21}$$

$$\frac{1}{C} \leq |G_{kk}^{(\mathbb{T})}| \leq C \qquad \text{for all } k \notin \mathbb{T}, \tag{3.22}$$

$$\max_{k\neq l} |G_{k\,l}^{(\mathbb{T})}| \leq C\Lambda_o, \tag{3.23}$$

$$\max_i |A_i| \leq \frac{C}{N} + C\Lambda_o^2 \tag{3.24}$$

for any fixed $|\mathbb{T}|$ and for any sufficiently large N. We recall that all quantities depend on the spectral parameter z and the estimates are uniform in $z = E + i\eta$ as long as $|E| \leq 5$, $0 < \eta \leq 10$.

Proof. For $\mathbb{T} = \emptyset$, the estimates (3.21) and (3.23) follow directly from the definitions (2.16). The bound (3.22) follows from (3.20) and that $|m_{sc}(z)| \sim 1$, see (3.12). Finally, (3.24) follows from inserting (3.22), (3.23), (2.2) and (2.4) into (3.15). The general case can be proved by induction on $|\mathbb{T}|$ and using the formulas (3.6) and (3.7) that guarantee that

$$|G_{k\ell}^{(\mathbb{T})} - G_{k\ell}^{(\mathbb{T}')}| \le C^* \Lambda_o^2 \tag{3.25}$$

holds for any $\mathbb{T}' = \mathbb{T} \cup \{m\}$, where C^* depends on the constant $C_{\mathbb{T}}$ for the induction hypothesis. In the set $\mathbf{B}^c$ and for sufficiently large N, depending on $|\mathbb{T}|$, the estimate (3.25) together with $|m_{sc}(z)| \sim 1$ guarantees that the lower bound in (3.22) continues to hold for $\mathbb{T}'$. The other estimates for $\mathbb{T}'$ follow from (3.25) directly.

□

3.2 Estimate of the exceptional events

The following lemma is a modification of Lemma 4.5 in [23]. It improves the estimate in the sense that the control parameter depends only on Λ but not on Λ_d and Λ_o (see (2.16) for definitions). Since Λ, being an average quantity, behaves better, this yields a stronger estimate.

For any $\ell > 0$ we define the key control parameter Ψ, which is random variable, by

$$\Psi(z) := (\log N)^{\ell} \sqrt{\frac{\Lambda(z) + \mathfrak{Im}\, m_{sc}(z)}{N\eta}}. \tag{3.26}$$

We also define the events

$$\begin{aligned} \Omega_h &:= \left\{ \max_{1\le i,j\le N} |h_{ij}| \ge (\log N)^{\ell/10} |\sigma_{ij}| \right\} \cup \left\{ \Big| \sum_{i=1}^N h_{ii} \Big| \ge (\log N)^{\ell/10} \right\} \\ \Omega_d(z) &:= \left\{ \max_i |Z_i(z)| \ge \frac{1}{2}\Psi(z) \right\} \\ \Omega_o(z) &:= \left\{ \max_{i\ne j} |Z_{ij}^{(ij)}(z)| \ge \frac{1}{2}\Psi(z) \right\}, \end{aligned} \tag{3.27}$$

and we let

$$\Omega(z) := \Omega_h \cup \Big[\big(\Omega_d(z) \cup \Omega_o(z)\big) \cap \mathbf{B}(z)^c \Big] \tag{3.28}$$

be the set of exceptional events. These definitions depend on the parameter ℓ that we omit from the notation.

The main reason that Ψ emerges as the key controlling parameter can be seen from the following consideration. In order to estimate the off-diagonal term G_{ij}, we need to bound (3.5) K_{ij} and thus Z_{ij}. By the large deviation estimate, (3.11), we have

$$|Z_{ij}^{(ij)}| \le C(\log N)^{\ell/3} \sqrt{\sum_{k,l\ne i,j} \Big|\sigma_{ik} G_{k\,l}^{(ij)} \sigma_{lj}\Big|^2} \le C(\log N)^{\ell/3} \sqrt{\frac{1}{N^2} \sum_{k,l\ne i,j} \Big|G_{k\,l}^{(ij)}\Big|^2} \tag{3.29}$$

holds with high probability. Here we have used that $\sigma_{il}^2 \le C_0/N$ from (2.4).

For any normal matrix A, we have

$$\sum_j |A_{ij}|^2 = (AA^*)_{ii} = (|A|^2)_{ii} \tag{3.30}$$

where $|A|^2 := AA^*$.

Applying this identity to the Green function $G = [H - z]^{-1}$, we obtain the following "Ward identity":

$$\sum_l |G_{k\,l}|^2 = \sum_\alpha \frac{|u_\alpha(k)|^2}{|\lambda_\alpha - z|^2} = \frac{\mathfrak{Im}\, G_{kk}}{\eta}, \tag{3.31}$$

where u_α and λ_α are the eigenvectors and eigenvalues of H. The term "Ward identity" comes from quantum field theory and it represents an identity derived from a conservation law or symmetry of a system. In our case, the symmetry is generated by the global phase multiplication $e^{i\theta}$, but this connection is not important for our purpose.

Applying (3.31) to estimate the last term in (3.29) and *neglecting the superscript* (ij), we can bound $Z_{ij}^{(ij)}$ by

$$|Z_{ij}^{(ij)}(z)| \le C(\log N)^{\ell/3} \sqrt{\frac{N^{-1}\sum_k \mathfrak{Im}\, G_{kk}}{N\eta}} \le C(\log N)^{\ell/3} \sqrt{\frac{\Lambda(z) + \mathfrak{Im}\, m_{sc}(z)}{N\eta}}$$

where we have used the definition of Λ in the last inequality. Notice that the control parameter Ψ appears naturally in this estimate. Furthermore, it is $\mathfrak{Im}\, m_{sc}(z)$ which appears in the numerator, not $m_{sc}(z)$. This is the fundamental reason that we are able to obtain optimal estimate up to the edges of the spectrum. Near the edges, $\mathfrak{Im}\, m_{sc}(z)$ is small while $|m_{sc}(z)|$ stays near 1.

Lemma 3.6 *There exist a constant $0 < \phi < 1$, depending on ϑ (2.17), and universal constants $C > 1$, $c > 0$, such that for any ℓ with $4/\phi \le \ell \le C \log N / \log\log N$ and for any $z \in \mathbf{S}_\ell$ we have*

$$\mathbb{P}(\Omega(z)) \le C \exp\big[-c(\log N)^{\phi\ell}\big], \tag{3.32}$$

and we also have the pointwise statement

$$(\log N)^{\ell/2}\Lambda_o(z) + \max_i |\Upsilon_i(z)| \le \Psi(z) \quad \text{in} \quad \Omega(z)^c \cap \mathbf{B}(z)^c \tag{3.33}$$

for any sufficiently large $N \ge N_0(\vartheta)$.

Proof. There exists $0 < \phi < 1$, depending on ϑ, such that the following two estimates hold for any $\ell \ge 4/\phi$:

$$\mathbb{P}\Big\{|h_{ij}| \ge (\log N)^{\ell/10}|\sigma_{ij}|\Big\} \le C \exp\big[-(\log N)^{\phi\ell}\big], \qquad \forall i,j$$

by (2.17), and

$$\mathbb{P}\Bigg\{\Big|\sum_{i=1}^N h_{ii}\Big| \ge (\log N)^{\ell/10}\Bigg\} \le C \exp\big[-(\log N)^{\phi\ell}\big]$$

by (2.4) and the large deviation principle for the sum of independent random variables (e.g., (3.9)). Thus

$$\mathbb{P}\left(\Omega_h\right) \le C \exp\big[-c(\log N)^{\phi\ell}\big], \tag{3.34}$$

so we can work on the complement set Ω_h^c. Note that

$$\Omega^c \cap \mathbf{B}^c = \Omega_h^c \cap \Omega_d^c \cap \Omega_o^c \cap \mathbf{B}^c. \tag{3.35}$$

Fix $z \in \mathbf{S}_\ell$ and we will prove, possibly with a smaller ϕ, that for $\ell \geq 4/\phi$ we have

$$\mathbb{P}\Big(\Omega_h^c \cap \Omega_d(z) \cap \mathbf{B}^c(z)\Big) \leq C \exp\big[-c(\log N)^{\phi\ell}\big] \tag{3.36}$$

and

$$\mathbb{P}\Big(\Omega_h^c \cap \Omega_o(z) \cap \mathbf{B}^c(z)\Big) \leq C \exp\big[-c(\log N)^{\phi\ell}\big], \tag{3.37}$$

and this will prove (3.32).

To prove the diagonal estimate (3.36), we can choose a sufficiently small $\phi > 0$ (depending on ϑ) and apply the large deviation bound (3.10) from Lemma 3.3 to obtain that for any fixed i

$$|Z_i| \leq (\log N)^{\ell/3} \sqrt{\sum_{k,l \neq i} \Big|\sigma_{ik} G^{(i)}_{k\,l} \sigma_{li}\Big|^2} \tag{3.38}$$

holds with a probability larger than $1 - C \exp\big[-c(\log N)^{\phi\ell}\big]$ for sufficiently large N. From the Ward identity (3.31) and $\sigma_{il}^2 \leq C_0/N$ (by (2.4) and (3.21)), we have

$$\sum_{k,l \neq i} \Big|\sigma_{ik} G^{(i)}_{k\,l} \sigma_{li}\Big|^2 \leq \frac{C_0}{N} \sum_{k \neq i} \frac{\mathfrak{Im}\, G^{(i)}_{kk}}{N\eta}. \tag{3.39}$$

Since we are in the set $\mathbf{B}^c$, we have $\Lambda_d + \Lambda_o \leq (\log N)^{-2}$. Thus from (3.6) and (3.22) we have that

$$0 < \mathfrak{Im}\, G^{(i)}_{kk} \leq \mathfrak{Im}\, G_{kk} + |G^{(i)}_{kk} - G_{kk}| \leq \mathfrak{Im}\, G_{kk} + C|G_{ik}|^2 \leq \mathfrak{Im}\, G_{kk} + C\Lambda_o^2. \tag{3.40}$$

The last term of (3.39) is bounded by

$$\frac{C_0^2}{N^2} \sum_{k \neq i} \frac{\mathfrak{Im}\, G^{(i)}_{kk}}{\eta} \leq C \frac{\Lambda + \Lambda_o^2 + \mathfrak{Im}\, m_{sc}}{N\eta} \qquad \text{in } \mathbf{B}^c. \tag{3.41}$$

We have thus proved that for any $z \in \mathbf{S}_\ell$

$$|Z_i(z)| \leq C(\log N)^{\ell/3} \sqrt{\frac{\Lambda(z) + \Lambda_o^2(z) + \mathfrak{Im}\, m_{sc}(z)}{N\eta}} \qquad \text{in } \mathbf{B}^c(z). \tag{3.42}$$

holds with a probability larger than $1 - C \exp\big[-c(\log N)^{\phi\ell}\big]$ for sufficiently large N.

Similarly, for the off-diagonal estimate (3.37), for any fixed $i \neq j$, we have from (3.11) that

$$|Z^{(ij)}_{ij}| \leq C(\log N)^{\ell/3} \sqrt{\sum_{k,l \neq i,j} \Big|\sigma_{ik} G^{(ij)}_{k\,l} \sigma_{lj}\Big|^2} \tag{3.43}$$

holds with a probability larger than $1 - C \exp\big[-c(\log N)^{\phi\ell}\big]$ for sufficiently large N. Similarly to the proof of (3.42) for Z_i, we have

$$|Z^{(ij)}_{ij}(z)| \leq C(\log N)^{\ell/3} \sqrt{\frac{\Lambda(z) + \Lambda_o^2(z) + \mathfrak{Im}\, m_{sc}(z)}{N\eta}} \qquad \text{in } \mathbf{B}^c(z) \tag{3.44}$$

holds for any $z \in \mathbf{S}_\ell$ with a probability larger than $1 - C\exp\big[-c(\log N)^{\phi\ell}\big]$ for sufficiently large N.

Using Lemma 3.5, we have $|G_{ii}| \le C$ and $|G_{jj}^{(i)}| \le C$ in the set $\mathbf{B}^c$. From (3.5), we can thus estimate the off-diagonal term G_{ij} by

$$|G_{ij}| = |G_{ii}||G_{jj}^{(i)}||K_{ij}^{(ij)}| \le C\Big(|h_{ij}| + |Z_{ij}^{(ij)}|\Big), \qquad i \ne j, \qquad \text{in } \mathbf{B}^c. \tag{3.45}$$

Hence we have that in the event $\mathbf{B}^c \cap \Omega_h^c$

$$\Lambda_o = \max_{i\ne j} |G_{ij}| \le \frac{C(\log N)^{\ell/10}}{\sqrt{N}} + C(\log N)^{\ell/3}\sqrt{\frac{\Lambda + \Lambda_o^2 + \mathfrak{Im}\, m_{sc}}{N\eta}} \tag{3.46}$$

holds with a probability larger than $1 - C\exp\big[-c(\log N)^{\phi\ell}\big]$ for sufficiently large N.

Recall that $N\eta \ge (\log N)^{10\ell}$ on the set $\mathbf{S}_\ell$ and since $\ell \ge 4/\phi \ge 4$, we have $(\log N)^{\ell/3} \ll \sqrt{N\eta}$, thus the Λ_o term on the right hand side of (3.46) can be absorbed into the left side for sufficiently large N. Furthermore, by (3.14), we have $\mathfrak{Im}\, m_{sc}(z) \ge c\eta$ with a universal positive constant c for any $z \in \mathbf{S}_\ell$. Thus the first term on the right hand side of (3.46) can be bounded by

$$\frac{C(\log N)^{\ell/10}}{\sqrt{N}} \le (\log N)^{\ell/3}\sqrt{\frac{\mathfrak{Im}\, m_{sc}(z)}{N\eta}}$$

for large enough N, and thus it can be absorbed into the second term. We conclude that

$$\mathbb{P}\Big\{\Lambda_o \le C(\log N)^{-2\ell/3}\Psi, \quad \mathbf{B}^c \cap \Omega_h^c\Big\} \ge 1 - C\exp\big[-c(\log N)^{\phi\ell}\big]. \tag{3.47}$$

Inserting this bound into (3.42) and (3.44), we have proved (3.36) and (3.37). Finally, the estimate (3.33) for Υ and Λ_o is a simple consequence of (3.47), the definition (3.17), the bound (3.24), the definition of Ω_d and that $\Omega^c \cap \mathbf{B}^c \subset \Omega_h^c$. This completes the proof of Lemma 3.6. $\square$

3.3 Analysis of the self-consistent equation

Now we start using the self-consistent equation (3.18). Since

$$\Big|\sum_j \sigma_{ij}^2 v_j - \Upsilon_i\Big| \le \Lambda_d + |\Upsilon_i|,$$

the bound (3.12) allows us to expand the denominator in (3.18) as long as $\Lambda_d + |\Upsilon_i| \le \frac{1}{2}$. In this case, using (2.12), we obtain the following equation for v_i

$$v_i = m_{sc}^2\Big(\sum_j \sigma_{ij}^2 v_j - \Upsilon_i\Big) + m_{sc}^3\Big(\sum_j \sigma_{ij}^2 v_j - \Upsilon_i\Big)^2 + O\Big(\sum_j \sigma_{ij}^2 v_j - \Upsilon_i\Big)^3. \tag{3.48}$$

Recall that B denotes the $N \times N$ matrix of covariances, $B = (\sigma_{ij}^2)$. Thus we can rewrite the last equation as

$$[(1 - m_{sc}^2 B)\mathbf{v}]_i = -m_{sc}^2\Upsilon_i + m_{sc}^3\Big((B\mathbf{v})_i - \Upsilon_i\Big)^2 + O\Big((B\mathbf{v})_i - \Upsilon_i\Big)^3.$$

We will first use this equation to estimate $v_i - [v]$, i.e. the deviation of v_i from its average (Lemma 3.8). In the second step, we will add up (3.48) for all i and obtain an equation for $[v]$ (Lemma 3.9). Finally, we use a dichotomy argument to estimate $\Lambda = |[v]|$ in Lemma 3.10.

By normalization assumption $\sum_j \sigma_{ij}^2 = 1$, the vector $\mathbf{e} = (1, 1, \ldots, 1)$ is the (unique) eigenvector of B with eigenvalue 1. We introduce the notation

$$q = q(z) := \max\{\delta_+, |1 - \mathfrak{Re}\, m_{sc}^2(z)|\}, \tag{3.49}$$

and we recall the following elementary lemma that was proven in [23, Lemma 4.8].

Lemma 3.7 *The matrix $\mathbb{I} - m_{sc}^2(z)B$ is invertible on the subspace orthogonal to $\mathbf{e}$. Let $\mathbf{u}$ be a vector which is orthogonal to $\mathbf{e}$ and let*

$$\mathbf{w} = (\mathbb{I} - m_{sc}^2(z)B)\mathbf{u},$$

then

$$\|\mathbf{u}\|_\infty \le \frac{C \log N}{q(z)} \|\mathbf{w}\|_\infty$$

for some constant C that only depends on δ_- in (2.3). □

The following lemma estimates the deviation of v_i from its average $[v]$:

Lemma 3.8 *Suppose that $4 \le \ell \le C \log N / \log\log N$. Fix the spectral parameter $z \in \mathbf{S}_\ell$ and we will omit it from the notations. Suppose that in some set Ξ it holds that*

$$\Lambda_d \le \frac{q}{(\log N)^{3/2}}, \tag{3.50}$$

then in the set $\Xi \cap \Omega^c \cap \mathbf{B}^c$ we have

$$\max_i |v_i - [v]| \le \frac{C \log N}{q} \left(\Lambda^2 + \Psi + \frac{(\log N)^2}{q^2} \Psi^2 \right) \le \frac{C \log N}{q^3} \left(\Lambda^2 + \Psi \right) \tag{3.51}$$

for some constant C depending only on δ_- and for sufficiently large N.

Proof. For $z \in \mathbf{S}_\ell$, $q(z)$ and $\mathfrak{Im}\, m_{sc}(z)$ are bounded. Combining (3.50) with the definitions of $\Psi(z)$, $\mathbf{S}_\ell$ and with $\ell \ge 4$, we obtain that $\Lambda_d(z)$, $\Lambda(z)$ are bounded by $C(\log N)^{-3/2}$ and $\Psi(z)$ is bounded by $C(\log N)^{-2}$. Thus the expansion (3.48) holds true in the set $\Xi \cap \Omega^c \cap \mathbf{B}^c$, by using (3.33). We can estimate the second and third order terms in (3.48) by $C(\Psi + \Lambda_d)^2$ and we obtain

$$v_i = m_{sc}^2 \sum_j \sigma_{ij}^2 v_j + \varepsilon_i, \qquad \text{with} \quad \varepsilon_i = O(\Psi) + O(\Lambda_d^2) \quad \text{in } \; \Xi \cap \Omega^c \cap \mathbf{B}^c. \tag{3.52}$$

Taking the average over i, we have

$$(1 - m_{sc})[v] = \frac{1}{N} \sum_i \varepsilon_i = O(\Psi) + O(\Lambda_d^2),$$

and thus it follows from (3.52) that

$$v_i - [v] = m_{sc}^2 \sum_j \sigma_{ij}^2 (v_j - [v]) + O(\Psi) + O(\Lambda_d^2).$$

Applying Lemma 3.7 for $u_i = v_i - [v]$, we obtain

$$\max_i |v_i - [v]| \le \frac{C \log N}{q} \left(\Lambda_d^2 + \Psi\right), \tag{3.53}$$

hence

$$\Lambda_d \le \Lambda + \frac{C \log N}{q} \left(\Lambda_d^2 + \Psi\right).$$

With (3.50), this inequality implies

$$\Lambda_d \le \Lambda + \frac{C \log N}{q} (\Lambda^2 + \Psi). \tag{3.54}$$

Using (3.54) to bound Λ_d^2 in (3.53), we have proved the first inequality of (3.51), the second one follows from $\Psi \le C(\log N)^{-2}$. This completes the proof of Lemma 3.8. □

In this paper we assumed that the positive constants $\delta_\pm$ are independent of N (see (2.3)), thus q is bounded and the condition (3.50) is automatically satisfied in the set $\mathbf{B}^c$, see (3.20), and therefore (3.51) can be written as

$$\max_i |v_i - [v]| \le C(\log N) \left(\Lambda^2 + \Psi\right) \qquad \text{in } \Omega^c \cap \mathbf{B}^c, \tag{3.55}$$

in particular,

$$\Lambda_d \le \Lambda + C(\log N)(\Lambda^2 + \Psi) \qquad \text{in } \Omega^c \cap \mathbf{B}^c, \tag{3.56}$$

with some constant C depending only on $\delta_\pm$.

Lemma 3.9 *Suppose that $4 \le \ell \le C \log N / \log\log N$. Fix the spectral parameter $z \in \mathbf{S}_\ell$ and we will omit it from the notations. Then in the set $\Omega^c \cap \mathbf{B}^c$ we have*

$$(1 - m_{sc}^2)[v] = m_{sc}^3 [v]^2 + m_{sc}^2 [Z] + O\Big(\frac{\Lambda^2}{\log N}\Big) + O\Big((\log N)\Psi^2\Big), \tag{3.57}$$

where $[Z] := N^{-1} \sum_{i=1}^N Z_i$. The implicit constants in the error terms depend only on $\delta_\pm$ and C_0.

Proof. From the choice $\ell \ge 4$, and from $\Lambda \le (\log N)^{-2}$ in the set $\mathbf{B}^c$, we have

$$\Psi \le (\log N)^{-8}. \tag{3.58}$$

Moreover, for $z \in \mathbf{S}_\ell$, we have $\mathfrak{Im}\, m_{sc}(z) \ge c\eta$ with some universal positive constant c (see Lemma 3.4), we also have

$$\Psi \ge \frac{(\log N)^\ell}{\sqrt{N}}. \tag{3.59}$$

By the definition of Υ_i (3.17), by the estimates (3.24) and (3.33), we have

$$\Upsilon_i = A_i + h_{ii} - Z_i = h_{ii} - Z_i + O(\Lambda_o^2 + N^{-1}) = h_{ii} - Z_i + O(\Psi^2) \qquad \text{in } \Omega^c \cap \mathbf{B}^c. \tag{3.60}$$

The size of the last term of (3.48) is less than $O(\Psi^3 + \Lambda_d^3)$ which is bounded by $O(\Psi^2 + \Lambda^3)$ using (3.56) and (3.58). Thus we have, from (3.33) and (3.48),

$$v_i = m_{sc}^2 \Big(\sum_j \sigma_{ij}^2 v_j + Z_i - h_{ii} + O(\Psi^2) \Big) + m_{sc}^3 \Big(\sum_j \sigma_{ij}^2 v_j + O(\Psi) \Big)^2 + O\Big(\Psi^2 + \Lambda^3\Big) \qquad \text{in } \Omega^c \cap \mathbf{B}^c. \tag{3.61}$$

Summing up i and dividing by N, we obtain

$$[v] = m_{sc}^2[v] + m_{sc}^2[Z] + O(\Psi^2 + \Lambda^3) + \frac{m_{sc}^3}{N}\sum_i \Big(\sum_j \sigma_{ij}^2 v_j + O(\Psi)\Big)^2 \qquad \text{in } \Omega^c \cap \mathbf{B}^c. \tag{3.62}$$

Here we used that in the set $\Omega^c \cap \mathbf{B}^c \subset \Omega_h^c$, we have $N^{-1}|\sum_i h_{ii}| \le (\log N)^{\ell/10} N^{-1} \le \Psi^2$ by (3.59). Writing $v_j = (v_j - [v]) + [v]$, the last term in (3.62) can be estimated using (3.55)

$$\frac{m_{sc}^3}{N}\sum_i \Big(\sum_j \sigma_{ij}^2 v_j + O(\Psi)\Big)^2 = m_{sc}^3 [v]^2 + O\big((\log N)\Psi(\Lambda^2 + \Psi)\big) + O(\Lambda\Psi) + O(\Psi^2).$$

Collecting the various error terms and using (3.58) and that $\Lambda \le (\log N)^{-2}$ in $\mathbf{B}^c$, we obtain (3.57) from (3.62). This completes the proof of Lemma 3.9. □

3.4 Dichotomy estimate for Λ

Throughout this section we fix the parameter ℓ with $4 \le \ell \le C\log N/\log\log N$. By Lemma 3.9 we have that in $\Omega^c \cap \mathbf{B}^c$

$$(1 - m_{sc}^2)[v] - m_{sc}^3[v]^2 = O(\Psi) + O(\Lambda^2)/\log N, \tag{3.63}$$

where we have used the simple bound $\Psi \le 1/\log N$ and that in the set $\Omega(z)^c \cap \mathbf{B}(z)^c$ all Z_i, hence $[Z]$ can be bounded by Ψ (see (3.35) and the definition of Ω_d).

We introduce the following notations:

$$\alpha := \Big|\frac{1 - m_{sc}^2}{m_{sc}^3}\Big| \qquad \beta := \frac{(\log N)^{2\ell}}{(N\eta)^{1/3}}, \quad \text{with} \quad \eta = \mathfrak{Im}\, z, \tag{3.64}$$

where $\alpha = \alpha(z)$ and $\beta = \beta(z)$ depend on the spectral parameter z. For any $z \in \mathbf{S}_\ell$ we have the bound $\beta(z) \le (\log N)^{-4}$, by $\ell \ge 4$. From Lemma 3.4 it also follows that there is a universal constant $K \ge 1$ such that

$$\frac{1}{K}\sqrt{\kappa + \eta} \le \alpha(z) \le K\sqrt{\kappa + \eta} \tag{3.65}$$

for any $z \in \mathbf{S}_\ell$.

By definition of $\Psi = \Psi(z)$ (3.26), we have

$$\begin{aligned}\Psi =& (\log N)^\ell \sqrt{\frac{\Lambda + \mathfrak{Im}\, m_{sc}}{N\eta}} \\ \le& (\log N)^\ell \frac{\Lambda + \mathfrak{Im}\, m_{sc}}{(N\eta)^{1/3}} + (\log N)^\ell (N\eta)^{-2/3} \le \beta\Lambda + \alpha\beta + \beta^2,\end{aligned} \tag{3.66}$$

where, in the last step, we have used that $\alpha(z) \sim \sqrt{\kappa + \eta}$, see (3.65), and thus $\mathfrak{Im}\, m_{sc}(z) \le C\alpha(z)$ (see Lemma 3.4). We conclude from (3.63) and $|m_{sc}| \sim 1$ that

$$\Big|\frac{1 - m_{sc}^2}{m_{sc}^3}[v] - [v]^2\Big| \le C^*\big(\beta\Lambda + \alpha\beta + \beta^2\big) + O(\Lambda^2)/\log N \qquad \text{in } \Omega^c \cap \mathbf{B}^c \tag{3.67}$$

with some constant C^*.

Neglecting the error term and replacing $[v]$ by Λ, we roughly have the equation

$$\Big|\alpha\Lambda - \Lambda^2\Big| \le C^*\big(\beta\Lambda + \alpha\beta + \beta^2\big). \tag{3.68}$$

This inequality provides certain estimates on Λ depending on whether $\alpha \lesssim \beta$ or not.

Since α and β are functions of z ($\beta(z)$ depends only on $\eta = \mathfrak{Im}\, z$), we will fix $E = \mathfrak{Re}\, z$ and vary $\eta = \mathfrak{Im}\, z$ from $\eta = 10$ down to $\eta = (\log N)^{10\ell}/N$. Thanks to (3.65), $\alpha(z)$ is essentially monotone increasing in η, up to universal constants. The function $\beta(z)$ is monotonically decreasing. Therefore there exists a threshold $\widetilde{\eta}$ such that for $\eta \le \widetilde{\eta}$ we have $\alpha \lesssim \beta$ and for $\eta \ge \widetilde{\eta}$ we have $\alpha \gtrsim \beta$. To implement precisely the idea of dividing the estimate according to the relative size of α and β, we will need to choose a large but fixed constant $U > 1$ depending only on C^*. Let $\widetilde{\eta} = \widetilde{\eta}(U, E)$ be the solution to $\sqrt{\kappa + \eta} = 2U^2 K\beta(z)$ where $\kappa = \big||E| - 2\big|$. Note that up to a constant factor, this equation is the same as $\alpha(z) = \beta(z)$. Since $\sqrt{\kappa+\eta}$ is increasing while $\beta(z)$ is decreasing in η, the solution is unique and one can easily prove that

$$\widetilde{\eta} \le N^{-1/3} \tag{3.69}$$

for sufficiently large N, depending on U. The implementation of this idea and precise estimates on Λ is given by the following Lemma:

Lemma 3.10 [Dichotomy Lemma] *Suppose that $4 \le \ell \le C \log N / \log\log N$. Then there is a constant $U_0 = U_0(\delta_\pm, C_0) \ge 1$ such that for any $U \ge U_0$, there exists a constant $C_1(U)$, depending only on U, such that for any spectral parameter $z \in \mathbf{S}_\ell$ the following estimates hold*

$$\Lambda(z) \le U\beta(z) \quad \text{or} \quad \Lambda(z) \ge \frac{\alpha(z)}{U} \qquad \text{if} \quad \mathfrak{Im}\, z \ge \widetilde{\eta}(U, \mathfrak{Re}\, z) \tag{3.70}$$

$$\Lambda(z) \le C_1(U)\beta(z) \qquad \text{if} \quad \mathfrak{Im}\, z < \widetilde{\eta}(U, \mathfrak{Re}\, z) \tag{3.71}$$

in the set $\Omega(z)^c \cap \mathbf{B}(z)^c$ and for any sufficiently large $N \ge N_0(\delta_\pm, C_0)$.

Proof. We will set $U_0 = 9(C^* + 1)$ and let $U \ge U_0$ where C^* is the constant appearing in (3.67). Depending on the relative size of β and α, which is determined by z, we will either express $[v]$ or $[v]^2$ from (3.67). This will correspond to the two cases in Lemma 3.10. Recalling that $|[v]| = \Lambda$, the last error term in (3.67) can be easily absorbed for sufficiently large N and we will get a quadratic inequality for Λ.

Case 1: $\eta = \mathfrak{Im}\, z \ge \widetilde{\eta}(U, E)$. By the definition of $\widetilde{\eta}$, in this case $\sqrt{\kappa+\eta} \ge 2U^2 K\beta(z)$, i.e.,

$$\alpha(z) \ge 2U^2\beta(z) \tag{3.72}$$

by (3.65). From the choice of U_0 and $U \ge U_0$ we get that $\alpha \ge \beta$ and $\frac{1}{2}\alpha \ge C^*\beta$. Expressing $[v]$ from (3.67) and absorbing the $C^*\beta\Lambda$ term into the left hand side, we obtain

$$\frac{1}{2}\alpha\Lambda \le 2\Lambda^2 + 2C^*\alpha\beta. \tag{3.73}$$

Thus either

$$\frac{1}{4}\alpha\Lambda \le 2\Lambda^2,$$

i.e. $\Lambda \geq \alpha/8$ which is larger than α/U, or

$$\frac{1}{4}\alpha\Lambda \leq 2C^*\alpha\beta,$$

i.e. $\Lambda \leq 8C^*\beta \leq U\beta$, which proves (3.70).

Case 2: $\eta = \mathfrak{Im}\, z < \widetilde{\eta}(U, E)$. In this case $\sqrt{\kappa+\eta} \leq 2U^2K\beta(z)$, i.e., $\alpha(z) \leq 2U^2K^2\beta(z)$. We express $[v]^2$ from (3.67) and we get

$$\Lambda^2 \leq 2\alpha\Lambda + 2C^*\big[\,\beta\Lambda + \beta\alpha + \beta^2\,\big] \leq C'\beta\Lambda + C'\beta^2 \tag{3.74}$$

with a constant C' depending on U. This quadratic inequality immediately implies that $\Lambda \leq C_1(U)\beta$ with some U-dependent constant $C_1(U)$. Hence we have proved Lemma 3.10.

□

3.5 Initial estimates for large η

In this section we show that Theorem 3.1 holds for $\eta = \mathfrak{Im}\, z = 10$, i.e. on the upper boundary of $\mathbf{S}_\ell$. This will serve as an initial step for the continuity argument. The proof for $\eta = 10$ is similar to the arguments in Sections 3.2 and 3.3 but much easier. In particular, no apriori assumption similar to (3.20) or no bad set $\mathbf{B}$ are necessary. We start with the analogue of Lemma 3.6 which actually holds uniformly for any z with $0 < \eta = \mathfrak{Im}\, z \leq 10$ and not only for $z \in \mathbf{S}_\ell$. Note that these estimates are very weak for small η, but we will use them only for $\eta = 10$.

Lemma 3.11 *For any $z \in \mathbb{C}$ with $0 < \eta = \mathfrak{Im}\, z \leq 10$, define the exceptional events*

$$\begin{aligned} \Theta_d(z) &:= \Big\{ \max_i |Z_i(z)| \geq \frac{(\log N)^\ell}{\sqrt{N}\eta} \Big\} \\ \Theta_o(z) &:= \Big\{ \max_{i\neq j} |Z_{ij}^{(ij)}(z)| \geq \frac{(\log N)^\ell}{\sqrt{N}\eta} \Big\} \\ \Theta(z) &:= \Omega_h \cup \Theta_d(z) \cup \Theta_o(z), \end{aligned} \tag{3.75}$$

where we recall the definition of Ω_h in (3.27). Then there exists constants $0 < \phi < 1$, $C > 1$, $c > 0$, depending on ϑ (2.17), such that for any ℓ with $4/\phi \leq \ell \leq C \log N/\log\log N$ and for any $z \in \mathbf{S}_\ell$ we have

$$\mathbb{P}(\Theta(z)) \leq C \exp\big[-c(\log N)^{\phi\ell}\big], \tag{3.76}$$

and the pointwise bound

$$\max_i |\Upsilon_i(z)| \leq CN^{-1/3}\eta^{-3} \quad \text{in} \quad \Theta(z)^c \tag{3.77}$$

for sufficiently large $N \geq N_0(\vartheta, C_0)$. Furthermore, for $\eta \geq 3$ we have the estimate

$$\Lambda_d(z) + \Lambda_o(z) \leq CN^{-1/3} \quad \text{in} \quad \Theta(z)^c. \tag{3.78}$$

for sufficiently large $N \geq N_0(\vartheta, C_0)$.

Proof. Given the estimate (3.34), for the proof of (3.76) it is sufficient to estimate the probability of Θ_d and Θ_o. The estimate (3.39) still holds, but we can now bound the last term in (3.39) simply by

$$\sum_{k,l\neq i} \left|\sigma_{ik} G^{(i)}_{k\,l} \sigma_{li}\right|^2 \leq \frac{1}{N}\sum_{k\neq i} \sigma_{ik}^2 \frac{\mathfrak{Im}\, G^{(i)}_{kk}(z)}{\eta} \leq \frac{1}{N\eta^2}, \tag{3.79}$$

for any z, using the trivial deterministic estimate

$$|G^{(\mathbb{T})}_{ij}| \leq \eta^{-1} \tag{3.80}$$

that holds for any i, j and for any $\mathbb{T}$. Combining (3.79) with the large deviation bound (3.10) from Lemma 3.3 as in (3.38), we obtain $\mathbb{P}(\Theta_d) \leq C\exp\big[-c(\log N)^{\phi\ell}\big]$. The same argument holds for the exceptional set Θ_o involving the off-diagonal elements and this proves (3.76).

From (3.5) and the trivial estimate (3.80), we can estimate the off-diagonal term G_{ij} in the set $\Theta(z)^c$ by

$$|G_{ij}| = |G_{ii}||G^{(i)}_{jj}||K^{(ij)}_{ij}| \leq \eta^{-2}\left(|h_{ij}| + |Z^{(ij)}_{ij}|\right) \leq (\log N)^{\ell}\left[\frac{1}{\sqrt{N}\eta^2} + \frac{1}{\sqrt{N}\eta^3}\right] \leq N^{-1/3}\eta^{-3}, \quad i\neq j, \tag{3.81}$$

for sufficiently large N. Moreover, the same argument gives

$$\frac{|G_{ij}|}{|G_{ii}|} = |G^{(i)}_{jj}||K^{(ij)}_{ij}| \leq N^{-1/3}\eta^{-2}, \quad i\neq j,$$

which can be inserted in the definition of A, (3.15), and with $N\eta \gg 1$, we get

$$|A_i| \leq \frac{C_0}{N\eta} + \frac{1}{N^{1/3}\eta^3} \leq \frac{2}{N^{1/3}\eta^3}$$

for sufficiently large N. In the set Θ^c a similar bound holds for h_{ii} and Z_i using $\eta \leq 10$. Recalling that $\Upsilon_i = A_i + h_{ii} - Z_i$, and this proves (3.77).

For the proof of (3.78) it is sufficient to bound only Λ_d, the necessary estimate for Λ_o is given in (3.81). We define $\Upsilon = \max_i |\Upsilon_i|$ and note that for $\eta \geq 3$ we have $\Upsilon \leq CN^{-1/3}$ in the set Θ^c by (3.77). From the self consistent equation (3.18) and the defining equation (2.12) of m_{sc}, we have

$$v_n = \frac{\sum_i \sigma^2_{ni} v_i + O(\Upsilon)}{(z + m_{sc} + \sum_i \sigma^2_{ni} v_i + O(\Upsilon))(z + m_{sc})}, \quad 1 \leq n \leq N. \tag{3.82}$$

Using $|G_{ii}| \leq \eta^{-1}$ from (3.80) and $|m_{sc}(z)| = \big|\int \varrho_{sc}(x)/(x-z)\mathrm{d}x\big| \leq \eta^{-1}$, we obtain for $\eta \geq 3$ that

$$\Lambda_d = \max_i |v_i| \leq 2/\eta \leq 2/3. \tag{3.83}$$

By (3.12), we have $|z + m_{sc}(z)| = |m_{sc}(z)|^{-1} \geq 3$. Together with (3.83), we obtain from (3.82) that

$$|v_n| \leq \frac{\max_i |v_i|}{|z + m_{sc}(z)| - \max_i |v_i|} + O(\Upsilon). \tag{3.84}$$

Maximizing over n, we have

$$\Lambda_d = \max_n |v_n| \leq \frac{\Lambda_d}{|z + m_{sc}| - \Lambda_d} + O(\Upsilon). \tag{3.85}$$

Since the denominator satisfies $|z + m_{sc}(z)| - \Lambda_d \geq 3 - 2/3 = 7/3$ by $\Lambda_d \leq 2/3$, the estimate (3.78) follows from (3.85) and (3.77). This completes the proof of Lemma 3.11. □

3.6 Continuity argument : conclusion of the proof of Theorem 3.1

Fix an energy E with $|E| \leq 5$ and choose a decreasing finite sequence $\eta_k \in \mathbf{S}_\ell$, $k = 1, 2, \ldots, k_0$, with $k_0 \leq CN^8$ such that $|\eta_k - \eta_{k+1}| \leq N^{-8}$ and $\eta_1 = 10$, $\eta_{k_0} = N^{-1}(\log N)^{10\ell}$. Denote by $z_k = E + i\eta_k$. We will first show that Theorem 3.1 holds for any $z = z_k$.

Throughout this section fix any $U \geq U_0$ from Lemma 3.10 and recall the definition of $\widetilde{\eta}(U, E)$ from before this lemma. Consider first the case of z_1. Since $\eta_1 \geq \widetilde{\eta}(U, E)$, see (3.69), we are in the first case (3.70) in Lemma 3.10. By Lemma 3.11, we have $\Lambda_d(z_1) + \Lambda_o(z_1) \leq CN^{-1/3}$ in the set $\Theta(z_1)^c$, in particular, $\Theta(z_1)^c \subset \mathbf{B}(z_1)^c$. Moreover, by $\Lambda(z_1) \leq CN^{-1/3}$ in the set $\Theta(z_1)^c$, and (3.65), the second alternative of (3.70) cannot hold and therefore $\Lambda(z_1) \leq U\beta(z_1)$ in the set $\Theta(z_1)^c \cap \Omega(z_1)^c \cap \mathbf{B}(z_1)^c = \Theta(z_1)^c \cap \Omega(z_1)^c$. Using the probability estimates (3.32) and (3.76), we have proved that

$$\mathbb{P}\Big[\Lambda(z_1) \geq U\beta(z_1)\Big] + \mathbb{P}\big(\mathbf{B}(z_1)\big) \leq C \exp\big[-c(\log N)^{\phi\ell}\big]. \tag{3.86}$$

For a general k we have the following:

Lemma 3.12 *There exist constants $0 < \phi < 1$, $C' > 1$, $c > 0$, depending on ϑ, such that if ℓ satisfies $4/\phi \leq \ell \leq C' \log N / \log\log N$ and U is chosen $U \geq U_0(\delta_\pm, C_0)$ (see Lemma 3.10) then the following hold for any $k \leq k_0$ and for any sufficiently large $N \geq N_0(\vartheta, \delta_\pm, C_0, U)$: Case 1. If $\eta_k \geq \widetilde{\eta}(U, E)$, then*

$$\mathbb{P}\Big[\Lambda(z_k) \geq U\beta(z_k)\Big] \leq C'k \exp\big[-c(\log N)^{\phi\ell}\big] \quad \textit{and} \quad \mathbb{P}\big(\mathbf{B}(z_k)\big) \leq C'k \exp\big[-c(\log N)^{\phi\ell}\big]. \tag{3.87}$$

Case 2. If $\eta_k < \widetilde{\eta}(U, E)$, then

$$\mathbb{P}\Big[\Lambda(z_k) \geq C_1(U)\beta(z_k)\Big] \leq C'k \exp\big[-c(\log N)^{\phi\ell}\big] \quad \textit{and} \quad \mathbb{P}\big(\mathbf{B}(z_k)\big) \leq C'k \exp\big[-c(\log N)^{\phi\ell}\big], \tag{3.88}$$

where $C_1(U)$ is given from Lemma 3.10.

Proof. We proceed by induction on k, the case $k = 1$ has been checked in (3.86). First consider Case 1, when $k < k_0$ is such that $\eta_k \geq \widetilde{\eta}(U, E)$, i.e. (3.87) holds by the induction hypothesis. By the definition of the sequence z_k, we have

$$\Big|G_{ij}(z_k) - G_{ij}(z_{k+1})\Big| \leq |z_k - z_{k+1}| \sup_{z \in \mathbf{S}_\ell} \Big|\frac{\partial G_{ij}(z)}{\partial z}\Big| \leq N^{-8} \sup_{z \in \mathbf{S}_\ell} \frac{1}{|\mathfrak{Im}\, z|^2} \leq N^{-6} \tag{3.89}$$

for any i, j. Hence $|\Lambda(z_k) - \Lambda(z_{k+1})| \leq N^{-6} \leq \frac{1}{2}U\beta(z_{k+1})$ and thus

$$\mathbb{P}\Big[\Lambda(z_{k+1}) \geq \frac{3}{2}U\beta(z_{k+1})\Big] \leq C'k \exp\big[-c(\log N)^{\phi\ell}\big]. \tag{3.90}$$

In other words, the estimate on $\Lambda(z_{k+1})$ is deteriorated by a factor 3/2, but it will be gained back by the dichotomy estimate in Lemma 3.10.

Using (3.89) we also have, in $\Omega(z_k)^c \cap \mathbf{B}(z_k)^c$,

$$\begin{aligned}\Lambda_d(z_{k+1}) + \Lambda_o(z_{k+1}) &\leq \Lambda_d(z_k) + \Lambda_o(z_k) + 2N^{-6} \\ &\leq (\log N)^\ell(\Lambda(z_k)^2 + \Psi(z_k)) + \Lambda(z_k) + 2N^{-6} \\ &\leq (\log N)^{2\ell}\sqrt{\frac{U\beta(z_k) + \mathfrak{Im}\, m_{sc}(z_k)}{N\eta_k}} + 2U\beta(z_k) + 2N^{-6}. \end{aligned} \tag{3.91}$$

Here in the second line we used the bounds (3.33) and (3.56) that hold on the set $\Omega(z_k)^c \cap \mathbf{B}(z_k)^c$, in the last line we used $\Lambda(z_k) \le U\beta(z_k) \le (\log N)^{-\ell}$. All these estimates hold on an event with probability at least $1 - C'(k+\frac{1}{2})\exp\big[-c(\log N)^{\phi\ell}\big]$ using (3.32) and the estimate on $\mathbb{P}(\mathbf{B}(z_k))$ from (3.87). Here we assumed that the constant C' is larger than twice the constant C in (3.32).

By the choice of $\ell \ge 4$ and the definition of β from (3.64), the last line of (3.91) is bounded by $(\log N)^{-2}$ and thus we have

$$\mathbb{P}(\mathbf{B}(z_{k+1})) \le C'\Big(k+\frac{1}{2}\Big)\exp\big[-c(\log N)^{\phi\ell}\big]. \tag{3.92}$$

Suppose now that $k+1$ falls into the first case, $\eta_{k+1} \ge \widetilde{\eta}(U,E)$, then, from (3.72),

$$\frac{3}{2}U\beta(z_{k+1}) < \frac{\alpha(z_{k+1})}{U},$$

so by the dichotomy estimate (3.70), $\Lambda(z_{k+1}) \le \frac{3}{2}U\beta(z_{k+1})$ from (3.90) implies $\Lambda(z_{k+1}) \le U\beta(z_{k+1})$ on the set $\Omega(z_{k+1})^c \cap \mathbf{B}(z_{k+1})^c$. Thus (3.32), (3.90) and (3.92) imply that

$$\mathbb{P}\Big[\Lambda(z_{k+1}) \ge U\beta(z_{k+1})\Big] \le C'(k+1)\exp\big[-c(\log N)^{\phi\ell}\big] \tag{3.93}$$

by using $C' \ge 2C$ where C is the constant from (3.32). This proves (3.87), i.e. the induction step if η_{k+1} is in the first case. If η_{k+1} falls into the second case, i.e., $\eta_{k+1} \le \widetilde{\eta}(U,E)$, then (3.90) gives directly the induction step, i.e. (3.88) for $k+1$.

So far we considered Case 1, i.e., we assumed that $\eta_k \ge \widetilde{\eta}(U,E)$. Now consider Case 2, when $\eta_k < \widetilde{\eta}(U,E)$ and therefore the induction hypothesis is (3.88). The argument is very similar to the previous case but $U\beta(z_k)$ is replaced with $C_1(U)\beta(z_k)$ everywhere in (3.90), (3.91) and we still obtain (3.92). Since $\eta_{k+1} < \eta_k \le \widetilde{\eta}(U,E)$, we can directly refer to (3.71) to obtain the induction step, i.e. (3.88) for $k+1$. This completes the proof of Lemma 3.12. □

Choosing a sufficiently large but fixed U, e.g. $U = U_0(\delta_\pm, C_0)$, we have thus proved that $\Lambda(z_k) \le C\beta(z_k)$ for all $k \le k_0$ with a constant depending on $\delta_\pm$ and C_0, in particular $\Psi(z_k) \le C\beta(z_k)$ by the definition of Ψ (3.26). Using (3.33) and (3.56) we have proved Theorem 3.1 for all z_k, $k \le k_0$ and any fixed energy E with $|E| \le 5$. For any $z = E + i\eta \in \mathbf{S}_\ell$ there is a $z_k = E + i\eta_k$ with $|z - z_k| \le N^{-8}$. Using the Lipschitz continuity of $G_{ij}(z)$ and $m_{sc}(z)$ with Lipschitz constant at most N^2, we easily conclude the proof of Theorem 3.1 for any $z \in \mathbf{S}_\ell$. Note that in order to accommodate the higher $(\log N)$-power in β and the additional logarithmic factors in (3.33) and (3.56) with the final formulation of the result in Theorem 3.1, we needed to redefine $\ell \to \ell/3$ which results in a decreased ϕ in the final statement. □

4 Optimal error bound in the strong local semicircle law

We have proved Theorem 3.1 which is weaker than the main result Theorem 2.1 but it will be used as an apriori bound for the improvement. The key ingredient for the stronger result is the following lemma which shows that $[Z]$, the average of Z_i's, is much smaller than the size of typical Z_i. (Notice that in the proof of Theorem 3.1, $[Z]$ was estimated in (3.63) by the same quantity, Ψ, as each individual Z_i.)

For $z \in \mathbf{S}_\ell$ define

$$\Gamma = \Gamma(z) := \Omega_h \cup \mathbf{B}(z), \qquad \Delta = \Delta(z) := \Omega(z) \cup \mathbf{B}(z), \tag{4.1}$$

where Ω_h, Ω were defined in (3.27)–(3.28) and $\mathbf{B}$ was given in (3.19). Recall that Ω_h and Ω depend on ℓ and thus Γ and Δ also depend on ℓ but we omit this fact from the notation. We remark that Theorem 3.1 shows that there exists a positive constant $\phi > 0$ such that for any $4/\phi \leq \ell \leq \log N/\log\log N$ we have

$$\mathbb{P}\big(\mathbf{B}(z)\big) = \mathbb{P}\Big(\Lambda_d(z) + \Lambda_o(z) \geq (\log N)^{-2}\Big) \leq C\exp\big[-c(\log N)^{\phi\ell}\big], \qquad z \in \mathbf{S}_\ell, \tag{4.2}$$

since the error bar $(\log N)^\ell/(N\eta)^{1/3}$ in Theorem 3.1 is much smaller than $(\log N)^{-2}$. Combining (4.2) with (3.32) and $\Gamma \subset \Delta$, we get that

$$\mathbb{P}\big(\Gamma(z)\big) \leq \mathbb{P}\big(\Delta(z)\big) \leq C\exp\big[-c(\log N)^{\phi\ell}\big], \qquad z \in \mathbf{S}_\ell, \tag{4.3}$$

with positive constants C, c depending only on ϑ in (2.17), $\delta_\pm$ from Assumption **(B)** and C_0 from Assumption **(C)**.

With this notation, and recalling that $\Lambda_o(z) = \max_{i\neq j}|G_{ij}(z)|$, we then have the following lemma whose proof will be given separately in Section 7.

Lemma 4.1 *There exist positive constants $D \geq 1$, $A_0 \geq 1$, and $\psi \leq \min\{1/10, \phi\}$, depending on ϑ, such that for any ℓ with*

$$A_0 \log\log N \leq \ell \leq \frac{\log N}{\log\log N}, \tag{4.4}$$

for any $p \leq (\log N)^{\psi\ell-2}$ positive even number and for any fixed $z \in \mathbf{S}_\ell$ we have

$$\mathbb{E}\left[\mathbf{1}\big(\Gamma^c(z)\big)\Big|\frac{1}{N}\sum_{i=1}^N Z_i(z)\Big|^p\right] \leq (Dp)^{Dp}\,\mathbb{E}\Big[\mathbf{1}\big(\Gamma^c(z)\big)\big[\Lambda_o(z)^2 + N^{-1}\big]^p\Big] \tag{4.5}$$

for any sufficiently large $N \geq N_0(A_0, \psi)$.

The first version of this lemma was presented in Lemma 5.2 of [23] where the p-dependence of the constant in (4.5) was not carefully tracked and the effect of the exceptional event Γ was estimated less precisely. This was sufficient since in [23] we applied the result for an exponent p independent of N; as a consequence, in particular, the probability estimates for the local semicircle law were only power law and not subexponential in N as here. In the current paper we allow p to depend on N which requires the more precise form as stated in Lemma 4.1. Furthermore, here we give a new proof that relies on a different organization of partially independent terms. The main difference is that here we separate dependences on individual matrix elements, while in [23] we separated entire rows and columns. The new method is therefore more robust, but combinatorially more demanding.

Recalling the notation

$$[Z] = [Z](z) = \frac{1}{N}\sum_{i=1}^N Z_i(z),$$

we will apply Lemma 4.1 in the following form:

Corollary 4.2 *There exist positive constants $D \geq 1$, $A_0 \geq 1$, and $\psi \leq \min\{1/10, \phi, 1/D\}$, depending on ϑ, such that for any ℓ satisfying* (4.4)*, for any $p \leq (\log N)^{\psi\ell-2}$ positive even number and for any fixed $z \in \mathbf{S}_\ell$* (2.20) *we have for any set Ξ in the probability space*

$$\mathbb{E}\Big[\mathbf{1}\big(\Gamma^c\big)|[Z](z)|^p\Big] \leq \mathbb{E}\Big[\mathbf{1}\big(\Gamma^c \cap \Xi^c\big)\Psi(z)^{2p}\Big] + (Dp)^{Dp}\big[\mathbb{P}(\Omega(z)) + \mathbb{P}(\Xi)\big] \tag{4.6}$$

where $\Omega(z)$ is defined in Lemma 3.6.

Proof. On the right hand side of (4.5) we can split the set Γ^c as

$$\Gamma^c = \Omega_h^c \cap \mathbf{B}^c = \big[\Omega^c \cap \mathbf{B}^c \cap \Xi^c\big] \cup \big[\Omega^c \cap \mathbf{B}^c \cap \Xi\big] \cup \big[(\Omega_h^c \setminus \Omega^c) \cap \mathbf{B}^c\big]$$

On the set $\big[\Omega^c \cap \mathbf{B}^c \cap \Xi\big] \cup \big[(\Omega_h^c \setminus \Omega^c) \cap \mathbf{B}^c\big] \subset \mathbf{B}^c$, we estimate Λ_o trivially by

$$\Lambda_o \le (\log N)^{-2} \le 1. \tag{4.7}$$

Since $\Omega_h^c \setminus \Omega^c \subset \Omega$, we have

$$\mathbb{E}\Big[\mathbf{1}\big(\Gamma^c\big)|[Z](z)|^p\Big] \le (Dp)^{Dp}\mathbb{E}\Big[\mathbf{1}\big(\Omega^c \cap \mathbf{B}^c\big)\big[\Lambda_o(z)^2 + N^{-1}\big]^p\Big] + (Dp)^{Dp}\big[\mathbb{P}(\Xi) + \mathbb{P}(\Omega)\big]. \tag{4.8}$$

Choosing $\psi \le 1/D$ we see that $(Dp)^D \le (\log N)^\ell$. Thus we can use $(\log N)^\ell N^{-1} \le C\Psi^2(z)$ for $z \in \mathbf{S}_\ell$ (by $\mathfrak{Im}\, m_{sc}(z) \ge c\eta$) and that $(\log N)^\ell \Lambda_o^2 \le \Psi^2$ on $\Omega^c \cap \mathbf{B}^c$, see (3.33), to absorb the $(Dp)^{Dp}$ prefactor in the first term in (4.8). This concludes the proof of Corollary 4.2. $\square$

Lemma 4.3 *Fix two numbers ℓ and L that satisfy $4 \le \ell \le L \le \frac{\log(10N)}{10\log\log N}$, in particular $\mathbf{S}_L \subset \mathbf{S}_\ell$, and let $0 < \tau \le 1$ be an arbitrary constant. For any $z = E + i\eta$ define*

$$\gamma = \gamma(z) := \frac{(\log N)^{3\ell+2}}{(N\eta)^\tau}. \tag{4.9}$$

Suppose that for all $z \in \mathbf{S}_L$ we have

$$\Lambda(z) \le \gamma(z) \tag{4.10}$$

and

$$|[Z](z)| \le (\log N)^{3\ell}\left(\frac{\gamma(z) + \mathfrak{Im}\, m_{sc}(z)}{N\eta}\right). \tag{4.11}$$

Suppose that $\Lambda(z) = o(1)$ for $\eta = 10$, $|E| \le 5$. Then in the set $\Omega^c \cap \mathbf{B}^c$ we have

$$\Lambda(z) \le (\log N)^{3\ell+2}(N\eta)^{-(\tau+1)/2} \tag{4.12}$$

for any $z \in \mathbf{S}_L$. Furthermore, if $\Lambda(z) \le \alpha(z)/2$ and (4.11) hold for some $z \in \mathbf{S}_L$, then

$$\Lambda(z) \le C(\log N)^{3\ell+1}\left(\frac{\gamma(z) + \mathfrak{Im}\, m_{sc}(z)}{\alpha(z)N\eta}\right), \tag{4.13}$$

in the set $\Omega^c \cap \mathbf{B}^c$, where α was defined in (3.64).

Proof: In the first part of the proof $z \in \mathbf{S}_L$ is fixed so we drop the z-dependence of various quantities. Recall (3.64), (3.65) and Lemma 3.4 for m_{sc} and $\alpha \sim \sqrt{\kappa + \eta}$. From Lemma 3.9 and using (4.11), in the set $\Omega^c \cap \mathbf{B}^c$ we have, with $w := [v]$, the estimate

$$\frac{(1 - m_{sc}^2)}{m_{sc}^3} w - w^2 = O\left(\frac{|w|^2}{\log N}\right) + O\Big[(\log N)^{3\ell+1}\left(\frac{\gamma + \mathfrak{Im}\, m_{sc}}{N\eta}\right)\Big] \tag{4.14}$$

where we have used (4.10), the definition of Ψ (3.26) and that $|w| = \Lambda$. We can complete the square of the left side and obtain the inequality

$$\Lambda \leq 2\alpha + C(\log N)^{\frac{3\ell+1}{2}} \left(\frac{\gamma + \alpha}{N\eta}\right)^{1/2}, \tag{4.15}$$

where we have used that $\mathfrak{Im}\, m_{sc} \leq C\alpha$. We claim that in fact

$$\Lambda \leq 2\alpha + C(\log N)^{\frac{3\ell+1}{2}} \left(\frac{\gamma}{N\eta}\right)^{1/2} \tag{4.16}$$

also holds; indeed this is trivial if $\Lambda \leq 2\alpha$, and if $\Lambda \geq 2\alpha$ then by assumption (4.10) $\gamma \geq \Lambda \geq 2\alpha$, so α can be absorbed into γ in (4.15).

Define

$$\alpha_0 = \alpha_0(z) := T(\log N)^{\frac{3\ell+1}{2}} \left(\frac{\gamma}{N\eta}\right)^{1/2} = T(\log N)^{3\ell+3/2}(N\eta)^{-\frac{1+\tau}{2}} \tag{4.17}$$

with a large parameter T (independent of N) to be specified later, and note that $\alpha_0 \leq \gamma$ for sufficiently large N.

Suppose that $\Lambda \leq \alpha/2$. In this case the w^2 terms are smaller than the leading term αw in the left hand side of (4.14), therefore we can express $|w| = \Lambda$ and estimate it by

$$\Lambda \leq C(\log N)^{3\ell+1} \left(\frac{\gamma + \mathfrak{Im}\, m_{sc}}{\alpha N\eta}\right) \leq C(\log N)^{3\ell+1} \left(\frac{\gamma}{\alpha N\eta} + \frac{1}{N\eta}\right). \tag{4.18}$$

In the second step also used $\mathfrak{Im}\, m_{sc} \leq C\alpha$. In particular, the first inequality proves (4.13).

Assume now that $\Lambda \leq \alpha/2$ and $\alpha \geq \alpha_0$. Plugging the lower bound (4.17) on α into (4.18) and using the definition of γ we obtain

$$\Lambda \leq CT^{-1}(\log N)^{\frac{3\ell+1}{2}} \left(\frac{\gamma}{N\eta}\right)^{1/2} = CT^{-2}\alpha_0. \tag{4.19}$$

Choosing T as a sufficiently large constant we obtain that

$$\Lambda \leq \frac{\alpha}{4} \tag{4.20}$$

under the condition that $\Lambda \leq \alpha/2$ and $\alpha \geq \alpha_0$. Therefore, as long as $\alpha \geq \alpha_0$, we have a dichotomy: either $\Lambda \geq \alpha/2$ or $\Lambda \leq \alpha/4$.

We now fix E and we continuously decrease η from $\eta = 10$ to $\eta = N^{-1}(\log N)^L$, the lower point in $\mathbf{S}_L$. Since $\Lambda(z) \ll 1$ and $\alpha(z)$ is bounded away from zero for $\eta = 10$, $|E| \leq 5$, we know that $\Lambda \leq \alpha/2$ holds for $\eta = 10$. Since $\Lambda(z)$ is continuous function, by the dichotomy we have that $\Lambda \leq \alpha/4$ for all η as long as $\alpha \geq \alpha_0$. In particular, $\Lambda \leq CT^{-2}\alpha_0$ from (4.19) which proves (4.12) in the case $\alpha \geq \alpha_0$.

Finally, for $\alpha \leq \alpha_0$, we can estimate Λ directly via (4.16) and this proves that

$$\Lambda \leq C(\log N)^{\frac{3\ell+1}{2}} \left(\frac{\gamma}{N\eta}\right)^{1/2} \tag{4.21}$$

from which (4.12) follows and we have thus completed the proof. □

Proof of Theorem 2.1. First we explain the idea. We will prove, by an induction on the exponent τ, that $\Lambda \leq (N\eta)^{-\tau}$ holds modulo logarithmic factors with a high probability. Notice that we proved this statement for $\tau = 1/3$ in Theorem 3.1. Lemma 4.3 asserts that if this statement is true for some τ, then it also holds for $\frac{1+\tau}{2}$ assuming a bound on $[Z]$. This bound can be obtained from Corollary 4.2 with a high probability. Repeating the induction step for $O(\log\log N)$ times, we will obtain that τ is essentially one, i.e. we get Theorem 2.1. However, we have to keep track of the increasing logarithmic factors and the deteriorating probability estimates of the exceptional sets.

Throughout the proof we fix L satisfying (2.18) with the constant A_0 obtained from Corollary 4.2 and we also fix ψ from the same Corollary. We will also use a moving exponent ℓ whose value will always satisfy $L/2 \leq \ell \leq L$, in particular $\mathbf{S}_L \subset \mathbf{S}_\ell$.

We recall the definition

$$\gamma = \gamma(z, \tau, \ell) = \frac{(\log N)^{3\ell+2}}{(N\eta)^\tau}, \tag{4.22}$$

where we now emphasize the dependence on τ and ℓ. Define the events

$$R_{\tau,\ell} := \bigcup_{z \in \mathbf{S}_L} R_{\tau,\ell}(z), \quad R_{\tau,\ell}(z) := \Big\{ \Lambda(z) \geq \gamma(z, \tau, \ell) \Big\}. \tag{4.23}$$

Then (3.1) in Theorem 3.1 states that there is a ψ with $0 < \psi < 1/10$ such that for any $\ell_0 := L$ we have

$$\mathbb{P}\left(R_{\tau,\ell_0}\right) \leq \exp\big[-(\log N)^{\psi\ell_0}\big], \tag{4.24}$$

with $\tau = 1/3$ and for any $N \geq N_0(\vartheta, \delta_\pm, C_0)$. Notice that we have used a weaker form of Theorem 3.1 by making the threshold γ larger, the restrictions for ℓ_0 stronger and reducing the exponent ϕ to ψ since this weaker form will be preserved in the iterative procedure. By setting a sufficiently large lower threshold on N, we could remove the constants C, c from (3.1). The general iteration step is included in the following lemma.

Lemma 4.4 *There exists a sufficiently large $N_0 = N_0(\vartheta, \delta_\pm, C_0)$ such that for any $N \geq N_0$ the following implication holds. If for some $0 < \tau < 1$ and for some ℓ with $L/2 \leq \ell \leq L$*

$$\mathbb{P}\left(R_{\tau,\ell}\right) \leq \exp\big[-(\log N)^{\psi\ell}\big], \tag{4.25}$$

then

$$\mathbb{P}\left(R_{\tau',\ell'}\right) \leq \exp\big[-(\log N)^{\psi\ell'}\big], \tag{4.26}$$

where

$$\tau' = \frac{\tau+1}{2}, \quad \ell' = \ell - \frac{3}{\psi}. \tag{4.27}$$

Proof. Define

$$\Phi = \Phi(z, \tau, \ell) := (\log N)^\ell \sqrt{\frac{\gamma(z, \tau, \ell) + \mathfrak{Im}\, m_{sc}(z)}{N\eta}}. \tag{4.28}$$

Fix $z \in \mathbf{S}_L$, then from Corollary 4.2 with the choice of $\Xi = R_{\tau,\ell}$ we have

$$\mathbb{E}\Big[\mathbf{1}\big(\Gamma^c\big)|[Z]|^p\Big] \leq \mathbb{E}\Big[\mathbf{1}\big(R^c_{\tau,\ell}\big)\Psi^{2p}\Big] + (Dp)^{Dp}\exp\Big[-c(\log N)^{\psi\ell}\Big] \tag{4.29}$$

$$\leq \Phi^{2p} + (Dp)^{Dp}\exp\Big[-c(\log N)^{\psi\ell}\Big], \tag{4.30}$$

where we have used (4.25) and (3.32) to bound the probability of Ξ and Ω and we used that $\Lambda \le \gamma$ on $R^c_{\tau,\ell}$ to estimate $\Psi \le \Phi$. We will choose $p = (\log N)^a$ with

$$a = \psi\ell - 3. \tag{4.31}$$

From Markov's inequality and (4.3) we obtain that

$$\begin{aligned}\mathbb{P}\Big(|[Z]| \ge \frac{1}{2}(\log N)\Phi^2\Big) &\le 2^p(\log N)^{-p}\Phi^{-2p}\Big[\Phi^{2p} + (Dp)^{Dp}\exp\big[-(\log N)^{\psi\ell}\big]\Big] + C\exp[-c(\log N)^{\phi\ell}]\\ &\le 2^p(\log N)^{-p} + \exp\big[Dp\log(2Dp) + p(\log N) - (\log N)^{a+3}\big] + C\exp[-c(\log N)^{\phi\ell}]\\ &\le \exp\Big[-3(\log N)^a\Big]. \end{aligned}\tag{4.32}$$

Here in the second line we used $\Phi \ge N^{-1/2}$ from $\mathfrak{Im}\, m_{sc}(z) \ge c\eta$ to estimate Φ^{-2p}. In the final estimate we used that $\log p = a\log\log N \le \psi\ell\log\log N \le \psi\log N$ and that $\psi \le \phi$. This estimate was for any fixed $z \in \mathbf{S}_L$. By choosing a grid of z-values in $\mathbf{S}_L$ with spacing of order N^{-c}, with some large c, we can use the Lipschitz continuity of $[Z](z)$ and $\Phi(z)$ to conclude that essentially the same estimate holds simultaneously for all $z \in \mathbf{S}_L$.

Combining this with (4.25), we have

$$|[Z]| \le (\log N)\Phi^2 \le (\log N)^{3\ell}\Big(\frac{\gamma + \mathfrak{Im}\, m_{sc}}{N\eta}\Big) \qquad \text{and} \qquad \Lambda \le \gamma, \tag{4.33}$$

for all $z \in \mathbf{S}_L$ with a probability at least $1 - \exp\big[-2(\log N)^a\big]$. We can now apply Lemma 4.3 so that

$$\Lambda(z) \le (\log N)^{3\ell+2}(N\eta)^{-(\tau+1)/2} \tag{4.34}$$

hold for any $z \in \mathbf{S}_L$ with a probability bigger than $1 - \exp\big[-(\log N)^a\big]$. Here we have used that $\mathbb{P}(\Omega \cup \mathbf{B}) \le \exp\big[-2(\log N)^a\big]$ from (4.3). We have thus proved (4.26) and Lemma 4.4. □

Returning to the proof of Theorem 2.1, we choose $\tau_0 = 1/3$ and $\ell_0 = L$ as the initial values of the iteration. The input condition (4.25) in Lemma 4.4 for the initial step has been checked in (4.24). Iterating Lemma 4.4 yields a sequence of (τ_n, ℓ_n) so that $\tau_{n+1} = \tau_n'$ and $\ell_{n+1} = \ell_n'$ via (4.27), more precisely

$$\tau_n = 1 - 2^{-n}\cdot\frac{2}{3} \ge 1 - 2^{-n}, \qquad \ell_n = L - 3n/\psi,$$

such that

$$\mathbb{P}\Big(\bigcup_{z\in\mathbf{S}_L}\Big\{\Lambda(z) \ge \frac{(\log N)^{3\ell_n+2}}{(N\eta)^{1-2^{-n}}}\Big\}\Big) \le \exp\big[-(\log N)^{\psi\ell_n}\big]. \tag{4.35}$$

We run the iteration until $n = 2\log\log N$ so that

$$(N\eta)^{2^{-n}} \le N^{2^{-n}} \le e.$$

If $A_0 = 20/\psi$, i.e. $L \ge (20/\psi)\log\log N$, then $\ell_n \ge 2L/3$ and thus

$$\mathbb{P}\Big(\bigcup_{z\in\mathbf{S}_L}\Big\{\Lambda(z) \ge \frac{e(\log N)^{3L+2}}{N\eta}\Big\}\Big) \le \exp\big[-(\log N)^{2\psi L/3}\big]. \tag{4.36}$$

This proves (2.19) after renaming $2\psi/3$ to a new ϕ. The proof of (2.21) follows from the estimate on Λ, from (3.33), (3.56) and (4.3).

Finally, to prove (2.22), we need the following Lemma.

Lemma 4.5 *Let $L \geq 4$ satisfy (2.18) and define the set*

$$\mathbf{U}_L := \Big\{ z = E + i\eta \; : \; 5 \geq |E| \geq 2 + N^{-2/3}(\log N)^{8L+8}, \quad \eta = N^{-2/3}(\log N)^{2L+1} \Big\}. \tag{4.37}$$

Then for A_0 large enough in (2.18), we have

$$\mathbb{P}\Big(\bigcup_{z \in \mathbf{U}_L} \Big\{ \Lambda(z) \leq (\log N)^{-1}(N\eta)^{-1} \Big\} \Big) \geq 1 - C \exp\big[- c(\log N)^{\psi L/2} \big]. \tag{4.38}$$

Proof of Lemma 4.5: For $z \in \mathbf{U}_L$ we have $\kappa = N^{-2/3}(\log N)^{8L+8} \geq \eta$ and thus we have (see (3.65))

$$\alpha(z) \geq c\sqrt{\kappa + \eta} \geq cN^{-1/3}(\log N)^{4L+4}.$$

Therefore $\Lambda \leq \alpha/2$ holds on the event $\Lambda(z) \leq \frac{e(\log N)^{3L+2}}{N\eta}$ for any $z \in \mathbf{U}_L$. Since $\mathbf{U}_L \subset \mathbf{S}_L$, the probability of this event is bigger than $1 - \exp\big[- (\log N)^{2\psi L/3} \big]$ by (4.36). Combining this bound on Λ with the estimate (3.32) for $\ell = L$, we know that

$$|[Z](z)| \leq (\log N)^{2L} \frac{\gamma(z) + \mathfrak{Im}\, m_{sc}(z)}{N\eta}$$

holds with a probability bigger than $1 - 2\exp\big[- (\log N)^{2\psi L/3} \big]$. Here we used $\gamma(z) = (\log N)^{3L+2}(N\eta)^{-1}$ with the choice of $\tau = 1$ and $\ell = L$, see (4.22).

We can now use (4.13) from Lemma 4.3 with $\ell = L$ and $\tau = 1$ to have

$$\Lambda \leq C(\log N)^{3L+1} \left(\frac{(\log N)^{3L+2}(N\eta)^{-1} + \frac{\eta}{\sqrt{\kappa}}}{\sqrt{\kappa} N\eta} \right) \tag{4.39}$$

with probability larger than $1 - 3\exp\big[- (\log N)^{2\psi L/3} \big]$. Here we used the probability estimate (4.3) on $P(\Omega \cup \mathbf{B})$ and the first bound in (3.14). Then using the values of κ and η in the set (4.37), we obtain

$$\Lambda \leq (\log N)^{-1}(N\eta)^{-1}$$

from (4.39) and this proves Lemma 4.5. □

We now prove (2.22). On the set $\mathbf{U}_L$ we have

$$\mathfrak{Im}\, m_{sc} = O(\frac{\eta}{\sqrt{\kappa}}) \leq (\log N)^{-1}(N\eta)^{-1}. \tag{4.40}$$

Combining it with (4.38), we obtain that

$$\mathbb{P}\Big(\bigcup_{z \in \mathbf{U}_L} \Big\{ \mathfrak{Im}\, m(z) \leq 2(\log N)^{-1}(N\eta)^{-1} \Big\} \Big) \geq 1 - C \exp\big[- c(\log N)^{\psi L/2} \big]. \tag{4.41}$$

Fix $z = E + i\eta \in \mathbf{U}_L$ and define the event

$$W(z) := \{ \exists j \; : \; |\lambda_j - E| \leq \eta \}.$$

Recalling the definition of m,

$$\mathfrak{Im}\, m(z) = \frac{1}{N} \sum_{j=1}^{N} \frac{\eta}{(E - \lambda_j)^2 + \eta^2}, \tag{4.42}$$

it is clear that $\mathfrak{Im}\, m(z) \geq \frac{1}{4}(N\eta)^{-1}$ on the set $W(z)$. Using (4.41) we obtain that

$$P\Big(\exists j \ :\ 2 + N^{-2/3}(\log N)^{8L+8} \leq |\lambda_j| \leq 5\Big) \leq C \exp\big[-c(\log N)^{\psi L/2}\big]. \tag{4.43}$$

Finally, we need to control the probability of a very large eigenvalue. For example, the following (not optimal) estimate was proved in, e..g, Lemma 7.2 of [22]. We formulate the results for the largest eigenvalue λ_N, but analogous results hold for the smallest eigenvalue λ_1 as well.

Lemma 4.6 *Let H satisfy Assumptions* **(A), (B), (C)** *and the subexponential decay condition* (2.17)*. Then for some $\varepsilon > 0$, depending on ϑ, we have*

$$\mathbb{P}(\lambda_N \geq K) \leq e^{-N^{\varepsilon} \log K} \tag{4.44}$$

for any $K \geq 3$.

Combining this lemma with (4.43) we completed the proof of (2.22). □

5 Estimates on the location of eigenvalues

Proof of Theorem 2.2. We now translate the information on the Stieltjes transform obtained in Theorem 2.1 to prove Theorem 2.2 on the location of the eigenvalues. We will need the following Lemma 5.1 which is a special case of Lemma 6.1 proved in [23] with the choice $A = 0$. The conditions (6.1) and (6.2) stated in Lemma 6.1 of [23] are not sufficient. Instead, the following slightly stronger assumption is necessary:

$$|m^{\Delta}(x + iy)| \leq \frac{CU}{y(\kappa_x + y)^A}, \qquad \text{for} \qquad 1 \geq y > 0, \quad |x| \leq K + 1, \tag{5.1}$$

i.e., it is not sufficient to control only the imaginary part of m^{Δ}. This stronger condition is needed in (6.7) of [23], where the imaginary part of m is changed to its real part after an integration by parts. With the condition (5.1), the proof of Lemma 6.1 in [23] remains otherwise unchanged. This immediately proves the following lemma as a special case:

Lemma 5.1 *Let ϱ^{Δ} be a signed measure on the real line with supp $\varrho^{\Delta} \subset [-K, K]$ for some fixed constant K. For any $E_1, E_2 \in [-3, 3]$ and $\eta > 0$ we define $f(\lambda) = f_{E_1,E_2,\eta}(\lambda)$ to be a characteristic function of $[E_1, E_2]$ smoothed on scale η, i.e., $f \equiv 1$ on $[E_1, E_2]$, $f \equiv 0$ on $\mathbb{R} \setminus [E_1 - \eta, E_2 + \eta]$ and $|f'| \leq C\eta^{-1}$, $|f''| \leq C\eta^{-2}$. Let m^{Δ} be the Stieltjes transform of ϱ^{Δ}. Suppose for some positive number U (may depend on N) we have*

$$|m^{\Delta}(x + iy)| \leq \frac{CU}{Ny} \qquad \textit{for} \qquad 1 \geq y > 0, \quad |x| + y \leq K. \tag{5.2}$$

Then

$$\Big|\int_{\mathbb{R}} f_{E_1,E_2,\eta}(\lambda)\varrho^{\Delta}(\lambda)\mathrm{d}\lambda\Big| \leq \frac{CU|\log \eta|}{N} \tag{5.3}$$

with some constant C depending on K. □

We will apply this lemma with the choice that the signed measure is the difference of the empirical density and the semicircle law,

$$\varrho^{\Delta}(\mathrm{d}\lambda) = \varrho(\mathrm{d}\lambda) - \varrho_{sc}(\lambda)\mathrm{d}\lambda, \qquad \varrho(\mathrm{d}\lambda) := \frac{1}{N}\sum_i \delta(\lambda_i - \lambda).$$

First we prove (2.26). Choose $L := A_0 \log\log N$, where A_0 is given in Theorem 2.1, and we define

$$T_N := (\log N)^L = (\log N)^{A_0 \log\log N}$$

for simplicity. By Theorem 2.1, the assumptions of Lemma 5.1 hold for the difference $m^{\Delta} = m - m_{sc}$ with $K = 10$ and $U = T_N^4$ if $y \geq y_0 := T_N^{10}/N$. For $y \leq y_0$, set $z = x + iy$, $z_0 = x + iy_0$ and estimate

$$|m(z) - m_{sc}(z)| \leq |m(z_0) - m_{sc}(z_0)| + \int_y^{y_0} \big|\partial_\eta\big(m(x+i\eta) - m_{sc}(x+i\eta)\big)\big|\mathrm{d}\eta. \tag{5.4}$$

Note that

$$\begin{aligned} |\partial_\eta m(x+i\eta)| = & \Big|\frac{1}{N}\sum_j \partial_\eta G_{jj}(x+i\eta)\Big| & (5.5)\\ \leq & \frac{1}{N}\sum_{jk} |G_{jk}(x+i\eta)|^2 = \frac{1}{N\eta}\sum_j \mathfrak{Im}\, G_{jj}(x+i\eta) = \frac{1}{\eta}\mathfrak{Im}\, m(x+i\eta), & (5.6)\end{aligned}$$

and similarly

$$|\partial_\eta m_{sc}(x+i\eta)| = \Big|\int \frac{\varrho_{sc}(s)}{(s-x-i\eta)^2}\mathrm{d}s\Big| \leq \int \frac{\varrho_{sc}(s)}{|s-x-i\eta|^2}\mathrm{d}s = \frac{1}{\eta}\mathfrak{Im}\, m_{sc}(x+i\eta).$$

Now we use the fact that the functions $y \to y\mathfrak{Im}\, m(x+iy)$ and $y \to y\mathfrak{Im}\, m_{sc}(x+iy)$ are monotone increasing for any $y > 0$ since both are Stieltjes transforms of a positive measure. Therefore the integral in (5.4) can be bounded by

$$\int_y^{y_0} \frac{\mathrm{d}\eta}{\eta}\big[\mathfrak{Im}\, m(x+i\eta) + \mathfrak{Im}\, m_{sc}(x+i\eta)\big] \leq y_0\big[\mathfrak{Im}\, m(x+iy_0) + \mathfrak{Im}\, m_{sc}(x+iy_0)\big]\int_y^{y_0}\frac{\mathrm{d}\eta}{\eta^2} \tag{5.7}$$

By definition, $\mathfrak{Im}\, m_{sc}(x+iy_0) \leq |m_{sc}(x+iy_0)| \leq C$. By the choice of y_0 and Theorem 2.1, we have

$$\mathfrak{Im}\ m(x+iy_0) \leq \mathfrak{Im}\, m_{sc}(x+iy_0) + \frac{T_N^4}{Ny_0} \leq C \tag{5.8}$$

with very high probability. Together with (5.7) and (5.4), this proves that (5.2) holds for $y \leq y_0$ as well if U is increased to $U = T_N^{10}$.

The application of Lemma 5.1 shows that for any $\eta \geq 1/N$

$$\Big|\int_{\mathbb{R}} f_{E_1,E_2,\eta}(\lambda)\varrho(\lambda)\mathrm{d}\lambda - \int_{\mathbb{R}} f_{E_1,E_2,\eta}(\lambda)\varrho_{sc}(\lambda)\mathrm{d}\lambda\Big| \leq \frac{C(\log N)T_N^{10}}{N}. \tag{5.9}$$

With the fact: $y \to y\mathfrak{Im}\, m(x+iy)$ is monotone increasing for any $y > 0$, (5.8) implies a crude upper bound on the empirical density. Indeed, for any interval $I := [x-\eta, x+\eta]$, with $\eta = 1/N$, we have

$$\mathfrak{n}(x+\eta) - \mathfrak{n}(x-\eta) \leq C\eta\, \mathfrak{Im}\ m\big(x+i\eta\big) \leq Cy_0\, \mathfrak{Im}\ m\big(x+iy_0\big) \leq \frac{CT_N^{10}}{N}. \tag{5.10}$$

This bound can be used to estimate the difference between the characteristic function of the interval $[E_1, E_2]$ and the smoothed function $f_{E_1,E_2,\eta}$.

Since the probability to have eigenvalues outside the interval $[-3, 3]$ are extremely small, we consider only the case that all eigenvalues are inside $[-3, 3]$. Let $E_1 = -4$ and $E_2 := E \in [-3, 3]$. Then from (5.9) and (5.10) we have that

$$\Big|\mathfrak{n}(E) - n_{sc}(E)\Big| \leq \frac{C(\log N)T_N^{10}}{N} \tag{5.11}$$

holds for any fixed $E \in [-3, 3]$ with an overwhelming probability. The supremium over E is a standard argument for extremely small events and we omit the details. This completes the proof of (2.26) after possibly increasing L (hence A_0) and decreasing ϕ in order to replace the $(\log N)T_N^{10}$ with $(\log N)^L$.

Now we turn to the proof of (2.25). Let L as before. Fix any $1 \leq j \leq N/2$ and let $E = \gamma_j$, $E' = \lambda_j$. Setting $t_N = (\log N)T_N^{10} = (\log N)^{10L+1}$ for simplicity, from (5.11) we have

$$n_{sc}(E) = \mathfrak{n}(E') = n_{sc}(E') + O(t_N/N). \tag{5.12}$$

Clearly $E \leq 1$, and using (5.11) $E' \leq 1$ also holds with an overwhelming probability. First, using (2.22) and

$$n_{sc}(x) \sim (x + 2)^{3/2}, \;\text{ for }\; -2 \leq x \leq 1, \tag{5.13}$$

i.e.

$$n_{sc}(E) = n_{sc}(\gamma_j) = \frac{j}{N} \sim (E + 2)^{3/2},$$

we know that (2.25) holds (with a possibly increased power of $\log N$ in the left hand side) if

$$E, E' \leq -2 + t_N N^{-2/3}. \tag{5.14}$$

The correct power $(\log N)^L$ can be restored by increasing L (hence A_0) and decreasing ϕ, as before.

Hence, we can assume that one of E and E' is in the interval $[-2 + t_N N^{-2/3}, 1]$. With (5.13), this assumption implies that at least one of $n_{sc}(E)$ and $n_{sc}(E')$ is larger than $t_N^{3/2}/N$. Inserting this information into (5.12), we obtain that both $n_{sc}(E)$ and $n_{sc}(E')$ are positive and

$$n_{sc}(E) = n_{sc}(E')\big[1 + O(t_N^{-1/2})\big],$$

in particular, $E+2 \sim E'+2$. Using that $n'_{sc}(x) \sim (x+2)^{1/2}$ for $-2 \leq x \leq 1$, we obtain that $n'_{sc}(E) \sim n'_{sc}(E')$, and in fact $n'_{sc}(E)$ is comparable with $n'_{sc}(E'')$ for any E'' between E and E'. Then with Taylor's expansion, we have

$$|n_{sc}(E') - n_{sc}(E)| \leq C|n'_{sc}(E)||E' - E|. \tag{5.15}$$

Since $n'_{sc}(E) = \rho_{sc}(E) \sim \sqrt{\kappa}$ and $n_{sc}(E) \sim \kappa^{3/2}$, moreover, by $E = \gamma_j$ we also have $n_{sc}(E) = j/N$, we obtain from (5.12) and (5.15) that

$$|E' - E| \leq \frac{C|n_{sc}(E') - n_{sc}(E)|}{n'_{sc}(E)} \leq \frac{Ct_N}{Nn'_{sc}(E)} \leq \frac{Ct_N}{N(n_{sc}(E))^{1/3}} \leq \frac{Ct_N}{N^{2/3}j^{1/3}},$$

which proves (2.25), again, after increasing L and decreasing ϕ to achieve the claimed $(\log N)^L$ prefactor. This concludes the proof of Theorem 2.2.

□

6 Edge Universality

In this section, we prove the edge universality, i.e., Theorem 2.4. At the end of Section 6.1 we will give a heuristic explanation why matching the second moments is sufficient but we first need some preparation and to introduce various notations. We will consider the largest eigenvalue λ_N, but the same argument applies to the lowest eigenvalue λ_1 as well.

For any $E_1 \le E_2$ let

$$\mathcal{N}(E_1, E_2) := \#\{E_1 \le \lambda_j \le E_2\}$$

denote the number of eigenvalues in $[E_1, E_2]$. By Theorem 2.2 (rigidity of eigenvalues), there exist positive constants A_0, ϕ, C and $c > 0$, depending only on ϑ, $\delta_\pm$ and C_0 such that with setting

$$L := A_0 \log\log N \tag{6.1}$$

we have

$$\mathbb{P}\left\{ \left| N^{2/3}(\lambda_N - 2) \right| \ge (\log N)^L \right\} \le C \exp\left[- c(\log N)^{\phi L} \right] \tag{6.2}$$

and

$$\mathbb{P}\left\{ \mathcal{N}\left(2 - \frac{2(\log N)^L}{N^{2/3}},\ 2 + \frac{2(\log N)^L}{N^{2/3}} \right) \ge (\log N)^L \right\} \le C \exp\left[- c(\log N)^{\phi L} \right] \tag{6.3}$$

for sufficiently large $N \ge N_0(\vartheta, \delta_\pm, C_0)$. These estimates hold for both the $\mathbf{v}$ and $\mathbf{w}$ ensembles. Using these estimates, we can assume that s in (2.41) satisfies

$$-(\log N)^L \le s \le (\log N)^L. \tag{6.4}$$

With L from (6.1), we set

$$E_L := 2 + 2(\log N)^L N^{-2/3}. \tag{6.5}$$

For any $E \le E_L$ let

$$\chi_E := \mathbf{1}_{[E, E_L]}$$

be the characteristic function of the interval $[E, E_L]$. For any $\eta > 0$ we define

$$\theta_\eta(x) := \frac{\eta}{\pi(x^2 + \eta^2)} = \frac{1}{\pi} \mathfrak{Im} \frac{1}{x - i\eta} \tag{6.6}$$

to be an approximate delta function on scale η. In the following elementary lemma we compare the sharp counting function $\mathcal{N}(E, E_L) = \operatorname{Tr} \chi_E(H)$ by its approximation smoothed on scale η.

Lemma 6.1 *Suppose that the assumptions of Theorem 2.4 hold and L, ϕ satisfy* (6.2) *and* (6.3)*. For any $\varepsilon > 0$, set $\ell_1 := N^{-2/3-3\varepsilon}$ and $\eta := N^{-2/3-9\varepsilon}$. Then there exist constants C, c such that for any E satisfying*

$$|E - 2| N^{2/3} \le \frac{3}{2} (\log N)^L \tag{6.7}$$

we have

$$\mathbb{P}\Big\{ |Tr\, \chi_E(H) - Tr\, \chi_E * \theta_\eta(H)| \le C \left(N^{-2\varepsilon} + \mathcal{N}(E - \ell_1, E + \ell_1) \right) \Big\} \ge 1 - C \exp[-c(\log N)^{\phi L}] \tag{6.8}$$

for sufficiently large N. This estimate holds for both the $\mathbf{v}$ and $\mathbf{w}$ ensembles.

Proof of Lemma 6.1. By (6.5) and (6.7) we have

$$\eta \ll \ell_1 \ll E_L - E \le CN^{-2/3}(\log N)^L. \tag{6.9}$$

Since χ_E is the characteristic function of $[E, E_L]$, for any $x \in \mathbb{R}$, we have

$$|\chi_E(x) - \chi_E * \theta_\eta(x)| = \Big| \Big(\int_{\mathbb{R}} \chi_E(x) - \int_{E-x}^{E_L - x} \Big) \theta_\eta(y) \mathrm{d}y \Big|.$$

Let $d = d(x) := |x - E| + \eta$ and $d_L = d_L(x) := |x - E_L| + \eta$. Using that $\int \theta_\eta = 1$ and the estimate

$$\int_\alpha^\infty \theta_\eta(y) \mathrm{d}y = \frac{1}{\pi} \int_\alpha^\infty \frac{\eta}{y^2 + \eta^2} \mathrm{d}y \le \frac{C\eta}{\alpha + \eta}, \qquad \alpha > 0,$$

an elementary calculation shows that

$$|\chi_E(x) - \chi_E * \theta_\eta(x)| \le C\eta \Big[\frac{E_L - E}{d_L(x) d(x)} + \frac{\chi_E(x)}{d_L(x) + d(x)} \Big] \tag{6.10}$$

for some constant $C > 0$. It is easy to check that if $\min\{d, d_L\} \le \ell_1$, then the right side of (6.10) is bounded by a constant and if $\min\{d, d_L\} \ge \ell_1$, then it is less than $O(\eta/\ell_1) = O(N^{-6\varepsilon})$. Hence we have

$$|\mathrm{Tr}\, \chi_E(H) - \mathrm{Tr}\, \chi_E * \theta_\eta(H)| \le C \left(\mathrm{Tr}\, f(H) + \frac{\eta}{\ell_1} \, \mathcal{N}(E, E_L) + \mathcal{N}(E - \ell_1, E + \ell_1) + \mathcal{N}(E_L - \ell_1, \infty) \right), \tag{6.11}$$

where

$$f(x) := \frac{\eta(E_L - E)}{d_L(x) d(x)} \mathbf{1}(x \le E - \ell_1). \tag{6.12}$$

With the assumption (6.7), $\mathcal{N}(E, E_L)$ and $\mathcal{N}(E_L - \ell_1, \infty)$ can be bounded by using (6.3) and (6.2). Hence it follows from (6.11) that

$$|\mathrm{Tr}\, \chi_E(H) - \mathrm{Tr}\, \chi_E * \theta_\eta(H)| \le C \left(\mathrm{Tr}\, f(H) + \mathcal{N}(E - \ell_1, E + \ell_1) + N^{-5\varepsilon} \right) \tag{6.13}$$

holds with a probability larger than $1 - C\exp[-c(\log N)^{\phi L}]$, for some constants C and c and for sufficiently large N, uniformly in E with (6.7). Set

$$g(y) := \frac{1}{y^2 + \ell_1^2}, \tag{6.14}$$

and notice that

$$\frac{1}{a^2} \le C(g * \theta_{\ell_1})(a) \qquad \text{if } |a| \ge \ell_1, \tag{6.15}$$

which implies

$$\frac{f(x)}{\eta(E_L - E)} = \frac{\mathbf{1}(x \le E - \ell_1)}{d_L(x) d(x)} \le \frac{C \cdot \mathbf{1}(x \le E - \ell_1)}{|E - x|^2} \le C(g * \theta_{\ell_1})(E - x). \tag{6.16}$$

Recalling from (2.11) and (6.6) that

$$\frac{1}{N} \mathrm{Tr}\, \theta_{\ell_1}(H - E) = \frac{1}{\pi N} \mathfrak{Im}\, \mathrm{Tr}\, \frac{1}{H - E - i\ell_1} = \frac{1}{\pi} \mathfrak{Im}\, m(E + i\ell_1),$$

we obtain

$$\begin{aligned}\operatorname{Tr} f(H) &\le CN\eta(E_L - E)\int_{\mathbb{R}} \frac{1}{y^2+\ell_1^2}\mathfrak{Im}\, m(E-y+i\ell_1)\mathrm{d}y \\ &\le CN^{1/3}\eta(\log N)^L \int_{\mathbb{R}} \frac{1}{y^2+\ell_1^2}\Big[\mathfrak{Im}\, m_{sc}(E-y+i\ell_1) + \frac{(\log N)^{CL}}{N\ell_1}\Big]\mathrm{d}y, \end{aligned} \tag{6.17}$$

where, by (2.19), the second inequality holds with a probability larger than $1-C\exp[-c(\log N)^{\phi L}]$ and we also used (6.9). The integral of the second term in the r.h.s is bounded by

$$CN^{1/3}\eta(\log N)^L\int_{\mathbb{R}}\frac{1}{y^2+\ell_1^2}\frac{(\log N)^{CL}}{N\ell_1}\mathrm{d}y \le N^{-2/3}\eta(\log N)^{CL}\ell_1^{-2} \le N^{-2\varepsilon}, \tag{6.18}$$

by using the definitions of ℓ_1 and η.

For the first term in the r.h.s of (6.17) we use the elementary estimate

$$\mathfrak{Im}\, m_{sc}(E-y+i\ell_1) \le C\sqrt{\ell_1 + \big||E-y|-2\big|}.$$

The integral in the region

$$A := \big\{\big||E-y|-2\big| \ge \ell_1\big\}$$

can be bounded by

$$\int_A \frac{\mathfrak{Im}\, m_{sc}(E-y+i\ell_1)}{y^2+\ell_1^2}\mathrm{d}y \le C\int_A \frac{\big||E-y|-2\big|^{1/2}}{y^2+\ell_1^2}\mathrm{d}y \le C\int_{\mathbb{R}}\frac{|y|^{1/2}+|E-2|^{1/2}}{y^2+\ell_1^2}\mathrm{d}y \le C\Big(\frac{1}{\sqrt{\ell_1}}+\frac{|E-2|^{1/2}}{\ell_1}\Big).$$

On the complementary region we have

$$\int_{A^c}\frac{1}{y^2+\ell_1^2}\mathfrak{Im}\, m_{sc}(E-y+i\ell_1)\mathrm{d}y \le C\sqrt{\ell_1}\int_{A^c}\frac{1}{y^2+\ell_1^2}\mathrm{d}y \le C\ell_1^{-1/2}.$$

Combining these estimates and using (6.7) together with the definitions of ℓ_1 and η we get

$$CN^{1/3}\eta(\log N)^L\int_{\mathbb{R}}\frac{1}{y^2+\ell_1^2}\mathfrak{Im}\, m_{sc}(E-y+i\ell_1)\mathrm{d}y \le N^{-2\varepsilon},$$

and therefore, together with (6.18), we have $\operatorname{Tr} f(H) \le 2N^{-2\varepsilon}$. Considering (6.13), we have thus proved Lemma 6.1. □

Let $q : \mathbb{R} \to \mathbb{R}_+$ be a smooth cutoff function such that

$$q(x) = 1 \quad \text{if} \quad |x| \le 1/9, \qquad q(x) = 0 \quad \text{if} \quad |x| \ge 2/9,$$

and we assume that $q(x)$ is decreasing for $x \ge 0$.

Corollary 6.2 *Suppose the assumptions of Lemma 6.1 hold and E satisfies*

$$|E-2|N^{2/3} \le (\log N)^L. \tag{6.19}$$

Let $\ell := \frac{1}{2}\ell_1 N^{2\varepsilon} = \frac{1}{2}N^{-2/3-\varepsilon}$. Then the inequality

$$Tr\,\chi_{E+\ell} * \theta_\eta(H) - N^{-\varepsilon} \le \mathcal{N}(E,\infty) \le Tr\,\chi_{E-\ell} * \theta_\eta(H) + N^{-\varepsilon} \tag{6.20}$$

holds with a probability bigger than $1 - C\exp[-c(\log N)^{\phi L}]$. Furthermore, we have

$$\mathbb{E}\, q\,(Tr\,\chi_{E-\ell} * \theta_\eta(H)) \le \mathbb{P}(\mathcal{N}(E,\infty) = 0) \le \mathbb{E}\, q\,(Tr\,\chi_{E+\ell} * \theta_\eta(H)) + C\exp\big[-c(\log N)^{\phi L}\big] \tag{6.21}$$

for sufficiently large N independent of E as long as (6.19) *holds. Notice that the directions in the inequalities* (6.20) *and* (6.21) *are opposite since q is decreasing for positive arguments.*

Proof. For any E satisfying (6.19) we have $E_L - E \gg \ell$ thus $|E - 2 - \ell| N^{2/3} \le \frac{3}{2}(\log N)^L$ (see (6.7)), therefore (6.8) holds for E replaced with $y \in [E-\ell, E]$ as well. We thus obtain

$$\begin{aligned}
\mathrm{Tr}\,\chi_E(H) &\le \ell^{-1}\int_{E-\ell}^{E} \mathrm{d}y\ \mathrm{Tr}\,\chi_y(H) \\
&\le \ell^{-1}\int_{E-\ell}^{E} \mathrm{d}y\ \mathrm{Tr}\,\chi_y * \theta_\eta(H) + C\ell^{-1}\int_{E-\ell}^{E} \mathrm{d}y\,\big[N^{-2\varepsilon} + \mathcal{N}(y-\ell_1, y+\ell_1)\big] \\
&\le \mathrm{Tr}\,\chi_{E-\ell} * \theta_\eta(H) + CN^{-2\varepsilon} + C\frac{\ell_1}{\ell}\mathcal{N}(E-2\ell, E+\ell)
\end{aligned}$$

with a probability larger than $1 - C\exp[-c(\log N)^{\phi L}]$. From (2.26), (6.19), $\ell_1/\ell = 2N^{-2\varepsilon}$ and $\ell \le N^{-2/3}$, we can bound

$$\frac{\ell_1}{\ell}\mathcal{N}(E-2\ell, E+\ell) \le N^{1-2\varepsilon}\int_{E-2\ell}^{E+\ell} \varrho_{sc}(x)\mathrm{d}x + N^{-2\varepsilon}(\log N)^{L_1} \le \frac{1}{2}N^{-\varepsilon}$$

with a very high probability, where we estimated the explicit integral using that the integration domain is in a $CN^{-2/3}(\log N)^L$-vicinity of the edge at 2. We have thus proved

$$\mathcal{N}(E, E_L) = \mathrm{Tr}\,\chi_E(H) \le \mathrm{Tr}\,\chi_{E-\ell} * \theta_\eta(H) + N^{-\varepsilon}.$$

By (6.2), we can replace $\mathcal{N}(E, E_L)$ by $\mathcal{N}(E,\infty)$ with a change of probability of at most $C\exp[-c(\log N)^{\phi L}]$. This proves the upper bound of (6.20) and the lower bound can be proved similarly.

On the event that (6.20) holds, the condition $\mathcal{N}(E,\infty) = 0$ implies that $\mathrm{Tr}\,\chi_{E+\ell} * \theta_\eta(H) \le 1/9$. Thus we have

$$\mathbb{P}\,(\mathcal{N}(E,\infty) = 0) \le \mathbb{P}\,(\mathrm{Tr}\,\chi_{E+\ell} * \theta_\eta(H) \le 1/9) + C\exp[-c(\log N)^{\phi L}]. \tag{6.22}$$

Together with the Markov inequality, this proves the upper bound in (6.21). For the lower bound, we use

$$\mathbb{E}\, q\big(\mathrm{Tr}\,\chi_{E-\ell} * \theta_\eta(H)\big) \le \mathbb{P}\big(\mathrm{Tr}\,\chi_{E-\ell} * \theta_\eta(H) \le 2/9\big) \le \mathbb{P}\big(\mathcal{N}(E,\infty) \le 2/9 + N^{-\varepsilon}\big) = \mathbb{P}\big(\mathcal{N}(E,\infty) = 0\big),$$

where we used the upper bound from (6.20) and that $\mathcal{N}$ is an integer. This completes the proof of the Corollary. □

6.1 Green Function Comparison Theorem

Recalling that $\theta_\eta(H) = \frac{1}{\pi}\mathfrak{Im}\, G(i\eta)$, Corollary 6.2 bounds the probability of $\mathcal{N}(E, \infty) = 0$ in terms of the expectations of two functionals of Green functions. In this subsection, we show that the difference between the expectations of these functionals w.r.t. two probability distributions $\mathbf{v}$ and $\mathbf{w}$ is negligible assuming their second moments match. The precise statement is the following Green function comparison theorem on the edges. All statements are formulated for the upper spectral edge 2, but with the same proof they hold for the lower spectral edge -2 as well.

Theorem 6.3 (Green function comparison theorem on the edge) *Suppose that the assumptions of Theorem 2.4, including* (2.40)*, hold. Let $F : \mathbb{R} \to \mathbb{R}$ be a function whose derivatives satisfy*

$$\max_x |F^{(\alpha)}(x)| \left(|x| + 1\right)^{-C_1} \le C_1, \qquad \alpha = 1,\ 2,\ 3,\ 4 \tag{6.23}$$

with some constant $C_1 > 0$. Then there exists $\varepsilon_0 > 0$ depending only on C_1 such that for any $\varepsilon < \varepsilon_0$ and for any real numbers E, E_1 and E_2 satisfying

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

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

$$\left| \mathbb{E}^{\mathbf{v}} F\left(N\eta \mathfrak{Im}\, m(z)\right) - \mathbb{E}^{\mathbf{w}} F\left(N\eta \mathfrak{Im}\, m(z)\right) \right| \le C N^{-1/6+C\varepsilon}, \qquad z = E + i\eta, \tag{6.24}$$

and

$$\left| \mathbb{E}^{\mathbf{v}} F\left(N \int_{E_1}^{E_2} \mathrm{d}y\, \mathfrak{Im}\, m(y+i\eta)\right) - \mathbb{E}^{\mathbf{w}} F\left(N \int_{E_1}^{E_2} \mathrm{d}y\, \mathfrak{Im}\, m(y+i\eta)\right) \right| \le C N^{-1/6+C\varepsilon} \tag{6.25}$$

for some constant C and large enough N depending only on C_1, ϑ, $\delta_\pm$ and C_0 (in (2.4)*).*

Theorem 6.3 holds in a much greater generality. We state the following extension which can be used to prove (2.42), the generalization of Theorem 2.4. The class of functions F in the following theorem can be enlarged to allow some polynomially increasing functions similar to (6.23). But for the application to prove (2.42), the following form is sufficient. The proof of Theorem 6.4 is similar to that of Theorem 6.3 and will be omitted.

Theorem 6.4 *Suppose that the assumptions of Theorem 2.4, including* (2.40)*, hold. Fix any $k \in \mathbb{N}_+$ and let $F : \mathbb{R}^k \to \mathbb{R}$ be a bounded smooth function with bounded derivatives. Then for any sufficiently small ε there exists a $\delta > 0$ such that for any sequence of real numbers $E_k < \ldots < E_1 < E_0$ with $|E_j - 2| \le N^{-2/3+\varepsilon}$, $j = 0, 1, \ldots, k$, we have*

$$\left| \left(\mathbb{E}^{\mathbf{v}} - \mathbb{E}^{\mathbf{w}}\right) F\left(N \int_{E_1}^{E_0} \mathrm{d}y \mathfrak{Im}\, m(y+i\eta), \ldots, N \int_{E_k}^{E_0} \mathrm{d}y \mathfrak{Im}\, m(y+i\eta)\right) \right| \le N^{-\delta}. \tag{6.26}$$

Assuming that Theorem 6.3 holds, we now prove Theorem 2.4.

Proof of Theorem 2.4. As we discussed in (6.2) and (6.3), we can assume that (6.4) holds for the parameter s. We define $E := 2 + sN^{-2/3}$ that satisfies (6.19). We define E_L as in (6.5) with the L such that (6.2) and (6.3) hold. For simplicity, we set $\xi = \phi L$ and note that $\xi \geq 2$ for sufficiently large N. With the left side of (6.21), for any sufficiently small $\varepsilon > 0$, we have

$$\mathbb{E}^{\mathbf{w}}\, q\left(\operatorname{Tr}\chi_{E-\ell} * \theta_\eta(H)\right) \leq \mathbb{P}^{\mathbf{w}}(\mathcal{N}(E,\infty) = 0) \tag{6.27}$$

with the choice

$$\ell := \frac{1}{2}N^{-2/3-\varepsilon}, \qquad \eta := N^{-2/3-9\varepsilon}.$$

The bound (6.25) applying to the case $E_1 = E - \ell$ and $E_2 = E_L$ shows that there exist $\delta > 0$, for sufficiently small $\varepsilon > 0$, such that

$$\mathbb{E}^{\mathbf{v}}\, q\left(\operatorname{Tr}\chi_{E-\ell} * \theta_\eta(H)\right) \leq \mathbb{E}^{\mathbf{w}}\, q\left(\operatorname{Tr}\chi_{E-\ell} * \theta_\eta(H)\right) + N^{-\delta} \tag{6.28}$$

(note that 9ε plays the role of the ε in the Green function comparison theorem). Then applying the right side of (6.21) in Lemma 6.2, with $\xi = \phi L \geq 2$, to the l.h.s of (6.28), we have

$$\mathbb{P}^{\mathbf{v}}(\mathcal{N}(E-2\ell,\infty) = 0) \leq \mathbb{E}^{\mathbf{v}}\, q\left(\operatorname{Tr}\chi_{E-\ell} * \theta_\eta(H)\right) + C\exp\big[-c(\log N)^2\big]. \tag{6.29}$$

Combining these inequalities, we have

$$\mathbb{P}^{\mathbf{v}}(\mathcal{N}(E-2\ell,\infty) = 0) \leq \mathbb{P}^{\mathbf{w}}(\mathcal{N}(E,\infty) = 0) + 2N^{-\delta} \tag{6.30}$$

for sufficiently small $\varepsilon > 0$ and sufficiently large N. Recalling that $E = 2 + sN^{-2/3}$, this proves the first inequality of (2.41) and, by switching the role of $\mathbf{v}, \mathbf{w}$, the second inequality of (2.41) as well. This completes the proof of Theorem 2.4. $\square$

Proof of Theorem 6.3. Notice that

$$N\int_{E_1}^{E_2} \mathrm{d}y\, \mathfrak{Im}\, m(y+i\eta) = \eta\int_{E_1}^{E_2} \mathrm{d}y\, \operatorname{Tr} G(z)\overline{G}(z), \quad z = y + i\eta. \tag{6.31}$$

We now set up notations to replace the matrix elements one by one. This step is identical for the proof of both (6.24) and (6.25), and we will use the notations of the case (6.24) which are less involved.

Fix a bijective ordering map on the index set of the independent matrix elements,

$$\phi : \{(i,j) : 1 \leq i \leq j \leq N\} \to \Big\{1, \ldots, \gamma(N)\Big\}, \qquad \gamma(N) := \frac{N(N+1)}{2}, \tag{6.32}$$

and denote by H_γ the generalized Wigner matrix whose matrix elements h_{ij} follow the v-distribution if $\phi(i,j) \leq \gamma$ and they follow the w-distribution otherwise; in particular $H_0 = H^{(v)}$ and $H_{\gamma(N)} = H^{(w)}$. The specific choice of the ordering map (6.32) is irrelevant; in the following argument, ϕ could be any bijective ordering map. With $\eta = N^{-2/3-\varepsilon}$, it was proved in (2.21) that for any constant $\xi > 0$,

$$\mathbb{P}\left(\max_{0\leq\gamma\leq\gamma(N)}\max_{1\leq k,l\leq N}\max_{E}\left|\left(\frac{1}{H_\gamma - E - i\eta}\right)_{kl} - \delta_{kl}m_{sc}(E+i\eta)\right| \leq N^{-1/3+2\varepsilon}\right) \geq 1 - C\exp[-c(\log N)^\xi] \tag{6.33}$$

with some constants C, c and large enough $N \geq N_0$ (may depend on ξ). The last maximum in the formula (6.33) runs over all E satisfying $|E-2| \leq N^{-2/3+\varepsilon}$. When applying (2.21), we have used $(\log N)^{4L}(N\eta)^{-1} \leq N^{-1/3+2\varepsilon}$ and that

$$\mathfrak{Im}\, m_{sc}(E+i\eta) \leq \sqrt{|E-2|+\eta} \leq CN^{-1/3+\varepsilon/2} \tag{6.34}$$

for $|E-2| \leq CN^{-2/3+\varepsilon}$.

We set $z = E + i\eta$ where $|E-2| \leq CN^{-2/3+\varepsilon}$ and $\eta = N^{-2/3-\varepsilon}$. From (6.33), (6.34) and the identity

$$\mathfrak{Im}\, m(z) = \frac{1}{N}\mathfrak{Im}\,\mathrm{Tr}\, G = \frac{\eta}{N}\sum_{ij} G_{ij}\overline{G_{ij}},$$

we have that

$$\Big|\eta^2 \sum_{ij} G_{ij}\overline{G_{ij}}\Big| = |N\eta\mathfrak{Im}\, m(z)| \leq CN^{2\varepsilon} \tag{6.35}$$

and

$$\Big|\eta^2 \sum_{i=j} G_{ij}\overline{G_{ij}}\Big| \leq N\eta^2\left(|m_{sc}| + CN^{-1/3+2\varepsilon}\right) \leq CN^{-1/3-2\varepsilon} \tag{6.36}$$

hold with a probability larger than $1 - C\exp[-c(\log N)^{\xi}]$. Since the derivative of F is bounded as in (6.23), there exists C depending on F, ϑ, $\delta_\pm$ and C_0 such that

$$\left|\mathbb{E}F\left(\eta^2\sum_{ij} G_{ij}\overline{G_{ij}}\right) - \mathbb{E}F\left(\eta^2\sum_{i\neq j} G_{ij}\overline{G_{ij}}\right)\right| \leq CN^{-1/3+C\varepsilon}. \tag{6.37}$$

This holds for both the $\mathbf{v}$ and the $\mathbf{w}$ ensembles.

To show (6.24), we only need to prove that for small enough ε, there exists C depending on F, ϑ, $\delta_\pm$ and C_0 such that

$$\left|\mathbb{E}^{\mathbf{v}}F\left(\eta^2\sum_{i\neq j} G^{(v)}_{ij}\overline{G^{(v)}_{ji}}\right) - \mathbb{E}^{\mathbf{w}}F\left(G^{(v)} \to G^{(w)}\right)\right| \leq CN^{-1/6+C\varepsilon}, \tag{6.38}$$

where $G^{(v)}$ and $G^{(w)}$ denote the Green functions of the $H^{(v)}$ and $H^{(w)}$, respectively. Here the shorthand notation $F\left(G^{(v)} \to G^{(w)}\right)$ means that we consider the same argument of F as in the first term in (6.38), but all $G^{(v)}$ terms are replaced with $G^{(w)}$. In fact, the upper index notation is slightly superfluous since the Green function is the same, only the underlying ensemble measure changes, but we wish to emphasize the difference between the two ensembles in this way as well.

Similarly, for (6.25), we only need to prove that for small enough ε, there exists C depending on F, ϑ, $\delta_\pm$ and C_0 such that

$$\left|\mathbb{E}^{\mathbf{v}}F\left(N\int_{E_1}^{E_2} \mathrm{d}y\left(\eta\sum_{i\neq j} G^{(v)}_{ij}\overline{G^{(v)}_{ji}}(y+i\eta)\right)\right) - \mathbb{E}^{\mathbf{w}}F\left(G^{(v)} \to G^{(w)}\right)\right| \leq CN^{-1/6+C\varepsilon}. \tag{6.39}$$

Consider the telescopic sum of differences of expectations

$$\mathbb{E}\, F\left(\eta^2 \sum_{i\neq j}\left(\frac{1}{H^{(v)}-z}\right)_{ij}\overline{\left(\frac{1}{H^{(v)}-z}\right)_{ji}}\right) - \mathbb{E}\, F\left(H^{(v)}\to H^{(w)}\right) \tag{6.40}$$
$$= \sum_{\gamma=1}^{\gamma(N)}\left[\mathbb{E}\, F\left(H^{(v)}\to H_\gamma\right) - \mathbb{E}\, F\left(H^{(v)}\to H_{\gamma-1}\right)\right].$$

Let $E^{(ij)}$ denote the matrix whose matrix elements are zero everywhere except at the (i,j) position, where it is 1, i.e., $E^{(ij)}_{k\ell} = \delta_{ik}\delta_{j\ell}$. Fix a $\gamma \geq 1$ and let (a,b) be determined by $\phi(a,b)=\gamma$. For simplicity to introduce the notation, we assume that $a\neq b$. The $a=b$ case can be treated similarly. We note the total number of the diagonal terms is N and the one of the off-diagonal terms is $O(N^2)$. We will compare $H_{\gamma-1}$ with H_γ for each γ and then sum up the differences according to (6.40).

Note that these two matrices differ only in the (a,b) and (b,a) matrix elements and they can be written as

$$H_{\gamma-1} = Q + \frac{1}{\sqrt{N}}V, \qquad V := v_{ab}E^{(ab)} + v_{ba}E^{(ba)} \tag{6.41}$$
$$H_{\gamma} = Q + \frac{1}{\sqrt{N}}W, \qquad W := w_{ab}E^{(ab)} + w_{ba}E^{(ba)},$$

with a matrix Q that has zero matrix element at the (a,b) and (b,a) positions and where we set $v_{ji} := \overline{v}_{ij}$ for $i<j$ and similarly for w. Define the Green functions

$$R := \frac{1}{Q-z}, \qquad S := \frac{1}{H_{\gamma-1}-z}, \qquad T := \frac{1}{H_\gamma - z}. \tag{6.42}$$

We first claim that the estimate (6.33) holds for the Green function R as well. More precisely, the probability of the event

$$\Omega_R := \max_{1\leq k,l\leq N}\max_E \left|R_{kl}(E+i\eta) - \delta_{kl}m_{sc}(E+i\eta)\right| \geq N^{-1/3+2\varepsilon} \tag{6.43}$$

(where $\max_E$ is the maximum over all E with $|E-2|\leq N^{-2/3+\varepsilon}$) satisfies

$$\mathbb{P}(\Omega_R) \leq C\exp\big[-c(\log N)^\xi\big] \tag{6.44}$$

for any fixed $\xi>0$. To see this, we use the resolvent expansion

$$R = S + N^{-1/2}SVS + N^{-1}(SV)^2S + \ldots + N^{-9/5}(SV)^9S + N^{-5}(SV)^{10}R. \tag{6.45}$$

Since V has only at most two nonzero elements, when computing the (k,ℓ) matrix element of this matrix identity, each term is a sum of finitely many terms (i.e. the number of summands is N-independent) that involve matrix elements of S or R and v_{ij}, e.g. $(SVS)_{k\ell} = S_{ki}v_{ij}S_{j\ell} + S_{kj}v_{ji}S_{i\ell}$. Using the bound (6.33) for the S matrix elements, the subexponential decay for v_{ij} and the trivial bound $|R_{ij}|\leq\eta^{-1}\leq N$, we obtain that the estimate (6.33) holds for R as well.

After having introduced these notations, we are in a position to give a heuristic power counting argument that is the core of the proof. In particular, we can explain the origin of the second moment matching condition. Take $F(x)=x$ for simplicity. A resolvent expansion analogous to (6.45) gives

$$\mathbb{E}\eta\sum_i \mathfrak{Im}\, S_{ii} = \eta\,\mathbb{E}\,\mathfrak{Im}\sum_i\Big[R_{ii} - N^{-1/2}(RVR)_{ii} + N^{-1}((RV)^2R)_{ii} + \ldots\Big] \tag{6.46}$$

which is an expansion in the order of $N^{-1/2}$ since the matrix V contains only a few nonzero elements of size $N^{-1/2}$. Notice that $\eta \sum_i \mathfrak{Im}\, S_{ii}$ estimates the number of eigenvalues near E in a window of size η. For the two ensembles to have the same local eigenvalue distribution on scale η, we need the error term to be less than order one even after performing the telescopic sum. In the bulk, η has to be chosen as $\eta \sim N^{-1}$ and we can view $\eta \sum_i$ as order one in the power counting. Since in the telescopic expansion we will have N^2 terms to sum up, we need that the error term of the expansion is $o(N^{-2})$ for each replacement step, i.e., for each fixed label (a, b). This explains the usual condition of four moments to be identical for the Green function comparison theorem in the bulk [22] since the first four terms in (6.46) has to be equal. Near the edges, i.e., at energies E with $|E-2| \lesssim N^{-2/3}$, the correct local scale is $\eta \sim N^{-2/3}$ and the strong local semicircle law (2.21) implies that the off-diagonal Green functions are of order $N^{-1/3}$ and the diagonal Green functions are bounded. Hence the size of the third order term $\eta\, \mathbb{E} \sum_i N^{-3/2}((RV)^3R)_{ii}$ is of order

$$\eta N N^{-3/2} N^{-2/3} = N^{-2+1/6}$$

where we used that, for a generic label (a, b), there are at least two off-diagonal resolvent terms in $((RV)^3R)_{ii}$. Notice that the error term is still larger than N^{-2}, required for summing over a, b (this argument would be sufficient if we had a matching of three moments and only the fourth order term in (6.46) needed to be estimated). The key observation is that the leading term, which gives this order $N^{-2+1/6}$, has actually almost zero expectation which improves the error to be less than $o(N^{-2})$. This is due to the fact that with the help of (6.33) we are able to follow the main term in the diagonal elements of the Green functions and thus compute the expectation fairly precisely. Notice that similar reasons apply to the proof of Lemma 4.1 in Section 7.

6.2 Main Lemma

The key step to the proof of Theorem 6.3 is the following lemma:

Lemma 6.5 *Fix an index γ, recall the definitions of Q, R and S from (6.42) and suppose first that $\gamma = \phi(a, b)$ with $a \neq b$. For any small $\varepsilon > 0$ and under the assumptions in Theorem 6.3 on F, E, E_1 and E_2, there exists C depending on F, ϑ, $\delta_\pm$ and C_0 (but independent of γ) and there exist constants A_N and B_N, depending on the distribution of the Green function Q, denoted by dist(Q), and on the second moments of v_{ab}, denoted by $m_2(v_{ab})$, such that*

$$\left| \mathbb{E}\, F\left(\eta^2 \sum_{i\neq j} S_{ij}\overline{S}_{ji}(z)\right) - \mathbb{E}\, F\left(\eta^2 \sum_{i\neq j} R_{ij}\overline{R}_{ji}(z)\right) - A_N\big(m_2(v_{ab}), dist(Q)\big)\right| \leq CN^{-13/6+C\varepsilon}, \tag{6.47}$$

with $z = E + i\eta$, $\eta = N^{-2/3-\varepsilon}$, and

$$\begin{aligned} &\left| \mathbb{E}F\left(\eta \int_{E_1}^{E_2} \mathrm{d}y \sum_{i\neq j} S_{ij}\overline{S}_{ji}(y+i\eta)\right) - \mathbb{E}F\left(\eta \int_{E_1}^{E_2} \mathrm{d}y \sum_{i\neq j} R_{ij}\overline{R}_{ji}(y+i\eta)\right)\right. \\ &\qquad\qquad\qquad \left. - B_N\big(m_2(v_{ab}), dist(Q)\big)\right| \leq CN^{-13/6+C\varepsilon} \end{aligned} \tag{6.48}$$

for large enough N (independent of γ). The constants A_N and B_N may also depend on F and on the parameters ϑ, $\delta_\pm$ and C_0, but they depend on the centered random variable v_{ab} only through its second moments.

Finally, if $a = b$, i.e. $\gamma = \phi(a, a)$, then the bounds (6.47) *and* (6.48) *hold with $CN^{-11/6+C\varepsilon}$ standing on their right hand side.*

The same estimates hold if S is replaced by T everywhere and note that Q is independent of v_{ab} and w_{ab}. Since $m_2(v_{ab}) = m_2(w_{ab})$, we obviously have that $A_N\big(m_2(v_{ab}), \mathrm{dist}(Q)\big) = A_N\big(m_2(w_{ab}), \mathrm{dist}(Q)\big)$. Thus we get from Lemma 6.5 that in case of $a \neq b$

$$\left| \mathbb{E}\, F\left(\eta^2 \sum_{i\neq j} S_{ij}\overline{S}_{ji}(z)\right) - \mathbb{E}\, F\left(\eta^2 \sum_{i\neq j} T_{ij}\overline{T}_{ji}(z)\right) \right| \leq CN^{-13/6+C\varepsilon} \tag{6.49}$$

and a similar bound for the quantity (6.48). In case of $a = b$, the estimate is only $CN^{-11/6+C\varepsilon}$. Recalling the definitions of S and T from (6.42), the bound (6.49) compares the expectation of a function of the resolvent of H_γ and that of $H_{\gamma-1}$. The telescopic summation then implies (6.38) and (6.39) since the number of summands with $a \neq b$ is of order N^2 but the number of summands with $a = b$ is only N. This completes the proof of Theorem 6.3. $\square$

Proof of Lemma 6.5. We will only prove the more complicated case (6.48); the proof can be adapted easily for (6.47) which will be omitted. Similarly to Ω_R from (6.43), define

$$\Omega_S := \max_{1\leq k,l\leq N} \max_E \big|S_{kl}(E + i\eta) - \delta_{kl} m_{sc}(E + i\eta)\big| \geq N^{-1/3+2\varepsilon},$$

where $\max_E$ is the maximum over all E with $|E - 2| \leq N^{-2/3+\varepsilon}$. Since S is the Green function of $H_{\gamma-1}$, we obtain from (6.33) directly that

$$\mathbb{P}(\Omega_S) \leq C \exp\big[-c(\log N)^\xi\big] \tag{6.50}$$

for any fixed $\xi > 0$. Finally, set

$$\Omega_{\mathbf{v}} := \{|v_{ab}| \geq N^\varepsilon \sigma_{ab}\}, \qquad \text{and} \qquad \Omega := \Omega_R \cup \Omega_S \cup \Omega_{\mathbf{v}}. \tag{6.51}$$

Using (6.44), (6.50) and the subexponential decay of v_{ab}, we obtain

$$\mathbb{P}\,(\Omega) \leq C \exp\big[-c(\log N)^\xi\big]. \tag{6.52}$$

for any fixed $\xi > 0$ and large enough N. Since the arguments of F in (6.48) are bounded by $CN^{2+2\varepsilon}$ and $F(x)$ increases at most polynomially, it is easy to see that the contribution of the set Ω to the expectations in (6.48) is negligible. We can thus concentrate on the set Ω^c.

Define x^S and x^R by

$$x^S := \eta \int_{E_1}^{E_2} \mathrm{d}y \sum_{i\neq j} S_{ij}\overline{S}_{ji}(y + i\eta), \quad x^R := \eta \int_{E_1}^{E_2} \mathrm{d}y \sum_{i\neq j} R_{ij}\overline{R}_{ji}(y + i\eta), \tag{6.53}$$

and decompose x^S into three parts

$$x^S = x_2^S + x_1^S + x_0^S, \quad x_k^S := \eta \int_{E_1}^{E_2} \mathrm{d}y \sum_{i\neq j,\ |\{i,j\}\cap\{a,b\}|=k} S_{ij}\overline{S}_{ji}(y + i\eta), \tag{6.54}$$

and x_k^R are defined similarly. Here $k = |\{i,j\} \cap \{a,b\}|$ is the number of times a and b appears among the summation indices i, j (if $a = b$ then we count it only once); clearly $k = 0, 1$ or 2. The number of the terms in the summation of x_k^S is $O(N^{2-k})$ since a and b are fixed. From the resolvent expansion, we have

$$S = R - N^{-1/2}RVR + N^{-1}(RV)^2R - N^{-3/2}(RV)^3R + N^{-2}(RV)^4S. \tag{6.55}$$

In the following formulas we will omit the spectral parameter from the notation of the resolvents. The spectral parameter is always $y + i\eta$ with $y \in [E_1, E_2]$, in particular $|y - 2| \le N^{-2/3+\varepsilon}$.

If $|\{i,j\} \cap \{a,b\}| = k$, using (6.55) and (6.33), we have in Ω^c

$$\left|N^{-m/2}\left[(RV)^m R\right]_{ij}\right| \le C_m N^{-m/2+3m\varepsilon} N^{-(2-k)/3}, \;\; m \in \mathbb{N}_+, \;\; k = 0, 1, 2 \tag{6.56}$$

for some constants C_m. Furthermore, we can replace the last R by S, i.e., we also have

$$\left|N^{-2}\left[(RV)^4 S\right]_{ij}\right| \le CN^{-2-(2-k)/3+C\varepsilon}. \tag{6.57}$$

Therefore, in Ω^c we have,

$$|x_k^S - x_k^R| \le CN^{-5/6-2k/3+C\varepsilon}, \quad k = 0, 1, 2. \tag{6.58}$$

Inserting these bounds into the Taylor expansion of F and keeping only the terms larger than $o(N^{-2})$, we obtain

$$\left|\mathbb{E}[F(x^S) - F(x^R)] - \mathbb{E}\left(F'(x^R)(x_0^S - x_0^R) + \frac{1}{2}F''(x^R)(x_0^S - x_0^R)^2 + F'(x^R)(x_1^S - x_1^R)\right)\right| \le CN^{-13/6+C\varepsilon}, \tag{6.59}$$

where we used the remark after (6.52) to treat the contribution on the event Ω. Since there is no x_2 appearing in (6.59), we can focus on the case $k = 0$ or 1.

For $k = 0$ or 1, we define $Q_\ell^{(k)}$ for $\ell = 1$, 2 or 3, as the sum of the terms in $x_k^S - x_k^R$ in which the total number of v_{ab} or v_{ba} is ℓ, i.e.,

$$Q_1^{(k)} := -N^{-1/2}\eta \int_{E_1}^{E_2} \mathrm{d}y \sum_{|\{i,j\}\cap\{a,b\}|=k} \Big(R_{ij}\overline{(RVR)_{ji}} + (RVR)_{ij}\overline{R_{ji}}\Big) \tag{6.60}$$

$$Q_2^{(k)} := N^{-1}\eta \int_{E_1}^{E_2} \mathrm{d}y \sum_{|\{i,j\}\cap\{a,b\}|=k} \left(R_{ij}\overline{((RV)^2R)_{ji}} + ((RV)^2R)_{ij}\overline{R_{ji}} + (RVR)_{ij}\overline{(RVR)_{ji}}\right) \tag{6.61}$$

$$\begin{aligned} Q_3^{(k)} := -N^{-3/2}\eta \int_{E_1}^{E_2} \mathrm{d}y \sum_{|\{i,j\}\cap\{a,b\}|=k} \Bigg(&R_{ij}\overline{((RV)^3R)_{ji}} + \overline{R_{ji}}((RV)^3R)_{ij} + ((RV)^2R)_{ij}\overline{(RVR)_{ji}} \\ &+ (RVR)_{ij}\overline{((RV)^2R)_{ji}}\Bigg). \end{aligned} \tag{6.62}$$

By these definitions and (6.56), we have

$$Q_\ell^{(k)} \le N^{-\ell/2-1/3-2k/3+C\varepsilon} \quad \text{in } \Omega^c. \tag{6.63}$$

Furthermore, with (6.56) and (6.57), we decompose $x_k^S - x_k^R$ as

$$x_k^S - x_k^R = Q_1^{(k)} + Q_2^{(k)} + Q_3^{(k)} + O(N^{-7/3+C\varepsilon}). \tag{6.64}$$

The last two terms in (6.62) can also be bounded by using (6.56), i.e.,

$$Q_3^{(k)} = O(N^{-13/6+C\varepsilon}) - N^{-3/2}\eta \int_{E_1}^{E_2} \mathrm{d}y \sum_{|\{i,j\}\cap\{a,b\}|=k} \Big(R_{ij}\overline{((RV)^3R)_{ji}} + \overline{R_{ji}}((RV)^3R)_{ij} \Big) \quad \text{in } \Omega^c. \tag{6.65}$$

Inserting (6.63) and (6.64) into the second term of the l.h.s of (6.59), with the bounds on the derivatives of F, we have

$$\begin{aligned} &\mathbb{E}\left(F'(x^R)(x_0^S - x_0^R) + F'(x^R)(x_1^S - x_1^R) + \frac{1}{2}F''(x^R)(x_0^S - x_0^R)^2 \right) \\ &= B + \mathbb{E}F'(x^R)Q_3^{(0)} + O\left(N^{-13/6+C\varepsilon}\right), \end{aligned} \tag{6.66}$$

where

$$\begin{aligned} B &:= \mathbb{E}\left(\sum_{k=0,1} F'(x^R)[Q_1^{(k)} + Q_2^{(k)}] + \frac{1}{2}F''(x^R)[Q_1^{(0)}]^2 \right) \\ &= \mathbb{E}\left(\sum_{k=0,1} F'(x^R)\mathbb{E}_{v_{ab}}[Q_1^{(k)} + Q_2^{(k)}] + \frac{1}{2}F''(x^R)\mathbb{E}_{v_{ab}}[Q_1^{(0)}]^2 \right) \end{aligned} \tag{6.67}$$

depends on v_{ab} only through its expectation (which is zero) and on its second moments.

First we give a trivial estimate on $Q_3^{(0)}$. In case i, j are distinct from a and b, it is easy to see by writing out terms in (6.65) that they contain at least three offdiagonal elements of resolvent; for example in the term $R_{ij}\overline{R_{ja}v_{ab}R_{ba}v_{ab}R_{ba}v_{ab}R_{bi}}$, appearing in $R_{ij}\overline{((RV)^3R)_{ji}}$, the resolvent matrix elements $R_{ij}\overline{R_{ja}R_{bi}}$ are off-diagonal. Each off-diagonal matrix element of R is bounded by $N^{-1/3+2\varepsilon}$ in Ω_R^c, while the diagonal terms can be estimated by $|m_{sc}|$, hence by a constant, at a negligible error in the set $\Omega^c \subset \Omega_R^c$. This shows that each term in the integrand in (6.65) is bounded by $C\big[N^{-1/3+2\varepsilon}\big]^3$. Note that every estimate is uniform in y, the real part of the spectral parameter, as long as $|y-2| \le N^{-2/3+\varepsilon}$. Estimating F' trivially, we thus obtain

$$\big|\mathbb{E}[F(x^S) - F(x^R)] - B\big| \le CN^{-11/6+C\varepsilon}.$$

This bound proves Lemma 6.5 for the case $a = b$.

For $a \ne b$ this estimate would not be sufficient since the number of pairs $a \ne b$ to sum up in the telescopic summation is of order N^2. However, we will show that in this case the expectation of the $Q_3^{(0)}$ term is of smaller order than the trivial estimate gives.

From now on we assume that $a \neq b$. By (6.65) we have, in Ω^c that

$$
\begin{aligned}
Q_3^{(0)} &= O(N^{-13/6+C\varepsilon}) - N^{-3/2}\eta \int_{E_1}^{E_2} \mathrm{d}y \sum_{j\neq a,b} \sum_{i\neq j,a,b} \\
&\qquad \left[\left(R_{ij}\overline{R_{ja}v_{ab}R_{bb}v_{ba}R_{aa}v_{ab}R_{bi}} + R_{ia}v_{ab}R_{bb}v_{ba}R_{aa}v_{ab}R_{bj}\overline{R_{ji}} \right) + (a \leftrightarrow b) \right] \\
&= O(N^{-13/6+C\varepsilon}) - N^{-3/2}\eta \int_{E_1}^{E_2} \mathrm{d}y \sum_{j\neq a,b} \sum_{i\neq j,a,b} \\
&\qquad \left[\left(\overline{m_{sc}^2} R_{ij}\overline{R_{ja}R_{bi}} + m_{sc}^2 R_{ia}R_{bj}\overline{R_{ji}} \right) |v_{ab}|^2 v_{ab} + (a \leftrightarrow b) \right].
\end{aligned} \tag{6.68}
$$

Note that we explicitly collected those terms that contain the most diagonal elements of R; these are the main terms of $Q_3^{(0)}$. There are several other terms, for example $R_{ij}\overline{R_{ja}v_{ab}R_{ba}v_{ab}R_{ba}v_{ab}R_{bi}}$, that appear in the expansion of $R_{ij}\overline{[(RV)^3R]_{ji}}$, but these are lower order terms and can be directly included in the error term. In the second step in (6.68) we estimated the diagonal terms by m_{sc} at a negligible error in the set $\Omega^c \subset \Omega_R^c$.

We note that v_{ab} is independent of R and $\mathbb{E}_{v_{ab}}|v_{ab}|^2 v_{ab} = O(1)$. Combining (6.68) with (6.66) and (6.59), we obtain

$$
\begin{aligned}
&\left|\mathbb{E}[F(x^S) - F(x^R)] - B\right| \\
\le& CN^{-13/6+C\varepsilon} + \left|\mathbb{E}F'(x^R)Q_3^{(0)}\right| \\
\le& CN^{-13/6+C\varepsilon} + CN^{-5/6+C\varepsilon} \max_y \max_{i\neq j: \{i,j\}\cap\{a,b\}=\emptyset} \Big[\left|\mathbb{E}F'(x^R)R_{ij}\overline{R_{ja}R_{bi}}\right| + \left|\mathbb{E}F'(x^R)R_{ia}R_{bj}\overline{R_{ji}}\right| + (a \leftrightarrow b) \Big],
\end{aligned} \tag{6.69}
$$

where we used the trivial bounds on F' and m_{sc} and we agsin used that every estimate is uniform in y, the real part of the spectral parameter, as long as $|y-2| \le N^{-2/3+\varepsilon}$. As before, $\max_y$ in the last line of (6.69) indicates maximum over all y with $|y-2| \le N^{-2/3+\varepsilon}$ and the spectral parameter of all resolvents is $y + i\eta$.

The following lemma shows that the expectation of the product of the off-diagonal terms in (6.69) is of smaller order than the trivial estimate gives.

Lemma 6.6 *Under the assumption of Lemma 6.5 and assuming that a, b, i, j are all different, we have*

$$
|\mathbb{E}F'(x^R)R_{ij}\overline{R_{ja}R_{bi}}(y+i\eta)| \le N^{-4/3+C\varepsilon} \tag{6.70}
$$

for any y with $|y-2| \le N^{-2/3+\varepsilon}$, and the same estimate holds for the other three terms in the r.h.s of (6.69).

If this lemma holds, then we have thus proved in the case $a \neq b$ that

$$
\left|\mathbb{E}[F(x^S) - F(x^R)] - B\right| \le N^{-13/6+C\varepsilon} \tag{6.71}
$$

where B is defined in (6.67). With the definitions of x's in (6.53), this completes the proof of Lemma 6.5 for the remaining $a \neq b$ case. □

Proof of Lemma 6.6. With the relation between R and S in (6.45) and (6.56), one can see that (6.70) is implied by

$$|\mathbb{E}F'(x^S)S_{ij}\overline{S_{ja}S_{bi}}| \le N^{-4/3+C\varepsilon}, \tag{6.72}$$

under the assumption that a, b, i, j are all different. This replacement is only a technical convenience when we apply the large deviation estimate (Lemma 3.3) below. Lemma 3.3 was formulated with random variables of equal variance, while the matrix elements of Q cannot all be normalized to have the same variance since two matrix elements are zero. The contribution of these two elements is negligible anyway, but the presentation of the argument is simpler if we do not have to carry them separately in the notation. Since S is the Green function of a usual generalized Wigner matrix with all variances being positive, it is easier to deal with (6.72) instead of (6.70).

From the identity (3.7) applied to the Green function S, we have for any different i, j and a

$$|S_{ij} - S_{ij}^{(a)}| = \Big|S_{ia}S_{aj}(S_{aa})^{-1}\Big| \le C(N\eta)^{-2} \le CN^{-2/3+C\varepsilon} \qquad \text{in } \Omega^c. \tag{6.73}$$

From (6.33) we have

$$|S_{ij}| \le N^{-1/3+C\varepsilon}, \qquad i \ne j, \quad \text{in } \Omega^c. \tag{6.74}$$

Combining (6.73) and (6.74), we have

$$|x^S - \widetilde{x}^S| \le N^{-1/3+C\varepsilon}, \tag{6.75}$$

where $\widetilde{x}^S$ is defined using the resolvent of the matrix $H_{\gamma-1}^{(a)}$ exactly as x^S was defined using the resolvent S of matrix $H_{\gamma-1}$. As usual, $H_{\gamma-1}^{(a)}$ denotes the matrix $H_{\gamma-1}$ with a-th row and column removed. Similarly, we have

$$\Big|S_{ij}\overline{S_{ja}S_{bi}} - S_{ij}^{(a)}\overline{S_{ja}S_{bi}^{(a)}}\Big| \le N^{-4/3+C\varepsilon}, \quad \text{in } \Omega^c. \tag{6.76}$$

Hence by these inequalities and the bounds on the derivatives of F, we have

$$|\mathbb{E}F'(x^S)S_{ij}\overline{S_{ja}S_{bi}}| \le \Big|\mathbb{E}[F'(\widetilde{x}^S)]S_{ij}^{(a)}\overline{S_{ja}S_{bi}^{(a)}}\Big| + O\left(N^{-4/3+C\varepsilon}\right). \tag{6.77}$$

Applying the identity (3.5) to S_{ja}, we have

$$S_{ja} = S_{jj}S_{aa}^{(j)}Z_{ja}^{(S)}, \quad \text{with} \qquad Z_{ja}^{(S)} := \sum_{s\, t\notin\{a,j\}} h_{js}S_{st}^{(ja)}h_{ta} - h_{ja}, \tag{6.78}$$

where $h_{\alpha\beta} = (H_{\gamma-1})_{\alpha\beta}$. With the bound on the matrix elements of S in (6.33) and the identity (3.7), in the set Ω^c we have

$$S_{jj} = m_{sc} + O(N^{-1/3+C\varepsilon}), \qquad S_{aa}^{(j)} = m_{sc} + O(N^{-1/3+C\varepsilon}), \qquad S_{ss}^{(ja)} = m_{sc} + O(N^{-1/3+C\varepsilon}). \tag{6.79}$$

Setting

$$\Omega_Z := \left\{|Z_{ja}^{(S)}| \ge N^{-1/3+C\varepsilon}\right\}$$

with a sufficiently large constant C, Lemma 3.3 implies that

$$\mathbb{P}(\Omega \cup \Omega_Z) \le C\exp\big[-c(\log N)^{\xi}\big],$$

for any fixed $\xi > 0$ since on the set Ω^c we have

$$\sum_{s,t\notin\{a,j\}} \sigma_{js}^2\sigma_{ta}^2|S_{st}^{(ja)}|^2 \le \frac{C_0^2}{N^2\eta}\sum_{s\neq a,j}\mathfrak{Im}\, S_{ss}^{(ja)} \le N^{-2/3+C\varepsilon}$$

using the last formula in (6.79). Therefore, with (6.78), in $\Omega^c\cap\Omega_Z^c$ we have

$$S_{ja} = m_{sc}^2 Z_{ja}^{(S)} + O(N^{-2/3+C\varepsilon}). \tag{6.80}$$

Combining (6.80) with (6.77), we see that

$$|\mathbb{E}F'(x^S)S_{ij}\overline{S_{ja}S_{bi}}| \le |m_{sc}^2| \left|\mathbb{E}[F'(\widetilde{x}^S)]S_{ij}^{(a)}S_{bi}^{(a)}\left(\sum_{s\,t\notin\{a,j\}} h_{js}S_{st}^{(ja)}h_{ta} - h_{ja}\right)\right| + O\left(N^{-4/3+C\varepsilon}\right). \tag{6.81}$$

Since $\widetilde{x}^S$, $S_{ij}^{(a)}S_{bi}^{(a)}$, h_{js} and $S_{st}^{(ja)}$ are all independent of the $a-$th row and column of $H_{\gamma-1}$, and the expectations of h_{ta} and h_{ja} are zero, the first term in r.h.s. of (6.81) equals to zero. This implies (6.72) and completes the proof of (6.70). The other terms in (6.69) can be bounded similarly. This completes the proof of Lemma 6.6. □

7 Proof of Lemma 4.1

7.1 Setup and notations

The p-th moment of $\sum_{i=1}^N Z_i$ is given by

$$\frac{1}{N^p}\mathbb{E}\mathbf{1}(\Gamma^c)\left|\sum_{q=1}^N Z_q\right|^p = \frac{1}{N^p}\mathbb{E}\sum_{\#}\sum_{q_1=1}^N\cdots\sum_{q_\alpha=1}^N\cdots\sum_{q_p=1}^N \mathbf{1}(\Gamma^c)Z_{q_1}^{\#}\dots Z_{q_p}^{\#}, \tag{7.1}$$

where the various #'s can be either 0 or the complex conjugate. The precise choice of # will be irrelevant for our argument and the summation over them yields an irrelevant overall factor 2^p.

We write up the definition of Z_{q_α} from (3.16) as follows:

$$Z_{q_\alpha} = \sum_{q_\alpha^2,q_\alpha^3=1}^N G_{q_\alpha^2\,q_\alpha^3}^{(q_\alpha)}\left[h_{q_\alpha,q_\alpha^2}h_{q_\alpha^3,q_\alpha} - \delta_{q_\alpha^3,q_\alpha^2}\sigma_{q_\alpha^2,q_\alpha}^2\right], \tag{7.2}$$

where the summation is over all $q_\alpha^2 \neq q_\alpha$ and $q_\alpha^3 \neq q_\alpha$. To bookkeep the indices in a uniform way, we denote q_α by q_α^1 and we organize the three indices $(q_\alpha^1, q_\alpha^2, q_\alpha^3)$ into a vector $\mathfrak{q}_\alpha$ for each $\alpha = 1,2,\dots,p$.

Furthermore, we organize these p vectors into a $3\times p$ matrix $\mathbf{q} = (q_\alpha^j)$, for $\alpha = 1,\dots,p$ and $j = 1,2,3$, with entries taking values in $\mathbb{N}_N := \{1,2,\dots,N\}$. The *slots* of the matrix $\mathbf{q}$, parametrized by (j,α), $\alpha = 1,2,\dots,p$, $j = 1,2,3$, are called *vertices*, since we will build a graph upon them. The element q_α^j will be called the *index* assigned to the vertex (j,α). The first entry q_α^1 in $\mathfrak{q}_\alpha$ will play a special role, it will be called *location index*, the other two indices, q_α^2, q_α^3 will be called *nonlocation indices*. Similarly, $(1,\alpha)$ will be called *location vertex* and $(2,\alpha)$, $(3,\alpha)$ will be called *nonlocation vertices*. A pair of indices is called *label*. We also define the set of labels in $\mathfrak{q}_\alpha$ that contain q_α^1:

$$Q_\alpha := \{(q_\alpha^1,q_\alpha^2),(q_\alpha^1,q_\alpha^3),(q_\alpha^2,q_\alpha^1),(q_\alpha^3,q_\alpha^1)\}, \qquad \alpha = 1,2\dots p,$$

and sometimes we will use a single letter ν or μ for labels, i.e. for elements of $\bigcup_{\alpha=1}^{p} Q_\alpha$. Note that Q_α contains any label ν together with its *transpose* ν^t, where $\nu^t := (p,q)$ if $\nu = (q,p)$. Carrying ν together with its transpose is necessary since $h_\nu = \bar{h}_{\nu^t}$, i.e. matrix elements with labels ν and ν^t are not independent.

With these notations, we have

$$\frac{1}{N^p}\mathbb{E}\mathbf{1}(\Gamma^c)\left|\sum_{i=1}^{N} Z_i\right|^p = \frac{1}{N^p}\sum_{\mathbf{q}} \Phi_{\mathbf{q}}, \tag{7.3}$$

where we defined

$$\Phi_{\mathbf{q}} := \mathbb{E}\mathbf{1}(\Gamma^c)\prod_{\alpha=1}^{p}\left[G^{(q_\alpha^1)}_{q_\alpha^2\, q_\alpha^3}\,\xi(\mathfrak{q}_\alpha)\right]^{\#}, \quad \xi(\mathfrak{q}_\alpha) := h_{q_\alpha^1,q_\alpha^2}h_{q_\alpha^3,q_\alpha^1} - \delta_{q_\alpha^3,q_\alpha^2}\sigma^2_{q_\alpha^2,q_\alpha^1}. \tag{7.4}$$

The summation in (7.3) runs over all $3\times p$ matrices $\mathbf{q}$ with elements from $\mathbb{N}_N$ and with the restriction that

$$q_\alpha^2 \neq q_\alpha^1, \quad \text{and} \quad q_\alpha^3 \neq q_\alpha^1. \tag{7.5}$$

Let

$$Q_{\mathbf{q}} = Q := \bigcup_{\alpha=1}^{p} Q_\alpha \tag{7.6}$$

denote the set of all possible labels of h-variables appearing in the $\xi(\mathfrak{q}_\alpha)$ factors and notice that its cardinality is bounded by $|Q|\le 4p$.

We would like to compute the expectation in (7.4) by first taking the expectation with respect to the h_ν-variables explicitly appearing in the ξ's. Recall $G^{(q)} = (H^{(q)}-z)^{-1}$ is the Green function of $H^{(q)}$ which is an $(N-1)\times(N-1)$ matrix after removing the q-th row and column from H. Thus $G^{(q_\alpha^1)}$ is independent of the random variables h_ν, $\nu\in Q_\alpha$, i.e. those h-variables that explicitly appear in $\xi(\mathfrak{q}_\alpha)$. There are, however, three complications. First, while each Green function $G^{(q_\alpha^1)}$, $\alpha = 1,2,\ldots,p$, is independent of h_ν, $\nu\in Q_\alpha$, by definition, it still depends on the other h_μ-variables, $\mu\in Q_\beta$, $\beta\neq\alpha$. Second, we have to deal with coincidences; the same h-variable may appear in $\xi(\mathfrak{q}_\alpha)$ and $\xi(\mathfrak{q}_\beta)$ with $\alpha\neq\beta$; in fact these terms give the non-zero contributions. We will develop a graphical scheme to bookkeep the structure of coincidences and estimate the number of off-diagonal resolvent elements. Finally, there is a small technical problem related to the factor $\mathbf{1}(\Gamma^c)$ that depends on all h-variables, but this factor equals one with a very high probability so a fairly easy argument can remove it.

To resolve the first problem, we use the resolvent expansion to express explicitly the dependence of $G^{(q_\alpha^1)}$ on the random variables h_ν with label $\nu\in Q_\beta$, $\beta\neq\alpha$. For $\mathbf{q}$ fixed, let $U^{\langle\alpha\rangle} = U^{\langle\alpha\rangle}_{\mathbf{q}}$ be the matrix

$$(U^{\langle\alpha\rangle})_{i,k} := (H^{(q_\alpha^1)})_{i,k}, \quad \text{for } (i,k)\in Q^{(\alpha)}_{\mathbf{q}} := Q^{(\alpha)} = \bigcup_{\beta\in\{1,\ldots,p\},\beta\neq\alpha} Q_\beta, \tag{7.7}$$

and $(U^{\langle\alpha\rangle})_{i,k} := 0$ otherwise. Note that the number of nonzero matrix elements of $U^{\langle\alpha\rangle}$ is bounded by $|Q|\le 4p$. Define

$$H^{[\alpha]} = H^{[\alpha]}_{\mathbf{q}} := H^{(q_\alpha^1)} - U^{\langle\alpha\rangle}, \qquad G^{[\alpha]}_{\mathbf{q}} = G^{[\alpha]} := (H^{(q_\alpha^1)} - U^{\langle\alpha\rangle} - z)^{-1}.$$

Notice that $G^{[\alpha]}_{\mathbf{q}}$ is independent of all the h-factors that explicitly appear in $\prod_\alpha \xi(\mathfrak{q}_\alpha)$. From the resolvent expansion, we have

$$G^{(q_\alpha^1)} = \sum_{n_\alpha=0}^{\infty}(-G^{[\alpha]}U^{\langle\alpha\rangle})^{n_\alpha}G^{[\alpha]}. \tag{7.8}$$

To estimate the size of these Green functions, we first note that there is a positive universal constant c such that on the set Γ^c we have

$$\max_{i\neq j}|G_{ij}| = \Lambda_o \le \frac{1}{(\log N)^2}, \qquad c \le |G_{ii}| \le 1 + \frac{1}{(\log N)^2}. \tag{7.9}$$

This follows from the fact that $c' \le |m_{sc}(z)| \le 1$ with some positive universal $c' > 0$ and for any $z \in \mathbf{S}_\ell$, see (3.13). By the perturbation formulas (3.6) and (3.7) we have $G^{(k)}_{ij} = G_{ij} - G_{ik}G_{kj}/G_{kk}$ for $i, j \neq k$, thus we also have

$$\max_{i\neq j}|G^{(k)}_{ij}| \le 2\Lambda_o \le \frac{C}{(\log N)^2}, \qquad c' \le |G^{(k)}_{ii}| \le 1 + \frac{C}{(\log N)^2}, \tag{7.10}$$

where $i, j \neq k$. In the good set Γ^c, the matrix elements of $U^{\langle\alpha\rangle}$ satisfy

$$|U^{\langle\alpha\rangle}_{ij}| \le \frac{(\log N)^{L/10}}{\sqrt{N}} \le N^{-1/4} \tag{7.11}$$

(here we used that $L \le \log N/\log\log N$), and $G^{[\alpha]}$ is bounded as

$$\max_{i\neq j}|G^{[\alpha]}_{ij}| \le 2\Lambda_o \le \frac{C}{(\log N)^2}, \quad |G^{[\alpha]}_{ii}| \le 1 + \frac{C}{(\log N)^2} \qquad \text{in } \Gamma^c. \tag{7.12}$$

To see (7.12), we expand

$$G^{[\alpha]} = \sum_{m=0}^{\infty}(G^{(q^1_\alpha)}U^{\langle\alpha\rangle})^m G^{(q^1_\alpha)} \qquad \text{in } \Gamma^c,$$

and use (7.11) and the bounds (7.10) on the matrix elements of $G^{(q)}$, $q \in \mathbb{N}_N$.

Using (7.11) and (7.12) and recalling that only finitely many matrix elements of U are non-zero, we easily see that the expansion (7.8) is convergent and it can be truncated at finite n_α so that the error term can be estimated. Thus there will be no convergence problem and we will focus on getting estimates.

We set

$$\mathbf{n} := (n_1, n_2, \ldots, n_p), \qquad |\mathbf{n}| = \sum_{\alpha=1}^{p} n_\alpha.$$

With this expansion, we can write (7.4) as

$$\Phi_{\mathbf{q}} = \sum_{n=0}^{\infty}\sum_{|\mathbf{n}|=n}\Phi^{\mathbf{n}}_{\mathbf{q}}$$

$$\Phi^{\mathbf{n}}_{\mathbf{q}} := \mathbb{E}\mathbf{1}(\Gamma^c)\prod_{\alpha=1}^{p}\Big[\mathcal{M}^{(n_\alpha)}\xi(\mathfrak{q}_\alpha)\Big]^{\#}, \tag{7.13}$$

$$\mathcal{M}^{(n_\alpha)} = \mathcal{M}^{(n_\alpha)}_{\mathbf{q}} := \big[(-G^{[\alpha]}U^{\langle\alpha\rangle})^{n_\alpha}G^{[\alpha]}\big]_{q^2_\alpha, q^3_\alpha} \tag{7.14}$$

$$= \sum_{\nu^\alpha_1, \nu^\alpha_2, \ldots, \nu^\alpha_{n_\alpha} \in Q^{(\alpha)}} V_{\mathbf{q}}(\underline{\mu}^\alpha, \underline{\nu}^\alpha, n^\alpha) \tag{7.15}$$

with $\underline{\nu}^\alpha := (\nu^\alpha_1, \nu^\alpha_2, \ldots, \nu^\alpha_{n_\alpha})$ and we have expanded $U^{\langle\alpha\rangle}$ appearing in $\big[(-G^{[\alpha]}U^{\langle\alpha\rangle})^{n_\alpha}G^{[\alpha]}\big]_{q^2_\alpha, q^3_\alpha}$ and used the notation

$$V_{\mathbf{q}}(\underline{\mu}^\alpha, \underline{\nu}^\alpha, n^\alpha) := (-1)^{n_\alpha}G^{[\alpha]}_{\mu^\alpha_1}h_{\nu^\alpha_1}G^{[\alpha]}_{\mu^\alpha_2}h_{\nu^\alpha_2}\ldots h_{\nu^\alpha_{n_\alpha}}G^{[\alpha]}_{\mu^\alpha_{n_\alpha+1}}. \tag{7.16}$$

The summation in (7.15) is over all possible ν-labels of the h factors in (7.16). The appearance of the μ^α-labels in (7.16) is just notational simplification, they are explicit functions of $\underline{\nu}^\alpha$ and $\mathfrak{q}_\alpha$ as follows:

$$\mu_1^\alpha = (q_\alpha^2, [\nu_1^\alpha]_1), \;\; \mu_2^\alpha = ([\nu_1^\alpha]_2, [\nu_2^\alpha]_1), \;\; \mu_3^\alpha = ([\nu_2^\alpha]_2, [\nu_3^\alpha]_1), \;\; \ldots \mu_{n_\alpha+1}^\alpha = ([\nu_{n_\alpha}^\alpha]_2, q_\alpha^3), \tag{7.17}$$

where $[\nu_j^\alpha]_1$ and $[\nu_j^\alpha]_2$ denotes the first and second element of the label ν_j^α. Notice that $G^{[\alpha]}$ is independent of all matrix elements h_{jk} explicitly appearing in the ξ-factors in (7.13).

Hence

$$\Phi_{\mathbf{q}}^{\mathbf{n}} := \mathbb{E}\mathbf{1}(\Gamma^c) \sum_{\boldsymbol{\nu}} \prod_{\alpha=1}^{p} \big[V_{\mathbf{q}}(\underline{\mu}^\alpha, \underline{\nu}^\alpha, n^\alpha) \xi(\mathfrak{q}_\alpha) \big]^{\#}, \tag{7.18}$$

where the summation is over all p-tuple of label sequences $\boldsymbol{\nu} = (\underline{\nu}^1, \underline{\nu}^2, \ldots, \underline{\nu}^p) \in A(\mathbf{q}, \mathbf{n}) := \prod_{\alpha=1}^p \big[Q^{(\alpha)} \big]^{n_\alpha}$. The number of different $\boldsymbol{\nu}$'s is bounded by $|A(\mathbf{q}, \mathbf{n})| \le (4p)^n$.

7.2 Strategy of the proof presented in the simplest example

In order to motivate the reader before we start the detailed estimates, we show our strategy via the simplest case $p = 2$,

$$\frac{1}{N^2} \mathbb{E}\mathbf{1}(\Gamma^c) \Big| \sum_{i=1}^N Z_i \Big|^2 = \frac{1}{N^2} \mathbb{E}\mathbf{1}(\Gamma^c) \sum_{i,j=1}^N Z_i \overline{Z_j}.$$

We write out

$$Z_i = \sum_{k,l \ne i} G_{kl}^{(i)} \big[h_{ik} h_{li} - \delta_{kl} \sigma_{ik}^2 \big], \qquad Z_j = \sum_{m,n \ne j} G_{mn}^{(j)} \big[h_{jm} h_{nj} - \delta_{mn} \sigma_{jm}^2 \big].$$

thus we have

$$\frac{1}{N^2} \mathbb{E}\mathbf{1}(\Gamma^c) \Big| \sum_{i=1}^N Z_i \Big|^2 = \frac{1}{N^2} \mathbb{E}\mathbf{1}(\Gamma^c) \sum_{ijklmn} G_{kl}^{(i)} \big[h_{ik} h_{li} - \delta_{kl} \sigma_{ik}^2 \big] \overline{G_{mn}^{(j)} \big[h_{jm} h_{nj} - \delta_{mn} \sigma_{jm}^2 \big]}. \tag{7.19}$$

With the general notation $\alpha = 1, 2$ and the six indices in the summation are organized into a 3x2 matrix with columns $\mathfrak{q}_1 = (q_1^1, q_1^2, q_1^3)$ and $\mathfrak{q}_2 = (q_2^1, q_2^2, q_2^3)$, i.e.

$$\mathbf{q} = \begin{pmatrix} q_1^1 & q_2^1 \\ q_1^2 & q_2^2 \\ q_1^3 & q_2^3 \end{pmatrix} = \begin{pmatrix} i & j \\ k & m \\ l & n \end{pmatrix}.$$

The only restriction for these indices is that the top element of each column is distinct from the other two below. The sets $Q_1 = \{(i,k), (k,i), (i,l), (l,i)\}$ and $Q_2 = \{(j,m), (m,j), (j,n), (n,j)\}$ contain the labels of the h factors that explicitly appear in Z_i and Z_j, respectively.

Now we expand $G^{(i)} = G^{(q_1^1)}$ in the variables h_ν labelled by $\nu \in Q_2$. We thus decompose the minor $H^{(q_1^1)} = H^{[1]} + U^{\langle 1 \rangle}$, where the matrix $U^{\langle 1 \rangle}$ contains only four non-zero entries h_{jm}, h_{mj}, h_{jn} and h_{nj} with labels from Q_2, and $H^{[1]}$ contains all other entries of $H^{(q_1^1)}$. The resolvent $G^{[1]} = (H^{[1]} - z)^{-1}$ is now independent of all expansion variables h_ν with $\nu \in Q = Q_1 \cup Q_2$. Note, however, that this decomposition depends on $\mathbf{q}$, i.e. it will be different for each summand in (7.19). Since $U^{\langle 1 \rangle}$ is small, we can expand

$$G^{(i)} = G^{(q_1^1)} = G^{[1]} - G^{[1]} U^{\langle 1 \rangle} G^{[1]} + G^{[1]} U^{\langle 1 \rangle} G^{[1]} U^{\langle 1 \rangle} G^{[1]} \ldots,$$

and a similar expansion holds for $G^{(j)} = G^{(q_2^1)}$.

We insert these expansions into (7.19) and organize the terms according to their number of the explicit h factors. Effectively, each h factor has a size $N^{-1/2}$ (neglecting logarithmic corrections). The centered random variable $\xi(\mathfrak{q}) = h_{q^1,q^2} h_{q^3,q^1} - \delta_{q^2,q^3}\sigma^2_{q^1,q^2}$ has size N^{-1} and the subtracted expectation $\delta_{q^2,q^3}\sigma^2_{q^1,q^2}$ is treated on the same footing as hh for the purpose of power counting.

Typically we need to show that terms with less than eight h factors have zero expectation to compensate for the sixfold summation of order N^6 with the prefactor N^{-2} in (7.19). Depending on certain coincidences among the summation indices, sometimes terms with less than eight h factors already give non-zero contribution, but then the combinatorial factor from the summation is smaller. Furthermore, we want to bookkeep the number of off-diagonal matrix elements since the final estimate is in terms of a power of Λ_o.

The leading term in (7.19),

$$G^{[1]}_{kl}\big[h_{ik}h_{li} - \delta_{kl}\sigma^2_{ik}\big]\overline{G^{[2]}_{mn}\big[h_{jm}h_{nj} - \delta_{mn}\sigma^2_{jm}\big]}, \tag{7.20}$$

has four h factors but its expectation vanishes unless at least two summation indices in (7.19) coincide, so the sixfold summation is effectively only fourfold. Here the key observation is that if at least one h factor appears linearly in the expansion, then the expectation is zero. However, since the quadratic factor $\xi(\mathfrak{q}_1) = \big[h_{ik}h_{li} - \delta_{kl}\sigma^2_{ik}\big]$ has zero expectation, it is not sufficient to set $k = l$ and $m = n$ to get a non-zero contribution; there must be coincidences between the h factors in $\big[h_{ik}h_{li} - \delta_{kl}\sigma^2_{ik}\big]$ and in $\big[h_{jm}h_{nj} - \delta_{mn}\sigma^2_{jm}\big]$. For example the case $i = j$, $k = m$, $l = n$ yields a nonzero contribution, i.e. the summation is only threefold. Moreover, if both resolvent elements in (7.20) are off-diagonal, then we get an estimate of order $N^{-2}N^3(N^{-1/2})^4\Lambda_o^2 = \Lambda_o^2 N^{-1}$. If one of the resolvent elements is diagonal, say $k = l$, then the other one has to be diagonal as well, $m = n$, otherwise the expectation is zero. This forces one more coincidence, i.e. either $i = j$ and $k = l = m = n$ or $i = m = n$, $j = k = l$. In both cases the summation in (7.19) gives only N^2 and the total estimate is of order N^{-2}.

The next order terms in the expansion are of the form

$$\begin{aligned}&\Big(G^{[1]}U^{\langle 1\rangle}G^{[1]}\Big)_{kl}\big[h_{ik}h_{li} - \delta_{kl}\sigma^2_{ik}\big]\overline{G^{[2]}_{mn}\big[h_{jm}h_{nj} - \delta_{mn}\sigma^2_{jm}\big]}\\ &\quad = \sum_{a,b\in Q_2} G^{[1]}_{ka}U^{\langle 1\rangle}_{ab}G^{[1]}_{bl}\big[h_{ik}h_{li} - \delta_{kl}\sigma^2_{ik}\big]\overline{G^{[2]}_{mn}\big[h_{jm}h_{nj} - \delta_{mn}\sigma^2_{jm}\big]}\end{aligned}$$

with five h factors. Notice that two new summation indices, a, b, have appeared, but their combinatorics is of order one and not of order N^2. In fact, $U^{\langle 1\rangle}_{ab}$ is just one of h_{jm}, h_{nj} or their transposes. Again, there should be at least three coincidences among the indices i, j, k, l, m, n to avoid that at least one h variable appears linearly or that at least one of the quadratic factors $\xi(\mathfrak{q}_1)$, $\xi(\mathfrak{q}_2)$ remains isolated leading to zero expectation. It is again easy to see that we collect at least Λ_o^2 (in fact, typically Λ_o^3) unless at least one additional index coincides.

The terms with six h factors are either of the form

$$\Big(G^{[1]}U^{\langle 1\rangle}G^{[1]}\Big)_{kl}\big[h_{ik}h_{li} - \delta_{kl}\sigma^2_{ik}\big]\overline{\Big(G^{[2]}U^{\langle 2\rangle}G^{[2]}\Big)_{mn}\big[h_{jm}h_{nj} - \delta_{mn}\sigma^2_{jm}\big]}$$

or of the form

$$\Big(G^{[1]}U^{\langle 1\rangle}G^{[1]}U^{\langle 1\rangle}G^{[1]}\Big)_{kl}\big[h_{ik}h_{li} - \delta_{kl}\sigma^2_{ik}\big]\overline{G^{[2]}_{mn}\big[h_{jm}h_{nj} - \delta_{mn}\sigma^2_{jm}\big]}$$

In both cases at least two h factor appears linearly, yielding zero expectation, unless there are two coincidences among i, j, k, l, m, n. Thus the summation in (7.19) is effectively reduced from N^6 to N^4. Since $h^6 \sim$

$(N^{-1/2})^6 = N^{-3}$, we obtain that (7.19) is of order N^{-1}. Moreover, in all cases there are at least two offdiagonal resolvent elements, unless an additional coincidence occurs. Thus the estimate is $N^{-1}(\Lambda_o^2+N^{-1})$. The seventh order terms can be dealt with similarly.

The lowest order non-zero terms with distinct i,j,k,l,m,n indices have eight h factors and they are of the form

$$\Big(G^{[1]}U^{\langle 1\rangle}G^{[1]}U^{\langle 1\rangle}G^{[1]}\Big)_{kl}\big[h_{ik}h_{li}-\delta_{kl}\sigma_{ik}^2\big]\overline{\Big(G^{[2]}U^{\langle 2\rangle}G^{[2]}U^{\langle 2\rangle}G^{[2]}\Big)_{mn}\big[h_{jm}h_{nj}-\delta_{mn}\sigma_{jm}^2\big]}.$$

We now have four U-factors, so they can ensure that all variables h_{ik}, h_{li}, h_{jm}, h_{nj} appear quadratically to prevent zero expectation. For example, the term

$$G^{[1]}_{km}h_{mj}G^{[1]}_{jj}h_{jn}G^{[1]}_{nl}\big[h_{ik}h_{li}-\delta_{kl}\sigma_{ik}^2\big]\,\overline{G^{[2]}_{mk}h_{ki}G^{[2]}_{ii}h_{il}G^{[2]}_{ln}\big[h_{jm}h_{nj}-\delta_{mn}\sigma_{jm}^2\big]}.$$

has non-zero expectation. Moreover, there are four resolvents in offdiagonal form, unless there is an index coincidence, so the size of this term is $N^{-2}N^6(N^{-1/2})^8\Lambda_o^4=\Lambda_o^4$.

The mechanism to estimate the term (7.18) for general p is the same, but the bookkeeping is more tedious. We will have to estimate the size of each non-vanishing term as powers of N and Λ_o^2.

The power counting in N is relatively straightforward. It is easy to see that if all indices in the matrix $\mathbf{q}$ are distinct, then at least $2p$ new h factors must come from the $V_{\mathbf{q}}$ factors to ensure that none of the h factors in $\prod_\alpha \xi(\mathfrak{q}_\alpha)$ appears linearly (otherwise the expectation would be zero). Thus the total number of h factors is at least $4p$ and their size is estimated by $(N^{-1/2})^{4p}=N^{-2p}$. Together with the N^{-p} prefactor in (7.3), this will compensate for the N^{3p} combinatorial factor coming from the summation over all $3\times q$ matrices. If some indices in $\mathbf{q}$ coincided, then the corresponding h factors could appear with a higher multiplicity in $\prod_\alpha \xi(\mathfrak{q}_\alpha)$, so their expectation would not necessarily vanish even without an additional h factor from $V_{\mathbf{q}}$. Each coincidence in $\mathbf{q}$ reduces the number of necessary h factors from $V_{\mathbf{q}}$ at most by two, hence keeping the overall balance of N-powers.

The power counting in Λ_o is more complicated and it is related to the fact that the expectation of each ξ is zero. This means that an index coincidence of the form $q_\alpha^2=q_\alpha^3$ does not imply non-vanishing expectation yet. The requirement of nonzero expection either forces coincidences of indices among h factors in *different* ξ terms, but then typically *two* indices have to match, so we gain an additional N^{-1}; or it forces matching h factors in the ξ-terms with U-factors in the expansion (7.8). The latter implies, however, that instead of a single resolvent $G^{[\alpha]}$ we consider a longer expansion of the form $G^{[\alpha]}U^{\langle\alpha\rangle}G^{[\alpha]}\ldots$ which typically has at least *two* off-diagonal resolvents instead of only one. These two scenarios yield an additional factor $(\Lambda_o^2+N^{-1})$ for each ξ-factor. This gives $(\Lambda_o^2+N^{-1})^p$ as a final estimate.

In the next section we give the precise details of this strategy.

7.3 Detailed proof of Lemma 4.1.

The proof will be divided into three parts. The first part is a technical preparation to deal with the very small probability event represented by the set Γ, where either h or a resolvent is too large. It can be skipped at the first reading. In the second part we organize the expansion by encoding the coincidence structure of various terms by a graph. Finally, in the third part we estimate the size of each term with the help of the graphical representation.

7.3.1 Cutoff of small probability events

Since $|h_{ij}| \le (\log N)^{L/10} N^{-1/2}$ in the set Γ^c and (7.12) also holds in Γ^c, we clearly have

$$\left|\mathbf{1}(\Gamma^c)\xi(\mathfrak{q}_\alpha)V_{\mathbf{q}}(\underline{\mu}^\alpha, \underline{\nu}^\alpha, n^\alpha)\right| \le C\left[\frac{C(\log N)^{L/10}}{\sqrt{N}}\right]^{n_\alpha} \frac{(\log N)^{L/10}}{N}. \tag{7.21}$$

Hence we have

$$|\Phi_{\mathbf{q}}^{\mathbf{n}}| \le (Cp)^n \left[\frac{(\log N)^{L/10}}{\sqrt{N}}\right]^n \Big(\frac{(\log N)^{L/10}}{N}\Big)^p, \tag{7.22}$$

where $(Cp)^n$ is the combinatorics of the summation over $\boldsymbol{\nu}$ in (7.18). Thus we have

$$\frac{1}{N^p}\sum_{\mathbf{q}} \Phi_{\mathbf{q}} = \frac{1}{N^p}\sum_{\mathbf{q}}\sum_{n=0}^{\infty}\sum_{|\mathbf{n}|=n} \Phi_{\mathbf{q}}^{\mathbf{n}} \le (\log N)^{pL/10} N^p \sum_{n=0}^{\infty}(Cp)^n \sum_{|\mathbf{n}|=n}\left[\frac{(\log N)^{L/10}}{\sqrt{N}}\right]^n, \tag{7.23}$$

where we used that the summation over all $\mathbf{q}$ yields a factor N^{3p}. Since the number of $\mathbf{n} = (n_1, n_2, \dots, n_p)$ with $|\mathbf{n}| = n$ is bounded by 2^{n+p}, the last term is bounded by

$$(CN(\log N)^{L/10})^p \sum_{n=0}^{\infty}\left[\frac{Cp(\log N)^{L/10}}{\sqrt{N}}\right]^n. \tag{7.24}$$

Since $p \le (\log N)^{L/10}$ and $L \le \log N/\log\log N$, the sum of the tail terms with $n \ge 6p$ is bounded by $CN^{-5p/2}$, for sufficiently large N, hence for the bound (4.5) we only have to estimate terms with $n \le 6p$.

We denote all independent random variables by $\mathbf{h} = (h_\nu)$ and split them according to the set Q (see (7.6)), i.e., we will write $\mathbf{h} = (\mathbf{h}_1, \mathbf{h}_2)$ with $\mathbf{h}_2 = (h_\nu : \nu \in Q)$ and $\mathbf{h}_1 = (h_\nu : \nu \notin Q)$. Denote the corresponding projection by $\pi_j, j = 1, 2$, i.e. $\pi_j \mathbf{h} = \mathbf{h}_j$. Define

$$(\Gamma^c)_1 := \pi_1(\Gamma^c), \qquad Y^c := \prod_{\nu\in Q}\{h_\nu : |h_\nu| \le (\log N)^{L/10}|\sigma_\nu|\} \subset \mathbb{C}^Q, \qquad Y := \mathbb{C}^Q \setminus Y^c. \tag{7.25}$$

By definition, $G^{[\alpha]}$ depends only on variables $\mathbf{h}_1$. Furthermore, for any $\mathbf{h}_1 \in (\Gamma^c)_1$ there exists $\mathbf{h}_2$ such that $\mathbf{h} = (\mathbf{h}_1, \mathbf{h}_2) \in \Gamma^c$, in particular, the estimates (7.12) hold for any $\mathbf{h}_1 \in (\Gamma^c)_1$. By definition of Γ^c, we have

$$\Gamma^c \subset (\Gamma^c)_1 \times Y^c \tag{7.26}$$

and (7.21) holds in the set $(\Gamma^c)_1 \times Y^c$. From the resolvent expansion, we have for $i \ne j$, and for $\mathbf{h}_1 \in (\Gamma^c)_1$,

$$G_{ij}^{[\alpha]}(\mathbf{h}_1) = G_{ij}^{[\alpha]}(\mathbf{h}) = (H^{(\alpha)} - U^{\langle\alpha\rangle} - z)^{-1}_{ij} = \sum_{n_\alpha=0}^{\infty}\Big[(G^{(\alpha)}U^{\langle\alpha\rangle})^{n_\alpha}G^{(\alpha)}\Big]_{ij} \tag{7.27}$$

$$= G_{ij}^{(\alpha)} + \sum_{n_\alpha=1}^{\infty}\Big[(G^{(\alpha)}U^{\langle\alpha\rangle})^{n_\alpha}G^{(\alpha)}\Big]_{ij}. \tag{7.28}$$

Using

$$\mathbf{1}(\Gamma^c) = \mathbf{1}((\Gamma^c)_1) - \mathbf{1}\Big((\Gamma^c)_1 \times Y^c \setminus \Gamma^c\Big) - \mathbf{1}((\Gamma^c)_1)\mathbf{1}(Y), \tag{7.29}$$

we can rewrite $\Phi_{\mathbf{q}}^{\mathbf{n}}$ as

$$\Phi_{\mathbf{q}}^{\mathbf{n}} := \widetilde{\Phi}_{\mathbf{q}}^{\mathbf{n}} + X_{\mathbf{q},1}^{\mathbf{n}} + X_{\mathbf{q},2}^{\mathbf{n}} \tag{7.30}$$

$$\widetilde{\Phi}_{\mathbf{q}}^{\mathbf{n}} := \sum_{\boldsymbol{\nu}\in A(\mathbf{q},\mathbf{n})} \widetilde{\Phi}_{\mathbf{q},\boldsymbol{\nu}}^{\mathbf{n}}$$

$$\widetilde{\Phi}_{\mathbf{q},\boldsymbol{\nu}}^{\mathbf{n}} := \mathbb{E}\mathbf{1}((\Gamma^c)_1)\prod_{\alpha=1}^{p} V_{\mathbf{q}}(\underline{\mu}^\alpha, \underline{\nu}^\alpha, n^\alpha)\xi(\mathfrak{q}_\alpha) \tag{7.31}$$

$$X_{\mathbf{q},1}^{\mathbf{n}} := -\mathbb{E}\mathbf{1}((\Gamma^c)_1)\mathbf{1}(Y)\sum_{\boldsymbol{\nu}}\prod_{\alpha=1}^{p} V_{\mathbf{q}}(\underline{\mu}^\alpha, \underline{\nu}^\alpha, n^\alpha)\xi(\mathfrak{q}_\alpha) \tag{7.32}$$

$$X_{\mathbf{q},2}^{\mathbf{n}} := -\mathbb{E}\mathbf{1}\Big((\Gamma^c)_1 \times Y^c \setminus \Gamma^c\Big)\sum_{\boldsymbol{\nu}}\prod_{\alpha=1}^{p} V_{\mathbf{q}}(\underline{\mu}^\alpha, \underline{\nu}^\alpha, n^\alpha)\xi(\mathfrak{q}_\alpha). \tag{7.33}$$

Analogously to (7.21)–(7.23), we can bound $X_{\mathbf{q},2}^{\mathbf{n}}$ as follows

$$\frac{1}{N^p}\sum_{\mathbf{q}}\sum_{n=0}^{6p}\sum_{|\mathbf{n}|=n}|X_{\mathbf{q},2}^{\mathbf{n}}| \le (Cp)^{6p}(N(\log N)^{L/10})^p\mathbb{P}(\Gamma) \le C\exp\big[-c(\log N)^{\phi L}\big], \tag{7.34}$$

using the fact that the estimate (7.21) holds even on $(\Gamma^c)_1 \times Y^c$ since all $G^{[\alpha]}$ appearing in $V_{\mathbf{q}}$ depend only on $\{h_\nu : \nu \notin Q\}$. In the last step we used (4.3), $n \le 6p \le 6(\log N)^{\phi L-2} \le 6(\log N)^{L/10}$. For the other error term we have

$$\frac{1}{N^p}\sum_{\mathbf{q}}\sum_{n=0}^{6p}\sum_{|\mathbf{n}|=n}|X_{\mathbf{q},1}^{\mathbf{n}}| \le (Cp)^{6p}(N(\log N)^{L/10})^p\,\exp\big[-c(\log N)^{\psi L}\big] \le C\exp\big[-c(\log N)^{\psi L}\big]. \tag{7.35}$$

Here we have used that for a sufficiently large L, the integration of $\mathbf{h}_2$ over the set Y, i.e. an $O(p)$-moment of the random variables $\mathbf{h}_\nu$, $\nu \in Q$, in the regime where $|h_\nu| \ge (\log N)^{L/10}\sigma_\nu$, is bounded by $C\exp\big[-c(\log N)^{\psi L}\big]$ with some positive ψ, depending on ϑ due to the subexponential decay (2.17) and due to the fact that $p \le (\log N)^{\psi L-2}$. In the estimate (7.35) we also used that (7.12) holds on $(\Gamma^c)_1$ to estimate the $G^{[\alpha]}$ factors remaining from the $V_{\mathbf{q}}$ terms after integrating out the random variables h_ν, $\nu \in Q$.

Collecting the estimates from (7.30), (7.34) and (7.35), we have

$$\frac{1}{N^p}\sum_{\mathbf{q}}\Phi_{\mathbf{q}} \le \frac{1}{N^p}\sum_{\mathbf{q}}\sum_{n=0}^{6p}\sum_{|\mathbf{n}|=n}\big|\widetilde{\Phi}_{\mathbf{q}}^{\mathbf{n}}\big| + C\exp\big[-c(\log N)^{\psi L}\big]. \tag{7.36}$$

The last error term can be absorbed into the N^{-p} term in (4.5) using that $p \le (\log N)^{\psi L-2}$. Hence we only have to estimate the contribution of $\widetilde{\Phi}_{\mathbf{q}}^{\mathbf{n}}$. The key observation is that

$$\mathbb{E}_{h_\nu}\mathbf{1}((\Gamma^c)_1)\xi(\mathfrak{q}_\alpha) = 0 \tag{7.37}$$

for any $\alpha = 1, 2, \ldots, p$ and for any $\nu \in Q$. Furthermore, any resolvent $G^{[\alpha]}$ appearing explicitly in

$$\prod_{\alpha=1}^{p} V_{\mathbf{q}}(\underline{\mu}^\alpha, \underline{\nu}^\alpha, n^\alpha) = \prod_{\alpha=1}^{p}(-1)^{n_\alpha}G_{\mu_1^\alpha}^{[\alpha]}h_{\nu_1^\alpha}G_{\mu_2^\alpha}^{[\alpha]}h_{\nu_2^\alpha}\ldots h_{\nu_{n_\alpha}^\alpha}G_{\mu_{n_\alpha+1}^\alpha}^{[\alpha]} \tag{7.38}$$

is independent of any h_ν, $\nu \in Q$. Therefore the expectation in (7.31) is nonzero only if for each $\nu \in Q$, either h_ν (or its transpose h_{ν^t}) appears explicitly in (7.38) or h_ν (or its transpose h_{ν^t}) appears in two different $\xi(\mathfrak{q}_\alpha)$ factors in (7.31). The first scenario imposes restrictions on the indices of the two resolvents $G^{[\alpha]}$ neighboring h_ν in (7.38) and we will infer that some of these resolvents must be off-diagonal that can be estimated by Λ_o. The second scenario restricts the total combinatorics of the summation over the $\mathbf{q}$ indices in (7.36), which gain can also be expressed as a power of $N^{-1/2}$. In the next step we set up a graphical representation to effectively bookkeep all possible situations.

7.3.2 Combinatorics

Recall that $\mathbf{q}$ is a $3 \times p$ matrix with $3p$ slots. The estimate of $\widetilde{\Phi}^{\mathbf{n}}_{\mathbf{q}}$ defined in the previous section depends on the structure of the indices $\mathbf{q} = (q^j_\alpha)$, more precisely, it depends on which of the indices q^j_α coincide. The relevant structure of these coincidences will be encoded by a graph, $\mathcal{G}(\mathbf{q})$, to be defined below. Roughly speaking (with some modifications specified below), the vertex set of $\mathcal{G}(\mathbf{q})$ will be the set of possible slots of the matrix $\mathbf{q}$; two vertices (j, α) and (i, β) are connected by an edge if the corresponding indices coincide, $q^j_\alpha = q^i_\beta$. Then the summation over $\mathbf{q}$ in the right side of (7.36) will be performed in two steps: first we sum over all possible graphs, then we sum over all possible $\mathbf{q}$'s compatible with this graph, i.e. we write

$$\sum_{\mathbf{q}} = \sum_{G} \sum_{\mathbf{q}\,:\,\mathcal{G}(\mathbf{q})=G}, \tag{7.39}$$

where the first summation is over all graphs with at most $3p$ vertices. In fact, only certain special graphs G will be compatible with a choice of indices $\mathbf{q}$ that occur in our expansion and their number will be bounded by p^{Cp}.

The reason for this resummation is that the size of $\widetilde{\Phi}^{\mathbf{n}}_{\mathbf{q}}$ is essentially given by the number of off-diagonal resolvents in the expansion (7.31), but considering only those terms which are not zero due to the expectation (see (7.51) below). This number can be estimated via the coincidence graph.

We now define the graph $\mathcal{G}(\mathbf{q})$, describing the relevant coincidence structure of $\mathbf{q}$, by performing the following four-step procedure. Strictly speaking, the graph is defined on a subset of the $3p$ vertices (or slots in the matrix) labelled by coordinates (j, α) with $1 \le j \le 3$ and $1 \le \alpha \le p$. We will say that a vertex (j, α) has the *value* r if $q^j_\alpha = r$, in other words, the index q^j_α assigned to the vertex (j, α) will be sometimes also referred to as the value of that vertex. If it does not lead to confusion, we will often simply refer to q^j_α instead of the vertex (j, α), e.g. we will say that two indices, q^j_α and q^i_β are connected by an edge, meaning that the vertices (j, α) and (i, β) are connected.

Let $\ell(\mathbf{q})$ denote the number of different location indices, i.e.,

$$\ell = \ell(\mathbf{q}) := \big|\, \{ q^1_\alpha : 1 \le \alpha \le p \} \,\big|, \tag{7.40}$$

where $|\cdot|$ denotes the cardinality of the set, disregarding multiplicity. We group together all columns with the same location indices; the union of these columns will be called *group*. Let $m_1, m_2, \dots m_\ell$ denote the multiplicity of the groups, i.e., the number of columns with the same location indices. We clearly have

$$\sum_{s=1}^{\ell} m_s = p. \tag{7.41}$$

We start with the matrix $\mathbf{q}$ and perform the following operations to obtain $\mathcal{G}(\mathbf{q})$. In Step 1 and 2 we specify the vertex-set of $\mathcal{G}(\mathbf{q})$ by removing some of the original $3p$ vertices. Step 3 and 4 specify the edges of $\mathcal{G}(\mathbf{q})$. After each step we give an intutive explanation.

Step 1. If $q_\alpha^2 = q_\alpha^3$, we replace q_α^3 by $*$ and the vertex $(3,\alpha)$ will not be part of the graph $\mathcal{G}(\mathbf{q})$. In the matrix, we put a $*$ in its location. We now call q_α^2 a *duplex* and put a subscript d to indicate it.

Explanation: If $q_\alpha^2 = q_\alpha^3$ then the two h factors in $\xi(\mathfrak{q}_\alpha)$ are the same. This coincidence has to be treated separately, since it does not automatically lead to non-zero expectation due to $\mathbb{E}\xi(\mathfrak{q}_\alpha) = 0$. It will thus be easier to merge the vertices $(2,\alpha)$ and $(3,\alpha)$ into one vertex.

Step 2. For $\alpha \neq \beta$ and any $i, j \in \{2,3\}$ we call the vertices (j,α) and (i,β) (and the corresponding indices q_α^j and q_β^i) *twin* if $q_\alpha^j = q_\beta^1$ and $q_\beta^i = q_\alpha^1$. We now replace q_α^j and q_β^i by t to indicate a twin but we do not make any change on location index. Vertices with t will not be part of the graph $\mathcal{G}(\mathbf{q})$. Notice that by the restriction (7.5), $q_\alpha^j \neq q_\alpha^1$ and thus $q_\alpha^1 \neq q_\beta^1$, i.e., twins can only be formed in different groups, i.e. in columns with different location indices.

Explanation: This is the situation where there is a coincidence among the h factors in two different

$$\xi(\mathfrak{q}_\alpha) = h_{q_\alpha^1, q_\alpha^2} h_{q_\alpha^3, q_\alpha^1} - \delta_{q_\alpha^3, q_\alpha^2} \sigma^2_{q_\alpha^2, q_\alpha^1} \quad \text{and} \quad \xi(\mathfrak{q}_\beta) = h_{q_\beta^1, q_\beta^2} h_{q_\beta^3, q_\beta^1} - \delta_{q_\beta^3, q_\beta^2} \sigma^2_{q_\beta^2, q_\beta^1}, \qquad \alpha \neq \beta,$$

e.g. $q_\alpha^2 = q_\beta^1$ and $q_\beta^2 = q_\alpha^1$. Such coincidence results in nonzero expectation with respect to $h_{q_\alpha^1, q_\alpha^2}$ *without forcing* $h_{q_\alpha^1, q_\alpha^2}$ to also appear somewhere in the resolvent expansions, i.e. in one of the $V_{\mathbf{q}}$ factors in (7.38). This means that $h_{q_\alpha^1, q_\alpha^2}$ may not generate an additional off-diagonal resolvent element. We will remove such vertices from the graph to allow a more uniform treatment for the rest and we will account for the twins separately.

Step 3. Two vertices are connected by an edge in $\mathcal{G}(\mathbf{q})$ if the indices assigned to them are the same, *except* if both vertices are in the first row of the matrix. I.e., edges connect vertices with identical indices, except that there is no edge between any two location indices.

Explanation. Since the location index plays a different role than the two non-location indices, their possible coincidence have separately been taken into account by the concept of groups.

Step 4. We add an edge between a duplex $(q_\alpha^2)_d$ and its location index q_α^1 if the multiplicity of the group that the duplex belongs to is one, i.e. if the duplex is *isolated.*

Explanation. This is a purely technical convenience. Later we will consider connected components of $\mathcal{G}(\mathbf{q})$. Isolated duplex will be treated separately (see Case 1. below in the proof of Proposition 7.1), but artificially making the two vertices of a duplex into one connected component will allow us to simplify the argument of Lemma 7.2.

We remark that the number of different graphs arising in via this procedure is bounded by p^{Cp}. This is because $\mathcal{G}(\mathbf{q})$ has the following special structure. Its vertices are partitioned into equivalence classes (according to the common value of their indices) and any two vertices within an equivalence class are connected by an edge, unless they are both location vertices. The number of partitions of the vertices is at most p^{Cp}. Furthermore, there are additional edges between duplexes and their location vertices if the corresponding location index appears only once in $\mathbf{q}$, but the possible combinatorics of these additional edges is at most a factor of 2^p.

Having defined $\mathcal{G}(\mathbf{q})$, the next step is to assign a *weight* to all vertices as follows.

Definition 7.1 (Weight of vertices and groups in $\mathcal{G}(\mathbf{q})$) *(i) In a group with multiplicity $m_s = 1$ each vertex has weight zero.*

(ii) In a group with multiplicity $m_s > 1$ we assign a weight 1 to each duplex in the group; all other non-location vertices in the group will have a weight $1/2$.

(iii) The total weight of a group is the sum of weights of its vertices.

(iv) The total weight $W = W(\mathbf{q})$ of the graph is the sum of the weights of all vertices.

Clearly, the total weight of each group is at most $m_s \le 2(m_s - 1)$. Thus the total weight of the graph satisfies, by (7.41),

$$W \le \sum_{s=1}^{\ell} 2(m_s - 1) = 2(p - \ell). \tag{7.42}$$

If all location indices are distinct, then all weights are zero. In this case, each nonlocation index in $\mathcal{G}(\mathbf{q})$ forces a new h term in $V_{\mathbf{q}}$, see (7.38); note that this statement used that twins are taken out of the graph. If some location indices coincide, i.e. we have a group with multiplicity larger than one, then the possible coincidences of non-location indices within the group may yield non-zero expectation without forcing a corresponding h factor in $V_{\mathbf{q}}$. This may shorten the expansion (7.38), hence reduce the total number of off-diagonal elements. The weight measures the maximal reduction of off-diagonal elements in (7.38) due to the larger multiplicity, compared with the multiplicity one case.

Definition 7.2 (Independent nonlocation indices) *Denote by N_{ind} the number of different nonlocation indices that do not coincide with any location index i.e.,*

$$N_{ind} = N_{ind}(\mathbf{q}) := \Big| \{q_\alpha^j : 2 \le j \le 3,\ 1 \le \alpha \le p\} \setminus \{q_\alpha^1 : 1 \le \alpha \le p\} \Big|, \tag{7.43}$$

where again $|\cdot|$ denotes the cardinality of the set, disregarding multiplicity. The elements of this set will be called independent nonlocation indices.

Note that N_{ind} gives the actual number of different q_α^2 and q_α^3 in the second sum in the right hand side of (7.39). Together with the number of groups ℓ, i.e. the number of different location indices, the number of terms in the $\sum_{\mathbf{q}}$ summation will be bounded by $N^{N_{ind}+\ell}$.

We show an example to illustrate this procedure and definitions. Let $p = 13$ and

$$\mathbf{q} = \begin{pmatrix} 1 & 2 & 3 & 3 & 4 & 4 & 5 & 5 & 5 & 5 & 6 & 7 & 8 \\ 10 & 1 & 2 & 9 & 7 & 15 & 9 & 9 & 9 & 9 & 2 & 4 & 14 \\ 10 & 11 & 5 & 6 & 7 & 12 & 9 & 9 & 13 & 13 & 2 & 12 & 14 \end{pmatrix} \tag{7.44}$$

Then after the first step, we get

$$\begin{pmatrix} 1 & 2 & 3 & 3 & 4 & 4 & 5 & 5 & 5 & 5 & 6 & 7 & 8 \\ (10)_d & 1 & 2 & 9 & 7_d & 15 & 9_d & 9_d & 9 & 9 & 2_d & 4 & 14_d \\ * & 11 & 5 & 6 & * & 12 & * & * & 13 & 13 & * & 12 & * \end{pmatrix} \tag{7.45}$$

After the second step we have

$$\begin{pmatrix} 1 & 2 & 3 & 3 & 4 & 4 & 5 & 5 & 5 & 5 & 6 & 7 & 8 \\ (10)_d & 1 & 2 & 9 & t & 15 & 9_d & 9_d & 9 & 9 & 2_d & t & 14_d \\ * & 11 & 5 & 6 & * & 12 & * & * & 13 & 13 & * & 12 & * \end{pmatrix} \tag{7.46}$$

In this example, the graph $\mathcal{G}(\mathbf{q})$ will have 31 vertices, identified with the slots of the matrix in (7.46) that contain numbers. The slots with stars and t's do not count as vertex of $\mathcal{G}(\mathbf{q})$. The different location indices are $1, 2, 3, 4, 5, 6, 7, 8$ and the different non-location indices are $9, 10, 11, 12, 13, 14, 15$, thus $\ell = 8$, $N_{ind} = 7$. The multiplicity of the groups with different location indices are $m_1 = m_2 = m_6 = m_7 = m_8 = 1$, $m_3 = m_4 = 2$, $m_5 = 4$. For simplicity, in this example, we chose $1, 2, 3, 4, 5, 6, 7, 8$ to be the eight different location indices and we used them to label the groups as well. We also used the consecutive seven numbers for non-location indices. In general, both the location and non-location indices can be arbitrary numbers between 1 and N.

For brevity, we will often use the index associated to a vertex to refer to a vertex, e.g., when we refer to the index 2_d in (7.46), we really mean the vertex $(2, 11)$ since $q^2_{11} = 2_d$. This sometimes creates confusion (e.g., there are two vertices 9) and in that case, we will be specific.

All vertices with identical indices are connected by an edge, except that there is never an edge between any two vertices in the first row. Furthermore, there is an edge between 2_d and 6 (more precisely, between the vertices $(2, 11)$ and $(1, 11)$); similarly for 14_d and 8, but there is no edge between the non-location indices 9_d and their location indices 5 since they belong to a group with multiplicity bigger than one (four) due to the four location indices 5. The vertices with $2, 5, 9, 6$ (with common location index 3) the vertices with $12, 15$ (with location index 4) and the two 9's and 13's (with common location index 5) all receive a weight $1/2$. The weight of both 9_d's is 1 and all other vertices have weight zero. Notice that the index pair $(5, 9)$ appears twice but they are not twins (there are no twins inside a group), similarly the two $(5, 9_d)$ are not twin indices.

We will consider connected components of this graph. Due to the special rule involving duplexes, a connected component may contain different indices, for example

$$C = \{(1, 2), (2, 3), (2, 11), (3, 4), (1, 11)\} \tag{7.47}$$

is a connected component in (7.46), since $q^1_2 = q^2_3 = q^2_{11} = 2$, $q^3_4 = q^1_{11} = 6$ and $q^1_{11} = 6$ is connected to $q^2_{11} = 2_d$. With as slight abuse of notation, encoding the elements of C only with the indices q^i_α instead of the vertices (i, α) we can write $C = \{2(\text{loc.}), 2, 2_d, 6, 6(\text{loc.})\}$, where (loc.) refers to location index. The list of all connected components in (7.46) is

$$\begin{aligned}
&\{1, 10_d, 1\}, \quad \{11\}, \quad \{2(\text{loc.}), 2, 2_d, 6, 6(\text{loc.})\}, \quad \{3\}, \quad \{3\}, \quad \{4\}, \quad \{4\}, \quad \{7\}; \\
&\{5, 5(\text{loc.}), 5(\text{loc.}), 5(\text{loc.}), 5(\text{loc.})\}, \quad \{15\}, \quad \{12, 12\}, \quad \{9_d, 9_d, 9, 9, 9\}, \quad \{13, 13\}, \quad \{8, 14_d\},
\end{aligned} \tag{7.48}$$

using the shorter and somewhat ambiguous index-notation.

7.3.3 Estimates on the integrals

We now estimate $\widetilde{\Phi}^{\mathbf{n}}_{\mathbf{q},\boldsymbol{\nu}}$ from (7.31). Let $O = O(\mathbf{q}, \mathbf{n}, \boldsymbol{\nu})$ be the number of the off-diagonal Green functions appearing in the expansion of the right hand side of (7.31), i.e., in

$$\prod_{\alpha=1}^{p} V_{\mathbf{q}}(\underline{\mu}^\alpha, \underline{\nu}^\alpha, n^\alpha)\xi(\mathfrak{q}_\alpha) = \prod_{\alpha=1}^{p} (-1)^{n_\alpha} G^{[\alpha]}_{\mu_1^\alpha} h_{\nu_1^\alpha} G^{[\alpha]}_{\mu_2^\alpha} h_{\nu_2^\alpha} \ldots h_{\nu^\alpha_{n_\alpha}} G^{[\alpha]}_{\mu^\alpha_{n_\alpha+1}} \xi(\mathfrak{q}_\alpha) \tag{7.49}$$

(see (7.16) and (7.17)). Define

$$\widetilde{\Lambda}_o = \max_{\alpha=1,\ldots p;\, i\neq j} \big|G^{[\alpha]}_{ij}\big| \tag{7.50}$$

to be the maximum of the off-diagonal elements of the Green functions $G^{[\alpha]}$. Note that $\widetilde{\Lambda}_o$ is independent of the random variables h_ν, $\nu \in Q$. In particular, the bound $\widetilde{\Lambda}_o \le C/(\log N)^2 \le 1$ from (7.12) holds not only on Γ^c but on $(\Gamma^c)_1$ as well. Then, with $O = O(\mathbf{q}, \mathbf{n}, \boldsymbol{\nu})$, and using $n \le 6p$, we have

$$\big|\widetilde{\Phi}^{\mathbf{n}}_{\mathbf{q}}\big| = \Big|\mathbb{E}\mathbf{1}((\Gamma^c)_1)\sum_{\boldsymbol{\nu}}\prod_{\alpha=1}^{p} V(\underline{\mu}^\alpha, \underline{\nu}^\alpha, n^\alpha)\xi(\mathfrak{q}_\alpha)\Big| \le N^{-n/2-p}(Cp)^{Cp}\sum_{\boldsymbol{\nu}}\mathbb{E}\big[\mathbf{1}((\Gamma^c)_1)\big(\widetilde{\Lambda}_o\big)^O\big], \tag{7.51}$$

where for the expectation of the random variables h_ν, $\nu \in Q$, we have used estimate of the form

$$\mathbb{E}|h_1|^{a_1}\dots|h_k|^{a_k} \le \big(Cm^C N^{-1/2}\big)^m, \quad m := \sum_j a_j \tag{7.52}$$

for any a_j nonnegative integers, where the constant C depends only on ϑ. The total number of h factors appearing in (7.49) is $n_1 + n_2 + \ldots + n_p + 2p = n + 2p$, and (7.52) shows that their expectation can be bounded in terms of their total number $\sum_j a_j$ irrespective of the precise distribution of the individual exponents $a_1, a_2, \ldots a_k$. Thus $N^{-1/2}$ appears to the power $n + 2p$ in (7.51).

We also recall that the number of terms in the summation over $\boldsymbol{\nu} \in A(\mathbf{q}, \mathbf{n})$ in (7.51) is bounded by $(4p)^n$, see remark below (7.18).

Since we have $\tilde{\Lambda}_o \le 1$ on the set $(\Gamma^c)_1$, we also have the trivial estimate

$$\widetilde{\Lambda}_o^O \le N^{[p-O/2]_+}\big[\widetilde{\Lambda}_o^2 + N^{-1}\big]^p$$

where $[\,]_+$ denotes the positive part. Thus the main term in (7.36) is estimated as

$$\begin{aligned}
&\frac{1}{N^p}\sum_{\mathbf{q}}\sum_{n=0}^{6p}\sum_{|\mathbf{n}|=n}\big|\widetilde{\Phi}^{\mathbf{n}}_{\mathbf{q}}\big| \\
&\le (Cp)^{Cp}\mathbb{E}\big[\mathbf{1}((\Gamma^c)_1)\widetilde{\Lambda}_o^2 + N^{-1}\big]^p\sum_{G}\sum_{\mathbf{q}\,:\,\mathcal{G}(\mathbf{q})=G}\sum_{n=0}^{6p}\sum_{|\mathbf{n}|=n}\sum_{\boldsymbol{\nu}\in A(\mathbf{q},\mathbf{n})} N^{-2p-n/2+[p-O/2]_+}\mathbf{1}(\widetilde{\Phi}^{\mathbf{n}}_{\mathbf{q},\boldsymbol{\nu}} \ne 0).
\end{aligned}\tag{7.53}$$

From (7.29) we have the decomposition

$$\mathbf{1}((\Gamma^c)_1) = \mathbf{1}(\Gamma^c) + \mathbf{1}\Big((\Gamma^c)_1 \times Y^c \setminus \Gamma^c\Big) + \mathbf{1}((\Gamma^c)_1)\mathbf{1}(Y). \tag{7.54}$$

Since $\widetilde{\Lambda}_o \le 1$ on the set $(\Gamma^c)_1$, the contributions from the sets $(\Gamma^c)_1 \times Y^c \setminus \Gamma^c$ and $(\Gamma^c)_1 \times Y$ can be estimated in the same way as in (7.34), (7.35) by $C\exp\big[-c(\log N)^{\psi L}\big]$. Finally, we can use

$$\widetilde{\Lambda}_o^O \le 2\Lambda_o^O$$

on the set $\Gamma^c \subset (\Gamma^c)_1 \times Y^c$ (see (7.12)) and thus we can replace $\mathbb{E}\big[\mathbf{1}((\Gamma^c)_1)\widetilde{\Lambda}_o^2 + N^{-1}\big]^p$ in (7.53) by $2^p\mathbb{E}\big[\mathbf{1}(\Gamma^c)\Lambda_o^2 + N^{-1}\big]^p$ with a negligible error $C\exp\big[-c(\log N)^{\psi L}\big]$.

By (7.40) and Definition 7.2, the total number of different summation indices $\mathbf{q}$ in (7.53) is $N_{ind} + \ell$. We will prove that

$$2p + n + O \ge 2N_{ind} + 2\ell \tag{7.55}$$

and

$$4p + n \ge 2N_{ind} + 2\ell \tag{7.56}$$

hold for any $\mathbf{q}$, $\mathbf{n}$ and $\boldsymbol{\nu}$ for which $\widetilde{\Phi}^{\mathbf{n}}_{\mathbf{q},\boldsymbol{\nu}} \neq 0$. Since the summations over G, n, $\mathbf{n}$ and $\boldsymbol{\nu}$ give a factor at most p^{Cp}, these two inequalities imply that (7.53) is bounded by the right hand side of (4.5). This proves Lemma 4.1 assuming (7.55) and (7.56).

We now we prove (7.55) and (7.56). Recalling the total weight of the graph W satisfies $W \le 2(p-\ell)$ by (7.42), the inequality (7.55) is a consequence of the following

Proposition 7.1 *For any $\mathbf{q}$, $\mathbf{n}$ and $\boldsymbol{\nu}$ such that $\widetilde{\Phi}^{\mathbf{n}}_{\mathbf{q},\boldsymbol{\nu}} \neq 0$, we have*

$$W(\mathbf{q}) + |\mathbf{n}| + O(\mathbf{q}, \mathbf{n}, \boldsymbol{\nu}) \ge 2N_{ind}(\mathbf{q}). \tag{7.57}$$

Proof. We consider connected components C of the graph $\mathcal{G}(\mathbf{q})$. If a connected component consists of only one location index, we call it *trivial*, and we will consider only non-trivial components. Nontrivial components always contain at least one nonlocation vertex since location indices are never connected directly by an edge. We will prove that (7.57) holds for each nontrivial connected components and then we will sum these inequalities.

To formulate the statement precisely, we need a few notations. We will fix $\mathbf{q}$, $\mathbf{n}$ and $\boldsymbol{\nu} \in A(\mathbf{q}, \mathbf{n})$; all quantities in the following notations will depend on these parameters.

For each nontrivial connected component C of $\mathcal{G}(\mathbf{q})$, let I_C denote the set of all nonlocation indices appearing in C, i.e.,

$$I_C := \{q^i_\alpha : (i, \alpha) \in C, \ i = 2, 3\}, \tag{7.58}$$

and for the purpose of I_C we do not distinguish between indices with or without a possible d (duplex) subscript. Let L_C denote the set of all labels associated with C together with their transposes ν^t, where $\nu^t = (q, p)$ if $\nu = (p, q)$, i.e.,

$$L_C := \{(q^1_\alpha, q^i_\alpha) : (i, \alpha) \in C\} \cup \{(q^i_\alpha, q^1_\alpha) : (i, \alpha) \in C\}. \tag{7.59}$$

For example, $L_C = \{(2,3), (3,2), (6,2), (2,6), (3,6), (6,3)\}$ for the connected component C from (7.47). Let

$$n(C) = n(C; \mathbf{q}, \mathbf{n}, \boldsymbol{\nu}) := \sum_{\alpha=1}^{p} \sum_{m=1}^{n_\alpha} \mathbf{1}(\nu^\alpha_m \in L_C)$$

be the total number of h_ν-factors with $\nu \in L_C$ appearing in the expansion (7.49) without the h factors from $\prod_\alpha \xi(\mathfrak{q}_\alpha)$. Finally, we define $W(C) = W(C; \mathbf{q})$ as the total weight of the component C, i.e. the sum of the weights of vertices in C.

The following key quantity will be used to count the number of offdiagonal resolvent matrix elements appearing in the expansion.

Definition 7.3 *For $\sigma \in I_C$, let*

$$2O(\sigma) := \sum_{\alpha=1}^{p} \sum_{m=1}^{n_\alpha+1} \Big[\mathbf{1}([\mu^\alpha_m]_1 = \sigma, [\mu^\alpha_m]_2 \neq \sigma) + \mathbf{1}([\mu^\alpha_m]_2 = \sigma, [\mu^\alpha_m]_1 \neq \sigma)\Big],$$

i.e., $2O(\sigma)$ is the number of times that σ appears as one of the two indices of an off-diagonal Green function in the expansion (7.49)*. Let*

$$O(C) = O(C; \mathbf{q}, \mathbf{n}, \boldsymbol{\nu}) := \sum_{\sigma \in I_C} O(\sigma) \tag{7.60}$$

i.e., $2O(C)$ is the number of times that an index associated with C appears in an off-diagonal Green function in (7.49)*.*

Note that we do not directly count the total number O of off-diagonal resolvent matrix elements, we rather count how often a fixed non-location index contributes to an off-diagonal Green function factor. In this way we can determine how much each non-location index contributes to off-diagonal matrix elements and we can perform our estimates for each component separately.

By definition of the edges in the graph, two different nontrivial components C_1, C_2 have disjoint sets of nonlocation indices; $I_{C_1} \cap I_{C_2} = \emptyset$. As a corollary, the sets L_C for different components are also disjoint since the twins are eliminated and for any fixed $\mathbf{q}, \mathbf{n}$ and $\boldsymbol{\nu}$ we have

$$\sum_C n(C; \mathbf{q}, \mathbf{n}, \boldsymbol{\nu}) \leq |\mathbf{n}|, \qquad \sum_C O(C; \mathbf{q}, \mathbf{n}, \boldsymbol{\nu}) \leq O(\mathbf{q}, \mathbf{n}, \boldsymbol{\nu}), \tag{7.61}$$

where the summations are over all nontrivial connected components. Strict inequality can happen as there are indices left out in twins. Moreover, we define

$$N_{ind}(C) = N_{ind}(C; \mathbf{q}) := \Big| \{q_\alpha^j : 2 \leq j \leq 3,\ 1 \leq \alpha \leq p\ ,\ (j, \alpha) \in C\} \setminus \{q_\alpha^1 : 1 \leq \alpha \leq p\} \Big|$$

to be the number of independent nonlocation indices in the component C. This is the same concept as $N_{ind}(\mathbf{q})$ defined in (7.43) but restricted to a fixed component C. We clearly have

$$\sum_C W(C) = W, \qquad \sum_C N_{ind}(C; \mathbf{q}) = N_{ind}(\mathbf{q}). \tag{7.62}$$

We will prove below that (7.57) holds in each nontrivial component C, i.e. for $\widetilde{\Phi}^{\mathbf{n}}_{\mathbf{q},\boldsymbol{\nu}} \neq 0$, we have

$$W(C; \mathbf{q}) + n(C; \mathbf{q}, \mathbf{n}, \boldsymbol{\nu}) + O(C; \mathbf{q}, \mathbf{n}, \boldsymbol{\nu}) \geq 2N_{ind}(C; \mathbf{q}). \tag{7.63}$$

then (7.57) will follow from (7.61) and (7.62).

Lemma 7.2 *Let C be a nontrivial connected component of $\mathcal{G}(\mathbf{q})$. Then $N_{ind}(C) \leq 1$.*

Proof. Suppose that C contains at least two different independent nonlocation indices $q_\alpha^j \neq q_\beta^i$ and consider a path in $\mathcal{G}(\mathbf{q})$ connecting their vertices $W_1 = (j, \alpha)$ and $W_2 = (i, \beta)$. Along this path there must be two subsequent vertices whose indices are different. Considering the construction of $\mathcal{G}(\mathbf{q})$, this can happen only along an edge created by the special rule in Step 4 in the definition of $\mathcal{G}(\mathbf{q})$, i.e. there is a duplex connected to its location vertex (any other edge connects identical indices). For definiteness, we may choose the notation W_1 and W_2 in such a way that along the path from W_1 to W_2 the first special edge created by Step 4 with different indices is reached at its non-location vertex (duplex vertex), call it U_1. Clearly U_1 and W_1 have the same index. Let now E be the edge connecting U_1 to its location vertex $V_1 \in C$, then by the choice of U_1 the index of V_1 differs from that of U_1. Let D be the set of all vertices with the same value as U_1 and let D_1 be the set of all vertices with the same value as V_1, then D and D_1 are disjoint subsets of C.

We claim that apart from V_1, D_1 consists of nonlocation vertices only. Suppose this is not the case. Then there is another location vertex V_1' taking the same value as V_1. But this implies that V_1 and V_1' belong to a group with multiplicity at least two. In this case, however, we did not connect the duplex V to its location vertex and this leads to contradiction.

The number of independent nonlocation indices in D is exactly one, namely the index of W_1. The number of independent nonlocation indices in D_1 is zero since they take the same value as a location index.

Suppose that $D \cup D_1$ did not exhaust C. In order that $D_1 \cup D$ is connected to another vertex with a different value, once again, there must be an edge E' connecting a duplex vertex to its location vertex; one of these two vertices must be $D_1 \cup D$, the other one must be in the complement. We claim that the duplex is in $D_1 \cup D$. Indeed, the location vertex cannot be in $D_1 \cup D$, since D has no location vertex at all (otherwise the index of U_1 would not be independent) and D_1 has only one location index, V_1, that is already connected within $D \cup D_1$ to its duplex.

Let U_2 denote the duplex in $D \cup D_1$ that is connected to its location index $V_2 \notin D \cup D_1$ and let D_2 denote the set of vertices with the same value as V_2. As before, we can establish that D_2 contains only non-location indices, apart from V_2, and there is no independent nonlocation index in D_2.

If $D \cup D_1 \cup D_2$ did not exhaust C, we continue the process by defining new sets D_3, D_4, etc. until C is exhausted, but we never get a new independent nonlocation index. This proves that $N_{ind}(C) \leq 1$.

□

We can start proving (7.63). We fix the parameters $\mathbf{q}, \mathbf{n}$ and $\boldsymbol{\nu}$ and omit them from the notation. We will distinguish the following cases that clearly cover all possibilities.

Case 1. C consists of a duplex $(q^2_\alpha)_d$ and its location index q^1_α.

Setting $\nu := (q^1_\alpha, q^2_\alpha)$, we know, in particular, that h_ν or h_{ν^t} do not appear in any other $\xi(\mathfrak{q}_\beta)$, $\beta \neq \alpha$ since C is an isolated component, not connected to any other vertices. Then, by the observation made in (7.37), h_ν (or h_{ν^t}) must explicitly appear in (7.38) and it clearly must appear in one of the following ways, with some $\beta \neq \alpha$,

$$(1): \qquad G^{[\beta]}_{f_1, q^1_\alpha} h_{q^1_\alpha q^2_\alpha} G^{[\beta]}_{q^2_\alpha, f_2}, \quad \text{or} \quad G^{[\beta]}_{f_2, q^2_\alpha} h_{q^2_\alpha q^1_\alpha} G^{[\beta]}_{q^2_\alpha, f_2}, \quad f_i \neq q^i_\alpha \tag{7.64}$$

$$(2): \qquad h_{q^1_\alpha q^2_\alpha} G^{[\beta]}_{q^2_\alpha, q^2_\alpha}\, h_{q^2_\alpha q^1_\alpha}, \quad \text{or} \quad h_{q^2_\alpha q^1_\alpha} G^{[\beta]}_{q^1_\alpha, q^1_\alpha} h_{q^1_\alpha q^2_\alpha}. \tag{7.65}$$

The main reason why only one of these possibilities occurs is because the indices $q^i_\alpha, i = 1, 2$, appear only in C. So either (1) both Green functions neighboring h_ν (or h_{ν^t}) are off-diagonal, or (2) either of the neighboring Green function is diagonal. In the latter case, however, the expansion must continue on the other side of this diagonal Green function with another factor $h_{q^1_\alpha q^2_\alpha}$ (or $h_{q^2_\alpha q^1_\alpha}$). The reason for this last statement is that the expansion cannot start or terminate with a diagonal Green function of the form $G^{[\beta]}_{q^1_\alpha, q^1_\alpha}$ or $G^{[\beta]}_{q^2_\alpha, q^2_\alpha}$ since that would entail that q^1_α (or q^2_α) equals to q^2_β or q^3_β, which would mean that C contained other elements as well.

In the first case, $n(C) \geq 1$ and we have identified two indices of off-diagonal Green functions associated with q^2_α, i.e. $O(C) \geq 1$. In the second case, we find that h_ν or h_{ν^t} appear altogether twice and hence $n(C) \geq 2$. Since $N_{ind}(C) = 1$ in this case, we have thus proved that in both cases

$$W(C) + n(C) + O(C) \geq 2 = 2N_{ind}(C). \tag{7.66}$$

Notice that we did not use weight $W(C)$ here.

Case 2. C is an isolated non-duplex vertex.

Since C is nontrivial, we can assume that C consists of a single vertex $(2, \alpha)$ (the case of $(3, \alpha)$ is identical). Let $\nu := (q^1_\alpha, q^2_\alpha)$. Consider the expansion of $G^{[\alpha]}$, see (7.16). The first and the last Green functions in this expansion will be called *extreme* Green functions; if $n_\alpha = 0$, then the single Green function $G^{(q^1_\alpha)}_{q^2_\alpha q^3_\alpha}$ will be called extreme. Since this expansion contains h_μ factors only with $\mu \in Q^{(\alpha)}$ and $q^2_\alpha \neq q^i_\beta$ for any $\beta \neq \alpha$ (since C is an isolated vertex), thus q^2_α cannot appear as an index of any

h_μ. Then the first Green function in (7.16) must be of the form $G^{[\alpha]}_{q^2_\alpha,f}$ with some $f \neq q^2_\alpha$, i.e. it must be off-diagonal, thus $O(C) \geq \frac{1}{2}$. Furthermore, h_ν or h_{ν^t} must appear as

$$(1): \qquad h_{q^1_\alpha q^2_\alpha} G^{[\beta]}_{q^2_\alpha,f} \quad \text{or} \quad G^{[\beta]}_{f,q^2_\alpha} h_{q^2_\alpha q^1_\alpha}, \quad f \neq q^2_\alpha \tag{7.67}$$

$$(2): \qquad h_{q^1_\alpha q^2_\alpha} G^{[\beta]}_{q^2_\alpha,q^2_\alpha} h_{q^2_\alpha q^1_\alpha} \quad \text{or} \quad h_{q^2_\alpha q^1_\alpha} G^{[\beta]}_{q^1_\alpha,q^1_\alpha} h_{q^1_\alpha q^2_\alpha} \tag{7.68}$$

In the first case (1), we have identified another index of off-diagonal Green function associated with q^2_α, so $O(C) \geq 1$ and $n(C) \geq 1$. In the second case (2), we find that h_ν and h_{ν^t} appears altogether twice and thus $n(C) \geq 2$. In both cases we have proved (7.63) since $N_{ind}(C) = 1$. Again, the weight $W(C)$ was not used.

Case 3. C has only one non-location vertex, (i, α), $i = 2, 3$ and at least one location vertex $(1, \beta)$ with $\beta \neq \alpha$.

In this case the non-location index q^i_α is equal to a location index, hence $N_{ind}(C) = 0$ and (7.63) is obvious.

Case 4. C has more than one non-location vertex.

Suppose the weight of a non-location vertex $(2, \alpha)$ in C is zero. Then $h_{q^1_\alpha q^2_\alpha}$ (or $h_{q^2_\alpha q^1_\alpha}$) must appear in (7.49) (apart from the ξ factors) and thus it contributes to $n(C)$ by one. Here we are using the following reason:

(†) If $h_{q^1_\alpha q^2_\alpha}$ and $h_{q^2_\alpha q^1_\alpha}$ appear in $\prod_\beta \xi(\mathfrak{q}_\beta)$ at least twice, then either $(2, \alpha)$ is a twin vertex or the multiplicity of the group containing $(2, \alpha)$ is more than one.

Both cases contradict our definitions; twins are not part of $\mathcal{G}(\mathbf{q})$, and non-location vertices in groups with higher multiplicity have nonzero weight. But if $h_{q^1_\alpha q^2_\alpha}$ and $h_{q^2_\alpha q^1_\alpha}$ appear only once in $\prod_\beta \xi(\mathfrak{q}_\beta)$ (namely, only in the factor $\xi(\mathfrak{q}_\alpha)$), then at least one of them need to appear at least one more times in (7.49) to make the expectation nonzero.

Hence if we have at least two weight zero non-location vertices in C, then $n(C) \geq 2$ and (7.63) holds. Note that each of these two vertices contribute to $n(C)$ by one, since together with their own location vertex they must form two different labels, otherwise they would be part of a twin or a group with multiplicity at least 1 and their weight would not be zero. We can also assume that the total weight $W(C)$ is less than 2 or, if there is a weight zero non-location vertex, hence $n(C) \geq 1$, then the total weight is at most $W(C) \leq 1/2$. In all other cases (7.63) follows trivially from $N_{ind}(C) \leq 1$.

So we only have to consider the following remaining cases:

1. *The non-location vertices of C consist of exactly two weight $1/2$ vertices v_1, v_2.*

   First notice that these two vertices must have the same index. Otherwise they could be in the same connected component only if one of them, say v_2, would be equal to a duplex $(q^2_\beta)_d$ with some $\beta \neq \alpha$ where $v_1 = q^j_\alpha$ ($j \in \{2, 3\}$), and this duplex would belong to a group with multiplicity one (a connecting edge between vertices with different indices can be provided only via a special edge from Step 4. between a duplex and its location vertex and only if the corresponding group has multiplicity one). But in this case the weight of the non-location vertex $(2, \beta)$ in C would be zero by (i) of Definition 7.1.

Thus the two vertices v_1, v_2 cannot be in the same column of the matrix (otherwise they formed a duplex), so without loss of generality we can assume that they are of the form $(2,\alpha)$ and $(2,\beta)$ with $\alpha \neq \beta$ and we know that $q_\alpha^2 = q_\beta^2$.

Consider first the case $q_\alpha^1 \neq q_\beta^1$. By the fact that the common value $q_\alpha^2 = q_\beta^2$ appears only twice in C, both factors $h_{q_\alpha^1 q_\alpha^2}$ and $h_{q_\beta^1 q_\beta^2}$ (or their transposes) have to appear in (7.49). Thus $n(C) \geq 2$ and (7.63) holds.

Finally, consider the case $q_\alpha^1 = q_\beta^1$. Since q_α^2 and q_β^2 have weight 1/2, they are not duplex. By construction, we have to expand the Green function $G^{(q_\alpha^1)}_{q_\alpha^2, q_\alpha^3}$. Since $q_\alpha^2 \neq q_\alpha^3$, in the expansion (7.49), the first Green function $G^{[\alpha]}_{\mu_1^\alpha}$ is off-diagonal (otherwise the beginning of the expansion were $G^{[\alpha]}_{q_\alpha^2, q_\alpha^2} h_{q_\alpha^2 q_\alpha^1} \ldots$, but $h_{q_\alpha^2 q_\alpha^1}$ cannot appear in the expansion of $G^{(q_\alpha^1)}_{q_\alpha^2, q_\alpha^3}$). Hence q_α^2 appears as an index of an extreme off-diagonal Green function. Similar statement holds for q_β^2. Hence we have identified two indices of off-diagonal Green functions associated with C so that $O(C) \geq 1$ and together with $W(C) \geq 1$ we obtain that (7.63) holds.

2. *The non-location vertices of C consist of exactly one weight 1/2 vertex, and one weight* 1 *vertex.*

   Since the weight 1 vertex is a duplex, these two vertices cannot be in the same column of $\mathbf{q}$. Without loss of generality, let $(2,\alpha)$ be the weight 1/2 vertex and let $(2,\beta)_d$ be the weight 1 vertex, $\alpha \neq \beta$. We can consider two cases: $q_\alpha^1 \neq q_\beta^1$ and $q_\alpha^1 = q_\beta^1$. As before, for the first case, $n(C) \geq 1$. For the second case, q_α^2 cannot appear as an index of any h_ν in any other $\xi(\mathfrak{q}_\gamma)$ for $\gamma \neq \alpha, \beta$ since C consist of exactly two columns, namely the columns α and β. Thus $h_{q_\alpha^1, q_\alpha^2}$ or its transpose must appear in the expansion of $G^{[\alpha]}$ and therefore we can find q_α^2 as one of the indices of an extreme off-diagonal Green function. Hence we have $O(C) \geq 1/2$ in the second case. Since $W(C) \geq 3/2$, we obtain in both cases that (7.63) holds.

3. *The non-location vertices of C consist of exactly one weight* $1/2$ *vertex one weight zero vertex.*

   Since the two vertices have different weights, they are in different columns of the matrix. Without loss of generality, we can assume that the weight 1/2 vertex is $(2,\alpha)$ and the weight zero vertex is $(2,\beta)$ with $\alpha \neq \beta$. In this case, both $h_{q_\alpha^1 q_\alpha^2}$ and $h_{q_\beta^1 q_\beta^2}$ (or their transposes) have to appear in the expansion, thus $n(C) \geq 2$ and (7.63) holds.

4. *The non-location vertices of C consist of exactly three weight* $1/2$ *vertices.*

   Similar arguments as in the first case, we can show that these three vertices are in different columns and we can thus assume that they are of the form $(2,\alpha)$, $(2,\beta)$ and $(2,\gamma)$ with different α, β, γ. If $q_\alpha^1 = q_\beta^1 = q_\gamma^1$, then q_α^2 appears as an index of an extreme off-diagonal Green function in the expansion of $G^{(q_\alpha^1)}_{q_\alpha^2, q_\alpha^3}$ and $O(C) \geq 1/2$. On the other hand, if one of the three location indices, say q_α^1, differed from the other two, then $h_{q_\alpha^1 q_\alpha^2}$ (or its transpose) have to appear in the expansion and $n(C) \geq 1$. In either case, together with $W(C) \geq 3/2$, we obtain (7.63).

The main reason of the previous proof is that any weight 1/2 vertex either associated with an index of an extreme off-diagonal Green function or there is an h factor associated with it. We have thus proved Proposition 7.1 □

Finally, we have to prove the inequality (7.56). Let d denote the number of duplexes. Let a_1 be the number of nontrivial components C that contain only one non-location vertex and let a_2 be the number of nontrivial components C that contain at least two nonlocation vertices. Since by Lemma 7.2 we have $N_{ind} = \sum_C N_{ind}(C) \leq a_1 + a_2$ and obviously $\ell \leq p$, it is sufficient to show that

$$2p + n \geq 2(a_1 + a_2).$$

Since we there are $2p - d$ nonlocation vertices, we have $2p - d \geq a_1 + 2a_2$, thus it is sufficient to show that $n + d \geq a_1$. But each component with a single non-location vertex, say $(2, \alpha)$, is either a duplex or it gives rise to a factor $h_{q^1_\alpha q^2_\alpha}$ (or its transpose) that must appear in the expansion, hence it contributes to n. This shows (7.56) and this completes the proof of Lemma 4.1. □

References

[1] Alon, N.; Krivelich, M.; Vu, V.: On the concentration of eigenvalues of random symmetric matrices. *Israel J. Math.* **131** (2002), 259-267.

[2] Anderson, G., Guionnet, A., Zeitouni, O.: An Introduction to Random Matrices. Studies in advanced mathematics, **118**, Cambridge University Press, 2009.

[3] Anderson, G.; Zeitouni, O. : A CLT for a band matrix model. *Probab. Theory Related Fields* **134** (2006), no. 2, 283–338.

[4] Auffinger, A., Ben Arous, G., Péché, S.: Poisson Convergence for the largest eigenvalues of heavy-taled matrices. *Ann. Inst. Henri Poincar??? Probab. Stat.* **45** (2009), no. 3, 589–610.

[5] Bai, Z. D., Miao, B., Tsay, J.: Convergence rates of the spectral distributions of large Wigner matrices. *Int. Math. J.* **1** (2002), no. 1, 65–90.

[6] Ben Arous, G., Péché, S.: Universality of local eigenvalue statistics for some sample covariance matrices. *Comm. Pure Appl. Math.* **LVIII.** (2005), 1–42.

[7] Biroli, G., Bouchaud,J.-P., Potters, M.: On the top eigenvalue of heavy-tailed random matrices, *Europhysics Letters*, **78** (2007), 10001.

[8] Bleher, P., Its, A.: Semiclassical asymptotics of orthogonal polynomials, Riemann-Hilbert problem, and universality in the matrix model. *Ann. of Math.* **150** (1999), 185–266.

[9] Brézin, E., Hikami, S.: Correlations of nearby levels induced by a random potential. *Nucl. Phys. B* **479** (1996), 697–706, and Spectral form factor in a random matrix theory. *Phys. Rev. E* **55** (1997), 4067–4083.

[10] Deift, P.: Orthogonal polynomials and random matrices: a Riemann-Hilbert approach. *Courant Lecture Notes in Mathematics* **3**, American Mathematical Society, Providence, RI, 1999

[11] Deift, P., Gioev, D.: Random Matrix Theory: Invariant Ensembles and Universality. *Courant Lecture Notes in Mathematics* **18**, American Mathematical Society, Providence, RI, 2009

[12] Deift, P., Kriecherbauer, T., McLaughlin, K.T-R, Venakides, S., Zhou, X.: Uniform asymptotics for polynomials orthogonal with respect to varying exponential weights and applications to universality questions in random matrix theory. *Comm. Pure Appl. Math.* **52** (1999):1335–1425.

[13] Deift, P., Kriecherbauer, T., McLaughlin, K.T-R, Venakides, S., Zhou, X.: Strong asymptotics of orthogonal polynomials with respect to exponential weights. *Comm. Pure Appl. Math.* **52** (1999): 1491–1552.

[14] Disertori, M., Pinson, H., Spencer, T.: Density of states for random band matrices. *Commun. Math. Phys.* **232**, 83–124 (2002)

[15] Dyson, F.J.: A Brownian-motion model for the eigenvalues of a random matrix. *J. Math. Phys.* **3**, 1191-1198 (1962)

[16] Erdős, L., Schlein, B., Yau, H.-T.: Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices. *Ann. Probab.* **37**, No. 3, 815–852 (2009)

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

[18] Erdős, L., Schlein, B., Yau, H.-T.: Wegner estimate and level repulsion for Wigner random matrices. *Int. Math. Res. Notices.* **2010**, No. 3, 436-479 (2010)

[19] Erdős, L., Schlein, B., Yau, H.-T.: Universality of random matrices and local relaxation flow. *Invent. Math.* **185** (2011), no.1, 75–119.

[20] Erdős, L., Péché, G., Ramírez, J., Schlein, B., and Yau, H.-T., Bulk universality for Wigner matrices. *Commun. Pure Appl. Math.* **63**, No. 7, 895–925 (2010)

[21] Erdős, L., Schlein, B., Yau, H.-T., Yin, J.: The local relaxation flow approach to universality of the local statistics for random matrices. To appear in Annales Inst. H. Poincaré (B), Probability and Statistics. Preprint arXiv:0911.3687

[22] Erdős, L., Yau, H.-T., Yin, J.: Bulk universality for generalized Wigner matrices. To appear in Prob. Theor. Rel. Fields. Preprint arXiv:1001.3453

[23] Erdős, L., Yau, H.-T., Yin, J.: Universality for generalized Wigner matrices with Bernoulli distribution. *J. of Combinatorics*, **1** (2011), no. 2, 15–85

[24] Guionnet, A.: Large deviation upper bounds and central limit theorems for band matrices, *Ann. Inst. H. Poincaré Probab. Statist* **38** , (2002), 341-384.

[25] Gustavsson, J.: Gaussian Fluctuations of Eigenvalues in the GUE, *Ann. Inst. H. Poincaré Probab. Statist.* **41** (2005), no. 2, 151-178

[26] Johansson, K.: Universality of the local spacing distribution in certain ensembles of Hermitian Wigner matrices. *Comm. Math. Phys.* **215** (2001), no.3. 683–705.

[27] Johansson, K.: Universality for certain hermitian Wigner matrices under weak moment conditions. Preprint arxiv.org/abs/0910.4467

[28] Mehta, M.L.: Random Matrices. Academic Press, New York, 1991.

[29] O'Rourke, S.: Gaussian Fluctuations of Eigenvalues in Wigner Random Matrices, *J. Stat. Phys.*, **138** (2009), no.6., pp.1045-1066

[30] Pastur, L., Shcherbina M.: Bulk universality and related properties of Hermitian matrix models. *J. Stat. Phys.* **130** (2008), no.2., 205-250.

[31] Péché, S., Soshnikov, A.: On the lower bound of the spectral norm of symmetric random matrices with independent entries. *Electron. Commun. Probab.* **13** (2008), 280–290.

[32] Péché, S., Soshnikov, A.: Wigner random matrices with non-symmetrically distributed entries. *J. Stat. Phys.* **129** (2007), no. 5-6, 857–884.

[33] Ruzmaikina, A.: Universality of the edge distribution of eigenvalues of Wigner random matrices with polynomially decaying distributions of entries, *Comm. Math. Phys.* **261** (2006), no. 2, 277–296.

[34] Sinai, Y. and Soshnikov, A.: A refinement of Wigner's semicircle law in a neighborhood of the spectrum edge. *Functional Anal. and Appl.* **32** (1998), no. 2, 114–131.

[35] Sodin, S.: The spectral edge of some random band matrices. *Ann. of Math.* **172** (2010), No. 3, 2223-2251

[36] Sodin, S.: The Tracy–Widom law for some sparse random matrices. *J. Stat. Phys.* **136** (2009), no. 5, 834-841

[37] Soshnikov, A.: Universality at the edge of the spectrum in Wigner random matrices. *Comm. Math. Phys.* **207** (1999), no.3. 697-733.

[38] A. Soshnikov, Poisson statistics for the largest eigenvalues of Wigner matrices with heavy tails, *Elect. Commun. in Probab.*, **9** (2004), 82 - 91

[39] Spencer, T.: Review article on random band matrices. Draft in preparation.

[40] Tao, T. and Vu, V.: Random matrices: Universality of the local eigenvalue statistics, to appear in *Acta Math.*, Preprint. arXiv:0906.0510.

[41] Tao, T. and Vu, V.: Random matrices: Universality of local eigenvalue statistics up to the edge. Preprint. arXiv:0908.1982

[42] Tao, T. and Vu, V.: Random matrices: localization of the eigenvalues and the necessity of four moments. Preprint. arXiv:1005.2901

[43] C. Tracy, H. Widom, Level-Spacing Distributions and the Airy Kernel, *Comm. Math. Phys.* **159** (1994), 151-174.

[44] C. Tracy, H. Widom, On orthogonal and symplectic matrix ensembles, *Comm. Math. Phys.* **177** (1996), no. 3, 727-754.

[45] Wigner, E.: Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math.* **62** (1955), 548-564.