The Annals of Applied Probability
2023, Vol. 33, No. 1, 447–489
https://doi.org/10.1214/22-AAP1820

FUNCTIONAL CENTRAL LIMIT THEOREMS FOR WIGNER MATRICES

BY GIORGIO CIPOLLONI[1,a], LÁSZLÓ ERDŐS[1,b] AND DOMINIK SCHRÖDER[2,c]

[1]*IST Austria,* [a]*gcipollo@ist.ac.at,* [b]*lerdos@ist.ac.at*
[2]*ETH Institute for Theoretical Studies, ETH Zürich,* [c]*dschoreder@ethz.ch*

We consider the fluctuations of regular functions f of a Wigner matrix W viewed as an entire matrix $f(W)$. Going beyond the well-studied tracial mode, $\operatorname{Tr} f(W)$, which is equivalent to the customary linear statistics of eigenvalues, we show that $\operatorname{Tr} f(W)A$ is asymptotically normal for any nontrivial bounded deterministic matrix A. We identify three different and asymptotically independent modes of this fluctuation, corresponding to the tracial part, the traceless diagonal part and the off-diagonal part of $f(W)$ in the entire mesoscopic regime, where we find that the off-diagonal modes fluctuate on a much smaller scale than the tracial mode. As a main motivation to study CLT in such generality on small mesoscopic scales, we determine the fluctuations in the eigenstate thermalization hypothesis (*Phys. Rev. A* **43** (1991) 2046–2049), that is, prove that the eigenfunction overlaps with any deterministic matrix are asymptotically Gaussian after a small spectral averaging. Finally, in the macroscopic regime our result also generalizes (*Zh. Mat. Fiz. Anal. Geom.* **9** (2013) 536–581, 611, 615) to complex W and to all crossover ensembles in between. The main technical inputs are the recent multiresolvent local laws with traceless deterministic matrices from the companion paper (*Comm. Math. Phys.* **388** (2021) 1005–1048).

1. Introduction. The eigenvalues $\{\lambda_i\}_{i=1}^N$ of large $N \times N$ Hermitian random matrices W form a strongly correlated system of random points on the real line. One manifestation of this feature is that their linear statistics, $\operatorname{Tr} f(W) = \sum_{i=1}^N f(\lambda_i)$ with a regular test function $f: \mathbf{R} \to \mathbf{R}$ has a variance of order one. In fact, it satisfies a central limit theorem (CLT) but without the customary $N^{-1/2}$ scaling factor. Note that Gaussian fluctuations normally emerge with the $N^{-1/2}$ factor as a cumulative effect of N independent or weakly dependent random variables. Thus it is quite remarkable that CLT holds for the strongly correlated eigenvalues and the anomalous scaling alone offsets all effects of these correlations, rendering the fluctuations of $\sum_i f(\lambda_i)$ still Gaussian.

What about the fluctuations of $f(W)$ viewed as a matrix and not just considering its trace? In this paper, we show that $f(W)$ tested against any bounded deterministic matrix A, $\|A\| \le 1$, is still asymptotically normal, provided that $\operatorname{Tr} AA^* \gtrsim N^{\epsilon}$. Our result holds in the macroscopic and in the entire mesoscopic regime, including spectral edges. More precisely, we consider the centered *functional linear statistics*

$$(1) \qquad L_N(f, A) := \operatorname{Tr}\big[f(W)A\big] - \mathbf{E}\operatorname{Tr}\big[f(W)A\big] = \sum_{i=1}^N f(\lambda_i)\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle - \mathbf{E}[\dots],$$

where $\boldsymbol{u}_i$ is the normalized eigenvector of W corresponding to λ_i. The statistics is called *macroscopic* if f is N-independent, and *mesoscopic on scale N^{-a}* with some exponent $a \in (0, 1)$ if f is of the form $f(x) = g(N^a(x - E))$ with some N-independent compactly

Received March 2021; revised February 2022.
MSC2020 subject classifications. Primary 60B20; secondary 15B52.
Key words and phrases. Quantum unique ergodicity, multiresolvent local law, multiscale Gaussian fluctuation, eigenfunction thermalization hypothesis.

supported function g, that is, if f lives on a scale N^{-a} around a fixed energy $E \in [-2, 2]$ in the spectrum.

One prominent motivation to study functional CLT on small mesoscopic scales is to understand the fluctuation in the eigenstate thermalization hypothesis in physics [18], also known as the strong quantum unique ergodicity (QUE) in mathematics [49]; see [16] for further references. QUE for Wigner matrices asserts that a law of large numbers holds for the eigenvector that overlaps with deterministic matrices A, that is, that $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i \rangle$ converges to the normalized trace of A as $N \to \infty$. In our companion paper [16], we established the optimal convergence rate of order $N^{-1/2+\epsilon}$, for any $\epsilon > 0$, with a very high probability. In Theorem 2.3 of the current paper, we prove that the overlaps $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i \rangle$ are asymptotically Gaussian after a small spectral averaging in the index i, which corresponds to the mesoscopic functional CLT for (1) when f is a characteristic function supported on a small spectral interval containing about N^{ϵ} eigenvalues for any arbitrary small $\epsilon > 0$. We remark that the Gaussian fluctuation of $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i \rangle$ is expected to hold for each i individually, but this result has only been proven for finite rank A using the Dyson Brownian motion for eigenvectors; see [9, 10, 43].

For $A = I$, the quantity $L_N(f, I)$ is the standard linear statistics of the eigenvalues that have been studied extensively both in the macroscopic regime by many authors [3, 5, 8, 27, 32, 33, 40, 51, 54, 55] and in the entire mesoscopic regime $a \in (0, 1)$ by He and Knowles; see also [1, 4, 6, 13, 19, 31, 35–37, 40] for related models on mesoscopic scales and [11, 12, 20, 21] for previous works on nonoptimal intermediate scales. It is therefore well known that $L_N(f, I)$ is asymptotically normal, that is, without a further $N^{-a/2}$ normalization it satisfies a central limit theorem with a variance given by essentially the $H^{1/2}$-norm of f; see (17). Note that the entire analysis of the special case $A = I$ is *tracial*; it relies only on the eigenvalues of W and is insensitive to its eigenvectors.

For the case of general observables, we decompose A as

$$A = \langle A \rangle I + \mathring{A}_{\mathrm{d}} + A_{\mathrm{od}}, \qquad \langle A \rangle := \frac{1}{N} \operatorname{Tr} A, \tag{2}$$

where $\mathring{A}_{\mathrm{d}} = A_{\mathrm{d}} - \langle A_{\mathrm{d}} \rangle$ is the traceless component of the diagonal part A_{d} of A and $A_{\mathrm{od}} := A - A_{\mathrm{d}}$ is the off-diagonal part of A. Following this decomposition, $L_N(f, A)$ has three different, mutually asymptotically independent Gaussian fluctuation modes, their expectations and variances are given in Theorem 2.4. On the macroscopic scale and for real symmetric Wigner matrices, this result was essentially obtained by Lytova in [38]. In Theorem 2.4, we extend [38] to complex Hermitian Wigner matrices including all crossover ensembles, that is, following the dependence on the real parameter $\sigma := N\mathbf{E}w_{12}^2$ in its entire range $\sigma \in [-1, 1]$ under the standard normalization $\mathbf{E}|w_{12}|^2 = \frac{1}{N}$, $\mathbf{E}w_{12} = 0$ for the off-diagonal matrix elements of W.

Our main contribution, however, is to establish a similar decomposition of fluctuations for the entire mesoscopic regime, $a \in (0, 1)$, since our Theorem 2.4 also allows for mesoscopic test functions. The corresponding limiting variances are computed in Propositions 2.9–2.10. For mesoscopic test functions f, the current paper contains the first results on the limiting distribution of $\operatorname{Tr}[f(W)A]$, with $A \neq I$. It turns out that the two traceless modes fluctuate on a scale of order $N^{-a/2}$ in the bulk and $N^{-3a/4}$ at the edge in contrast to the $\mathcal{O}(1)$ fluctuation scale of $L_N(f, I)$. Hence we not only need to explore the genuine off-diagonal fluctuations involving eigenvectors, but we also need to work at a much higher accuracy to detect the relevant fluctuations that are subleading compared with the previously explored regimes. This is a major new complication not present in the $a = 0$ macroscopic scale in [38]. Furthermore, we also show that mesoscopic linear statistics living on different scales are asymptotically independent (Theorem 2.13).

We explain the phenomenon of different fluctuation scales on the standard example of the resolvents, $G = G(z) = (W - z)^{-1}$ with spectral parameter $z \in \mathbf{C} \setminus \mathbf{R}$, that can be viewed as a function f of W living on scale $\eta := \Im z > 0$ around the point $E := \Re z$. To understand $\langle GA \rangle$ for a deterministic matrix A, we decompose A into its tracial and traceless parts as $A =: \langle A \rangle + \mathring{A}$ and write

$$\langle GA \rangle = m\langle A \rangle + \langle A \rangle \langle G - m \rangle + \langle G\mathring{A} \rangle, \tag{3}$$

where $m = m(z)$ is the Stieltjes transform of the semicircle law. The first term is deterministic, the second one is asymptotically Gaussian on scale $\langle (G - m)(z) \rangle \sim (N\eta)^{-1}$ by [30]. We prove that the last term in (3) is also Gaussian, independent of the first one, and it has size $\langle G\mathring{A} \rangle \sim \langle \mathring{A}\mathring{A}^* \rangle^{1/2}/(N\eta^{1/2})$, provided that $\langle \mathring{A}\mathring{A}^* \rangle \gg (N\eta)^{-1}$. In fact, it can be further split into a diagonal and off-diagonal part following (2). Thus the fluctuation of the tracial part is much bigger than that of the traceless part in the small η regime, however, the latter determines the fluctuation of $\langle GA \rangle$ for traceless observables $\langle A \rangle = 0$.

We now mention a few related works on general Gaussian fluctuations in Wigner matrices. In contrast to the extensively studied linear eigenvalue statistics, this question received much less attention in the random matrix community, although a Wigner matrix contains many other physically or mathematically relevant random modes and most of them are expected to be Gaussian (notable exception is the eigenvalue gaps that follow the Wigner–Dyson statistics). Besides Lytova's work [38], tracial CLTs for certain minors were obtained in [25]. Special functional CLTs have been proven for Haar distributed matrices [52, 53], and for partial traces of invariant ensembles [45]. The free probability community has systematically studied Gaussian fluctuations of traces of products of a Wigner matrix and deterministic matrices via the concept of *second-order freeness* [17, 44]. This theory has recently been extended to polynomials in several independent Wigner matrices ([42], Theorems 3–4). However, these results rely on the moment method and handle only *polynomials* of Wigner matrices. It is yet unclear if the moment approach can be extended to general functions on the macroscopic scale; mesoscopic scales seem inaccessible.

Finally, we mention that the fluctuation of certain specific observables may be non-Gaussian. For example, the fluctuation of matrix entries $f(W)_{ij}$ of $f(W)$ for regular test functions f is a linear combination of w_{ij} and an independent Gaussian of size $N^{-1/2}$; see [24, 40, 41, 46, 47]. In contrast, our result shows that $\operatorname{Tr} f(W)A$ is always asymptotically Gaussian whenever $\|A\| \sim 1$ and $\langle AA^* \rangle \gtrsim N^{-1+\epsilon}$. Hence, the non-Gaussian components of $f(W)$ are only visible for very low rank observables A.

The paper is structured as follows. After this Introduction, we present the main results in the next Section 2. We start with our motivating Theorem 2.3 on the Gaussian fluctuation of the overlaps $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i \rangle$ after a small spectral averaging in i. Then we formulate our functional CLT (Theorem 2.4) in full generality in the bulk and at the edge of the spectrum of W, from the macroscopic scale down to the smallest possible mesoscopic scale just above the local eigenvalue spacing. Our formulation exhibits the three distinguished fluctuation modes with their own scaling factors. Simplified formulas in the mesoscopic regime for the expectations and the variances of the limit Gaussian processes are given in Proposition 2.9 in the bulk and in Proposition 2.10, respectively. We also include all the additional effects of the fourth cumulant $\kappa_4 = N^2\mathbf{E}|w_{12}|^4 - 2 - \sigma^2$ of the off-diagonal matrix element w_{12}, the parameter $\sigma = N\mathbf{E}w_{12}^2$ describing the crossover regime between complex and real symmetry class and the size of the diagonal element $w_2 = N\mathbf{E}w_{11}^2$. These three parameters appear in the exact form of the limiting expectations and variances of the three different modes of $L_N(f, A)$. Some earlier works assumed special values of these parameters, for example, $\sigma = 0, 1$ and $w_2 = 1 + \sigma$ is a typical choice in certain more restricted definition of the Wigner ensemble. Consequently, some explicit terms did not always appear. We also identify the cases when some of these three limiting modes have vanishing variance and explain their algebraic origin in Appendix A. Finally, in Theorem 2.13 we show that fluctuations on different scales are asymptotically independent. In Section 3, we present the necessary multiresolvent local laws: some of them have already been proven in; some others, especially the ones involving three resolvents, are shown here with some proofs deferred to the Supplementary Material [14], Appendix D.2. The main technical input for all these cases is [16], Theorem 5, and its slight extension in Theorem 3.5, proven in [14], Appendix D.1, that control the most critical fluctuation term (the so-called renormalized "underlined" term) in the self-consistent equation for products of resolvents and deterministic matrices. Some additional technical estimates are deferred to [14], Appendix B. In Section 4, we prove a general CLT for resolvents; this section is the technical centerpiece of the current paper. Finally, in Section 5 we convert the resolvents into general functions by using Helffer–Sjöstrand-type formulas, and thus prove our general functional CLTs. The proof of Theorem 2.3 is given in full detail in Section 5, while several technical calculations for the proof of the very general Theorem 2.4 are deferred to [14], Appendix F.

Notation and conventions. We introduce some notation we use throughout the paper. For integers $k \in \mathbf{N}$, we use the notation $[k] := \{1, \dots, k\}$. For positive quantities f, g, we write $f \lesssim g$ and $f \sim g$ if $f \le Cg$ or $cg \le f \le Cg$, respectively, for some constants $c, C > 0$, which depend only on the moments of the matrix elements, that is, on the constants appearing in (5). We denote vectors by bold-faced lower case Roman letters $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{C}^k$, for some $k \in \mathbf{N}$. Vector and matrix norms, $\|\boldsymbol{x}\|$ and $\|A\|$, indicate the usual Euclidean norm and the corresponding induced matrix norm. For any $N \times N$ matrix A, we use the notation $\langle A \rangle := N^{-1} \operatorname{Tr} A$ to denote the normalized trace of A. Moreover, for vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{C}^N$ we define

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \sum \overline{x}_i y_i, \qquad A_{\boldsymbol{x}\boldsymbol{y}} := \langle \boldsymbol{x}, A\boldsymbol{y} \rangle,$$

with $A \in \mathbf{C}^{N \times N}$.

We will use the concept of "with very high probability" meaning that for any fixed $D > 0$ the probability of the N-dependent event is bigger than $1 - N^{-D}$ if $N \ge N_0(D)$. Moreover, we use the convention that $\xi > 0$ denotes an arbitrary small constant which is independent of N.

2. Main results. Let W be an $N \times N$ real or complex Wigner matrix with eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_N$ and corresponding orthonormal eigenvectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_N$. The eigenvalue density profile is described by the semicircular law

$$\rho(x) = \rho_{\mathrm{sc}}(x) := \frac{\sqrt{4 - x^2}}{2\pi}. \tag{4}$$

On the entries of W, we formulate the following assumptions.

ASSUMPTION 2.1. The matrix elements w_{ab} of W are independent up to Hermitian symmetry $w_{ab} = \overline{w_{ba}}$. We assume identical distribution in the sense that $w_{ab} \overset{\mathrm{d}}{=} N^{-1/2} \chi_{\mathrm{od}}$, for $a < b$, $w_{aa} \overset{\mathrm{d}}{=} N^{-1/2} \chi_{\mathrm{d}}$, with χ_{d} being a real, and χ_{od} being either a real or complex random variable such that $\mathbf{E}\chi_{\mathrm{od}} = \mathbf{E}\chi_{\mathrm{d}} = 0$, $\mathbf{E}|\chi_{\mathrm{od}}|^2 = 1$ and $\sigma := \mathbf{E}\chi_{\mathrm{od}}^2 \in \mathbf{R}$. In addition, we assume the existence of the high moments of χ_{od}, χ_{d}, that is, that there exist constants $C_p > 0$, for any $p \in \mathbf{N}$, such that

$$\mathbf{E}|\chi_{\mathrm{d}}|^p + \mathbf{E}|\chi_{\mathrm{od}}|^p \le C_p. \tag{5}$$

Notice that $\sigma \in [-1, 1]$; the case $\sigma = 0$ corresponds to complex Hermitian Wigner matrices with $\mathbf{E}w_{ab}^2 = 0$, the case $\sigma = 1$ corresponds to real symmetric matrices, and the case $\sigma = -1$ corresponds Wigner matrices $W = D + \mathrm{i}O$, with D being a diagonal matrix and O being real skew-symmetric, that is, $O^t = -O$.

Finally, in order to state our results compactly, we introduce the following notation to indicate that two random vectors have asymptotically equal moments.

NOTATION 2.2. For two random vectors $X = (X_1, \dots, X_k)$, $Y = (Y_1, \dots, Y_k)$, with $k \in \mathbf{N}$, of N-dependent random variables we define of the concept of *closeness in the sense of moments* and we denote it as

$$X \stackrel{\mathrm{m}}{=} Y + \mathcal{O}_{\mathrm{m}}(N^{-c})$$

for some $c > 0$, if for any polynomial $p(x_1, \dots, x_k)$ it holds that

$$\mathbf{E}p(X_1, \dots, X_k) = \mathbf{E}p(Y_1, \dots, Y_k) + \mathcal{O}(N^{-c+\xi}),$$

for any small $\xi > 0$, where the implicit constant in $\mathcal{O}(\cdot)$ depends on k, ξ, the polynomial p and the constants in Assumption 2.1.

2.1. *CLT for eigenvector overlaps.* As explained in the Introduction, the Gaussian fluctuation of the eigenvector overlaps $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle$ with a deterministic matrix A is a fundamental question since it describes the fluctuation in the strong quantum unique ergodicity for Wigner matrices. This problem has only been solved for finite rank A; see [9, 10, 43]. Our first theorem establishes an averaged version of this CLT for general A.

THEOREM 2.3 (CLT for averages of eigenvector overlaps). *Let A be a deterministic $N \times N$ matrix with $\|A\| \le 1$ and let $\mathring{A} := A - \langle A\rangle$ denote its traceless part. Let $\epsilon > 0$ and $K \in \mathbf{N}$ with $N^{\epsilon} \le K \le N^{1-\epsilon}$. Then for some $\omega = \omega(\epsilon) > 0$ we have the CLT at the edge*:

$$\begin{aligned}&\frac{1}{\sqrt{K}} \sum_{i=N-K+1}^{N} \sqrt{N}\big[\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle - \langle A\rangle\big]\\ &\quad\stackrel{\mathrm{m}}{=} \mathcal{N}\left(0, \frac{\sqrt{2}}{3}\big[\langle \mathring{A}\mathring{A}^*\rangle + \mathbf{1}(\sigma = 1)\langle \mathring{A}\overline{\mathring{A}}\rangle\big]\right) + \mathcal{O}_{\mathrm{m}}(N^{-\omega}).\end{aligned} \tag{6}$$

Moreover, for any $\delta > 0$ and $\delta N < i_0 < (1-\delta)N$ and $\sigma > -1$, we have CLT in the bulk:

$$\begin{aligned}&\frac{1}{\sqrt{2K}} \sum_{|i-i_0|\le K} \sqrt{N}\big[\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle - \langle A\rangle\big]\\ &\quad\stackrel{\mathrm{m}}{=} \mathcal{N}\left(0, \frac{\langle \mathring{A}\mathring{A}^*\rangle + \mathbf{1}(\sigma = 1)\langle \mathring{A}\overline{\mathring{A}}\rangle}{2}\right) + \mathcal{O}_{\mathrm{m}}(N^{-\omega}),\end{aligned} \tag{7}$$

where the implicit constant in $\mathcal{O}_{\mathrm{m}}(\cdot)$ depends on δ. Finally, in case $\sigma = -1$ for any fixed $c \in (\delta, 1-\delta)$ we have a slightly different CLT in the bulk:

$$\begin{aligned}&\frac{1}{\sqrt{2K}} \sum_{|i-cN|\le K} \sqrt{N}\big[\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle - \langle A\rangle\big]\\ &\quad\stackrel{\mathrm{m}}{=} \mathcal{N}\left(0, \frac{\langle \mathring{A}\mathring{A}^*\rangle + \mathbf{1}(c = 1/2)\langle \mathring{A}\overline{\mathring{A}}\rangle}{2}\right) + \mathcal{O}_{\mathrm{m}}(N^{-\omega}).\end{aligned} \tag{8}$$

In the next subsection, we formulate the CLT for the functional linear statistics (1) in full generality for regular test functions f. Theorem 2.3 is a special case of such CLT on mesoscopic scales with f essentially being the characteristic function of an interval. While this sharp cut-off test function formally does not satisfy the regularity condition imposed on f in Theorem 2.4 below, in Section 5 we will show how to cover this special case as well.

2.2. *General functional CLT.* Let $g \in H_0^2(\mathbf{R})$ be a compactly supported real valued test function, then for $0 \le a < 1$ and $|E| \le 2$ we define the test function rescaled to a scale N^{-a} around E as

$$f(x) := g\big(N^a(x - E)\big). \tag{9}$$

The scale $a = 0$ corresponds to the *macroscopic regime*. The scales $0 < a < 1$ in the bulk and $0 < a < 2/3$ at the edges, $|E| = 2$, correspond to the *mesoscopic regime*. Our result holds uniformly in E, that is, it also covers the entire transitionary regime between bulk and edge.

For deterministic $N \times N$ matrices A, and test functions[1] $f : \mathbf{R} \to \mathbf{R}$ defined as in (9), we define the centered linear statistics

$$L_N(f, A) := \sum_{i=1}^N f(\lambda_i)\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle - \mathbf{E}\sum_{i=1}^N f(\lambda_i)\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle. \tag{10}$$

For the general CLT, it is natural to decompose the space of matrices in three mutually orthogonal subspaces. We will write a general matrix A as

$$A = A_{\mathrm{d}} + A_{\mathrm{od}} = \langle A\rangle I + \mathring{A}_{\mathrm{d}} + A_{\mathrm{od}}, \qquad \mathring{A} := A - \langle A\rangle,$$

that is, as the sum of a constant multiple of the identity matrix, a diagonal traceless matrix $\mathring{A}_{\mathrm{d}}$, and an off-diagonal matrix A_{od}. Given the decomposition of A, the linear statistics has three modes

$$L_N(f, A) = \langle A\rangle L_N(f, I) + L_N(f, \mathring{A}_{\mathrm{d}}) + L_N(f, A_{\mathrm{od}}), \tag{11}$$

which we prove to be asymptotically independent Gaussians.

For the sake of shorter notation, we denote the expectation of a function f with respect to the semicircular density and its inverse by

$$\langle f\rangle_{\mathrm{sc}} := \int_{-2}^2 f(x)\frac{\sqrt{4 - x^2}}{2\pi}\,\mathrm{d}x, \qquad \langle f\rangle_{1/\mathrm{sc}} := \int_{-2}^2 \frac{f(x)}{\pi\sqrt{4 - x^2}}\,\mathrm{d}x.$$

We also define the Stieltjes transform of the semicircle law

$$m(z) = m_{\mathrm{sc}}(z) := \int_{-2}^2 \frac{\rho_{\mathrm{sc}}(x)}{x - z}\,\mathrm{d}x, \quad z \in \mathbf{C} \setminus \mathbf{R}. \tag{12}$$

We set $\rho(z) := \frac{1}{\pi}|\Im m_{\mathrm{sc}}(z)|$ and note that $\rho(x + \mathrm{i}0) = \rho_{\mathrm{sc}}(x)$.

Finally, we introduce some notation related to the distribution of the matrix elements of W. We denote the normalized fourth cumulant of the off-diagonal entries, the expectation of the square of the off-diagonal entries and variance of the diagonal entries of W and a certain frequently used combination of them by

$$\begin{aligned} &\kappa_4 := \mathbf{E}|\chi_{\mathrm{od}}|^4 - 2 - \sigma^2, \qquad \sigma := \mathbf{E}\chi_{\mathrm{od}}^2, \\ &w_2 := \mathbf{E}\chi_{\mathrm{d}}^2, \qquad \widetilde{w_2} := w_2 - 1 - \sigma, \end{aligned} \tag{13}$$

respectively.

We now state our main result, the functional CLT in both the macroscopic $a = 0$ and mesoscopic $a > 0$ regimes. In Theorem 2.4, we rescale the traceless diagonal linear statistics $L_N(f, \mathring{A}_{\mathrm{d}})$ and the off-diagonal linear statistics $L_N(f, A_{\mathrm{od}})$ in such a way the limiting processes are, to leading order, N-independent except for the explicit dependence on $\langle|\mathring{A}_{\mathrm{d}}|^2\rangle$ and $\langle A_{\mathrm{od}}A_{\mathrm{od}}^*\rangle$, irrespective of the scaling parameter a for test functions of the form (9). In Section 2.3 below, we provide explicit formulas for the mesoscopic limits of the processes in terms of g, demonstrating the N-independence of $L_N(f, \mathring{A}_{\mathrm{d}})$, $L_N(f, A_{\mathrm{od}})$ to leading order.

[1] By linearity, there is no loss in generality by assuming the test function to be real valued.

THEOREM 2.4 (Macroscopic and mesoscopic functional CLT). *Let* $0 \le a < 1$ *and define the scaling factor for any, possibly* N*-dependent,* $E = E_N \in [-2, 2]$ *as*

$$C_N = C_N^{a,E} := \frac{N^a}{\rho_N^{a,E}}, \quad \textit{where } \rho_N = \rho_N^{a,E} := \begin{cases} \rho(E + \mathrm{i}N^{-a}) & a > 0, \\ 1 & a = 0. \end{cases}$$

Note that $C_N = 1$ *for the macroscopic* $a = 0$ *case. Let* $g \in H_0^2(\mathbf{R})$ *be a compactly supported function and set* $f(x) := g(N^a(x - E))$. *Let* A *be a deterministic matrix with* $\|A\| \le 1$. *Then, in the limiting the regime* $C_N \ll N$, *the three centered linear statistics* (11) *are approximately distributed* (*in the sense of moments*)

$$\big(L_N(f, I), \sqrt{C_N} L_N(f, \mathring{A}_{\mathrm{d}}), \sqrt{C_N} L_N(f, A_{\mathrm{od}})\big)$$

$$\overset{\mathrm{m}}{=} \big(\xi_{\mathrm{tr}}(f), \xi_{\mathrm{d}}(f, \mathring{A}_{\mathrm{d}}), \xi_{\mathrm{od}}(f, A_{\mathrm{od}})\big) + \mathcal{O}_{\mathrm{m}}\bigg(\sqrt{\frac{C_N}{N}}\bigg)$$

as three independent centered N*-dependent Gaussian processes* $\xi_{\mathrm{tr}}(f)$, $\xi_{\mathrm{d}}(f, \mathring{A}_{\mathrm{d}})$, $\xi_{\mathrm{od}}(f, A_{\mathrm{od}})$ *whenever* $\langle |\mathring{A}_{\mathrm{d}}|^2 \rangle$, $\langle A_{\mathrm{od}} A_{\mathrm{od}}^* \rangle \gtrsim C_N N^{-1+\epsilon}$ *for some* $\epsilon > 0$. *Their variances are given by*[2]

$$\mathbf{E}\big|\xi_{\mathrm{tr}}(f)\big|^2 = V_{\mathrm{tr}}^1(f) + V_{\mathrm{tr}}^2(f, \sigma) + \frac{\kappa_4}{2}\big\langle(2 - x^2) f\big\rangle_{1/\mathrm{sc}}^2 + \frac{\widetilde{w_2}}{4}\langle x f\rangle_{1/\mathrm{sc}}^2, \tag{14}$$

$$\begin{aligned} \mathbf{E}\big|\xi_{\mathrm{d}}(f, \mathring{A}_{\mathrm{d}})\big|^2 = {} & C_N \langle |\mathring{A}_{\mathrm{d}}|^2 \rangle \big(V_{\mathrm{d}}^1(f) + V_{\mathrm{d}}^2(f, \sigma) \\ & + \widetilde{w_2}\langle f x\rangle_{\mathrm{sc}}^2 + \kappa_4 \big\langle (x^2 - 1) f\big\rangle_{\mathrm{sc}}^2\big), \end{aligned} \tag{15}$$

$$\mathbf{E}\big|\xi_{\mathrm{od}}(f, A_{\mathrm{od}})\big|^2 = C_N\big(\langle A_{\mathrm{od}} A_{\mathrm{od}}^*\rangle V_{\mathrm{d}}^1(f) + \langle A_{\mathrm{od}} \overline{A_{\mathrm{od}}}\rangle V_{\mathrm{d}}^2(f, \sigma)\big), \tag{16}$$

with

$$V_{\mathrm{tr}}^1(f) := \frac{1}{4\pi^2} \iint_{-2}^{2} \bigg(\frac{f(x) - f(y)}{x - y}\bigg)^2 \frac{4 - xy}{\sqrt{(4 - x^2)(4 - y^2)}} \,\mathrm{d}x\,\mathrm{d}y, \tag{17}$$

$$\begin{aligned} V_{\mathrm{tr}}^2(f, \sigma) := {} & \frac{1}{4\pi^2} \iint_{-2}^{2} f(x) f(y) \partial_x \partial_y \\ & \times \log\bigg[\frac{(x - \sigma y)^2 + (\sqrt{4 - x^2} + \sigma\sqrt{4 - y^2})^2}{(x - \sigma y)^2 + (\sqrt{4 - x^2} - \sigma\sqrt{4 - y^2})^2}\bigg] \,\mathrm{d}x\,\mathrm{d}y, \end{aligned} \tag{18}$$

$$V_{\mathrm{d}}^1(f) := \langle f^2 \rangle_{\mathrm{sc}} - \langle f \rangle_{\mathrm{sc}}^2, \tag{19}$$

$$\begin{aligned} V_{\mathrm{d}}^2(f, \sigma) := {} & \frac{1}{4\pi^2} \iint_{-2}^{2} f(x) f(y) \\ & \times \frac{(1 - \sigma^2)\sqrt{(4 - x^2)(4 - y^2)}}{\sigma^2(x^2 + y^2) + (1 - \sigma^2)^2 - xy\sigma(1 + \sigma^2)} \,\mathrm{d}x\,\mathrm{d}y - \langle f \rangle_{\mathrm{sc}}^2, \end{aligned} \tag{20}$$

for $|\sigma| < 1$, *and* V_{tr}^2, V_{d}^2 *are extended to* $\sigma = \pm 1$ *by continuity,* $V_{\mathrm{tr/d}}^2(f, \pm 1) := \lim_{\sigma \to \pm 1} V_{\mathrm{tr/d}}^2(f, \sigma)$. *Moreover, for any* $\xi > 0$, *for the expectation of the linear statistics*

[2]The Gaussians are scaled such that $\xi_{\mathrm{tr}}(f)$, $\xi_{\mathrm{d}}(f, \mathring{A}_{\mathrm{d}})/\langle |\mathring{A}_{\mathrm{d}}|^2\rangle^{1/2}$, $\xi_{\mathrm{od}}(f, A_{\mathrm{od}})/\langle A_{\mathrm{od}} A_{\mathrm{od}}^*\rangle^{1/2}$ are of order one. The N-dependence of C_N is exactly offset by the N-dependence of $V_{\mathrm{d}}^1(f)$ etc. in the mesoscopic regime; see Section 2.3.

we have the expansions

$$\begin{aligned}\mathbf{E}\sum_i f(\lambda_i) = N\langle f\rangle_{\mathrm{sc}} + \frac{\kappa_4}{2}\langle (x^4-4x^2+2)f\rangle_{1/\mathrm{sc}} - \frac{\widetilde{w_2}}{2}\langle (2-x^2)f\rangle_{1/\mathrm{sc}} \\ - \frac{E_{\mathrm{tr}}(f,\sigma)}{2} + \mathcal{O}\Big(N^\xi\sqrt{\frac{C_N}{N}}\Big),\end{aligned} \tag{21}$$

$$\Big|\mathbf{E}\sum_i f(\lambda_i)\langle \boldsymbol{u}_i, A_{\mathrm{od}}\boldsymbol{u}_i\rangle\Big| + \Big|\mathbf{E}\sum_i f(\lambda_i)\langle \boldsymbol{u}_i, \mathring{A}_{\mathrm{d}}\boldsymbol{u}_i\rangle\Big| = \mathcal{O}\Big(\frac{N^\xi}{\sqrt{N}}\Big), \tag{22}$$

where

$$E_{\mathrm{tr}}(f,\sigma) := \Big\langle f\Big(1 - \frac{1-\sigma^2}{(1+\sigma)^2-\sigma x^2}\Big)\Big\rangle_{1/\mathrm{sc}}, \quad |\sigma|<1, \tag{23}$$

and $E_{\mathrm{tr}}(f,\pm 1) := \lim_{\sigma\to\pm 1} E_{\mathrm{tr}}(f,\sigma)$.[3] *The implicit constants in* $\mathcal{O}(\cdot)$ *in all error terms above depend only on the model parameters in Assumptions 2.1 and on* $\|g\|_{H_0^2}$, $|\operatorname{supp} g|$ (*additionally in* (22) *the constant also depends on* ξ), *in particular they are independent of* E.

Theorem 2.4 is only meaningful in the regime where $C_N \ll N$, equivalently, when N^{-a} is above the local eigenvalue spacing around E, by using that $\rho(E+iN^{-a}) \sim (||E|-2| + N^{-a})^{1/2}$. Thus our result covers the entire mesoscopic range uniformly for any $|E|\le 2$. In particular, we allow for the range $a\in[0,1)$ in the bulk regime, $|E|\le 2-\epsilon$, and $a\in[0,2/3)$ in the edge regime, $|E|=2$.

We note that the expectation of $\operatorname{Tr} f(W)$ is typically of order N, hence much larger than its fluctuation. However, Theorem 2.4 identifies the leading term of $\mathbf{E}\operatorname{Tr} f(W)$ to an accuracy beyond its fluctuation size. For both $\operatorname{Tr} f(W)\mathring{A}_{\mathrm{d}}$ and $\operatorname{Tr} f(W)A_{\mathrm{od}}$, their expectations are much smaller than their fluctuation.

For simplicity, we formulated Theorem 2.4 for linear statistics with one test function f only. Our method, however, can handle linear combinations of test functions living on different scales since the main input of Theorem 2.4, the resolvent CLT in Theorem 4.1, allows for each involved resolvent to be evaluated at its own spectral parameter with possibly very different imaginary parts. Hence, by standard polarization, a multivariate variant of Theorem 2.4 directly follows.

COROLLARY 2.5 (Multivariate CLT). *Let* $p\in\mathbf{N}$, $E_1,\dots,E_p\in[-2,2]$, $0\le a_1,\dots,a_p<1$, *and let* $g_1,\dots,g_p\in H_0^2(\mathbf{R})$ *be compactly supported test functions and set*

$$f_i(x) := g_i\big(N^{a_i}(x-E_i)\big).$$

Then for deterministic matrices $A_1,\dots,A_p$ *of bounded norms,* $\|A_i\|\lesssim 1$ *the joint linear statistics* (11) *are approximately distributed* (*in the sense of moments*)

$$\begin{aligned}&\Big(L_N(f_i,I), \sqrt{C_N^{a_i,E_i}}L_N\big(f_i,(\mathring{A}_i)_{\mathrm{d}}\big), \sqrt{C_N^{a_i,E_i}}L_N\big(f_i,(A_i)_{\mathrm{od}}\big)\Big)_{i\in[p]} \\ &\overset{\mathrm{m}}{=} \big(\xi_{\mathrm{tr}}(f_i), \xi_{\mathrm{d}}\big(f_i,(\mathring{A}_i)_{\mathrm{d}}\big), \xi_{\mathrm{od}}\big(f_i,(A_i)_{\mathrm{od}}\big)\big)_{i\in[p]} + \mathcal{O}_{\mathrm{m}}\Big(\sqrt{\frac{\max_i C_N^{a_i,E_i}}{N}}\Big)\end{aligned}$$

[3]Note that $E_{\mathrm{tr}}(f,\pm 1)$ is defined as a limit $\sigma\to\pm 1$, which is different from plugging $\sigma=\pm 1$ into (23). In particular, $\sigma\to E_{\mathrm{tr}}(f,\sigma)$ is continuous on the closed interval $[-1,1]$.

as centered Gaussian processes ξ_{tr}, ξ_{d}, ξ_{od} *of covariances obtained from the variances in Theorem* 2.4 *by polarization, uniformly in* $E_i \in [-2,2]$. *The implicit constant in the* $\mathcal{O}(\cdot)$ *error term above only depends on the model parameters in Assumption* 2.1 *and on* g *via* $\|g_i\|_{H_0^2}$ *and* $|\operatorname{supp} g_i|$, *in particular it is independent of* E_i.

REMARK 2.6 (Alternative representation of the variances in Theorem 2.4 via Chebyshev polynomials). By a direct computation using the geometric series, we find

$$V_{\mathrm{tr}}^1(f) = \sum_{k\geq 1} k\langle f t_k\rangle_{1/\mathrm{sc}}^2, \qquad V_{\mathrm{tr}}^2(f,\sigma) = \sum_{k\geq 1} k\sigma^k\langle f t_k\rangle_{1/\mathrm{sc}}^2, \tag{24}$$

$$V_{\mathrm{d}}^1(f) = \sum_{k\geq 1} \langle f u_k\rangle_{\mathrm{sc}}^2, \qquad V_{\mathrm{d}}^2(f,\sigma) = \sum_{k\geq 1} \sigma^k\langle f u_k\rangle_{\mathrm{sc}}^2, \tag{25}$$

where $t_k(x) := T_k(x/2)$, $u_k(x) := U_k(x/2)$ and T_k, U_k are the kth Chebyshev polynomial of the first and second kind, that is, $T_k(\cos\theta) = \cos(k\theta)$, $U_k(\cos\theta) = \sin[(k+1)\theta]/\sin\theta$. In particular, we can recover the representation of $\mathbf{E}|\xi_{\mathrm{tr}}(f)|^2$ obtained in [5], equation (1.5), and write

$$\mathbf{E}|\xi_{\mathrm{tr}}(f)|^2 = \sum_{k\geq 3} k(1+\sigma^k)\langle f t_k\rangle_{1/\mathrm{sc}}^2 + 2(\kappa_4+1+\sigma^2)\langle f t_2\rangle_{1/\mathrm{sc}}^2 + w_2\langle f t_1\rangle_{1/\mathrm{sc}}^2. \tag{26}$$

Similarly, for the ξ_{d}, ξ_{od} we obtain

$$\begin{aligned} \mathbf{E}|\xi_{\mathrm{d}}(f,\mathring{A}_{\mathrm{d}})|^2 = \langle|\mathring{A}_{\mathrm{d}}|^2\rangle\Bigg[&\sum_{k\geq 3}(1+\sigma^k)\langle f u_k\rangle_{\mathrm{sc}}^2 + w_2\langle f u_1\rangle_{\mathrm{sc}}^2 \\ &+ (\kappa_4+1+\sigma^2)\langle f u_2\rangle_{\mathrm{sc}}^2\Bigg], \end{aligned} \tag{27}$$

$$\mathbf{E}|\xi_{\mathrm{od}}(f,A_{\mathrm{od}})|^2 = \sum_{k\geq 1}\langle f u_k\rangle_{\mathrm{sc}}^2\big(\langle A_{\mathrm{od}}A_{\mathrm{od}}^*\rangle + \sigma^k\langle A_{\mathrm{od}}\overline{A_{\mathrm{od}}}\rangle\big). \tag{28}$$

Note, that (26)–(27) are sums of nonnegative terms since $w_2\geq 0$ and $\kappa_4 = \mathbf{E}|\chi_{\mathrm{od}}|^4 - 2 - \sigma^2 \geq -1-\sigma^2$ due to $\mathbf{E}|\chi_{\mathrm{od}}|^4 \geq (\mathbf{E}|\chi_{\mathrm{od}}|^2)^2 = 1$. Similarly, (28) is a sum of nonnegative terms since $\sigma^k\langle A_{\mathrm{od}}\overline{A_{\mathrm{od}}}\rangle \geq -|\langle A_{\mathrm{od}}\overline{A_{\mathrm{od}}}\rangle| \geq -\langle A_{\mathrm{od}}A_{\mathrm{od}}^*\rangle$.

REMARK 2.7 (Explicit formulas for $\sigma = \pm 1$). The limits of (23) are explicitly given by

$$E_{\mathrm{tr}}(f,1) = \langle f\rangle_{1/\mathrm{sc}} - \frac{f(2)+f(-2)}{2}, \qquad E_{\mathrm{tr}}(f,-1) = \langle f\rangle_{1/\mathrm{sc}} - f(0). \tag{29}$$

For the variances in case $\sigma = 1$, we have $V_{\mathrm{tr}}^2(f,1) = V_{\mathrm{tr}}^1(f)$ and $V_{\mathrm{d}}^2(f,1) = V_{\mathrm{d}}^1(f)$, while for $\sigma = -1$ we have

$$\begin{aligned} V_{\mathrm{tr}}^2(f,-1) &= \frac{1}{4\pi^2}\iint_{-2}^{2}\frac{(f(x)-f(y))(f(-x)-f(-y))}{(x-y)^2} \\ &\qquad\times\frac{4-xy}{\sqrt{(4-x^2)(4-y^2)}}\,\mathrm{d}x\,\mathrm{d}y, \\ V_{\mathrm{d}}^2(f,-1) &= \langle f(x)f(-x)\rangle_{\mathrm{sc}} - \langle f\rangle_{\mathrm{sc}}^2. \end{aligned} \tag{30}$$

REMARK 2.8 (Cases of vanishing variance in Theorem 2.4). From the Chebyshev representation in Remark 2.6, we can easily identify the necessary and sufficient conditions for the processes ξ_{tr}, ξ_{d}, ξ_{od} to vanish:[4]

[4]Note that in case $\sigma = -1$ the condition on f differs for the three processes. For ξ_{od} and symmetric $A_{\mathrm{od}} = A_{\mathrm{od}}^t$ any odd function f results in $\xi_{\mathrm{od}}(f,A_{\mathrm{od}}) = 0$, while for ξ_{d} and ξ_{tr} only odd functions f orthogonal to $x\mapsto x$ with

(a) $\xi_{\mathrm{tr}}(f)=0$ if and only if f is of the form

$$f(x)=\mathbf{1}(\sigma=-1)\Big(\phi(x)-\frac{\langle\phi x\rangle_{1/\mathrm{sc}}}{2}x\Big)+b+\mathbf{1}(w_2=0)cx + \mathbf{1}\big(\kappa_4=-1-\sigma^2\big)\,dx^2 \tag{31}$$

for some odd function $\phi(-x)=-\phi(x)$ and $b,c,d\in\mathbf{R}$.

(b) For each fixed[5] N, we have $\xi_{\mathrm{d}}(f,\mathring{A}_{\mathrm{d}})=0$ if and only if either (i) $\mathring{A}_{\mathrm{d}}=0$, or (ii) f is of the form

$$f(x)=\mathbf{1}(\sigma=-1)\big(\phi(x)-\langle\phi x\rangle_{\mathrm{sc}}x\big)+b+\mathbf{1}(w_2=0)cx+\mathbf{1}\big(\kappa_4=-1-\sigma^2\big)\,dx^2 \tag{32}$$

for some odd function ϕ and $b,c,d\in\mathbf{R}$.

(c) For fixed N, we have $\xi_{\mathrm{od}}(f,A_{\mathrm{od}})=0$ if and only if either (i) $A_{\mathrm{od}}=0$, or (ii) $A_{\mathrm{od}}=-A_{\mathrm{od}}^t$, $\sigma=1$, or (iii) $A_{\mathrm{od}}=A_{\mathrm{od}}^t$, $\sigma=-1$ and $f(x)=b+\phi(x)$ for some odd function ϕ.

In Appendix A, we will comment on why these cases naturally yield vanishing variances.

2.3. *Computation of the expectations and variances in the mesoscopic regime.* Theorem 2.4 identified the expectations and the variances of the limiting processes ξ_{tr}, ξ_{d}, ξ_{od} in terms of the test function $f\in H_0^2(\mathbf{R})$. In case of mesoscopic test functions of the form $f(x)=g(N^a(x-E))$ with some scaling exponent $a>0$, reference energy $E\in[-2,2]$ and a compactly supported function $g\in H_0^2(\mathbf{R})$, we may compute the leading terms of the variances in terms of g. The result is different in the bulk ($|E|<2-\epsilon$ for any $\epsilon>0$ independent of N) and at the edge ($E=\pm 2$), therefore here we explicitly distinguish these two regimes. We note, however, that all error terms in our main Theorem 2.4 are valid uniformly in E, so this distinction is made here only in order to obtain simple limiting formulas. The variances can be conveniently expressed in terms of the L^2 and $\dot{H}^{1/2}$ inner products on real valued functions

$$\langle f,g\rangle_{L^2}:=\int_{\mathbf{R}}f(x)g(x)\,\mathrm{d}x,\qquad \langle f,g\rangle_{\dot{H}^{1/2}}:=\int_{\mathbf{R}^2}\frac{f(x)-f(y)}{x-y}\frac{g(x)-g(y)}{x-y}\,\mathrm{d}x\,\mathrm{d}y.$$

PROPOSITION 2.9 (Bulk scaling asymptotics). *Fix an $a\in(0,1)$ and an $\epsilon>0$ (independent of N) and recall $f(x)=g(N^a(x-E))$. Then for any $|E|\le 2-\epsilon$, the variances and expectation in Theorem 2.4 have the following large N asymptotic behavior*:

$$\begin{aligned}
V_{\mathrm{tr}}^1(f)&=\frac{\|g\|_{\dot{H}^{1/2}}^2}{4\pi^2}+\mathcal{O}(N^{-a}),\\
V_{\mathrm{tr}}^2(f,\sigma)&=\mathbf{1}(\sigma=1)\frac{\|g\|_{\dot{H}^{1/2}}^2}{4\pi^2}\\
&\quad+\mathbf{1}(\sigma=-1)\mathbf{1}(E=0)\frac{\langle g(x),g(-x)\rangle_{\dot{H}^{1/2}}}{4\pi^2}+\mathcal{O}(N^{-a}),\\
C_N V_{\mathrm{d}}^1(f)&=\|g\|_{L^2}^2+\mathcal{O}(N^{-a}),\\
C_N V_{\mathrm{d}}^2(f,\sigma)&=\mathbf{1}(\sigma=1)\|g\|_{L^2}^2+\mathbf{1}(\sigma=-1)\mathbf{1}(E=0)\big\langle g(x),g(-x)\big\rangle_{L^2}+\mathcal{O}(N^{-a}),\\
E_{\mathrm{tr}}(f,\sigma)&=\mathbf{1}(\sigma=-1)\mathbf{1}(E=0)\frac{g(0)}{2}+\mathcal{O}(N^{-a}).
\end{aligned}\tag{33}$$

respect to $\langle\cdot\rangle_{\mathrm{sc}}$ and $\langle\cdot\rangle_{1/\mathrm{sc}}$, respectively, result in ξ_{d}, ξ_{tr} to vanish. Thus, for example, $\xi_{\mathrm{od}}(x^3)=0$, $\xi_{\mathrm{d}}(x^3-2x)=0$ and $\xi_{\mathrm{tr}}(x^3-3x)=0$.

[5] Recall that the processes ξ_{d}, ξ_{od} depend on N through $\mathring{A}_{\mathrm{d}}$, A_{od}.

The implicit constants in the error terms depend only on a, ϵ, $\|g\|_{H_0^2}$, $|\operatorname{supp} g|$ *and on* C_p *from* (5) *and they are uniform in* E, σ *in a specific sense explained in Remark* 2.11.

PROPOSITION 2.10 (Edge scaling asymptotics). *For*[6] $E = 2$, *that is*, *at the right edge*, *and any* $0 < a < 2/3$ *the variances and expectation in Theorem* 2.4 *we have the scaling asymptotics*

$$
\begin{aligned}
V_{\mathrm{tr}}^1(f) &= \frac{\|g(-x^2)\|_{\dot{H}^{1/2}}^2}{8\pi^2} + \mathcal{O}(N^{-a/2}),\\
V_{\mathrm{tr}}^2(f,\sigma) &= \mathbf{1}(\sigma=1)\frac{\|g(-x^2)\|_{\dot{H}^{1/2}}^2}{8\pi^2} + \mathcal{O}(N^{-a/2}),\\
C_N V_{\mathrm{d}}^1(f) &= \frac{\|g(-x^2)x\|_{L^2}^2}{\pi} + \mathcal{O}(N^{-a/2}),\\
C_N V_{\mathrm{d}}^2(f,\sigma) &= \mathbf{1}(\sigma=1)\frac{\|g(-x^2)x\|_{L^2}^2}{\pi} + \mathcal{O}(N^{-a/2}),\\
E_{\mathrm{tr}}(f,\sigma) &= \mathbf{1}(\sigma=1)\frac{g(0)}{4} + \mathcal{O}(N^{-a/2}).
\end{aligned}
\tag{34}
$$

The implicit constants in the error terms depend only on a, $\|g\|_{H_0^2}$, $|\operatorname{supp} g|$ *and on* C_p *from* (5) *and on* σ *in a specific sense explained in Remark* 2.11.

REMARK 2.11. Our proof also gives uniformity of the dependence on the constants E, σ in the error terms in (33)–(34) in the following sense. In those formulas among (33)–(34) that contain $\mathbf{1}(\sigma = 1)$, the error is uniform in $\sigma \le 1 - \epsilon$ for any fixed $\epsilon > 0$ when $\sigma \ne 1$. Similarly, the presence of a factor $\mathbf{1}(\sigma = -1)$ in the formula comes with uniformity for any $\sigma \ge -1 + \epsilon$ whenever $\sigma \ne -1$. Finally, in terms with $\mathbf{1}(E = 0)$ in (33) we have uniformity for any $|E| \ge \epsilon$, whenever $E \ne 0$. In all other terms, our result is uniform for all $|E| \le 2 - \epsilon$. See also Remark 2.15.

REMARK 2.12. In contrast to the macroscopic scale, note that on the mesoscopic scale the limits in Propositions 2.9–2.10 are independent on κ_4 and w_2 and their σ-dependence is via a very simple characteristic function. This shows that the mesoscopic fluctuations are less sensitive to the details of the ensemble, in agreement with the general paradigm that more local statistics are more universal. In fact, for $\sigma > -1$ the appearance of $\mathbf{1}(\sigma = 1)$ in the variance V_{tr}^2, V_{d}^2 corresponds to a factor of 2 difference between real symmetric and complex Hermitian symmetry classes. Furthermore, for $\sigma = -1$, assuming $w_2 = 0$, we have $W = \mathrm{i}O$, where $O = -O^t$ is a real skew-symmetric matrix, in particular the spectrum of W is symmetric with respect to zero, that is, the eigenvalues around some energy E and $-E$ are strongly dependent. On the mesoscopic scale, this feature is relevant only for $E = 0$ and it changes the expectation and the variance. In particular, for antisymmetric test functions, $g(x) = -g(-x)$, we have $L_N(f, A) = 0$, and indeed, the variances in Proposition 2.9 add up to zero in this case.

Additionally, we prove that the linear statistics for test functions living on different scales are asymptotically independent. The proof of the following theorem follows by standard arguments completely analogous to the proof of Theorem 2.4 and is presented in [14], Section F.3.

[6]The case of the left edge, $E = -2$ is completely analogous upon replacing $-x^2$ by x^2 in the formulas (34).

THEOREM 2.13. *Let $\epsilon > 0$ and $E_1, E_2 \in [-2+\epsilon, 2-\epsilon]$, $0 \le a_1 \neq a_2 < 1$ and let $g_1, g_2 \in H_0^2(\mathbf{R})$ be compactly supported functions and set $f_i(x) := g_i(N^{a_i}(x-E_i))$. Then the limiting Gaussian processes $\xi_{\mathrm{tr}}(f_1), \xi_{\mathrm{tr}}(f_2)$ from Theorem 2.4 are asymptotically independent in the sense*

$$|\mathrm{Cov}(\xi_{\mathrm{tr}}(f_1), \xi_{\mathrm{tr}}(f_2))| \lesssim N^{-|a_1-a_2|}. \tag{35}$$

Similarly, for bounded deterministic matrices A_1, A_2 the processes $\xi_{\mathrm{d}}, \xi_{\mathrm{od}}$ are asymptotically independent in the sense

$$|\mathrm{Cov}(\xi_{\mathrm{d}}(f_1), \xi_{\mathrm{d}}(f_2))| + |\mathrm{Cov}(\xi_{\mathrm{od}}(f_1, A_1), \xi_{\mathrm{od}}(f_2, A_2))| \lesssim N^{-|a_1-a_2|/2}. \tag{36}$$

To make our presentation simpler, we stated this result only in the bulk, but our proof naturally yields the independence of linear statistics living on different scales uniformly in the spectrum. Moreover, the same argument also yields independence of linear statistics living on the same scale at distant energies, that is, for $a_1 = a_2 = a$ and $|E_1 - E_2| \gg N^{-a}$.

Theorem 2.13 together with Theorem 2.4 imply the asymptotic independence of linear statistics living on different scales $0 \le a_1 \neq a_2 < 1$ in the sense

$$|\mathrm{Cov}(L_N(f_1, I), L_N(f_2, I))| \lesssim N^{-|a_1-a_2|} + N^{(a_1-1)/2} + N^{(a_2-1)/2}, \tag{37}$$

and similarly for $\sqrt{C_N} L_N(f_i, \mathring{A}_{\mathrm{d}})$, $\sqrt{C_N} L_N(f_i, A_{\mathrm{od}})$. We note, however, that for large $|a_1 - a_2|$ the estimate on the covariance of linear statistics in (37) may be larger than that of the limiting processes in (35) owing to the error terms from Theorem 2.4.

2.4. *Related earlier results and miscellaneous remarks*. The linear eigenvalue statistics $\sum_i f(\lambda_i)$ have been extensively studied, and a CLT has been proven for macroscopic test functions as well as for mesoscopic test functions down to the optimal scale both in the bulk and at the edge, hence our results on $\xi_{\mathrm{tr}}(f)$ are not new, and we only listed them for completeness. More precisely, the explicit form of the variance $\mathbf{E}|\xi_{\mathrm{tr}}(f)|^2$ for macroscopic test functions in (14) exactly agrees with [39], equation (3.92), for $w_2 = 2/\beta$ and with [51], equation (1.10), for the case when $w_2 \neq 2/\beta$. Note that the parameter β, customary in random matrix theory distinguishing between the real symmetric ($\beta = 1$) and complex Hermitian ($\beta = 2$) symmetry classes, corresponds to $\beta = 2/(1+\sigma)$ with our notation in the cases $\sigma = 0, 1$.

For mesoscopic test functions, the variance $\mathbf{E}|\xi_{\mathrm{tr}}(f)|^2$ in (14) with (33) in the bulk and with (34) at the edge exactly agree with [36], equation (2.22), and [36], equation (2.23), [29], equation (2.6), respectively, in case of $\sigma = 0, 1$. Our formulas for general σ agree with the results in [28] for $\sigma \in (-1, 1]$; however, the final formula for the variance in case $\sigma = -1$ appears to be wrong in [28] (probably the error stems from [28], equation (6.25), overlooking that $|\widetilde{T}|$ is not far away from zero, in fact $|\widetilde{T}| \sim \eta$ in this case).

As far as the expectation (density of states) is concerned, the explicit formula for $\mathbf{E}\sum_i f(\lambda_i)$ in (22) with (23) exactly agrees with the formula given in [2], Theorem 1.1, for $\sigma \in \{0, 1\}$ and with [5], equation (1.4), for the general case. We also mention that for the Gaussian case explicit N-dependent formulas are obtained in [50] on the density of states by supersymmetric methods.

The joint linear statistics of eigenvalues and eigenvectors with observable $A \neq I$, that is, quantities $\mathrm{Tr}[f(W)A] = \sum_i f(\lambda_i)\langle u_i, Au_i\rangle$ are much less studied. For macroscopic test functions f, the variances $\mathbf{E}|\xi_{\mathrm{d}}(f, \mathring{A}_d)|^2$, $\mathbf{E}|\xi_{\mathrm{od}}(f, A_{\mathrm{od}})|^2$ in (14) exactly agree with [38], equation (4.16), equation (4.19), in the real symmetric case. For mesoscopic test functions f, the current paper achieves the first results on the limiting distribution of $\mathrm{Tr}[f(W)A]$, with $A \neq I$, in particular, explicit formulas for $\mathbf{E}|\xi_{\mathrm{d}}(f, \mathring{A}_d)|^2$ and $\mathbf{E}|\xi_{\mathrm{od}}(f, A_{\mathrm{od}})|^2$ in (15)–(16), with their limiting behavior in (33) and (34), are new.

REMARK 2.14. In (9), we assumed that $g \in H_0^2(\mathbf{R})$ is compactly supported to make the proof cleaner. The proof of the functional CLT (Theorem 2.4) on the macroscopic scale ($a = 0$) presented in [14], Appendix F, would work exactly in the same way if $f \in H^2(\mathbf{R})$. The only difference is that throughout the proof we have to replace f by its cut-off version, $f_\chi := f\chi$, with χ a smooth cut-off function that is equal to one on $[-5, 5]$ and equal to zero on $[-10, 10]^c$.

REMARK 2.15. The formulas in Propositions 2.9–2.10 indicate a somewhat different limiting expectation and variance when $\sigma = \pm 1$ in contrast to the $|\sigma| < 1$ case. With our methods, it is also possible to study the transitional regime, where $1 - |\sigma|$ vanishes as an N-power, as it was done for the tracial part in [28], but we refrained from doing so in order to keep the paper more transparent.

3. Local laws for multiple resolvents. Given a Wigner matrix W, we define its resolvent by $G(z) := (W - z)^{-1}$, with $z \in \mathbf{C} \setminus \mathbf{R}$. In this paper, we consider resolvents allowing the spectral parameter z to have a positive or negative imaginary part, in order to conveniently account for possible adjoints of the resolvent since $G(z)^* = G(\overline{z})$.

In this section, we prove local laws for one resolvent and for certain products of two or three resolvents that will be used as an input to prove the central limit theorem for resolvents in Section 4. These local laws are stated in Propositions 3.2–3.4. Additionally, in Lemma 3.6 we present an improvement for the bound of $\langle \boldsymbol{x}, \underline{GAWG}\boldsymbol{y}\rangle^2$ in (46), which we need only in a second moment sense. The main inputs for the proof of these local laws are the bounds in [16], Theorem 5.

As $N \to \infty$, the resolvent G becomes approximately deterministic (local laws). Its deterministic approximation is given by $m(z) = m_{\mathrm{sc}}(z)$, with $m_{\mathrm{sc}}(z)$ being the Stieltjes transform (12) of the semicircular law ρ_{sc} defined in (4). In particular, $m = m(z)$ is given by the unique solution of the quadratic equation

$$-\frac{1}{m} = z + m, \qquad \Im m(z)\Im z > 0. \tag{38}$$

Recall that the density $\rho(z)$ is defined as $\rho(z) := \pi^{-1}|\Im m(z)|$.

In order to formulate the local laws concisely, we introduce the commonly used notion of *stochastic domination*.

DEFINITION 3.1 (Stochastic Domination). If

$$X = \big(X^{(N)}(u) \mid N \in \mathbf{N}, u \in U^{(N)}\big) \quad \text{and} \quad Y = \big(Y^{(N)}(u) \mid N \in \mathbf{N}, u \in U^{(N)}\big)$$

are families of nonnegative random variables indexed by N, and possibly some parameter u, then we say that X is stochastically dominated by Y, if for all $\xi, D > 0$ we have

$$\sup_{u \in U^{(N)}} \mathbf{P}\big[X^{(N)}(u) > N^\xi Y^{(N)}(u)\big] \le N^{-D}$$

for large enough $N \ge N_0(\xi, D)$. In this case, we use the notation $X \prec Y$ and $X = \mathcal{O}_\prec(Y)$.

In addition to the $\mathcal{O}_\prec(\cdot)$ notation indicating a stochastic domination in the sense of arbitrary high moments, in this proof we introduce two related new notations, $\mathcal{O}_\prec^1(\cdot)$, $\mathcal{O}_\prec^2(\cdot)$, indicating domination only in first and second moment sense. More precisely, we write $X = \mathcal{O}_\prec^2(\psi)$ and $X = \mathcal{O}_\prec^1(\psi)$ if $\mathbf{E}|X|^2 \lesssim N^\xi \psi^2$ and $\mathbf{E}|X| \lesssim N^\xi \psi$, respectively, for any $\xi > 0$

and some deterministic ψ. We note that we trivially have the following product estimates:[7]

$$X = \mathcal{O}^2_\prec(\phi), Y = \mathcal{O}^2_\prec(\psi) \quad \Rightarrow \quad XY = \mathcal{O}^1_\prec(\phi\psi), \tag{39a}$$

$$X = \mathcal{O}^1_\prec(\phi), Y = \mathcal{O}_\prec(\psi) \quad \Rightarrow \quad XY = \mathcal{O}^1_\prec(\phi\psi), \tag{39b}$$

$$X = \mathcal{O}^2_\prec(\phi), Y = \mathcal{O}_\prec(\psi) \quad \Rightarrow \quad XY = \mathcal{O}^2_\prec(\phi\psi), \tag{39c}$$

so that by (39a), in particular, $X = \mathcal{O}^2_\prec(\psi)$ implies $X = \mathcal{O}^1_\prec(\psi)$.

We start with the statement of the local laws for single resolvents.

PROPOSITION 3.2 (Single G local laws). *Let $z \in \mathbf{C} \setminus \mathbf{R}$ with[8] $N\eta\rho \geq N^\delta$ for some $\delta > 0$, where we use notation $\eta := |\Im z|$, $\rho = \rho(z)$, $m = m_{\mathrm{sc}}(z)$. Then for any deterministic matrix A with $\|A\| \lesssim 1$ and $\langle A \rangle = 0$, we have the averaged local laws*

$$\big|\langle G - m\rangle\big| \prec \frac{1}{N\eta}, \qquad \big|\langle GA\rangle\big| \prec \frac{\sqrt{\rho}}{N\sqrt{\eta}}. \tag{40}$$

Additionally, for any deterministic vectors $\boldsymbol{x}$, $\boldsymbol{y}$ such that $\|\boldsymbol{x}\| + \|\boldsymbol{y}\| \lesssim 1$, we have the isotropic law

$$\big|\langle \boldsymbol{x}, (G - m)\boldsymbol{y}\rangle\big| \prec \sqrt{\frac{\rho}{N\eta}}. \tag{41}$$

The averaged and isotropic law for $G - m$ have been proven in [7, 26, 34]. The local law for $\langle GA \rangle$ in (40) is quite recently proven in [16], Theorem 3.

Next we state averaged and isotropic local laws for products of two resolvents.

PROPOSITION 3.3 (Local laws for two G's). *Fix any $\delta > 0$. Let $z_1, z_2 \in \mathbf{C} \setminus \mathbf{R}$ with $L := \min_i(\eta_i\rho_i) \geq N^\delta$ where $\eta_i := |\Im z_i|$, $\rho_i := \rho(z_i)$, $m_i := m_{\mathrm{sc}}(z_i)$. We introduce the notation $G_i := G(z_i)$ and $K := N\eta_*\rho^* \geq L$, $\eta_* := \eta_1 \wedge \eta_2$, $\rho^* := \rho_1 \vee \rho_2$. Then for any deterministic matrices A, A' with $\|A\| + \|A'\| \lesssim 1$ and $\langle A \rangle = \langle A' \rangle = 0$ we have the averaged local laws*[9]

$$\begin{aligned} \langle G_1G_2\rangle &= \frac{m_1m_2}{1 - m_1m_2} + \mathcal{O}_\prec\Big(\frac{1}{N\eta_1\eta_2}\Big), \\ \big\langle G_1AG_2A'\big\rangle &= m_1m_2\langle AA'\rangle + \mathcal{O}_\prec\Big(\frac{\rho^*}{\sqrt{K}}\Big), \end{aligned} \tag{42}$$

$$\big\langle G_1G_2^t\big\rangle = \frac{m_1m_2}{1 - \sigma m_1m_2} + \mathcal{O}_\prec\Big(\frac{\mathbf{1}(\sigma = \pm 1)}{N\eta_1\eta_2} + \mathbf{1}(|\sigma| < 1)\Big[\frac{1}{N\eta_*^2} \wedge \frac{\rho^*}{\sqrt{K}}\Big]\Big), \tag{43}$$

$$\big\langle G_1AG_2^tA'\big\rangle = m_1m_2\langle AA'\rangle + \mathcal{O}_\prec\Big(\frac{\rho^*}{\sqrt{K}}\Big). \tag{44}$$

We also have the following bound:

$$\big|\langle G_1G_2A\rangle\big| + \big|\langle G_1G_2^tA\rangle\big| = \mathcal{O}^2_\prec\Big(\frac{\sqrt{\rho_1\rho_2}}{\sqrt{NL\eta_1\eta_2}}\Big). \tag{45}$$

[7]In (39b)–(39c), we additionally need to assume that $|Y| \leq N^C$, for some large $C > 0$, which will hold in all our applications.

[8]Actually the tracial local law $|\langle G - m\rangle| \prec (N\eta)^{-1}$ holds true for any $\eta > N^{-100}$, and so does the isotropic local law $|\langle \boldsymbol{x}, (G - m)\boldsymbol{y}\rangle| \prec \sqrt{\rho/N\eta} + 1/N\eta$ with the additional error term $1/N\eta$.

[9]The second error term in (43) is uniform in σ as long as $|\sigma| \leq 1 - \epsilon'$ for any fixed $\epsilon' > 0$.

Moreover, for any deterministic vectors $\boldsymbol{x}, \boldsymbol{y}$ *such that* $\|\boldsymbol{x}\| + \|\boldsymbol{y}\| \lesssim 1$ *we have the isotropic laws*

$$\langle \boldsymbol{x}, G_1 G_2 \boldsymbol{y} \rangle = \frac{m_1 m_2}{1 - m_1 m_2} \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \mathcal{O}_\prec\Big(\frac{\sqrt{\rho^*}}{\sqrt{N\eta_*}\eta^*}\Big),$$

$$\big|\langle \boldsymbol{x}, G_1 A G_2 \boldsymbol{y} \rangle\big| \prec \sqrt{\frac{\rho^*}{\eta_*}}, \tag{46}$$

where $\eta^* := \eta_1 \vee \eta_2$.

Now we state averaged laws for certain products of three resolvents.

PROPOSITION 3.4 (Local laws for three G's). *Fix any* $\delta > 0$. *Let* $z_1, z_2 \in \mathbf{C} \setminus \mathbf{R}$ *with* $L := \min_i(\eta_i \rho_i) \geq N^\delta$ *where* $\eta_i := |\Im z_i|$, $\rho_i := \rho(z_i)$, $m_i := m_{\mathrm{sc}}(z_i)$. *We introduce the notation* $G_i := G(z_i)$ *and* $K := N\eta_* \rho^* \geq L$, $\eta_* := \eta_1 \wedge \eta_2$, $\rho^* := \rho_1 \vee \rho_2$. *Then for any deterministic matrices* A, A' *with* $\|A\| + \|A'\| \lesssim 1$ *and* $\langle A \rangle = \langle A' \rangle = 0$ *we have the averaged local laws*

$$\langle G_1 G_2^2 \rangle = \frac{m_1 m_2'}{(1 - m_1 m_2)^2} + \mathcal{O}_\prec\Big(\frac{\rho_*}{L\eta_1\eta_2}\Big), \tag{47}$$

$$\langle G_1 G_2 A G_1 A' \rangle = \frac{m_1^2 m_2 \langle AA' \rangle}{1 - m_1 m_2} + \mathcal{O}_\prec^2\Big(\frac{\sqrt{\rho_1\rho_2}}{\sqrt{L\eta_1\eta_2}}\Big), \tag{48}$$

$$\langle G_1 (G_2^t)^2 \rangle = \frac{m_1 m_2'}{(1 - \sigma m_1 m_2)^2} + \mathcal{O}_\prec\Big(\frac{\rho_1}{\sqrt{L}\eta_1\eta_2}\Big), \tag{49}$$

$$\langle G_1 G_2^t A G_1 A' \rangle = \frac{m_1^2 m_2 \langle AA' \rangle}{1 - \sigma m_1 m_2} + \mathcal{O}_\prec^2\Big(\frac{\sqrt{\rho_1\rho_2}}{\sqrt{L\eta_1\eta_2}}\Big). \tag{50}$$

Additionally, we have the following bound:

$$|\langle G_1 G_2^2 A \rangle| + |\langle G_1^t G_2^2 A \rangle| = \mathcal{O}_\prec^2\Big(\frac{\sqrt{\rho_*}}{L\sqrt{\eta_1}\eta_2}\Big). \tag{51}$$

The local laws and bounds in (42)–(46) and (47)–(51) all have the structure that the first term in the right-hand side is the explicit leading term. The error term in the right-hand side is smaller than the typical size of the leading term using $L \gg 1$, the fact that $|m| \sim 1$, $|m'| \sim \rho^{-1}$, and the bound

$$\Big|\frac{1}{1 - m_1 m_2}\Big| \lesssim \begin{Bmatrix} 1/\rho^*, & \mathrm{sgn}(\Im z_1) = \mathrm{sgn}(\Im z_2), \\ \sqrt{\rho_1\rho_2/\eta_1\eta_2}, & \text{else}, \end{Bmatrix} \lesssim \sqrt{\frac{\rho_1\rho_2}{\eta_1\eta_2}}, \tag{52}$$

which follow from elementary calculus. In the sequel, we will often use these local laws in their weaker form just as an upper bound for the left-hand side in terms of the upper estimate on the leading term on the right-hand side. For example, (42) together with (52) implies

$$|\langle G_1 G_2 \rangle| \prec \sqrt{\frac{\rho_1\rho_2}{\eta_1\eta_2}},$$

and similarly for all the other local laws.

For any given functions f, g of the Wigner matrix W we define the *renormalization* of the product $g(W)Wf(W)$ (denoted by underline) as follows:

$$\underline{g(W)Wf(W)} := g(W)Wf(W) - \widetilde{\mathbf{E}} g(W)\widetilde{W}(\partial_{\widetilde{W}} f)(W) - \widetilde{\mathbf{E}}(\partial_{\widetilde{W}} g)(W)\widetilde{W} f(W), \tag{53}$$

where $\partial_{\widetilde{W}} f(W)$ denotes the directional derivative of the function f in the direction $\widetilde{W}$ at the point W, and $\widetilde{W}$ is an independent copy of W. The definition is chosen such that it subtracts the variance term in the cumulant expansion, in particular if all entries of W were Gaussian then we had $\mathbf{E}\underline{g(W)Wf(W)} = 0$. Note that the definition (53) only makes sense if it is clear to which W the underline refers, that is, it would be ambiguous if $f(W) = W$. In our applications, however, each underlined term contains exactly a single W factor, and hence such ambiguities will not arise.

The key inputs for the proof of the local laws with two or three G's are strong bounds for renormalized products of the form $\langle \underline{WG_1B_1G_2 \dots G_lB_l} \rangle$. In Theorem 5 of our companion paper [16], we proved such estimates but they are in terms η_*, the minimal of all η's, that is, no distinction among different η's is made. To remedy this situation, in the following Theorem 3.5 we prove a generalization of [16], Theorem 5, which allows for the proof of the local laws for two and three G's with distinguished η-dependencies as stated above. This distinction is necessary when we insert these local laws into the Helffer–Sjöstrand formula that involves integrals of all spectral parameters. Furthermore, for a few specific terms we need a somewhat stronger bound than our general Theorem 3.5 gives, but we need them only in variance sense in contrast to the high probability bounds in Theorem 3.5. These specific bounds are listed separately in Lemma 3.6. The proof of Theorem 3.5 is presented in [14], Appendix D.1, and the proof of Lemma 3.6 in [14], Appendix D.2.

THEOREM 3.5. *Fix $\delta > 0$, let $l, n_1, \dots, n_l \in \mathbf{N}$, $z_{1,1}, \dots, z_{1,n_1}, z_{2,1}, \dots z_{l,n_l} \in \mathbf{C} \setminus \mathbf{R}$ and for $k \in [l]$, $j \in [n_k]$ let*

$$\mathcal{G}_k \in \{G_{k,1}G_{k,2} \cdots G_{k,n_k}, (G_{k,1}G_{k,2} \cdots G_{k,n_k})^t\}, \qquad G_{k,j} \in \{G(z_{k,j}), \Im G(z_{k,j})\}$$

and let B_k be deterministic $N \times N$ matrices, and $\boldsymbol{x}$, $\boldsymbol{y}$ be deterministic vectors with bounded norms $\|B_k\| \lesssim 1$, $\|\boldsymbol{x}\| + \|\boldsymbol{y}\| \lesssim 1$. Set

$$L := N \min_{k,i}(\eta_{k,i}\rho_{k,i}), \qquad \rho^* := \max_{k,i} \rho_{k,i}, \qquad \eta_* := \min_{k,i} \eta_{k,i}, \tag{54}$$

with $\eta_{k,i} := |\Im z_{k,i}|$, $\rho_{k,i} := \rho(z_{k,i}) = |\Im m(z_{k,i})|/\pi$ and assume $L \geq N^\delta$ and $\eta_* \lesssim 1$. Let $\mathfrak{a}$, $\mathfrak{t}$ denote disjoint sets of indices, $\mathfrak{a} \cap \mathfrak{t} = \varnothing$, such that for each $k \in \mathfrak{a}$ we have $\langle B_k \rangle = 0$, and for each $k \in \mathfrak{t}$ exactly one of $\mathcal{G}_k$, $\mathcal{G}_{k+1}$ is transposed, where in the averaged case and $k = l$ it is understood that $\mathcal{G}_{l+1} = \mathcal{G}_1$. Then with $a := |\mathfrak{a}|$, $t := |\mathfrak{t}|$, we have the following bounds:

(av1) *For $\mathfrak{a} = \mathfrak{t} = \varnothing$ we have*

$$\left|\langle \underline{W\mathcal{G}_1B_1\mathcal{G}_2B_2 \cdots \mathcal{G}_lB_l} \rangle\right| \prec \frac{\rho^*}{N\eta_*^l} \prod_{k \in [l]} \frac{\min_i \eta_{k,i}}{\eta_{k,1} \cdots \eta_{k,n_k}}. \tag{55}$$

(av2) *For $\mathfrak{a}, \mathfrak{t} \subset [l]$, $|\mathfrak{a} \cup \mathfrak{t}| \geq 1$, we have the bound*

$$\left|\langle \underline{W\mathcal{G}_1B_1\mathcal{G}_2B_2 \cdots \mathcal{G}_lB_l} \rangle\right| \prec \frac{(\sqrt{N}\eta_*)^{a+t}}{N\eta_*^l} \sqrt{\frac{\rho^*}{N\eta_*}} \prod_{k \in [l]} \frac{\min_i \eta_{k,i}}{\eta_{k,1} \cdots \eta_{k,n_k}}. \tag{56}$$

(iso) *For $\mathfrak{a}, \mathfrak{t} \subset [l-1]$ and for any $0 \leq j < l$, we have the bound*

$$\begin{aligned} &\left|\langle \boldsymbol{x}, \underline{\mathcal{G}_1B_1 \cdots \mathcal{G}_jB_jW\mathcal{G}_{j+1}B_{j+1} \cdots B_{l-1}\mathcal{G}_l}\boldsymbol{y} \rangle\right| \\ &\quad \prec \frac{(\sqrt{N}\eta_*)^{a+t}}{\eta_*^{l-1}} \sqrt{\frac{\rho^*}{N\eta_*}} \prod_{k \in [l]} \frac{\min_i \eta_{k,i}}{\eta_{k,1} \cdots \eta_{k,n_k}}, \end{aligned} \tag{57}$$

where the $j = 0$ case is understood as $\langle \boldsymbol{x}, \underline{W\mathcal{G}_1B_1 \cdots B_{l-1}\mathcal{G}_l}\boldsymbol{y} \rangle$.

Moreover, for any $\eta_* \geq 1$ *we have the bounds*

$$(58)\qquad \begin{aligned} \big|\langle \underline{W\mathcal{G}_1 B_1 \mathcal{G}_2 B_2 \cdots \mathcal{G}_l B_l}\rangle\big| &\prec \frac{1}{N\eta_*^l}\prod_{k\in[l]}\frac{\min_i \eta_{k,i}}{\eta_{k,1}\cdots\eta_{k,n_k}},\\ \big|\langle \boldsymbol{x}, \underline{\mathcal{G}_1 B_1 \cdots \mathcal{G}_j B_j W \mathcal{G}_{j+1} B_{j+1}\cdots B_{l-1}\mathcal{G}_l \boldsymbol{y}}\rangle\big| &\prec \frac{1}{N^{1/2}\eta_*^l}\prod_{k\in[l]}\frac{\min_i \eta_{k,i}}{\eta_{k,1}\cdots\eta_{k,n_k}}. \end{aligned}$$

LEMMA 3.6. *Let* $z, z_1, z_2 \in \mathbf{C}\setminus\mathbf{R}$ *with* $L := N\min_i(\eta_i\rho_i) \geq N^\delta$ *for some* $\delta > 0$ *and let* $G = G(z)$, $G_i = G(z_i)$. *Then, for any fixed deterministic vectors* $\boldsymbol{x}$, $\boldsymbol{y}$ *and matrix* A *with* $\langle A\rangle = 0$ *and* $\|A\| + \|\boldsymbol{x}\| + \|\boldsymbol{y}\| \lesssim 1$, *we have*

$$(59)\qquad \big|\langle \boldsymbol{x}, \underline{GAWG\boldsymbol{y}}\rangle\big| = \mathcal{O}_\prec^2\bigg(\frac{\rho}{N^{1/2}\eta}\bigg),$$

and

$$(60)\qquad \big|\langle \underline{WG_1G_2}A\rangle\big| + \big|\langle \underline{WG_1G_2^t}A\rangle\big| = \mathcal{O}_\prec^2\bigg(\frac{\rho_*^{1/2}}{N\eta_*^{1/2}}\frac{1}{\sqrt{\eta_1\eta_2}}\bigg),$$

$$(61)\qquad \big|\langle \underline{WG_1G_1G_2}A\rangle\big| + \big|\langle \underline{WG_1G_1G_2^t}A\rangle\big| = \mathcal{O}_\prec^2\bigg(\frac{\rho_*^{1/2}}{N\eta_*^{1/2}}\frac{1}{\sqrt{\eta_1\eta_2}}\frac{1}{\sqrt{\eta_*\eta_1}}\bigg),$$

$$(62)\qquad \big|\langle \underline{WG_1G_2AG_1}A\rangle\big| + \big|\langle \underline{WG_1G_2^tAG_1}A\rangle\big| = \mathcal{O}_\prec^2\bigg(\frac{\rho_*^{1/2}}{N^{1/2}\eta_*^{1/2}}\frac{1}{\sqrt{\eta_1\eta_2}}\bigg).$$

Notice that the bound in (59) is better by a factor $\sqrt{\rho/N\eta}$ compared to (57). The bounds (60)–(62) improve upon (56) in two aspects: First, they depend on ρ_* rather than ρ^*, and second, in the cases including transposes the bounds distinguish different η's (note that in Theorem 3.5 it is not allowed to have both G, G^t within one $\mathcal{G}$-block, hence the bound is purely in terms η_*).

We conclude this section with the proof of Propositions 3.3–3.4.

PROOF OF PROPOSITION 3.3. The local laws for $\langle G_1AG_2A'\rangle$, $\langle G_1AG_2^tA'\rangle$, in (42), (44), respectively, and the bound for G_1AG_2 in (46) follow by [16], Proposition 2, together with [16], Theorem 4. The bounds in (45) follow by exactly the same proof of [16], equation (22), but using the new bound (60) instead of [16], equation (62), for the underlined term. Also the local law for $\langle G_1G_2^t\rangle$ with error term $\rho^*K^{-1/2}$ for $|\sigma| < 1$ in (43) follows by [16], Theorem 4, Proposition 2. Hence, in order to conclude the proof of Proposition 3.3 we are left with the averaged and isotropic law for G_1G_2 in (42) and (46), respectively, and with the proof of the remaining cases for the local law for $\langle G_1G_2^t\rangle$ in (43).

We first consider the local laws that involve no transposes, then at the end of the proof of Proposition 3.3 we explain the necessary changes when the transposes are considered.

By the self consistent equation for m in (38), and by $G(W - z) = I$, we have

$$(63)\qquad G = m - mWG - m\langle G\rangle G + m\langle G - m\rangle G.$$

As a special case of (53), we have that

$$(64)\qquad \underline{WG} = WG + \langle G\rangle G + \sigma\frac{G^tG}{N} + \frac{\widetilde{w_2}}{N}\operatorname{diag}(G)G,$$

where for any matrix R in this section we let $\mathrm{diag}(R)$ denote the matrix of its diagonal that was denoted by R_{d} earlier. We recall the parameters σ, $\widetilde{w_2}$ from (13). Then by (63) and and (64), it follows that

$$G = m - m\underline{WG} + \frac{m\sigma}{N}G^tG + m\frac{\widetilde{w_2}}{N}\mathrm{diag}(G)G + m\langle G - m\rangle G. \tag{65}$$

We now start writing the equation for generic products of two resolvents $G_1B_1G_2B_2$, where $G_i = (W - z_i)^{-1}$ and B_1, B_2 are deterministic matrices. Using the equation (65) for G_1B_1 and writing $G_2 = m_2 + (G_2 - m_2)$, we obtain

$$\begin{aligned} G_1B_1G_2B_2 = m_1m_2B_1B_2 &+ m_1B_1(G_2 - m_2)B_2 \\ &- m_1\underline{WG_1B_1G_2}B_2 + m_1\langle G_1B_1G_2\rangle G_2B_2 \\ &+ m_1\langle G_1 - m_1\rangle G_1B_1G_2B_2 + \frac{m_1\sigma}{N}G_1^tG_1B_1G_2B_2 \\ &+ \frac{m_1\sigma}{N}(G_1B_1G_2)^tG_2B_1 \\ &+ \frac{m_1\widetilde{w_2}}{N}\mathrm{diag}(G_1)G_1B_1G_2B_2 + \frac{m_1\widetilde{w_2}}{N}\mathrm{diag}(G_1B_1G_2)G_2B_2, \end{aligned} \tag{66}$$

where we used that

$$\begin{aligned} \underline{WG_1B_1G_2} = \underline{WG_1}B_1G_2 &+ \langle G_1B_1G_2\rangle G_2 + \frac{\sigma}{N}(G_1B_1G_2)^tG_2 \\ &+ \frac{\widetilde{w_2}}{N}\mathrm{diag}(G_1B_1G_2)G_2, \end{aligned} \tag{67}$$

with $\underline{WG_1}$ from (64). The identity in (67) follows by the definition of underline in (53).

PROOF OF THE LOCAL LAW FOR $\langle G_1G_2\rangle$ IN (42). We divide the proof of this local law into two cases: (i) $\Im z_1\Im z_2 < 0$, (ii) $\Im z_1\Im z_2 > 0$. The difference in these two cases is that in (ii) the stability factor $1 - m_1m_2$ is bounded from below by ρ^*, while in case (i) the stability factor is bounded from below only by η^* and so it is not affordable to invert it.

We start with $\Im z_1\Im z_2 < 0$, in this case we can use resolvent identity and the local law $|\langle G_i - m_i\rangle| \prec (N\eta_i)^{-1}$ from (40):

$$\begin{aligned} \langle G_1G_2\rangle &= \frac{G_1 - G_2}{z_1 - z_2} \\ &= \frac{m_1 - m_2}{z_1 - z_2} + \mathcal{O}_\prec\left(\frac{1}{N\eta_*|z_1 - z_2|}\right) \\ &= \frac{m_1m_2}{1 - m_1m_2} + \mathcal{O}_\prec\left(\frac{1}{N\eta_*|z_1 - z_2|}\right), \end{aligned} \tag{68}$$

where we used that the self-consistent equation (38) for m_1, m_2 in the third equality. This concludes the proof of the local law for $\langle G_1G_2\rangle$ when $\Im z_1\Im z_2 < 0$.

We now consider the case $\Im z_1\Im z_2 > 0$. Choosing $B_1 = B_2 = I$ in (66), and using $|\langle G_i - m_i\rangle| \prec (N\eta_i)^{-1}$, we find that

$$\begin{aligned} &\left[1 - m_1m_2 + \mathcal{O}_\prec\left(\frac{1}{N\eta_*}\right)\right]\langle G_1G_2\rangle \\ &= m_1m_2 - m_1\langle \underline{WG_1G_2}\rangle + m_1\langle (G_2 - m_2)\rangle \\ &\quad + \frac{m_1\sigma}{N}\langle G_1^tG_1G_2\rangle + \frac{m_1\sigma}{N}\langle (G_1G_2)^tG_2\rangle \\ &\quad + \frac{m_1\widetilde{w_2}}{N}\langle \mathrm{diag}(G_1)G_1G_2\rangle + \frac{m_1\widetilde{w_2}}{N}\langle \mathrm{diag}(G_1G_2)G_2\rangle. \end{aligned} \tag{69}$$

Using a Schwarz inequality, we readily conclude that

$$\frac{1}{N}|\langle G_1^t G_1 G_2\rangle| \le \frac{1}{N}\langle G_1 G_1^*\rangle^{1/2}\langle G_1 G_2 G_2^* G_1^*\rangle^{1/2} \prec \frac{\rho_1}{N\eta_1\eta_2}, \tag{70}$$

where we used that $\langle \Im G_i\rangle \prec \rho_i$, that $\eta_1 G_1 G_1^* = \Im G_1$ by Ward identity, that $|\langle G_1 G_2 G_2^* G_1^*\rangle| \le \|G_2 G_2^*\|\langle G_1 G_1^*\rangle$, and that $\|G_i\| \le \eta_i^{-1}$. We also prove that $|\langle (G_1 G_2)^t G_2\rangle| \prec \rho_2(N\eta_1\eta_2)^{-1}$ using exactly the same computations. Additionally, we get that

$$\frac{1}{N}|\langle \operatorname{diag}(G_1)G_1G_2\rangle| = \left|\frac{1}{N^2}\sum_i (G_1)_{ii}(G_1G_2)_{ii}\right| \prec \frac{\sqrt{\rho_1\rho_2}}{N\sqrt{\eta_1\eta_2}}, \tag{71}$$

where we used that $|G_{ii}| \prec 1$, and that $|(G_1G_2)_{ii}| \prec \sqrt{\rho_1\rho_2/(\eta_1\eta_2)}$ by a Schwarz inequality and Ward identity. The bound for $|\langle \operatorname{diag}(G_1G_2)G_2\rangle|$ is completely analogous and so omitted. Combining (69) with (70)–(71) and using that $|\langle G_2 - m_2\rangle| \prec (N\eta_2)^{-1}$, we finally conclude that

$$\langle G_1G_2\rangle = \frac{m_1m_2}{1-m_1m_2} - \frac{m_1}{1-m_1m_2}\langle \underline{WG_1G_2}\rangle + \mathcal{O}_\prec\left(\frac{1}{N\eta_1\eta_2} + \frac{1}{N\eta_*\rho^*}\right),$$

where we used that by easy computations we have $|1-m_1m_2| \ge \rho^*$ and that $\rho^* \gg \frac{1}{N\eta_*}$, by $K = N\eta_*\rho^* \gg 1$, to divide through the multiplicative factor in the left-hand side of (69). Finally, using that $|\langle \underline{WG_1G_2}\rangle| \prec \rho^*(N\eta_1\eta_2)^{-1}$ by (55), $|1-m_1m_2| \ge \rho^*$ once again, and that $(\rho^*)^2 \gtrsim \eta^*$, $\eta_*\eta^* = \eta_1\eta_2$, we conclude that

$$\langle G_1G_2\rangle = \frac{m_1m_2}{1-m_1m_2} + \mathcal{O}_\prec\left(\frac{1}{N\eta_1\eta_2}\right). \tag{72}$$

□

PROOF OF THE LOCAL LAW FOR $\langle \boldsymbol{x}, G_1G_2\boldsymbol{y}\rangle$ IN (73). The proof of the isotropic law for G_1G_2 is very similar to the proof of the averaged law above, hence we explain only the minor differences. Similar to the averaged local law, the case $\Im z_1 \Im z_2 < 0$ trivially follows by resolvent identity. In the opposite case, choosing $B_1 = B_2 = I$ in (66), and that $|\langle G_i - m_i\rangle| \prec (N\eta_i)^{-1}$, we find that

$$\begin{aligned}\left[1+\mathcal{O}_\prec\left(\frac{1}{N\eta_*}\right)\right]&\langle \boldsymbol{x}, G_1G_2\boldsymbol{y}\rangle \\ &= m_1m_2\langle \boldsymbol{x},\boldsymbol{y}\rangle - m_1\langle \boldsymbol{x}, \underline{WG_1G_2}\boldsymbol{y}\rangle + m_1\langle G_1G_2\rangle\langle \boldsymbol{x}, G_2\boldsymbol{y}\rangle \\ &\quad + \mathcal{O}_\prec\left(\frac{\rho^*}{N\eta_1\eta_2}\right),\end{aligned} \tag{73}$$

where we used that the terms with a prefactor σ or $w_2 - 1 - \sigma$ can be estimated by $N^{-1}\rho^*(\eta_1\eta_2)^{-1}$ using a Schwarz inequality similarl to (70)–(71). Then using that

$$\langle G_1G_2\rangle = \frac{m_1m_2}{1-m_1m_2} + \mathcal{O}_\prec\left(\frac{1}{N\eta_1\eta_2}\right),$$

by (72), and that $|\langle \boldsymbol{x}, \underline{WG_1G_2}\boldsymbol{y}\rangle| \prec \sqrt{\rho^*}(N\eta_*)^{-1/2}(\eta^*)^{-1}$ by Theorem 3.5, we finally conclude that

$$\langle \boldsymbol{x}, G_1G_2\boldsymbol{y}\rangle = \frac{m_1m_2}{1-m_1m_2}\langle \boldsymbol{x},\boldsymbol{y}\rangle + \mathcal{O}_\prec\left(\frac{\sqrt{\rho^*}}{\sqrt{N\eta_*}\eta^*} + \frac{1}{N\eta_*\rho^*}\right).$$

□

In order to conclude the proof of Proposition 3.3, we are left with considering transposes.

PROOF OF THE LOCAL LAW FOR $\langle G_1G_2^t\rangle$ IN (43). The proof of this local law is divided into three cases: (i) $\sigma = 1$, (ii) $\sigma = -1$, (iii) $|\sigma| < 1$. The main difference compared to the proof of $\langle G_1 G_2\rangle$ is that the two body stability factor is now given by $1-\sigma m_1 m_2$ instead of $1-m_1m_2$.

For $\sigma=1$, there is nothing else to prove since in this case W is real symmetric and so $G_2^t = G_2$.

The proof of the local law for $|\sigma|<1$ is completely analogous to the proof of (72), modulo the bound for the underline term that is now given by $|\langle \underline{WG_1G_2^t}\rangle| \prec \rho^*(N\eta_*^2)^{-1}$, since the only thing we used in this proof is that the stability factor $1-m_1m_2$ is bounded from below by ρ^*. This is also the case for $1-\sigma m_1 m_2$ when $|\sigma|<1$, since $|1-\sigma m_1m_2|\ge 1-|\sigma|$.

We are now left with the case $\sigma=-1$, when the stability factor is given by $1+m_1m_2$. Note that when $\sigma=-1$ we can write $W=D+\mathrm{i}O$ with D being a diagonal matrix and O being a real skew-symmetric matrix, that is, $O^t=-O$. If $D=0$, and either $\Im z_1\Im z_2>0$ or $\Im z_1\Im z_2<0$ and $|z_1+z_2|\ge\eta^*$, using the notation $R(z_i):=(\mathrm{i}O-z_i)^{-1}$ and that $R(z_i)^t = -R(-z_i)$, by resolvent identity we conclude

$$
\begin{aligned}
\langle G(z_1)G(z_2)^t\rangle &= -\langle R(z_1)R(-z_2)\rangle \\
&= -\frac{\langle R(z_1)\rangle - \langle R(-z_2)\rangle}{z_1+z_2} \\
&= \frac{m_1m_2}{1+m_1m_2} + \mathcal{O}_\prec\left(\frac{1}{N\eta_*|z_1+z_2|}\right),
\end{aligned}
\tag{74}
$$

where we used that $m(-z_i)=-m(z_i)$, and the local law for R_i, that holds even for Wigner matrices with zero diagonal. For $\Im z_1\Im z_2<0$ and $|z_1+z_2|<\eta^*$, using that $R(z_2)^t = -R(-z_2)$ we proceed exactly as in the proof of the local law for $\langle G_1G_2\rangle$ above in case (69)–(72). This gives the local law for $\langle G_1G_2^t\rangle$ in (43). In order to conclude the proof, we are now left only with the case $D\neq 0$. In this case, we use the following lemma whose proof is postponed to [14], Appendix B.

LEMMA 3.7. *Fix $\epsilon>0$. Let $W=D+\mathrm{i}O$ be a Wigner matrix with D being diagonal and O real skew-symmetric. Denote $G_i=(W-z_i)^{-1}$ and $R_i=(\mathrm{i}O-z_i)^{-1}$, with $z_1,z_2\in\mathbf{C}\setminus\mathbf{R}$ such that $\eta_*:=|\Im z_1|\wedge|\Im z_2|\ge N^{-1+\epsilon}$ and $\eta^*:=|\Im z_1|\vee|\Im z_2|$, then for $\sigma=-1$ it holds*

$$
\langle G_1G_2^t\rangle = \langle R_1R_2^t\rangle + \mathcal{O}_\prec\left(\frac{1}{N\eta_*}\left[\frac{1}{|z_1+z_2|}\wedge\frac{1}{\eta^*}\right]\right).
\tag{75}
$$

Moreover, we also have

$$
\langle G_1^tG_2^2\rangle = \langle R_1^tR_2^2\rangle + \mathcal{O}_\prec\left(\frac{1}{N\eta_*|\Im z_2|}\left[\frac{1}{|z_1+z_2|}\wedge\frac{1}{\eta^*}\right]\right).
\tag{76}
$$

In Lemma 3.7, we stated the result for three G's as well (76) even if not needed for the proof of the local law of $\langle G_1G_2^t\rangle$. We will use (76) later in (C.3).

Finally, combining (75) with (74), we conclude the proof of the local law for $\langle G_1G_2^t\rangle$. □

This concludes the proof of Proposition 3.3, modulo the proof of Lemma 3.7, which is postponed to [14], Appendix B. □

We conclude this section with the proof of the local laws for certain products of three resolvents. We will prove the estimates without transposed resolvents, the analogous results with transposes are proven in [14], Appendix C.

PROOF OF PROPOSITION 3.4. We start writing the equation for general products of three different resolvents G_1, G_2, G_3 and deterministic matrices B_1, B_2, B_3:

$$
\begin{aligned}
G_1B_1G_2B_2G_3B_3 = {} & m_1B_1G_2B_2G_3B_3 - m_1\underline{WG_1B_1G_2B_2G_3}B_3 \\
& + m_1\langle G_1B_1G_2\rangle G_2B_2G_3B_3 \\
& + m_1\langle G_1 - m_1\rangle G_1B_1G_2B_2G_3B_3 \\
& + m_1\langle G_1B_1G_2B_2G_3\rangle G_3B_3 \\
& + \frac{m_1\sigma}{N}G_1^tG_1B_1G_2B_2G_3B_3 + \frac{m_1\sigma}{N}(G_1B_1G_2)^tG_2B_2G_3B_3 \\
& + \frac{m_1\sigma}{N}(G_1B_1G_2B_2G_3)^tG_3B_3 \\
& + \frac{m_1\widetilde{w_2}}{N}\operatorname{diag}(G_1)G_1B_1G_2B_2G_3B_3 \\
& + \frac{m_1\widetilde{w_2}}{N}\operatorname{diag}(G_1B_1G_2)G_2B_2G_3B_3 \\
& + \frac{m_1\widetilde{w_2}}{N}\operatorname{diag}(G_1B_1G_2B_2G_3)G_3B_3,
\end{aligned}
\tag{77}
$$

where we used that

$$
\begin{aligned}
\underline{WG_1B_1G_2B_2G_3} = {} & \underline{WG_1B_1G_2}B_2G_3 + \langle G_1B_1G_2B_2G_3\rangle G_3 \\
& + \frac{\sigma}{N}(G_1B_1G_2B_2G_3)^tG_3 \\
& + \frac{\widetilde{w_2}}{N}\operatorname{diag}(G_1B_1G_2B_2G_3)G_3,
\end{aligned}
\tag{78}
$$

with $\underline{WG_1B_1G_2}$ from (67). The identity in (78) follows by the definition of the renormalization (denoted by underline) in (53).

PROOF OF THE LOCAL LAW FOR $\langle G_1G_2^2\rangle$ IN (47). Similar to the proof of the local law for $\langle G_1G_2\rangle$, the proof of the local law for $\langle G_1G_2^2\rangle$ is divided into two cases: (i) $\Im z_1\Im z_2 < 0$ or $\Im z_1\Im z_2 > 0$ and $|z_1 - z_2| \geq \eta^*$ (ii) $\Im z_1\Im z_2 > 0$ and $|z_1 - z_2| < \eta^*$, where we recall that $\eta^* := \eta_1 \vee \eta_2$, $\eta_i := |\Im z_i|$.

Similar to (68), if either $\Im z_1\Im z_2 < 0$ or $\Im z_1\Im z_2 > 0$ and $|z_1 - z_2| \gtrsim \eta^*$ we use the resolvent identity twice to get

$$
\begin{aligned}
\langle G_1G_2^2\rangle & = \frac{\langle G_2^2\rangle - \langle G_1G_2\rangle}{z_2 - z_1} \\
& = \frac{m_2'}{z_2 - z_1} + \frac{\langle G_1\rangle - \langle G_2\rangle}{(z_1 - z_2)^2} + \mathcal{O}_\prec\left(\frac{1}{N\eta_2^2|z_1 - z_2|}\right) \\
& = \frac{m_2'}{z_2 - z_1} + \frac{m_1 - m_2}{(z_1 - z_2)^2} + \mathcal{O}_\prec\left(\frac{1}{N\eta_2^2|z_1 - z_2|} + \frac{1}{N\eta_*|z_1 - z_2|^2}\right) \\
& = \frac{m_1m_2'}{(1 - m_1m_2)^2} + \mathcal{O}_\prec\left(\frac{\rho_*}{L\eta_1\eta_2}\right),
\end{aligned}
\tag{79}
$$

where in the first line we used the local law for $\langle G_2^2\rangle$ in (42), and the identity $m_2' = m_2^2(1 - m_2^2)^{-1}$, and to go from the second to the third line we again used the equation of m_1, m_2. We remark that to estimate the error terms to go from the second to the third line we also used (80) below. The proof of this bound is postponed to [14], Appendix B.

LEMMA 3.8. *Let $z_1, z_2 \in \mathbf{C}\setminus\mathbf{R}$ such that $|z_1-z_2|\gtrsim \eta^*$, then it holds*

$$\frac{1}{N\eta_*^2|z_1-z_2|}\lesssim \frac{\rho_*}{L\eta_1\eta_2}, \tag{80}$$

where $\eta_i=|\Im z_i|$, $\eta_*=\eta_1\wedge\eta_2$, $\rho_*=\rho_1\wedge\rho_2$, $\eta^*=\eta_1\vee\eta_2$.

Next we consider the last remaining case $\Im z_1\Im z_2>0$ and $|z_1-z_2|<\eta^*$. By (77) with $B_1=B_2=B_3=I$ and $G_3=G_2$, we have that

$$\begin{aligned}(1-m_1m_2)\langle G_1G_2^2\rangle &= m_1\langle G_2^2\rangle - m_1\langle \underline{WG_1G_2G_2}\rangle \\ &\quad + m_1\big[\langle G_1-m_1\rangle+\langle G_2-m_2\rangle\big]\langle G_1G_2^2\rangle \\ &\quad + m_1\langle G_1G_2\rangle\langle G_2^2\rangle + \mathcal{O}_\prec\Big(\frac{\rho^*}{N\eta_2^2\eta_1}\Big) \\ &= \frac{m_1m_2'}{1-m_1m_2}+\mathcal{O}_\prec\Big(\frac{\rho^*}{N\eta_2^2\eta_1}\Big),\end{aligned}\tag{81}$$

where to go from the second to the third line we used the local laws for $\langle G_1G_2\rangle$ and $\langle G_2^2\rangle$ in (42), together with (52), and we also used that $|\langle G_i-m_i\rangle|\prec(N\eta_i)^{-1}$ by the first local law in (40), and that $|\langle \underline{WG_1G_2^2}\rangle|\prec\rho^*(N\eta_2\eta_1)^{-1}$ by (55). We remark that to go from (77) to (81) we used that all the terms with a prefactor N^{-1} in (77) are bounded by $\rho^*(N\eta_2^2\eta_1)^{-1}$. To make this clearer, we show this bound for two representative terms:

$$\begin{aligned}\frac{1}{N}\big|\langle G_1^tG_1G_2G_2\rangle\big| &\le \frac{1}{N}\langle G_1G_1^*\rangle^{1/2}\langle G_1G_2G_2G_2^*G_2^*G_1^*\rangle^{1/2}\prec\frac{\rho^*}{N\eta_1\eta_2^2},\\ \frac{1}{N}\big|\langle \mathrm{diag}(G_1)G_1G_2G_2\rangle\big| &= \Big|\frac{1}{N}\sum_i (G_1)_{ii}(G_1G_2G_2)_{ii}\Big|\prec\frac{\rho^*}{N\eta_1^{1/2}\eta_2^{3/2}},\end{aligned}$$

where we used a Schwarz inequality, the norm bound $\|G_2G_2G_2^*G_2^*\|\le\eta_2^{-4}$, and that $|(G_1)_{ii}|\prec 1$, $|(G_1G_2G_2)_{ii}|\prec\rho^*\eta_1^{-1/2}\eta_2^{-3/2}$.

Note that the error term in the right-hand side of (81) is smaller than our goal $\rho_*(L\eta_1\eta_2)^{-1}$ in (47), since for $|z_1-z_2|<\eta^*$ we have that

$$\frac{1}{N\eta_2^2\eta_1}\lesssim\frac{\rho_*}{L\eta_1\eta_2}. \tag{82}$$

This concludes the proof of the local law for $\langle G^2G'\rangle$. □

PROOF OF THE LOCAL LAW FOR $\langle G_1G_2AG_1A'\rangle$ IN (48). Consider the equation in (77) for $G_3=G_1$, and $B_1=I$, $B_2=A$, $B_3=A'$, with $\langle A\rangle=\langle A'\rangle=0$. Before proceeding with writing the equation for $\langle G_1G_2AG_1A'\rangle$, we bound two representative terms with a prefactor N^{-1} in (77):

$$\begin{aligned}\big|\langle G_1^tG_1G_2AG_1A'\rangle\big| &\le \langle G_1G_1^*\rangle^{1/2}\langle G_2AG_1A'G_1^t(G_1^t)^*A'G_1^*AG_2^*\rangle^{1/2}\\ &\prec\sqrt{\frac{\rho_1}{\eta_1}}\frac{1}{\sqrt{\eta_1\eta_2}}\langle\Im G_2AG_1A'\Im G_1^tA'G_1^*A\rangle^{1/2}\prec\frac{\sqrt{N}\rho_1\sqrt{\rho_2}}{\sqrt{\eta_2}\eta_1},\\ \big|\langle\mathrm{diag}(G_1)G_1G_2AG_1A'\rangle\big| &\prec\frac{1}{N}\sum_a\big|(G_1)_{aa}(G_1G_2AG_1A')_{aa}\big|\prec\sqrt{\frac{N\rho_1\rho_2}{\eta_1\eta_2}},\end{aligned}\tag{83}$$

where in the first estimate we used [16], Lemma 5, to bound

$$\left|\langle \Im G_2 A G_1 A' \Im G_1 A' G_1^* A\rangle\right| \prec N\rho_1\rho_2,$$

and in the second estimate we used that

$$\left|\langle \boldsymbol{x}, G_1 G_2 A G_1 \boldsymbol{y}\rangle\right| \le \langle \boldsymbol{x}, G_1 G_1^* \boldsymbol{x}\rangle^{1/2}\langle \boldsymbol{y}, G_1^* A G_2^* G_2 A G_1 \boldsymbol{y}\rangle^{1/2} \prec \sqrt{\frac{N\rho_1\rho_2}{\eta_1\eta_2}},$$

by [16], Lemma 5, again. The bound of all the other terms with a prefactor N^{-1} is analogous and so omitted. Then, by (83) and (77), we conclude that

$$\begin{aligned}\left[1+\mathcal{O}_\prec\left(\frac{1}{N\eta_*}\right)\right]\langle G_1G_2AG_1A'\rangle &= m_1\langle G_2AG_1A'\rangle - m_1\langle \underline{WG_1G_2AG_1}A'\rangle \\ &\quad + m_1\langle G_1G_2\rangle\langle G_2AG_1A'\rangle \\ &\quad + m_1\langle G_1G_2AG_1\rangle\langle G_1A'\rangle + \mathcal{O}_\prec\left(\frac{\sqrt{\rho_1\rho_2}}{\sqrt{L\eta_1\eta_2}}\right) \\ &= \frac{m_1^2m_2}{1-m_1m_2}\langle AA'\rangle - m_1\langle \underline{WG_1G_2AG_1}A'\rangle \\ &\quad + \mathcal{O}_\prec\left(\frac{\sqrt{\rho_1\rho_2}}{\sqrt{L\eta_1\eta_2}}\right),\end{aligned} \tag{84}$$

where we used that

$$\begin{aligned}\langle G_2AG_1A'\rangle &= m_1m_2\langle AA'\rangle + \mathcal{O}_\prec\left(\frac{\rho^*}{\sqrt{L}}\right), \\ \langle G_1G_2\rangle &= \frac{m_1m_2}{1-m_1m_2} + \mathcal{O}_\prec\left(\frac{1}{N\eta_1\eta_2}\right),\end{aligned} \tag{85}$$

and

$$\left|\langle G_1A'\rangle\right| \prec \frac{\sqrt{\rho_1}}{N\sqrt{\eta_1}}. \tag{86}$$

The local laws in (85) follow by (42), while the bound in (86) follows by (40).

Finally, using the bound $|\langle \underline{WG_1G_2AG_1}A\rangle| \prec \rho^* K^{-1/2}(\eta^*)^{-1}$ from [16], Theorem 5, we conclude that

$$\langle G_1G_2AG_1A'\rangle = \frac{m_1^2m_2}{1-m_1m_2}\langle AA'\rangle + \mathcal{O}_\prec\left(\frac{\rho_*}{\sqrt{L\eta_1\eta_2}}\right), \tag{87}$$

where we recall that $K = N\eta_*\rho^*$, and $L = N(\rho_1\eta_1 \wedge \rho_2\eta_2)$. This concludes the proof of the local law (48) for $\langle G_1G_2AG_1A'\rangle$. □

PROOF OF THE BOUND FOR $\langle G_1G_2G_2A\rangle$ IN (51). Consider the equation in (77) for $G_3 = G_2$ and $B_1 = B_2 = I$, $B_3 = A$, with $\langle A\rangle = 0$. Proceeding similar to (83) to estimate the error terms with a prefactor N^{-1}, we conclude that

$$\begin{aligned}\langle G_1G_2G_2A\rangle &= m_1\langle G_2G_2A\rangle - m_1\langle \underline{WG_1G_2G_2}A\rangle + m_1\langle G_1-m_1\rangle\langle G_1G_2G_2A\rangle \\ &\quad + m_1\langle G_1G_2\rangle\langle G_2G_2A\rangle + m_1\langle G_1G_2G_2\rangle\langle G_2A\rangle + \mathcal{O}_\prec\left(\frac{\sqrt{\rho_*}}{L\sqrt{\eta_1}\eta_2}\right) \\ &= \mathcal{O}_\prec^2\left(\frac{\sqrt{\rho_*}}{L\sqrt{\eta_1}\eta_2} + \frac{\rho^*\sqrt{\rho_*}}{K\sqrt{\eta_1}\eta_2}\right) \\ &= \mathcal{O}_\prec^2\left(\frac{\sqrt{\rho_*}}{L\sqrt{\eta_1}\eta_2}\right),\end{aligned} \tag{88}$$

where to go from the second to the third line we used that

$$\big|\langle \underline{WG_1G_2G_2}A\rangle\big| = \mathcal{O}_\prec^2\left(\frac{\rho^*\sqrt{\rho^*}}{K\sqrt{\eta_1}\eta_2}\right)$$

by (61), and that

$$\begin{aligned} &\big|\langle G_2G_2A\rangle\big| = \mathcal{O}_\prec^2\left(\frac{\sqrt{\rho_2}}{N\eta_2^{3/2}}\right), \qquad \big|\langle G_1G_2\rangle\big| \prec \sqrt{\frac{\rho_1\rho_2}{\eta_1\eta_2}}, \\ &\big|\langle G_1G_2G_2\rangle\big| \prec \frac{\sqrt{\rho_1\rho_2}}{\sqrt{\eta_1}\eta_2^{3/2}}. \end{aligned} \tag{89}$$

The first bound in (89) follows by (45), while the second and the third bound follow by a simple Schwarz inequality. This concludes the proof of the bound for $\langle G_1G_2G_2A\rangle$. □

The remaining statements of Proposition 3.4 involving transposed resolvents are proved similarly. For completeness we included those proofs in [14], Appendix C. This concludes the proof of Proposition 3.4. □

4. CLT for resolvents. We now formulate the resolvent CLT identifying the joint distribution of $\langle G - \mathbf{E}G\rangle$, $\langle (G-\mathbf{E}G)A\rangle$ for multiple z's and traceless A's. An analogous result for only $\langle G-\mathbf{E}G\rangle$ factors was proven in [29]. Let $p \le q \in \mathbf{N}$ and let $A_1,\dots,A_p$ be matrices with $\langle A_i\rangle = 0$, $\|A_i\| \lesssim 1$ and let $\boldsymbol{a}_i$ denote the vector of the diagonal elements of A_i. Let $z_1,\dots,z_q \in \mathbf{C}\setminus\mathbf{R}$ be spectral parameters. We then set

$$\begin{aligned} &G_i := G(z_i), \qquad m_i := m(z_i), \qquad \rho_i := \frac{1}{\pi}|\Im m_i|, \qquad \eta_i := |\Im z_i|, \\ &X_i := \big\langle [G_i - \mathbf{E}G_i]A_i\big\rangle, \qquad Y_i := \langle G_i - \mathbf{E}G_i\rangle, \\ &X_S := \prod_{i\in S} X_i, \qquad Y_S := \prod_{i\in S} Y_i, \end{aligned} \tag{90}$$

so that from the local laws in (40) we have the *a priori* bounds:

$$|X_S| \prec \Psi_S, \qquad |Y_S| \prec \Psi_S, \qquad \Psi_S := \prod_{i\in S}\Psi_i,$$

$$\Psi_i := \frac{\rho_i^{1/2}}{N\eta_i^{1/2}}\mathbf{1}(i\le p) + \frac{1}{N\eta_i}\mathbf{1}\quad (i>p).$$

The following theorem identifies the leading terms of the joint moments of X's and Y's up to an error term that is smaller than the *a priori* bounds. For notational simplicity, we write $[p] := \{1,\dots,p\}$ and $(p,q] := \{p+1,\dots,q\}$.

THEOREM 4.1. *For any $\xi,\delta > 0$ and $N^\delta \le L := \min_i N\eta_i\rho_i$, we have that*

$$\mathbf{E}X_{[p]}Y_{(p,q]} = \frac{1}{N^q}\sum_{\substack{P\in \mathrm{Pair}([p])\\ Q\in\mathrm{Pair}((p,q])}}\prod_{\{i,j\}\in P} V_{ij}^\circ(A_i,A_j)\prod_{\{i,j\}\in Q} V_{ij} + \mathcal{O}\left(\frac{N^\xi\Psi}{\sqrt{L}}\right), \tag{91}$$

where we set $\Psi := \Psi_{[q]}$, define

$$\begin{aligned} V_{ij}^\circ(A_i,A_j) &:= \frac{m_i^2m_j^2\langle A_iA_j\rangle}{1-m_im_j} + \frac{\sigma m_i^2m_j^2\langle A_iA_j^t\rangle}{1-\sigma m_im_j} + \big[\kappa_4m_i^3m_j^3 + \widetilde{w}_2m_i^2m_j^2\big]\langle \boldsymbol{a}_i\boldsymbol{a}_j\rangle, \\ V_{ij} &:= \frac{m_i'm_j'}{(1-m_im_j)^2} + \frac{\sigma m_i'm_j'}{(1-\sigma m_im_j)^2} + \frac{\kappa_4}{2}(m_i^2)'(m_j^2)' + \widetilde{w}_2m_1'm_i', \end{aligned} \tag{92}$$

and Pair(S) *denotes the set of pairings of a base set* S. *Moreover, the expectation* $\mathbf{E}G$ *is given by*

$$(93)\qquad \begin{aligned}\langle \mathbf{E}G_i\rangle &= m_i + \frac{1}{N}\Big(\frac{m_i'}{m_i}\frac{\sigma m_i^2}{1-\sigma m_i^2} + \widetilde{w_2} m_i' m_i + \kappa_4 m_i' m_i^3\Big) + \mathcal{O}\Big(\frac{N^\xi \Psi_i}{L^{1/2}}\Big),\\ \langle \mathbf{E}G_i A_i\rangle &= \mathcal{O}\Big(\frac{N^\xi \Psi_i}{L^{1/2}}\Big).\end{aligned}$$

Note that the first (leading) term in (91) has a natural size of order Ψ whenever $(\Im z_i)(\Im z_j) < 0$ for every $\{i,j\}$ in the pairings P, Q.

Within the proof of Theorem 4.1, we use the classical cumulant expansion in the form

$$(94)\qquad \mathbf{E} w_{ab} f(W) = \sum_{k=1}^{R-1} \sum_{\boldsymbol{\alpha}\in\{ab,ba\}^k} \frac{\kappa(ab,\boldsymbol{\alpha})}{k!}\mathbf{E}\partial_{\boldsymbol{\alpha}} f(W) + \Omega_R,$$

where $\kappa(ab,\boldsymbol{\alpha})$ denotes the joint cumulant of $w_{ab}, w_{\alpha_1},\dots,w_{\alpha_k}$ for $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_k)$. Here, for any cut-off index R the error term Ω_R has an explicit integral representation [23], Proposition 3.2. For our application, where $f(W)$ is a product of resolvents, the error term can easily be estimated by $|\Omega_R| \lesssim N^{-(R+1)/2}$. This is due to the fact that the kth cumulant scales like $N^{-k/2}$, and each derivative creates an additional resolvent entry, which can be estimated by $\mathcal{O}(1)$ due to the single resolvent local law. In the sequel, we will omit the cutoff R from the formulas and we simply write a cumulant expansion with a formally infinite sum over k, but technically we always estimate the truncated sum.

PROOF OF THEOREM 4.1. Recalling from (65) that

$$(95)\qquad G = m - m\underline{WG} + m\langle G-m\rangle G + m\sigma\frac{G^tG}{N} + m\frac{\widetilde{w_2}}{N}\operatorname{diag}(G)G,$$

it follows that (with $\eta = |\Im z|$, $\rho = \rho(z)$)

$$(96)\qquad \begin{aligned}(1-m^2)\langle G-m\rangle &= -m\langle \underline{WG}\rangle + m\langle G-m\rangle^2 + \frac{\sigma m}{N}\langle G^tG\rangle + m\frac{\widetilde{w_2}}{N}\langle \operatorname{diag}(G)G\rangle\\ &= -m\langle \underline{WG}\rangle + \frac{m}{N}\Big(\frac{\sigma m^2}{1-\sigma m^2} + \widetilde{w_2}m^2\Big) + \mathcal{O}_\prec\Big(\frac{\rho}{N\eta L^{1/2}}\Big)\end{aligned}$$

from (43) and $\langle \operatorname{diag}(G)G\rangle = m^2 + \mathcal{O}_\prec(\sqrt{\rho/N\eta})$ due to (41). Using the cumulant expansion (94), we prove below in Section 4.1 that

$$(97)\qquad \begin{aligned}\mathbf{E}\langle \underline{WG}\rangle &= \frac{1}{N}\sum_{k\ge 2}\sum_{ab}\sum_{\boldsymbol{\alpha}\in\{ab,ba\}^k}\frac{\kappa(ab,\boldsymbol{\alpha})}{k!}\mathbf{E}\partial_{\boldsymbol{\alpha}} G_{ba}\\ &= -\kappa_4\frac{m^4}{N} + \mathcal{O}\Big(N^\xi\Big(\frac{\rho}{N^2\eta} + \frac{\rho^{3/2}}{N^{3/2}\eta^{1/2}}\Big)\Big),\end{aligned}$$

where we ignored the irrelevant error term Ω_R. Note that the $k=1$ summand has been canceled by the variance term which is included in the definition of the "underline" renormalization. Then the first claim in (93) follows immediately from (97) together with (96).

Similarly, from (95) we obtain

$$(98)\qquad \begin{aligned}\big(1-m\langle G-m\rangle\big)\langle GA\rangle &= -m\langle \underline{WGA}\rangle + \frac{\sigma m}{N}\langle G^tGA\rangle + m\frac{\widetilde{w_2}}{N}\langle \operatorname{diag}(G)GA\rangle\\ &= -m\langle \underline{WGA}\rangle + \mathcal{O}_\prec\Big(\frac{\rho}{N^{3/2}\eta}\Big)\end{aligned}$$

from (45) and

$$\langle \operatorname{diag}(G)GA\rangle = m\langle GA\rangle + \langle \operatorname{diag}(G-m)GA\rangle = \mathcal{O}_\prec\Big(\sqrt{\frac{\rho}{N\eta}}\Big).$$

For the underlined term in (98), it follows exactly as in (97) that

$$\begin{aligned} \mathbf{E}\langle \underline{WG}A\rangle &= \frac{1}{N}\sum_{k\geq 2}\sum_{ab}\sum_{\boldsymbol{\alpha}\in\{ab,ba\}^k}\frac{\kappa(ab,\boldsymbol{\alpha})}{k!}\mathbf{E}\partial_{\boldsymbol{\alpha}}(GA)_{ba} \\ &= \mathcal{O}\Big(N^\xi\Big(\frac{\rho}{N^2\eta}+\frac{\rho^{3/2}}{N^{3/2}\eta^{1/2}}\Big)\Big), \end{aligned} \tag{99}$$

where the term corresponding to κ_4 vanishes due to $\langle A\rangle = 0$. Together with (98), the second claim in (93) is also proven. This concludes the computation of the expectation.

We will now prove an asymptotic Wick theorem and explicitly compute the variance. Using (40) and (98)–(99), we replace $X_1 = \langle [G_1 - \mathbf{E}G_1]A_1\rangle$ with its leading term $\langle \underline{WG_1A_1}\rangle$, that is,

$$\begin{aligned} \langle [G_1-\mathbf{E}G_1]A_1\rangle &= -m\langle \underline{WG_1}A_1\rangle + m\mathbf{E}\langle \underline{WG_1}A_1\rangle + \mathcal{O}_\prec\Big(\frac{\rho}{N^{3/2}\eta}\Big) \\ &= -m\langle \underline{WG_1}A_1\rangle + \mathcal{O}_\prec\Big(\frac{\rho}{N^{3/2}\eta}\Big). \end{aligned} \tag{100}$$

Then for

$$\mathbf{E}X_{[p]}Y_{(p,q]} = -m_1\mathbf{E}\langle \underline{WG_1A_1}\rangle X_{(1,p]}Y_{(p,q]} + \mathcal{O}\Big(\frac{N^\xi\Psi}{L^{1/2}}\Big) \tag{101}$$

we perform a cumulant expansion of $\langle \underline{WG_1A_1}\rangle$ to obtain

$$\begin{aligned} \mathbf{E}X_{[p]}Y_{(p,q]} = &\sum_{i\in[p]\setminus\{1\}} m_1\mathbf{E}\widetilde{\mathbf{E}}\langle \widetilde{W}G_1A_1\rangle\langle G_i\widetilde{W}G_iA_i\rangle X_{[p]\setminus\{1,i\}}Y_{(p,q]} + \mathcal{O}\Big(\frac{N^\xi\Psi}{L^{1/2}}\Big) \\ &+ \sum_{i\in(p,q]} m_1\mathbf{E}\widetilde{\mathbf{E}}\langle \widetilde{W}G_1A_1\rangle\langle G_i\widetilde{W}G_i\rangle X_{[p]\setminus\{1\}}Y_{(p,q]\setminus\{i\}} \\ &- \sum_{k\geq 2}\sum_{ab}\sum_{\boldsymbol{\alpha}\in\{ab,ba\}^k}\frac{\kappa(ab,\boldsymbol{\alpha})}{k!N}\mathbf{E}\partial_{\boldsymbol{\alpha}}\big[m_1(G_1A_1)_{ba}X_{[p]\setminus\{1\}}Y_{(p,q]}\big]. \end{aligned} \tag{102}$$

We note that for arbitrary matrices U, V independent of $\widetilde{W}$ we have

$$\widetilde{\mathbf{E}}\langle \widetilde{W}U\rangle\langle \widetilde{W}V\rangle = \frac{1}{N^2}\big(\langle UV\rangle + \sigma\langle UV^t\rangle + \widetilde{w_2}\langle \operatorname{diag}(U)\operatorname{diag}(V)\rangle\big), \tag{103}$$

so that it follows that

$$\begin{aligned} &N^2\widetilde{\mathbf{E}}\langle \widetilde{W}G_1A_1\rangle\langle G_i\widetilde{W}G_iA_i\rangle \\ &= \langle G_1A_1G_iA_iG_i\rangle + \sigma\langle G_1A_1G_i^tA_i^tG_i^t\rangle + \widetilde{w_2}\langle \operatorname{diag}(G_1A_1)\operatorname{diag}(G_iA_iG_i)\rangle \\ &= \frac{m_1m_i^2\langle A_1A_i\rangle}{1-m_1m_i} + \frac{\sigma m_1m_i^2\langle A_1A_i^t\rangle}{1-\sigma m_1m_i} + m_1m_i^2\widetilde{w_2}\langle \boldsymbol{a}_1\boldsymbol{a}_i\rangle + \mathcal{O}_\prec^2\Big(\frac{N^2\Psi_{\{1,i\}}}{\sqrt{L}}\Big) \end{aligned} \tag{104}$$

from (48), (50) and

$$\begin{aligned} \langle \operatorname{diag}(G_1A_1)\operatorname{diag}(G_iA_iG_i)\rangle &= m_1\langle \operatorname{diag}(\boldsymbol{a}_1)G_iA_iG_i\rangle + \mathcal{O}_\prec\Big(\frac{1}{\sqrt{L}}\frac{\sqrt{\rho_i}}{\sqrt{\eta_i}}\Big) \\ &= m_1m_i^2\langle \boldsymbol{a}_1\boldsymbol{a}_i\rangle + \mathcal{O}_\prec\Big(\frac{1}{\sqrt{L}}\frac{\sqrt{\rho_i}}{\sqrt{\eta_i}}\Big) \end{aligned}$$

due to the second bound in (46) and the second local law in (42). Similarly, for the second line on the right-hand side of (102) we obtain

$$
\begin{aligned}
&N^2\widetilde{\mathbf{E}}\langle \widetilde{W} G_1 A_1\rangle\langle G_i \widetilde{W} G_i\rangle \\
&\quad = \langle G_1 A_1 G_i^2\rangle + \sigma\langle G_1 A_1 (G_i^2)^t\rangle + \widetilde{w_2}\langle \operatorname{diag}(G_1 A_1)\operatorname{diag}(G_i^2)\rangle \\
&\quad = \mathcal{O}_\prec^2\Big(\frac{N^2\Psi_{\{1,i\}}}{\sqrt{L}}\Big)
\end{aligned}
\tag{105}
$$

from (51) and

$$
\langle \operatorname{diag}(G_1 A_1)\operatorname{diag}(G_i^2)\rangle = m_1\langle \operatorname{diag}(\boldsymbol{a}_1) G_i^2\rangle + \mathcal{O}_\prec\Big(\frac{1}{\sqrt{L}\eta_i}\Big) = \mathcal{O}_\prec\Big(\frac{1}{\sqrt{L}\eta_i}\Big)
$$

due to (45) and the first local law in (46).

It remains to consider the third line in (102) where due to the Leibniz rule many terms can arise from the derivative. For $k \ge 2$, the derivative may act on G_1 or any of the X_i, Y_i and we consider the corresponding terms separately as in

$$
\begin{aligned}
&\sum_{k\ge 2}\sum_{ab}\sum_{\boldsymbol{\alpha}\in\{ab,ba\}^k} \frac{\kappa(ab,\boldsymbol{\alpha})}{k!N}\partial_{\boldsymbol{\alpha}}\big[m_1(G_1A_1)_{ba} X_{[p]\setminus\{1\}} Y_{(p,q]}\big] \\
&\quad = \sum_{k\ge 2}\sum_{|P_X\cup P_Y|\le k} X_{(1,p]\setminus S(P_X)} Y_{(p,q]\setminus S(P_Y)}\,\Xi_k(P_X,P_Y), \\
&\Xi_k(P_X,P_Y) \\
&\quad := \sum_{ab}\sum_{\boldsymbol{\alpha}}\kappa(ab,\boldsymbol{\alpha})\Big(\partial_{\boldsymbol{\alpha}_1}\frac{m_1(G_1A_1)_{ba}}{k_1!N}\Big)\Big(\prod_{i\in S(P_X)}\frac{\partial_{\boldsymbol{\alpha}_i} X_i}{k_i!}\Big)\Big(\prod_{i\in S(P_Y)}\frac{\partial_{\boldsymbol{\alpha}_i} Y_i}{k_i!}\Big),
\end{aligned}
\tag{106}
$$

where P_X, P_Y are unordered multisets with *support* $S(P_X) \subset (1,p]$, $S(P_Y) \subset (p,q]$. The last summation $\sum_{\boldsymbol{\alpha}}$ indicates the summation over tuples $\boldsymbol{\alpha}_1 \in \{ab,ba\}^{k_1}$, $\boldsymbol{\alpha}_i \in \{ab,ba\}^{k_i}$ with $k_i \ge 1$ denoting the multiplicity of i in $S(P_X \cup P_Y)$ and $k_1 := k - |P_X \cup P_Y| \ge 0$. We will prove below that

$$
\begin{aligned}
\Xi_k(P_X,P_Y) = {}& \mathcal{O}_\prec^1\Big(\frac{\Psi_{\{1\}\cup S(P_X\cup P_Y)}}{\sqrt{L}}\Big) \\
&- m_1^3 m_i^3\langle \boldsymbol{a}_1\boldsymbol{a}_i\rangle\frac{\kappa_4}{N^2}\mathbf{1}\big((k,P_X,P_Y) = (3,\{i,i\},\varnothing), i\in(1,p]\big).
\end{aligned}
\tag{107}
$$

By combining (102), (104), (105), (106) and (107), we obtain from induction on the number of X-factors

$$
\mathbf{E}X_{[p]}Y_{(p,q]} = \mathbf{E}Y_{(p,q]}\frac{1}{N^p}\sum_{P\in\operatorname{Pair}([p])}\prod_{\{i,j\}\in P} V_{i,j}^\circ(A_i,A_j) + \mathcal{O}\Big(\frac{N^\xi\Psi}{L^{1/2}}\Big),
\tag{108}
$$

where we recall that $V_{i,j}^\circ$ defined in (92) is given by the sum of the leading terms in (104) and (107). Here, we used that for $X_{(1,p]\setminus S(P_X)}$ and $Y_{(p,q]\setminus S(P_Y)}$ we have the high probability a priori bounds $|X_{(1,p]\setminus S(P_X)}| \prec \Psi_{(1,p]\setminus S(P_X)}$ and $|Y_{(p,q]\setminus S(P_Y)}| \prec \Psi_{(p,q]\setminus S(P_Y)}$ and, therefore,

$$
\mathbf{E}\Big|X_{(1,p]\setminus S(P_X)} Y_{(p,q]\setminus S(P_Y)}\mathcal{O}_\prec^1\Big(\frac{\Psi_{\{1\}\cup S(P_X\cup P_Y)}}{\sqrt{L}}\Big)\Big| = \mathbf{E}\mathcal{O}_\prec^1\Big(\frac{\Psi_{[1,q]}}{\sqrt{L}}\Big) \lesssim \frac{N^\xi\Psi}{\sqrt{L}}.
$$

In order to complete the proof of the theorem, it remains to compute $\mathbf{E}Y_{(p,q]}$. For convenience of notation, we relabel $Y_{(p,q]}$ to $Y_{[r]}$ with $r=q-p+1$ and obtain, analogously to (102),

$$
\begin{aligned}
\mathbf{E}Y_{[r]} = {} & \sum_{i\in[2,r]} \frac{m_1'}{m_1}\mathbf{E}\widetilde{\mathbf{E}}\langle \widetilde{W}G_1\rangle\langle G_i\widetilde{W}G_i\rangle Y_{[r]\setminus\{1,i\}} - \frac{\kappa_4}{N}m_1'm_1^3\mathbf{E}Y_{[2,r]} \\
& - \sum_{k\geq 2}\sum_{ab}\sum_{\boldsymbol{\alpha}\in\{ab,ba\}^k}\frac{\kappa(ab,\boldsymbol{\alpha})}{k!N}\mathbf{E}\partial_{\boldsymbol{\alpha}}\Big[\frac{m_1'}{m_1}(G_1)_{ba}Y_{(1,r]}\Big].
\end{aligned}
\tag{109}
$$

For the first term on the right-hand side of (109), we obtain with (103) that

$$
\begin{aligned}
& N^2\widetilde{\mathbf{E}}\langle\widetilde{W}G_1\rangle\langle G_i\widetilde{W}G_i\rangle \\
& = \langle G_1G_i^2\rangle + \sigma\langle G_1(G_i^t)^2\rangle + \widetilde{w_2}\langle \operatorname{diag}(G_1)\operatorname{diag}(G_i^2)\rangle \\
& = \frac{m_1m_i'}{(1-m_1m_i)^2} + \frac{\sigma m_1 m_i'}{(1-\sigma m_1m_i)^2} + \widetilde{w_2}m_1m_i' + \mathcal{O}_\prec\Big(\frac{\rho_1}{\sqrt{L}\eta_1\eta_i}\Big),
\end{aligned}
\tag{110}
$$

where in the last step we used (47), (49) and

$$
\langle \operatorname{diag}(G_1)\operatorname{diag}(G_i^2)\rangle = m_1\langle G_i^2\rangle + \mathcal{O}_\prec\Big(\frac{\rho_1}{L^{1/2}\eta_i}\Big) = m_1m_i' + \mathcal{O}_\prec\Big(\frac{\rho_1}{L^{1/2}\eta_i}\Big)
$$

due to (41), the first local law in (42) and the first local law in (46).

For the second line of (109) we distribute the derivative according to the Leibniz rule as

$$
\begin{aligned}
& \sum_{ab}\sum_{\boldsymbol{\alpha}\in\{ab,ba\}^k}\frac{\kappa(ab,\boldsymbol{\alpha})}{k!N}\mathbf{E}\partial_{\boldsymbol{\alpha}}\Big[\frac{m_1'}{m_1}(G_1)_{ba}Y_{(1,r]}\Big] \\
& \quad = \sum_{|P_Y|\leq k}\mathbf{E}Y_{(1,r]\setminus S(P_Y)}\Phi_k(P_Y), \\
& \Phi_k(P_Y) := \sum_{ab}\sum_{\boldsymbol{\alpha}}\frac{\kappa(ab,\boldsymbol{\alpha})}{N}\Big(\frac{m_1'}{m_1}\partial_{\boldsymbol{\alpha}_1}\frac{(G_1)_{ba}}{k_1!}\Big)\Big(\prod_{i\in S(P_Y)}\frac{\partial_{\boldsymbol{\alpha}_i}Y_i}{k_i!}\Big),
\end{aligned}
\tag{111}
$$

where P_Y is a multiset with support $S(P_Y)\subset(1,r]$, and the summation $\sum_{\boldsymbol{\alpha}}$ indicates the summation over tuples $\boldsymbol{\alpha}_1\in\{ab,ba\}^{k_1}$, $\boldsymbol{\alpha}_i\in\{ab,ba\}^{k_i}$ with $k_i\geq 1$ denoting the multiplicity of $i\in S(P_Y)$ and $k_1:=k-|P_Y|\geq 0$. Similar to (107) (but in high probability sense), we prove below that

$$
\begin{aligned}
\Phi_k(P_Y) = {} & \mathcal{O}_\prec\Big(\frac{\Psi_{\{1\}\cup S(P_Y)}}{\sqrt{L}}\Big) - (m_1^2)'(m_i^2)'\frac{\kappa_4}{2N^2}\mathbf{1}((k,P_Y)=(3,\{i,i\}), i\in(1,r]) \\
& - \frac{\kappa_4}{N}m_1'm_1^3\mathbf{1}((k,P_Y)=(3,\varnothing)).
\end{aligned}
\tag{112}
$$

By combining (109), (110), (111) and (112), we conclude

$$
\mathbf{E}Y_{[r]} = \frac{1}{N^r}\sum_{Q\in\operatorname{Pair}([r])}\prod_{\{i,j\}\in Q}V_{ij} + \mathcal{O}\Big(\frac{N^\xi\Psi_{[r]}}{L^{1/2}}\Big).
\tag{113}
$$

Therefore, the claim (91) follows immediately from combining (108) and (113) and the proof of the Theorem 4.1 is complete, modulo the proofs of (97), (107) and (112), which we present below. □

4.1. *Auxiliary calculations*: *Proof of* (97), (107) *and* (112).

PROOF OF (97). For $k=2$, the summation in (97) has either two or none diagonal G's and from (41) we estimate the corresponding terms by

$$N^{-5/2}\sum_{ab}|G_{ba}|^3 \prec N^{-3/2}+\frac{\rho^{3/2}}{N^2\eta^{3/2}}$$

and

$$N^{-5/2}\left|\sum_{ab}G_{bb}G_{aa}G_{ba}\right| \prec N^{-3/2}+\frac{\rho^{3/2}}{N^2\eta^{3/2}}.$$

Here, the first estimate uses only $|G_{ab}|\prec \mathbf{1}(a=b)+\sqrt{\rho/N\eta}$, while the second one uses $G=m+(G-m)$ and the *isotropic resummation* procedure. More precisely, by this we mean the idea of summing up the free indices into constant vectors, that is,

$$\begin{aligned}
&\sum_{ab}G_{bb}G_{aa}G_{ba}\\
&\quad=\sum_{ab}\Big[m^2+m(G-m)_{aa}+m(G-m)_{bb}+\mathcal{O}\Big(\frac{\rho}{N\eta}\Big)\Big]G_{ba}\\
&\quad=m^2G_{\mathbf{11}}+m\sum_a G_{\mathbf{1}a}\mathcal{O}\Big(\sqrt{\frac{\rho}{N\eta}}\Big)+m\sum_b G_{b\mathbf{1}}\mathcal{O}\Big(\sqrt{\frac{\rho}{N\eta}}\Big)+\sum_{ab}\mathcal{O}\Big(\frac{\rho}{N\eta}\Big)G_{ba} \qquad (114)\\
&\quad=\mathcal{O}\Big(|G_{\mathbf{11}}|+N^{1/2}\sqrt{\frac{\rho}{N\eta}}\sqrt{(GG^*)_{\mathbf{11}}}+N\frac{\rho}{N\eta}\sqrt{\operatorname{Tr}GG^*}\Big)\\
&\quad=\mathcal{O}\Big(N+\frac{N^{1/2}\rho^{3/2}}{\eta^{3/2}}\Big),
\end{aligned}$$

where we used a Schwarz inequality, and in the last step the isotropic local law for the all-one vector $\mathbf{1}=(1,\dots,1)$ of norm $\|\mathbf{1}\|=\sqrt{N}$. Thus we obtain a bound $N^{-3/2}+N^{-2}\rho^{3/2}/\eta^{3/2}$ for the $k=2$ terms in (97).

Next we consider the $k=3$ terms which give a contribution of $\rho N^{-2}\eta^{-1}$ whenever there are at least two off-diagonal G's. In order to achieve only diagonal G's, $\boldsymbol{\alpha}$ is necessarily one of (ab,ba,ba), (ba,ab,ba), or (ba,ba,ab), for which we obtain $\kappa(ab,ba,ba,ab)=\kappa_4/N^2$ for $a\neq b$. The derivative then is given by

$$\partial_{\boldsymbol{\alpha}}G_{ba}=-\partial_{ba}\partial_{ab}G_{bb}G_{aa}=2\partial_{ba}G_{ba}G_{bb}G_{aa}=-2G_{aa}^2G_{bb}^2+\cdots, \qquad (115)$$

where the neglected terms contain two off-diagonal G's and can hence be neglected. Therefore, the $k=3$ contribution of (97) is given by

$$-2\frac{3}{3!}\kappa_4\frac{1}{N^3}\sum_{ab}G_{aa}^2G_{bb}^2=-\frac{\kappa_4}{N}m^4+\mathcal{O}_\prec\Big(\frac{\rho}{N^2\eta}+\frac{\rho^{3/2}}{N^{3/2}\eta^{1/2}}\Big).$$

By estimating the $k\geq 4$ contribution trivially via $|G_{ab}|\prec 1$, this concludes the proof of (97). □

LEMMA 4.2 (Auxiliary a priori estimates). *For X_i, Y_i from* (90) *and their derivatives $\partial_{\boldsymbol{\alpha}}X_i$, $\partial_{\boldsymbol{\alpha}}Y_i$ for any multiindex $\boldsymbol{\alpha}$, we have the high probability a priori estimates*

$$|\partial_{\boldsymbol{\alpha}}X_i|\prec\frac{\rho_i^{1/2}}{N\eta_i^{1/2}},\qquad |\partial_{\boldsymbol{\alpha}}Y_i|\prec\frac{1}{N\eta_i}, \qquad (116)$$

and the more precise expansions for the first and second-order derivatives

$$
\begin{aligned}
\partial_{ab} Y_i &= -m_i' \frac{\delta_{ba}}{N} + \mathcal{O}_\prec\Big(\frac{\rho_i^{3/2}}{(N\eta_i)^{3/2}}\Big), \\
\partial_{ab}\partial_{cd} Y_i &= 2 m_i m_i' \frac{\delta_{da}\delta_{bc}}{N} + \mathcal{O}_\prec\Big(\frac{\rho_i^{3/2}}{(N\eta_i)^{3/2}}\Big).
\end{aligned} \tag{117}
$$

Moreover, we have the expansions

$$
\begin{aligned}
\partial_{ab} X_i &= -m_i^2 \frac{(A_i)_{ba}}{N} + \mathcal{O}_\prec^2\Big(\frac{\rho_i}{N^{3/2}\eta_i}\Big), \\
\partial_{ab}\partial_{cd} X_i &= m_i^3 \frac{(A_i)_{da}\delta_{bc} + (A_i)_{bc}\delta_{da}}{N} + \mathcal{O}_\prec^2\Big(\frac{\rho_i}{N^{3/2}\eta_i}\Big),
\end{aligned} \tag{118}
$$

in variance sense.

PROOF. We first establish an isotropic local law in variance sense using (59) of the form

$$
(GAG)_{\boldsymbol{xy}} = m^2 A_{\boldsymbol{xy}} + \mathcal{O}_\prec^2\Big(\frac{\rho}{N^{1/2}\eta}\Big) \tag{119}
$$

which is proved analogously to (73). The claims (117)–(118) then follow directly from (119), the first local law in (46), and

$$
\partial_{ab}\langle GB\rangle = -\frac{(GBG)_{ba}}{N} \tag{120}
$$

and

$$
\partial_{ab}\partial_{cd}\langle GB\rangle = -\partial_{cd}\frac{(GBG)_{ba}}{N} = \frac{G_{bc}(GBG)_{da} + G_{da}(GBG)_{bc}}{N}. \tag{121}
$$

The claim (116) follows inductively by the second local law in (46) since each additional derivative just adds an additional factor of G which is at most of order 1. □

PROOF OF (107). We prove (107) by considering the following five cases, which cover all possibilities: (i) odd k_1, $k \geq 4$, (ii) even k_1, $k \geq 3$, (iii) $k = 3$, $k_1 = 1$, (iv) $k = 3$, $k_1 = 3$ and (v) $k = 2$. Before considering each case separately, we outline a few ideas that are used repeatedly in the argument. The first idea is that we often replace diagonal resolvents G_{aa} and $(GA)_{aa}$ using the isotropic local law $\langle \boldsymbol{x}, G\boldsymbol{y}\rangle = m\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \mathcal{O}_\prec(\sqrt{\frac{\rho}{N\eta}})$ in order to make the leading term independent of the summation index a, as was done in (114). For example, for $\sum_a G_{aa}G_{a\boldsymbol{x}}$ this allows us to sum up the index a into the constant vector $\mathbf{1} = (1, \dots, 1)$ of norm $\sqrt{N}$ (isotropic resummation), effectively gaining a factor of $\sqrt{\rho/N\eta}$ over the naive estimate since

$$
\sum_a G_{aa}G_{a\boldsymbol{x}} = mG_{\mathbf{1}\boldsymbol{x}} + \mathcal{O}_\prec\Big(\sqrt{\frac{\rho}{N\eta}}\sum_a |G_{a\boldsymbol{x}}|\Big) = \mathcal{O}_\prec\Big(N\sqrt{\frac{\rho}{N\eta}}\Big).
$$

The second idea is that for off-diagonal resolvents we use a Schwarz inequality, followed by the Ward identity to effectively also gain a factor of $\sqrt{\rho/N\eta}$ over the naive estimate, for example,

$$
\Big|\sum_a |G_{a\boldsymbol{x}}|\Big| \leq \sqrt{N}\sqrt{\sum_a |G_{a\boldsymbol{x}}|^2} = \sqrt{N}\sqrt{(G^*G)_{\boldsymbol{xx}}} = \sqrt{\frac{N}{\eta}}\sqrt{(\Im G)_{\boldsymbol{xx}}} \prec N\sqrt{\frac{\rho}{N\eta}}.
$$

Finally, we also frequently use a simple parity consideration to count off-diagonal resolvents since the local law gives a stronger estimate for them. For an odd number of G's, each evaluated in one of the entries aa, bb, ab, ba with $a \neq b$ in total occurring equally often, at least one of the G's has to be off-diagonal.

Case (*i*), k_1 *is odd*, $k \geq 4$. In this case, we estimate $|\partial_{\boldsymbol{\alpha}_1}(G_1A_1)_{ba}| \prec 1$ by the isotropic local law $|\langle \boldsymbol{x}, GA_1\boldsymbol{y}\rangle| \prec \|\boldsymbol{x}\|\|A_1\boldsymbol{y}\| \lesssim \|\boldsymbol{x}\|\|\boldsymbol{y}\|$ in the definition of Ξ in (106) and obtain from (116) that

$$\bigl|\Xi_k(P_X, P_Y)\bigr| \prec N^{-(k+3)/2}N^2\Psi_{S(P_X\cup P_Y)} \lesssim N^{-(k-3)/2}\Psi_{\{1\}\cup S(P_X\cup P_Y)},$$

from $N^{-1} \lesssim \rho_1^{1/2}N^{-1}\eta_1^{-1/2}$, confirming (107).

Case (*ii*), k_1 *is even*, $k \geq 3$. Since $k_1 = |\boldsymbol{\alpha}_1|$ is even, it follows by parity that at least one G or GA factor is off-diagonal, hence by the local law we have that

$$\bigl|\partial_{\boldsymbol{\alpha}_1}(G_1A_1)_{ba}\bigr| \prec \bigl|(G_1)_{ab}\bigr| + \bigl|(G_1A_1)_{ba}\bigr|$$

and, therefore, by (116) and a Ward-estimate it follows that

$$\begin{aligned}\bigl|\Xi_k(P_X, P_Y)\bigr| &\prec N^{-(k+3)/2}\sum_{ab}\bigl(\bigl|(G_1A_1)_{ba}\bigr| + \bigl|(G_1)_{ab}\bigr|\bigr)\Psi_{S(P_X\cup P_Y)} \\ &\prec N^{-(k+3)/2}N^2\frac{\sqrt{\rho_1}}{\sqrt{N\eta_1}}\Psi_{S(P_X\cup P_Y)} \\ &\lesssim N^{-(k-2)/2}\Psi_{\{1\}\cup S(P_X\cup P_Y)},\end{aligned}$$

confirming (107).

Case (*iii*), $k = 3$, $k_1 = 3$. The three derivatives acting on $(G_1A_1)_{ab}$ results in one G_1A and three G_1 factors with a total of four a and four b indices. By using the local law, we replace each G_1 by m_1 and obtain

$$\begin{aligned}\bigl|\Xi_3(\varnothing, \varnothing)\bigr| &\lesssim N^{-3}\biggl|\sum_{ab}m_1^4(A_1)_{aa}\biggr| + N^{-3}\biggl|\sum_{ab}m_1^4(A_1)_{ab}\biggr| + \mathcal{O}_\prec\biggl(\frac{\Psi_{\{1\}}}{\sqrt{L}}\biggr) \\ &\prec N^{-2}\sqrt{\sum_{ab}\bigl|(A_1)_{ab}\bigr|^2} + \frac{\Psi_{\{1\}}}{\sqrt{L}} \\ &= N^{-3/2}\sqrt{\langle A_1A_1^*\rangle} + \frac{\Psi_{\{1\}}}{\sqrt{L}},\end{aligned}$$

again confirming (107).

Case (*iv*), $k = 3$, $k_1 = 1$. For $k_1 = 1$, the derivative of $(G_1A_1)_{ba}$ is given by

$$\begin{aligned}\sum_{\boldsymbol{\alpha}_1}\partial_{\boldsymbol{\alpha}_1}(G_1A_1)_{ba} &= -(G_1)_{ba}(G_1A_1)_{ba} - (G_1)_{bb}(G_1A_1)_{aa} \\ &= -m_1^2(1+\delta_{ba})(A_1)_{aa} + \mathcal{O}_\prec\biggl(\sqrt{\frac{\rho_1}{N\eta_1}}\biggr).\end{aligned}$$

If $P_X = \{i, i\}$ for some $i \in (1, p]$, then we obtain from (118) that

$$\partial_{ab,ba}X_i = \partial_{ba,ab}X_i = m_i^3\frac{(A_i)_{bb} + (A_i)_{aa}}{N} + \mathcal{O}_\prec^2\biggl(\frac{\rho_i}{N^{3/2}\eta_i}\biggr),$$

while both the $\partial_{ab,ab}$ and $\partial_{ba,ba}$ derivatives lead to delta functions δ_{ab} and are therefore lower order after summation, and thus

$$\begin{aligned}\Xi_3(\{i,i\},\varnothing) &= -m_1^3 m_i^3 \sum_{ab} \kappa(ab,ba,ab,ba) \frac{(A_1)_{aa}}{N} \frac{(A_i)_{aa}+(A_i)_{bb}}{N} + \mathcal{O}_\prec^2\Big(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\Big) \\ &= -\frac{\kappa_4}{N^2} m_1^3 m_i^3 \big(\langle \boldsymbol{a}_1 \boldsymbol{a}_i\rangle + \langle A_1\rangle\langle A_i\rangle\big) + \mathcal{O}_\prec^2\Big(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\Big) \\ &= -\frac{\kappa_4}{N^2} m_1^3 m_i^3 \langle \boldsymbol{a}_1 \boldsymbol{a}_i\rangle + \mathcal{O}_\prec^2\Big(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\Big),\end{aligned}$$

giving the leading contribution to (107).

If $P_Y = \{i,i\}$ for some $i \in (p,q]$, then we obtain from the leading term of (117) that

$$\partial_{ab,ba} Y_i = \partial_{ba,ab} Y_i = m_i m_i' \frac{2}{N} + \mathcal{O}_\prec\Big(\frac{\rho_i^{1/2}}{(N\eta_i)^{3/2}}\Big),$$

with the other derivatives again being lower order, hence

$$\begin{aligned}\Xi_3(\varnothing,\{i,i\}) &= -m_1^3 m_i m_i' \sum_{ab} \kappa(ab,ba,ab,ba) \frac{(A_1)_{aa}}{N} \frac{2}{N} + \mathcal{O}_\prec\Big(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\Big) \\ &= -2\frac{\kappa_4}{N^2} m_1^3 m_i^3 \langle A_1\rangle + \mathcal{O}_\prec\Big(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\Big) \\ &= \mathcal{O}_\prec\Big(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\Big).\end{aligned}$$

Finally, if $P_X \cup P_Y = \{i,j\}$ for some $1 < i < j$, then we either obtain a δ_{ab} from (117) or a $(A_i)_{ba}$ from (118) and, therefore, due to

$$\sum_{ab} |(A_i)_{ba}| \le N \sqrt{\sum_{ab} |(A_i)_{ba}|^2} = N^{3/2} \sqrt{\langle A_i A_i^*\rangle}$$

the leading term is at most of size $\Psi_{\{1,i,j\}} L^{-1/2}$ and we obtain

$$|\Xi_3(\{i,j\},\varnothing)| + |\Xi_3(\varnothing,\{i,j\})| + |\Xi_3(\{i\},\{j\})| = \mathcal{O}_\prec^1\Big(\frac{\Psi_{\{1,i,j\}}}{L^{1/2}}\Big).$$

Here, we used (39a), so that in case $P_X = \{i,j\}$ with $i \neq j$, the error terms from (118) can be multiplied. This concludes the proof of (107) for the case $k_1 = 1$, $k = 3$.

Case (v), $k = 2$. In case $k = 2$, there are five subcases to consider; $k_1 = 2$, $P_X = \{i,i\}$, $P_Y = \{i,i\}$, $k_1 = 1$ or $|S(P_X \cup P_Y)| = 2$.

If $k_1 = 2$, then the derivative is given by

$$\partial^2 (G_1 A_1)_{ba} = (G_1)(G_1)(G_1 A_1)$$

with three a and three b indices, so that by parity either all three matrices have indices ab, ba, or only one with the remaining two having aa, bb. For all three matrices having ab, ba indices, we can gain two factors of $\sqrt{\rho_1/N\eta_1}$ via Ward-estimates over the naive size $N^{-1/2}$ in order to obtain $\rho_1 N^{-3/2}\eta_1^{-1} \le \Psi_{\{1\}} L^{-1/2}$. If two matrices have indices aa, bb, then we perform an isotropic resummation, that is, replace one diagonal resolvent by m as we did in (114) and estimate, for example,

$$\begin{aligned}N^{-5/2} \sum_{ab} (G_1)_{bb}(G_1 A_1)_{aa}(G_1)_{ba} &= m_1 N^{-5/2} \sum_a (G_1 A_1)_{aa} (G_1)_{\mathbf{1}a} + \mathcal{O}_\prec\Big(\frac{\rho_1}{N^{3/2}\eta_1}\Big) \\ &= \mathcal{O}_\prec\Big(\frac{\rho_1^{1/2}}{N^{3/2}\eta_1^{1/2}}\Big) = \mathcal{O}_\prec\Big(\frac{\Psi_{\{1\}}}{L^{1/2}}\Big),\end{aligned}$$

and similarly for all other index distributions. Thus we obtain

$$\big|\Xi_2(\varnothing, \varnothing)\big| \prec \frac{\Psi_{\{1\}}}{L^{1/2}}.$$

Next, if $P_X = \{i, i\}$, then we obtain from (118) that

$$\begin{aligned}\Xi_2\big(\{i, i\}, \varnothing\big) &= \sum_{ab} \kappa(ab, ab, ba) \frac{m_1(G_1A_1)_{ba}}{N} m_i^3 \frac{(A_i)_{aa} + (A_i)_{bb}}{N} + \mathcal{O}_\prec^2\Big(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\Big) \\ &= m_1 m_i^3 \frac{\kappa_3}{N^{7/2}}\Big(\sum_b (G_1A_1)_{b\boldsymbol{a}_i} + \sum_a (G_1A_1)_{\boldsymbol{a}_i a}\Big) + \mathcal{O}_\prec^2\Big(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\Big) \\ &= \mathcal{O}_\prec^2\Big(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\Big),\end{aligned}$$

again using isotropic resummation and that

$$\|\boldsymbol{a}_i\| := \|\operatorname{diag} A_i\| \le N^{1/2}\sqrt{\langle A_i^* A_i\rangle}.$$

The case $P_Y = \{i, i\}$ is completely analogous, except that using (117) the constant **1** vector is summed up instead of $\boldsymbol{a}_i$ and we obtain

$$\big|\Xi_2\big(\varnothing, \{i, i\}\big)\big| \prec \frac{\Psi_{\{1,i\}}}{L^{1/2}}.$$

The case $k_1 = 1$ can be estimated by

$$\begin{aligned}\big|\Xi_2\big(\{i\}, \varnothing\big)\big| &\lesssim N^{-7/2}\Big|\sum_{ab}\big[(G_1)_{bb}(G_1A_1)_{aa} + (G_1)_{ba}(G_1A_1)_{ba}\big](G_iA_iG_i)_{ba}\Big| \\ &\lesssim N^{-7/2}\sum_a \big|(G_iA_iG_i)_{\mathbf{1}a}\big| + N^{-7/2}\sum_{ab}\sqrt{\frac{\rho_1}{N\eta_1}}\big|(G_iA_iG_i)_{ba}\big| \\ &\lesssim \frac{\rho_1^{1/2}}{N\eta_1^{1/2}}\frac{\rho_i}{N^{3/2}\eta_i} \\ &\lesssim \frac{\Psi_{\{1,i\}}}{L^{1/2}}\end{aligned}$$

and similarly

$$\begin{aligned}\big|\Xi_2\big(\varnothing, \{i\}\big)\big| &\lesssim N^{-7/2}\Big|\sum_{ab}\big[(G_1)_{bb}(G_1A_1)_{aa} + (G_1)_{ba}(G_1A_1)_{ba}\big](G_i^2)_{ba}\Big| \\ &\lesssim N^{-7/2}\sum_a \big|(G_i^2)_{\mathbf{1}a}\big| + N^{-7/2}\sum_{ab}\sqrt{\frac{\rho_1}{N\eta_1}}\big|(G_i^2)_{ba}\big| \\ &\lesssim \frac{\rho_1^{1/2}}{N\eta_1^{1/2}}\frac{\rho_i^{3/2}}{N^{3/2}\eta_i^{3/2}} \\ &\lesssim \frac{\Psi_{\{1,i\}}}{L^{1/2}}\end{aligned}$$

from (120).

For the final case $|S(P_X \cup P_Y)| = 2$, we estimate for $i \neq j$,

$$\begin{aligned}\big|\Xi_2(\{i,j\},\varnothing)\big| &\lesssim \frac{1}{N^{5/2}}\sum_{ab}|(G_1A_1)_{ba}|\bigg(\frac{|(A_i)_{ba}|}{N} + \mathcal{O}^2_\prec\Big(\frac{\rho_i}{N^{3/2}|\eta_i|}\Big)\bigg)\\ &\quad\times\bigg(\frac{|(A_j)_{ba}|}{N} + \mathcal{O}^2_\prec\Big(\frac{\rho_j}{N^{3/2}|\eta_j|}\Big)\bigg)\\ &= \mathcal{O}^1_\prec\bigg(N^{-1/2}\sqrt{\frac{\rho_1}{N|\eta_1|}}\Big(\frac{1}{N^{3/2}} + \frac{\rho_i}{N^{3/2}|\eta_i|}\Big)\frac{\rho_j^{1/2}}{N|\eta_j|^{1/2}}\bigg)\\ &= \mathcal{O}^1_\prec\Big(\frac{\Psi_{\{1,i,j\}}}{L^{1/2}}\Big),\\ \big|\Xi_2(\varnothing,\{i,j\})\big| &\lesssim N^{-5/2}\sum_{ab}|(G_1A_1)_{ba}|\bigg(\frac{\delta_{ba}}{N} + \mathcal{O}_\prec\Big(\frac{\rho_i^{1/2}}{(N|\eta_i|)^{3/2}}\Big)\bigg)\\ &\quad\times\bigg(\frac{\delta_{ba}}{N} + \mathcal{O}_\prec\Big(\frac{\rho_j^{1/2}}{(N|\eta_j|)^{3/2}}\Big)\bigg)\\ &\prec N^{-5/2}\frac{1}{N|\eta_j|} + N^{-5/2}N^2\sqrt{\frac{\rho_1}{N|\eta_1|}}\frac{\rho_i^{1/2}}{(N|\eta_i|)^{3/2}}\frac{1}{N|\eta_j|}\\ &\prec \frac{\Psi_{\{1,i,j\}}}{L^{1/2}}\end{aligned}$$

and similarly

$$\big|\Xi_2(\{i\},\{j\})\big| = \mathcal{O}^2_\prec\Big(\frac{\Psi_{\{1,i,j\}}}{L^{1/2}}\Big).$$

This concludes the proof of (107). □

PROOF OF (112). The proof of (112) is very similar to that of (107) and we again consider the cases (i) odd k_1, $k \geq 4$, (ii) even k_1, $k \geq 3$, (iii) $k = 3$, $k_1 = 1$, (iv) $k = 3$, $k_1 = 3$ separately.

Case (*i*), k_1 *is odd*, $k \geq 4$. In this case, we estimate $|\partial_{\boldsymbol{\alpha}_1}(G_1)_{ba}| \prec 1$ and obtain from (116) that

$$\big|\Phi_k(P_Y)\big| \prec \rho_1^{-1}N^{-(k+3)/2}N^2\Psi_{S(P_Y)} \lesssim N^{-(k-3)/2}\Psi_{\{1\}\cup S(P_Y)},$$

from $N^{-1} \lesssim \rho_1 N^{-1}\eta_1^{-1}$, confirming (112).

Case (*ii*), k_1 *is even*, $k \geq 3$. Since k_1 is even it follows by parity and the local law that $|\partial_{\boldsymbol{\alpha}_1}(G_1)_{ba}| \prec |(G_1)_{ab}|$ and, therefore, by (116) and a Ward-estimate it follows that

$$\big|\Phi_k(P_Y)\big| \prec \rho_1^{-1}N^{-(k+3)/2}N^2\frac{\sqrt{\rho_1}}{\sqrt{N\eta_1}}\Psi_{S(P_Y)} \lesssim N^{-(k-2)/2}\Psi_{\{1\}\cup S(P_Y)},$$

confirming (112).

Case $k = 3$, $k_1 = 3$. The derivatives acting on $(G_1)_{ab}$ results in four G_1 factors with a total of four a and four b indices and we obtain

$$\begin{aligned}\Phi_3(\varnothing) &= -\kappa_4 \frac{m_1'}{m_1} N^{-3} \sum_{ab} (G_1)_{aa}^2 (G_1)_{bb}^2 + \mathcal{O}\Big(\rho_1^{-1} N^{-3} \sum_{ab} \big|(G_1)_{ab}\big|\Big) \\ &= -\frac{\kappa_4}{N} m_1^3 m_1' + \mathcal{O}_\prec\Big(\frac{1}{N^{3/2}\sqrt{\eta_1\rho_1}}\Big) \\ &= -\frac{\kappa_4}{N} m_1^3 m_1' + \mathcal{O}_\prec\Big(\frac{\Psi_{\{1\}}}{\sqrt{L}}\Big),\end{aligned}$$

which gives one of the leading terms in (112).

Case (*iii*), $k = 3$, $k_1 = 1$. For $k_1 = 1$, the derivative of $(G_1)_{ba}$ is given by

$$\sum_{\boldsymbol{\alpha}_1} \partial_{\boldsymbol{\alpha}_1} (G_1)_{ba} = -(G_1)_{ba}^2 - (G_1)_{bb}(G_1)_{aa} = -m_1^2(1 + \delta_{ba}) + \mathcal{O}_\prec\Big(\sqrt{\frac{\rho_1}{N\eta_1}}\Big).$$

If $P_Y = \{i, i\}$ for some $i \in (1, r]$, then we obtain from (117) that

$$\partial_{ab,ba} Y_i = \partial_{ba,ab} Y_i = m_i m_i' \frac{2}{N} + \mathcal{O}_\prec\Big(\frac{\rho_i^{1/2}}{(N\eta_i)^{3/2}}\Big),$$

while both the $\partial_{ab,ab}$ and $\partial_{ba,ba}$ derivatives lead to delta functions δ_{ab} and are therefore lower order, and thus

$$\begin{aligned}\Phi_3\big(\{i, i\}, \varnothing\big) &= -m_1 m_1' m_i m_i' \sum_{ab} \kappa(ab, ba, ab, ba) \frac{1}{N}\frac{2}{N} + \mathcal{O}_\prec\Big(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\Big) \\ &= -\frac{2\kappa_4}{N^2} m_1 m_1' m_i m_i' + \mathcal{O}_\prec\Big(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\Big) \\ &= -\frac{\kappa_4}{2N^2} \big(m_1^2\big)'\big(m_i^2\big)' + \mathcal{O}_\prec\Big(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\Big),\end{aligned}$$

giving the other leading term in (112). On the other hand, if $P_Y = \{i, j\}$ for some $1 < i < j$, then we obtain a δ_{ab} from (117) and, therefore, the leading term is at most of size $\Psi_{\{1,i,j\}} L^{-1/2}$ and we obtain

$$\big|\Phi_3\big(\{i, j\}\big)\big| = \mathcal{O}_\prec\Big(\frac{\Psi_{\{1,i,j\}}}{L^{1/2}}\Big).$$

This concludes the proof of (112) for the case $k_1 = 1$, $k = 3$.

Case (*iv*), $k = 2$. In case $k = 2$, there are four subcases to consider; $k_1 = 2$, $P_Y = \{i, i\}$, $k_1 = 1$ or $P_Y = \{i, j\}$ for $i \neq j$.

If $k_1 = 2$, then the derivative is given by $\partial^2 (G_1)_{ba} = (G_1)^3$ with three a and three b indices, so that by parity either all three matrices have indices ab, ba, or only one with the remaining two having aa, bb. For all three matrices having ab, ba indices, we can gain two factors of $\sqrt{\rho_1/N\eta_1}$ via Ward estimates over the naive size $N^{-1/2}$ in order to obtain $N^{-3/2}\eta_1^{-1} \le \Psi_{\{1\}} L^{-1/2}$. If two matrices have indices aa, bb, then using isotropic resummation we estimate

$$\begin{aligned}\rho_1^{-1} N^{-5/2} \sum_{ab} (G_1)_{bb}(G_1)_{aa}(G_1)_{ba} &= m_1 N^{-5/2} \sum_a (G_1)_{aa}(G_1)_{\mathbf{1}a} + \mathcal{O}_\prec\Big(\frac{\rho_1}{N^{3/2}\eta_1}\Big) \\ &= \mathcal{O}_\prec\Big(\frac{\rho_1^{1/2}}{N^{3/2}\eta_1^{1/2}}\Big) = \mathcal{O}_\prec\Big(\frac{\Psi_{\{1\}}}{L^{1/2}}\Big),\end{aligned}$$

in order to conclude

$$\left|\Phi_2(\varnothing)\right| \prec \frac{\Psi_{\{1\}}}{L^{1/2}}.$$

Next, if $P_Y = \{i, i\}$, then we obtain from (117) that

$$\begin{aligned}\Phi_2(\{i,i\}) &= \frac{m_1' m_i m_i'}{m_1} \sum_{ab} \kappa(ab, ab, ba) \frac{(G_1)_{ba}}{N} \frac{2}{N} + \mathcal{O}_\prec\left(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\right) \\ &= \frac{m_1' m_i m_i'}{m_1} \frac{2\kappa_3}{N^{7/2}} \langle \mathbf{1}, G_1 \mathbf{1} \rangle + \mathcal{O}_\prec\left(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\right) \\ &= \mathcal{O}_\prec\left(\frac{\Psi_{\{1,i\}}}{L^{1/2}}\right).\end{aligned}$$

The case $k_1 = 1$ can be estimated again by isotropic resummation as

$$\begin{aligned}\left|\Phi_2(\{i\})\right| &\lesssim \frac{N^{-7/2}}{\rho_1 \rho_i} \left| \sum_{ab} [(G_1)_{bb}(G_1)_{aa} + (G_1)_{ba}^2](G_i^2)_{ba} \right| \\ &\lesssim \frac{N^{-7/2}}{\rho_1 \rho_i} \sum_a |(G_i^2)_{\mathbf{1}a}| + \frac{N^{-7/2}}{\rho_1 \rho_i} \sum_{ab} \sqrt{\frac{\rho_1}{N\eta_1}} |(G_i^2)_{ba}| \\ &\lesssim \frac{\rho_1^{1/2}}{N\eta_1^{1/2}} \frac{\rho_i}{N^{3/2}\eta_i} \lesssim \frac{\Psi_{\{1,i\}}}{L^{1/2}}.\end{aligned}$$

For the final case $P_Y = \{i, j\}$ with $i \neq j$, we obtain from (117) that

$$\begin{aligned}\left|\Phi_2(\{i,j\})\right| &\lesssim \frac{N^{-5/2}}{\rho_1 \rho_i \rho_j} \sum_{ab} |(G_1)_{ba}| \left(\frac{\delta_{ba}}{N} + \mathcal{O}_\prec\left(\frac{\rho_i^{1/2}}{(N\eta_i)^{3/2}}\right)\right)\left(\frac{\delta_{ba}}{N} + \mathcal{O}_\prec\left(\frac{\rho_j^{1/2}}{(N\eta_j)^{3/2}}\right)\right) \\ &\prec \frac{\Psi_{\{1,i,j\}}}{L^{1/2}}.\end{aligned}$$

This concludes the proof of (112). □

5. Functional CLT: Proof of Theorems 2.3–2.4. In this section, we prove our main results, the functional central limit theorems using the resolvent CLT, Theorem 4.1. Via standard representation formulas this involves fairly standard but tedious calculations. We first give a detailed calculation for the case sharp cut-off case in Section 5.1 using the less-known Pleijel's formula which proves Theorems 2.3. The proof of Theorem 2.4 in Section 5.2 relies on similar calculations using the more conventional Helffer–Sjöstrand formula; the details are deferred to [14], Appendix F.

5.1. *Proof of the functional CLT for the sharp cut-off.* PROOF OF THEOREM 2.3. Let $\mathring{A} := A - \langle A \rangle$, and define

(122) $$\mathbf{1}_{K,i_0}(i) := \mathbf{1}\big(|i - i_0| \le K\big).$$

We recall the rigidity bound (see, e.g., [22], Lemma 7.1, Theorem 7.6, or [26], Section 5):

(123) $$|\lambda_i - \gamma_i| \prec \frac{1}{N^{2/3}\hat{i}^{1/3}},$$

where $\widehat{i} := i \wedge (N+1-i)$. Here, γ_i are the classical eigenvalue locations (*quantiles*) defined by

$$\int_{-\infty}^{\gamma_i} \rho(x)\,\mathrm{d}x = \frac{i}{N}, \quad i \in [N], \tag{124}$$

where we recall $\rho(x) = \rho_{\mathrm{sc}}(x) = (2\pi)^{-1}\sqrt{(4-x^2)_+}$. We now present the proof in the bulk regime, the edge is completely analogous and so omitted. Define $\eta_K(\gamma_{i_0})$ implicitly by

$$\eta_{i_0} = \eta_K(\gamma_{i_0}) := \frac{K}{N\rho(\gamma_{i_0} + \mathrm{i}\eta_K(\gamma_{i_0}))}; \tag{125}$$

which is the local scale around γ_{i_0} containing roughly K eigenvalues. Then, by (122) and (123), we readily conclude

$$\begin{aligned} \sqrt{\frac{N}{K}} \sum_{|i-i_0|\le K} \langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle &= \sqrt{\frac{N}{K}} \sum_{i=1}^{N} \mathbf{1}_{K,i_0}(i)\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle \\ &= \frac{N^{3/2}}{\sqrt{K}}\langle P(W)\mathring{A}\rangle + \mathcal{O}_\prec\left(\frac{1}{\sqrt{K}}\right), \end{aligned} \tag{126}$$

where we defined the spectral projection

$$P(W) = \mathbf{1}(\gamma_{i_0} - \eta_{i_0} \le W \le \gamma_{i_0} + \eta_{i_0}), \tag{127}$$

and used that $|\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle| \prec N^{-1/2}$ by [16], Theorem 1.

Using Pleijel's representation formula of the spectral projection of a Hermitian matrix in terms of contour integral of its resolvent in [24], equation (13), (see also [48], equation (5)), we find that (see Appendix E for more details)

$$\frac{N^{3/2}}{\sqrt{2K}}\langle P(W)\mathring{A}\rangle = \frac{N^{3/2}}{2\pi\mathrm{i}\sqrt{2K}} \int_{\Gamma_{K,i_0}} \langle G(z)\mathring{A}\rangle\,\mathrm{d}z + \mathcal{O}_\prec\left(\frac{N\eta_0}{\sqrt{K}}\right), \tag{128}$$

with Γ_{K,i_0} the contour oriented counterclockwise and defined by

$$\begin{aligned} \Gamma_{K,i_0} &:= \{z \in \mathbf{C} \mid \Re z \in [\gamma_{i_0} - \eta_{i_0}, \gamma_{i_0} + \eta_{i_0}] \text{ and } |\Im z| = M\} \\ &\quad \cup \{z \in \mathbf{C} \mid \Re z \in \{\gamma_{i_0} - \eta_{i_0}, \gamma_{i_0} + \eta_{i_0}\} \text{ and } |\Im z| \in [\eta_0, M]\}, \end{aligned} \tag{129}$$

for any $M > 0$, and with some parameter η_0 such that $N^{-1} \ll \eta_0 \ll K^{1/2}/N$.

By Young's inequality, for any $p \ge 2$ and $\delta > 0$, we get from (128) that

$$\begin{aligned} &\mathbf{E}\left|\frac{N^{3/2}}{\sqrt{2K}}\langle P(W)\mathring{A}\rangle\right|^p \\ &\quad = (1+\mathcal{O}(N^{-\delta}))\mathbf{E}\left|\frac{N^{3/2}}{\sqrt{2K}}\int_{\Gamma_{K,i_0}}\langle G(z)\mathring{A}\rangle\,\mathrm{d}z\right|^p + \mathcal{O}\left(\left(\frac{N^{1+\delta}\eta_0}{\sqrt{K}}\right)^p\right). \end{aligned} \tag{130}$$

By (91), it follows that for even p (for odd p the leading term is zero) we have

$$\begin{aligned} &\mathbf{E}\left|\frac{N^{3/2}}{\sqrt{2K}}\langle P(W)\mathring{A}\rangle\right|^p \\ &\quad = ((1+\mathcal{O}(N^{-\delta}))\left(\frac{N}{2K}\right)^{p/2} \sum_{P\in \mathrm{Pair}([p])} \prod_{\{i,j\}\in P} \frac{1}{4\pi^2}\int_{\Gamma_{K,i_0}}\mathrm{d}z_i \int_{\Gamma_{K,i_0}}\mathrm{d}z_j \\ &\qquad \times \left(\frac{m_i^2 m_j^2\langle \mathring{A}^2\rangle}{1-m_i m_j} + \frac{\sigma m_i^2 m_j^2 \langle \mathring{A}\mathring{A}^t\rangle}{1-\sigma m_i m_j} + \widetilde{w}_2 m_i^2 m_j^2 \langle \mathring{\boldsymbol{a}}_i \mathring{\boldsymbol{a}}_j\rangle \right. \end{aligned} \tag{131}$$

$$+\kappa_4 m_i^3 m_j^3 \langle \mathring{\boldsymbol{a}}_i \mathring{\boldsymbol{a}}_j \rangle\Big)$$
$$+\mathcal{O}\bigg(\frac{N^\xi}{\sqrt{NM}}\bigg[\bigg(\frac{MN}{K}\bigg)^{p/2}+\bigg(\frac{MN}{K}\bigg)^{-p/2}\bigg]+\bigg(\frac{N^{1+\delta}\eta_0}{\sqrt{K}}\bigg)^p\bigg),$$

for any $\xi, \delta > 0$, with $\mathring{\boldsymbol{a}} := \operatorname{diag}(\mathring{A})$. In estimating the error term coming from (91), we used that

$$\bigg(\prod_{i\in[p]}\int_{\Gamma_{K,i_0}} \mathrm{d}z_i\bigg)\frac{1}{\sqrt{N\eta_*}}\prod_{i\in[p]}\frac{N^\xi}{N\sqrt{|\eta_i|}} \lesssim \frac{N^\xi}{\sqrt{NM}}\bigg[\bigg(\frac{MN}{K}\bigg)^{p/2}+\bigg(\frac{MN}{K}\bigg)^{-p/2}\bigg],$$

where $(MN/K)^{p/2}$ comes from the vertical lines of the contour Γ_{K,i_0} and $(MN/K)^{-p/2}$ from the horizontal ones. Note that in order to apply (91) we had to choose $\eta_0 \gg N^{-1}$, which ensures $L = N\min_i(|\eta_i|\rho_i) \sim N\eta_* \gg 1$ (since we are in the bulk regime), with $\eta_* := \min_i |\eta_i|$. It is clear that we can choose ξ, δ and $\eta_0 \ll K^{1/2}/N$, $M \ll K/N$ so that the error term in (131) is bounded by $N^{-c(p,\epsilon)}$ for some constant $c(p,\epsilon) > 0$, where $N^\epsilon \le K \le N^{1-\epsilon}$. In the following, $\eta_0 \ll M \ll K/N$ ensures that only the horizontal lines of Γ_{K,i_0} contribute to the integral, the vertical lines are negligible giving a contribution MN/K.

We start computing

$$\begin{aligned}
&\frac{2N}{K}\int\int_{\gamma_{i_0}-\eta_{i_0}}^{\gamma_{i_0}+\eta_{i_0}} \mathrm{d}x\,\mathrm{d}y\,\Re\bigg[\frac{\sigma m_1^2 m_2^2}{1-\sigma m_1 m_2}-\frac{\sigma m_1^2 \overline{m_2}^2}{1-\sigma m_1\overline{m_2}}\bigg]\\
&\quad= -\frac{N}{K}\int\int_{\gamma_{i_0}-\eta_{i_0}}^{\gamma_{i_0}+\eta_{i_0}}\sqrt{(4-x^2)_+(4-y^2)_+}\,\mathrm{d}x\,\mathrm{d}y\\
&\qquad+\mathbf{1}(\sigma=\pm 1)\frac{2N\pi}{K}\int\int_{\gamma_{i_0}-\eta_{i_0}}^{\gamma_{i_0}+\eta_{i_0}}\sqrt{(4-x^2)_+}\,\delta_{x-\sigma y}\,\mathrm{d}x\,\mathrm{d}y\\
&\qquad-\frac{N}{K}\int\int_{\gamma_{i_0}-\eta_{i_0}}^{\gamma_{i_0}+\eta_{i_0}}\frac{(1-\sigma^2)\sqrt{(4-x^2)_+(4-y^2)_+}}{\sigma^2(x^2+y^2)+(1-\sigma^2)^2-xy\sigma(1+\sigma^2)}+\mathcal{O}(\sqrt{M}).
\end{aligned}\tag{132}$$

In particular, for $K \ll N$ we get that the right-hand side of (132) is equal to

$$8\pi^2\mathbf{1}(\sigma=1)+\mathbf{1}(\sigma=-1)\frac{2N\pi}{K}\int_{I_{i_0}}\sqrt{4-x^2}\,\mathrm{d}x+\mathcal{O}\bigg(\sqrt{M}+\frac{K}{N}\bigg),$$

with $I_{i_0} := [\gamma_{i_0}-\eta_{i_0}, \gamma_{i_0}+\eta_{i_0}] \cap [-\gamma_{i_0}-\eta_{i_0}, -\gamma_{i_0}+\eta_{i_0}]$. Note that for $i_0 = \lceil cN \rceil$ we have

$$\frac{2N\pi}{K}\int_{I_{i_0}}\sqrt{4-x^2}\,\mathrm{d}x = \frac{4N\pi}{K}|I_{i_0}|+\mathcal{O}\bigg(\frac{K}{N}\bigg).$$

We now distinguish two cases: (i) $c \neq 1/2$, (ii) $c = 1/2$. If $c \neq 1/2$, then $|\gamma_{i_0}| \gtrsim |c-1/2|$ and so I_{i_0} is empty, that is,

$$\frac{2N\pi}{K}\int_{I_{i_0}}\sqrt{4-x^2}\,\mathrm{d}x = 0.$$

On the other hand, for $c = 1/2$ we have

$$\frac{2N\pi}{K}\int_{I_{i_0}}\sqrt{4-x^2}\,\mathrm{d}x = \frac{4N\pi}{K}|I_{i_0}|+\mathcal{O}\bigg(\frac{K}{N}\bigg) = 8\pi^2+\mathcal{O}\bigg(\frac{K}{N}\bigg),$$

where we used that for $c = 1/2$ we have $|I_{i_0}| = 2\eta_{i_0}+\mathcal{O}(N^{-1})$, as a consequence of $\gamma_{i_0} = \mathcal{O}(N^{-1})$, and that $\rho(\mathrm{i}\eta_{i_0}) = \pi^{-1}+\mathcal{O}(\eta_{i_0})$, with $\eta_{i_0} \lesssim KN^{-1}$ in the bulk. By analogous computations, we conclude that in the edge regime the right-hand side of (132) is equal to $\mathbf{1}(\sigma=1)8\sqrt{2}\pi^2/3$.

For mesoscopic scales, the third and of the fourth term are negligible. This concludes the proof of Theorem 2.3. □

5.2. *Proof of the functional CLT for smooth cut-off.* PROOF OF THEOREM 2.4. For simplicity, we present the proof in the macroscopic scale and in the mesoscopic scale in the bulk. The computation of the leading term and the estimate of the error terms at the edge are completely analogous and so omitted.

For any $z = x + \mathrm{i}\eta \in \mathbf{C}$, we define the almost analytic extension of $f \in H^2$ by

$$f_{\mathbf{C}}(z) = f_{\mathbf{C}}(x+\mathrm{i}\eta) := \left[f(x) + \mathrm{i}\eta\partial_x f(x)\right]\chi\left(N^a\eta\right), \tag{133}$$

where χ is a smooth cut-off equal to 1 for $\eta \in [-5, 5]$ and equal to 0 for $\eta \in [-10, 10]^c$. Note that

$$|\partial_{\overline{z}} f_{\mathbf{C}}| \lesssim N^{2a}|g''||\eta| + N^a\left(|g| + N^a|g'||\eta|\right)|\chi'|, \tag{134}$$

where $2\partial_{\overline{z}} := \partial_x + \mathrm{i}\partial_\eta$.

By the Helffer–Sjöstrand formula, we have that

$$f(\lambda) = \frac{1}{\pi}\int_{\mathbf{C}} \frac{\partial_{\overline{z}} f_{\mathbf{C}}(z)}{\lambda - z}\,\mathrm{d}^2 z = \frac{2}{\pi}\Re\int_{\mathbf{R}}\int_{\mathbf{R}_+} \frac{\partial_{\overline{z}} f_{\mathbf{C}}(z)}{\lambda - z}\,\mathrm{d}\eta\,\mathrm{d}x, \tag{135}$$

where $\mathrm{d}^2 z = \mathrm{d}x\,\mathrm{d}\eta$ is the Lebesgue measure on $\mathbf{R}^2$. We recall the following notation:

$$L_N(f, I) = \sum_{i=1}^N f(\lambda_i) - \mathbf{E}\sum_{i=1}^N f(\lambda_i) \tag{136}$$

$$\begin{aligned} L_N(f, \mathring{A}) &= L_N(f, \mathring{A}_{\mathrm{d}}) + L_N(f, A_{\mathrm{od}}), \\ &= \sum_{i=1}^N f(\lambda_i)\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle - \mathbf{E}\sum_{i=1}^N f(\lambda_i)\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle. \end{aligned} \tag{137}$$

Using (135), we write

$$\begin{aligned} L_N(f, I) &= \frac{2N}{\pi}\Re\int_{\mathbf{R}}\int_{\mathbf{R}_+} \partial_{\overline{z}} f_{\mathbf{C}}(z)\langle G(x+\mathrm{i}\eta) - \mathbf{E}G(x+\mathrm{i}\eta)\rangle\,\mathrm{d}\eta\,\mathrm{d}x, \\ L_N(f, \mathring{A}) &= \frac{2N}{\pi}\Re\int_{\mathbf{R}}\int_{\mathbf{R}_+} \partial_{\overline{z}} f_{\mathbf{C}}(z)\langle\left(G(x+\mathrm{i}\eta) - \mathbf{E}G(x+\mathrm{i}\eta)\right)\mathring{A}\rangle\,\mathrm{d}\eta\,\mathrm{d}x. \end{aligned} \tag{138}$$

Using that $|\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle| \prec N^{-1/2}$ writing G in spectral decomposition, we conclude

$$\left|\langle G(x+\mathrm{i}\eta)\mathring{A}\rangle\right| \le \frac{1}{N}\sum_{i=1}^N \frac{|\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle|}{|\lambda_i - z|} \prec \frac{1}{\sqrt{N}}\left(1 + \frac{1}{N\eta}\right) \tag{139}$$

for any $\eta \ge N^{-100}$, where we used that the local law for $|\langle G - \mathbf{E}G\rangle| \prec (N\eta)^{-1}$ holds for any $\eta \ge N^{-100}$ (e.g., see [15], Appendix A).

Then, using (139), (134) and the local law $|\langle G - \mathbf{E}G\rangle| \prec (N\eta)^{-1}$, we readily conclude that

$$\begin{aligned} L_N(f, I) &= \frac{2N}{\pi}\Re\int_{\mathbf{R}}\int_{\eta_0}^{\eta_a} \partial_{\overline{z}} f_{\mathbf{C}}(z)\langle G(x+\mathrm{i}\eta) - \mathbf{E}G(x+\mathrm{i}\eta)\rangle\,\mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_\prec\left(\eta_0 N^a\right), \\ L_N(f, \mathring{A}) &= \frac{2N}{\pi}\Re\int_{\mathbf{R}}\int_{\eta_0}^{\eta_a} \partial_{\overline{z}} f_{\mathbf{C}}(z)\langle\left(G(x+\mathrm{i}\eta)\right. \\ &\qquad \left.- \mathbf{E}G(x+\mathrm{i}\eta)\right)\mathring{A}\rangle\,\mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_\prec\left(\eta_0^2 N^{1/2+a}\right), \end{aligned} \tag{140}$$

where we defined

$$\eta_0 := N^{-1+\epsilon}, \qquad \eta_a := 10N^{-a},$$

for some small $\epsilon > 0$ such that $\eta_0 \ll \eta_a$. Note that in (140) we used that $\chi'(N^a\eta) = 0$ on $\eta \in [0, \eta_0]$ since $\eta_0 \ll N^{-a}$, and so that $|\partial_{\bar{z}} f_{\mathbf{C}}| \lesssim N^{2a}\eta$ by (134). We remark that the regime $\eta \leq N^{-100}$ in (140) is bounded trivially by $N^{-100+2a}$ using that $|\langle GA\rangle| \leq \eta^{-1}$ and that $|\partial_{\bar{z}} f_{\mathbf{C}}| \lesssim N^{2a}\eta$.

With the formulas (140), we thus reduced the proof of the functional CLT for general test function f to the CLT for resolvents as given in Theorem 4.1, modulo detailed calculations of the leading terms. These calculations are deferred to [14], Appendix F, and with their help we conclude the proof of Theorem 2.4. □

APPENDIX A: CASE OF VANISHING VARIANCES IN THEOREM 2.4

In this section, we give a short explanation for the cases of vanishing variances in Theorem 2.4 as listed in Remark 2.8.

The fact that for constant f the limiting processes vanish is obvious since in this case $\operatorname{Tr} f(W)A$ is deterministic. Similarly, for linear $f(x) = bx$ and $w_2 = 0$ the diagonal processes ξ_{tr}, ξ_{d} vanish since $\operatorname{Tr} f(W)A_{\mathrm{d}} = b \operatorname{Tr} WA_{\mathrm{d}} = 0$ almost surely if $w_2 = 0$. For the case of quadratic $f(x) = cx^2$ and $\kappa_4 = -1 - \sigma^2$, that is, $|w_{12}| = 1/\sqrt{N}$ almost surely, we have

$$\operatorname{Tr} W^2 A_{\mathrm{d}} = \sum_a (A_{\mathrm{d}})_{aa}\Big(\sum_b w_{ab}w_{ba}\Big) = \sum_a (A_{\mathrm{d}})_{aa}\Big(\frac{N-1}{N} + w_{aa}^2\Big)$$

almost surely, so that $\operatorname{Var}(\operatorname{Tr} W^2 A_{\mathrm{d}}) \lesssim \langle |A_{\mathrm{d}}|^2\rangle/N$. For $\sigma = 1$ and real skew-symmetric $A_{\mathrm{od}} = -A_{\mathrm{od}}^t$ is clear that $2 \operatorname{Tr} f(W)A_{\mathrm{od}} = \operatorname{Tr} f(W)A_{\mathrm{od}} + \operatorname{Tr}(f(W)A_{\mathrm{od}})^t = \operatorname{Tr} f(W)(A_{\mathrm{od}} + A_{\mathrm{od}}^t) = 0$ due to W, $f(W)$ being almost surely real-symmetric.

It remains to consider the cases of vanishing variances for $\sigma = -1$, that is, when $W = D + \mathrm{i}R$ for some real diagonal D and some real skew-symmetric $R = -R^t$. If $D = 0$, then by the exact symmetry of the spectrum, $\lambda_i = -\lambda_{N+1-i}$ and $\boldsymbol{u}_i = \overline{\boldsymbol{u}_{N+1-i}}$ (up to phase), we immediately see that all three linear statistics are constant for odd functions f. In case $D \neq 0$, the variance is not algebraically zero but it is vanishing for large N. To illustrate this mechanism, we consider the odd function $\phi(x) = x^3$. Then (a)–(c) in Remark 2.8 are saying that $\operatorname{Tr}(W^3 - 3W)$, $\operatorname{Tr}(W^3 - 2W)\mathring{A}_{\mathrm{d}}$ and $\operatorname{Tr} W^3 A_{\mathrm{od}}$ fluctuate on a scale $\ll 1$. Indeed,

$$\begin{aligned}\operatorname{Tr} W^3 A = {} & \operatorname{Tr} D^3 A + \mathrm{i} \operatorname{Tr}\big(D^2R + DRD + RD^2\big)A \\ & - \operatorname{Tr}\big(DR^2 + RDR + R^2D\big)A - \mathrm{i} \operatorname{Tr} R^3 A\end{aligned}$$

so that

$$\operatorname{Tr}\big(W^3 - 3W\big) = \operatorname{Tr} D^3 - 3 \operatorname{Tr} D\big(1 + R^2\big)$$

is $\ll 1$ since $(R^2)_{aa} = -\sum_{ab} R_{ab}^2 \approx 1$. Similarly,

$$\operatorname{Tr}\big(W^3 - 2W\big)A_{\mathrm{d}} = \operatorname{Tr} D^3 A_{\mathrm{d}} - \operatorname{Tr}\big(2D\big(1 + R^2\big) + RDR\big)A_{\mathrm{d}}$$

since $\operatorname{Tr} R^3 A_{\mathrm{d}} = 0$ due to $R = -R^t$ and $A_{\mathrm{d}} = A_{\mathrm{d}}^t$ and the right-hand side is $\ll 1$ since

$$|\operatorname{Tr} RDRA_{\mathrm{d}}| = \Big|\sum_{ab} R_{ab}^2 D_{bb}(A_{\mathrm{d}})_{aa}\Big| \lesssim N^{-1/2}\big\langle |A_{\mathrm{d}}|^2\big\rangle.$$

Finally,

$$\begin{aligned}\operatorname{Tr} W^3 A_{\mathrm{od}} = {} & \mathrm{i} \operatorname{Tr}\big(D^2R + DRD + RD^2\big)A_{\mathrm{od}} \\ & - \operatorname{Tr}\big(DR^2 + RDR + R^2D\big)A_{\mathrm{od}} - \mathrm{i} \operatorname{Tr} R^3 A_{\mathrm{od}}\end{aligned}$$

is $\lesssim N^{-1/2}\sqrt{A_{\mathrm{od}}A_{\mathrm{od}}^*}$ due to A_{od} being off-diagonal. A similar argument works for any odd polynomial.

Funding. The second author is partially funded by the ERC Advanced Grant "RMT-BEYOND" No. 101020331. The third author is supported by Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zürich Foundation.

SUPPLEMENTARY MATERIAL

Supplement to "Functional central limit theorems for Wigner matrices" (DOI: 10.1214/22-AAP1820SUPP; .pdf). Appendix B: Proofs of Lemmata 3.7–3.8. In this supplement, we prove two technical lemmata. Lemma 3.7 is relevant only for the very special $\sigma = -1$ case where the local laws with two and three G's with one transpose require a separate argument that is outside of the main methods of the paper. Lemma 3.8 is a simple calculus exercise. Appendix C: Proof of remaining estimates for Proposition 3.4. In this supplement, we present the proofs of the local laws and estimates for chains of three resolvents involving some transpose G^t in Proposition 3.4. These computations are similar to those presented in Section 3 without transposes. Appendix D: Proof of the refined bounds on renormalized alternating chains: Theorem 3.5 and Lemma 3.6. In this supplement, we prove two simple extensions of the general high probability bound [16], Theorem 5, on renormalized alternating chains of resolvents and deterministic matrices. Theorem 3.5 refines [16], Theorem 5, by distinguishing different η's in the estimates by essentially bookkeeping them more carefully. Lemma 3.6 gives a slightly stronger bound albeit only in variance sense in a very special case of two resolvents. We present its proof in full details mostly for pedagogical reasons, since the actual novelty is only a somewhat different estimate of one single term. Appendix E: Proof of the Pleijel's representation formula in (128). In this supplement, we give a detailed proof of integral representation (128) for $\langle P(W)\mathring{A}\rangle$, with $P(W)$ defined in (127). Appendix F: Calculations for the functional CLT for smooth test functions. Starting from (140) and using the explicit formulas in Theorem 4.1 in this supplement we complete the proof of Theorem 2.4 by computing the expectations and variances of the limiting Gaussian processes explicitly. The calculations are mostly mechanical albeit delicate since singular integrands need careful regularizations.

REFERENCES

[1] ADHIKARI, A. and HUANG, J. (2020). Dyson Brownian motion for general β and potential at the edge. *Probab. Theory Related Fields* **178** 893–950. MR4168391 https://doi.org/10.1007/s00440-020-00992-9

[2] BAI, Z., WANG, X. and ZHOU, W. (2009). CLT for linear spectral statistics of Wigner matrices. *Electron. J. Probab.* **14** 2391–2417. MR2556016 https://doi.org/10.1214/EJP.v14-705

[3] BAI, Z. D. and YAO, J. (2005). On the convergence of the spectral empirical process of Wigner matrices. *Bernoulli* **11** 1059–1092. MR2189081 https://doi.org/10.3150/bj/1137421640

[4] BAO, Z., SCHNELLI, K. and XU, Y. (2022). Central limit theorem for mesoscopic eigenvalue statistics of the free sum of matrices. *Int. Math. Res. Not. IMRN* **7** 5320–5382. MR4403963 https://doi.org/10.1093/imrn/rnaa210

[5] BAO, Z. and XIE, J. (2016). CLT for linear spectral statistics of Hermitian Wigner matrices with general moment conditions. *Theory Probab. Appl.* **60** 187–206. MR3568772 https://doi.org/10.1137/S0040585X97T987624

[6] BEKERMAN, F. and LODHIA, A. (2018). Mesoscopic central limit theorem for general β-ensembles. *Ann. Inst. Henri Poincaré Probab. Stat.* **54** 1917–1938. MR3865662 https://doi.org/10.1214/17-AIHP860

[7] BLOEMENDAL, A., ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2014). Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.* **19** 33. MR3183577 https://doi.org/10.1214/ejp.v19-3054

[8] BOROT, G. and GUIONNET, A. (2013). Asymptotic expansion of beta matrix models in the multi-cut regime. Preprint. Available at arXiv:1303.1045.

[9] BOURGADE, P. and YAU, H.-T. (2017). The eigenvector moment flow and local quantum unique ergodicity. *Comm. Math. Phys.* **350** 231–278. MR3606475 https://doi.org/10.1007/s00220-016-2627-6

[10] BOURGADE, P., YAU, H.-T. and YIN, J. (2020). Random band matrices in the delocalized phase I: Quantum unique ergodicity and universality. *Comm. Pure Appl. Math.* **73** 1526–1596. MR4156609 https://doi.org/10.1002/cpa.21895

[11] BOUTET DE MONVEL, A. and KHORUNZHY, A. (1999). Asymptotic distribution of smoothed eigenvalue density. I. Gaussian random matrices. *Random Oper. Stoch. Equ.* **7** 1–22. MR1678012 https://doi.org/10.1515/rose.1999.7.1.1

[12] BOUTET DE MONVEL, A. and KHORUNZHY, A. (1999). Asymptotic distribution of smoothed eigenvalue density. II. Wigner random matrices. *Random Oper. Stoch. Equ.* **7** 149–168. MR1689027 https://doi.org/10.1515/rose.1999.7.2.149

[13] BREUER, J. and DUITS, M. (2016). Universality of mesoscopic fluctuations for orthogonal polynomial ensembles. *Comm. Math. Phys.* **342** 491–531. MR3459158 https://doi.org/10.1007/s00220-015-2514-6

[14] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2023). Supplement to "Functional central limit theorems for Wigner matrices." https://doi.org/10.1214/22-AAP1820SUPP

[15] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2021). Edge universality for non-Hermitian random matrices. *Probab. Theory Related Fields* **179** 1–28. MR4221653 https://doi.org/10.1007/s00440-020-01003-7

[16] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2021). Eigenstate thermalization hypothesis for Wigner matrices. *Comm. Math. Phys.* **388** 1005–1048. MR4334253 https://doi.org/10.1007/s00220-021-04239-z

[17] COLLINS, B., MINGO, J. A., ŚNIADY, P. and SPEICHER, R. (2007). Second order freeness and fluctuations of random matrices. III. Higher order freeness and free cumulants. *Doc. Math.* **12** 1–70. MR2302524

[18] DEUTSCH, J. M. (1991). Quantum statistical mechanics in a closed system. *Phys. Rev. A* **43** 2046–2049. https://doi.org/10.1103/physreva.43.2046

[19] DUITS, M. and JOHANSSON, K. (2018). On mesoscopic equilibrium for linear statistics in Dyson's Brownian motion. *Mem. Amer. Math. Soc.* **255** v+118. MR3852256 https://doi.org/10.1090/memo/1222

[20] ERDŐS, L. and KNOWLES, A. (2015). The Altshuler–Shklovskii formulas for random band matrices I: The unimodular case. *Comm. Math. Phys.* **333** 1365–1416. MR3302637 https://doi.org/10.1007/s00220-014-2119-5

[21] ERDŐS, L. and KNOWLES, A. (2015). The Altshuler–Shklovskii formulas for random band matrices II: The general case. *Ann. Henri Poincaré* **16** 709–799. MR3311888 https://doi.org/10.1007/s00023-014-0333-5

[22] ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2013). The local semicircle law for a general class of random matrices. *Electron. J. Probab.* **18** 59. MR3068390 https://doi.org/10.1214/EJP.v18-2473

[23] ERDŐS, L., KRÜGER, T. and SCHRÖDER, D. (2019). Random matrices with slow correlation decay. *Forum Math. Sigma* **7** e8. MR3941370 https://doi.org/10.1017/fms.2019.2

[24] ERDŐS, L. and SCHRÖDER, D. (2016). Fluctuations of functions of Wigner matrices. *Electron. Commun. Probab.* **21** 86. MR3600514 https://doi.org/10.1214/16-ECP38

[25] ERDŐS, L. and SCHRÖDER, D. (2018). Fluctuations of rectangular Young diagrams of interlacing Wigner eigenvalues. *Int. Math. Res. Not. IMRN* **10** 3255–3298. MR3805203 https://doi.org/10.1093/imrn/rnw330

[26] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Rigidity of eigenvalues of generalized Wigner matrices. *Adv. Math.* **229** 1435–1515. MR2871147 https://doi.org/10.1016/j.aim.2011.12.010

[27] GUIONNET, A. (2002). Large deviations upper bounds and central limit theorems for non-commutative functionals of Gaussian large random matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **38** 341–384. MR1899457 https://doi.org/10.1016/S0246-0203(01)01093-7

[28] HE, Y. (2019). Mesoscopic linear statistics of Wigner matrices of mixed symmetry class. *J. Stat. Phys.* **175** 932–959. MR3959983 https://doi.org/10.1007/s10955-019-02266-8

[29] HE, Y. and KNOWLES, A. (2017). Mesoscopic eigenvalue statistics of Wigner matrices. *Ann. Appl. Probab.* **27** 1510–1550. MR3678478 https://doi.org/10.1214/16-AAP1237

[30] HE, Y. and KNOWLES, A. (2020). Mesoscopic eigenvalue density correlations of Wigner matrices. *Probab. Theory Related Fields* **177** 147–216. MR4095015 https://doi.org/10.1007/s00440-019-00946-w

[31] HUANG, J. and LANDON, B. (2019). Rigidity and a mesoscopic central limit theorem for Dyson Brownian motion for general β and potentials. *Probab. Theory Related Fields* **175** 209–253. MR4009708 https://doi.org/10.1007/s00440-018-0889-y

[32] JOHANSSON, K. (1998). On fluctuations of eigenvalues of random Hermitian matrices. *Duke Math. J.* **91** 151–204. MR1487983 https://doi.org/10.1215/S0012-7094-98-09108-6

[33] KHORUNZHY, A. M., KHORUZHENKO, B. A. and PASTUR, L. A. (1996). Asymptotic properties of large random matrices with independent entries. *J. Math. Phys.* **37** 5033–5060. MR1411619 https://doi.org/10.1063/1.531589

[34] KNOWLES, A. and YIN, J. (2013). The isotropic semicircle law and deformation of Wigner matrices. *Comm. Pure Appl. Math.* **66** 1663–1750. MR3103909 https://doi.org/10.1002/cpa.21450

[35] LANDON, B., SOSOE, P. and YAU, H.-T. (2019). Fixed energy universality of Dyson Brownian motion. *Adv. Math.* **346** 1137–1332. MR3914908 https://doi.org/10.1016/j.aim.2019.02.010

[36] LI, Y., SCHNELLI, K. and XU, Y. (2021). Central limit theorem for mesoscopic eigenvalue statistics of deformed Wigner matrices and sample covariance matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **57** 506–546. MR4255183 https://doi.org/10.1214/20-aihp1086

[37] LI, Y. and XU, Y. (2021). On fluctuations of global and mesoscopic linear statistics of generalized Wigner matrices. *Bernoulli* **27** 1057–1076. MR4255226 https://doi.org/10.3150/20-bej1265

[38] LYTOVA, A. (2013). On non-Gaussian limiting laws for certain statistics of Wigner matrices. *Zh. Mat. Fiz. Anal. Geom.* **9** 536–581, 611, 615. MR3155024

[39] LYTOVA, A. and PASTUR, L. (2009). Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *Ann. Probab.* **37** 1778–1840. MR2561434 https://doi.org/10.1214/09-AOP452

[40] LYTOVA, A. and PASTUR, L. (2009). Fluctuations of matrix elements of regular functions of Gaussian random matrices. *J. Stat. Phys.* **134** 147–159. MR2489497 https://doi.org/10.1007/s10955-008-9665-1

[41] LYTOVA, A. and PASTUR, L. (2011). Non-Gaussian limiting laws for the entries of regular functions of the Wigner matrices. Preprint. Available at arXiv:1103.2345.

[42] MALE, C., MINGO, J. A., PÉCHÉ, S. and SPEICHER, R. (2022). Joint global fluctuations of complex Wigner and deterministic matrices. *Random Matrices Theory Appl.* **11** 2250015. MR4414193 https://doi.org/10.1142/S2010326322500150

[43] MARCINEK, J. (2020). *High Dimensional Normality of Noisy Eigenvectors*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Harvard University. MR4272266

[44] MINGO, J. A. and SPEICHER, R. (2017). *Free Probability and Random Matrices*. *Fields Institute Monographs* **35**. Springer, New York. MR3585560 https://doi.org/10.1007/978-1-4939-6942-5

[45] NAGEL, J. (2021). A functional CLT for partial traces of random matrices. *J. Theoret. Probab.* **34** 953–974. MR4259455 https://doi.org/10.1007/s10959-019-00982-1

[46] O'ROURKE, S., RENFREW, D. and SOSHNIKOV, A. (2013). On fluctuations of matrix entries of regular functions of Wigner matrices with non-identically distributed entries. *J. Theoret. Probab.* **26** 750–780. MR3090549 https://doi.org/10.1007/s10959-011-0396-x

[47] PIZZO, A., RENFREW, D. and SOSHNIKOV, A. (2012). Fluctuations of matrix entries of regular functions of Wigner matrices. *J. Stat. Phys.* **146** 550–591. MR2880032 https://doi.org/10.1007/s10955-011-0404-7

[48] PLEIJEL, Å. (1963). On a theorem by P. Malliavin. *Israel J. Math.* **1** 166–168. MR0167751 https://doi.org/10.1007/BF02759705

[49] RUDNICK, Z. and SARNAK, P. (1994). The behaviour of eigenstates of arithmetic hyperbolic manifolds. *Comm. Math. Phys.* **161** 195–213. MR1266075

[50] SHAMIS, M. (2013). Density of states for Gaussian unitary ensemble, Gaussian orthogonal ensemble, and interpolating ensembles through supersymmetric approach. *J. Math. Phys.* **54** 113505. MR3137043 https://doi.org/10.1063/1.4830013

[51] SHCHERBINA, M. (2011). Central limit theorem for linear eigenvalue statistics of the Wigner and sample covariance random matrices. *Zh. Mat. Fiz. Anal. Geom.* **7** 176–192, 197, 199. MR2829615

[52] SILVERSTEIN, J. W. (1990). Weak convergence of random functions defined by the eigenvectors of sample covariance matrices. *Ann. Probab.* **18** 1174–1194. MR1062064

[53] SILVERSTEIN, J. W. (2020). Weak convergence of a collection of random functions defined by the eigenvectors of large dimensional random matrices. Preprint. Available at arXiv:2012.12950.

[54] SINAĬ, Y. G. and SOSHNIKOV, A. B. (1998). A refinement of Wigner's semicircle law in a neighborhood of the spectrum edge for random symmetric matrices. *Funktsional. Anal. i Prilozhen.* **32** 56–79, 96. MR1647832 https://doi.org/10.1007/BF02482597

[55] SOSOE, P. and WONG, P. (2013). Regularity conditions in the CLT for linear eigenvalue statistics of Wigner matrices. *Adv. Math.* **249** 37–87. MR3116567 https://doi.org/10.1016/j.aim.2013.09.004