Probab. Theory Relat. Fields (2015) 163:1–59
DOI 10.1007/s00440-014-0586-4

Bounds for the Stieltjes transform and the density of states of Wigner matrices

Claudio Cacciapuoti · Anna Maltsev · Benjamin Schlein

Received: 12 December 2013 / Revised: 15 October 2014 / Published online: 2 November 2014

Abstract We consider ensembles of Wigner matrices, whose entries are (up to the symmetry constraints) independent and identically distributed random variables. We show the convergence of the Stieltjes transform towards the Stieltjes transform of the semicircle law on optimal scales and with the optimal rate. Our bounds improve previous results, in particular from Erdős et al. (Adv Math 229(3):1435–1515, 2012; Electron J Probab 18(59):1–58, 2013), by removing the logarithmic corrections. As applications, we establish the convergence of the eigenvalue counting functions with the rate $(\log N)/N$ and the rigidity of the eigenvalues of Wigner matrices on the same scale. These bounds improve the results of Erdős et al. (Adv Math 229(3):1435–1515, 2012; Electron J Probab 18(59):1–58, 2013), Götze and Tikhomirov (2013).

Keywords Random matrices · Wigner matrices · Rigidity of the eigenvalues · Rate of convergence

C. Cacciapuoti · B. Schlein
Hausdorff Center for Mathematics, Institute for Applied Mathematics, University of Bonn, Endenicher Allee 60, 53115 Bonn, Germany

B. Schlein
e-mail: benjamin.schlein@hcm.uni-bonn.de

Present Address:
C. Cacciapuoti (✉)
Dipartimento di Scienza e Alta Tecnologia, Università dell'Insubria, Via Valleggio 11, 22100 Como, Italy
e-mail: claudio.cacciapuoti@uninsubria.it

A. Maltsev
Department of Mathematics, University of Bristol, University Walk, Clifton, Bristol BS8 1TW, UK
e-mail: annavmaltsev@gmail.com

Mathematics Subject Classification 60B20 · 60B12 · 47B80

1 Introduction and main results

Let H be an $N \times N$ Wigner matrix, whose entries are independent (up to the symmetry constraints) and identically distributed random variables with zero mean and a fixed variance. We are interested in the statistical properties of the eigenvalues of H in the limit of large N. As already shown by Wigner in [37], the density of the eigenvalues of H converges, after appropriate normalization, towards the famous semicircle law

$$\rho_{sc}(x) = \begin{cases} \frac{1}{\pi}\sqrt{1-\frac{x^2}{4}} & \text{if } |x| \leq 2 \\ 0 & \text{if } |x| > 2 \end{cases}$$

Wigner's result concerns the density of states of H on intervals with length of order one, containing typically order N eigenvalues (we normalize Wigner matrices so that the typical distance between eigenvalues in the bulk of the spectrum is of the order $1/N$). It is natural to ask what happens on smaller intervals, in which the typical number of eigenvalues is large, but not macroscopically large. This question was first addressed in [15–17], where the density of states of H was proven to converge towards the semicircle law on microscopic intervals, containing a large but fixed number of eigenvalues (independent of N). In order to establish the convergence towards the semicircle law, it is useful to study the Stieltjes transform $m(z)$ of the Wigner matrix H, defined for Im z > 0 by

$$m(z) = \frac{1}{N} \text{Tr} \; \frac{1}{H-z} = \frac{1}{N} \sum_{\alpha=1}^{N} \frac{1}{\lambda_\alpha - z} \tag{1.1}$$

where $\lambda_\alpha, \alpha = 1, \ldots, N$, denote the eigenvalues of H. We observe that

$$\text{Im}\, m(E+i\eta) = \frac{1}{N} \sum_{\alpha=1}^{N} \frac{\eta}{(\lambda_\alpha - E)^2 + \eta^2}. \tag{1.2}$$

From (1.2) it is clear that the density of states in an interval of size η around E is closely related to the imaginary part of the Stieltjes transform at the point $z = E + i\eta$. To get information on the density of states on a small scale of order η, we need to control the Stieltjes transform m at distances of order η from the real axis. In the limit of large N, $m(z)$ approaches the Stieltjes transform of the semicircle law, given for all $x \in \mathbb{C}\backslash[-2; 2]$ by

$$m_{sc}(z) = \int dx \, \frac{\rho_{sc}(x)}{x-z}. \tag{1.3}$$

There are many reasons for analyzing the Stieltjes transform instead of looking directly at the density of states. Since every eigenvalue contributes to m, it seems that m should be more stable with respect to small fluctuations of the eigenvalues. More importantly, the Stieltjes transform m_{sc} of the semicircle law satisfies the simple fixed point equation

$$m_{sc}(z) + \frac{1}{z + m_{sc}(z)} = 0. \tag{1.4}$$

As a consequence, to prove that m is close to m_{sc}, it is enough to show that m is an approximate solution of (1.4). Using this strategy and assuming the entries of H to have subgaussian tails, it was shown in [17] (improving previous results from [15,16]) that, for every $\kappa > 0$, there exist constants $C, c > 0$ such that

$$\mathbb{P}\left(|m(E + i\eta) - m_{sc}(E + i\eta)| \geq \delta\right) \leq Ce^{-c\delta\sqrt{N\eta}} \tag{1.5}$$

for all $|E| \leq 2-\kappa$ (i.e. for E in the bulk of the spectrum). The estimate (1.5) establishes the convergence of the Stieltjes transform $m(E + i\eta)$ for all $\eta \gg 1/N$ and implies therefore the convergence of the density of states on all intervals of size $\eta \gg 1/N$. It is important to observe that the scale $\eta \simeq 1/N$ is optimal. For $\eta \lesssim 1/N$, the number of eigenvalues contained in an interval of size η is small and therefore the fluctuations of the density of states cannot be neglected. It is clear therefore that on scales $\eta \ll 1/N$ it is impossible to have convergence in probability for the Stieltjes transform and for the density of states (one has, however, convergence in expectation of the density of states, if the probability density of the matrix entries is sufficiently regular; see [28]). While (1.5) is optimal in the sense that it gives convergence for all $\eta \gg 1/N$, it turns out that it is not optimal in the size of the fluctuations around m_{sc}. From (1.5), fluctuations are at most of the order $(N\eta)^{-1/2}$. This bound was substantially improved in [22], where the authors showed that, with high probability,

$$|m(E + i\eta) - m_{sc}(E + i\eta)| \lesssim \frac{(\log N)^c}{N\eta} \tag{1.6}$$

uniformly in E. In fact, the results of [22] also apply to generalized Wigner matrices, whose entries are not required to be identically distributed. Hence, up to logarithmic corrections, the fluctuations around m_{sc} are of the order $(N\eta)^{-1}$. An important observation which was used in [22] to show (1.6) is the fact that not only the Stieltjes transform, given by the average of the diagonal entries $G_{jj}(z) = (H - z)^{-1}$ of the resolvent, converges towards m_{sc}, but that also each $G_{jj}(z)$ approaches the same limit. Making use of a vectorial fixed point equations for the diagonal entries of the resolvent, it was proved in [22] that

$$|G_{jj}(E + i\eta) - m_{sc}(E + i\eta)| \lesssim \frac{(\log N)^c}{\sqrt{N\eta}}. \tag{1.7}$$

For the diagonal entries G_{jj} of the resolvent, the rate predicted by (1.7) turns out to be optimal, up to the logarithmic corrections. In order to improve the bound (1.7) to the stronger rate on the r.h.s. of (1.6), two important observations are needed (which go back to [22]); first of all, the expectation of $G_{jj} - m_{sc}$ turns out to be of order $(N\eta)^{-1}$, hence much smaller than the typical size of $|G_{jj} - m_{sc}|$. On the other hand, the different resolvent entries are only weakly dependent on each other. As a consequence, the average

$$m(E+i\eta) - m_{sc}(E+i\eta) = \frac{1}{N}\sum_{j=1}^{N}\big(G_{jj}(E+i\eta) - m_{sc}(E+i\eta)\big)$$

exhibits substantial cancellations, which, similarly to what happens in the central limit theorem, make its typical size much smaller than that of the single summands (in fact the argument used in [22] is slightly different, controlling $m - m_{sc}$ in terms of the average of certain error terms Z_j which are not exactly the same as $G_{jj} - m_{sc}$; while each Z_j has a typical size of order $(N\eta)^{-1/2}$, the average is much smaller because they are only weakly dependent random variables). To establish the validity of this picture, terms arising in high moments of $m - m_{sc}$ are organized in [22] in contributions which are either small or independent of an increasing number of variables. More recently, a simpler expansion algorithm leading to bounds of the form (1.6) has been developed in [10]; this paper is the basis for our analysis.

Our goal is to obtain estimates of the form (1.6) but without logarithmic corrections. To simplify the analysis, we restrict our attention to Wigner matrices with identically distributed entries. We believe, however, that similar arguments would also work for generalized Wigner matrices, as considered in [10,22]. The bounds in Theorem 1 below establish the convergence of the imaginary part of the Stieltjes transform of Wigner matrices towards the imaginary part of the Stieltjes transform of the semicircle law on the optimal scale $\eta \simeq 1/N$ and with the optimal rate $(N\eta)^{-1}$, for arbitrary energies $E \in \mathbb{R}$. For energies inside the spectrum of H, Theorem 1 also gives convergence of the real part of the Stieltjes transform on the optimal scale and with the optimal rate; more precisely, it shows that

$$|m(E+i\eta) - m_{sc}(E+i\eta)| \lesssim \frac{1}{N\eta}$$

for all $|E| \leq 2 + \eta$. The reason why we cannot prove the convergence of the real part of the Stieltjes transform on the optimal scale and with the optimal rate outside the spectrum of H (i.e. for $|E| > 2 + \eta$) is an instability of the equation for the difference $\Lambda(E+i\eta) = m(E+i\eta) - m_{sc}(E+i\eta)$; we will discuss this point after Proposition 2.2. If we assume E to be well inside the bulk of the spectrum of H (i.e. $|E| \leq 2 - \kappa$, for an arbitrary but fixed $\kappa > 0$) then Theorem 2 strengthens the result of Theorem 1, showing a faster decay of the probability. To prove Theorem 1 and Theorem 2, we follow the general strategy of [10,22]. Since we do not allow logarithmic corrections, however, we need to modify several steps of the proof and, in particular, of the expansion algorithm. Since our small probabilities are independent of N, we cannot estimate the probability of unions of many bad events with the trivial union bound. Compared with [10,22], we have to work with high moments, instead of taking probabilities; similar ideas were used in [17] to prove the convergence towards the semicircle law on the optimal scale (but, in that case, not with the optimal rate).

As an application of the optimal convergence of the Stieltjes transform, we obtain in Theorem 3 bounds on the rate of convergence of the eigenvalue counting function $n(E) = N^{-1}|\{\alpha = 1, \ldots, N : \lambda_\alpha \leq E\}|$ and of the density of states. More precisely we show that

$$|n(E) - n_{sc}(E)| \lesssim \frac{\log N}{N} \tag{1.8}$$

for all $E \in \mathbb{R}$. Notice that, to prove (1.8), convergence of the imaginary part of $m(E + i\eta)$ is not enough, we really need convergence of the full Stieltjes transform, including its real part. Fortunately, Theorem 1 does imply convergence of $\mathrm{Re}\, m(E + i\eta)$, for $|E| \leq 2$. This allows us to prove that, with high probability, there are at most $K \log N$ eigenvalues outside the interval $[-2; 2]$; hence, we can establish convergence of the eigenvalue counting function $n(E)$ for all $E \in \mathbb{R}$.

Another application of the convergence of the Stieltjes transform is the rigidity of the eigenvalues of Wigner matrices. In Theorem 4 we show that, with high probability, the distance between an eigenvalue λ_α of the Wigner matrix H and the location γ_α of the same eigenvalue as predicted by the semicircle law is smaller than $(\log N)/N$. The results that we obtain in Theorem 3 and in Theorem 4 on the eigenvalue counting function, the density of states and the rigidity of the eigenvalues improve similar results from [10,22] by removing all logarithmic factors but one. While it is known that these bounds cannot hold true without logarithmic corrections, one expects eigenvalues to fluctuate on the scale $\sqrt{\log N}/N$, rather than on the scale $(\log N)/N$ obtained in Theorems 3 and 4 (for GUE and for hermitian ensembles whose entries have four moments matching GUE, this follows from the results of [24] and, respectively, [32], where the eigenvalues are shown to be, asymptotically, gaussian). We remark, moreover, that bounds for the rate of convergence of the eigenvalue counting function, as well as rigidity estimates for the eigenvalues of Wigner matrices, have been recently obtained in [23]. Roughly speaking, the authors of [23] show with methods which are different from the ones of [10,22] and of the present paper, that, with high probability, $|n(E) - n_{sc}(E)| \lesssim (\log N)^4/N$.

In the last years, there has been a lot of progress in the mathematical understanding of the statistical properties of the eigenvalues of Wigner matrices. As indicated above, the validity of the semicircle law for the density of states on microscopic scales was established in [15–17]. In the same works, the local semicircle law was applied to show the complete delocalization of the eigenvectors of Wigner matrices and to prove the repulsion among the eigenvalues. In [12], the validity of the semicircle law on microscopic intervals was used (in combination with [26]) to prove the (bulk) universality of the local eigenvalue statistics of hermitian Wigner matrices; independently of the choice of the law of the entries, the eigenvalue correlation functions always converge towards the same Wigner–Dyson distribution observed for the Gaussian Unitary Ensemble. At the same time, a different proof of universality for hermitian Wigner matrices was obtained in [32] (the two results have been combined in [14]). Soon after, a new argument, based on the introduction of a relaxation flow approximating Dyson Brownian motion, was introduced in [18] to prove bulk universality of Wigner matrices with arbitrary symmetry. In all these proofs of universality, the local convergence of the density of states was a crucial ingredient. In [21], local convergence towards the semicircle law and universality were extended to generalized Wigner matrices. Universality at the edge of the spectrum was proven in [31] and more recently in [22,33];

a necessary and sufficient condition on the decay of the matrix entries to obtain edge universality has been proven in [27]. For sample covariance matrices, local convergence towards the Marchenko–Pastur law [29] and universality of the local eigenvalue statistics were determined in the bulk [19,34] and at the soft edge [30,36]. Local convergence towards the Marchenko–Pastur law at the hard edge of the spectrum was proven in [3] and, implicitly, in [1,35]. The local convergence towards the circular law for the density of states of non-symmetric random matrices with independent entries was established in [1,2,35,38]. We notice that the local convergence of the density of states and delocalization of eigenvectors have also been obtained, in the last years, for more structured ensembles, such as the adjacency matrices of Erdős–Rényi graphs [9,11] and band matrices [5–8]. A nice overview on these and other results can be found in [4].

Next, we present our results in more details. Let $H = (h_{\ell j})$ be an $N \times N$ hermitian Wigner matrix, with entries $h_{jj} = x_{jj}/\sqrt{N}$ on the diagonal and

$$h_{\ell j} = \frac{1}{\sqrt{N}}(x_{\ell j} + i y_{\ell j})$$

for all $\ell < j$ where $x_{jj}, x_{\ell j}, y_{\ell j}$ are independent random variables. The off-diagonal variables $x_{\ell j}, y_{\ell j}$ are identically distributed with $\mathbb{E}\, x_{\ell j} = 0$ and $\mathbb{E}\, x_{\ell j}^2 = 1/2$. The diagonal entries x_{ii} are also identically distributed with $\mathbb{E}\, x_{ii} = 0$ and $\mathbb{E} x_{ii}^2 = 1$ (in fact the variance of the diagonal entries is not important, it should only be finite). We assume that the common distribution ν of $x_{\ell j}, x_{jj}$ and $y_{\ell j}$ has subgaussian decay, i.e., there exists $\delta_0 > 0$ such that

$$\int_{\mathbb{R}} e^{\delta_0 x^2} d\nu(x) < \infty. \tag{1.9}$$

This implies that $\mathbb{E}|x_{ij}|^{2q} \leq (Cq)^q$ for any $q \geq 1$. We define the Stieltjes transform $m(z)$ of H by (1.1) and the Stieltjes transform $m_{sc}(z)$ of the semicircle law by (1.3).

Theorem 1 *Assume* (1.9) *and fix* $\tilde{\eta} > 0$.

(i) *There exist constants* $M_0, N_0, C, c, c_0 > 0$ *such that*

$$\mathbb{P}\left(|m(E + i\eta) - m_{sc}(E + i\eta)| \geq \frac{K}{N\eta} \right) \leq \frac{(Cq)^{cq^2}}{K^q} \tag{1.10}$$

for all $\eta \leq \tilde{\eta}$, $|E| \leq 2 + \eta$, $K > 0$, $N > N_0$ *such that* $N\eta \geq M_0$, $q \in \mathbb{N}$ *with* $q \leq c_0 (N\eta)^{1/8}$.

(ii) *For any* $\tilde{E} > 0$ *there exist constants* $M_0, N_0, C, c, c_0 > 0$ *such that*

$$\mathbb{P}\left(|\operatorname{Im}\, m(E + i\eta) - \operatorname{Im}\, m_{sc}(E + i\eta)| \geq \frac{K}{N\eta} \right) \leq \frac{(Cq)^{cq^2}}{K^q} \tag{1.11}$$

for all $\eta \leq \tilde{\eta}$, $|E| \leq \tilde{E}$, $K > 0$, $N > N_0$ *such that* $N\eta \geq M_0$, $q \in \mathbb{N}$ *with* $q \leq c_0 (N\eta)^{1/8}$.

In the bulk of the spectrum, away from the edges, we can improve the bound (1.11), showing that it holds for all $q \in \mathbb{N}$ [in contrast to (1.11), which we only prove under the additional condition $q \leq c(N\eta)^{1/8}$].

Theorem 2 *Assume* (1.9), *fix* $\tilde{\eta} > 0$ *and* $\kappa > 0$. *Then there exist constants* $M_0, N_0, C, c > 0$ *such that*

$$\mathbb{P}\left(|m(E+i\eta) - m_{sc}(E+i\eta)| \geq \frac{K}{N\eta}\right) \leq \frac{(Cq)^{cq^2}}{K^q}$$

for all $E \in [-2+\kappa; 2-\kappa]$, $K > 0$, $N > N_0$, $\eta \leq \tilde{\eta}$ *such that* $N\eta \geq M_0$ *and* $q \in \mathbb{N}$.

As an application of Theorem 1, we prove the convergence of the counting function of the eigenvalues. We define

$$n(E) = \frac{1}{N}|\{\alpha : \lambda_\alpha \leq E\}|$$

and compare it with the cumulative distribution of the semicircle law, defined by

$$n_{sc}(E) = \int_{-\infty}^{E} \rho_{sc}(x)dx.$$

Theorem 3 *Assume* (1.9). *Then there exists constants* $N_0, C, c > 0$ *such that*

$$\mathbb{P}\left(|n(E) - n_{sc}(E)| \geq \frac{K \log N}{N}\right) \leq \frac{(Cq)^{cq^2}}{K^q} \tag{1.12}$$

for all $E \in \mathbb{R}$, $K > 0$, $N > N_0$, $q \in \mathbb{N}$.

Remark Theorem 3 immediately implies the convergence of the density of states. Let $\mathcal{N}[a; b]$ denote the number of eigenvalues in the interval $[a; b]$. From (1.12), we find

$$\mathbb{P}\left(\left|\frac{\mathcal{N}[E-\frac{\xi}{2N}; E+\frac{\xi}{2N}]}{\xi} - \rho_{sc}(E)\right| \geq \frac{K \log N}{\xi}\right) \leq \frac{(Cq)^{cq^2}}{K^q} \tag{1.13}$$

for all $\xi > 0$.

Another application of Theorem 1 is the rigidity of the eigenvalues of H, as stated in the following theorem. Recall that the classical location of the α-th eigenvalue, here denoted by γ_α, is defined by

$$\int_{-\infty}^{\gamma_\alpha} \rho_{sc}(x)dx = \frac{\alpha}{N} \quad 1 \leq \alpha \leq N$$

Theorem 4 *Assume* (1.9)*. For* $\alpha = 1, \ldots, N$*, let* $\hat{\alpha} = \min\{\alpha, N+1-\alpha\}$*. Then there exist constants* $C, c, N_0, \varepsilon > 0$ *such that*

$$\mathbb{P}\left(|\lambda_\alpha - \gamma_\alpha| \geq \frac{K \log N}{N} \left(\frac{N}{\hat{\alpha}}\right)^{\frac{1}{3}} \right) \leq \frac{(Cq)^{cq^2}}{K^q} \tag{1.14}$$

for all $N > N_0$, $K > 0$, $q \in \mathbb{N}$ *with* $q \leq N^\varepsilon$.

Remark In contrast with (1.10) and (1.11), the bounds (1.12), (1.13) and (1.14) are not optimal. By the results of [24] for GUE, the eigenvalues are expected to fluctuate on the scale $\sqrt{\log N}/N$.

The plan of the paper is as follows. In Sect. 2 we collect some preliminary results which will be needed in our analysis. In particular, we prove some bounds on the Stieltjes transform of the semicircle law, and we discuss certain algebraic identities concerning the entries of the resolvent $G(z) = (H - z)^{-1}$. In Sect. 3, we derive a first non-optimal estimate on the fluctuations of the Stieltjes transform m around m_{sc}, see, in particular, Lemma 3.5. In Sect. 4, we prove Theorem 1, assuming a bound for the high moments of certain error terms, as stated in Lemma 4.1. Section 5 is then devoted to the proof of Lemma 4.1; it is here that, to obtain optimal bounds on $m - m_{sc}$, we make use of an expansion algorithm similar to the one introduced in [10]. In Sect. 6, we prove Theorem 2. Finally, in Sect. 7 and Sect. 8, we prove Theorem 3 and, respectively, Theorem 4.

2 Preliminary results

2.1 Properties of the Stieltjes transform of the semicircle law

In the next proposition, we give simple proofs of several useful bounds on the Stieltjes transform m_{sc} of the semicircle law, as defined in (1.3).

Proposition 2.1 *Let* $z = E + i\eta$*. For arbitrary* $E \in \mathbb{R}$ *and* $\eta > 0$*, we have* $(1 + |z|)^{-1} < |m_{sc}(z)| < 1$ *and the bounds*

$$\frac{\text{Im } m_{sc}(z)}{|m_{sc}^2(z) - 1|} \leq 1 \quad \text{and} \quad \frac{\eta}{|m_{sc}^2(z) - 1|} \leq \sqrt{\eta}\,(1 + \sqrt{\eta}). \tag{2.1}$$

Proof We rewrite the equation for $m_{sc}(z)$ as

$$z + m_{sc}(z) = -\frac{1}{m_{sc}(z)}. \tag{2.2}$$

Writing explicitly the real and imaginary part we get

$$(\text{Re } m_{sc}(z)) \left(1 + \frac{1}{|m_{sc}(z)|^2}\right) + i\,(\text{Im } m_{sc}(z)) \left(1 - \frac{1}{|m_{sc}(z)|^2}\right) = -E - i\eta.$$

Therefore, we get

$$(1-|m_{sc}(z)|^2)\,\text{Im}\ m_{sc}(z) = \eta|m_{sc}(z)|^2. \tag{2.3}$$

Since Im $m_{sc}(z) > 0$ for $\eta > 0$, it follows that $|m_{sc}(z)| < 1$. The bound $|m_{sc}(z)| > (1+|z|)^{-1}$ follows then immediately from (2.2).

To prove the first bound in (2.1), we rewrite

$$|m_{sc}^2(z) - 1| = |m_{sc}(z) - 1||m_{sc}(z) + 1|.$$

We distinguish two cases. If $|\,\text{Re}\ m_{sc}(z) - 1| \geq 1$, we use $|m_{sc}(z) - 1| \geq 1$ and $|m_{sc}(z)+1| \geq \text{Im}\ m_{sc}(z)$. If $|\,\text{Re}\ m_{sc}(z)-1| \leq 1$ we use $|m_{sc}(z)+1| \geq |\,\text{Re}\ m_{sc}(z)+1| \geq 2 - |\,\text{Re}\ m_{sc}(z) - 1| \geq 1$ and $|m_{sc}(z) - 1| \geq \text{Im}\ m_{sc}(z)$. In both cases, we find $|m_{sc}^2(z) - 1| \geq \text{Im}\ m_{sc}(z)$.

Also to prove the second bound in (2.1), we distinguish two cases. If Im $m_{sc}(z) \geq \sqrt{\eta}$, we use

$$|m_{sc}^2(z) - 1| \geq \text{Im}\ m_{sc}(z) \geq \sqrt{\eta} \geq \frac{\sqrt{\eta}}{1+\sqrt{\eta}}.$$

If, on the other hand, Im $m_{sc}(z) \leq \sqrt{\eta}$, Eq. (2.3) gives

$$\eta\,|m_{sc}(z)|^2 = (1-|m_{sc}(z)|^2)\,\text{Im}\, m_{sc}(z) \leq \sqrt{\eta}\,(1-|m_{sc}(z)|^2)$$

and therefore

$$|m_{sc}(z)|^2 \leq \frac{1}{1+\sqrt{\eta}}.$$

The desired bound follows then from $|m_{sc}^2(z) - 1| \geq 1 - |m_{sc}(z)|^2$. □

2.2 Identities for resolvent entries and for the Stieltjes transform

We use the following notation

$$G(z) = (H-z)^{-1}$$

for the resolvent of the Wigner matrix H, so that

$$m(z) = \frac{1}{N}\,\text{Tr}\,G(z).$$

To shorten the notation, we will often omit from the notation the dependence on z from the resolvent $G(z)$ and from the Stieltjes transform $m(z)$ (and also from the Stieltjes transform of the semicircle law $m_{sc}(z)$), writing $G \equiv G(z)$, $m \equiv m(z)$ (and $m_{sc} \equiv m_{sc}(z)$).

We denote by $H^{(j)}$ the j-th minor of the matrix H and by $G^{(j)}$ the matrix $G^{(j)} = (H^{(j)} - z)^{-1}$. For any $j = 1, \ldots, N$, we can express the (j, j)-entry of the resolvent G as

$$G_{jj} = \frac{1}{h_{jj} - z - \mathbf{a}_j^* G^{(j)} \mathbf{a}_j} \tag{2.4}$$

where $\mathbf{a}_j \in \mathbb{C}^{N-1}$ is a vector, whose components are the off-diagonal entries of the j-th column of H. We will label the components of $\mathbf{a}_j$ with indices in $\{1, \ldots, N\} \backslash \{j\}$.

To compare the diagonal entries of G with the diagonal entries of $G^{(j)}$, we will make use of the identity

$$G_{kk} = G_{kk}^{(j)} + \frac{G_{jk} G_{kj}}{G_{jj}} \tag{2.5}$$

valid for all $k \neq j$. A proof of (2.4) and of (2.5) can be found, for example in [4].

Let

$$\Lambda(z) = m(z) - m_{sc}(z).$$

Also in this case, we will often omit the z dependence from the notation, writing just $\Lambda \equiv \Lambda(z)$. Next, we are going to obtain a self-consistent equation for Λ; we follow here Section 5.1 in [10]. From (2.4), we have

$$G_{jj} = -\frac{1}{-h_{jj} + z + (\mathbb{I} - \mathbb{E}_j)\, \mathbf{a}_j^* G^{(j)} \mathbf{a}_j + \mathbb{E}_j\, \mathbf{a}_j^* G^{(j)} \mathbf{a}_j}$$

where $\mathbb{E}_j$ denotes the expectation with respect to $\mathbf{a}_j$. Since the components of $\mathbf{a}_j$ are independent of $G^{(j)}$, we find

$$\mathbb{E}_j\, \mathbf{a}_j^* G^{(j)} \mathbf{a}_j = \frac{1}{N} \sum_{k \neq j} G_{kk}^{(j)} = m + \frac{1}{N} \sum_{k \neq j} (G_{kk}^{(j)} - G_{kk}) - \frac{1}{N} G_{jj}.$$

Hence

$$G_{jj} = -\frac{1}{z + m_{sc} + \Lambda + \Upsilon_j} \tag{2.6}$$

with

$$\begin{aligned} \Upsilon_j &= -h_{jj} - \frac{1}{N} G_{jj} - \frac{1}{N} \sum_{k \neq j} (G_{kk} - G_{kk}^{(j)}) + (\mathbb{I} - \mathbb{E}_j) \mathbf{a}_j^* G^{(j)} \mathbf{a}_j \\ &= -h_{jj} - \frac{1}{G_{jj}} \frac{1}{N} \sum_k G_{jk} G_{kj} + (\mathbb{I} - \mathbb{E}_j) \mathbf{a}_j^* G^{(j)} \mathbf{a}_j. \end{aligned} \tag{2.7}$$

Setting

$$g_j = G_{jj} - m_{sc},$$

we find from (2.6) that

$$g_j = m_{sc}(\Lambda + \Upsilon_j)G_{jj}. \tag{2.8}$$

Averaging over j, we find

$$\begin{aligned}\Lambda = \frac{1}{N}\sum_{j=1}^{N} g_j &= \frac{m_{sc}\Lambda}{N}\sum_{j=1}^{N} G_{jj} + \frac{m_{sc}}{N}\sum_{j=1}^{N}\Upsilon_j G_{jj} \\ &= m_{sc}^2\Lambda + m_{sc}\Lambda^2 + \frac{m_{sc}}{N}\sum_{j=1}^{N}\Upsilon_j G_{jj}\end{aligned}$$

and hence, rearranging the terms and dividing by m_{sc},

$$\Lambda^2 + \frac{m_{sc}^2 - 1}{m_{sc}}\Lambda + R = 0 \tag{2.9}$$

where we set

$$R = \frac{1}{N}\sum_{j=1}^{N}\Upsilon_j G_{jj}. \tag{2.10}$$

Using Eq. (1.4), we can rewrite (2.9) as

$$\Lambda^2 + (2m_{sc} + z)\Lambda + R = 0 \tag{2.11}$$

This quadratic equation for Λ has two solutions. However, only one of them satisfies the condition $\text{Im}\,\Lambda > -\text{Im}\,m_{sc}$. We find

$$\Lambda = -m_{sc} - \frac{z}{2} + \sqrt{\left(m_{sc} + \frac{z}{2}\right)^2 - R} \tag{2.12}$$

where, for any $w \in \mathbb{C}$, we denote by $\sqrt{w}$ the square root of w with $\text{Im}\,w \geq 0$ (for w real and positive, we choose $\sqrt{w} \geq 0$). Notice that, since

$$m_{sc}(z) = -\frac{z}{2} + \sqrt{\frac{z^2}{4} - 1}.$$

for any z with $\text{Im}\, z > 0$, we can write

$$\Lambda = -\sqrt{\frac{z^2}{4} - 1} + \sqrt{\frac{z^2}{4} - 1 - R} \tag{2.13}$$

In the following proposition we use the algebraic identity (2.12) to bound $|\Lambda|$ and $|\text{Im}\,\Lambda|$ in terms of $|R|$. These bounds are crucial in the proof of Theorem 1.

Proposition 2.2 *Let* $z = E + i\eta$. *As in* (2.10) *and* (2.12), (2.13) *we use the shorthand notations* $\Lambda = \Lambda(z)$ *and* $R = R(z)$. *There exists a constant* $C > 0$ *such that*

$$|\Lambda| \leq C \min\left\{ \frac{|R|}{|m_{sc}^2 - 1|}, \sqrt{|R|} \right\} \tag{2.14}$$

for all $\eta > 0$ *and* $|E| \leq 2 + \eta$ *and such that*

$$|\,\text{Im}\ \Lambda| \leq C \min\left\{ \frac{|R|}{|m_{sc}^2 - 1|}, \sqrt{|R|} \right\} \tag{2.15}$$

$$\min\left(|\Lambda|, |\Lambda + (2m_{sc} + z)|\right) \leq C\sqrt{|R|} \tag{2.16}$$

for all $\eta > 0$ *and* $E \in \mathbb{R}$.

Remark Note that, for $|E| > 2 + \eta$, Proposition 2.2 only controls the imaginary part of Λ in terms of R. This is the origin of the distinction between part (i) and part (ii) of Theorem 1. The reason why we cannot control the real part of Λ for $|E| > 2+\eta$ is clear from (2.13). Recall that $\sqrt{w}$ denotes the square root of w with positive imaginary part, which is discontinuous across the positive real axis. For $|E| > 2+\eta$, $\text{Re}\,(z^2/4-1) > 0$ and $|\text{Im}\,(z^2/4 - 1)| = |E|\eta/2$ can be very small (for small η). For this reason even for very small $|R|$ it is possible that $z^2/4 - 1$ and $z^2/4 - 1 - R$ are on opposite sides of the positive real axis. In this case, $\Lambda \simeq -2\sqrt{z^2/4 - 1}$ would not be small. Notice, however, that the discontinuity of the square root concerns only its real part; the imaginary part is continuous on the whole complex plane. This is why we still get control of the imaginary part of Λ in (2.15) also for $|E| > 2 + \eta$.

Proof As before we denote by $\sqrt{w}$ the square root of $w \in \mathbb{C}$ with $\text{Im}\, w \geq 0$ (and with $\sqrt{w} > 0$ for $w > 0$ real). We claim that

- for any fixed $c > 0$, there exists a constant $C > 0$ such that

$$\left|\sqrt{a+b} - \sqrt{a}\right| \leq C \frac{|b|}{\sqrt{|a| + |b|}} \tag{2.17}$$

  for all $a, b \in \mathbb{C}$ with $|\text{Im}\, a| \geq c\,\text{Re}\, a$.
- there exists a constant $C > 0$ such that

$$\left|\text{Im}\ \sqrt{a+b} - \text{Im}\ \sqrt{a}\right| \leq C \frac{|b|}{\sqrt{|a| + |b|}} \tag{2.18}$$

  for all $a, b \in \mathbb{C}$.

The bounds (2.14) and (2.15) follow directly from (2.12), (2.13), applying (2.18) and, respectively, (2.17) with $a = (m_{sc} + z/2)^2 = (m_{sc}^2 - 1)^2/m_{sc}^2$ and $b = -R$ (since $|m_{sc}| \leq 1$; see Proposition 2.1). Here, to prove (2.14), we use the fact that $a = z^2/4 - 1$, and therefore, with $z = E + i\eta$, that

$$\text{Re}\, a = \frac{E^2 - \eta^2}{4} - 1 \quad \text{and} \quad \text{Im}\, a = \frac{E\eta}{2}$$

This implies that $|\text{Im}\, a| \geq c\, \text{Re}\, a$ for all $|E| \leq 2 + \eta$ (for $|E| \leq 2$, $\text{Re}\, a < 0$ and the bound $|\text{Im}\, a| \geq c\, \text{Re}\, a$ is trivial).

To prove (2.16), we observe from (2.11) that $\Lambda(\Lambda + (2m_{sc} + z)) = -R$. This implies that

$$|\Lambda + (2m_{sc} + z)| = \frac{|R|}{|\Lambda|}$$

and hence that either $|\Lambda| \leq \sqrt{|R|}$ or $|\Lambda + (2m_{sc} + z)| \leq \sqrt{|R|}$.

We conclude with the proof of the bounds (2.17) and (2.18). Both estimates clearly hold when $|b| > |a|/2$, since then

$$\left|\sqrt{a+b} - \sqrt{a}\right| \leq C|b|^{1/2} \leq C\frac{|b|}{\sqrt{|a| + |b|}}$$

Hence, we can assume that $|b| \leq |a|/2$. If $|\text{Im}\, a| \geq c\, \text{Re}\, a$, we have $\text{Im}\, \sqrt{a} \geq \widetilde{c}\, |a|^{1/2}$ for an appropriate constant $\widetilde{c} > 0$. This implies that

$$\left|\sqrt{a+b} - \sqrt{a}\right| = \left|\frac{b}{\sqrt{a+b} + \sqrt{a}}\right| \leq \frac{|b|}{\text{Im}\, \sqrt{a} + \text{Im}\, \sqrt{a+b}} \leq C\frac{|b|}{|a|^{1/2}} \leq C\frac{|b|}{\sqrt{|a| + |b|}}$$

from the assumption $|b| \leq |a|/2$ and because $\text{Im}\, \sqrt{a+b} \geq 0$.

It remains to prove (2.18) for $|\text{Im}\, a| < c\, \text{Re}\, a$ and $|b| \leq |a|/2$. To this end, we consider first the case $a > 0$ real. If $\text{Im}\, b \geq 0$, we find

$$\begin{aligned}\left|\text{Im}\, \sqrt{a+b} - \text{Im}\, \sqrt{a}\right| &\leq \left|\sqrt{a+b} - \sqrt{a}\right| \leq \frac{|b|}{|\sqrt{a+b} + \sqrt{a}|} \\ &\leq \frac{|b|}{\text{Re}\, \sqrt{a+b} + \sqrt{a}} \leq \frac{|b|}{\sqrt{a}} \leq C\frac{|b|}{\sqrt{|a| + |b|}}\end{aligned}$$

since $\text{Re}\, \sqrt{a+b} \geq 0$ and $|b| < c\, |a|$. If instead $\text{Im}\, b < 0$, we have (again, for $a > 0$ real)

$$\begin{aligned}\left|\text{Im}\, \sqrt{a+b} - \text{Im}\, \sqrt{a}\right| &= \left|\text{Im}\, \sqrt{a+b} + \text{Im}\, \sqrt{a}\right| \leq \left|\sqrt{a+b} + \sqrt{a}\right| \\ &\leq \frac{|b|}{|\sqrt{a} - \sqrt{a+b}|} \leq \frac{|b|}{\sqrt{|a|}} \leq \frac{|b|}{\sqrt{|a| + |b|}}\end{aligned}$$

because, in this case, $\text{Re}\, \sqrt{a+b} < 0$.

Now, we consider general $a, b \in \mathbb{C}$ with $|\text{Im}\, a| < c\,\text{Re}\, a$ and $|b| \leq c\,|a|$. Again, we distinguish two cases. If $\text{Im}\,(a+b) > 0$, we have $\text{Re}\,\sqrt{a+b} > 0$ and therefore

$$\left|\text{Im}\,\sqrt{a+b} - \text{Im}\,\sqrt{a}\right| \leq \frac{|b|}{\text{Re}\,\sqrt{a}} \leq \frac{|b|}{\sqrt{|a|}} \leq \frac{|b|}{\sqrt{|a|+|b|}}$$

since the assumption $|\text{Im}\, a| < c\,\text{Re}\, a$ implies that $\text{Re}\,\sqrt{a} \geq c\sqrt{|a|}$. Finally, let us assume that $\text{Im}\,(a+b) < 0$. Then we find $\lambda \in (0,1)$ such that $a + \lambda b = d > 0$ is real. Hence, we can estimate

$$\begin{aligned}\left|\text{Im}\,\sqrt{a+b} - \text{Im}\,\sqrt{a}\right| &\leq \left|\text{Im}\,\sqrt{a+b} - \text{Im}\,\sqrt{d}\right| + \left|\text{Im}\,\sqrt{d} - \text{Im}\,\sqrt{a}\right| \\ &= \left|\text{Im}\,\sqrt{d+(1-\lambda)b} - \text{Im}\,\sqrt{d}\right| + \left|\text{Im}\,\sqrt{d-\lambda b} - \text{Im}\,\sqrt{d}\right| \\ &\leq C\frac{|b|}{\sqrt{|d|+|b|}} \leq C\frac{|b|}{\sqrt{|a|+|b|}}\end{aligned}$$

where we applied the bounds obtained above for $a > 0$ real. □

The inequalities in Proposition 2.2 are the starting point for all our estimates on the random variable Λ. In the next section, we will prove a non-optimal bound on Λ, based on the second bound in (2.14), proportional to $\sqrt{|R|}$ [for $z = E + i\eta$ with $|E| > 2+\eta$, we cannot apply (2.14); instead, we will control the imaginary part of Λ using (2.15) and we will get some bound on its real part using (2.16)]. Afterwards, in Sect. 4, we will prove Theorem 1, which gives an optimal estimate on $|\Lambda|$. To achieve this goal, we will apply the first bound in (2.14), proportional to $|R|$ in the bulk of the spectrum (where the denominator $|m_{sc}^2 - 1|$ is of order one) and we will use the second bound in (2.14), proportional to $\sqrt{|R|}$, close to the edges of the spectrum (where $|m_{sc}^2 - 1|$ is small). Also in this case, for $|E| > 2 + \eta$ we cannot apply (2.14); instead, in this region we will apply (2.15) (getting only a bound for the imaginary part of Λ).

3 Non-optimal bound on $\Lambda = m - m_{sc}$

3.1 Bound on moments of Υ_j

According to (2.15) and (2.14), in order to show smallness of Λ we need bounds on the quantity $R = N^{-1}\sum_{j=1}^{N} \Upsilon_j G_{jj}$, introduced in (2.10). We will prove in Lemma 3.4 that the diagonal entries G_{jj} of the resolvent are bounded with high probability, hence it is possible to bound R by controlling the coefficients Υ_j. This is the goal of the next lemma.

Lemma 3.1 *Assume* (1.9). *Let* Υ_1 *be defined as in* (2.7). *Then there exists a universal constant* C *such that*

$$\mathbb{E}|\Upsilon_1|^{2q} \leq (Cq)^{2q}\left(\frac{(\text{Im}\, m_{sc})^q + \mathbb{E}|\,\text{Im}\,\Lambda|^q}{(N\eta)^q} + \frac{1}{(N\eta)^{2q}} + \frac{1}{N^q}\right) \qquad (3.1)$$

for all $E \in \mathbb{R}$, $N \geq 1$, $\eta \geq 1/N$ *and* $q \in \mathbb{N}$.

Remarks Of course, the same bound is valid for Υ_j, for any $j = 2, \ldots, N$, since all these variables have the same law. From (3.1), we conclude that the typical size of $|\Upsilon_j|$ is of the order $(N\eta)^{-1/2}$ (assuming that Im $\Lambda|$ is at least bounded by a constant). Combined with an upper bound on the diagonal entries of the resolvent, this would allow us to show that $R = N^{-1} \sum_j \Upsilon_j G_{jj}$ is at most of the order $(N\eta)^{-1/2}$. From (2.14), this would imply (at least in the bulk of the spectrum) that $|\Lambda| \lesssim (N\eta)^{-1/2}$, with high probability. This result, however, is still far from the optimal bound $|\Lambda| \lesssim (N\eta)^{-1}$ that we are aiming at. To obtain the optimal bound, it is crucial to use the cancellations among the different terms in the average defining R. We will do so in Sects. 4 and 5.

Proof We observe that

$$\left| \frac{1}{G_{11}} \frac{1}{N} \sum_{k=1}^{N} G_{1k} G_{k1} \right| = \frac{|(G^2)_{11}|}{N|G_{11}|}.$$

Since

$$|(G^2(z))_{11}| = |\langle e_1, G^2(z) e_1 \rangle| \leq \|G^*(z) e_1\| \|G(z) e_1\| \leq |G(z)|^2_{11}$$

and

$$|G(z)|^2 = G(z) G(\bar{z}) = \frac{\text{Im } G(z)}{\eta} \tag{3.2}$$

we obtain

$$|(G^2(z))_{11}| \leq \frac{\text{Im } G_{11}(z)}{\eta}$$

and therefore

$$\left| \frac{1}{G_{11}} \frac{1}{N} \sum_{k=1}^{N} G_{1k} G_{k1} \right| \leq \frac{1}{N\eta}$$

and hence, using also the assumption (1.9), we find

$$\mathbb{E}\, |\Upsilon_j|^{2q} \leq \frac{(Cq)^q}{N^q} + \frac{1}{(N\eta)^{2q}} + C^q\, \mathbb{E} |Z_j|^{2q}$$

where we defined

$$Z_j = (\mathbb{I} - \mathbb{E}_j)\, \mathbf{a}_j^* G^{(j)} \mathbf{a}_j. \tag{3.3}$$

The claim now follows from Proposition 3.2 below, where we bound the moments of Z_j. □

In the next proposition, we bound the moments of the variable Z_1, defined in (3.3), in terms of the moments of $\text{Im }\Lambda$.

Proposition 3.2 *Assume* (1.9) *and let* Z_1 *be defined as in* (3.3). *Then there exists a constant* $C > 0$ *such that*

$$\mathbb{E}\,|Z_1|^{2q} \le (Cq)^{2q} \left(\frac{(\text{Im }m_{sc})^q + \mathbb{E}|\,\text{Im }\Lambda|^q}{(N\eta)^q} + \frac{1}{(N\eta)^{2q}} \right), \tag{3.4}$$

for all $E \in \mathbb{R}$, $N \ge 1$, $\eta \ge 1/N$ *and* $q \in \mathbb{N}$.

Proof We keep the randomness in the entries of the minor $H^{(1)}$ fixed. By the Hanson–Wright large deviation estimate, the fluctuations of the quadratic form around its mean can be controlled by the Hilbert–Schmidt norm of $G^{(1)}$. In fact, it follows from Proposition 9.1 that

$$\mathbb{P}_1 \left(\left| (\mathbb{I} - \mathbb{E}_1)\, \mathbf{a}_1^* G^{(1)} \mathbf{a}_1 \right| \ge t \right) \le C \exp \left(- \frac{ctN}{\left(\text{Tr }|G^{(1)}|^2\right)^{1/2}} \right)$$

and therefore that

$$\mathbb{E}\,|Z_1|^{2q} = \mathbb{E} \int_0^{+\infty} \mathbb{P}_1 \left(|(\mathbb{I} - \mathbb{E}_1)\mathbf{a}_1^* G^{(1)} \mathbf{a}_1|^{2q} \ge t \right) dt \le (Cq)^{2q}\, \mathbb{E} \left(\frac{\text{Tr }|G^{(1)}|^2}{N^2} \right)^q.$$

To conclude the proof of (3.4), we notice that, similarly to (3.2),

$$|G^{(1)}(z)|^2 = \frac{\text{Im }G^{(1)}(z)}{\eta} \tag{3.5}$$

and hence that

$$\begin{aligned} \frac{\text{Tr }|G^{(1)}|^2}{N^2} &= \frac{1}{N} \frac{1}{N\eta} \text{Im} \sum_{k \ne 1} G^{(1)}_{kk} \\ &= -\frac{1}{N} \frac{1}{N\eta} \text{Im } \frac{1}{G_{11}} \sum_k G_{k1} G_{1k} + \frac{\text{Im }m}{N\eta} \\ &\le \frac{1}{(N\eta)^2} + \frac{\text{Im }m_{sc} + |\,\text{Im }\Lambda|}{N\eta}. \end{aligned} \tag{3.6}$$

□

3.2 Bound on moments of Λ in terms of moments of G_{11}

In the next lemma, we estimate the moments of $|\text{Im }\Lambda|$ in terms of the moments of the diagonal entries of the resolvent. In the region $|E| \le 2 + \eta$, the lemma also gives a bound on the moments of $|\Lambda|$ (including also the real part of Λ).

Lemma 3.3 *Assume* (1.9), *fix* $\tilde{\eta} > 0$ *and set* $z = E + i\eta$.

(i) *There exists a constant* $C > 0$ *such that*

$$\mathbb{E}|\Lambda|^{2q} \leq \frac{(Cq)^{4q/3}}{(N\eta)^{q/2}} \left[(\mathbb{E}|G_{11}|^{2q})^{2/3} + 1 \right], \tag{3.7}$$

for all $|E| \leq 2 + \eta$, $\eta \leq \tilde{\eta}$, $N \geq 1$ *such that* $N\eta \geq 1$, *and for all* $q \in \mathbb{N}$.

(ii) *There exists a constant* $C > 0$ *such that*

$$\mathbb{E}|\text{ Im } \Lambda|^{2q} \leq \frac{(Cq)^{4q/3}}{(N\eta)^{q/2}} \left[(\mathbb{E}|G_{11}|^{2q})^{2/3} + 1 \right], \tag{3.8}$$

for all $E \in \mathbb{R}$, $\eta \leq \tilde{\eta}$, $N \geq 1$ *such that* $N\eta \geq 1$, *and for all* $q \in \mathbb{N}$.

Proof We show (3.8). To this end, we use the bound (2.15) proportional to $\sqrt{|R|}$. We find

$$\begin{aligned}
\mathbb{E}|\text{Im } \Lambda|^{2q} &\leq C^q \mathbb{E}|R|^q \leq C^q\, \mathbb{E} \left| \frac{1}{N} \sum_j \Upsilon_j G_{jj} \right|^q \leq C^q\, \mathbb{E}|\Upsilon_1 G_{11}|^q \\
&\leq C^q \sqrt{\mathbb{E}|\Upsilon_1|^{2q}\, \mathbb{E}|G_{11}|^{2q}} \\
&\leq (Cq)^q \sqrt{\mathbb{E}|G_{11}|^{2q}} \sqrt{\frac{(\text{Im } m_{sc})^q + \mathbb{E}|\text{ Im } \Lambda|^q}{(N\eta)^q} + \frac{1}{N^q} + \frac{1}{(N\eta)^{2q}}}.
\end{aligned}$$

By Cauchy–Schwarz we find

$$\mathbb{E}|\text{ Im } \Lambda|^{2q} \leq \frac{(Cq)^q \sqrt{\mathbb{E}|G_{11}|^{2q}}}{(N\eta)^{q/2}} \left(\sqrt[4]{\mathbb{E}|\text{ Im } \Lambda|^{2q}} + 1 \right).$$

Since $x^{1/4} \leq \delta x + \delta^{-1/3}$ for all $x, \delta > 0$, we find

$$\begin{aligned}
\mathbb{E}|\text{ Im } \Lambda|^{2q} &\leq (Cq)^{4q/3} \left[\left(\frac{\mathbb{E}|G_{11}|^{2q}}{(N\eta)^q} \right)^{2/3} + \left(\frac{\mathbb{E}|G_{11}|^{2q}}{(N\eta)^q} \right)^{1/2} \right] \\
&\leq \frac{(Cq)^{4q/3}}{(N\eta)^{q/2}} \left[(\mathbb{E}|G_{11}|^{2q})^{2/3} + 1 \right].
\end{aligned}$$

The bound (3.7) follows analogously, using (2.14) instead of (2.15). □

3.3 Bound on moments of G_{11}

To obtain a (non-optimal) bound on the moments of $|\text{Im}\Lambda|$ and (in the region $|E| \leq 2+\eta$) of $|\Lambda|$, we still need control on the moments of the diagonal entries of the resolvent. This is the content of the following lemma, which makes use of the bound (2.16) (in the region $|E| \leq 2$, we could also apply (2.14) to arrive at the same conclusions).

Lemma 3.4 *Assume* (1.9), *fix* $\tilde{E} > 0$ *and* $\tilde{\eta} > 0$. *We set* $z = E + i\eta$ *and use the shorthand notation* $G_{jj} = G_{jj}(z)$. *Then there exist constants* $C_0, M > 0$ *such that*

$$\mathbb{E}\,|G_{11}|^q \le C_0^q, \tag{3.9}$$

for all $|E| < \tilde{E}, \eta \le \tilde{\eta}, N \ge 1$ *such that* $N\eta \ge M$, *and for all* $q \in \mathbb{N}$ *with* $q \le (N\eta)^{1/4}$.

Proof We show first how to relate $G_{11}(E + i\eta/s)$ to $G_{11}(E + i\eta)$. We consider $G_{11}(E + i\nu)$ as a function of ν, with fixed E. We find

$$\frac{d}{d\nu} \log G_{11}(E + i\nu) = \frac{i}{G_{11}(E + i\nu)} \left\langle e_1, \frac{1}{(H - E - i\nu)^2} e_1 \right\rangle.$$

Taking absolute value, and arguing as in (3.5), we find

$$\left| \frac{d}{d\nu} \log G_{11}(E + i\nu) \right| \le \frac{1}{|G_{11}(E + i\nu)|} \|G(E + i\nu)e_1\| \, \|G(E - i\nu)e_1\| \le \frac{1}{\nu}.$$

Hence

$$\begin{aligned} |\log G_{11}(E + i\eta) - \log G_{11}(E + i\eta/s)| &= \left| \int_{\eta/s}^{\eta} d\nu \, \frac{d}{d\nu} \log G_{11}(E + i\nu) \right| \\ &\le \int_{\eta/s}^{\eta} d\nu \, \frac{1}{\nu} = \log s. \end{aligned}$$

This gives

$$|\mathrm{Re}\, \log G_{11}(E + i\eta) - \mathrm{Re}\, \log G_{11}(E + i\eta/s)| \le \log s$$

and therefore

$$\log |G_{11}(E + i\eta/s)| \le \log |G_{11}(E + i\eta)| + \log s.$$

Exponentiating, we find

$$|G_{11}(E + i\eta/s)| \le s\,|G_{11}(E + i\eta)|. \tag{3.10}$$

The proof of (3.9) now proceeds by induction in the distance η from the real axis. We fix $|E| \le \tilde{E}$ and we use the notation $G_{11}(E + i\eta) \equiv G_{11}(\eta)$, and similarly for the other quantities depending on z. To start the induction, we recall that for $\eta \ge 0.1$, $|G_{11}(\eta)| \le 10$, so $\mathbb{E}|G_{11}(\eta)|^q \le 10^q$. Next, we assume that, for some $\eta_0 > 0$,

$$\mathbb{E}|G_{11}(\eta_0)|^q \le C_0^q \quad \text{for all } 1 \le q \le (N\eta_0)^{1/4}. \tag{3.11}$$

We show that the same bounds hold true, if we replace η_0 by $\eta_1 = \eta_0/16$.

We set $s = 16$ in (3.10); from the assumption (3.11) we get that

$$\mathbb{E}|G_{11}(\eta_1)|^q \leq (50C_0)^q \tag{3.12}$$

for any $q \leq (N\eta_0)^{1/4}$. To iterate further this bound, we need to improve it and get back to the constant C_0^q. To this end, we recall the identity $G_{11} = m_{sc} + m_{sc}(\Lambda + \Upsilon_1)G_{11}$ [see (2.8)], valid for all $z \in \mathbb{C}$. On the one hand, this implies that

$$|G_{11}| \leq 1 + |\Lambda||G_{11}| + |\Upsilon_1||G_{11}|. \tag{3.13}$$

for all $z \in \mathbb{C}$. On the other hand, it also implies that

$$G_{11} = m_{sc} + m_{sc}(\Lambda + 2m_{sc} + z + \Upsilon_1)G_{11} - m_{sc}(2m_{sc} + z)G_{11}$$

Since $1 + m_{sc}(2m_{sc} + z) = m_{sc}^2$, we conclude that

$$|G_{11}| \leq \frac{1}{|m_{sc}|} + \frac{1}{|m_{sc}|}|\Lambda + 2m_{sc} + z||G_{11}| + \frac{1}{|m_{sc}|}|\Upsilon_1||G_{11}|.$$

Using the bound $|m_{sc}|^{-1} \leq 1 + |z|$, and combining with (3.13), we obtain that there exists a constant C depending on $\tilde{E}$ and $\tilde{\eta}$ such that

$$\begin{aligned} |G_{11}| &\leq C\left(1 + \min\left(|\Lambda|, |\Lambda + 2m_{sc} + z|\right)|G_{11}| + |\Upsilon_1||G_{11}|\right) \\ &\leq C\left(1 + \sqrt{|R|}\,|G_{11}| + |\Upsilon_1||G_{11}|\right) \end{aligned}$$

for all $z = E + i\eta$ with $|E| \leq \tilde{E}$ and $0 < \eta < \tilde{\eta}$. In the second inequality, we used (2.16). Taking the q-th moment, using the definition (2.10) and Cauchy–Schwarz, we find

$$\mathbb{E}|G_{11}|^q \leq (3C)^q \left(1 + (\mathbb{E}|G_{11}|^{2q})^{3/4}(\mathbb{E}|\Upsilon_1|^{2q})^{1/4} + (\mathbb{E}|G_{11}|^{2q})^{1/2}(\mathbb{E}|\Upsilon_1|^{2q})^{1/2}\right) \tag{3.14}$$

for all $z = E + i\eta$ with $|E| \leq \tilde{E}$ and $0 < \eta < \tilde{\eta}$. Plugging (3.8) into (3.1), we find

$$\mathbb{E}|\Upsilon_1|^{2q} \leq \frac{(\widetilde{C}q)^{8q/3}}{(N\eta)^q}\left(1 + (\mathbb{E}|G_{11}|^{2q})^{1/3}\right)$$

for all $z = E + i\eta$ with $|E| \leq \tilde{E}$ and $0 < \eta < \tilde{\eta}$. Next, we insert the last bound into (3.14). We specialize to $z = E + i\eta_1$ and to $q \leq (N\eta_1)^{1/4}$ [which implies that $2q \leq (N\eta_0)^{1/4}$ and therefore that we can use (3.12) to bound $\mathbb{E}|G_{11}|^{2q}$]. We find

$$\begin{aligned}\mathbb{E}|G_{11}(\eta_1)|^q &\le (3C)^q \left(1+\frac{(\widetilde{C}q)^{2q/3}}{(N\eta_1)^{q/4}}(1+(\mathbb{E}|G_{11}(\eta_1)|^{2q})^{1/12})(\mathbb{E}|G_{11}(\eta_1)|^{2q})^{3/4}\right.\\ &\qquad\left.+\frac{(\widetilde{C}q)^{4q/3}}{(N\eta_1)^{q/2}}(1+(\mathbb{E}|G_{11}(\eta_1)|^{2q})^{1/6})(\mathbb{E}|G_{11}(\eta_1)|^{2q})^{1/2}\right)\\ &\le (3C)^q \left(1+\frac{(\widetilde{C}q)^{2q/3}}{(N\eta_1)^{q/4}}(\mathbb{E}|G_{11}(\eta_1)|^{2q})^{5/6}\right.\\ &\qquad\left.+\frac{(\widetilde{C}q)^{4q/3}}{(N\eta_1)^{q/2}}(\mathbb{E}|G_{11}(\eta_1)|^{2q})^{2/3}\right)\\ &\le (3C)^q \left(1+\frac{(\widetilde{C}q)^{2q/3}}{(N\eta_1)^{q/4}}(50C_0)^{5q/3}+\frac{(\widetilde{C}q)^{4q/3}}{(N\eta_1)^{q/2}}(50C_0)^{4q/3}\right)\\ &\le (3C)^q \left(1+\frac{K^q}{(N\eta_1)^{q/12}}\right).\end{aligned}$$

for a constant $K > 0$ depending only on C_0 (and on the universal constant $\widetilde{C}$). Choosing first $C_0 > 6C$ and then $M > K^{12}$, it follows that, for $N\eta_1 > M$, $K^q/(N\eta_1)^{q/12} < 1$ and therefore that

$$\mathbb{E}|G_{11}(\eta_1)|^q \le C_0^q.$$

This concludes the proof of Lemma 3.4. □

3.4 Non-optimal bound on moments of Λ

Combining Lemma 3.3 with Lemma 3.4, we obtain a non-optimal upper bound on the moments of Λ.

Lemma 3.5 *Assume* (1.9) *and fix* $\tilde{\eta} > 0$. *As usual, we set* $z = E + i\eta$ *and denote* $\Lambda = \Lambda(z)$.

(i) *There exist constants* $C, M > 0$ *such that*

$$\mathbb{E}|\Lambda|^q \le \frac{(Cq)^{\frac{2q}{3}}}{(N\eta)^{\frac{q}{4}}}$$

for all $0 < \eta \le \tilde{\eta}$, $|E| \le 2+\eta$, $N \ge 1$ *such that* $N\eta \ge M$, *and for all* $q \in \mathbb{N}$ *with* $q \le (N\eta)^{1/4}$.

(ii) *Fix* $\tilde{E} > 0$. *Then there exist constants* $C, M > 0$ *such that*

$$\mathbb{E}|\mathit{Im}\ \Lambda|^q \le \frac{(Cq)^{\frac{2q}{3}}}{(N\eta)^{\frac{q}{4}}}$$

for all $0 < \eta \le \tilde{\eta}$, $|E| \le \tilde{E}$, $N \ge 1$ *such that* $N\eta \ge M$, *and for all* $q \in \mathbb{N}$ *with* $q \le (N\eta)^{1/4}$.

4 Optimal bound on Λ; proof of Theorem 1

To prove Theorem 1, we will control the moments of the error term (2.10) through the control parameters

$$\mathcal{E}_q = \max\left\{\frac{1}{(N\eta)^{2q}}, \frac{(\mathrm{Im}\, m_{sc})^q + \mathbb{E}|\mathrm{Im}\, \Lambda|^q}{(N\eta)^q}\right\} + \frac{1}{N^q}.$$

We notice that, by definition, $\mathcal{E}_q \geq (N\eta)^{-2q}$. Moreover, assuming that $\mathbb{E}\,|\,\mathrm{Im}\ \Lambda|^{2q} \leq 1$ and $\mathbb{E}\,|\,\mathrm{Im}\ \Lambda|^p \leq 1$ (by Lemma 3.5, we know that these assumptions hold true, for $\max(p, 2q) \leq (N\eta)^{1/4}$), there exist universal constants $C, c > 0$ such that

$$c\mathcal{E}_p \leq \mathcal{E}_q \leq C\mathcal{E}_p^{q/p} \tag{4.1}$$

for all $1 \leq q \leq p$. In fact, the second inequality is a consequence of $\mathbb{E}\,|\,\mathrm{Im}\ \Lambda|^q \leq (\mathbb{E}\,|\,\mathrm{Im}\ \Lambda|^p)^{q/p}$. The first inequality in (4.1), on the other hand, can be proven as follows. If $(\mathrm{Im}\, m_{sc})^p + \mathbb{E}|\,\mathrm{Im}\ \Lambda|^p \leq (N\eta)^{-p}$, then

$$\mathcal{E}_p = \frac{1}{(N\eta)^{2p}} + \frac{1}{N^p} \leq \frac{1}{(N\eta)^{2q}} + \frac{1}{N^q} \leq \mathcal{E}_q.$$

If instead $(\mathrm{Im}\, m_{sc})^p + \mathbb{E}|\,\mathrm{Im}\ \Lambda|^p > (N\eta)^{-p}$, then

$$\mathcal{E}_p = \frac{(\mathrm{Im}\, m_{sc})^p + \mathbb{E}|\,\mathrm{Im}\ \Lambda|^p}{(N\eta)^p} + \frac{1}{N^p}.$$

For $p \geq 2q$, we find (under the assumption that $\mathbb{E}\,|\,\mathrm{Im}\ \Lambda|^p \leq 1$)

$$\mathcal{E}_p \leq \frac{2}{(N\eta)^{2q}} + \frac{1}{N^{2q}} \leq 2\mathcal{E}_q.$$

For $q \leq p < 2q$, we write $p = 2\alpha q + (1-\alpha)q$ (with $\alpha = (p-q)/q$) and we observe that, by Hölder's inequality (under the assumption $\mathbb{E}|\,\mathrm{Im}\ \Lambda|^{2q} \leq 1$)

$$\mathbb{E}|\,\mathrm{Im}\ \Lambda|^p \leq (\mathbb{E}|\,\mathrm{Im}\ \Lambda|^q)^{1-\alpha}\,(\mathbb{E}|\,\mathrm{Im}\ \Lambda|^{2q})^\alpha \leq (\mathbb{E}|\,\mathrm{Im}\ \Lambda|^q)^{1-\alpha}.$$

This gives

$$\begin{aligned}\frac{(\mathrm{Im}\, m_{sc})^p + \mathbb{E}|\,\mathrm{Im}\ \Lambda|^p}{(N\eta)^p} &\leq C\left(\frac{(\mathrm{Im}\, m_{sc})^q + \mathbb{E}|\,\mathrm{Im}\ \Lambda|^q}{(N\eta)^q}\right)^{1-\alpha}\left(\frac{1}{(N\eta)^{2q}}\right)^\alpha \\ &\leq C\max\left\{\frac{1}{(N\eta)^{2q}}, \frac{(\mathrm{Im}\, m_{sc})^q + \mathbb{E}|\,\mathrm{Im}\ \Lambda|^q}{(N\eta)^q}\right\}\end{aligned}$$

and proves that $\mathcal{E}_p \leq C\mathcal{E}_q$.

The proof of Theorem 1 is based on the following lemma.

Lemma 4.1 *Assume* (1.9), *fix* $\tilde{E} > 0$ *and* $\tilde{\eta} > 0$. *Set* $z = E + i\eta$. *There exist constants* $C, M, c_0 > 0$ *such that*

$$\mathbb{E}\left|\frac{1}{N}\sum_k Z_k G_{kk}\right|^{2q} \leq (Cq)^{cq^2}\mathcal{E}_{4q}^{\frac{1}{2}},$$

for all $|E| \leq \tilde{E}, 0 < \eta \leq \tilde{\eta}, N \geq 1$ *such that* $N\eta \geq M$, *and for all* $q \in \mathbb{N}$ *with* $q \leq c_0(N\eta)^{1/8}$.

Technically, the proof of this lemma is the main part of the paper. We defer it to the next section. Assuming the result of Lemma 4.1, we can now proceed with the proof of Theorem 1.

Proof of Theorem 1 We prove first the bound (1.11) for the imaginary part of Λ. According to (2.15), we need to control the moments of $R = N^{-1}\sum_{j=1}^N \Upsilon_j G_{jj}$. Recalling the definition (2.7) of the coefficients Υ_j, we obtain

$$\mathbb{E}|R|^{2q} \leq C^q\left(\mathbb{E}\left|\frac{1}{N}\sum_j h_{kk}\,G_{kk}\right|^{2q} + \mathbb{E}\left|\frac{1}{N^2}\sum_{k,j} G_{kj}G_{jk}\right|^{2q} + \mathbb{E}\left|\frac{1}{N}\sum_k Z_k G_{kk}\right|^{2q}\right) \tag{4.2}$$

where $Z_k = (\mathbb{I} - \mathbb{E}_k)\,\mathbf{a}_k^* G^{(k)}\mathbf{a}_k$. From the assumption (1.9) and Lemma 3.4, the first term in the parenthesis can be bounded by

$$\mathbb{E}\left|\frac{1}{N}\sum_j h_{kk}\,G_{kk}\right|^{2q} \leq \mathbb{E}|h_{11}G_{11}|^{2q} \leq \frac{(Cq)^q}{N^q}(\mathbb{E}|G_{11}|^{4q})^{1/2} \leq \frac{(Cq)^q}{N^q} \tag{4.3}$$

for all $1 \leq q \leq (N\eta)^{1/4}$. As for the second term on the r.h.s. of Eq. (4.2), we get

$$\begin{aligned}\mathbb{E}\left|\frac{1}{N^2}\sum_{k,j} G_{kj}G_{jk}\right|^{2q} &\leq \mathbb{E}\left(\frac{1}{N^2}\sum_{k,j}|G_{jk}|^2\right)^{2q} \leq \mathbb{E}\left(\frac{\text{Im } m}{N\eta}\right)^{2q} \\ &\leq C^q\,\frac{(\text{Im } m_{sc})^{2q} + \mathbb{E}|\text{ Im } \Lambda|^{2q}}{(N\eta)^{2q}}.\end{aligned} \tag{4.4}$$

Combining (4.3) and (4.4) with Lemma 4.1, we find

$$\mathbb{E}|R|^{2q} \leq (Cq)^{cq^2}\mathcal{E}_{4q}^{\frac{1}{2}} \tag{4.5}$$

for all $1 \leq q \leq c_0(N\eta)^{1/8}$. Next, we fix N, η and q with $1 \leq q \leq c_0(N\eta)^{1/8}$, and we insert the last bound on the r.h.s. of (2.15). We distinguish several cases. We can assume that

$$(\text{Im } m_{sc})^{4q} + \mathbb{E}|\text{ Im } \Lambda|^{4q} \geq \frac{1}{(N\eta)^{4q}}$$

since otherwise there is nothing to prove. In this case,

$$\mathcal{E}_{2q} = \frac{(\text{Im } m_{sc})^{2q} + \mathbb{E}|\text{ Im } \Lambda|^{2q}}{(N\eta)^{2q}} + \frac{1}{N^{2q}} = \frac{(\text{Im } m_{sc})^{2q} + \eta^{2q} + \mathbb{E}|\text{ Im } \Lambda|^{2q}}{(N\eta)^{2q}}.$$

If $\mathbb{E}|\text{ Im } \Lambda|^{2q} \leq (\text{Im } m_{sc})^{2q} + \eta^{2q}$, we use the bound proportional to $|R|$ in (2.15). We find

$$\begin{aligned}\mathbb{E}|\text{ Im } \Lambda|^{q} &\leq \frac{C^q\, \mathbb{E}|R|^q}{|m_{sc}^2 - 1|^q} \leq \frac{(Cq)^{cq^2}}{|m_{sc}^2 - 1|^q} \left(\frac{(\text{Im } m_{sc})^{2q} + \eta^{2q}}{(N\eta)^{2q}} \right)^{\frac{1}{2}} \\ &\leq \frac{(Cq)^{cq^2}}{(N\eta)^q} \left[\frac{\text{Im}\, m_{sc}}{|m_{sc}^2 - 1|} + \frac{\eta}{|m_{sc}^2 - 1|} \right]^q.\end{aligned}$$

From Proposition 2.1, we conclude

$$\mathbb{E}\,|\text{ Im } \Lambda|^q \leq \frac{(Cq)^{cq^2}}{(N\eta)^q}. \tag{4.6}$$

If, on the other hand, $(\text{Im } m_{sc})^{2q} + \eta^{2q} \leq \mathbb{E}|\text{ Im } \Lambda|^{2q}$, we use the bound proportional to $|R|^{1/2}$ on the r.h.s. of (2.15). We find

$$\mathbb{E}|\text{ Im } \Lambda|^{2q} \leq C^q\, \mathbb{E}|R|^q \leq (Cq)^{cq^2} \left(\frac{\mathbb{E}|\text{ Im } \Lambda|^{2q}}{(N\eta)^{2q}} \right)^{\frac{1}{2}}.$$

This implies that $\mathbb{E}|\text{ Im } \Lambda|^{2q} \leq (Cq)^{2cq^2}/(N\eta)^{2q}$ and therefore that

$$\mathbb{E}|\text{ Im } \Lambda|^q \leq \sqrt{\mathbb{E}|\text{ Im } \Lambda|^{2q}} \leq \frac{(Cq)^{cq^2}}{(N\eta)^q}.$$

Combined with (4.6), this implies that

$$\mathbb{P}\left(|\text{ Im } m(z) - \text{Im } m_{sc}(z)| \geq \frac{K}{N\eta} \right) \leq \frac{(N\eta)^q}{K^q} \mathbb{E}|\text{ Im } \Lambda|^q \leq \frac{(Cq)^{cq^2}}{K^q}.$$

for all $1 \leq q \leq c_0 (N\eta)^{1/8}$ and concludes the proof of (1.11).

To prove (1.10), we proceed similarly, using however (2.15) instead of (2.14). □

5 Proof of Lemma 4.1

We rewrite the quantity we are interested in in a more convenient form. Set

$$W_k = Z_k G_{kk},$$

then

$$\frac{1}{N}\sum_k Z_k G_{kk} = \frac{1}{N}\sum_k W_k = \frac{1}{N}\sum_k (\mathbb{I} - \mathbb{E}_k) W_k + \frac{1}{N}\sum_k \mathbb{E}_k W_k.$$

By Hölder inequality we get

$$\mathbb{E}\left|\frac{1}{N}\sum_k Z_k G_{kk}\right|^{2q} \le C^q \mathbb{E}\left|\frac{1}{N}\sum_k (\mathbb{I} - \mathbb{E}_k) W_k\right|^{2q} + C^q \mathbb{E}\left|\mathbb{E}_1 W_1\right|^{2q}. \tag{5.1}$$

Next we claim that under the assumptions of Theorem 1

$$\mathbb{E}\left|\mathbb{E}_1 W_1\right|^{2q} \le (Cq)^{4q} \mathcal{E}_{4q}^{\frac{1}{2}} \tag{5.2}$$

and

$$\mathbb{E}\left|\frac{1}{N}\sum_k (\mathbb{I} - \mathbb{E}_k) W_k\right|^{2q} \le (Cq)^{cq^2} \mathcal{E}_{4q}^{\frac{1}{2}}. \tag{5.3}$$

Lemma 4.1 then follows by inserting the bounds (5.2) and (5.3) in (5.1). The rest of this section is devoted to the proof of (5.2) and (5.3).

5.1 Proof of Eq. (5.2)

Recalling that

$$Z_k = -(\mathbb{I} - \mathbb{E}_k)\frac{1}{G_{kk}},$$

we find

$$\mathbb{E}_k W_k = \mathbb{E}_k \frac{G_{kk}}{(\mathbb{E}_k \frac{1}{G_{kk}})}\left((\mathbb{I} - \mathbb{E}_k)\frac{1}{G_{kk}}\right)^2 = \frac{\mathbb{E}_k G_{kk} Z_k^2}{(\mathbb{E}_k \frac{1}{G_{kk}})}. \tag{5.4}$$

for any $k = 1, \ldots, N$. To bound the denominator on the r.h.s. of the last equation, we make use of the following lemma.

Lemma 5.1 *Assume* (1.9) *and fix* $\tilde{E} > 0$, $\tilde{\eta} > 0$. *Set* $z = E + i\eta$. *There exist constants* $c, C_1, M > 0$ *such that*

$$\mathbb{E}\left|\frac{1}{\mathbb{E}_1 \frac{1}{G_{11}}}\right|^q \leq C_1^q, \tag{5.5}$$

for all $|E| \leq \tilde{E}, \eta \leq \tilde{\eta}, N \geq 1$ *such that* $N\eta \geq M$, *and for all* $q \in \mathbb{N}$ *with* $q \leq c(N\eta)^{1/4}$.

Proof Define

$$\widetilde{G_{11}} \equiv \frac{1}{\mathbb{E}_1 \frac{1}{G_{11}}} = -\frac{1}{\frac{N-1}{N} m^{(1)} + z - h_{11}}$$

with $m^{(1)} = \frac{1}{N-1} \operatorname{Tr} G^{(1)}$. We notice that

$$\left|\frac{d}{d\nu} \log \widetilde{G_{11}}(E + i\nu)\right| = \left|\frac{i + \frac{N-1}{N} \frac{d}{d\nu} m^{(1)}}{\frac{N-1}{N} m^{(1)} + z - h_{11}}\right| \leq \frac{\nu + \frac{N-1}{N} \operatorname{Im} m^{(1)}}{\nu |\frac{N-1}{N} m^{(1)} + z - h_{11}|} \leq \frac{1}{\nu}.$$

As argued in the proof of Lemma 3.4, this implies that

$$|\widetilde{G_{11}}(E + i\eta/s)| \leq s|\widetilde{G_{11}}(E + i\eta)|.$$

Next we observe that

$$\widetilde{G_{11}} = G_{11} + G_{11}\widetilde{G_{11}} Z_1$$

where we recall the definition $Z_1 = (\mathbb{I} - \mathbb{E}_1)\mathbf{a}_1^* G^{(1)} \mathbf{a}_1 = (\mathbb{I} - \mathbb{E}_1) G_{11}^{-1}$. From Proposition 3.2 and Lemma 3.4 we get

$$\mathbb{E}|\widetilde{G_{11}}|^q \leq C^q + \frac{(Cq)^q}{(N\eta)^{q/2}} (\mathbb{E}|\widetilde{G_{11}}|^{3q})^{1/3},$$

provided $3q \leq (N\eta)^{1/4}$. Here we used the fact that $|m_{sc}| < 1$ and $\mathbb{E}| \operatorname{Im} \Lambda|^{3q} \leq 1$ by Lemma 3.5. At this point, we can use a bootstrap argument similar to the one used in Lemma 3.4 to conclude the proof. □

Applying (5.5) to (5.4) and using also Lemma 3.4 and Proposition 3.2 we conclude that

$$\begin{aligned}
\mathbb{E}\,|\mathbb{E}_1 W_1|^{2q} &\leq (\mathbb{E}|G_{11}|^{8q})^{\frac{1}{4}} \left(\mathbb{E}\left|\frac{1}{\mathbb{E}_1 \frac{1}{G_{11}}}\right|^{8q}\right)^{\frac{1}{4}} (\mathbb{E}|Z_1|^{8q})^{\frac{1}{2}} \\
&\leq (Cq)^{4q} \left(\frac{(\operatorname{Im} m_{sc})^{4q} + \mathbb{E}|\operatorname{Im} \Lambda|^{4q}}{(N\eta)^{4q}} + \frac{1}{(N\eta)^{8q}}\right)^{\frac{1}{2}},
\end{aligned}$$

This immediately implies (5.2).

5.2 Preliminaries to the proof of Eq. (5.3): the expansion algorithm

In order to prove the bound (5.3), we will follow Theorems 4.6 and 4.7 in [10] to expand the expectation on the l.h.s. in a sum over indices $k_1, \ldots, k_{2q}$ of products of terms of the form

$$(\mathbb{I} - \mathbb{E}_{k_j}) W_{k_j} = (\mathbb{I} - \mathbb{E}_{k_j}) \left[(\mathbb{I} - \mathbb{E}_{k_j}) \frac{1}{G_{k_j k_j}} \right] G_{k_j k_j}. \tag{5.6}$$

In each one of these factors, we will expand further the resolvent entries (both in the numerator and in the denominator) using the relation (2.5). The gain here is that we either gain independence [the first term on the r.h.s. of (2.5) does not depend on the randomness in the j-th row and column of the original matrix] or, alternatively, we gain smallness [the second term on the r.h.s. of (2.5) has two more off-diagonal entries, which are typically of size $(N\eta)^{-1/2} \ll 1$]. Recursively, we continue to expand the resulting terms using either (2.5) or a similar formula for the off-diagonal entries [see (5.7) below], either until there are no more variables over which we can expand (in which case we reached maximal independence) or until there are sufficiently many off-diagonal terms [to show that (2.5) is of the order $(N\eta)^{-2q}$]. This type of expansions for resolvent entries have first been applied in the analysis of Wigner matrices in [20–22]; we follow here the more recent work [10].

Next, we give a precise and detailed definition of the expansion algorithm. Afterwards, we apply it to the resolvent entries $G_{k_j k_j}$ and $1/G_{k_j k_j}$ appearing in (5.6).

5.2.1 A general description of the expansion algorithm

Let $\mathbb{T} \subset \{1, \ldots, N\}$ be a set of indices. We denote by $H^{(\mathbb{T})}$ the $(N - |\mathbb{T}|) \times (N - |\mathbb{T}|)$ minor of the matrix H obtained by deleting the rows and columns of H corresponding to the indices in $\mathbb{T}$; the entries of the matrix $H^{(\mathbb{T})}$ are h_{ij} with $i, j \in \{1, \ldots, N\} \backslash \mathbb{T}$. We denote by $G^{(\mathbb{T})}$ the matrix

$$G^{(\mathbb{T})} = \left(H^{(\mathbb{T})} - z \right)^{-1}.$$

Rows and columns of $G^{(\mathbb{T})}$ are also labelled with indices $i, j \in \{1, \ldots, N\} \backslash \mathbb{T}$.

Similarly to (2.5), we have (see, e.g., Eq. (4.6) in [10])

$$G_{ij}^{(\mathbb{T})} = G_{ij}^{(\mathbb{T}k)} + \frac{G_{ik}^{(\mathbb{T})} G_{kj}^{(\mathbb{T})}}{G_{kk}^{(\mathbb{T})}} \quad \forall i, j, k \notin \mathbb{T} \text{ and } i, j \neq k. \tag{5.7}$$

Setting $i = j$, we find

$$G_{ii}^{(\mathbb{T})} = G_{ii}^{(\mathbb{T}k)} + \frac{G_{ik}^{(\mathbb{T})} G_{ki}^{(\mathbb{T})}}{G_{kk}^{(\mathbb{T})}} \quad \forall i, k \notin \mathbb{T} \text{ and } i \neq k \tag{5.8}$$

and thus

$$\frac{1}{G_{ii}^{(\mathbb{T})}} = \frac{1}{G_{ii}^{(\mathbb{T}k)}} - \frac{G_{ik}^{(\mathbb{T})}G_{ki}^{(\mathbb{T})}}{G_{ii}^{(\mathbb{T})}G_{ii}^{(\mathbb{T}k)}G_{kk}^{(\mathbb{T})}} \quad \forall i, k \notin \mathbb{T} \text{ and } i \neq k. \tag{5.9}$$

For any $\mathbf{k} = (k_1, \ldots, k_{2q})$, with $k_s \in \{1, \ldots, N\}$ for $s = 1, \ldots, 2q$, let $\mathcal{P}(\mathbf{k})$ be the partition of $\{1, \ldots, 2q\}$ induced by the coincidences in $\mathbf{k}$. In other words, $\mathcal{P}(\mathbf{k})$ is the partition induced by the equivalence relation on $\{1, \ldots, 2q\}$ defined by $r \sim s$ if and only if $k_r = k_s$. We denote, moreover, by $\mathfrak{P}_{2q}$ the set of all partitions of $\{1, \ldots, 2q\}$.

Following [10], we can write the expectation on the l.h.s. of Eq. (5.3) as

$$\mathbb{E}\left|\frac{1}{N}\sum_k(\mathbb{I} - \mathbb{E}_k)W_k\right|^{2q} = \frac{1}{N^{2q}}\sum_{P\in\mathfrak{P}_{2q}}\sum_{\mathbf{k}}\mathbf{1}(\mathcal{P}(\mathbf{k}) = P)V(\mathbf{k}) \tag{5.10}$$

where each coordinate of $\mathbf{k} = (k_1, \ldots, k_{2q})$ is summed over the set $\{1, \ldots, N\}$, and

$$V(\mathbf{k}) = \mathbb{E}(\mathbb{I} - \mathbb{E}_{k_1})W_{k_1} \cdots (\mathbb{I} - \mathbb{E}_{k_q})W_{k_q}\overline{(\mathbb{I} - \mathbb{E}_{k_{q+1}})W_{k_{q+1}}} \cdots \overline{(\mathbb{I} - \mathbb{E}_{k_{2q}})W_{k_{2q}}}. \tag{5.11}$$

For any fixed $\mathbf{k}$ we say that $s \in \{1, \ldots, 2q\}$ is a *lone* label if $k_s \neq k_r$ for any $r \in \{1, \ldots, 2q\}\backslash\{s\}$, i.e., if s is the only element in its equivalence class in the partition $\mathcal{P}(\mathbf{k})$. We denote by $L(\mathbf{k})$ the set of lone indices and by $\mathbf{k}_L \subset \mathbf{k}$ the set of coordinates of $\mathbf{k}$ associated with lone labels.

For a fixed $\mathbf{k}$, we say that a resolvent entry $G_{ij}^{(\mathbb{T})}$ with $i, j \notin \mathbb{T}$ is *maximally expanded* if $\mathbf{k}_L \subseteq \mathbb{T} \cup \{i, j\}$. Similarly, we say that a factor $G_{ii}^{(\mathbb{T})}$ or $1/G_{ii}^{(\mathbb{T})}$ with $i \notin \mathbb{T}$ is maximally expanded, if $\mathbf{k}_L \subset \mathbb{T} \cup \{i\}$.

Let Q be a product of diagonal and/or off-diagonal resolvent entries of the form

$$Q = G_{ii}^{(\mathbb{T})} \quad \text{or} \quad Q = \frac{1}{G_{i_1i_1}^{(\mathbb{T}_1)}} \cdots \frac{1}{G_{i_li_l}^{(\mathbb{T}_l)}} G_{i_{l+1}j_{l+1}}^{(\mathbb{T}_{l+1})} \cdots G_{i_{l+m}j_{l+m}}^{(\mathbb{T}_{l+m})} \tag{5.12}$$

for arbitrary non-negative integers l, m and for $i_r \neq j_r$ and $i_r, j_r \notin \mathbb{T}_r$, and $i \notin \mathbb{T}$. We assume the resolvent entries on the r.h.s. of Eq. (5.12) to be ordered in some way.

For fixed $\mathbf{k}$, we define, similarly to [10], the operation w on monomials of the form of Q.

- The operation $w(Q)$ can be performed only if at least one resolvent entry in Q (either in the numerator or in the denominator) is not maximally expanded.
- If at least one of the resolvent entries of Q is not maximally expanded, w acts only on the first (according to the previously chosen order) resolvent entry of Q which is not maximally expanded. We set then $w(Q) = w_0(Q) + w_1(Q)$ where w_0 and w_1 are defined by the following rules:
  - If the first not maximally expanded resolvent entry is $G_{ij}^{(\mathbb{T})}$, we use (5.7) to define w_0 and w_1 in the following way:

$$G_{ij}^{(\mathbb{T})} \xrightarrow{w_0} G_{ij}^{(\mathbb{T}u)} \quad \text{and} \quad G_{ij}^{(\mathbb{T})} \xrightarrow{w_1} \frac{1}{G_{uu}^{(\mathbb{T})}} G_{iu}^{(\mathbb{T})} G_{uj}^{(\mathbb{T})} \tag{5.13}$$

where u is the smallest index such that $u \in \mathbf{k}_L \backslash (\mathbb{T} \cup \{i, j\})$.

- the first not maximally expanded resolvent entry is $G_{ii}^{(\mathbb{T})}$, we use (5.8) to define w_0 and w_1 in the following way:

$$G_{ii}^{(\mathbb{T})} \xrightarrow{w_0} G_{ii}^{(\mathbb{T}u)} \quad \text{and} \quad G_{ii}^{(\mathbb{T})} \xrightarrow{w_1} \frac{G_{iu}^{(\mathbb{T})} G_{ui}^{(\mathbb{T})}}{G_{uu}^{(\mathbb{T})}} \tag{5.14}$$

where u is the smallest index such that $u \in \mathbf{k}_L \backslash (\mathbb{T} \cup \{i\})$.

- If the first not maximally expanded resolvent entry is $1/G_{ii}^{(\mathbb{T})}$, we use (5.9) to define w_0 and w_1 in the following way:

$$\frac{1}{G_{ii}^{(\mathbb{T})}} \xrightarrow{w_0} \frac{1}{G_{ii}^{(\mathbb{T}u)}} \quad \text{and} \quad \frac{1}{G_{ii}^{(\mathbb{T})}} \xrightarrow{w_1} -\frac{1}{G_{ii}^{(\mathbb{T})}} \frac{1}{G_{uu}^{(\mathbb{T})}} \frac{1}{G_{ii}^{(\mathbb{T}u)}} G_{iu}^{(\mathbb{T})} G_{ui}^{(\mathbb{T})} \tag{5.15}$$

where u is the smallest index such that $u \in \mathbf{k}_L \backslash (\mathbb{T} \cup \{i\})$.

- With this definition, we rewrite Q as $w(Q) = w_0(Q) + w_1(Q)$ where $w_0(Q)$ and $w_1(Q)$ are two new monomials in diagonal and off-diagonal resolvent entries of the form (5.12).

For any starting monomial Q, we use the operation w repeatedly; we decompose Q in a sum of terms having the form (5.12). To keep track of all the resulting terms, let σ be a (ordered, finite) string of 0 and 1. For any fixed σ we construct, analogously to [10], the monomial Q_σ through the following recursion

$$Q_0 = w_0(Q) \quad \text{and} \quad Q_1 = w_1(Q)$$

and

$$Q_{\sigma' 0} = w_0(Q_{\sigma'}) \quad \text{and} \quad Q_{\sigma' 1} = w_1(Q_{\sigma'}).$$

To write Q as an appropriate sum of factors Q_σ we apply recursively the operation w as many times as allowed by the following stopping rule:

($\mathcal{SR}$) Continue to apply w to Q_σ until either all the resolvent entries of Q_σ are maximally expanded or the number of off-diagonal resolvent entries in Q_σ is greater than $2q$.

In this way we obtain

$$Q = \sum_{\sigma \in \mathcal{L}_Q} Q_\sigma$$

where $\mathcal{L}_Q$ is the set of all strings σ such that Q_σ satisfies the stopping rule.

5.2.2 Application to terms of the form (5.6)

We use now the expansion algorithm to expand terms of the form (5.6), which arise from (5.3). We consider theinitial monomials $A^r = 1/G_{k_r k_r}$ and $B^r = G_{k_r k_r}$ (in [10] there were only terms of the type A^r; this is the reason why the operation (5.14) was not used there). Applying the expansion algorithm presented in the previous section, we find

$$A^r := \frac{1}{G_{k_r k_r}} := \sum_{\sigma \in \mathcal{L}_r} A^r_\sigma \tag{5.16}$$

and

$$B^r := G_{k_r k_r} := \sum_{\rho \in \mathcal{M}_r} B^r_\rho \tag{5.17}$$

where the sets $\mathcal{L}_r$ and $\mathcal{M}_r$ are chosen such that A^r_σ and B^r_ρ satisfy the stopping rule $(\mathcal{SR})$.

We denote by $\gamma(A^r_\sigma)$ (respectively $\gamma(B^r_\rho)$) the number of off-diagonal resolvent entries in A^r_σ (respectively B^r_ρ). From the stopping rule $(\mathcal{SR})$, it follows that for all $\sigma \in \mathcal{L}_r$, we can have only two possibilities: either all resolvent entries in A^r_σ are maximally expanded and $\gamma(A^r_\sigma) \leq 2q$, or alternatively, $2q+1 \leq \gamma(A^r_\sigma) \leq 2q+2$. This follows because, by (5.13)–(5.15), the operation w increases the number of off-diagonal resolvent entries by at most two. Analogously, we find for B^r_ρ that, for all $\rho \in \mathcal{M}_r$, either all resolvent entries are maximally expanded and $\gamma(B^r_\rho) \leq 2q$, or, alternatively, $2q+1 \leq \gamma(B^r_\rho) \leq 2q+2$.

We denote by $\delta(A^r_\sigma)$ (respectively $\delta(B^r_\rho)$) the number of diagonal resolvent entries in A^r_σ (respectively B^r_ρ), appearing either in the numerator or in the denominator. We note that the difference $\delta(A^r_\sigma) - \gamma(A^r_\sigma)$ is invariant with respect to the operation w [because, when dealing with terms arising from the initial A^r, we only apply (5.13) or (5.15), never (5.14)]. This implies that

$$\delta(A^r_\sigma) = \gamma(A^r_\sigma) + 1. \tag{5.18}$$

On the other hand, the difference $\delta(B^r_\rho) - \gamma(B^r_\rho)$ is not always invariant. In fact, the operation w_1 applied to a resolvent entry in the numerator, according to (5.14), decreases it by two. We remark, however, that this operation can take place at most once (because w_1 in (5.14) removes the diagonal entry from the numerator). As a consequence, we have either $\delta(B^r_\rho) = 1$ and $\gamma(B^r_\rho) = 0$ (if the operation w_1 in (5.14) never occurs) or $\delta(B^r_\rho) - \gamma(B^r_\rho) = -1$ (if the operation w_1 in (5.14) takes place once). We conclude that

$$\delta(B^r_\rho) = \max\{\gamma(B^r_\rho) - 1, 1\}. \tag{5.19}$$

It follows from (5.18) and (5.19) and from the previous bounds on the number of off-diagonal entries, that the number of diagonal entries in A^r_σ and B^r_ρ is always bounded by $(2q+3)$ (even by $(2q+1)$ in the B^r_ρ terms). Hence the total number of resolvent entries in A^r_σ and in B^r_ρ (diagonal and off-diagonal, in the numerator and in the denominator) is bounded by $4q+5$.

Following [10], we denote by $b(\sigma)$ and $b(\rho)$ the number of ones in the strings σ and ρ, respectively. For any $r \in \{1, \ldots, 2q\}$ and for any $\sigma \in \mathcal{L}_r$ and $\rho \in \mathcal{M}_r$, we have:

1. If $b(\sigma) \geq 1$ then $\gamma(A^r_\sigma) \geq b(\sigma) + 1$. Analogously, if $b(\rho) \geq 1$ then $\gamma(B^r_\rho) \geq b(\rho)+1$. This follows from the observation that the first application of w_1 generates [according to (5.14) and (5.15)] two off-diagonal entries while all further applications create [according to (5.13), (5.14), (5.15)] at least one additional off-diagonal entry.
2. $b(\sigma) \leq 2q$ and $b(\rho) \leq 2q$. The first application of w_1 generates two off-diagonal terms. Each subsequent application of w_1 generates at least one more off-diagonal term. Hence $2q$ applications of w_1 create at least $(2q + 1)$ off diagonal terms, which are enough to satisfy the stopping rule ($\mathcal{SR}$).
3. The number of zeros in the strings σ and in ρ is bounded by $2q(4q+5)$. This follows because every application of w_0 produces one additional "top" index (indicating that the generated entry is independent of on additional row and column). The total number of "top" indices is bounded, however, by the number of resolvent entries (which, as discussed above, is at most $(4q + 5)$) times $2q$ (the number of coordinates of $\mathbf{k}$).
4. The length of σ and ρ is bounded by $4q(2q + 3)$ (this follows combining the bounds for the number of zeros and the number of ones in the strings).
5. Considering that the length of σ and ρ is at most $4q(2q + 3)$ and that the number of ones is at most $2q$, we can bound the cardinality of $\mathcal{L}_r$ and $\mathcal{M}_r$ by

$$|\mathcal{L}_r|, |\mathcal{M}_r| \leq \sum_{k=0}^{2q} \binom{4q(2q+3)}{k} \leq (Cq)^{2q}. \tag{5.20}$$

5.3 Preliminaries to the proof of Eq. (5.3): bounds on resolvent entries

We are going to use (5.16) and (5.17) to expand the initial resolvent entries $A^r = 1/G_{k_r k_r}$ and $B^r = G_{k_r k_r}$. To prove a bound of the form (5.3), we need to estimate the resolvent entries appearing in the expanded terms A^r_σ and B^r_ρ. More precisely, after applying Hölder's inequality to separate the many factors in the products A^r_σ, B^r_ρ, we will need control on high moments of quantities of the form

$$|G^{(\mathbb{T})}_{kk}|, \quad \frac{1}{|G^{(\mathbb{T})}_{kk}|}, \quad \left|(\mathbb{I} - \mathbb{E}_k)\frac{1}{G^{(\mathbb{T})}_{kk}}\right|, \quad |G^{(\mathbb{T})}_{kl}|. \tag{5.21}$$

While the first two terms in (5.21) are typically of order one, the last two are expected to be small. To show (5.3), it is important to extract the correct small factor in the bounds for these quantities.

Bounds for moments of $|G_{kk}|$ have already been obtained in Lemma 3.9. Similarly one can also estimate moments of $|G^{(\mathbb{T})}_{kk}|$, for $\mathbb{T} \subset \{1, \ldots, N\}$ with $|\mathbb{T}| \leq 2q$ and $k \notin \mathbb{T}$.

Lemma 5.2 *Assume* (1.9), *fix* $\tilde{E} > 0$, $\tilde{\eta} > 0$. *Set* $z = E + i\eta$. *There exist constants* $c, C, M_1, M_2 > 0$ *such that*

$$\mathbb{E}\frac{1}{|G_{11}^{(\mathbb{T})}|^{2q}} \le C^q$$

for all $|E| \le \tilde{E}$, $\eta \le \tilde{\eta}$, $N > M_1$ *such that* $N\eta \ge M_2$, $q \in \mathbb{N}$ *with* $q \le c(N\eta)^{1/4}$ *and* $\mathbb{T} \subset \{1, \ldots, N\}$ *with* $|\mathbb{T}| \le 2q$.

Remark For $\mathbb{T} \subset \{1, \ldots, N\}$ with $|\mathbb{T}| \le 2q$, let $\Lambda^{(\mathbb{T})} = m^{(\mathbb{T})} - m_{sc}$, where $m^{(\mathbb{T})}$ is the Stieltjes transform of $H^{(\mathbb{T})}$. By the interlacing properties of the eigenvalues of H and $H^{(\mathbb{T})}$ it is easy to check that $|m - m^{(\mathbb{T})}| \le C|\mathbb{T}|/(N\eta)$, which implies

$$\mathbb{E}|\Lambda^{(\mathbb{T})}|^q \le C^q\mathbb{E}|\Lambda|^q + \frac{(Cq)^q}{(N\eta)^q} \quad \text{and} \quad \mathbb{E}|\operatorname{Im}\Lambda^{(\mathbb{T})}|^q \le C^q\mathbb{E}|\operatorname{Im}\Lambda|^q + \frac{(Cq)^q}{(N\eta)^q}. \tag{5.22}$$

We will use this observation to reduce our analysis to the case $\mathbb{T} = \emptyset$.

Proof By the above remark, we can take $\mathbb{T} = \emptyset$. We have

$$\begin{aligned}\mathbb{E}\frac{1}{|G_{11}|^{2q}} &= \mathbb{E}|h_{11} - z - \mathbf{a}_1^* G^{(1)}\mathbf{a}_1|^{2q} \le \frac{(Cq)^q}{N^q} + C^q + C^q\,\mathbb{E}\,|\mathbf{a}_1^* G^{(1)}\mathbf{a}_1|^{2q} \\ &\le C^q + C^q\,\mathbb{E}\,|\mathbf{a}_1^* G^{(1)}\mathbf{a}_1|^{2q}.\end{aligned}$$

We write

$$\begin{aligned}\mathbb{E}|\mathbf{a}_1^* G^{(1)}\mathbf{a}_1|^{2q} &= \mathbb{E}|\mathbf{a}_1^* G^{(1)}\mathbf{a}_1 - \mathbb{E}_1\mathbf{a}_1^* G^{(1)}\mathbf{a}_1 + \mathbb{E}_1\mathbf{a}_1^* G^{(1)}\mathbf{a}_1|^{2q} \\ &\le C^q(\mathbb{E}|\mathbf{a}_1^* G^{(1)}\mathbf{a}_1 - \mathbb{E}_1\mathbf{a}_1^* G^{(1)}\mathbf{a}_1|^{2q} + \mathbb{E}|\mathbb{E}_1\mathbf{a}_1^* G^{(1)}\mathbf{a}_1|^{2q}).\end{aligned} \tag{5.23}$$

For the first term we use the Hanson–Wright large deviation estimate (see Proposition 3.2). We find

$$\mathbb{E}|\mathbf{a}_1^* G^{(1)}\mathbf{a}_1 - \mathbb{E}_1\mathbf{a}_1^* G^{(1)}\mathbf{a}_1|^{2q} \le (Cq)^{2q}\left(\frac{(\operatorname{Im} m_{sc})^q + \mathbb{E}|\operatorname{Im}\Lambda|^q}{(N\eta)^q} + \frac{1}{(N\eta)^{2q}}\right) \le 1$$

since, by Lemma 3.5, $\mathbb{E}|\operatorname{Im}\Lambda|^q < 1$ for all $q \le (N\eta)^{1/4}$, and N large enough.

As for the second term on the r.h.s. of (5.23), we find

$$\mathbb{E}|\mathbb{E}_1\mathbf{a}_1^* G^{(1)}\mathbf{a}_1|^{2q} = \mathbb{E}\left|\frac{1}{N}\operatorname{Tr} G^{(1)}\right|^{2q} \le C^q\left(\mathbb{E}\left|\frac{1}{N}\operatorname{Tr} G\right|^{2q} + \frac{1}{(N\eta)^{2q}}\right) \le C^q$$

from $\mathbb{E}\left|\frac{1}{N}\operatorname{Tr} G\right|^{2q} \le \mathbb{E}|G_{11}|^{2q} \le C^q$ (by Lemma 3.4). □

To estimate the third term in (5.21), we recall that [similarly to (3.3)]

$$(\mathbb{I}-\mathbb{E}_k)\frac{1}{G_{kk}^{(\mathbb{T})}} = -Z_k^{(\mathbb{T})},$$

where we set $Z_k^{(\mathbb{T})} = (\mathbb{I}-\mathbb{E}_k)\mathbf{a}_k^{(\mathbb{T})*}G^{(\mathbb{T})}\mathbf{a}_k^{(\mathbb{T})}$, and where $\mathbf{a}_l^{(\mathbb{T})}$ is the l-th column of the matrix H without the elements corresponding to the labels in $\mathbb{T}$. Applying Proposition 3.2 to $Z_k^{(\mathbb{T})}$ and using the remark after Lemma 5.2, we find that (for $q \le N/2$)

$$\begin{aligned}\mathbb{E}\left|(\mathbb{I}-\mathbb{E}_k)\frac{1}{G_{kk}^{(\mathbb{T})}}\right|^{2q} &\le (Cq)^{2q}\left(\frac{(\operatorname{Im}\, m_{sc})^q + \mathbb{E}|\operatorname{Im}\,\Lambda^{(\mathbb{T})}|^q}{((N-|\mathbb{T}|)\eta)^q} + \frac{1}{((N-|\mathbb{T}|)\eta)^{2q}}\right)\\ &\le (Cq)^{3q}\mathcal{E}_q.\end{aligned}$$

Finally, we show how to estimate the last term in (5.21). To this end, we use the formula (see, e.g., Eq. (2.8) in [4])

$$G_{kl}^{(\mathbb{T})} = G_{ll}^{(\mathbb{T})}G_{kk}^{(\mathbb{T}l)}(h_{kl} - \mathbf{a}_k^{(\mathbb{T}l)*}G^{(\mathbb{T}kl)}\mathbf{a}_l^{(\mathbb{T}k)}) = G_{ll}^{(\mathbb{T})}G_{kk}^{(\mathbb{T}l)}K_{kl}^{(\mathbb{T})} \tag{5.24}$$

valid for any $k \ne l$, $k, l \in \{1, \dots, N\}\backslash\mathbb{T}$, with

$$K_{kl}^{(\mathbb{T})} = (h_{kl} - \mathbf{a}_k^{(\mathbb{T}l)*}G^{(\mathbb{T}kl)}\mathbf{a}_l^{(\mathbb{T}k)}). \tag{5.25}$$

High moments of the diagonal entries $G_{ll}^{(\mathbb{T})}$ and $G_{kk}^{(\mathbb{T}l)}$ can be bounded with Lemma 3.9. High moments of $K_{kl}^{(\mathbb{T})}$, on the other hand, are controlled (and shown to be small) in the next lemma.

Lemma 5.3 *Assume* (1.9), *set* $z = E + i\eta$ *and let* $K_{kl}^{(\mathbb{T})}$ *be defined as in* (5.25) (*with* $\mathbb{T} = \emptyset$). *Then there exist constants* $c, c_0, C, M_1, M_2 > 0$ *such that*

$$\mathbb{E}|K_{kl}^{(\mathbb{T})}|^{2q} \le (Cq)^{cq}\mathcal{E}_q, \tag{5.26}$$

for all $E \in \mathbb{R}$, $N > M_1$, $\eta > 0$ *with* $N\eta > M_2$, $k \ne l \in \{1, \dots, N\}$, $q \in \mathbb{N}$ *with* $q \le c_0 N$.

Proof By the remark following Lemma 5.2 we can restrict our attention to the case $\mathbb{T} = \emptyset$.

By the definition of K_{kl} we get

$$\mathbb{E}|K_{kl}|^{2q} \le C^q\left(\mathbb{E}|h_{kl}|^{2q} + \mathbb{E}|\mathbf{a}_k^{(l)*}G^{(kl)}\mathbf{a}_l^{(k)}|^{2q}\right) \le C^q\left(\frac{(Cq)^q}{N^q} + \mathbb{E}|\mathbf{a}_k^{(l)*}G^{(kl)}\mathbf{a}_l^{(k)}|^{2q}\right).$$

Let $\mathbf{b}^{(k)} = G^{(kl)}\mathbf{a}_l^{(k)}$. Then

$$\mathbb{E}|\mathbf{a}_k^{(l)*}G^{(kl)}\mathbf{a}_l^{(k)}|^{2q}$$
$$= \mathbb{E}|(\mathbf{a}_k^{(l)}, \mathbf{b}^{(k)})|^{2q} \leq C^q \left(\mathbb{E}\left| |(\mathbf{a}_k^{(l)}, \mathbf{b}^{(k)})|^2 - \frac{\|\mathbf{b}^{(k)}\|^2}{N} \right|^q + \mathbb{E}\frac{\|\mathbf{b}^{(k)}\|^{2q}}{N^q} \right).$$

Noticing that $\mathbb{E}_k|(\mathbf{a}_k^{(l)}, \mathbf{b}^{(k)})|^2 = N^{-1}\|\mathbf{b}^{(k)}\|^2$, the Hanson–Wright large deviation estimate (see Proposition 9.1) implies that

$$\mathbb{E}\left| |(\mathbf{a}_k^{(l)}, \mathbf{b}^{(k)})|^2 - \frac{\|\mathbf{b}^{(k)}\|^2}{N} \right|^q \leq (Cq)^q\, \mathbb{E}\,\frac{\|\mathbf{b}^{(k)}\|^{2q}}{N^q}. \tag{5.27}$$

Hence

$$\mathbb{E}\,|\mathbf{a}_k^{(l)*}G^{(kl)}\mathbf{a}_l^{(k)}|^{2q} \leq (Cq)^q\, \mathbb{E}\frac{\|G^{(kl)}\mathbf{a}_l^{(k)}\|^{2q}}{N^q}. \tag{5.28}$$

Noticing that $\mathbb{E}_l\|G^{(kl)}\mathbf{a}_l^{(k)}\|^2 = N^{-1}\,\mathrm{Tr}\,|G^{(kl)}|^2$, applying again Proposition 9.1, we conclude that

$$\begin{aligned}\mathbb{E}\,\frac{\|G^{(kl)}\mathbf{a}_l^{(k)}\|^{2q}}{N^q} &\leq C^q\left((Cq)^q\,\mathbb{E}\left(\frac{\mathrm{Tr}\,|G^{(kl)}|^4}{N^4}\right)^{q/2} + \mathbb{E}\left(\frac{\mathrm{Tr}\,|G^{(kl)}|^2}{N^2}\right)^q \right) \\ &\leq (Cq)^q\left(\frac{(\mathrm{Im}\ m_{sc})^q + \mathbb{E}|\,\mathrm{Im}\ \Lambda^{(kl)}|^q}{(N\eta)^q} + \frac{(\mathrm{Im}\ m_{sc})^{\frac{q}{2}} + \mathbb{E}|\,\mathrm{Im}\ \Lambda^{(kl)}|^{\frac{q}{2}}}{(N\eta)^{\frac{3}{2}q}} \right),\end{aligned} \tag{5.29}$$

where we used the bound $\mathrm{Tr}\,|G^{(kl)}|^4 \leq \eta^{-2}\,\mathrm{Tr}\,|G^{(kl)}|^2$ and the estimate

$$\frac{1}{N^2}\,\mathrm{Tr}\ |G^{(kl)}|^2 = \frac{1}{N\eta}\,\mathrm{Im}\ m^{(kl)} \leq \frac{1}{(N\eta)^2} + \frac{\mathrm{Im}\, m_{sc} + |\,\mathrm{Im}\ \Lambda^{(k\ell)}|}{N\eta}$$

proven as in (3.6). Inserting (5.29) into (5.28) and then into (5.3), we find

$$\mathbb{E}|K_{kl}|^{2q} \leq (Cq)^{2q}\left(\frac{(\mathrm{Im}\ m_{sc})^q + \mathbb{E}|\,\mathrm{Im}\ \Lambda^{(kl)}|^q}{(N\eta)^q} + \frac{(\mathrm{Im}\ m_{sc})^{\frac{q}{2}} + \mathbb{E}|\,\mathrm{Im}\ \Lambda^{(kl)}|^{\frac{q}{2}}}{(N\eta)^{\frac{3}{2}q}} + \frac{1}{N^q} \right). \tag{5.30}$$

From the interlacing properties of the eigenvalues of H and of its minors, it is easy to check that $|m - m^{(kl)}| \leq C/(N\eta)$. This implies that

$$\begin{aligned}\mathbb{E}|\operatorname{Im}\,\Lambda^{(kl)}|^q &\le C^q\left(\mathbb{E}|\operatorname{Im}\,\Lambda|^q+\mathbb{E}|\operatorname{Im}\,m-\operatorname{Im}\,m^{(kl)}|^q\right)\\ &\le C^q\left(\mathbb{E}|\operatorname{Im}\,\Lambda|^q+\frac{1}{(N\eta)^q}\right),\end{aligned}$$

and hence, from (5.30), that

$$\mathbb{E}|K_{kl}|^{2q}\le(Cq)^{2q}\left(\frac{1}{(N\eta)^{2q}}+\frac{(\operatorname{Im}\,m_{sc})^q+\mathbb{E}|\operatorname{Im}\,\Lambda|^q}{(N\eta)^q}+\frac{(\operatorname{Im}\,m_{sc})^{\frac{q}{2}}+\mathbb{E}|\operatorname{Im}\,\Lambda|^{\frac{q}{2}}}{(N\eta)^{\frac{3}{2}q}}+\frac{1}{N^q}\right).$$

To conclude the proof, we observe that

$$\begin{aligned}\frac{(\operatorname{Im}\,m_{sc})^{\frac{q}{2}}+\mathbb{E}|\operatorname{Im}\,\Lambda|^{\frac{q}{2}}}{(N\eta)^{\frac{3}{2}q}}&\le\frac{1}{(N\eta)^q}\sqrt{\frac{(\operatorname{Im}\,m_{sc})^q+\mathbb{E}\,|\operatorname{Im}\,\Lambda|^q}{(N\eta)^q}}\\ &\le\frac{1}{(N\eta)^{2q}}+\frac{(\operatorname{Im}\,m_{sc})^q+\mathbb{E}|\operatorname{Im}\,\Lambda|^q}{(N\eta)^q}.\end{aligned}$$

□

5.4 Proof of Eq. (5.3)

For a fixed $\mathbf{k}$ we use (5.16) and (5.17) to expand $V(\mathbf{k})$ in Eq. (5.11) as

$$\begin{aligned}V(\mathbf{k})=\sum_{\substack{\sigma_1\in\mathcal{L}_1\\ \rho_1\in\mathcal{M}_1}}\cdots\sum_{\substack{\sigma_{2q}\in\mathcal{L}_{2q}\\ \rho_{2q}\in\mathcal{M}_{2q}}}&\mathbb{E}((\mathbb{I}-\mathbb{E}_{k_1})((\mathbb{I}-\mathbb{E}_{k_1})A^1_{\sigma_1})B^1_{\rho_1})\\ &\cdots\overline{((\mathbb{I}-\mathbb{E}_{k_{2q}})((\mathbb{I}-\mathbb{E}_{k_{2q}})A^{2q}_{\sigma_{2q}})B^{2q}_{\rho_{2q}})}.\end{aligned}$$

We claim that

$$\begin{aligned}&\left|\mathbb{E}((\mathbb{I}-\mathbb{E}_{k_1})((\mathbb{I}-\mathbb{E}_{k_1})A^1_{\sigma_1})B^1_{\rho_1})\cdots\overline{((\mathbb{I}-\mathbb{E}_{k_{2q}})((\mathbb{I}-\mathbb{E}_{k_{2q}})A^{2q}_{\sigma_{2q}})B^{2q}_{\rho_{2q}})}\right|\\ &\quad\le(Cq)^{cq^2}\mathcal{E}^{\frac{1}{2}}_{2q+|L(\mathbf{k})|}\end{aligned}\tag{5.31}$$

for any $\mathbf{k}$, $\sigma_1\in\mathcal{L}_1$, $\rho_1\in\mathcal{M}_1,\ldots,\sigma_{2q}\in\mathcal{L}_{2q}$, $\rho_{2q}\in\mathcal{M}_{2q}$ and for all $q\le c_0(N\eta)^{1/8}$ and $(N\eta)$ and N large enough. Here $|L(\mathbf{k})|$ is the number of lone labels associated with the vector $\mathbf{k}$. Using the bounds (5.20) on the cardinality of the sets $\mathcal{L}_j$, $\mathcal{M}_j$, (5.31) implies that

$$|V(\mathbf{k})|\le(Cq)^{cq^2}\mathcal{E}^{\frac{1}{2}}_{2q+|L(\mathbf{k})|}.$$

From Eq. (5.10), we get

$$\mathbb{E}\left|\frac{1}{N}\sum_k(\mathbb{I}-\mathbb{E}_k)W_k\right|^{2q} \leq \frac{(Cq)^{cq^2}}{N^{2q}}\sum_{P\in\mathfrak{P}_{2q}}\sum_{\mathbf{k}}\mathbf{1}(\mathcal{P}(\mathbf{k})=P)\,\mathcal{E}_{2q+|L(\mathbf{k})|}^{\frac{1}{2}}$$
$$\leq (Cq)^{cq^2}\sum_{P\in\mathfrak{P}_{2q}}\frac{1}{N^{2q-|P|}}\,\mathcal{E}_{2q+|L(P)|}^{\frac{1}{2}}. \tag{5.32}$$

Here we wrote $|L(P)|$ to indicate the number of lone labels associated with any vector $\mathbf{k}$ such that $\mathcal{P}(\mathbf{k}) = P$ (it is clear that the number of lone labels only depends on the partition P), and we used the fact that

$$|\{\mathbf{k} : \mathcal{P}(\mathbf{k}) = P\}| \leq N^{|P|}$$

where $|P|$ denotes the size of the partition P (the number of sets in which $\{1, \ldots, 2q\}$ is divided). Since, by definition, $\mathcal{E}_q \geq N^{-q}$, we have

$$\frac{1}{N^{2q-|P|}}\,\mathcal{E}_{2q+|L|}^{\frac{1}{2}} \leq \mathcal{E}_{2q-|P|}\,\mathcal{E}_{2q+|L|}^{\frac{1}{2}} \leq C\mathcal{E}_{4q}^{\frac{2q-|P|}{4q}}\,\mathcal{E}_{4q}^{\frac{1}{2}\left(\frac{2q+|L|}{4q}\right)} = C\mathcal{E}_{4q}^{\frac{1}{2}+\frac{q+\frac{|L|}{2}-|P|}{4q}} \leq C\mathcal{E}_{4q}^{\frac{1}{2}} \tag{5.33}$$

where we used (4.1) and the bound

$$|P| \leq |L| + \frac{2q-|L|}{2} \quad\Rightarrow\quad q + \frac{|L|}{2} - |P| \geq 0.$$

Inserting (5.33) into (5.32), and using the fact that the number of partitions of $2q$ elements is bounded by $(Cq)^q$, we obtain that

$$\mathbb{E}\left|\frac{1}{N}\sum_k(\mathbb{I}-\mathbb{E}_k)W_k\right|^{2q} \leq (Cq)^{cq^2}\mathcal{E}_{4q}^{\frac{1}{2}}$$

which concludes the proof of (5.3).

Next, we prove Eq. (5.31), for fixed $\mathbf{k}$, $\sigma_1, \rho_1, \ldots, \sigma_{2q}, \rho_{2q}$. We distinguish two cases.

Case 1 For any $r \in \{1, \ldots, 2q\}$ all resolvent entries in $A^r_{\sigma_r}$ and $B^r_{\rho_r}$ are maximally expanded.

The l.h.s. of (5.31) vanishes if there exists a lone label r such that, for all $s \in \{1, \ldots, 2q\}\backslash\{r\}$, the monomials $A^s_{\sigma_s}$ and $B^s_{\rho_s}$ do not depend on r, i.e. if there exists a lone label r such that k_r appears as an upper index in all resolvent entries in $A^s_{\sigma_s}$, $B^s_{\rho_s}$, for all $s \in \{1, \ldots, 2q\}\backslash\{r\}$.

For $r \in L(\mathbf{k})$, we say, following [10], that $s \in \{1, \ldots, 2q\}\backslash\{r\}$ is a *partner label* of r if k_r appears as lower index in one of the resolvent entries of $A^s_{\sigma_s}$ or of $B^s_{\rho_s}$. Since, by the definition of the expansion algorithm, the same index cannot appear both as a lower and as an upper index in the same resolvent entry, we conclude that non-zero terms on the l.h.s. of Eq. (5.31) are characterized by the fact that for every lone label $r \in L(\mathbf{k})$ there exists at least one partner label s.

We denote by $\ell(s)$ the number of lone labels having s as a partner. In order for the l.h.s. of (5.31) not to vanish, we must have (as remarked in [10])

$$\sum_{s=1}^{2q} \ell(s) \geq |L(\mathbf{k})|.$$

This follows because all resolvent entries are maximally expanded and therefore, for every lone label r, and every $s \in \{1, 2, \ldots, 2q\}\backslash\{r\}$, k_r either appears as lower label in at least one resolvent entry in $A^s_{\sigma_s}$ or $B^s_{\rho_s}$ (in which case, s has r as a partner), or every resolvent entry in $A^s_{\sigma_s}$ or $B^s_{\rho_s}$ has r as an upper index. This is the point where the assumption that all entries are maximally expanded is used.

The combined number of lower indices different from k_s in $A^s_{\sigma_s}$ and $B^s_{\rho_s}$ is at least $\ell(s)$. Since the operation w_1 produces only one additional lower index, while w_0 does not produce new lower labels, we find that

$$b(\sigma_s) + b(\rho_s) \geq \ell(s).$$

(Recall that $b(\sigma)$ denotes the number of ones occurring in the string σ). We notice, moreover, that for any $\mathbb{T} \subseteq \{1, \ldots, 2q\}$ the operation w_1, applied on a diagonal entry like $G^{(\mathbb{T})}_{kk}$ or $1/G^{(\mathbb{T})}_{kk}$, adds a new lower index, associated with a lone label from $L(\mathbf{k})$. Since any subsequent operation will not remove the new lower index, $\ell(s) = 0$ if and only if the strings σ_s and ρ_s contain only zeros. Since we assumed all entries to be maximally expanded, we have that

$$\ell(s) = 0 \quad \text{if and only if} \quad A^s_{\sigma_s} = \frac{1}{G^{(L(\mathbf{k})\backslash\{s\})}_{ss}} \text{ and } B^s_{\rho_s} = G^{(L(\mathbf{k})\backslash\{s\})}_{ss}. \tag{5.34}$$

Let us assume that in the term on the l.h.s. of (5.31) there are t labels $s_1, \ldots, s_t$ with $\ell(s_j) = 0$. Then there will be $u = 2q - t$ labels $r_1, \ldots, r_u$ with $\ell(r_j) \geq 1$. Without loss of generality we can assume the first t labels in (5.31) to be $s_1, \ldots, s_t$ and the last u labels to be $r_1, \ldots, r_u$ (the fact that some terms are complex conjugated is irrelevant for our argument). For $s = 1, \ldots, t$ (so that $\ell(s) = 0$), we observe that

$$\begin{aligned}
&|(\mathbb{I} - \mathbb{E}_s)((\mathbb{I} - \mathbb{E}_s)A^s_{\sigma_s})B^s_{\rho_s}| \\
&\quad \leq |((\mathbb{I} - \mathbb{E}_s)A^s_{\sigma_s})B^s_{\rho_s}| + \mathbb{E}_s|(\mathbb{I} - \mathbb{E}_s)A^s_{\sigma_s})B^s_{\rho_s}| \\
&\quad \leq |(\mathbb{I} - \mathbb{E}_s)A^s_{\sigma_s}||B^s_{\rho_s}| + (\mathbb{E}_s|(\mathbb{I} - \mathbb{E}_s)A^s_{\sigma_s}|^2)^{1/2}(\mathbb{E}_s|B^s_{\rho_s}|^2)^{1/2} \\
&\quad \leq C m_s \widetilde{m}_s
\end{aligned} \tag{5.35}$$

where we defined the random variables

$$\begin{aligned}
m_s &= |(\mathbb{I} - \mathbb{E}_s)A^s_{\sigma_s}| + (\mathbb{E}_s|(\mathbb{I} - \mathbb{E}_s)A^s_{\sigma_s}|^2)^{\frac{1}{2}}, \\
\widetilde{m}_s &= |B^s_{\rho_s}| + (\mathbb{E}_s|B^s_{\rho_s}|^2)^{\frac{1}{2}}
\end{aligned} \tag{5.36}$$

for all $s = 1, \ldots, t$. We will estimate high moments of $m_s, \widetilde{m}_s$ using the bounds established in Sect. 5.3. Notice that, while $\widetilde{m}_s$ is only bounded, the quantity m_s is expected to be small. It is important that the estimates that we will use reflect this smallness.

To bound the terms on the l.h.s. of (5.31) with $s = t+1, \ldots, 2q$ we proceed as follows. Since here $\ell(s) \geq 1$, either $A^s_{\sigma_s}$ or $B^s_{\rho_s}$ must contain at least one off-diagonal entry. Using (5.24), we can express the off-diagonal entries in terms of the variable $K^{(\mathbb{T})}_{kl}$, defined in Eq. (5.25). We collect all the factors $K^{(\mathbb{T})}_{kl}$ appearing in $A^s_{\sigma_s}$ (one for every off-diagonal resolvent entry) in a single monomial O_s and all the diagonal entries (appearing in $A^s_{\sigma_s}$, either in the numerator or in the denominator) in a monomial P_s; accordingly $A^s_{\sigma_s} = O_s P_s$. We proceed similarly for $B^s_{\rho_s}$ and we obtain $B^s_{\rho_s} = \widetilde{O}_s \widetilde{P}_s$. We estimate

$$\begin{aligned}
&|(\mathbb{I} - \mathbb{E}_s)((\mathbb{I} - \mathbb{E}_s) A^s_{\sigma_s}) B^s_{\rho_s}| \\
&\quad = \left|(\mathbb{I} - \mathbb{E}_s)((\mathbb{I} - \mathbb{E}_s) O_s P_s) \widetilde{O}_s \widetilde{P}_s\right| \\
&\quad \leq |O_s||\widetilde{O}_s||P_s||\widetilde{P}_s| + |\widetilde{O}_s \widetilde{P}_s \, \mathbb{E}_s(O_s P_s)| + |\mathbb{E}_s(O_s P_s \widetilde{O}_s \widetilde{P}_s)| + |\mathbb{E}_s(O_s P_s)\, \mathbb{E}_s(\widetilde{O}_s \widetilde{P}_s)| \\
&\quad \leq |O_s||\widetilde{O}_s||P_s||\widetilde{P}_s| + |\widetilde{O}_s||\widetilde{P}_s|(\mathbb{E}_s|O_s|^2)^{1/2}(\mathbb{E}_s|P_s|^2)^{1/2} \\
&\qquad + (\mathbb{E}_s|O_s\widetilde{O}_s|^2)^{1/2}(\mathbb{E}_s|P_s\widetilde{P}_s|^2)^{1/2} + (\mathbb{E}_s|O_s|^2)^{1/2}(\mathbb{E}_s|P_s|^2)^{1/2}(\mathbb{E}_s|\widetilde{O}_s|^2)^{1/2}(\mathbb{E}_s|\widetilde{P}_s|^2)^{1/2} \\
&\quad \leq C M_s \widetilde{M}_s
\end{aligned} \tag{5.37}$$

for all $s = t+1, \ldots, 2q$. Here we defined the random variables

$$\begin{aligned}
M_s &= |O_s||\widetilde{O}_s| + (\mathbb{E}_s|O_s|^2)^{\frac{1}{2}}|\widetilde{O}_s| + (\mathbb{E}_s|O_s\widetilde{O}_s|^2)^{\frac{1}{2}} + (\mathbb{E}_s|O_s|^2)^{\frac{1}{2}}(\mathbb{E}_s|\widetilde{O}_s|^2)^{\frac{1}{2}} \\
\widetilde{M}_s &= |P_s||\widetilde{P}_s| + (\mathbb{E}_s|P_s|^2)^{\frac{1}{2}}|\widetilde{P}_s| + (\mathbb{E}_s|P_s\widetilde{P}_s|^2)^{\frac{1}{2}} + (\mathbb{E}_s|P_s|^2)^{\frac{1}{2}}(\mathbb{E}_s|\widetilde{P}_s|^2)^{\frac{1}{2}}.
\end{aligned}$$

Combining (5.35) with (5.37), and applying Cauchy–Schwarz's inequality, we find

$$\begin{aligned}
&\left|\mathbb{E}((\mathbb{I} - \mathbb{E}_{k_1})((\mathbb{I} - \mathbb{E}_{k_1})A^1_{\sigma_1})B^1_{\rho_1}) \cdots \overline{((\mathbb{I} - \mathbb{E}_{k_{2q}})((\mathbb{I} - \mathbb{E}_{k_{2q}})A^{2q}_{\sigma_{2q}})B^{2q}_{\rho_{2q}})}\right| \\
&\quad \leq C\,(\mathbb{E}\, m_1^2 \cdots m_t^2 M_{t+1}^2 \cdots M_{2q}^2)^{\frac{1}{2}}\,(\mathbb{E}\, \widetilde{m}_1^2 \cdots \widetilde{m}_t^2 \widetilde{M}_{t+1}^2 \cdots \widetilde{M}_{2q}^2)^{\frac{1}{2}}.
\end{aligned} \tag{5.38}$$

We bound the first factor on the r.h.s. of the last equation. Recall that O_s and $\widetilde{O}_s$ are the product of respectively $\gamma_s := \gamma(A^s_{\sigma_s})$ and $\widetilde{\gamma}_s := \gamma(B^s_{\sigma_s})$ entries of the form $K^{(\mathbb{T})}_{kl}$. With Hölder's inequality we obtain

$$\begin{aligned}
\mathbb{E}\, m_1^2 \cdots m_t^2 M_{t+1}^2 \cdots M_{2q}^2 \leq & \prod_{s=1}^{t} \left[\mathbb{E}\, m_s^{2(t+\sum_s(\gamma_s+\widetilde{\gamma}_s))}\right]^{\frac{1}{t+\sum_s(\gamma_s+\widetilde{\gamma}_s)}} \\
& \times \prod_{s=t+1}^{2q} \left[\mathbb{E}\, M_s^{2\frac{t+\sum_s(\gamma_s+\widetilde{\gamma}_s)}{\gamma_s+\widetilde{\gamma}_s}}\right]^{\frac{\gamma_s+\widetilde{\gamma}_s}{t+\sum_s(\gamma_s+\widetilde{\gamma}_s)}}.
\end{aligned} \tag{5.39}$$

According to (5.34) we have, for $s = 1, \ldots, t$,

$$m_s^2 \le C \left| (\mathbb{I} - \mathbb{E}_s) \frac{1}{G_{k_s k_s}^{(\mathbb{T}_s)}} \right|^2 + C\, \mathbb{E}_s \left| (\mathbb{I} - \mathbb{E}_s) \frac{1}{G_{k_s k_s}^{(\mathbb{T}_s)}} \right|^2 \tag{5.40}$$

where we set $\mathbb{T}_s = L(\mathbf{k}) \backslash \{s\}$. This implies that

$$\mathbb{E}\, m_s^{2(t+\sum_s (\gamma_s + \widetilde{\gamma}_s))} \le C^{t+\sum_s (\gamma_s+\widetilde{\gamma}_s)}\, \mathbb{E} \left| (\mathbb{I} - \mathbb{E}_s) \frac{1}{G_{k_s k_s}^{(\mathbb{T}_s)}} \right|^{2(t+\sum_s (\gamma_s+\widetilde{\gamma}_s))} . \tag{5.41}$$

To bound the M_s variables, we observe that, for any $s = t+1, \ldots, 2q$,

$$M_s^2 \le C(|O_s|^2 |\widetilde{O}_s|^2 + (\mathbb{E}_s |O_s|^2)|\widetilde{O}_s|^2 + (\mathbb{E}_s |O_s \widetilde{O}_s|^2) + (\mathbb{E}_s |O_s|^2)(\mathbb{E}_s |\widetilde{O}_s|^2)). \tag{5.42}$$

Inserting the contribution of

$$|O_s|^2 |\widetilde{O}_s|^2 = \prod_{j=1}^{\gamma_s + \widetilde{\gamma}_s} |K_{i_j l_j}^{(\mathbb{T}_j)}|^2$$

(for appropriate indices i_j, l_j and sets $\mathbb{T}_j$) in the expectation on the r.h.s. of (5.39), we find, again by Hölder's inequality,

$$\left[\mathbb{E}\, (|O_s||\widetilde{O}_s|)^{2 \frac{t+\sum_s (\gamma_s+\widetilde{\gamma}_s)}{\gamma_s+\widetilde{\gamma}_s}} \right]^{\frac{\gamma_s+\widetilde{\gamma}_s}{t+\sum_s (\gamma_s+\widetilde{\gamma}_s)}} \le \prod_{j=1}^{\gamma_s+\widetilde{\gamma}_s} \left[\mathbb{E}\, |K_{i_j l_j}^{(\mathbb{T}_j)}|^{2(t+\sum_s (\gamma_s+\widetilde{\gamma}_s))} \right]^{\frac{1}{t+\sum_s (\gamma_s+\widetilde{\gamma}_s)}} .$$

The contribution of the second term on the r.h.s. of (5.42), having the form

$$\begin{aligned} (\mathbb{E}_s |O_s|^2)|\widetilde{O}_s|^2 &= \left(\mathbb{E}_s \prod_{j=1}^{\gamma_s} |K_{i_j l_j}^{(\mathbb{T}_j)}|^2 \right) \prod_{j=\gamma_s+1}^{\gamma_s+\widetilde{\gamma}_s} |K_{i_j l_j}^{(\mathbb{T}_j)}|^2 \\ &\le \prod_{j=1}^{\gamma_s} \left(\mathbb{E}_s |K_{i_j l_j}^{(\mathbb{T}_j)}|^{2\gamma_s} \right)^{\frac{1}{\gamma_s}} \prod_{j=\gamma_s+1}^{\gamma_s+\widetilde{\gamma}_s} |K_{i_j l_j}^{(\mathbb{T}_j)}|^2, \end{aligned}$$

to (5.39) can be estimated by

$$\begin{aligned}
&\mathbb{E}\Big[(\mathbb{E}_s|O_s|^2)^{\frac{t+\sum_s(\gamma_s+\widetilde{\gamma}_s)}{\gamma_s+\widetilde{\gamma}_s}}|\widetilde{O}_s|^{2\frac{t+\sum_s(\gamma_s+\widetilde{\gamma}_s)}{\gamma_s+\widetilde{\gamma}_s}}\Big]\\
&\le \mathbb{E}\prod_{j=1}^{\gamma_s}\Big(\mathbb{E}_s|K_{i_jl_j}^{(\mathbb{T}_j)}|^{2\gamma_s}\Big)^{\frac{t+\sum_s(\gamma_s+\widetilde{\gamma}_s)}{\gamma_s(\gamma_s+\widetilde{\gamma}_s)}}\prod_{j=\gamma_s+1}^{\gamma_s+\widetilde{\gamma}_s}|K_{i_jl_j}^{(\mathbb{T}_j)}|^{2\frac{t+\sum_s(\gamma_s+\widetilde{\gamma}_s)}{\gamma_s+\widetilde{\gamma}_s}}\\
&\le \prod_{j=1}^{\gamma_s}\Big[\mathbb{E}(\mathbb{E}_s|K_{i_jl_j}^{(\mathbb{T}_j)}|^{2\gamma_s})^{\frac{t+\sum_s(\gamma_s+\widetilde{\gamma}_s)}{\gamma_s}}\Big]^{\frac{1}{\gamma_s+\widetilde{\gamma}_s}}\prod_{j=\gamma_s+1}^{\gamma_s+\widetilde{\gamma}_s}\Big[\mathbb{E}\,|K_{i_jl_j}^{(\mathbb{T}_j)}|^{2(t+\sum_s(\gamma_s+\widetilde{\gamma}_s))}\Big]^{\frac{1}{\gamma_s+\widetilde{\gamma}_s}}\\
&\le \prod_{j=1}^{\gamma_s+\widetilde{\gamma}_s}\Big[\mathbb{E}\,|K_{i_jl_j}^{(\mathbb{T}_j)}|^{2(t+\sum_s(\gamma_s+\widetilde{\gamma}_s))}\Big]^{\frac{1}{\gamma_s+\widetilde{\gamma}_s}}.
\end{aligned}$$

The contributions from the last two terms on the r.h.s. (5.42) can also be bounded similarly. We obtain

$$\Big[\mathbb{E}\,M_s^{2\frac{t+\sum_s(\gamma_s+\widetilde{\gamma}_s)}{\gamma_s+\widetilde{\gamma}_s}}\Big]^{\frac{\gamma_s+\widetilde{\gamma}_s}{t+\sum_s(\gamma_s+\widetilde{\gamma}_s)}}\le C\prod_{j=1}^{\gamma_s+\widetilde{\gamma}_s}\Big[\mathbb{E}\,|K_{i_jl_j}^{(\mathbb{T}_j)}|^{2(t+\sum_s(\gamma_s+\widetilde{\gamma}_s))}\Big]^{\frac{1}{t+\sum_s(\gamma_s+\widetilde{\gamma}_s)}}.$$

Inserting (5.41) and the last equation into (5.39), we conclude that

$$\begin{aligned}
\mathbb{E}\,m_1^2\cdots m_t^2M_1^2\dots M_u^2 &\le \prod_{s=1}^{t}\left[\mathbb{E}\left|(\mathbb{I}-\mathbb{E}_s)\frac{1}{G_{k_sk_s}^{(\mathbb{T}_s)}}\right|^{2(t+\sum_s(\gamma_s+\widetilde{\gamma}_s))}\right]^{\frac{1}{t+\sum_s(\gamma_s+\widetilde{\gamma}_s)}}\\
&\quad\times\prod_{s=t+1}^{2q}\prod_{j=1}^{\gamma_s+\widetilde{\gamma}_s}\Big[\mathbb{E}\,|K_{i_{j,s}l_{j,s}}^{(\mathbb{T}_{j,s})}|^{2(t+\sum_s(\gamma_s+\widetilde{\gamma}_s))}\Big]^{\frac{1}{t+\sum_s(\gamma_s+\widetilde{\gamma}_s)}}\\
&\le (Cq)^{cq^2}\mathcal{E}_{t+\sum_{s=t+1}^{2q}(\gamma_s+\widetilde{\gamma}_s)}\\
&\le (Cq)^{cq^2}\mathcal{E}_{2q+|L(\mathbf{k})|}
\end{aligned}\tag{5.43}$$

where we used Lemma 5.3, Eq. (3.4), the bound $t+\sum_{s=t+1}^{2q}(\gamma_s+\tilde{\gamma}_s)\le cq^2$ (which follows from $t,\tilde{\gamma}_s,\gamma_s\le 2q$), and the estimate

$$t+\sum_{s=t+1}^{2q}(\gamma_s+\tilde{\gamma}_s)\ge 2q+\sum_{s=1}^{u}\ell(s)\ge 2q+|L(\mathbf{k})|,$$

which follows from Remark (1) in Sect. 5.2.2.

To bound the second factor on the r.h.s. of (5.38), we observe that all resolvent entries in $\widetilde{m}_s$ and $\widetilde{M}_s$ are diagonal and that their total number is at most cq^2. Applying Lemma 3.9 and Lemma 5.2 (using the assumption $q^2\le c(N\eta)^{1/4}$), we find

$$\mathbb{E}\,\widetilde{m}_1^2 \cdots \widetilde{m}_t^2 \widetilde{M}_{t+1}^2 \cdots \widetilde{M}_{2q}^2 \leq (Cq)^{cq^2}. \tag{5.44}$$

Inserting in (5.38), we obtain

$$\left|\mathbb{E}((\mathbb{I}-\mathbb{E}_{k_1})((\mathbb{I}-\mathbb{E}_{k_1})A^1_{\sigma_1})B^1_{\rho_1})\cdots\overline{((\mathbb{I}-\mathbb{E}_{k_{2q}})((\mathbb{I}-\mathbb{E}_{k_{2q}})A^{2q}_{\sigma_{2q}})B^{2q}_{\rho_{2q}})}\right| \leq (Cq)^{cq^2}\mathcal{E}^{1/2}_{2q+|L(\mathbf{k})|}$$

which concludes the proof of (5.31) in Case 1.

Case 2 For some $r \in \{1,\ldots,2q\}$ there is at least one resolvent entry, either in $A^r_{\sigma_r}$ or in $B^r_{\rho_r}$, which is not maximally expanded. Among the other $2q-1$ indices in $\{1,\ldots,2q\}\backslash\{r\}$ let us say that for t of them (denoted by $s_1,\ldots,s_t$), the strings σ_{s_j} and ρ_{s_j} only contain zeros, while for the other $u = 2q-1-t$ labels, $s_{t+1},\ldots,s_{2q-1}$, either the string σ_{s_j} or the string ρ_{s_j} contain at least a one. Without loss of generality, we can assume that $s_1 = 1,\ldots,s_t = t, s_{t+1} = t+1,\ldots,s_{2q-1} = 2q-1$ and $r = 2q$.

For $s = 1,\ldots,t$, we use the estimate

$$|(\mathbb{I}-\mathbb{E}_s)((\mathbb{I}-\mathbb{E}_s)A^s_{\sigma_s})B^s_{\rho_s}| \leq C m_s \widetilde{m}_s$$

established in (5.35) with m_s and $\widetilde{m}_s$ defined as in (5.36). Since, for $s = 1,\ldots,t$, the strings σ_s and ρ_s only contain zeros, we have

$$A^s_{\sigma_s} = \frac{1}{G^{(\mathbb{T}_s)}_{ss}}, \quad \text{and} \quad B^s_{\rho_s} = G^{(\mathbb{T}'_s)}_{ss}$$

for appropriate sets $\mathbb{T}_s$ and $\mathbb{T}'_s$. This implies that [like in (5.40)]

$$m_s^2 \leq C\left|(\mathbb{I}-\mathbb{E}_s)\frac{1}{G^{(\mathbb{T}_s)}_{k_s k_s}}\right|^2 + C\,\mathbb{E}_s\left|(\mathbb{I}-\mathbb{E}_s)\frac{1}{G^{(\mathbb{T}_s)}_{k_s k_s}}\right|^2. \tag{5.45}$$

For $s = t+1,\ldots,2q-1$, either $A^s_{\sigma_s}$ or $B^s_{\rho_s}$ contains at least two off-diagonal resolvent entries. Let us assume, for example, that $A^s_{\sigma_s}$ contains two off-diagonal entries (it is easy to invert the roles of $A^s_{\sigma_s}$ and $B^s_{\rho_s}$, if it is $B^s_{\rho_s}$ which contains the off-diagonal entries). In this case, we write $A^s_{\sigma_s} = O_s P_s$, where the monomial O_s is given by the product of $\gamma_s = \gamma(A^s_{\sigma_s}) \geq 2$ terms of the form $K^{(\mathbb{T}_j)}_{i_j l_j}$ while P_s is a product of diagonal entries (appearing either in the numerator or in the denominator). We bound (for any $s = t+1,\ldots,2q-1$)

$$\left|(\mathbb{I}-\mathbb{E}_s)((\mathbb{I}-\mathbb{E}_s)A^s_{\sigma_s})B^s_{\rho_s}\right| \leq C L_s \widetilde{L}_s$$

where we defined

$$\begin{aligned} L_s &= |O_s| + (\mathbb{E}_s|O_s|^2)^{\frac{1}{2}} \\ \widetilde{L}_s &= (|P_s| + (\mathbb{E}_s|P_s|^4)^{\frac{1}{4}})(|B^s_{\rho_s}| + (\mathbb{E}_s|B^s_{\rho_s}|^4)^{\frac{1}{4}}). \end{aligned}$$

We observe that

$$L_s^2 \leq |O_s|^2 + \mathbb{E}_s|O_s|^2 = \prod_{j=1}^{\gamma_s} |K_{i_{j,s}l_{j,s}}^{(\mathbb{T}_{j,s})}|^2 + \prod_{j=1}^{\gamma_s} \left[\mathbb{E}_s |K_{i_{j,s}l_{j,s}}^{(\mathbb{T}_{j,s})}|^{2\gamma_s}\right]^{\frac{1}{\gamma_s}} \tag{5.46}$$

for appropriate indices $i_{j,s}$, $l_{j,s}$ and appropriate sets $\mathbb{T}_{j,s}$, and for all $s = t+1, \ldots, 2q-1$.

Finally, we consider $s = 2q$, for which one of the resolvent entries, either in $A^{2q}_{\sigma_{2q}}$ or in $B^{2q}_{\rho_{2q}}$ is not maximally expanded. Let us assume that the entry which is not maximally expanded is in $A^{2q}_{\sigma_{2q}}$. In this case, we write

$$A^{2q}_{\sigma_{2q}} = OP,$$

where O is a monomial given by the product of $\gamma \equiv \gamma(A^{2q}_{\sigma_{2q}})$ terms of the form $K_{kl}^{(\mathbb{T})}$ and P contains only diagonal entries. Note that, since we assumed $A^{2q}_{\sigma_{2q}}$ not to be maximally expanded, we must have $2q+1 \leq \gamma \leq 2q+2$. We estimate

$$\left|(\mathbb{I}-\mathbb{E}_1)((\mathbb{I}-\mathbb{E}_1)OP)B^{2q}_{\rho_{2q}}\right| \leq CL\widetilde{L}$$

where we set

$$\begin{aligned} L &= |O| + (\mathbb{E}_{2q}|O|^2)^{\frac{1}{2}} \\ \widetilde{L} &= (|P| + (\mathbb{E}_{2q}|P|^4)^{\frac{1}{4}})(|B^1_{\rho_1}| + (\mathbb{E}_{2q}|B^1_{\rho_1}|^4)^{\frac{1}{4}}). \end{aligned}$$

Similarly to (5.46), we find

$$L^2 \leq \prod_{j=1}^{\gamma} |K_{i_jl_j}^{(\mathbb{T}_j)}|^2 + \prod_{j=1}^{\gamma} \left[\mathbb{E}_s |K_{i_jl_j}^{(\mathbb{T}_j)}|^{2\gamma_s}\right]^{\frac{1}{\gamma_s}}. \tag{5.47}$$

By Cauchy–Schwarz, we can bound the l.h.s. of Eq. (5.31) by

$$\begin{aligned} &\left|\mathbb{E}((\mathbb{I}-\mathbb{E}_{k_1})((\mathbb{I}-\mathbb{E}_{k_1})A^1_{\sigma_1})B^1_{\rho_1})\cdots\overline{((\mathbb{I}-\mathbb{E}_{k_{2q}})((\mathbb{I}-\mathbb{E}_{k_{2q}})A^{2q}_{\sigma_{2q}})B^{2q}_{\rho_{2q}})}\right| \\ &\quad \leq C(\mathbb{E}\, m_1^2\cdots m_t^2 L_{t+1}^2\ldots L_{2q-1}^2 L^2)^{\frac{1}{2}}\,(\mathbb{E}\,\widetilde{m}_1^2\cdots\widetilde{m}_t^2\widetilde{L}_{t+1}^2\cdots\widetilde{L}_{2q-1}^2\widetilde{L}^2)^{\frac{1}{2}}. \end{aligned} \tag{5.48}$$

Proceeding as in (5.43), and using the bounds (5.45), (5.46), (5.47), we find

$$\begin{aligned} &\mathbb{E}\, m_1^2\cdots m_t^2 L_{t+1}^2\cdots L_{2q-1}^2 L^2 \\ &\leq (Cq)^{2q}\prod_{s=1}^{t}\left[\mathbb{E}\left|(\mathbb{I}-\mathbb{E}_1)\frac{1}{G_{k_sk_s}^{(\mathbb{T}_s)}}\right|^{2(t+\gamma+\sum_s\gamma_s))}\right]^{\frac{1}{2(t+\gamma+\sum_s\gamma_s))}} \end{aligned}$$

$$
\begin{aligned}
&\times \prod_{s=t+1}^{2q-1} \prod_{j=1}^{\gamma_s} \left[\mathbb{E}\,|K^{(\mathbb{T}_{j,s})}_{i_{j,s},l_{j,s}}|^{2(t+\gamma+\sum_s \gamma_s)}\right]^{\frac{1}{2(t+\gamma+\sum_s \gamma_s)}} \prod_{j=1}^{\gamma} \left[\mathbb{E}\,|K^{(\mathbb{T})}_{i_j l_j}|^{2(t+\gamma+\sum_s \gamma_s)}\right]^{\frac{1}{2(t+\gamma+\sum_s \gamma_s)}} \\
&\leq (Cq)^{cq^2} \mathcal{E}_{t+\gamma+\sum_s \gamma_s} \\
&\leq (Cq)^{cq^2}\, \mathcal{E}_{4q}.
\end{aligned} \tag{5.49}
$$

Here we used the fact that $t+\gamma+\sum_s \gamma_s \leq cq^2$ and the bound

$$
t+\gamma+\sum_s \gamma_s \geq t+2q+1+u \geq 4q,
$$

because $\gamma \geq 2q+1$ and $\gamma_s \geq 1$ for all $s = t+1, \ldots, 2q-1$. On the other hand we have that, similarly to (5.44),

$$
\mathbb{E}\, \widetilde{m}_1^2 \ldots \widetilde{m}_t^2\, \widetilde{L}_{t+1}^2 \cdots \widetilde{L}_{2q-1}^2 \widetilde{L}^2 \leq (Cq)^{cq^2}.
$$

Inserting (5.49) and the last inequality into (5.48), we obtain

$$
\left|\mathbb{E}((\mathbb{I}-\mathbb{E}_{k_1})((\mathbb{I}-\mathbb{E}_{k_1})A^1_{\sigma_1})B^1_{\rho_1})\cdots \overline{((\mathbb{I}-\mathbb{E}_{k_{2q}})((\mathbb{I}-\mathbb{E}_{k_{2q}})A^{2q}_{\sigma_{2q}})B^{2q}_{\rho_{2q}})}\right| \leq (Cq)^{cq^2} \mathcal{E}_{4q}^{1/2}
$$

Equation (5.31) follows because $2q + |L(\mathbf{k})| \leq 4q$ [and because of (4.1)].

6 Improved optimal bound in the bulk

In this section, we prove Theorem 2, which gives a stronger bound on the fluctuations of the Stieltjes transform in the bulk of the spectrum. Compared with Theorem 2, here there is no upper bound on the size of the power q. Looking at the proof of Theorem 1, we observe that there are two results, in which the restriction $q \leq c(N\eta)^{1/8}$ played an important role, namely the proof of Lemma 3.4, containing a bound for $\mathbb{E}|G_{11}|^q$, and of Lemma 5.1, giving an estimate for $\mathbb{E}|(\mathbb{E}_1 G_{11}^{-1})^{-1}|^q$. In Lemma 6.4 and Lemma 6.3 below, we show that, in the bulk of the spectrum, at distances of order one from the edges, these moments can be bounded for arbitrary $q \in \mathbb{N}$. Using these estimates, Theorem 2 can then be shown exactly as we proved Theorem 1.

In order to bound the moments of $|G_{11}|$ and $|(\mathbb{E}_1 G_{11}^{-1})^{-1}|$ in the bulk of the spectrum, we will need, first of all, an upper bound on the density of states, as stated in the next proposition.

Proposition 6.1 *Assume (1.9), fix $\tilde{\eta} > 0$ and $0 < \kappa < 2$ and set $z = E + i\eta$. Then there exist constants $c, C, M, K_0 > 0$ such that*

$$
\mathbb{P}\left(\frac{\mathcal{N}[E-\frac{\eta}{2}, E+\frac{\eta}{2}]}{N\eta} \geq K\right) \leq Ce^{-c\sqrt{KN\eta}}
$$

for all $E \in (-2+\kappa, 2-\kappa)$, $N > M$, $K > K_0$, $\eta > 1/N$. Here $\mathcal{N}[a; b]$ denotes the number of eigenvalues in the interval $[a; b]$.

Proof We set $I = [E - \frac{\eta}{2}, E + \frac{\eta}{2}]$ and denote by $\mathcal{N}_I$ the number of eigenvalues in the interval I. For $N\eta \geq (\log N)^4$, the claim follows from Theorem 4.6 in [17]. So, we can assume that $1 \leq N\eta \leq (\log N)^4$. We distinguish two cases.

- If $K \leq CN^\beta/(N\eta)$, for a $0 < \beta < 1/16$, we use Theorem 5.1 of [17], with $\nu = 1/4$, to estimate

$$\mathbb{P}\left(\frac{\mathcal{N}_I}{N\eta} \geq K\right) \leq \left(\frac{C}{K}\right)^{KN\eta/4} \leq e^{-cKN\eta}$$

  which holds true for any interval $I = [E - \eta/2, E + \eta/2]$ with $1 \leq N\eta \leq CN^\beta$ (and in particular holds for all $1 \leq N\eta \leq (\log N)^4$).
- If $K \geq CN^\beta/(N\eta)$ we use the bound in Theorem 4.6 of [17]. We consider a new interval $\widetilde{I} = [E - \widetilde{\eta}/2, E + \widetilde{\eta}/2]$ with $(\log N)^4 \leq N\widetilde{\eta} \leq 2(\log N)^4$. Since obviously $\mathcal{N}_{\widetilde{I}} \geq \mathcal{N}_I$, we find

$$\mathbb{P}\left(\frac{\mathcal{N}_I}{N\eta} \geq K\right) = \mathbb{P}\left(\mathcal{N}_I \geq KN\eta\right) \leq \mathbb{P}\left(\mathcal{N}_{\widetilde{I}} \geq KN\eta\right) = \mathbb{P}\left(\frac{\mathcal{N}_{\widetilde{I}}}{N\widetilde{\eta}} \geq \frac{K\eta}{\widetilde{\eta}}\right) \leq Ce^{-c\sqrt{KN\eta}}$$

  where we used Theorem 4.6 of [17], and the fact that $K\eta/\widetilde{\eta} \geq CN^\beta/(\log N)^4$ is large. □

Secondly, we will need an upper bound on the size of the gap between eigenvalues; more precisely, we need to know that, if we consider an interval larger than Kp/N the probability of finding fewer than p eigenvalues tends to zero, as $K \to \infty$.

Proposition 6.2 *Assume* (1.9), *fix* $\tilde{\eta} > 0$ *and* $0 < \kappa < 2$ *and set* $z = E + i\eta$. *Then there exist constants* $c, C, M, K_0 > 0$ *such that*

$$\mathbb{P}\left(\left|\left\{\alpha : |\lambda_\alpha - E| \leq \frac{Kp}{N}\right\}\right| \leq p\right) \leq Ce^{-cK^{1/4}}$$

for any $E \in [-2 + \kappa, 2 - \kappa]$, $p \geq 1$, $N > M$, $K > K_0$.

Proof Set $\eta = p\sqrt{K}/N$ and $z = E + i\eta$. We consider the intervals

$$I_0 = [E - \sqrt{K}\eta, E + \sqrt{K}\eta] \quad \text{and}$$
$$I_j = [E - 2^j\sqrt{K}\eta, E - 2^{j-1}\sqrt{K}\eta] \cup [E + 2^{j-1}\sqrt{K}\eta, E + 2^j\sqrt{K}\eta] \quad \text{for all } j \geq 1.$$

We denote by $\mathcal{N}_j$ the number of eigenvalues in I_j. For a large constant M, we define the event

$$\Omega := \left\{\max_j \frac{\mathcal{N}_j}{N|I_j|} \geq M\right\}.$$

From the upper bound in Proposition 6.1, we find

$$\mathbb{P}(\Omega) \leq \sum_{j=0}^{\infty} Ce^{-c\sqrt{MN|I_j|}} \leq \sum_{j=0}^{\infty} Ce^{-c\sqrt{MKp\,2^j}} \leq Ce^{-cK^{1/2}}. \qquad (6.1)$$

On Ω^c, and assuming that $|\{\alpha : |\lambda_\alpha - E| \leq Kp/N\}| \leq p$, we find

$$\begin{aligned} \text{Im } m(E+i\eta) &= \frac{1}{N}\sum_{\alpha=1}^{N} \frac{\eta}{(\lambda_\alpha - E)^2 + \eta^2} \\ &= \frac{1}{N}\sum_{j=0}^{\infty}\sum_{\alpha:\,\lambda_\alpha \in I_j} \frac{\eta}{(\lambda_\alpha - E)^2 + \eta^2} \\ &\leq \frac{p}{N\eta} + \frac{C}{\sqrt{K}}\sum_{j=1}^{\infty}\frac{1}{2^j} \leq \frac{C}{\sqrt{K}}. \end{aligned} \tag{6.2}$$

On the other hand, since we are away from the edges, we have $c_0 = \text{Im } m_{sc}(z) > 0$. Hence, for K large enough, (6.2) implies that the fluctuations of $m(z)$ from the Stieltjes transform of the semicircle law are larger than, say, $c_0/2$. Using Theorem 3.1 of [17], we know that the probability for such large fluctuations is small. More precisely, we find

$$\begin{aligned} &\mathbb{P}\left(\Omega^c \text{ and } \left|\left\{\alpha : |\lambda_\alpha - E| \leq \frac{Kp}{N}\right\}\right| \leq p\right) \leq \mathbb{P}\left(\text{Im } m(E+i\eta) \leq \frac{C}{\sqrt{K}}\right) \\ &\leq \mathbb{P}\left(|m(z) - m_{sc}(z)| \geq \frac{c_0}{2}\right) \\ &\leq Ce^{-c\sqrt{N\eta}} \leq Ce^{-cK^{1/4}}. \end{aligned}$$

Together with (6.1), this concludes the proof of the proposition. □

We are now ready to prove upper bounds for arbitrary moments of diagonal resolvent entries. The next lemma gives an improvement, in the bulk, of Lemma 3.4.

Lemma 6.3 *Assume* (1.9), *fix* $\tilde{\eta} > 0$ *and* $0 < \kappa < 2$ *and set* $z = E + i\eta$. *Then there exist constants* $c, C, M > 0$ *such that*

$$\mathbb{E}|G_{11}|^q \leq (Cq)^{cq}$$

for all $E \in [-2+\kappa, 2-\kappa]$, $\eta > 1/N$, $N > M$, $q \in \mathbb{N}$.

Proof We fix $K \geq 1$. From (2.4), estimating $|1/w| \leq 1/|\text{Im } w|$, we find

$$|G_{11}| \leq \left[\frac{\eta}{N}\sum_{\alpha=1}^{N-1}\frac{\xi_\alpha}{(\lambda_\alpha^{(1)} - E)^2 + \eta^2}\right]^{-1} \leq \left[\frac{1}{NK^2\eta}\sum_{\alpha:\,|\lambda_\alpha^{(1)} - E| \leq K\eta} \xi_\alpha\right]^{-1}$$

where $\xi_\alpha = N|\mathbf{a}_1 \cdot \mathbf{u}_\alpha^{(1)}|^2$ and $\lambda_\alpha^{(1)}, \mathbf{u}_\alpha^{(1)}$ are the eigenvalues and the eigenvectors of the minor obtained by removing the first row and column from the original Wigner matrix. We find

$$\mathbb{P}\left(|G_{11}| \geq K^3\right) \leq \mathbb{P}\left(|\{\alpha : |\lambda_\alpha^{(1)} - E| \leq K\eta\}| \leq \sqrt{K}N\eta\right) + \mathbb{P}\left(\frac{1}{\sqrt{K}N\eta}\sum_{j=1}^{\sqrt{K}N\eta} \xi_{\alpha_j} \leq \frac{1}{K^{3/2}}\right). \tag{6.3}$$

From Proposition 6.2, with $p = \sqrt{K}N\eta$, we obtain

$$\mathbb{P}\left(|\{\alpha : |\lambda_\alpha^{(1)} - E| \leq K\eta\}| \leq \sqrt{K}N\eta\right) \leq Ce^{-cK^{1/8}} \tag{6.4}$$

for all $K > K_0$ large enough. On the other hand, the large deviation estimates in Proposition 9.1 imply (see for example Lemma 4.7 in [17]) that

$$\mathbb{P}\left(\frac{1}{\sqrt{K}N\eta}\sum_{j=1}^{\sqrt{K}N\eta} \xi_{\alpha_j} \leq \frac{1}{K^{3/2}}\right) \leq e^{-c\sqrt{\sqrt{K}N\eta}} \leq e^{-cK^{1/4}}. \tag{6.5}$$

Inserting (6.5) and (6.4) into (6.3), we obtain that there exists a constant t_0 such that

$$\mathbb{P}\left(|G_{11}| \geq t\right) \leq Ce^{-ct^{1/24}}$$

for all $t \geq t_0$. This implies that

$$\mathbb{E}|G_{11}|^q = \int_0^\infty \mathbb{P}(|G_{11}| \geq t^{1/q})dt \leq t_0 + C\int_{t_0}^\infty e^{-ct^{1/(24q)}}dt \leq (Cq)^{cq}.$$

□

Finally, we establish also bounds for arbitrary moments of $|(\mathbb{E}_1 G_{11}^{-1})^{-1}|$ which improve, in the bulk of the spectrum, the results of Lemma 5.1.

Lemma 6.4 *Assume* (1.9), *fix* $\tilde{\eta} > 0$ *and* $0 < \kappa < 2$ *and set* $z = E + i\eta$. *Then there exist constants* $c, C, M > 0$ *such that*

$$\mathbb{E}\left|\frac{1}{\mathbb{E}_1 \frac{1}{G_{11}}}\right|^q \leq (Cq)^{cq} \tag{6.6}$$

for all $E \in [-2+\kappa, 2-\kappa]$, $\eta > 1/N$, $N > M$, $q \in \mathbb{N}$.

Proof Let $K \geq 1$. Using (2.4) and estimating $|1/w| \leq 1/|\operatorname{Im} w|$, we find

$$\left|\frac{1}{\mathbb{E}_1 \frac{1}{G_{11}}}\right| \leq \left[\frac{\eta}{N}\sum_{\alpha=1}^{N-1}\frac{1}{(\lambda_\alpha^{(1)} - E)^2 + \eta^2}\right]^{-1} \leq \left[\frac{|\{\alpha : |\lambda_\alpha^{(1)} - E| \leq K\eta\}|}{NK^2\eta}\right]^{-1}$$

where $\lambda_\alpha^{(1)}$ are the eigenvalues of the minor obtained by removing the first row and column from the original Wigner matrix. Using Proposition 6.2 with $p = N\eta$, we conclude that, for all $K > K_0$ large enough,

$$\mathbb{P}\left(\left|\frac{1}{\mathbb{E}_1 \frac{1}{G_{11}}}\right| \geq K^2\right) \leq \mathbb{P}\left(|\{\alpha : |\lambda_\alpha^{(1)} - E| \leq K\eta\}| \leq N\eta\right) \leq Ce^{-cK^{\frac{1}{4}}},$$

which clearly implies the claim (6.6). □

7 Convergence of the eigenvalue counting function

In this section we apply Theorem 1 to prove the convergence of the counting function of the eigenvalues, as stated in Theorem 3. We adapt here the approach developed in [10,20,22].

Proof of Theorem 3 We are going to show that there exist constants $C, c > 0$ such that

$$\mathbb{E}\,|(n(E_2) - n(E_1)) - (n_{sc}(E_2) - n_{sc}(E_1))|^q \leq (Cq)^{cq^2} \frac{(\log N)^q}{N^q} \tag{7.1}$$

for all $-2 \leq E_1 < E_2 \leq 2$, $N > 1$ and for all $q \in \mathbb{N}$. This gives

$$\mathbb{P}\left(|(n(E_2) - n(E_1)) - (n_{sc}(E_2) - n_{sc}(E_1))| \geq \frac{K \log N}{N}\right) \leq \frac{(Cq)^{cq^2}}{K^q} \tag{7.2}$$

for all $-2 \leq E_1 < E_2 \leq 2$, $K > 0$ and $q \in \mathbb{N}$. In particular,

$$\mathbb{P}\left(|(n(2) - n(-2)) - 1| \geq \frac{K \log N}{N}\right) \leq \frac{(Cq)^{cq^2}}{K^q}$$

for all $K > 0$ and $q \in \mathbb{N}$. This implies that, with high probability, there are at most $K \log N$ eigenvalues outside the interval $[-2; 2]$; this observation allows us to conclude the proof of Theorem 3. Indeed, for $E \leq 2$ we have $n_{sc}(E) = 0$ and therefore

$$\begin{aligned}\mathbb{P}\left(|n(E) - n_{sc}(E)| \geq \frac{K \log N}{N}\right) &= \mathbb{P}\left(n(E) \geq \frac{K \log N}{N}\right) \\ &\leq \mathbb{P}\left(|n(2) - n(-2) - 1| \geq \frac{K \log N}{N}\right) \\ &\leq \frac{(Cq)^{cq^2}}{K^q}\end{aligned} \tag{7.3}$$

because, for $E \leq -2$, $n(E) + n(2) - n(-2) \leq 1$. For $E \geq 2$, we have $n_{sc}(E) = 1$ and hence

$$\mathbb{P}\left(|n(E) - n_{sc}(E)| \geq \frac{K \log N}{N}\right) = \mathbb{P}\left(1 - n(E) \geq \frac{K \log N}{N}\right) \leq \mathbb{P}\left(|n(2) - n(-2) - 1| \geq \frac{K \log N}{N}\right) \leq \frac{(Cq)^{cq^2}}{K^q}$$

because, for $E \geq 2$, $1 - n(E) + n(2) - n(-2) \leq 1$. Finally, for $-2 < E < 2$, we have $|n(E) - n_{sc}(E)| \leq |n(E) - n(-2) - n_{sc}(E)| + n(-2)$. Since $n_{sc}(-2) = 0$, we find

$$\begin{aligned} &\mathbb{P}\left(|n(E) - n_{sc}(E)| \geq \frac{K \log N}{N}\right) \\ &\leq \mathbb{P}\left(|(n(E) - n(-2)) - (n_{sc}(E) - n_{sc}(-2))| \geq \frac{K \log N}{2N}\right) \\ &\quad + \mathbb{P}\left(n(-2) \geq \frac{K \log N}{2N}\right) \\ &\leq \frac{(Cq)^{cq^2}}{K^q} \end{aligned}$$

from (7.2) and (7.3).

To conclude the proof of Theorem 3, we show (7.1), following [10]. We define the empirical eigenvalue distribution

$$\rho(\lambda) = \frac{1}{N} \sum_{\alpha=1}^{N} \delta(\lambda - \lambda_\alpha)$$

and we denote by $\mathbf{1}_{[E_1, E_2]}$ the characteristic function of the interval $[E_1, E_2]$. With this notation we have

$$(n(E_2) - n(E_1)) - (n_{sc}(E_2) - n_{sc}(E_1)) = \int_{\mathbb{R}} \mathbf{1}_{[E_1, E_2]}(\lambda) \, (\rho(\lambda) - \rho_{sc}(\lambda)) \, d\lambda.$$

Next, we approximate $\mathbf{1}_{[E_1, E_2]}$ with a smooth function f. For a given $q \in \mathbb{N}$, we choose $M = Cq^8$ for a sufficiently large constant $C > 0$ and we set $\tilde{\eta} = M/N$ (we choose the constant $C > 0$ large enough and consider $N \geq q^8$ large enough, so that we can apply Theorem 1 to bound $\Lambda(E + i\tilde{\eta})$; if $N < q^8$, (7.1) is trivial since, by definition, $0 \leq n(E) \leq 1$ for all $E \in \mathbb{R}$). We choose $f \in C_0^\infty(\mathbb{R})$ such that $\text{supp } f \subset (E_1, E_2)$ and $f = 1$ in $[E_1 + \tilde{\eta}, E_2 - \tilde{\eta}]$. Notice that, since we are considering $-2 \leq E_1 < E_2 \leq 2$, we can apply part (i) of Theorem 1 to bound $|\Lambda(E + i\tilde{\eta})|$ for all $E \in \text{supp } f$. We also assume that $|f'| \leq C/\tilde{\eta}$ and $|f''| \leq C/\tilde{\eta}^2$.

From the boundedness of ρ_{sc}, we find

$$\left|\int_{\mathbb{R}} (\mathbf{1}_{[E_1,E_2]}(\lambda) - f(\lambda))\rho_{sc}(\lambda)d\lambda\right|^q \leq C^q \tilde{\eta}^q = \frac{C^q M^q}{N^q}.$$

Moreover, recalling the bound

$$|n(E+\eta) - n(E-\eta)| \leq C\eta \,\mathrm{Im}\, m(E+i\eta) \leq C\eta\,(1 + |\,\mathrm{Im}\,\Lambda(E+i\eta)|)$$

we find

$$\left|\int_{\mathbb{R}} (\mathbf{1}_{[E_1,E_2]}(\lambda) - f(\lambda))\rho(\lambda)d\lambda\right| \leq C \sum_{j=1,2} \left|n(E_j+\tilde{\eta}) - n(E_j-\tilde{\eta})\right| \leq C\tilde{\eta}(1+|\,\mathrm{Im}\,\Lambda(E+i\tilde{\eta})|)$$

where, as usual, $\Lambda(z) = m(z) - m_{sc}(z)$. Taking the q-th moment, we obtain

$$\mathbb{E}\left|\int_{\mathbb{R}} (\mathbf{1}_{[E_1,E_2]}(\lambda) - f(\lambda))\rho(\lambda)d\lambda\right|^q \leq C^q\tilde{\eta}^q(1+\mathbb{E}|\,\mathrm{Im}\,\Lambda(E+i\tilde{\eta})|^q) \leq \frac{C^q M^q}{N^q}$$

where we applied the bound $\mathbb{E}|\,\mathrm{Im}\,\Lambda(E+i\tilde{\eta})|^q \leq 1$ from Lemma 3.5 (using that $q \leq c(N\tilde{\eta})^{1/8}$). We conclude that

$$\begin{aligned}
&\mathbb{E}|(n(E_2) - n(E_1)) - (n_{sc}(E_2) - n_{sc}(E_1))|^q \\
&\quad \leq \frac{C^q}{N^q} + C^q\mathbb{E}\left|\int_{\mathbb{R}} f(\lambda)(\rho(\lambda) - \rho_{sc}(\lambda))d\lambda\right|^q. \qquad (7.4)
\end{aligned}$$

Next, we use Helffer–Sjöstrand functional calculus. We introduce the notation $I = [E_1, E_2]$ and $\tilde{I} = [E_1, E_1+\tilde{\eta}] \cup [E_2-\tilde{\eta}, E_2]$. Notice that $|\tilde{I}| = 2\tilde{\eta}$. For a $\chi \in C_0^\infty(\mathbb{R})$ with $\chi(y) = 1$ on $[-1, 1]$ and $\chi(y) = 0$ on $[-2, 2]^c$, we can define the almost analytic extension of f on $\mathbb{C}$ given by

$$\tilde{f}(x+iy) = (f(x) + iyf'(x))\chi(y).$$

Then, we have, see Eq. (B.12) in [13],

$$\begin{aligned}
f(\lambda) &= \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{\partial_{\bar{z}}\tilde{f}(x+iy)}{\lambda - x - iy} dx\,dy \\
&= \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{iyf''(x)\chi(y) + i(f(x)+iyf'(x))\chi'(y)}{\lambda - x - iy} dx\,dy.
\end{aligned}$$

Since f is real, we find

$$\begin{aligned}
&\int_{\mathbb{R}} f(\lambda)(\rho(\lambda) - \rho_{sc}(\lambda))d\lambda \\
&\quad = \frac{1}{2\pi}\,\mathrm{Re}\int_{\mathbb{R}^2} [iyf''(x)\chi(y) + i(f(x)+iyf'(x))\chi'(y)]\Lambda(x+iy)dx\,dy.
\end{aligned}$$

Taking the q-th moment, and recalling that f and χ are compactly supported, we find, similarly to Eq. (7.33) in [10],

$$\begin{aligned}
&\mathbb{E}\Big|\int_{\mathbb{R}} f(\lambda)(\rho(\lambda)-\rho_{sc}(\lambda))d\lambda\Big|^q \\
&\quad\le C^q\,\mathbb{E}\Bigg\{\Bigg|\int_I dx\int_1^2 dy f(x)\chi'(y)\Lambda(x+iy)\Bigg|^q \\
&\qquad+\Bigg|\int_{\tilde I} dx\int_1^2 dy f'(x)y\chi'(y)\Lambda(x+iy)\Bigg|^q \\
&\qquad+\Bigg|\int_{\tilde I} dx\int_0^{\tilde\eta} dy f''(x)y\chi(y)\,\mathrm{Im}\;\Lambda(x+iy)\Bigg|^q \\
&\qquad+\Bigg|\int_{\tilde I} dx\int_{\tilde\eta}^2 dy f''(x)y\chi(y)\,\mathrm{Im}\;\Lambda(x+iy)\Bigg|^q\Bigg\}. \qquad (7.5)
\end{aligned}$$

Recalling that $|f|\le 1$ and $\chi'\le C$, the first term on the right hand side of (7.5) is bounded by

$$\mathbb{E}\Bigg|\int_I dx\int_1^2 dy f(x)\chi'(y)\Lambda(x+iy)\Bigg|^q \le C^q\int_I dx\int_1^2 dy\,\mathbb{E}\,|\Lambda(x+iy)|^q\,.$$

From Theorem 1, part (i), we find

$$\mathbb{E}\Bigg|\int_I dx\int_1^2 dy f(x)\chi'(y)\Lambda(x+iy)\Bigg|^q \le \frac{(Cq)^{cq^2}}{N^q}. \qquad (7.6)$$

The second term on the r.h.s. of Eq. (7.5) can be bounded similarly. In this case $|f'|\le C/\tilde\eta$ is large but it is compensated by the volume factor $|\tilde I|\le 2\tilde\eta$. We find, again by part (i) of Theorem 1, that

$$\begin{aligned}
&\mathbb{E}\Bigg|\int_{\tilde I} dx\int_1^2 dy f'(x)y\chi'(y)\Lambda(x+iy)\Bigg|^q \\
&\quad\le C^q\tilde\eta^{-1}\int_{\tilde I} dx\int_1^2 dy\,\mathbb{E}\,|\Lambda(x+iy)|^q \le \frac{(Cq)^{cq^2}}{N^q}. \qquad (7.7)
\end{aligned}$$

The third term on the r.h.s. of Eq. (7.5) can be estimated by

$$\begin{aligned}
&\mathbb{E}\Bigg|\int_{\tilde I} dx\int_0^{\tilde\eta} dy f''(x)\,y\,\chi(y)\,\mathrm{Im}\;\Lambda(x+iy)\Bigg|^q \\
&\quad\le \frac{C^q}{\tilde\eta^2}\int_{\tilde I} dx\int_0^{\tilde\eta} dy\,y^q\,\mathbb{E}\,|\mathrm{Im}\;\Lambda(x+iy)|^q \qquad (7.8)
\end{aligned}$$

because of the assumption $|f''| \leq C/\tilde{\eta}^2$. Here we cannot apply directly Theorem 1, because we are arbitrarily close to the real axis. We follow instead an argument of Theorem 2.2 in [22]. For any $x \in \mathbb{R}$, we note that the functions $y \to y \operatorname{Im} m(x+iy)$ and $y \to y \operatorname{Im} m_{sc}(x+iy)$ are monotone increasing on $\mathbb{R}_+$ (they are Stieltjes transforms of positive measures). This implies that

$$y|\operatorname{Im} \Lambda(x+iy)| \leq \tilde{\eta}(\operatorname{Im} m(x+i\tilde{\eta}) + \operatorname{Im} m_{sc}(x+i\tilde{\eta})) \leq \tilde{\eta}\,(2+|\operatorname{Im} \Lambda(x+i\tilde{\eta})|).$$

From (7.8), we get therefore

$$\begin{aligned}
&\mathbb{E}\Big|\int_{\tilde{I}} dx \int_0^{\tilde{\eta}} dy f''(x) y\,\chi(y) \operatorname{Im} \Lambda(x+iy)\Big|^q \\
&\quad \leq \frac{C^q}{\tilde{\eta}^2}\int_{\tilde{I}} dx \int_0^{\tilde{\eta}} dy\, \tilde{\eta}^q\,(1+\mathbb{E}|\operatorname{Im} \Lambda(x+i\tilde{\eta})|^q) \leq C^q \tilde{\eta}^q \leq \frac{C^q M^q}{N^q} \qquad (7.9)
\end{aligned}$$

because, by Lemma 3.5, $\mathbb{E}|\operatorname{Im} \Lambda(x+i\tilde{\eta})|^q \leq 1$ (since $\tilde{\eta}$ was chosen so that $q \leq c(N\tilde{\eta})^{1/8}$).

To estimate the fourth term on the r.h.s. of Eq. (7.5), we integrate by parts, first in x and then in y. We obtain

$$\begin{aligned}
&\mathbb{E}\left|\int_{\tilde{I}} dx \int_{\tilde{\eta}}^2 dy f''(x) y\chi(y) \operatorname{Im} \Lambda(x+iy)\right|^q \\
&\leq \mathbb{E}\left|\int_{\tilde{I}} dx \int_{\tilde{\eta}}^2 dy f''(x) y\chi(y)\Lambda(x+iy)\right|^q \\
&\leq C^q\,\mathbb{E}\left\{\left|\int_{\tilde{I}} dx f'(x)\tilde{\eta}\Lambda(x+i\tilde{\eta})\right|^q + \left|\int_{\tilde{I}} dx \int_{\tilde{\eta}}^2 dy f'(x)\chi(y)\Lambda(x+iy)\right|^q \right. \\
&\qquad \left. + \left|\int_{\tilde{I}} dx \int_1^2 dy f'(x) y\chi'(y)\Lambda(x+iy)\right|^q\right\} \qquad (7.10)
\end{aligned}$$

where we used the observation that $\chi(\tilde{\eta}) = 1$ and $\chi'(x) = 0$ for $x \notin [1;2]$ in the first and, respectively, in the third term. By part (i) of Theorem 1 (and using that $|f'(x)| \leq C/\tilde{\eta}$) the first term on the r.h.s. of (7.10) can be bounded by

$$\mathbb{E}\left|\int_{\tilde{I}} dx f'(x)\tilde{\eta}\Lambda(x+i\tilde{\eta})\right|^q \leq \frac{C^q}{\tilde{\eta}}\int_{\tilde{I}} dx\, \tilde{\eta}^q\,\mathbb{E}\,|\Lambda(x+i\tilde{\eta})|^q \leq \frac{(Cq)^{cq^2}}{N^q}. \qquad (7.11)$$

The third term on the r.h.s. of (7.10), on the other hand, can be controlled by

$$\begin{aligned}
&\mathbb{E}\left|\int_{\tilde{I}} dx \int_1^2 dy f'(x) y\chi'(y)\Lambda(x+iy)\right|^q \leq \frac{C^q}{\tilde{\eta}}\int_{\tilde{I}} dx \int_1^2 dy\,\mathbb{E}\,|\Lambda(x+iy)|^q \\
&\quad \leq \frac{(Cq)^{cq^2}}{N^q}. \qquad (7.12)
\end{aligned}$$

As for the second term on the r.h.s. of Eq. (7.10), we find that

$$\begin{aligned}
&\mathbb{E}\left|\int_{\tilde{I}} dx \int_{\tilde{\eta}}^{2} dy f'(x)\chi(y)\Lambda(x+iy)\right|^q \\
&\leq \int_{\tilde{I}} dx_1 \cdots \int_{\tilde{I}} dx_q \int_{\tilde{\eta}}^{2} dy_1 \cdots \int_{\tilde{\eta}}^{2} dy_q \prod_{j=1}^{q} |f'(x_j)||\chi(y_j)| \,\mathbb{E} \prod_{j=1}^{q} |\Lambda(x_j+iy_j)| \\
&\leq \prod_{j=1}^{q} \int_{\tilde{I}} dx_j \int_{\tilde{\eta}}^{2} dy_j\, |f'(x_j)||\chi(y_j)| \left(\mathbb{E}|\Lambda(x_j+iy_j)|^q\right)^{1/q}
\end{aligned}$$

applying Hölder's inequality to the expectation. With the usual bounds on $|f'|$ and $|\chi|$, we obtain

$$\begin{aligned}
&\mathbb{E}\left|\int_{\tilde{I}} dx \int_{\tilde{\eta}}^{2} dy f'(x)\chi(y)\Lambda(x+iy)\right|^q \\
&\leq \frac{(Cq)^{cq^2}}{\tilde{\eta}^q}\left[\int_{\tilde{I}} dx \int_{\tilde{\eta}}^{2} dy\, \frac{1}{Ny}\right]^q \leq (Cq)^{cq^2}\frac{(\log\tilde{\eta})^q}{N^q} \leq (Cq)^{cq^2}\frac{(\log N)^q}{N^q}.
\end{aligned} \tag{7.13}$$

Inserting (7.11), (7.12) and (7.13) in (7.10) we find

$$\mathbb{E}\left|\int_{\tilde{I}} dx \int_{\tilde{\eta}}^{2} dy f''(x) y\chi(y) \,\mathrm{Im}\;\Lambda(x+iy)\right|^q \leq (Cq)^{cq^2}\frac{(\log N)^q}{N^q}.$$

Inserting last equation, together with (7.6), (7.7) and (7.9), into (7.5), we get

$$\mathbb{E}\left|\int_{\mathbb{R}} f(x)(\rho(x)-\rho_{sc}(x))dx\right|^q \leq (Cq)^{cq^2}\frac{(\log N)^q}{N^q}.$$

Combined with (7.4), this implies (7.1). □

8 Rigidity of the semicircle law

The goal of this section is to prove Theorem 4, giving precise bounds on the fluctuation of eigenvalues of a Wigner matrix with respect to the locations predicted by the semicircle law. We will need, here, bounds on the location of the largest and smallest eigenvalues. As a first rough bound, we will use Lemma 7.2 in [21], which states that, for any $x > 3$ and for an appropriate $\varepsilon > 0$,

$$\mathbb{E}\, n(-x) \leq e^{-N^{\varepsilon}\log x}, \quad \text{and} \quad \mathbb{E}\, n(x) \geq 1 - e^{-N^{\varepsilon}\log x}. \tag{8.1}$$

To obtain better estimates for the extremal eigenvalues, we make use of the following bounds for the Stieltjes transform of the semicircle law. For $E \in \mathbb{R}$, we set $\kappa =$

$||E|-2|$. For any fixed $E_0 > 2$ and $\eta_0 > 0$, there exist constants $c, C > 0$ such that (see Eq. (4.3) in [10])

$$c\sqrt{\kappa+\eta} \leq |1 - m_{sc}^2(E+i\eta)| \leq C\sqrt{\kappa+\eta} \quad \forall\, |E| \leq E_0,\ 0 < \eta \leq \eta_0$$

and

$$c\frac{\eta}{\sqrt{\kappa+\eta}} \leq \text{Im}\ m_{sc}(E+i\eta) \leq C\frac{\eta}{\sqrt{\kappa+\eta}} \quad \forall\, 2 \leq |E| \leq E_0,\ 0 < \eta \leq \eta_0. \tag{8.2}$$

From Theorem 1 part (ii), we have

$$\mathbb{E}|\text{Im}\ \Lambda|^{2q} \leq \frac{(Cq)^{cq^2}}{(N\eta)^{2q}},$$

for all $q \leq c(N\eta)^{1/8}$. Inserting this estimate and the bound (8.2) into the r.h.s. of (4.5), we conclude that, for $2 \leq |E| \leq E_0$ and $\eta \leq \eta_0$ with $N\eta \geq M$ large enough, we have

$$\begin{aligned}
\mathbb{E}|R|^q &\leq (Cq)^{cq^2}\left(\max\left\{\frac{1}{(N\eta)^{4q}}, \frac{(\text{Im}\ m_{sc})^{2q} + \mathbb{E}|\,\text{Im}\ \Lambda|^{2q}}{(N\eta)^{2q}}\right\} + \frac{1}{N^{2q}}\right)^{\frac{1}{2}} \\
&\leq (Cq)^{cq^2}\left(\frac{1}{(N\eta)^{4q}} + \frac{1}{(N\sqrt{\kappa+\eta})^{2q}} + \frac{1}{N^{2q}}\right)^{\frac{1}{2}} \\
&\leq (Cq)^{cq^2}\left(\frac{1}{(N\eta)^{2q}} + \frac{1}{(N\sqrt{\kappa+\eta})^{q}}\right)
\end{aligned}$$

for all $q \leq c(N\eta)^{1/8}$. Plugging back into (2.15), we obtain a stronger bound for the imaginary part of Λ, given by

$$\mathbb{E}|\,\text{Im}\ \Lambda|^q \leq (Cq)^{cq^2}\left(\frac{1}{(N(\kappa+\eta))^q} + \frac{1}{((N\eta)^2\sqrt{\kappa+\eta})^q}\right), \tag{8.3}$$

and valid for $2 \leq |E| \leq E_0$, $\eta \leq \eta_0$ with $N\eta \geq M$ large enough, $1 \leq q \leq c(N\eta)^{1/8}$ (a similar bound is given in Eq. (2.20) in [10]).

The next lemma (which is similar to Theorem 7.3 in [10]) estimates the probability of having eigenvalues outside the interval $[-2, 2]$, making use of the bound (8.3).

Lemma 8.1 *Assume* (1.9). *Then there exist* $C, c, \varepsilon > 0$ *such that*

$$\mathbb{P}\left(\max_\alpha |\lambda_\alpha| \geq 2 + \frac{K}{N^{2/3}}\right) \leq \frac{(Cq)^{cq^2}}{K^q} \tag{8.4}$$

for all $K > 0$ *and* $q \in \mathbb{N}$ *with* $q \leq N^\varepsilon$.

Proof We prove that

$$\mathbb{P}\left(\lambda_1 \leq -2 - \frac{K}{N^{2/3}}\right) \leq \frac{(Cq)^{cq^2}}{K^q},$$

where λ_1 is the smallest eigenvalue of H. The proof of the analogous bound for the largest eigenvalue λ_N is similar and we omit it.

We can assume that K is large enough, since otherwise (8.4) is trivial. We can also assume that $K \leq N$, since otherwise we can apply (8.1) to estimate

$$\mathbb{P}\left(\lambda_1 \leq -2 - \frac{K}{N^{2/3}}\right) \leq \mathbb{P}\left(n\left(-\frac{K}{N^{2/3}}\right) \geq \frac{1}{N}\right) \leq N\,\mathbb{E}\,n\left(-\frac{K}{N^{2/3}}\right) \leq e^{-N^{\varepsilon}\log K} \leq \frac{1}{K^q}$$

for all $q \leq N^{\varepsilon}$.

We fix $E_0 > 3$. From (8.1) we find that, for any small $\varepsilon < 0$, $K \leq N$, $q \leq N^{\varepsilon}$

$$\mathbb{P}(\lambda_1 \leq -E_0) \leq \mathbb{P}\left(n(-E_0) \geq \frac{1}{N}\right) \leq N\mathbb{E}n(-E_0) \leq e^{-N^{2\varepsilon}\log E_0} \leq e^{-N^{\varepsilon}\log K} \leq \frac{1}{K^q}. \tag{8.5}$$

We still have to bound the probability that $-E_0 < \lambda_1 \leq -2 - K/N^{2/3}$. We define

$$\kappa_j = \frac{(K+j)}{N^{2/3}}; \quad \eta_j = \frac{(K+j)^{\frac{1}{8}}}{N^{2/3}}. \tag{8.6}$$

Moreover we define the intervals $I_j = \left[-2 - \kappa_{j+1}, -2 - \kappa_j\right]$ for $j = 0, \ldots, j_{max}$, where j_{max} is the smallest integer with $2 + \kappa_{j+1} \geq E_0$ (clearly $j_{max} \leq E_0 N^{2/3}$). We denote by $x_j = -2 - \kappa_j$ the endpoints of the intervals I_j.

We observe that

$$\mathbb{P}\left(-E_0 \leq \lambda_1 \leq -2 - \frac{K}{N^{\frac{2}{3}}}\right) \leq \sum_{j=0}^{j_{max}} \mathbb{P}\left(\lambda_1 \in I_j\right). \tag{8.7}$$

We note that $|I_j| = N^{-2/3}$, so that if $\lambda_1 \in I_j$ one has $|\lambda_1 - x_j| \leq |I_j| \leq \eta_j$. Hence, setting $z_j = x_j + i\eta_j$, the event $\lambda_1 \in I_j$ implies that

$$\text{Im } m(z_j) = \frac{1}{N}\sum_{\alpha} \frac{\eta}{(\lambda_\alpha - x_j)^2 + \eta_j^2} \geq \frac{1}{2N\eta_j}.$$

On the other hand, since $2 \leq |x_j| \leq 2E_0$, the bound (8.2) gives

$$\text{Im } m_{sc}(z_j) \leq \frac{C\eta_j}{\sqrt{\kappa_j}}.$$

Since, by (8.6),

$$\frac{C\eta_j}{\sqrt{\kappa_j}} \le \frac{1}{4N\eta_j}$$

for $K > K_0$ large enough (depending only on the value of C), we conclude that, if $\lambda_1 \in I_j$,

$$\text{Im } m(z_j) - \text{Im } m_{sc}(z_j) \ge \frac{1}{2N\eta_j} - \frac{C\eta_j}{\sqrt{\kappa_j}} \ge \frac{1}{4N\eta_j}. \tag{8.8}$$

Next we observe that, again by the definition of κ_j and η_j, and for $K > K_0$ large enough,

$$\frac{1}{N\eta_j} \ge \frac{C(K+j)^{\frac{1}{2}}}{N\kappa_j} \ge \frac{C(K+j)^{\frac{1}{2}}}{N(\kappa_j+\eta_j)}$$

and

$$\frac{1}{N\eta_j} \ge \frac{C(K+j)^{\frac{1}{2}}}{(N\eta_j)^2\sqrt{\kappa_j}} \ge \frac{C(K+j)^{\frac{1}{2}}}{(N\eta_j)^2\sqrt{\kappa_j+\eta_j}}.$$

Hence, (8.8) implies that

$$\text{Im } m(z_j) - \text{Im } m_{sc}(z_j) \ge C(K+j)^{\frac{1}{2}} \left(\frac{1}{N(\kappa_j+\eta_j)} + \frac{1}{(N\eta_j)^2\sqrt{\kappa_j+\eta_j}} \right).$$

With Eq. (8.7) and by the bound (8.3) we conclude that

$$\begin{aligned}
\mathbb{P}\left(-E_0 \le \lambda_1 \le -2 - \frac{K}{N^{\frac{2}{3}}}\right) &\le \sum_{j=0}^{j_{max}} \mathbb{P}\left(\lambda_1 \in I_j\right) \\
&\le \sum_{j=0}^{j_{max}} \mathbb{P}\Big(|\text{ Im } \Lambda(z_j)| \ge C(K+j)^{\frac{1}{2}} \\
&\quad \times \left(\frac{1}{N(\kappa_j+\eta_j)} + \frac{1}{(N\eta_j)^2\sqrt{\kappa_j+\eta_j}} \right)\Big) \\
&\le \sum_{j=0}^{j_{max}} \frac{(Cq)^{cq^2}}{(K+j)^{q/2}} \le \frac{(Cq)^{cq^2}}{K^{q/2-2}} \sum_{j=0}^{j_{max}} \frac{1}{j^2} \le \frac{(Cq)^{cq^2}}{K^{q/2-2}}.
\end{aligned}$$

Changing $q/2 - 2 \to q$, this concludes, together with (8.5), the proof of the lemma. □

We are now ready to prove Theorem 4; we proceed here similarly as in the proof of Theorem 7.6 in [10].

Proof of Theorem 4 We prove the theorem only for $\alpha \leq N/2$, i.e., $\hat{\alpha} = \alpha$, the case $\alpha > N/2$ is similar. We will make use of the inequalities

$$c(2+x)^{\frac{3}{2}} \leq n_{sc}(x) \leq C(2+x)^{\frac{3}{2}}, \quad \text{and} \quad cn_{sc}(x)^{\frac{1}{3}} \leq \rho_{sc}(x) \leq Cn_{sc}(x)^{\frac{1}{3}} \tag{8.9}$$

which hold true for every $x \in [-2, 1]$ (see Eq. (7.31) in [10]). In particular, the second inequality implies that

$$c(\alpha/N)^{\frac{1}{3}} \leq \rho_{sc}(\gamma_\alpha) \leq C(\alpha/N)^{\frac{1}{3}}, \tag{8.10}$$

for any $\alpha \leq N/2$.

We have

$$\begin{aligned}\mathbb{P}\left(|\lambda_\alpha - \gamma_\alpha| \geq \frac{K\log N}{N}\left(\frac{N}{\alpha}\right)^{1/3}\right) &\leq \mathbb{P}\left(|\lambda_\alpha - \gamma_\alpha| \geq \frac{K\log N}{N}\left(\frac{N}{\alpha}\right)^{1/3} \text{ and } \lambda_\alpha \geq \gamma_\alpha\right)\\ &\quad + \mathbb{P}\left(|\lambda_\alpha - \gamma_\alpha| \geq \frac{K\log N}{N}\left(\frac{N}{\alpha}\right)^{1/3} \text{ and } \lambda_\alpha < \gamma_\alpha\right)\\ &= \mathrm{A} + \mathrm{B}.\end{aligned}$$

We consider first the term A. We set

$$\ell = \frac{K\log N}{N}\left(\frac{N}{\alpha}\right)^{1/3}.$$

From $|\lambda_\alpha - \gamma_\alpha| \geq \ell$ and $\lambda_\alpha \geq \gamma_\alpha$ we find $\lambda_\alpha \geq \gamma_\alpha + \ell$. This implies that $n\left(\gamma_\alpha + \ell\right) \leq n_{sc}(\gamma_\alpha)$. Therefore, by the mean value theorem, there exists $x^* \in [\gamma_\alpha; \gamma_\alpha + \ell]$ such that

$$\begin{aligned}&n_{sc}\left(\gamma_\alpha + \ell\right) - n\left(\gamma_\alpha + \ell\right)\\ &\quad = n_{sc}(\gamma_\alpha) + \rho_{sc}(x^*)\ell - n\left(\gamma_\alpha + \ell\right) \geq \rho_{sc}(x^*)\frac{K\log N}{N}\left(\frac{N}{\alpha}\right)^{1/3}.\end{aligned} \tag{8.11}$$

If $\gamma_\alpha + \ell > 1$, then $\lambda_\alpha \geq \gamma_\alpha + \ell$ implies $\lambda_\alpha > 1$. On the other hand, if $\gamma_\alpha + \ell \leq 1$, then $x^* \in [\gamma_\alpha, 1]$ and (8.10) implies that

$$\rho_{sc}(x^*) \geq \min(\rho_{sc}(\gamma_\alpha), \rho_{sc}(1)) \geq c\left(\frac{\alpha}{N}\right)^{1/3}.$$

From (8.11), we conclude that

$$\mathrm{A} \leq \mathbb{P}(\lambda_\alpha > 1) + \mathbb{P}\left(\left|n\left(\gamma_\alpha + \ell\right) - n_{sc}\left(\gamma_\alpha + \ell\right)\right| \geq c\frac{K\log N}{N}\right).$$

The event $\lambda_\alpha > 1$ implies that $n(1) < 1/2$ and therefore that $|n(1) - n_{sc}(1)| > n_{sc}(1) - 1/2 > c$ for some $c > 0$. Theorem 3 implies therefore that

$$\begin{aligned} \mathrm{A} &\le \mathbb{P}(|n(1) - n_{sc}(1)| > c) + \mathbb{P}\left(|n\,(\gamma_\alpha + \ell) - n_{sc}\,(\gamma_\alpha + \ell)| \ge c\frac{K \log N}{N}\right) \\ &\le (Cq)^{cq^2}\left[\left(\frac{\log N}{N}\right)^q + \frac{1}{K^q}\right]. \end{aligned}$$

If, say, $K \le 10N/(\log N)$, this implies that

$$\mathrm{A} \le \frac{(Cq)^{cq^2}}{K^q}. \tag{8.12}$$

If, on the other hand, $K > 10N/(\log N)$, then (8.12) follows from (8.1) since

$$\begin{aligned} &\mathbb{P}\left(|\lambda_\alpha - \gamma_\alpha| \ge \frac{K \log N}{N}\left(\frac{N}{\alpha}\right)^{1/3} \text{ and } \lambda_\alpha > \gamma_\alpha\right) \\ &\quad \le \mathbb{P}(\lambda_N > 8) \le \mathbb{P}(1 - n(8) > 1/N) \le N\,(1 - \mathbb{E}\,n(8)) \le N e^{-CN^\varepsilon} \le \frac{(Cq)^{cq}}{N^q} \end{aligned}$$

for all $q \in \mathbb{N}$.

Next, we estimate the term B. We distinguish two cases, $\alpha \le \tilde{c}K \log N$ and $\alpha > \tilde{c}K \log N$, for some sufficiently small constant $\tilde{c} > 0$.

Case 1 We assume here that $\alpha \le \tilde{c}K \log N$. From (8.9), we get $\gamma_\alpha \le -2 + C(\alpha/N)^{1/3}$ Hence, $|\lambda_\alpha - \gamma_\alpha| \ge \ell$ and $\lambda_\alpha < \gamma_\alpha$ imply that

$$\begin{aligned} \lambda_\alpha &< \gamma_\alpha - \ell \\ &\le -2 + C\left(\frac{\alpha}{N}\right)^{1/3} - \left(\frac{K \log N}{N}\right)\left(\frac{N}{\alpha}\right)^{1/3} \\ &\le -2 + \frac{1}{N^{2/3}}\left(C\alpha^{2/3} - \frac{K \log N}{\alpha^{1/3}}\right) \le -2 - \frac{1}{2}\left(\frac{K \log N}{N}\right)^{2/3} \end{aligned}$$

if the constant $\tilde{c}$ is small enough. From Lemma 8.1, we conclude that

$$\mathbb{P}(|\lambda_\alpha - \gamma_\alpha| > \ell,\ \lambda_\alpha < \gamma_\alpha \text{ and } \alpha \le \tilde{c}K \log N) \le \frac{(Cq)^{cq^2}}{K^q} \tag{8.13}$$

for all $q \le N^\varepsilon$.

Case 2 Assume now that $\alpha > \tilde{c}K \log N$. From (8.9), we have $\gamma_\alpha \ge -2 + c\,(\alpha/N)^{2/3}$, for some small constant $c > 0$. We define

$$y = -2 + \frac{c}{2}\left(\frac{\alpha}{N}\right)^{2/3}$$

and consider the cases $\gamma_\alpha - \ell > y$ and $\gamma_\alpha - \ell \leq y$ separately. Let us first assume $\gamma_\alpha - \ell > y$. Then $\lambda_\alpha < \gamma_\alpha - \ell$ implies that $n(\gamma_\alpha - \ell) \geq n_{sc}(\gamma_\alpha)$. Hence, from the mean value theorem, we find $x^* \in [\gamma_\alpha - \ell; \gamma_\alpha] \subset [y; \gamma_\alpha]$ such that

$$n(\gamma_\alpha - \ell) - n_{sc}(\gamma_\alpha - \ell) = n(\gamma_\alpha - \ell) - n_{sc}(\gamma_\alpha) + \rho_{sc}(x^*)\ell \geq \rho_{sc}(x^*) \frac{K \log N}{N} \left(\frac{N}{\alpha}\right)^{1/3}.$$

Since ρ_{sc} is increasing on $(-\infty; 0]$, we have

$$\rho_{sc}(x^*) \geq \rho_{sc}(y) \geq c\sqrt{2+y} \geq c\left(\frac{\alpha}{N}\right)^{1/3}.$$

From Theorem 3, we conclude that

$$\begin{aligned} &\mathbb{P}\Big(|\lambda_\alpha - \gamma_\alpha| > \ell,\ \lambda_\alpha < \gamma_\alpha,\ \alpha > \tilde{c} K \log N \text{ and } \gamma_\alpha - \ell > y\Big) \\ &\quad \leq \mathbb{P}\left(n(\gamma_\alpha - \ell) - n_{sc}(\gamma_\alpha - \ell) \geq c\frac{K \log N}{N}\right) \leq \frac{(Cq)^{cq^2}}{K^q} \end{aligned} \tag{8.14}$$

for any $q \in \mathbb{N}$. Finally, we consider the case $\gamma_\alpha - \ell \leq y$. Then $\lambda_\alpha < \gamma_\alpha - \ell$ also implies $\lambda_\alpha < y$ and therefore $n(y) - n_{sc}(\gamma_\alpha) \geq 0$. Hence, from the mean value theorem, there exists $x^* \in [y, \gamma_\alpha]$ with

$$n(y) - n_{sc}(y) = n(y) - n_{sc}(\gamma_\alpha) + \rho_{sc}(x^*)(\gamma_\alpha - y) \geq c\rho_{sc}(x^*)\left(\frac{\alpha}{N}\right)^{2/3} \tag{8.15}$$

by the very definition of y. Since $x^* > y$, we find again

$$\rho_{sc}(x^*) \geq c\left(\frac{\alpha}{N}\right)^{1/3}.$$

With (8.15), we conclude from Theorem 3 that

$$\begin{aligned} &\mathbb{P}\Big(|\lambda_\alpha - \gamma_\alpha| > \ell,\ \lambda_\alpha < \gamma_\alpha,\ \alpha > \tilde{c} K \log N \text{ and } \gamma_\alpha - \ell \leq y\Big) \\ &\quad \leq \mathbb{P}\left(n(y) - n_{sc}(y) > c\left(\frac{\alpha}{N}\right)\right) \leq \mathbb{P}\left(n(y) - n_{sc}(y) > c\frac{K \log N}{N}\right) \leq \frac{(Cq)^{cq^2}}{K^q}. \end{aligned} \tag{8.16}$$

Combining (8.13), (8.14) and (8.16), we obtain that $\mathrm{B} \leq (Cq)^{cq^2} K^{-q}$. □

Acknowledgments The work of B.S. has been partially supported by ERC Starting Grant MAQD-240518. C.C. acknowledges the support of the FIR 2013 project "Condensed Matter in Mathematical Physics" (Code RBFR13WAET) in the final stages of the development of the paper. A.M. acknowledges the support of the Leverhulme Trust Early Career Fellowship (ECF 2013-613). We would like to thank the referee for pointing out a mistake in a previous version of this paper.

Appendix: Large deviation estimates for quadratic forms

In the following proposition we recall an inequality for the fluctuations of quadratic forms, this is a well known result due to Hanson and Wright. For the proof we refer to [17, Prop. 4.5], see also [21, App. B] and [25].

Proposition 9.1 *For* $j = 1, \ldots, N$ *let* $x_j = \text{Re } x_j + i \text{ Im } x_j$, *where* $\{\text{Re } x_j, \text{Im } x_j\}_{j=1}^N$ *is a sequence of* $2N$ *real iid random variables, whose common distribution* ν *has sub-gaussian decay. Let* $A = (a_{ij})$ *be a* $N \times N$ *complex matrix. Then there exist constants* $c, C > 0$ *such that, for any* $\delta > 0$

$$\mathbb{P}\left(\left|\sum_{i,j=1}^{N} a_{ij}\left(x_i\bar{x}_j - \mathbb{E}x_i\bar{x}_j\right)\right| \geq \delta\sqrt{\text{Tr } A^*A}\right) \leq Ce^{-c\min\{\delta,\delta^2\}}.$$

References

1. Bourgade, P., Yau, H.-T., Yin, J.: Local circular law for random matrices. Probab. Theory. Relat. Fields. **159**(3–4), 545–595 (2014)
2. Bourgade, P., Yau, H.-T., Yin, J.: The local circular law II: the edge case. Probab. Theory. Relat. Fields. **159**(3–4), 619–660 (2014)
3. Cacciapuoti, C., Maltsev, A., Schlein, B.: Local Marchenko–Pastur law at the hard edge of sample covariance matrices. J. Math. Phys. **54**(4), 043302 (2013)
4. Erdős, L.: Universality of Wigner random matrices: a survey of recent results. Russ. Math. Surv. **66**(3), 507–626 (2011)
5. Erdős, L., Knowles, A.: Quantum diffusion and delocalization for band matrices with general distribution. Ann. Henri Poincaré **12**(7), 1227–1319 (2011)
6. Erdős, L., Knowles, A.: Quantum diffusion and eigenfunction delocalization in a random band matrix model. Commun. Math. Phys. **303**(2), 509–554 (2011)
7. Erdős, L., Knowles, A., Yau, H.-T.: Averaging fluctuations in resolvents of random band matrices. Ann. Henri Poincaré **14**(8), 1837–1926 (2013)
8. Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Delocalization and diffusion profile for random band matrices. Commun. Math. Phys. **323**(1), 367–416 (2013)
9. Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Spectral statistics of Erdős–Rényi graphs II: eigenvalue spacing and the extreme eigenvalues. Commun. Math. Phys. **314**(3), 587–640 (2012)
10. Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: The local semicircle law for a general class of random matrices. Electron. J. Probab. **18**(59), 1–58 (2013)
11. Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Spectral statistics of Erdős–Rényi graphs I: local semicircle law. Ann. Probab. **41**(3B), 2279–2375 (2013)
12. Erdős, L., Péché, S., Ramírez, J.A., Schlein, B., Yau, H.-T.: Bulk universality for Wigner matrices. Commun. Pure Appl. Math. **63**(7), 895–925 (2010)
13. Erdős, L., Ramirez, J., Schlein, B., Yau, H.-T.: Universality of sine-kernel for Wigner matrices with a small Gaussian perturbation. Electron. J. Probab. **15**(18), 526–604 (2010)
14. Erdős, L., Ramírez, J., Schlein, B., Tao, T., Vu, V., Yau, H.-T.: Bulk universality for Wigner Hermitian matrices with subexponential decay. Math. Res. Lett. **17**(4), 667–674 (2010)
15. Erdős, L., Schlein, B., Yau, H.-T.: Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices. Ann. Probab. **37**(3), 815–852 (2009)
16. Erdős, L., Schlein, B., Yau, H.-T.: Local semicircle law and complete delocalization for Wigner random matrices. Commun. Math. Phys. **287**(2), 641–655 (2009)
17. Erdős, L., Schlein, B., Yau, H.-T.: Wegner estimate and level repulsion for Wigner random matrices. Int. Math. Res. Not. IMRN **3**, 436–479 (2010)
18. Erdős, L., Schlein, B., Yau, H.-T.: Universality of random matrices and local relaxation flow. Invent. Math. **185**(1), 75–119 (2011)

19. Erdős, L., Schlein, B., Yau, H.-T., Yin, J.: The local relaxation flow approach to universality of the local statistics for random matrices. Ann. Inst. H. Poincaré Probab. Stat. **48**, 1–46 (2012)
20. Erdős, L., Yau, H.-T., Yin, J.: Universality for generalized Wigner matrices with Bernoulli distribution. J. Comb. **2**(1), 15–81 (2011)
21. Erdős, L., Yau, H.-T., Yin, J.: Bulk universality for generalized Wigner matrices. Probab. Theory Relat. Fields **154**(1–2), 341–407 (2012)
22. Erdős, L., Yau, H.-T., Yin, J.: Rigidity of eigenvalues of generalized Wigner matrices. Adv. Math. **229**(3), 1435–1515 (2012)
23. Götze, F., Tikhomirov, A.: On the rate of convergence to the semi-circular law. In: High Dimensional Probability VI. Progress in probability, vol. 66, pp. 139–165. Springer, Basel (2013)
24. Gustavsson, J.: Gaussian fluctuations of eigenvalues in the GUE. Ann. Inst. H. Poincaré Probab. Stat. **41**(2), 151–178 (2005)
25. Hanson, D.L., Wright, F.T.: A bound on tail probabilities for quadratic forms in independent random variables. Ann. Math. Stat. **42**, 1079–1083 (1971)
26. Johansson, K.: Universality of the local spacing distribution in certain ensembles of Hermitian Wigner matrices. Commun. Math. Phys. **215**(3), 683–705 (2001)
27. Lee, J.O., Yin, J.: A necessary and sufficient condition for edge universality of Wigner matrices. Duke Math. J. **163**(1), 117–173 (2014)
28. Maltsev, A., Schlein, B.: Average density of states of Hermitian Wigner matrices. Adv. Math. **228**(5), 2797–2836 (2011)
29. Marčenko, V.A., Pastur, L.A.: Distribution of eigenvalues in certain sets of random matrices. Mat. Sb. (N.S.) **72**(114), 507–536 (1967)
30. Pillai, N., Yin, J.: Universality of covariance matrices. Ann. Appl. Probab. **24**(3), 935–1001 (2014)
31. Soshnikov, A.: Universality at the edge of the spectrum in Wigner random matrices. Commun. Math. Phys. **207**(3), 697–733 (1999)
32. Tao, T., Vu, V.: Random matrices: universality of local eigenvalue statistics. Acta Math. **206**(1), 127–204 (2011)
33. Tao, T., Vu, V.: Random matrices: universality of local eigenvalue statistics up to the edge. Commun. Math. Phys. **298**(2), 549–572 (2010)
34. Tao, T., Vu, V.: Random covariance matrices: universality of local statistics of eigenvalues. Ann. Probab. **40**(3), 1285–1315 (2012)
35. Tao, T., Vu, V.: Random matrices: universality of local spectral statistics of non-Hermitian matrices. Ann. Probab. (2012, to appear). arXiv:1206.1893 [math.PR]
36. Wang, K.: Random covariance matrices: universality of local statistics of eigenvalues up to the edge. Random Matrices Theory Appl. **1**(1), 1150005 (2012)
37. Wigner, E.P.: Characteristic vectors of bordered matrices with infinite dimensions. Ann. Math. **2**(62), 548–564 (1955)
38. Yin, J.: The local circular law III: general case. Probab. Theory. Relat. Fields (2013). doi:10.1007/s00440-013-0539-3