Eigenstate Thermalization Hypothesis for Wigner Matrices

Giorgio Cipolloni[1], **László Erdős**[1] , **Dominik Schröder**[2]

[1] IST Austria, Am Campus 1, 3400 Klosterneuburg, Austria.
   E-mail: giorgio.cipolloni@ist.ac.at; lerdos@ist.ac.at
[2] Institute for Theoretical Studies, ETH Zurich, Clausiusstr. 47, 8092 Zurich, Switzerland.
   E-mail: dschroeder@ethz.ch

Abstract: We prove that any deterministic matrix is approximately the identity in the eigenbasis of a large random Wigner matrix with very high probability and with an optimal error inversely proportional to the square root of the dimension. Our theorem thus rigorously verifies the Eigenstate Thermalisation Hypothesis by Deutsch (Phys Rev A 43:2046–2049, 1991) for the simplest chaotic quantum system, the Wigner ensemble. In mathematical terms, we prove the strong form of Quantum Unique Ergodicity (QUE) with an optimal convergence rate for all eigenvectors simultaneously, generalizing previous probabilistic QUE results in Bourgade and Yau (Commun Math Phys 350:231–278, 2017) and Bourgade et al. (Commun Pure Appl Math 73:1526–1596, 2020).

1. Introduction

Since the groundbreaking discovery of E. Wigner [53] postulating that Hermitian random matrices can effectively model the universal statistics of gaps between energy levels of large atomic nuclei, simple random matrices have been routinely used to replace more complicated quantum Hamilton operators for many other physically relevant problems, especially in disordered or chaotic quantum systems. A fundamental phenomenon of such systems is *Quantum Ergodicity (QE)*, stating that the eigenvectors tend to become uniformly distributed in the phase space.

   In this paper we study an enhanced version of this question, the *Quantum Unique Ergodicity (QUE)*, for real or complex Wigner matrices and for general observables. We recall that the Wigner matrix ensemble consists of $N \times N$ random Hermitian matrices $W = W^*$ with centred, independent, identically distributed *(i.i.d.)* entries up to the symmetry constraint $w_{ab} = \overline{w_{ba}}$. Let $\{u_i\}_{i=1}^N$ be an orthonormal eigenbasis of W. Our main Theorem 1 asserts that for any deterministic matrix A with $\|A\| \leq 1$ we have the limit $\langle u_i, A u_j \rangle \to \delta_{ij} \langle A \rangle$ with very high probability (and hence uniform in i, j) with

optimal speed of convergence $1/\sqrt{N}$, i.e.

$$\max_{ij} \left| \langle \boldsymbol{u}_i, A\boldsymbol{u}_j \rangle - \langle A \rangle \delta_{ij} \right| \lesssim \frac{N^\epsilon}{\sqrt{N}}. \tag{1}$$

Here we introduced the shorthand notation $\langle R \rangle := \frac{1}{N} \operatorname{Tr} R$ for the normalized trace of any $N \times N$ matrix. In other words, (1) establishes the *QUE in strong form* (i.e. uniformly in i, j) for any Wigner matrix, and shows that the action of any bounded traceless deterministic matrix on the eigenbasis $\{\boldsymbol{u}_i\}_{i=1}^N$ makes it *asymptotically orthogonal* to itself (up to an optimal error $N^{-1/2}$). For genuinely complex Wigner matrices our second main Theorem 2 asserts that

$$\max_{ij} \left| \langle \boldsymbol{u}_i, \overline{\boldsymbol{u}_j} \rangle \right| \lesssim \frac{N^\epsilon}{\sqrt{N}}, \tag{2}$$

again with very high probability, showing that the eigenbases of W and $W^t = \overline{W}$ are asymptotically orthogonal.

The question of ergodicity for general observables is also known as the *Eigenstate Thermalization Hypothesis (ETH)* in the physics literature since the seminal papers of Deutsch [22] and Srednicki [51], see also [20] and [21] for reviews and further references. Our result thus proves ETH with an optimal speed of convergence as predicted, e.g., in [20, Eqs. (20)] for the simplest chaotic quantum system, the Wigner ensemble.

Historically, the most prominent model for quantum ergodicity is the natural quantization of a chaotic classical dynamical system in the semiclassical or in the high-energy regimes. The first mathematical result on QE was obtained by Shnirelman [49]. It asserts that for *most* high energy (normalized) eigenfunctions ψ_i of the Laplace-Beltrami operator on a surface with ergodic geodesic flow the measures $|\psi_i(x)|^2 \mathrm{d}x$ become completely flat as $i \to \infty$. This result was later extended by Colin de Verdiére [19] and Zelditch [56] for much larger classes of observables showing that if A is an appropriate pseudod-ifferential operator with symbol $\sigma(A)$, then $\langle \psi_i, A\psi_j \rangle \to \delta_{ij} \int_{S^*} \sigma(A)$ for *most* index pairs as $i, j \to \infty$, where S^* is the unit cotangent bundle of the surface. The analogous result on large regular graphs was obtained by Anantharaman and Le Masson [2]. The celebrated Quantum Unique Ergodicity (QUE) conjecture, formulated by Rudnick and Sarnak [46] in 1994, is a natural strengthening of these results stating that the same limits hold for *all* indices excluding that exceptional sequences may exhibit exotic behaviour (*scarring*). QUE in this form is still an outstanding open question; only certain special cases have been proven, e.g. on arithmetic surfaces for the joint eigenfunctions of the Laplacian and the Hecke operator by Lindenstrauss [42], with Soundararajan's extension [50], see also [15,36].

The speed of convergence in quantum ergodicity has been a fundamental question in the theory of quantum chaos, see e.g. [54] for a review and [5] for numerical results. For strongly chaotic (hyperbolic) systems the general physics prediction is that the variance of $\langle \psi_i, A\psi_j \rangle$ is proportional with the inverse of the Heisenberg time, roughly speaking the local eigenvalue spacing (see e.g. [23, Eq. (24)] building upon earlier results by Feingold and Peres [32]). For the hyperbolic geodesic flow on general Riemannian manifolds only inverse logarithmic decay has been proven by Zelditch [55] and Schubert [48] which is even optimal for a special highly degenerate eigenbasis of the quantization of Arnold's cat map [47], see also [40] for surfaces of high genus. For similar quantitative QUE results on large deterministic graphs see [1,3,16]. Much stronger polynomial bounds hold for special arithmetic surfaces proven by Luo and Sarnak [43] and for linear

maps on the torus [39,45] and toral eigenfunctions [35]. For random d-regular graph optimal polynomial speed of convergence for QUE with diagonal observables has been obtained in [6,7].

For large Wigner matrices, using the Dyson Brownian motion (DBM) for eigenvectors, Bourgade and Yau [12] showed that for any fixed deterministic unit vector q and any i in the bulk spectrum or close to the edge, the squared overlaps $N|\langle u_i, q \rangle|^2$ converge in distribution to the square of a standard Gaussian as $N \to \infty$ (see also [8] for deformed Wigner matrices and [37,52] for the same result under four moment matching condition in the bulk). The corresponding a priori bound, asserting that $N|\langle u_i, q \rangle|^2 \lesssim N^\epsilon$ with very high probability for any $\epsilon > 0$, has been known beforehand as the *complete delocalisation of eigenvectors* [11,29,30,38]. DBM methods allow to obtain optimal delocalisation estimates [10]. We mention that [12] also obtains asymptotic normality for the joint distribution of finitely many eigenvectors tested against one fixed vector q and for the joint distribution of a single eigenvector with finitely many test vectors $q_1, q_2, \ldots q_K$. Very recently the joint normality of finitely many eigenvectors and finitely many test vectors has also been achieved [44].

These results based upon DBM establish the universality of fluctuation for individual eigenvectors tested against *finite rank* observables A, i.e. for $A = \frac{N}{K} \sum_{k \le K} a_k |q_k\rangle\langle q_k|$, with K being N-independent and $a_k \in [-1, 1]$. The key mechanism of QUE for general observables is the self-averaging (ergodic) property of this sum as the rank $K = K(N)$ tends to infinity. As a simple corollary of the fluctuation results, QUE in a weak form was also obtained asserting that $\langle u_i, A u_i \rangle \to \langle A \rangle$ in probability for any fixed i in the bulk if rank(A) grows with N, see [12, Corollary 1.4] (this result was stated only for diagonal matrices A, but it directly generalizes to any A by spectral decomposition). However, the effective probabilistic estimates in [12] were not sufficient to prove the strong form of QUE, i.e. to guarantee that the limit holds for all eigenvectors simultaneously. This uniformity was proven in [13, Theorem 2.5] but only for random matrices with a Gaussian component of size $t \gg 1/N$ with an error of order $1/\sqrt{Nt}$. An off-diagonal version, $\langle u_i, A u_j \rangle \to 0$ for $i \ne j$, coined as quantum weak mixing, was also obtained in [13] and strengthened in [9]. Standard Green function comparison arguments may be used to remove the large Gaussian component but only with a considerably suboptimal error or under the extra assumption of matching the first several (in fact more than four) moments of the matrix elements of W with those of the Gaussian GOE/GUE ensemble.

Summarizing, our Theorem 1 generalizes the probabilistic QUE proven in [12, Corollary 1.4] and in [13, Theorem 2.5] to general Wigner ensembles in three aspects: (i) the speed of convergence is optimal (up to an N^ϵ factor); (ii) the limit is controlled in very high probability, and (iii) it holds uniformly throughout the spectrum including bulk, edge and the intermediate regime. For any deterministic Hermitian observable written in spectral decomposition $A = \sum_{k=1}^{N} a_k |q_k\rangle\langle q_k|$, our main result,

$$\left| \langle u_i, A u_j \rangle - \delta_{ij} \langle A \rangle \right| = \left| \frac{1}{N} \sum_{k=1}^{N} a_k \Big(N \langle u_i, q_k \rangle\langle q_k, u_j \rangle - \delta_{ij} \Big) \right| \lesssim \frac{N^\epsilon}{\sqrt{N}}, \qquad (3)$$

shows that the fluctuations of $N \langle u_i, q_k \rangle\langle q_k, u_j \rangle$ are so strongly asymptotically independent for different k's that their average has the expected $1/\sqrt{N}$ fluctuation scaling reminiscent to the central limit theorem, up to an N^ϵ factor. In fact, in our companion paper [18, Theorem 2.3] we also show that the diagonal overlaps $\langle u_i, A u_i \rangle$, after a small averaging in the index i, satisfy a CLT@.

Next we outline the novel ideas of our proof. We consider a spectrally averaged version of the overlaps

$$\Lambda^2 := \max_{i_0, j_0} \frac{1}{(2J)^2} \sum_{\substack{|i-i_0|<J \\ |j-j_0|<J}} N|\langle \boldsymbol{u}_i, A\boldsymbol{u}_j \rangle|^2 \tag{4}$$

for bounded traceless observables, $\langle A \rangle = 0$, $\|A\| \le 1$, where $J = N^\epsilon$ with some tiny $\epsilon > 0$. Our goal is to show that Λ is essentially of order one, with high probability. Denoting by $G = G(z) = (W - z)^{-1}$ the resolvent at $z \in \mathbf{H}$, notice that, by spectral decomposition,

$$\Lambda^2 \sim \sup_{E, E' \in [-2,2]} (\rho\rho')^{-1} \langle \Im G(E + \mathrm{i}\eta) A \Im G(E' + \mathrm{i}\eta') A \rangle, \tag{5}$$

where η is slightly above the local eigenvalue spacing at E and ρ is the semicircular density at E smoothed out on scale η; the primed quantities defined analogously. We note that a relation analogous to (5) has previously been used in [4, Section 5]. The main work consists in proving a high probability optimal bound on the quadratic functional of the resolvent $\langle GAGA \rangle$, with possible imaginary parts and at different spectral parameters. Note that for overlaps with a rank one observable, $A = |\boldsymbol{q}\rangle\langle\boldsymbol{q}|$, it is sufficient to control $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i \rangle = |\langle \boldsymbol{q}, \boldsymbol{u}_i \rangle|^2$. After a mild local averaging in the index i this becomes comparable with $\langle \boldsymbol{q}, (\Im G)\boldsymbol{q} \rangle$ whose control is equivalent to a conventional single-G isotropic local law. This served as a natural input for the DBM proofs on eigenvectors in [12,13]. For traceless observables, however, $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i \rangle$ does not have a sign, so we need to consider $|\langle \boldsymbol{u}_i, A\boldsymbol{u}_i \rangle|^2$ to understand its size, hence the relevant quantity is Λ^2 containing two G factors (5), i.e. single-G local laws are not sufficient.

For estimating (5) we face a combination of two serious difficulties. First, we need to gain an additional cancellation from the fact that A is traceless; second, we need to handle local laws for products of several G's. The first issue already arises on the level of a single-G local law: In Theorem 3 we will prove that the resolvent approximation $G \approx m$ by the Stieltjes transform $m = m(z)$ of the Wigner semicircular density, commonly referred to as a *local law*, holds to a higher accuracy when tested against a traceless observable. More precisely for the decomposition $\langle GA \rangle = \langle A \rangle \langle G \rangle + \langle G(A - \langle A \rangle) \rangle$ and with $\rho := |\Im m| / \pi$ we have that

$$\langle G \rangle = m + \mathcal{O}\left(\frac{1}{N\eta}\right), \quad \langle G(A - \langle A \rangle) \rangle = \mathcal{O}\left(\frac{\rho^{1/2}}{N\eta^{1/2}}\right), \tag{6}$$

with both errors being optimal, in fact they identify the scale of the asymptotic Gaussian fluctuation of $\langle G \rangle$ and $\langle G(A - \langle A \rangle) \rangle$, respectively [18,34]. Note that the error term for the traceless part is much smaller than that for $\langle G \rangle$ in the relevant small η regime. For $\langle GAG^*A \rangle$ the discrepancy is even bigger; without zero trace assumption $\langle GAG^*A \rangle \sim 1/\eta$ (e.g. for $A = I$), while for $\langle A \rangle = 0$ we will show that $\langle GAG^*A \rangle \sim 1$ even for very small η.

The second issue touches upon the basic mechanism of the standard proof of the local laws. It consists in deriving an approximate self-consistent equation for the quantity in question, e.g. $\langle GAGA \rangle$, and compare it with the corresponding deterministic equation *(Dyson equation)* without approximation error. The main error term $\langle \underline{WGAGA} \rangle$, a renormalized version of $\langle WGAGA \rangle$, see (42), is expected to be smaller than $\langle GAGA \rangle$, but when estimating its high moments by a cumulant expansion many terms with traces

of more than two G-factors emerge. Trivial a priori bounds using $\|G\| \leq 1/\eta$ are not affordable, so one has to continue expanding, resulting in higher and higher degree monomials in G; reminiscent to the notoriously difficult closure problem in the BBGKY hierarchy for the correlation functions of interacting particle dynamics. In the proof of the conventional local law $\langle G \rangle = m + \mathcal{O}(1/N\eta)$, the expansion is stopped by using the Ward identity $GG^* = \Im G/\eta$, reducing the number of G factors by one. However, with a deterministic matrix in between, as in GAG^*, Ward identity is not applicable. A trivial Schwarz bound followed by Ward identity,

$$|\langle GA^*A \rangle| \leq \langle GAA^*G^* \rangle = \frac{1}{\eta}\langle (\Im G)AA^* \rangle, \tag{7}$$

is available, but at the expense of replacing the traceless matrix A with the non-zero trace matrix AA^*, hence losing the main cancellation effect that we cannot afford. Our main idea is to use Λ from (4) as the basic control quantity and derive a stochastic Gronwall inequality for it. In doing so, we use the spectral decomposition of G to estimate traces of products of many $G's$ and A's by the lower degree term $\langle GAGA \rangle$. Technically, this requires to extract sufficiently many Λ-factors in the cumulant expansion, which we achieve by a subtle Feynman graph analysis to estimate all high moments of $|\langle \underline{WGAGA} \rangle|$.

Feynman diagrams have been systematically used to organize cumulant expansions and their estimates come on different levels of sophistication, see e.g. [17,27,28], but also related expansions in random matrices, e.g. [24,25,33]. For the proof of (6) via cumulant expansion (e.g. following [28]), it is sufficient to monitor the number of N-factors (from the size of the cumulants and from the summation of intermediate indices) and the number of ρ/η factors from the Ward identity. In the current analysis we additionally need to monitor the Λ factors. While the number of traceless A-factors is preserved along the expansion, but the cancellation effect of some of them may be lost as in (7). Our proof has to carefully offset all such losses by the gains from higher order cumulants that typically accompany the loss of effective A-factors. In particular, since we are aiming at an optimal bound, in the expansion terms that involve only second order cumulants we need to gain from *all* A-factors. We used a similar but much simpler expansion in our work on CLT for non-Hermitian random matrices, see [17, Prop. 5.3, Eq. (5.10c)], where the additional smallness came from the large distance between two (non-Hermitian) spectral parameters z_1, z_2. However, in [17] it was sufficient to gain only a small proportion of all possible smallness factors since we did not aim at the optimal bound. In the current paper, using a refined combinatorics we manage to extract the zero trace orthogonality effect to the maximal extent; this is the key to obtain the optimal error bound in (3). Similarly, for the proof of (2) we manage to extract the *asymptotic orthogonality* effect between the eigenvectors $\boldsymbol{u}_i$ and their complex conjugates $\overline{\boldsymbol{u}_i}$ optimally, resulting in the bound $|\langle GG^t \rangle| \lesssim 1$, gaining a full power of η over e.g. $\langle GG^* \rangle \sim 1/\eta$.

After this introduction and presenting the main results in the next Sect. 2, we prove the local laws involving two resolvents in Sect. 3. The main inputs for them are the improved bounds on renormalized ("underlined") monomials in several G's in Theorem 5 that are proven in Sect. 5. Note that even though we are interested in local laws only with two G's, due to the cumulant expansion we need to control arbitrary long monomials involving a product of G's and A's.

Notations and conventions. We introduce some notations we use throughout the paper. For integers $k \in \mathbf{N}$ we use the notation $[k] := \{1, \ldots, k\}$. We write **H** for the upper

half-plane $\mathbf{H} := \left\{ z \in \mathbf{C} \big| \Im z > 0 \right\}$. For positive quantities f, g we write $f \lesssim g$ and $f \sim g$ if $f \leq Cg$ or $cg \leq f \leq Cg$, respectively, for some constants $c, C > 0$ which depend only on the constants appearing in (8). 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 $N \geq N_0(D)$. Moreover, we use the convention that $\xi > 0$ denotes an arbitrary small constant which is independent of N.

2. Main Results

We consider real symmetric or complex Hermitian $N \times N$ Wigner matrices W. We formulate the following assumptions on the entries of W.

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

$$\mathbf{E} |\chi_{\mathrm{d}}|^p + \mathbf{E} |\chi_{\mathrm{od}}|^p \leq C_p. \tag{8}$$

In this paper we use the notations $w_2 := \mathbf{E} \chi_{\mathrm{d}}^2, \sigma := \mathbf{E} \chi_{\mathrm{od}}^2$ and their commonly occurring combination $\widetilde{w_2} := w_2 - 1 - \sigma$, and note that $w_2, \widetilde{w_2}, \sigma \in \mathbf{R}$.

Our first main result is the proof of the *Eigenstate Thermalisation Hypothesis*, that in mathematical terms is the proof of an optimal convergence rate of the strong *Quantum Unique Ergodicity (QUE)* for general observables uniformly in the spectrum of W.

Theorem 1 (Eigenstate Thermalization Hypothesis). *Let W be a Wigner matrix satisfying Assumption 1, and denote by $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$ its orthonormal eigenvectors. Then for any deterministic matrix A with $\|A\| \lesssim 1$ it holds*

$$\max_{i,j \in [N]} \left| \langle \boldsymbol{u}_i, A\boldsymbol{u}_j \rangle - \langle A \rangle \delta_{ij} \right| + \max_{i,j \in [N]} \left| \langle \boldsymbol{u}_i, A\overline{\boldsymbol{u}_j} \rangle - \langle A \rangle \langle \boldsymbol{u}_i, \overline{\boldsymbol{u}_j} \rangle \right| \leq \frac{N^\xi}{\sqrt{N}}, \tag{9}$$

with very high probability for any arbitrary small $\xi > 0$.

The first relation in Theorem 1 states that any deterministic observable is essentially diagonal in the eigenbasis of W, in other words the eigenvectors remain *asymptotically orthogonal* when tested against any traceless observable $\langle A \rangle = 0$. The second relation shows the same phenomenon between the eigenbasis $\{\boldsymbol{u}_i\}_{i \in [N]}$ of W and the eigenbasis $\{\overline{\boldsymbol{u}_i}\}_{i \in [N]}$ of W^t. The scalar products $\langle \boldsymbol{u}_i, \overline{\boldsymbol{u}_j} \rangle$ appearing in (9) when $\langle A \rangle \neq 0$ can also be identified. Indeed, the next theorem shows that these two eigenbases are also essentially orthogonal apart from the extreme cases $\sigma = \pm 1$ (see Remark 1 below).

Theorem 2. *Let W be a Wigner matrix satisfying Assumption 1, and denote by $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$ its orthonormal eigenvectors. Recall $\sigma := \mathbf{E}\,\chi^2_{\mathrm{od}}$ and assume $|\sigma| < 1$, then there is a constant $C_\sigma < \infty$ such that*

$$\max_{i,j\in[N]} \left|\langle \boldsymbol{u}_i, \overline{\boldsymbol{u}}_j\rangle\right| \le C_\sigma \frac{N^\xi}{\sqrt{N}}, \tag{10}$$

with very high probability for any arbitrary small $\xi > 0$.

Remark 1. Theorem 2 does not hold for $\sigma = \pm 1$. Indeed, for $\sigma = 1$ the matrix W is real symmetric hence the eigenvectors can be chosen real so $\left|\langle \boldsymbol{u}_i, \overline{\boldsymbol{u}}_j\rangle\right| = \delta_{ij}$. On the other hand for $\sigma = -1$ and $w_2 = 0$ the spectrum is symmetric, i.e. the eigenvalues $\lambda_1 \le \lambda_2 \le \ldots$ and the corresponding eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots$ come in pairs, $\lambda_{N-i+1} = -\lambda_i$ and $\boldsymbol{u}_{N-i+1} = \overline{\boldsymbol{u}}_i$ (up to phase) and thus $\left|\langle \boldsymbol{u}_i, \overline{\boldsymbol{u}}_j\rangle\right| = \delta_{i,N-j+1}$.

The main inputs to prove Theorems 1–2 are the local laws for one and two resolvents (and their transposes) tested against matrices A with $\langle A\rangle = 0$. We recall that in the limit $N \to \infty$ the resolvent $G = G(z) = (W - z)^{-1}$ becomes approximately deterministic. Its deterministic approximation is given by $m = m_{\mathrm{sc}}$, the Stieltjes transform of the Wigner semicircular law, which is given by the unique solution of the quadratic equation

$$-\frac{1}{m} = z + m, \qquad \Im m(z)\Im z > 0. \tag{11}$$

We note that $|m| \le 1$ for any z. In this paper we allow spectral parameters with $\Im z < 0$, in order to conveniently account for possible adjoints of the resolvent since $G(z)^* = G(\bar{z})$. Therefore, in contrast with most papers on local laws, $\Im m_{\mathrm{sc}}$ may be negative and we define $\rho = \rho_{\mathrm{sc}}(z) := \pi^{-1}|\Im m_{\mathrm{sc}}|$ and $\eta := |\Im z|$.

The classical local law (see e.g. in [11,30,38]) for a single resolvent in *averaged* and *isotropic form* states that in the spectral regime $\{z\,|\,N\rho\eta \ge 1\}$ we have

$$|\langle (G - m)A\rangle| \prec \frac{1}{N\eta}, \qquad |\langle \boldsymbol{x}, (G - m)\boldsymbol{y}\rangle| \prec \sqrt{\frac{\rho}{N\eta}} \tag{12}$$

for any deterministic matrix A and vectors $\boldsymbol{x}, \boldsymbol{y}$, with $\|A\|, \|\boldsymbol{x}\|, \|\boldsymbol{y}\| \lesssim 1$. Here $\prec$ indicates the commonly used concept of stochastic domination (see, e.g. [26]) indicating a bound with very high probability up to a factor N^ϵ for any small $\epsilon > 0$, uniformly in $A, \boldsymbol{x}, \boldsymbol{y}$ and in the spectral parameter z as long as $N\rho\eta \ge 1$. The precise definition is as follows:

Definition 1 (*Stochastic Domination*). If

$$X = \left(X^{(N)}(u)\,\middle|\,N \in \mathbf{N}, u \in U^{(N)}\right) \quad \text{and} \quad Y = \left(Y^{(N)}(u)\,\middle|\,N \in \mathbf{N}, u \in U^{(N)}\right)$$

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

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

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

Our key new insight is that whenever the deterministic matrix A in (12) is traceless, then $\langle GA \rangle$ is considerably smaller (by a factor of $\sqrt{\rho\eta}$, in the interesting small η regime) than the general bound (12) predicts. There is no such improvement for the isotropic law.

Theorem 3 (Traceless single G local law). *Fix $\epsilon > 0$, let W be a Wigner matrix satisfying Assumption 1, let $z \in \mathbf{C} \setminus \mathbf{R}$, and let $G(z) = (W - z)^{-1}$. We use the notation $\eta := |\Im z|$, $\rho = \rho_{sc}(z)$, $m = m_{sc}(z)$. Then for $N\eta\rho \geq N^\epsilon$ and for any deterministic matrix A, with $\langle A \rangle = 0$ and $\|A\| \lesssim 1$, we have*

$$|\langle (G - m)A \rangle| = |\langle GA \rangle| \prec \frac{\sqrt{\rho}}{N\sqrt{\eta}}. \tag{13}$$

We prove a similar drastic improvement owing to the traceless observables for local laws involving two resolvents, like $\langle GAGA \rangle$, as well as local laws involving a resolvent and its transpose, $\langle GG^t \rangle$. The isotropic laws are also improved in this case. The precise statements will be given in Remark 3 in Sect. 3. We close the current section by a remark indicating the optimality of the new local law (13).

Remark 2. The local law for $\langle GA \rangle$ in (13) is optimal for G, G^*, as well as $\Im G$. Indeed a simple calculation from [18, Theorem 4.1] shows, for $\Im z > 0$, that

$$\mathbf{E} |\langle \Im GA \rangle|^2 \approx \frac{\langle AA^* \rangle}{2N^2} \left(\frac{\Im m}{\eta} - \Re\partial_z m \right) \sim \frac{\Im m}{N^2\eta}. \tag{14}$$

In fact, in our companion paper we prove that $\langle \Im GA \rangle$ is asymptotically Gaussian with zero expectation and variance given in (14) (see [18, Eq. (94)]). This variance is much smaller than the one without traceless observable, $\text{Var}\langle \Im G \rangle \sim (N\eta)^{-2}$ (see [34]).

3. Quantum Unique Ergodicity: Proof of Theorems 1–2

For integers $J \in \mathbf{N}$ and self-adjoint matrices $B = B^*$ we introduce the J-averaged observables

$$\Xi_J^B := \left(\max_{i_0, j_0} \frac{N}{(2J)^2} \sum_{|i - i_0| < J} \sum_{|j - j_0| < J} |\langle \mathbf{u}_i, B\mathbf{u}_j \rangle|^2 \right)^{1/2}, \tag{15}$$

$$\bar{\Xi}_J^B := \left(\max_{i_0, j_0} \frac{N}{(2J)^2} \sum_{|i - i_0| < J} \sum_{|j - j_0| < J} |\langle \mathbf{u}_i, B\overline{\mathbf{u}_j} \rangle|^2 \right)^{1/2}. \tag{16}$$

The following theorem shows that both averaged overlaps Ξ_J^B, $\bar{\Xi}_J^B$ are essentially bounded if B is traceless, and that $\bar{\Xi}_J^I$ is essentially bounded for $|\sigma| < 1$ and for choosing the identity matrix $B = I$.

Theorem 4. *Fix $\epsilon > 0$, let $J \geq N^\epsilon$, and let $A = A^*$ be a deterministic Hermitian matrix[1] such that $\langle A \rangle = 0$, $\|A\| \lesssim 1$, then it holds*

$$\Lambda_J^A \prec 1 \quad for \quad \Lambda_J^A := \Xi_J^A + \overline{\Xi}_J^A. \tag{17}$$

---

[1] We use the notational convention that the letter A denotes traceless matrices, while B denotes arbitrary matrices.

Similarly, for any $|\sigma| < 1$ it holds

$$\Pi_J \prec 1 \quad for \quad \Pi_J := \bar{\Xi}_J^I, \tag{18}$$

where the error is uniform in $|\sigma| \le 1 - \epsilon'$, for any fixed $\epsilon' > 0$.

Hence, up to a J-averaging, we have the *asymptotic orthogonality* of $\boldsymbol{u}_i$ and $A\boldsymbol{u}_j$, $A\overline{\boldsymbol{u}_j}$ for any i, j and for any traceless A. Similarly, for $|\sigma| < 1$ we have the *asymptotic orthogonality* of $\boldsymbol{u}_i$ and $\overline{\boldsymbol{u}_j}$. Note that in case $\sigma = \pm 1$ the bound (18) does not hold since then $|\Pi_J| \gtrsim \sqrt{N/J}$, as easily seen. Using Theorem 4 we immediately conclude Theorems 1–2 by removing J-averaging with a small J.

Proof. (Theorems 1–2) For the proof of Theorem 1 we may assume by linearity that A is traceless and self-adjoint. For any i, $j \in [N]$, by (17), we have that

$$|\langle \boldsymbol{u}_i, A\boldsymbol{u}_j \rangle|^2 + |\langle \boldsymbol{u}_i, A\overline{\boldsymbol{u}_j} \rangle|^2 \le \frac{J^2}{N}(\Lambda_J^A)^2 \prec \frac{J^2}{N}. \tag{19}$$

Since $J = N^\epsilon$ with $\epsilon > 0$ arbitrary small, together with the definition of $\prec$ in Definition 1, the bound in (19) implies that $|\langle \boldsymbol{u}_i, A\boldsymbol{u}_j \rangle|^2 + |\langle \boldsymbol{u}_i, A\overline{\boldsymbol{u}_j} \rangle|^2 \prec N^{-1}$, concluding the proof of Theorem 1. The proof of Theorem 2 is completely analogous and so omitted. □

As a first step towards the proof of Theorem 4, we first show that Ξ_J^B, $\bar{\Xi}_J^B$ are comparable with $\langle \Im G_1 B \Im G_2 B \rangle$, $\langle \Im G_1 B \Im G_2^t B \rangle$ for suitably chosen spectral parameters in the resolvents $G_i = G(z_i)$. For any $J \in \mathbf{N}$ and $E \in [-2, 2]$ we define $z = z(E, J) = E + i\eta(E, J) \in \mathbf{H}$ implicitly via the equation $N\eta(E, J)\rho(z(E, J)) = J$. Note that this equation has a unique solution $\eta(E, J) > 0$ since the function $\eta \to \eta \Im m(E + i\eta)$ is strictly increasing from 0 to 1. The following simple lemma will be proven at the end of this section.

Lemma 1. *Let $\epsilon > 0$, fix some $J \ge N^\epsilon$ and let $B = B^*$ be an arbitrary deterministic self-adjoint matrix. For $E_1, E_2 \in [-2, 2]$ let $z_i = z(E_i, J)$ and set $G_i = G(z_i)$, $\rho_i = \rho(z_i)$, then we have*

$$\begin{aligned} (\Xi_J^B)^2 &\lesssim \sup_{E_1, E_2} \frac{\langle \Im G_1 B \Im G_2 B \rangle}{\rho_1 \rho_2} \lesssim (\Xi_J^B)^2, \\ (\bar{\Xi}_J^B)^2 &\lesssim \sup_{E_1, E_2} \frac{\langle \Im G_1 B \Im G_2^t B \rangle}{\rho_1 \rho_2} \lesssim (\bar{\Xi}_J^B)^2, \end{aligned} \tag{20}$$

with very high probability.

As the main inputs for Theorem 4, we now state the various local laws and bounds for products of G's, their transposes and deterministic matrices in the following Propositions 1–2. These bounds still involve the key control quantities Λ and Π. Using these bounds, we will prove Theorem 4 by a Gronwall argument on Λ and Π that will immediately imply Theorem 3. Finally, for completeness, we also state a few representative local laws involving two resolvents in Remark 3. The key technical Propositions 1–2 will be proven in Sect. 4.

Proposition 1. *Let $A = A^*$ be a deterministic matrix with $\langle A \rangle = 0$. Fix $\epsilon > 0$ and consider $z \in \mathbf{C} \setminus \mathbf{R}$ such that $L := N\eta\rho \ge N^\epsilon$. Then for $G = G(z)$ we have that*

$$|\langle GA \rangle| \prec \frac{\sqrt{\rho}\Lambda_+^A}{N\sqrt{\eta}}, \tag{21}$$

with $\Lambda_+^A := \Lambda_L^A + \|A\|$.

Proposition 2. *Let $A = A^*$, $A' = (A')^*$ be a deterministic matrix with $\langle A \rangle = 0 = \langle A' \rangle$. Fix $\epsilon > 0$, let W be a Wigner matrix satisfying Assumption 1, let $z_1, z_2 \in \mathbf{C} \setminus \mathbf{R}$, and let $G_i = G(z_i)$, for $i \in \{1, 2\}$. We use the notations $\eta_i := |\Im z_i|$, $\rho_i = \rho_{sc}(z_i)$, $m_i = m_{sc}(z_i)$, and set $L := N \min_i (\eta_i \rho_i)$, $\eta_* := \eta_1 \wedge \eta_2$ and $\rho^* = \rho_1 \vee \rho_2$. Then for $L \geq N^\epsilon$ and setting $\Lambda_+^A = \Lambda_L^A + \|A\|$, $\Pi_+ := \Pi_L + 1$, we have the averaged local laws*

$$|\langle G_1 G_2^{(t)} A \rangle| \prec \frac{\sqrt{\rho^*} \Lambda_+^A}{N \eta_*^{3/2}}, \quad |\langle \Im G_1 A G_2^{(t)} \rangle| \prec \frac{\rho_1 \Lambda_+^A}{L \sqrt{\eta_*}}, \quad |\langle \Im G_1 A \Im G_2^{(t)} \rangle| \prec \frac{\rho_1 \rho_2 \Lambda_+^A}{L \sqrt{\eta_*}}, \tag{22}$$

$$\langle G_1 A G_2^{(t)} A' \rangle = m_1 m_2 \langle A A' \rangle + \mathcal{O}_\prec \left(\Lambda_+^A \Lambda_+^{A'} \sqrt{\frac{\rho^*}{N \eta_*}} \right), \tag{23}$$

$$\langle \Im G_1 A \Im G_2^{(t)} A' \rangle = \Im m_1 \Im m_2 \langle A A' \rangle + \mathcal{O}_\prec \left(\frac{\rho_1 \rho_2 \Lambda_+^A \Lambda_+^{A'}}{\sqrt{L}} \right), \tag{24}$$

where $G^{(t)}$ indicates that the bounds are valid for both choices G or G^t. Moreover, for any deterministic vectors $\mathbf{x}$, $\mathbf{y}$ such that $\|\mathbf{x}\| + \|\mathbf{y}\| \lesssim 1$ we have the isotropic law

$$|\langle \mathbf{x}, G_1 A G_2 \mathbf{y} \rangle| \prec \Lambda_+^A \sqrt{\frac{\rho^*}{\eta_*}}. \tag{25}$$

Additionally, for $|\sigma| < 1$ we have that

$$\langle G_1 G_2^t \rangle = \frac{m_1 m_2}{1 - \sigma m_1 m_2} + \mathcal{O}_\prec \left(\Pi_+^2 \sqrt{\frac{\rho^*}{N \eta_*}} \right), \tag{26}$$

and

$$\langle \Im G_1 \Im G_2^t \rangle = \frac{\Im m_1 \Im m_2 (1 - \sigma^2 |m_1|^2 |m_2|^2)}{|1 - \sigma m_1 m_2|^2 |1 - \sigma \overline{m_1} m_2|^2} + \mathcal{O}_\prec \left(\frac{\rho_1 \rho_2 \Pi_+^2}{\sqrt{L}} \right), \tag{27}$$

where the error is uniform in $|\sigma| \leq 1 - \epsilon'$, for any fixed $\epsilon' > 0$.

Using Lemma 1 and Propositions 1–2 as an input, we now conclude the proof of Theorem 4.

Theorem 4. We start with the proof of (17). Choose $J = N^\epsilon$ with a fixed arbitrary small $\epsilon > 0$, and $E_1, E_2 \in [-2, 2]$. Then by the definition of $z(E_i, J) = E_i + i\eta(E_i, J)$ above Lemma 1 it follows that

$$J = N \eta(E_1, J) \rho(z(E_1, J)) = N \eta(E_2, J) \rho(z(E_2, J)),$$

and thus we obtain from (24) that

$$\frac{\left| \langle \Im G(z(E_1, J)) A \Im G(z(E_2, J))^{(t)} A \rangle \right|}{\rho(z(E_1, J)) \rho(z(E_2, J))} \prec 1 + \frac{(\Lambda_+^A)^2}{J^{1/2}}. \tag{28}$$

By a standard grid argument using the Lipschitz continuity of the resolvent we conclude that (28) remains valid after taking the supremum over $E_1, E_2 \in [-2, 2]$ and therefore from the lower bound in Lemma 1 we conclude

$$(\Lambda_J^A)^2 \prec 1 + \frac{(\Lambda_+^A)^2}{J^{1/2}}. \tag{29}$$

Finally, by (29) it follows that $(\Lambda_J^A)^2 \prec 1$, concluding the proof of (17). The proof of (18) is completely analogous to the proof of (17) above
using the local law (27).
    This concludes the proof of Theorem 4.   □

Theorem 3. This theorem immediately follows from Propositions 1 together with $\Lambda_L^A \prec 1$ obtained in Theorem 4.   □

Remark 3. Proposition 2 combined with the $\Lambda_+^A + \mathbf{1}(|\sigma| < 1)\Pi_+ \prec 1$ obtained in Theorem 4 also provides local laws involving two resolvents as counterparts of the single G local law stated in Theorem 3. For example, with the notations of Proposition 2, we have

$$\langle G_1 A G_2^{(t)} A' \rangle = m_1 m_2 \langle A A' \rangle + \mathcal{O}_\prec \left(\sqrt{\frac{\rho^*}{N\eta_*}} \right), \qquad |\langle \boldsymbol{x}, G_1 A G_2 \boldsymbol{y} \rangle| \prec \sqrt{\frac{\rho^*}{\eta_*}}, \tag{30}$$

and for $|\sigma| < 1$ we also have

$$\langle G_1 G_2^t \rangle = \frac{m_1 m_2}{1 - \sigma m_1 m_2} + \mathcal{O}_\prec \left(\sqrt{\frac{\rho^*}{N\eta_*}} \right). \tag{31}$$

We stated only the local laws for two resolvents where the asymptotic orthogonality mechanism is detected, i.e. if a traceless deterministic matrix is present or if G and G^t appear next to each other and $|\sigma| < 1$. Note that when both mechanisms are simultaneously present, as in the terms with transposes in (22), one may gain from both effects simultaneously, but we refrain from doing this here.
    For comparison, we also list some local laws without exploiting this mechanism:

$$\begin{aligned}
\langle G_1 G_2 \rangle &= \frac{m_1 m_2}{1 - m_1 m_2} + \mathcal{O}_\prec \left(\frac{1}{N\eta_1 \eta_2} \right), \\
\langle \boldsymbol{x}, G_1 G_2 \boldsymbol{y} \rangle &= \frac{m_1 m_2 \langle \boldsymbol{x}, \boldsymbol{y} \rangle}{1 - m_1 m_2} + \mathcal{O}_\prec \left(\frac{\sqrt{\rho^*}}{\sqrt{N\eta_*}\eta^*} \right),
\end{aligned} \tag{32}$$

and for any $|\sigma| \le 1$

$$\langle G_1 G_2^t \rangle = \frac{m_1 m_2}{1 - \sigma m_1 m_2} + \mathcal{O}_\prec \left(\frac{1}{N\eta_1 \eta_2} \right); \tag{33}$$

these relations are proven in our companion paper [18] using Theorem 5 of the present paper. In the most interesting critical case $z := z_1 = \bar{z}_2$ with $\eta = |\Im z| \ll 1$, the leading term in (32) is of order $1/\eta$ with a large error $1/N\eta^2$, while the leading term in (30) is bounded (even zero in the isotropic case) with a negligble error term. The leading terms in (31) and (33) are of course the same, but the error term in (33) is much bigger since it ignores the asymptotic orthogonality mechanism.
    Note that using the decomposition $B = \langle B \rangle + \mathring{B}$, where $\mathring{B}$ is the traceless part of B, a combination of (30)–(33) trivially gives local laws for any product of the form $GBG^{(t)}B'$ for arbitrary deterministic matrices B and B'.

    Finally, we close this section by proving Lemma 1.

Proof. (Lemma 1) We only consider $\langle \Im G_1 B \Im G_2 B \rangle$, the proof of the bounds for $\langle \Im G_1 B \Im G_2^t B \rangle$ is completely analogous and so omitted.

We recall that by the averaged local law for single resolvent in (13) it follows the rigidity of the eigenvalues (see e.g. [26, Theorem 7.6] or [30]):

$$|\lambda_i - \gamma_i| \prec \frac{1}{N^{2/3}\widehat{i}^{1/3}}, \tag{34}$$

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

$$\int_{-\infty}^{\gamma_i} \rho(x) \, dx = \frac{i}{N}, \qquad i \in [N], \tag{35}$$

where we recall $\rho(x) = \rho_{\rm sc}(x) = (2\pi)^{-1}\sqrt{(4 - x^2)_+}$.

For $E_1, E_2 \in [-2, 2]$, we recall that $J = N\eta(E_i, J)\rho(z(E_i, J))$, for $i \in \{1, 2\}$, by the definition of $z(E_i, J) = E_i + i\eta(E_i, J)$ above Lemma 1, and thus, together with (35) we conclude that there is constant C such that for any a, a_0 it holds that

$$|\gamma_a - \gamma_{a_0}| \le \eta(\gamma_{a_0}) \Rightarrow |a - a_0| \le CJ, \quad |a - a_0| \le J \Rightarrow |\gamma_a - \gamma_{a_0}| \le C\eta(\gamma_{a_0}). \tag{36}$$

With a slight abuse of notation we will write this relation as

$$|\gamma_a - \gamma_{a_0}| \lesssim \eta(\gamma_{a_0}) \Leftrightarrow |a - a_0| \lesssim J, \tag{37}$$

since the implicit constants in the $\lesssim$ relation will be irrelevant for our analysis. With the short-hand notations $\Xi = \Xi_J^B$, $z_i = z(E_i, J)$, $G_i = G(z_i)$, $\eta_i = \eta(E_i, J)$, and $\rho_i = \rho(z_i)$, then by (34), and writing $\langle \Im G_1 B \Im G_2 B \rangle$ in spectral decomposition

$$\langle \Im G_1 B \Im G_2 B \rangle = \sum_{ab} \frac{R_{ab} S_{ab}}{N},$$
$$R_{ab} := \frac{\eta_1 \eta_2}{|\lambda_a - z_1|^2 |\lambda_b - z_2|^2}, \qquad S_{ab} := |\langle \boldsymbol{u}_a, B\boldsymbol{u}_b \rangle|^2, \tag{38}$$

it follows that

$$\Xi^2 \lesssim \sup_{E_1, E_2 \in [-2,2]} \frac{\langle \Im G_1 B \Im G_2 B \rangle}{\rho_1 \rho_2} \lesssim \Xi^2, \tag{39}$$

with very high probability on the set where the rigidity bound (34) holds. The lower bound in (39) is trivial by choosing $E_1 = \gamma_{a_0}$ and $E_2 = \gamma_{b_0}$. To prove the upper bound in (39) we use the local averaging formula

$$\sum_{ab} R_{ab} S_{ab} \sim \sum_{a_0 b_0} \frac{1}{(2J)^2} \sum_{\substack{|a-a_0|<J, \\ |b-b_0|<J}} R_{ab} S_{ab} \sim \sum_{a_0 b_0} R_{a_0 b_0} \frac{1}{(2J)^2} \sum_{\substack{|a-a_0|<J, \\ |b-b_0|<J}} S_{ij} \tag{40}$$

for general non-negative R_{ab}, S_{ab} such that $R_{ab} \sim R_{a_0 b_0}$ whenever $|a - a_0| \vee |b - b_0| \le J$ which is applicable for R_{ab} in (38) as a consequence of the rigidity bound in (34), the relation in (37), and the fact that we can choose the $\xi > 0$ so that N^ξ, coming from the rigidity high probability bound, is much smaller than $J \ge N^\epsilon$. Finally we note that $N^{-2} \sum_{ab} R_{ab} = \langle \Im G_1 \rangle \langle \Im G_2 \rangle \sim \rho_1 \rho_2$ by $|\langle \Im G_i - \Im m_i \rangle| \prec \rho_i$ from (13), concluding the proof of the upper bound in (39). □

4. Local Laws for One and Two Resolvents

In this section we prove the local laws in Propositions 1–2. By the self consistent equation for m in (11), and by $G(W - z) = I$, we have

$$G = m - mWG - m\langle G\rangle G + m\langle G - m\rangle G. \tag{41}$$

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

$$\underline{g(W)Wf(W)} := g(W)Wf(W) - \widetilde{\mathbf{E}}g(W)\widetilde{W}(\partial_{\widetilde{W}}f)(W) - \widetilde{\mathbf{E}}(\partial_{\widetilde{W}}g)(W)\widetilde{W}f(W), \tag{42}$$

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 second order 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 (42) only makes sense if it is clear to which W the underline refers, i.e. 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. As a special case we have that

$$\underline{WG} = WG + \langle G\rangle G + \sigma\frac{G^t G}{N} + \frac{\widetilde{w_2}}{N}\,\mathrm{diag}(G)G, \tag{43}$$

recalling the notation $\widetilde{w_2} = w_2 - 1 - \sigma$ from Assumption 1. Then by (41) and and (43), it follows that

$$G = m - m\underline{WG} + \frac{m\sigma}{N}G^t G + \frac{\widetilde{w_2}}{N}\,\mathrm{diag}(G)G + m\langle G - m\rangle G. \tag{44}$$

From (44) one can already see that in order to get a local law for G it is essential to estimate the underlined term $\underline{WG}$ in averaged and isotropic sense. In order to prove Proposition 2 we need bounds for underlined terms involving not only one G but also two G's (see e.g. (56) below). We now state the bound for these terms and for longer products of resolvents and deterministic matrices both in an averaged and in an isotropic sense, since the proof for products with more than two resolvents is very similar to the cases we need.

For $l \in \mathbf{N}$ we consider renormalised alternating products of resolvents $G_1, \ldots, G_l$ and deterministic matrices $B_1, \ldots, B_l$ in averaged and isotropic form,

$$\langle\underline{WG_1B_1G_2B_2\cdots G_lB_l}\rangle, \quad \langle x, \underline{WG_1B_1G_2\cdots B_{l-1}G_l}y\rangle. \tag{45}$$

Each resolvent G_k is evaluated at a (potentially) different spectral parameter $z_k \in \mathbf{C}\setminus\mathbf{R}$ and other than $G_k = G(z_k)$ we allow each resolvent to be transposed and/or being the imaginary part, i.e.

$$G_k \in \left\{G(z_k), G(z_k)^t, \Im G(z_k), (\Im G(z_k))^t\right\}. \tag{46}$$

Note that adjoints of resolvents can be included in the products by conjugating the spectral parameter since $G(z)^* = G(\overline{z})$. For a given product of the form (45) we consider three sets $\mathfrak{i}, \mathfrak{a}, \mathfrak{t}$ of indices, recording special structural properties of G_k, B_k. By definition, the set $\mathfrak{i} \subset [l]$ collects those indices $k \in [l]$ for which $G_k \in \left\{\Im G(z_k), (\Im G(z_k))^t\right\}$. For the choice of the sets $\mathfrak{a}, \mathfrak{t}$ we allow a certain freedom described in the theorem.

Theorem 5. *Fix $\epsilon > 0$, let $l \in \mathbf{N}$, $z_1, \ldots, z_l \in \mathbf{C} \setminus \mathbf{R}$ and for $k \in [l]$ let G_k be as in (46) and 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*

$$L := N \min_k (\eta_k \rho_k), \quad \rho^* := \max_k \rho_k, \quad \eta_* := \min_k \eta_k, \tag{47}$$

with $\eta_k := |\Im z_k|$, $\rho_k := \rho(z_k) = |\Im m(z_k)| / \pi$ and assume $L \geq N^\epsilon$ and $\eta_* \lesssim 1$. Let $\mathfrak{a}, \mathfrak{t}$ denote disjoint sets of indices, $\mathfrak{a} \cap \mathfrak{t} = \emptyset$, 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 G_k, G_{k+1} is transposed, where in the averaged case and $k = l$ it is understood that $G_{l+1} = G_1$. Recall the notations $\Pi_+ := \Pi_L + 1$, $\Lambda_+^B := \Lambda_L^B + \|B\|$. Then with $a := |\mathfrak{a}|$, $t := |\mathfrak{t}|$, we have the following bounds:

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

$$\left| \langle \underline{W G_1 B_1 G_2 B_2 \cdots G_l B_l} \rangle \right| \prec \frac{\rho^*}{N \eta_*^l}. \tag{48}$$

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

$$\left| \langle \underline{W G_1 B_1 G_2 B_2 \cdots G_l B_l} \rangle \right| \prec \frac{(\sqrt{N} \eta_*)^{a+t}}{N \eta_*^l} \sqrt{\frac{\rho^*}{N \eta_*}} \Pi_+^t \prod_{k \in \mathfrak{a}} \Lambda_+^{B_k}. \tag{49}$$

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

$$\left| \langle \boldsymbol{x}, \underline{G_1 B_1 \cdots G_j B_j W G_{j+1} B_{j+1} \cdots B_{l-1} G_l} \boldsymbol{y} \rangle \right|$$

$$\prec \frac{(\sqrt{N} \eta_*)^{a+t}}{\eta_*^{l-1}} \sqrt{\frac{\rho^*}{N \eta_*}} \Pi_+^t \prod_{k \in \mathfrak{a}} \Lambda_+^{B_k}, \tag{50}$$

where the $j = 0$ case is understood as $\langle \boldsymbol{x}, \underline{W G_1 B_1 \cdots B_{l-1} G_l} \boldsymbol{y} \rangle$.

In case $\prod_{k \in \mathfrak{i}} \rho_k \lesssim (\rho^*)^{b+1}$, the bounds (48)–(50) remain valid if the rhs. are multiplied by the factor $(\rho^*)^{-b-1} \prod_{k \in \mathfrak{i}} \rho_k$, where $b := l$ in case of (48), $b := l - a - t$ in case of (49), and $b := l - a - t - 1$ in case of (50). Moreover, for any $\eta_* \geq 1$ we have the bounds

$$\left| \langle \underline{W G_1 B_1 G_2 B_2 \cdots G_l B_l} \rangle \right| \prec \frac{1}{N \eta_*^l},$$

$$\left| \langle \boldsymbol{x}, \underline{G_1 B_1 \cdots G_j B_j W G_{j+1} B_{j+1} \cdots B_{l-1} G_l} \boldsymbol{y} \rangle \right| \prec \frac{1}{N^{1/2} \eta_*^l}. \tag{51}$$

Remark 4. (Asymptotic orthogonality effect) The main result of Theorem 5 are (49) and its isotropic counterpart (50). The essential part is the factor $(\sqrt{N} \eta_*)^{a+t}$ in (49) since the additional factors Π_+ and Λ_+ are a posteriori shown to be essentially $\mathcal{O}(1)$, c.f. Theorem 4. Compared with the robust bound (48) in the relevant small η_* regime the bound (49) represents an improvement of $\sqrt{N} \eta_*$ for each occurrence when the asymptotic orthogonality can be exploited, either due to a traceless matrix B or to a switch between a resolvent and its transpose. In addition, compared to the robust averaged bound (48) there is a further improvement of $\sqrt{\rho^*/N \eta_*}$ in (49) if at least one orthogonality effect is exploited, enabling the optimal GA local law in Theorem 3. We note that in case when $\sqrt{N} \eta_* \gg 1$ the robust bounds (48) and (50) with $a + t = 0$ are always available also in the presence of traceless deterministic matrices and alternating G, G^t simply by choosing the sets $\mathfrak{a}$, $\mathfrak{t}$ to be empty.

Remark 5. (Alternative renormalisation) In (42) we defined the renormalisation with respect to an independent copy of W while in some previous papers [17] the same notation was used to denote the renormalisation with respect to a suitable reference ensemble (e.g. the GUE-ensemble in the present paper or the complex Ginibre ensemble in case of [17]). However, mostly these two possible definitions only differ in some subleading terms. For example, denoting the renormalisation with respect to an independent GUE-matrix by

$$\underline{Wf(W)}_{\mathrm{GUE}} := Wf(W) - \widetilde{\mathbf{E}}_{\mathrm{GUE}}\widetilde{W}(\partial_{\widetilde{W}} f)$$

we have trivially

$$\langle \underline{WG} \rangle - \langle \underline{WG}_{\mathrm{GUE}} \rangle = \sigma \frac{\langle G^t G \rangle}{N} + \widetilde{w}_2 \frac{\langle \mathrm{diag}(G)G \rangle}{N} = \mathcal{O}_\prec \left(\frac{\rho}{N\eta} \right).$$

The difference between the two renormalisations becomes relevant in Theorem 5 only whenever at least one transposed resolvent occurs since then for example

$$\langle \underline{WG^t} \rangle - \langle \underline{WG^t}_{\mathrm{GUE}} \rangle = \sigma \langle G \rangle^2 + \widetilde{w}_2 \frac{\langle \mathrm{diag}(G)G \rangle}{N} \sim 1.$$

However, this is the only relevant case and the statement of Theorem 5 holds true verbatim if $\underline{Wf(W)}$ is replaced by $\underline{Wf(W)}_{\mathrm{GUE}}$ in case no resolvents are transposed, c.f. Remark 7 in Sect. 5.

Using the bounds for the underlined terms in Theorem 5 we conclude the proof of Proposition 1–2. We start with the proof of the local law for $\langle GA \rangle$ and then we prove the local laws and bounds for two G's.

Proof. (Proposition 1) Using the equation for G in (44), we start writing the equation for GA:

$$GA = mA - m\underline{WG}A + m\langle G - m \rangle GA + \frac{m\sigma}{N} G^t GA + \frac{\widetilde{w}_2}{N} \mathrm{diag}(G)GA, \quad (52)$$

where we recall the definition of $\underline{WG}$ in (42). Then, taking the average in (52), using that $\langle A \rangle = 0$ and that $|\langle G - m \rangle| \prec (N\eta)^{-1}$ by the first bound in (13), we conclude

$$\left[1 + \mathcal{O}_\prec \left(\frac{1}{N\eta} \right) \right] \langle GA \rangle = -m\langle \underline{WG}A \rangle + \frac{m\sigma}{N} \langle G^t GA \rangle + \frac{\widetilde{w}_2}{N} \langle \mathrm{diag}(G)GA \rangle.$$

Next we notice that

$$\frac{1}{N} |\langle G^t GA \rangle| \leq \frac{1}{N} \langle |G| \rangle^{1/2} \langle GA|G^t|AG^* \rangle^{1/2} \prec \frac{\Lambda_+ \sqrt{\rho}}{N\sqrt{\eta}}, \tag{53}$$

where in the last inequality we used Lemma 5, and the notation $\Lambda_+ := \Lambda_L^A + \|A\|$. Additionally, we also have that

$$\frac{1}{N} |\langle \mathrm{diag}(G)GA \rangle| = \left| \frac{1}{N^2} \sum_i G_{ii}(GA)_{ii} \right| \prec \frac{1}{N}, \tag{54}$$

where we used that $|G_{ii}| + |(GA)_{ii}| \prec 1$ by (12). Combining (53)–(54) we finally conclude that

$$\langle GA \rangle = -m \langle \underline{WG} A \rangle + \mathcal{O}_\prec \left(\frac{\Lambda_+ \sqrt{\rho}}{N \sqrt{\eta}} \right) = \mathcal{O}_\prec \left(\frac{\Lambda_+ \sqrt{\rho}}{N \sqrt{\eta}} \right), \tag{55}$$

where we used $|\langle \underline{WG} A \rangle| \prec \Lambda_+ \rho^{1/2} N^{-1} \eta^{-1/2}$ by (49). This concludes the proof of (21). $\square$

Next, using the local law for single resolvent proven above, we proceed with the proof of the bounds for for products of two resolvents and deterministic traceless matrices.

Proof. (Proposition 2) We start writing the equation for generic products of two G's $G_1 B_1 G_2 B_2$, where $G_i = (W - z_i)^{-1}$ and B_1, B_2 are deterministic matrices. Using the equation (52) for $G_1 B_1$, writing $G_2 = m_2 + (G_2 - m_2)$, we obtain

$$\begin{aligned}
G_1 B_1 G_2 B_2 = {} & m_1 m_2 B_1 B_2 + m_1 B_1 (G_2 - m_2) B_2 - m_1 \underline{WG_1 B_1 G_2} B_2 \\
& + m_1 \langle G_1 - m_1 \rangle G_1 B_1 G_2 B_2 + m_1 \langle G_1 B_1 G_2 \rangle G_2 B_2 \\
& + \frac{m_1 \sigma}{N} G_1^t G_1 B_1 G_2 B_2 + \frac{m_1 \sigma}{N} (G_1 B_1 G_2)^t G_2 B_1 \\
& + \frac{m_1 \widetilde{w_2}}{N} \operatorname{diag}(G_1) G_1 B_1 G_2 B_2 + \frac{m_1 \widetilde{w_2}}{N} \operatorname{diag}(G_1 B_1 G_2) G_2 B_2, \tag{56}
\end{aligned}$$

where we used that

$$\begin{aligned}
\underline{WG_1 B_1 G_2} = {} & \underline{WG_1} B_1 G_2 + \langle G_1 B_1 G_2 \rangle G_2 + \frac{\sigma}{N} (G_1 B_1 G_2)^t G_2 \\
& + \frac{\widetilde{w_2}}{N} \operatorname{diag}(G_1 B_1 G_2) G_2, \tag{57}
\end{aligned}$$

with $\underline{WG_1}$ from (43). The identity in (57) follows by the definition of underline in (42).

Remark 6. For notational simplicity, throughout the proof of Proposition 2 we use the notations

$$\Lambda_+ := \Lambda_+^A \vee \Lambda_+^{A'},$$

rather than distinguishing the different Λ's. However, the proof naturally yields in fact a factor of Λ_+^A for each traceless A giving the bounds in Propositions 2.

Proof. (Eq. (22)) We focus only on the proof of the bound for $\langle G_1 G_2 A \rangle$, the bounds for $\langle G_1^t G_2 A \rangle$, $\langle \Im G_1 A G_2 \rangle$, $\langle \Im G_1^t A G_2 \rangle$, $\langle \Im G_1 A \Im G_2 \rangle$, and $\langle \Im G_1^t A \Im G_2 \rangle$ are completely analogous and so omitted, modulo the bound for the underlined term. In particular, the bound in (62) has to be replaced by

$$|\langle \underline{WG_2 \Im G_1} A \rangle| \prec \frac{\rho_1 \Lambda_+}{\sqrt{N K \eta_*}}, \quad |\langle \underline{W \Im G_2 \Im G_1} A \rangle| \prec \frac{\rho_1 \rho_2 \Lambda_+}{\sqrt{N K \eta_*}},$$

for $\langle \Im G_1 A G_2 \rangle$, $\langle \Im G_1^t A G_2 \rangle$ and $\langle \Im G_1 A \Im G_2 \rangle$, $\langle \Im G_1^t A \Im G_2 \rangle$, respectively, where $K := N \eta_* \rho^*$.

Choosing $B_1 = I$ and $B_2 = A$ in (56), with $\langle A \rangle = 0$, using $|\langle G_1 - m_1 \rangle| \prec (N\eta_1)^{-1}$ we find that

$$
\left[1 + \mathcal{O}_\prec \left(\frac{1}{N\eta_2} \right) \right] \langle G_1 G_2 A \rangle = -m_1 \langle \underline{W G_1 G_2} A \rangle + m_1 \langle G_2 A \rangle
$$
$$
+ m_1 \langle G_1 G_2 \rangle \langle G_2 A \rangle + \frac{m_1 \sigma}{N} \langle G_1^t G_1 G_2 A \rangle
$$
$$
+ \frac{m_1 \sigma}{N} \langle (G_1 G_2)^t G_2 A \rangle + \frac{m_1 \widetilde{w_2}}{N} \langle \mathrm{diag}(G_1) G_1 G_2 A \rangle
$$
$$
+ \frac{m_1 \widetilde{w_2}}{N} \langle \mathrm{diag}(G_1 G_2) G_2 A \rangle. \tag{58}
$$

Then using Cauchy-Schwarz we have that

$$
\frac{1}{N} |\langle G_1 G_2 A G_1^t \rangle| \leq \frac{1}{N} \langle G_1 G_1^* \rangle^{1/2} \langle G_2 A G_1^t (G_1^t)^* A G_2^* \rangle^{1/2}
$$
$$
\prec \frac{\sqrt{\rho_1}}{N\eta_1 \sqrt{\eta_2}} \langle \Im G_2 A \Im G_1^t A \rangle^{1/2} \tag{59}
$$
$$
\prec \frac{\rho_1 \sqrt{\rho_2} \Lambda_+}{N\eta_1 \sqrt{\eta_2}} \leq \frac{\rho^* \sqrt{\rho_*} \Lambda_+}{\sqrt{NK}\eta_1\eta_2},
$$

where we used the Ward identity, that $\langle \Im G_1 \rangle \prec \rho_1$, and that $K = N\eta_* \rho^*$. In the penultimate inequality of (59) we also used Lemma 5 to prove that $\langle \Im G_2 A \Im G_1^t A \rangle \prec \rho_1 \rho_2 \Lambda_+^2$. Using exactly the same computations we conclude the same bound for $\langle (G_1 G_2)^t G_2 A \rangle$ as well. Now we show that the terms with a pre-factor $\widetilde{w_2}$ are negligible. We start with

$$
\frac{1}{N} |\langle \mathrm{diag}(G_1) G_1 G_2 A \rangle| = \left| \frac{1}{N^2} \sum_i G_{ii} (G_1 G_2 A)_{ii} \right| \prec \frac{\sqrt{\rho_1 \rho_2}}{N\sqrt{\eta_1 \eta_2}}, \tag{60}
$$

obtained using that $|G_{ii}| \prec 1$, by the isotropic law (12), and $|(G_1 G_2 A)_{ii}| \prec \sqrt{\rho_1 \rho_2 / \eta_1 \eta_2}$ by a simple Schwarz inequality. The bound for $|\langle \mathrm{diag}(G_1 G_2) G_2 A \rangle|$ is analogous and so omitted. Combining (58) with (59)–(60), using the bound $|\langle G_2 A \rangle| \prec \sqrt{\rho_2} N^{-1} \eta_2^{-1/2}$ by (21), and dividing by the factor in the lhs. of (58), we conclude that

$$
\langle G_1 G_2 A \rangle = -m_1 \langle \underline{W G_1 G_2} A \rangle + m_1 \langle G_1 G_2 \rangle \langle G_2 A \rangle + \mathcal{O}_\prec \left(\frac{\rho^* \Lambda_+}{\sqrt{NK}\eta_*} \right)
$$
$$
= \mathcal{O}_\prec \left(\frac{\rho^* \Lambda_+}{\sqrt{NK}\eta_*} \right), \tag{61}
$$

where to go from the first to the second line we used that $|\langle G_2 A \rangle| \Lambda_+ \prec \rho_2^{1/2} N^{-1} \eta_2^{-1/2}$ by (55), that $|\langle G_1 G_2 \rangle| \prec \sqrt{\rho_1 \rho_2 / (\eta_1 \eta_2)}$ by a Schwarz inequality, and that

$$
\left| \langle \underline{W G_1 G_2} A \rangle \right| \prec \frac{\rho^* \Lambda_+}{\sqrt{NK}\eta_*}, \tag{62}
$$

by (49). This concludes the proof of the bound of $|\langle G_1 G_2 A \rangle|$. $\qquad \square$

Proof. (Eq.(25) for $\langle x, G_1 A G_2 y \rangle$) Using the bound for $\langle G_1 G_2 A \rangle$ and the estimates in Lemma 5 below as an input, the proof of the bound

$$|\langle x, G_1 A G_2 y \rangle| \prec \Lambda_+ \sqrt{\frac{\rho^*}{\eta_*}}, \tag{63}$$

follows by exactly the same computations as in the proof of the bound for $|\langle G_1 G_2 A \rangle|$. □

Proof. (Local laws for two resolvents: Eqs. (23) and (26)) We focus only on the proof of the local law for $\langle G_1 A G_2 A' \rangle$, the proof of the local law for $\langle G_1^t A G_2 A' \rangle$ is exactly the same. The prof of the local law for $\langle G_1 G_2^t \rangle$ is also analogous to the proof of the local law for $\langle G_1 A G_2 A' \rangle$ with the only difference that the multiplicative factor in the rhs. of (64) has to be replaced by

$$1 - \sigma m_1 m_2 + \mathcal{O}_{\prec} \left(\frac{1}{N \eta_*} \right).$$

This difference does not create any change since for $|\sigma| < 1$ the stability factor $1 - \sigma m_1 m_2$ is bounded from below by $1 - |\sigma|$.

Choosing $B_1 = A$ and $B_2 = A'$ in (56), with $\langle A \rangle = \langle A' \rangle = 0$, and using that $|\langle G_1 - m_1 \rangle| \prec (N \eta_1)^{-1}$, we find that

$$\begin{aligned}
\left(1 + \mathcal{O} \left(\frac{1}{N \eta_2} \right) \right) &\langle G_1 A G_2 A' \rangle \\
= &\; m_1 m_2 \langle A A' \rangle - m_1 \underline{\langle W G_1 A G_2 A' \rangle} + \langle G_1 A G_2 \rangle \langle G_2 A' \rangle \\
&+ \frac{m_1 \sigma}{N} \langle G_1^t G_1 A G_2 A' \rangle + \frac{m_1 \sigma}{N} \langle (G_1 A G_2)^t G_2 A' \rangle \\
&+ \frac{m_1 \widetilde{w_2}}{N} \left[\langle \mathrm{diag}(G_1) G_1 A G_2 A' \rangle + \langle \mathrm{diag}(G_1 A G_2) G_2 A' \rangle \right].
\end{aligned} \tag{64}$$

We start with the bound

$$\begin{aligned}
\frac{1}{N} |\langle G_1^t G_1 A G_2 A' \rangle| &\leq \frac{1}{N} \langle G_1 A | G_2 | A G_1^* \rangle^{1/2} \langle (G_1^t)^* A' | G_2 | A' (G_1)^t \rangle^{1/2} \\
&= \frac{1}{N \eta_1} \langle \Im G_1 A | G_2 | A \rangle^{1/2} \langle \Im G_1^t A' | G_2 | A' \rangle^{1/2} \prec \frac{\rho_1 \Lambda_+^2}{N \eta_1}, \quad (65)
\end{aligned}$$

where we used a Schwarz inequality and that $|\langle \Im G_1 A | G_2 | A \rangle| \prec \rho_1 \Lambda_+^2$ by Lemma 5 below. Following exactly the same computations we conclude that $|\langle (G_1 A G_2)^t G_2 A' \rangle| \prec \Lambda_+^2 \rho_2 \eta_2^{-1}$. Similarly, we also bound

$$\frac{1}{N} |\langle \mathrm{diag}(G_1) G_1 A G_2 A' \rangle| = \left| \frac{1}{N^2} \sum_i (G_1)_{ii} (G_1 A G_2 A')_{ii} \right| \prec \frac{\sqrt{\rho_1 \rho_2}}{N \sqrt{\eta_1 \eta_2}}, \tag{66}$$

where we used that $|(G_1)_{ii}| \prec 1$ and that $|(G_1 A G_2 A')_{ii}| \prec \sqrt{\rho_1 \rho_2 / (\eta_1 \eta_2)}$ by a Schwarz inequality. The bound for $|\langle \mathrm{diag}(G_1 A G_2) G_2 A' \rangle|$ is completely analogous and so omitted.

Combining (64) with (65)–(66), and using that

$$|\langle G_2 A' \rangle| \prec \frac{\sqrt{\rho_2} \Lambda_+}{N \sqrt{\eta_2}}, \qquad |\langle G_2 G_1 A \rangle| \prec \frac{\rho^* \Lambda_+}{\sqrt{N K \eta_*}},$$

by (21) and (22), respectively, we conclude that

$$\langle G_1 A G_2 A' \rangle = m_1 m_2 \langle A A' \rangle - m_1 \langle \underline{W G_1 A G_2} A' \rangle + \mathcal{O}_\prec \left(\frac{\rho^* \Lambda_+^2}{K^2} \right)$$

$$= m_1 m_2 \langle A A' \rangle + \mathcal{O}_\prec \left(\frac{\rho^* \Lambda_+^2}{\sqrt{K}} \right). \tag{67}$$

To go from the first to the second line of (67) we used that $|\langle \underline{W G_1 A G_2 A'} \rangle| \prec \Lambda_+^2 \rho^* K^{-1/2}$ by (49). This concludes the proof of the local law for $\langle G_1 A G_2 A' \rangle$. □

In order to conclude the proof of Proposition 2 we are only left with the averaged local laws for $\Im G_1 A \Im G_2 A'$ and $\Im G_1 A \Im G_2^t A'$ in (24) and for $\Im G_1^t \Im G_2$ in (27).

Proof. (Eq. (24)) We present only the proof of the local law for $\langle \Im G_1 A \Im G_2 A' \rangle$, the proof for $\langle \Im G_1 A \Im G_2^t A' \rangle$ is identical and so omitted. We start with the formula analogous to (56) but with $\Im G$'s instead of G's, generating altogether twelve terms with a $1/N$ pre-factor. Ten of them can be estimated by $\Lambda_+^2 \rho_1 \rho_2 L^{-1}$ exactly as in (65)–(66) by writing out $2i\Im G_i = G_i - G_i^*$. Note that whenever the analogue of (65) is used, but with $G_1^{(t)} G_2$ instead of $G_1^{(t)} G_1$, we could gain the necessary factor $\sqrt{\rho_1 \rho_2 / (\eta_1 \eta_2)}$ instead of only ρ_1 / η_1 in the first Schwarz inequality in (65). Keeping the two special $1/N$ terms, this gives the expansion

$$\langle \Im G_1 A \Im G_2 A' \rangle + \mathcal{O}_\prec \left(\frac{\Lambda_+^2 \rho_1 \rho_2}{L} \right)$$

$$= \Im m_1 \Im m_2 \langle A A' \rangle + \Im m_1 \langle \Im (G_2 - m_2) A' A \rangle + \overline{m_1} \overline{\langle G_1 - m_1 \rangle} \langle \Im G_1 A \Im G_2 A' \rangle$$

$$+ \Im [m_1 \langle G_1 - m_1 \rangle] \langle G_1 A \Im G_2 A' \rangle - \Im m_1 \langle \underline{W G_1 A \Im G_2 A'} \rangle$$

$$- \overline{m_1} \langle \underline{W \Im G_1 A \Im G_2} A' \rangle + \Im m_1 \langle G_1 A G_2 \rangle \langle \Im G_2 A' \rangle$$

$$+ \Im m_1 \langle G_1 A \Im G_2 \rangle \langle G_2^* A' \rangle + \overline{m_1} \langle \Im G_1 A G_2 \rangle \langle \Im G_2 A' \rangle$$

$$+ \overline{m_1} \langle \Im G_1 A \Im G_2 \rangle \langle G_2^* A' \rangle + \frac{\sigma \Im m_1}{N} \langle G_1 A \Im G_2 A' G_1^t \rangle$$

$$+ \frac{\sigma \overline{m_1}}{N} \langle \Im [G_1^t G_1] A \Im G_2 A' \rangle. \tag{68}$$

The two remaining $1/N$ terms, where $\Im G_2$ is separated from G_1 by A's, are estimated as follows:

$$|\langle G_1 A \Im G_2 A' G_1^t \rangle| \leq \langle G_1 A \Im G_2 A' G_1^* \rangle^{1/2} \langle (G_1^t)^* A \Im G_2 A' G_1^t \rangle^{1/2} \prec \frac{N \rho_1 \rho_2 \Lambda_+^2}{L}, \tag{69}$$

where in the last inequality we used the Ward identity and Lemma 5 below. Inserting (69), the local law $|\langle G_2 - m_2 \rangle| \prec (N\eta_2)^{-1}$ and (21)–(22) into (68) we conclude that

$$\langle \Im G_1 A \Im G_2 A' \rangle = \Im m_1 \Im m_2 \langle A A' \rangle - \Im m_1 \langle \underline{W G_1 A \Im G_2 A'} \rangle$$

$$- \overline{m_1} \langle \underline{W \Im G_1 A \Im G_2 A'} \rangle + \mathcal{O}_\prec \left(\frac{\Lambda_+^2 \rho_1 \rho_2}{L} \right). \tag{70}$$

Finally, combining (70) with the bound for $\langle \underline{W G_1 A \Im G_2 A'} \rangle$, $\langle \underline{W \Im G_1 A \Im G_2 A'} \rangle$ in (49), we conclude that

$$\langle \Im G_1 A \Im G_2 A' \rangle = \Im m_1 \Im m_2 \langle A A' \rangle + \left(\frac{\Lambda_+^2 \rho_1 \rho_2}{\sqrt{L}} \right). \tag{71}$$

□

Proof. (Local law for $\langle \Im G_1 \Im G_2^t \rangle$: Eq. (27)) We closely follow the proof of $\langle \Im G_1 A \Im G_2 A' \rangle$, hence we only explain the differences. As each traceless A, A' between two resolvents gave rise to a factor Λ_+ in the proof of $\langle \Im G_1 A \Im G_2 A' \rangle$, here the fact that a resolvent is followed by its transpose gives rise to a factor Π_+. Keeping this modification in mind, in the basic equation for $\langle \Im G_1 \Im G_2^t \rangle$ we can again estimate all the $1/N$ terms as in (65)–(66) and (69) by $(1 + \Pi^2)\rho_1\rho_2 L^{-1}$. Then, using the local law $|\langle G_i - m_i \rangle| \prec (N\eta_i)^{-1}$, similarly to (68), we conclude that

$$\langle \Im G_1 \Im G_2^t \rangle = \Im m_1 \Im m_2 - \Im m_1 \underline{\langle W G_1 \Im G_2^t \rangle} - \overline{m_1} \underline{\langle W \Im G_1 \Im G_2^t \rangle}$$
$$+ \sigma \Im m_1 \langle G_1 G_2^t \rangle \langle \Im G_2 \rangle + \sigma \Im m_1 \langle G_1 \Im G_2^t \rangle \langle G_2^* \rangle$$
$$+ \sigma \overline{m_1} \langle \Im G_1 G_2^t \rangle \langle \Im G_2 \rangle$$
$$+ \sigma \overline{m_1} \langle \Im G_1 \Im G_2^t \rangle \langle G_2^* \rangle + \mathcal{O}_\prec \left(\frac{\Pi_+^2 \rho_1 \rho_2}{L} \right), \tag{72}$$

where we used $\Pi_+ := 1 + \Pi$. Note that several "large" terms remained in (72) in contrast to (70) since the analogues of $\langle \Im G_2 A \rangle$ and $\langle G_2^* A \rangle$ in (68) are now not small. Then using the bounds in (49) for the underlined terms in (72), and the local laws

$$\langle G_1 G_2^t \rangle = \frac{m_i \Im m_j}{1 - \sigma m_1 m_2} + \mathcal{O}_\prec \left(\frac{\rho^* \Pi_+^2}{\sqrt{K}} \right),$$
$$\langle G_i \Im G_j^t \rangle = \frac{m_i \Im m_j}{(1 - \sigma m_i m_j)(1 - \sigma m_i \overline{m_j})} + \mathcal{O}_\prec \left(\frac{\rho_j \Pi_+^2}{\sqrt{K}} \right), \tag{73}$$

we conclude

$$\langle \Im G_1 \Im G_2^t \rangle = \frac{\Im m_1 \Im m_2 (1 - \sigma^2 |m_1 m_2|^2)}{|1 - \sigma \overline{m_1} m_2|^2 (1 - \sigma m_1 m_2)} + \sigma \overline{m_1} \langle \Im G_1 \Im G_2^t \rangle \langle G_2^* \rangle$$
$$+ \mathcal{O}_\prec \left(\frac{\Pi_+^2 \rho_1 \rho_2}{\sqrt{L}} \right). \tag{74}$$

We remark that the second local law in (73) follows analogously to (26). Finally, writing $\langle G_2^* \rangle$ in the last term in the rhs. of (74) as $\langle G_2^* \rangle = \overline{m_2} + \langle (G_2 - m_2)^* \rangle$, we conclude (27). □

This concludes the proof of Proposition 2.   □

5. Feynman Diagrams: Proof of Theorem 5

For the sake of simpler notations we abbreviate

$$\eta := \eta_* = \min_k \eta_k, \quad \rho := \rho^* = \max_k \rho_k, \quad K := N\eta_*\rho^* \geq L = N \min_k (\eta_k \rho_k) \tag{75}$$

and within this section write ρ^i and Λ_+^a with $i := |\mathfrak{i}|$, $a = |\mathfrak{a}|$ and $\Lambda_+ := \max_{k \in \mathfrak{a}} \Lambda_+^{B_k}$ rather than carrying the products $\prod_{k \in \mathfrak{i}} \rho_i$ and $\prod_{k \in \mathfrak{a}} \Lambda_+^{B_k}$. Within the formal proof of Theorem 5 we argue, however, that the proof naturally yields the latter. In order to present the main body of the proof of Theorem 5 more concisely we make four simplifying assumptions, the removal of which are routine modifications.[2]

---

[2] In Appendix A of the arXiv:2012.13215 version of the present paper the necessary minor technical changes to remove each of these four simplifying assumptions are addressed in details.

(A-i) we assume that $w_2 = 1 + \sigma$,

(A-ii) we consider the regime $\eta \lesssim 1$,

(A-iii) for the averaged case we assume that $l \in \mathfrak{a} \cup \mathfrak{t}$ whenever $|\mathfrak{a} \cup \mathfrak{t}| \neq 0$,

(A-iv) in the isotropic bound we only consider $j \geq 1$.

5.1. Graphical representation of the cumulant expansion. Using multiple cumulant expansions we expand the high moments

$$\mathbf{E} \left| \langle \underline{W G_1 B_1 G_2 B_2 \cdots G_l B_l} \rangle \right|^{2p}$$

and

$$\left| \langle \boldsymbol{x}, \underline{G_1 B_1 \cdots G_j B_j W G_{j+1} B_{j+1} \cdots B_{l-1} G_l} \boldsymbol{y} \rangle \right|^{2p}$$

as a polynomial of resolvent entries for any $p \in \mathbf{N}$. More precisely, we iteratively use the expansion

$$\mathbf{E}\, w_{ab} f(W) = \sum_{k=1}^{R} \sum_{\boldsymbol{\alpha} \in \{ab, ba\}^k} \frac{\kappa(ab, \boldsymbol{\alpha})}{k!} \mathbf{E}\, \partial_{\boldsymbol{\alpha}} f(W) + \Omega_R \tag{76}$$

with some explicit error term Ω_R (see [28, Proposition 3.2 and Appendix C] applied with $\mathcal{N}(ab) = \{ab, ba\}$, or the previous works with slightly different error terms [14] and [34, Lemmata 3.1, 7.1]) which for our application can easily be seen to be $\mathcal{O}\left(N^{-2p}\right)$ if $R = 12p$. Here for a k-tuple of double indices $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ we use the shorthand notation $\kappa(ab, (\alpha_1, \ldots, \alpha_k)) := \kappa(w_{ab}, w_{\alpha_1}, \ldots, w_{\alpha_k})$ for the joint cumulant of the random variables $w_{ab}, w_{\alpha_1}, \ldots, w_{\alpha_k}$ and set $\partial_{\boldsymbol{\alpha}} := \partial_{w_{\alpha_1}} \cdots \partial_{w_{\alpha_k}}$, $\partial_{ab} := \partial_{w_{ab}}$. We wish to express the cumulant factors in (76) as a matrix with a, b matrix elements. To encode the fact the that cumulants have slightly different combinatorics for $a = b$ and $a \neq b$, we rewrite (76) as

$$\mathbf{E}\, w_{ab} f(W) = \sum_{k=1}^{R} \left(\mathbf{1}(a = b) \frac{\kappa(\{aa\}^{k+1})}{k!} \mathbf{E}\, \partial_{aa}^k f(W) \right.$$

$$\left. + \mathbf{1}(a \neq b) \sum_{q+q'=k} \binom{k}{q} \frac{\kappa(\{ab\}^{q+1}, \{ba\}^{q'})}{k!} \mathbf{E}\, \partial_{ab}^p \partial_{ba}^{q'} f(W) \right) + \Omega_R, \tag{77}$$

where we used that cumulants are invariant under reordering their entries, and thus $\kappa(ab, \boldsymbol{\alpha})$ can be expressed as the cumulant of $q + 1$ copies $\{ab\}^{q+1}$ of ab and q' copies $\{ba\}^{q'}$ of ba. In order to simplify notations we introduce the matrices $\kappa^{q+1,q'}$ for integers $q, q' \geq 0$ with $q + q' \geq 1$ with matrix elements

$$\kappa_{ab}^{1,1} := 1, \qquad \kappa_{ab}^{2,0} := \sigma,$$

$$\frac{\kappa_{ab}^{q+1,q'}}{N^{(k+1)/2}} := \mathbf{1}(a = b) \frac{\kappa(\{aa\}^{k+1})}{(k+1)k!} + \mathbf{1}(a \neq b) \binom{k}{q} \frac{\kappa(\{ab\}^{q+1}, \{ba\}^{q'})}{k!}, \tag{78}$$

with $k = q + q' \geq 2$, so that (76) can be rewritten as

$$
\begin{aligned}
\mathbf{E}\, w_{ab} f(W) &= \sum_{k=1}^{R} \sum_{q+q'=k} \frac{\kappa_{ab}^{q+1,q'}}{N^{(k+1)/2}} \, \mathbf{E}\, \partial_{ab}^{q} \partial_{ba}^{q'} f(W) + \Omega_R \\
&= \mathbf{E}\, \frac{\partial_{ba} f(W) + \sigma\, \partial_{ab} f(W)}{N} \\
&\quad + \sum_{k=2}^{R} \sum_{q+q'=k} \frac{\kappa_{ab}^{q+1,q'}}{N^{(k+1)/2}} \, \mathbf{E}\, \partial_{ab}^{q} \partial_{ba}^{q'} f(W) + \Omega_R,
\end{aligned}
\tag{79}
$$

where we used that due to (A-i) we have that $\kappa\big(\big\{\sqrt{N}\, w_{aa}\big\}^2\big) = w_2 = 1+\sigma = \kappa_{aa}^{1,1}+\kappa_{aa}^{2,0}$.

We begin with some examples before describing the general structure of the expansion. We consider the case $p = 1$ and $l = 2$ and perform a cumulant expansion

$$
\begin{aligned}
\mathbf{E}\, & \big|\underline{\langle WGA\Im GA\rangle}\big|^2 \\
&= \mathbf{E}\langle\underline{WGA\Im GA}\rangle\langle\underline{A\Im GAG^*W}\rangle \\
&= N^{-1} \sum_{ab} \mathbf{E}\Big(\langle\Delta^{ab}GA\Im GA\rangle \partial_{ba}\langle\underline{A\Im GAG^*W}\rangle \\
&\quad + \sigma\,\langle\Delta^{ab}GA\Im GA\rangle \partial_{ab}\langle\underline{A\Im GAG^*W}\rangle\Big) \\
&\quad + \sum_{k=2}^{R} \sum_{q+q'=k} \frac{\kappa_{ab}^{q+1,q'}}{N^{(k+1)/2}} \, \mathbf{E}\, \partial_{ab}^{q}\partial_{ba}^{q'}\Big[\langle\Delta^{ab}GA\Im GA\rangle\langle\underline{A\Im GAG^*W}\rangle\Big],
\end{aligned}
\tag{80}
$$

where $(\Delta^{ab})_{cd} = \delta_{ac}\delta_{bd}$. In order to compute the derivative of $\Im G$ we write

$$
\partial_{ba}\Im G = \eta\partial_{ba}GG^* = -\eta(G\Delta^{ba}GG^* + GG^*\Delta^{ba}G^*) = -(G\Delta^{ba}\Im G + \Im G\Delta^{ba}G^*).
$$

By distributing the derivatives according to Leibniz' rule we can write (80) as

$$
\begin{aligned}
&\mathbf{E}\sum_{ab} \frac{\kappa_{ab}^{1,1}}{N}\langle\Delta^{ab}GA\Im GA\rangle\langle\underline{A\Im GAG^*\Delta^{ba} - A(G\Delta^{ba}\Im G + \Im G\Delta^{ba}G^*)AG^*W}\rangle \\
&+ \mathbf{E}\sum_{ab} \frac{\kappa_{ab}^{2,0}}{N}\langle\Delta^{ab}GA\Im GA\rangle\langle\underline{A\Im GAG^*\Delta^{ab} - A(G\Delta^{ab}\Im G + \Im G\Delta^{ab}G^*)AG^*W}\rangle \\
&- \mathbf{E}\sum_{ab} \frac{\kappa_{ab}^{2,1}}{N^{3/2}}\langle\Delta^{ab}G\Delta^{ba}GA\Im GA\rangle\langle\underline{A\Im GAG^*\Delta^{ab} - A\Im GAG^*\Delta^{ab}G^*W}\rangle + \cdots
\end{aligned}
\tag{81}
$$

where we chose two representative terms for $k = 1$ and $k = 2$ each. By performing another cumulant expansion for the remaining underlined terms in (81) we obtain

$$
\begin{aligned}
N^2 \, &\mathbf{E} \left| \langle \underline{WGA\Im GA} \rangle \right|^2 \\
&= \mathbf{E} \sum_{ab} \frac{\kappa_{ab}^{1,1}}{N} (GA\Im GA)_{ba} \left[(A\Im GAG^*)_{ab} + \sigma (A\Im GAG^*)_{ba} \right] \\
&\quad - \mathbf{E} \sum_{ab} \frac{\kappa_{ab}^{2,1}}{N^{3/2}} G_{bb} (GA\Im GA)_{aa} (A\Im GAG^*)_{ab} \\
&\quad + \mathbf{E} \sum_{abcd} \frac{\kappa_{ab}^{1,1}}{N} \frac{\kappa_{cd}^{1,1}}{N} G_{bd} (GA\Im GA)_{ca} \left[(A\Im G)_{db} (G^*AG^*)_{ac} + (A\Im G)_{da} (G^*AG^*)_{bc} \right] \\
&\quad - \mathbf{E} \sum_{abcd} \frac{\kappa_{ab}^{2,1}}{N^{3/2}} \frac{\kappa_{ab}^{1,1}}{N} G_{bb} (GA\Im G)_{ad} (G^*A)_{ca} (A\Im GAG^*)_{da} G_{bc}^* \\
&\quad + \mathbf{E} \sum_{abcd} \frac{\kappa_{ab}^{2,1}}{N^{3/2}} \frac{\kappa_{cd}^{2,1}}{N^{3/2}} G_{bb} G_{ac} G_{dd} (GA\Im GA)_{ca} (A\Im GAG^*)_{da} G_{bc}^* + \cdots , \quad (82)
\end{aligned}
$$

where we again selected representative terms. We notice that the rhs. can be written as a polynomial in the entries of two types of matrices; the κ-matrices representing cumulants like $\kappa^{2,1}$, and the G-matrices representing resolvents like $\Im G$ or G^*, or their multiples with A like $A\Im G$, G^*A. In order to achieve this representation we introduce additional internal summation indices to expand longer products e.g. we write $(A\Im GAG^*)_{da} = \sum_e (A\Im G)_{de} (AG^*)_{ea}$. The *value* of any given graph is the numerical result of summing up all indices. The precise definition will be given later in (89); here, as an example, the first term in (82) with indicated summation indices reads

$$
\sum_{ab} \sum_{ij} \frac{1}{N} \, \mathbf{E}(GA)_{bi} (\Im GA)_{ia} (A\Im G)_{aj} (AG^*)_{jb} = \mathbf{E} \operatorname{Val} \left(\begin{array}{c} \end{array} \right),
$$

where the (directed) edges represent matrices and the vertices represent summation indices. The edge orientation indicates the order of indices of the represented matrix which for the G-edges is uniquely determined from the expansion, while for κ-edges it may be chosen arbitrarily, as long as the represented matrix is defined consistently with the orientation, see (88) later. Here we drew the internal vertices as empty, and the κ-*vertices* as filled nodes, the κ-matrices as dashed, and the G-matrices as solid edges. Both internal- and κ-vertices correspond to independent summations over the index set $[N]$.

Thus, graphically we can represent (82) as

$$N^2 \, \mathbf{E} |\langle \underline{WGA\Im GA} \rangle|^2$$

$$= \mathbf{E} \left[\mathrm{Val} \left(\vcenter{\hbox{}} \right) + \mathrm{Val} \left(\vcenter{\hbox{}} \right) - \mathrm{Val} \left(\vcenter{\hbox{}} \right) + \mathrm{Val} \left(\vcenter{\hbox{}} \right) \right.$$

$$\left. - \mathrm{Val} \left(\vcenter{\hbox{}} \right) - \mathrm{Val} \left(\vcenter{\hbox{}} \right) + \mathrm{Val} \left(\vcenter{\hbox{}} \right) + \cdots \right].$$

$$(83)$$

Note that the dashed edges connect only filled nodes and they form a perfect matching. The number of G-edges adjacent to each filled vertex is equal to the order of the corresponding cumulant expansion.

Similarly, for the isotropic case we obtain, for example the polynomials

$$\mathbf{E} \left| \langle x, \underline{GAWGy} \rangle \right|^2 = \mathbf{E} \langle x, \underline{GAWGy} \rangle \langle y, \underline{G^*WAG^*x} \rangle$$

$$= \mathbf{E} \sum_{ab} \frac{\kappa_{ab}^{1,1}}{N} (GA)_{xa} G_{by} G_{yb}^* (AG^*)_{ax}$$

$$+ \mathbf{E} \sum_{abcd} \frac{\kappa_{ab}^{1,1}}{N} \frac{\kappa_{cd}^{1,1}}{N} (GA)_{xa} G_{bd} G_{cy} \left[G_{yb}^* G_{ac}^* (AG^*)_{dx} + G_{yc}^* (AG^*)_{db} G_{ax}^* \right]$$

$$- \mathbf{E} \sum_{abcd} \frac{\kappa_{ab}^{2,1}}{N^{3/2}} \frac{\kappa_{cd}^{1,1}}{N} (GA)_{xa} G_{bb} G_{ad} G_{cy} G_{ya}^* G_{bc}^* (AG^*)_{dx} + \cdots \qquad (84)$$

which we represent graphically as

$$\mathbf{E} |\langle x, \underline{GAWGy} \rangle|^2 = \mathbf{E} \, \mathrm{Val} \left(\vcenter{\hbox{}} \right) + \mathbf{E} \, \mathrm{Val} \left(\vcenter{\hbox{}} \right)$$

$$+ \mathbf{E} \, \mathrm{Val} \left(\vcenter{\hbox{}} \right) - \mathbf{E} \, \mathrm{Val} \left(\vcenter{\hbox{}} \right) + \cdots,$$

$$(85)$$

with external vertices drawn as squares. Note that the vectors x and y are naturally represented by *external vertices* that are drawn as solid squares.

After these examples we now explain the general structure of the graphs and give a precise definition of graphs and their *graph value* used in (83) and (85).

Definition 2. We define the class $\mathcal{G}$ of oriented graphs used within this paper by the following requirements. Each $\Gamma = (V, E) \in \mathcal{G}$ has three types of vertices, κ-vertices V_κ, *internal* vertices V_i and external vertices V_e, so that $V = V_\kappa \dot\cup V_i \dot\cup V_e$, and two types of edges, κ-*edges* E_κ and G-*edges* E_g, so that $E = E_\kappa \dot\cup E_g$. For each vertex $v \in V$ we define its G-in- and out-degree $d_g^{in}(v), d_g^{out}(v)$ as the number of incoming and outgoing G-edges. The total degree $d_g(v)$ is defined as the sum $d_g(v) := d_g^{in}(v) + d_g^{out}(v)$ of in- and out-degrees and the three vertex classes satisfy

$$d_g(v) = \begin{cases} 1, & v \in V_e \\ 2, & v \in V_i, \end{cases}, \qquad d_g(v) \geq 2, \quad v \in V_\kappa. \tag{86}$$

We can partition $V_\kappa = \bigcup_{k \geq 2} V_\kappa^k$ with $V_\kappa^k := \left\{ v \in V_\kappa \,\middle|\, d_g(v) = k \right\}$.

Within the graphs Γ each external vertex $v \in V_e$ carries some $\boldsymbol{x}(v) \in \mathbf{C}^N$ as a vector-valued label recording which vector the vertex represents. Each κ-edge $e \in E_\kappa$ carries two integer-valued labels $r(e) \geq 1$, $s(e) \geq 0$ recording the cumulant type. Each G-edge $e \in E_g$ carries six labels. The binary labels $i(e), t(e), *(e) \in \{0, 1\}$ indicate whether e represents the imaginary part, the transpose and/or the adjoint of a resolvent. The scalar label $z(e)$ records the spectral parameter of the resolvent and the matrix-valued labels $L(e), R(e)$ record deterministic matrices which are multiplied with the resolvent from the left/right.

We now relate the graphs to the polynomials they represent. Each internal vertex or κ-vertex v corresponds to an independent summation $a_v \in [N]$. In order to unify notations we define a *labelling map*

$$\boldsymbol{x}: V \to \mathbf{C}^N, \qquad v \mapsto \boldsymbol{x}_v := \begin{cases} \boldsymbol{x}(e), & v \in V_e, \\ \boldsymbol{e}_{a_v}, & v \in V_i \cup V_\kappa, \end{cases} \tag{87}$$

where $\boldsymbol{e}_a$ is the a-th unit vector in the standard basis, and for $v \in V_e$, the vector $\boldsymbol{x}(v)$ is the label of v from Definition 2. The G-edges $e \in E_g$ represent resolvents defined via the labels of e from Definition 2. We define the matrix $\mathcal{G}^e$ as the resolvent $G(z(e))$ modified according to $i(e), t(e), *(e)$ and multiplied by $L(e), R(e)$ from the left/right. As an example, we set

$$\mathcal{G}^e = B(\Im G(z))^t \quad \text{for } e \in E_g$$

with

$$\Big(i(e), t(e), *(e), z(e), L(e), R(e)\Big) = (1, 1, 0, z, B, I).$$

We remark that for all G-edges e considered in this paper at most one of the matrices $L(e), R(e)$ is different from the identity matrix I. The κ-edges $e \in E_\kappa$ represent $N \times N$ cumulant matrices κ^e which are determined by the two integers $r(e), s(e)$ from Definition 2 such that for $a \neq b$,

$$\kappa_{ab}^{(uv)} := \kappa_{ab}^{r((uv)),s((uv))}, \tag{88}$$

where on the rhs. κ was defined in (78). We note that $\left|\kappa_{ab}^{(uv)}\right| \lesssim 1$ by (8). Finally, we define the graph value

$$\text{Val}(\Gamma) := \sum_{\substack{a_v \in [N] \\ v \in V_i \cup V_\kappa}} \left[\prod_{(uv) \in E_\kappa} \left(N^{-d_g(u)/2} \kappa_{a_u a_v}^{(uv)} \right) \right] \left(\prod_{(uv) \in E_g} \mathcal{G}_{\boldsymbol{x}_u \boldsymbol{x}_v}^{(uv)} \right). \tag{89}$$

Among the degree-2 vertices the ones between edges representing matrices whose eigenvectors are asymptotically orthogonal are of particular importance. There are two different mechanism for such orthogonality; (a) two resolvents, one with and one without transpose stand next to each other, e.g. GG^t or $G^*(A(\Im G)^t)$, (b) a traceless matrix A stands between two resolvents, e.g. $(GA)G^*$ or $G(A(\Im G)^t)$. Note that in some cases, e.g. $(GA)(\Im G)^t$, both mechanism can be present simultaneously, and hence a vertex can be 0tr- and t-vertex at the same time.

Definition 3 (*Orthogonality vertices*).

(a) A vertex $v \in V_\kappa^2 \cup V_i$ is called a *t-orthogonality vertex*, or short *t-vertex* if the two unique G-edges $e_1, e_2 \in E_g$ adjacent to v satisfy $(t(e_1), t(e_2)) \in \{(0, 1), (1, 0)\}$, i.e. if exactly one of the two G-edges adjacent to v is transposed.

(b) A vertex $v \in V_i \cup V_\kappa^2$ is called an *zero-trace-orthogonality vertex*, or short *0tr-vertex* if exactly one of the two edges adjacent to v represents a resolvent (which is allowed to be the imaginary part, transposed, or adjoint) multiplied by a traceless matrix on the side of v, while the other adjacent edge represents a resolvent matrix multiplied by the identity matrix on the side of v. More precisely, using the labels $L(e)$, $R(e)$ of the edges, v is defined to be an 0tr-vertex if one of the following three conditions is satisfied:

(b.i) there are incoming/outgoing edges $(uv), (vw) \in E_g$ such that either $\langle R((uv)) \rangle = 0, L((vw)) = I$ or $\langle L((vw)) \rangle = 0, R((uv)) = I$,

(b.ii) there are two outgoing edges $(vu), (vw) \in E_g$ such that either $\langle L((vu)) \rangle = 0, L((vw)) = I$ or $\langle L((vw)) \rangle = 0, L((vu)) = I$,

(b.iii) there are two incoming edges $(uv), (wv) \in E_g$ such that either $\langle R((uv)) \rangle = 0, R((wv)) = I$ or $\langle R((wv)) \rangle = 0, R((uv)) = I$.

Proposition 3 (Cumulant expansion). *Let $\mathfrak{a}, \mathfrak{t}, \mathfrak{i}$ be fixed sets as in Theorem 5 of sizes $a := |\mathfrak{a}|, t := |\mathfrak{t}|, i := |\mathfrak{i}|$. Then for any $p \in \mathbf{N}$ there exists a finite (N-independent) family of graphs $\mathcal{G}_p = \mathcal{G}_p^{\text{av}} \cup \mathcal{G}_p^{\text{iso}} \subset \mathcal{G}$ such that*

$$\mathbf{E} \left| \text{Tr } \underline{WG_1 B_1 G_2 B_2 \cdots G_l B_l} \right|^{2p} = \sum_{\Gamma \in \mathcal{G}_p^{\text{av}}} \mathbf{E} \, \text{Val}(\Gamma) + \mathcal{O}\left(N^{-2p}\right),$$

$$\tag{90}$$

$$\mathbf{E} \left| \langle \boldsymbol{x}, \underline{G_1 B_1 \cdots G_j B_j W G_{j+1} B_{j+1} \cdots B_{l-1} G_l \boldsymbol{y}} \rangle \right|^{2p} = \sum_{\Gamma \in \mathcal{G}_p^{\text{iso}}} \mathbf{E} \, \text{Val}(\Gamma) + \mathcal{O}\left(N^{-2p}\right),$$

$$\tag{91}$$

and for each graph Γ we may select two disjoint subsets $V_o^t \dot\cup V_o^{\text{0tr}} =: V_o$ of t- and 0tr-vertices, respectively, such that the following properties are satisfied:

(P1) The graph (V_κ, E_κ) is a perfect matching, in particular, $|V_\kappa| = 2|E_\kappa|$.
(P2) The number of κ-edges satisfies $1 \le |E_\kappa| \le 2p$.

(P3) The number of G-edges satisfies

$$\left|\left\{e \in E_g \,\middle|\, i(e) = 1\right\}\right| = 2ip \tag{92a}$$

$$\left|E_g\right| = \sum_{e \in E_\kappa} d_g(e) + 2(l-1)p \geq 2p. \tag{92b}$$

(P4) For $(uv) \in E_\kappa$ the G-degrees of $u, v \in V_\kappa$ satisfy $d_g^{\text{in}}(u) = d_g^{\text{out}}(v)$, $d_g^{\text{in}}(v) = d_g^{\text{out}}(u)$ and $d_g(u) = d_g(v) \geq 2$. Therefore we may define the G-degree of (uv) as $d_g((uv)) := d_g(u) = d_g(v)$ and partition $E_\kappa = \dot{\bigcup}_{k \geq 2} E_\kappa^k$ into $E_\kappa^k := \left\{e \in E_\kappa \,\middle|\, d_g(e) = k\right\}$.

(P5) Every E_g-cycle on $V_\kappa^2 \cup V_i$ must contain at least two V_κ^2-vertices, and in particular there cannot exist isolated loop edges, and there are at most $\left|E_\kappa^2\right|$ cycles.

(P6) Denoting the number of isolated cycles in $(V_\kappa \cup V_i, E_g)$ with k vertices in V_o by $n_{\text{cyc}}^{o=k}$, we have

$$2 n_{\text{cyc}}^{o=0} + n_{\text{cyc}}^{o=1} \leq 2\left|E_\kappa^2\right| - \left|V_o \cap V_\kappa^2\right|.$$

(P7) The numbers of selected internal 0tr- and t-vertices are

$$\left|V_i \cap V_o^{0\text{tr}}\right| = \begin{cases} 2p(a-1), & j \in \mathfrak{a} \\ 2pa, & j \notin \mathfrak{a}, \end{cases}, \quad \left|V_i \cap V_o^t\right| = \begin{cases} 2p(t-1), & j \in \mathfrak{t} \\ 2pt, & j \notin \mathfrak{t}, \end{cases}$$

where in the averaged case $j := l$ and j is determined by the lhs. of (91) in the isotropic case.

(P8) If $j \in \mathfrak{a}$ (with again $j := l$ in the averaged case), then the set of selected 0tr-vertices $V_o^{0\text{tr}}$ satisfies

$$2\left|E_\kappa^2\right| + \left|E_\kappa^{\geq 3}\right| - 2p \leq \left|V_o^{0\text{tr}} \cap V_\kappa^2\right| \leq 2p,$$

while otherwise $V_o^{0\text{tr}} \cap V_\kappa^2 = \emptyset$. Similarly, if $j \in \mathfrak{t}$, then the set of selected t-vertices V_o^t satisfies

$$2\left|E_\kappa^2\right| + \left|E_\kappa^{\geq 3}\right| - 2p \leq \left|V_o^t \cap V_\kappa^2\right| \leq 2p,$$

while otherwise $V_o^t \cap V_\kappa^2 = \emptyset$.

The graphs $\Gamma \in \mathcal{G}_p^{\text{av}}$ satisfy (P1)-(P8) and in addition:

($P^{\text{av}}9$) There are no external vertices, i.e. $V_e = \emptyset$

($P^{\text{av}}10$) The number of internal vertices satisfies $|V_i| = 2(l-1)p$. The graphs $\Gamma \in \mathcal{G}_p^{\text{iso}}$ satisfy (P1)-(P8) and in addition:

($P^{\text{iso}}9$) The number of external vertices is $|V_e| = 4p$ each $v \in V_e$ has degree $d_g(v) = 1$ and the unique connected vertex $u \in V$ with $(uv) \in E_g$ or $(vu) \in E_g$ satisfies $u \in V_\kappa$.

($P^{\text{iso}}10$) The number of internal vertices satisfies $|V_i| = 2p(l-2)$.

Definition 4. For some parameters $a, t, l, i, p \in \mathbf{N}$ we call graphs $\Gamma \in \mathcal{G}$ together with their selected V_{o}^t, $V_{\mathrm{o}}^{0\mathrm{tr}}$ sets satisfying (P1)-(P8) and $(\mathrm{P}^{\mathrm{av}}9)$-$(\mathrm{P}^{\mathrm{av}}10)$ *av-graphs*, while we call graphs $\Gamma \in \mathcal{G}$ (together with the sets V_{o}^t, $V_{\mathrm{o}}^{0\mathrm{tr}}$ and the extra parameter $j \in [l-1]$) satisfying (P1)-(P8) and $(\mathrm{P}^{\mathrm{iso}}9)$-$(\mathrm{P}^{\mathrm{iso}}10)$ *iso-graphs*.

Proof. (Proposition 3) In order to obtain (90) we iteratively perform cumulant expansions exactly as in the examples (80) and (84) until no underlined terms remain. Each cumulant expansion removes at least one underlined term, hence this process terminates.

We now explain which kinds of G-edges are created through this cumulant expansion procedure for the averaged case (90), the isotropic case (91) being very similar. Initially, the graph representing the lhs. of (90) after writing out $|\mathrm{Tr}\, X|^{2p} = (\mathrm{Tr}\, X)^p (\mathrm{Tr}\, X^*)^p$ consists of $2p$ cycles each with a W factor and l G-edges representing $\mathcal{G}$-*factors* $G_k B_k$ or $B_k^* G_k^*$ for $k \in [l]$. Each of these $\mathcal{G}$-factors can be fully described via the labels $i(e), t(e), *(e), z(e), L(e), R(e)$ from Definition 2, the first four being determined by the form of G_k while the latter two encode the multiplication from the left/right by deterministic matrices, e.g. $L(e) = I$, $R(e) = B_k$ for $G_k B_k$. While performing cumulant expansions of some $W = \sum_{ab} w_{ab} \Delta^{ab}$ using (76) these G-edges are modified and new G edges are created via the action of derivatives, and κ-edges representing $\kappa(ab, \boldsymbol{\alpha})$ are also created. This process creates creates (finitely) many different graphs for every cumulant expansion, both through the explicit summation over cumulants in (76) and the Leibniz rule for the derivative $\partial_{\boldsymbol{\alpha}}$ acting on the product of all remaining W's and G's. We note that for resolvent derivatives we have

$$\partial_{ab} G = -G \Delta^{ab} G, \quad \partial_{ab} G^* = -G^* \Delta^{ab} G^*,$$
$$\partial_{ab} G^t = -G^t \Delta^{ba} G^t, \quad \partial_{ab} (G^*)^t = -(G^*)^t \Delta^{ba} (G^*)^t$$

and

$$\partial_{ab} \Im G = -G \Delta^{ab} \Im G - \Im G \Delta^{ab} G^*,$$
$$\partial_{ab} (\Im G)^t = -(\Im G)^t \Delta^{ba} G^t - (G^*)^t \Delta^{ba} (\Im G)^t.$$

Hence, a derivative action on e representing the $\mathcal{G}$-factor $\mathcal{G}^e = G_k B_k$ (or similarly $B_k^* G_k^*$) creates two G-edges e_1, e_2, such that only the resolvent representing e_2 is multiplied from the right by $R(e_2) = B_k$ while $L(e_2) = L(e_1) = R(e_1) = I$. The labels $t(e), z(e)$ indicating the transposition status and spectral parameter are directly inherited to both e_1, e_2, while the label $i(e)$ is inherited to exactly one of e_1, e_2, $i(e_1) = 1$ the other one satisfying $i(e_2) = 0$, $*(e_2) \in \{0, 1\}$. If $*(e) = 1, i(e) = 0$, then both e_1, e_2 satisfy $*(e_1) = *(e_2) = 1$. It follows inductively that each $\mathcal{G}$-factor encountered in the expansion can be represented by an edge e with six labels $i(e), t(e), *(e), z(e), L(e), R(e)$, with $L(e) = I$, or $L(e) = B_k^*$ for some k, while $R(e) = I$ or $R(e) = B_k$ for some k, with for each e at least one of $L(e), R(e)$ being the identity. The spectral parameter label $z(e)$ satisfies $z(e) \in \{z_1, \ldots, z_l\}$ for each e. For example, the ab-derivative of the $\mathcal{G}$-factor $\mathcal{G}^e = B(\Im G(z))^t$ described by the edge e with labels $(1, 1, 0, z, B, I)$ yields a sum of two terms and hence the two new graphs given by
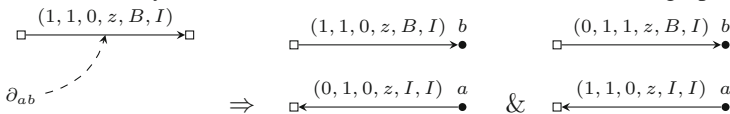
or, in formulas,

$$\partial_{ab}\left[B(\Im G(z))^t \right] = -B(\Im G(z))^t \Delta^{ba} G^t - B(G^*)^t \Delta^{ba} (\Im G(z))^t.$$

We now describe the selection of the orthogonality vertices V_o^t, V_o^{0tr} which is done in two steps. To unify notations we set $j := l$ in the averaged case.

(orth-1) For each $k \in \mathfrak{t} \setminus \{j\}$, $\mathfrak{a} \setminus \{j\}$ we collect $2p$ distinct vertices from V_i into the sets V_o^t and V_o^{0tr}, respectively.

(orth-2) If $j \in \mathfrak{t}$ or $j \in \mathfrak{a}$, then we select one vertex from V_κ^2 into V_o^t or V_o^{0tr}, respectively, for each W acting as a degree-2 cumulant on some resolvent.

Regarding (orth-1) for $k \in \mathfrak{a} \cup \mathfrak{t} \setminus \{j\}$ the initial graphs representing the lhs. of (90)–(91) contain p internal vertices $v_1, \ldots v_p$ between G-edges representing $(G_k B_k)$, $(G_{k+1} B_{k+1})$ and p internal vertices $v_{p+1}, \ldots v_{2p}$ between G-edges representing $(B_{k+1}^* G_{k+1}^*)$, $(B_k^* G_k^*)$. The G-edges adjacent to these internal vertices may change due to derivative actions along the cumulant expansions, however in case $k \in \mathfrak{a}$, due to the derivative rules explained in the paragraph above it is ensured that all times the two unique G-edges e_1, e_2 adjacent to v_k satisfy $R(e_1) = B_k, L(e_2) = I$ for $k \le p$ and $R(e_1) = I, L(e_2) = B_k^*$, so that v_k is guaranteed to remain an 0tr-vertex. Similarly, for $k \in \mathfrak{t}$ it is ensured that the two unique G-edges e_1, e_2 adjacent to v_k satisfy $t(e_1) = 1, t(e_2) = 0$, so that v_k is guaranteed to remain an t-vertex.

Regarding (orth-2) we note that while performing the cumulant expansion for $W = \sum_{ab} w_{ab} \Delta^{ab}$ in $G_j B_j W G_{j+1}$ we obtain the degree-2 cumulant term as

$$\sum_{ab} G_j B_j \Delta^{ab} G_{j+1} (\partial_{ba} + \sigma \partial_{ab})$$

the derivatives ∂_{ab} or ∂_{ba} acting on some resolvent G result in $G\Delta^{ab}G$ or $G\Delta^{ba}G$. In case $j \in \mathfrak{a}$ the κ-vertex corresponding to the summation index a satisfies the definition of 0tr-vertex since $\langle B_j \rangle = 0$ and the other resolvent is not multiplied by some additional matrix in the a-direction. Similarly, in case $j \in \mathfrak{t}$ either both or none of the two G's in $G\Delta^{ab}G$ or $G\Delta^{ba}G$ are transposed, while, by definition, exactly one of G_j, G_{j+1} is transposed. Thus exactly one of the κ-vertices corresponding to the a or b-summations satisfies the definition of being a t-vertex.

We note that the condition $\mathfrak{a} \cap \mathfrak{t} = \emptyset$ ensures the sets V_o^t, V_o^{0tr} constructed in this way to be disjoint. We now check that the properties (P1)-(P8), as well as (P^{av}9)-(P^{av}10) and (P^{iso}9)-(P^{iso}10) also hold for these graphs.

The properties (P1)-(P2) are obvious by construction since each cumulant expansion comes with two κ-vertices, and in total there are $2p$ underlined terms and thereby at most $2p$ cumulant expansions. The properties (P^{av}10)-(P^{iso}10) follow from the fact that for each factor of Tr $\underline{W G_1 B_1 \cdots G_l B_l}$ and

$$\langle x, \underline{G_1 B_1 \cdots G_j B_j W G_{j+1} B_{j+1} \cdots B_{l-1} G_l} y \rangle$$

there are $l-1$ and respectively $l-2$ internal vertices of in- and out-degree 1 and that these properties remain invariant under cumulant expansions. Similarly, the properties (P^{av}9) and (P^{iso}9) hold true trivially for the initial terms and remain invariant under cumulant expansions.

For (P4) note that the cumulant $\kappa(ab, (\alpha_1, \ldots, \alpha_k))$ comes together with matrices

$$\Delta^{ab}, (\Delta^{\alpha_1})^{(t)}, \ldots, (\Delta^{\alpha_k})^{(t)}$$

after derivative action, where the transpose is taken in case the derivative acts on a transposed resolvent. In all cases the in-degree of the vertex associated with a is equal to the out-degree of the vertex associated with b.

For (P5) note that by the definition of the underline-renormalisation it follows that for degree two edges the corresponding ∂_{ba} derivative cannot act on its own trace and therefore cycles have to involve at least two V_κ^2 vertices.

For (P6) we note that $\left| V_\kappa^2 \setminus V_o \right| = 2 \left| E_\kappa^2 \right| - \left| V_o \cap V_\kappa^2 \right|$, while due to (P5) each cycle with zero V_o-vertices contains at least two $V_\kappa^2 \setminus V_o$-vertices and each cycle with one V_o-vertex contains at least one $V_\kappa^2 \setminus V_o$-vertex.

The claim (P7) follows immediately from the construction (orth-1). Similarly, claim (P8) follows from the construction (orth-2) together with the observation that because $|E_\kappa|$ is the total number of cumulant expansions, a total of $2p - |E_\kappa|$ derivatives have acted on some W, and thus the number n of W's acting on as degree-2 cumulants on some G satisfies

$$n \geq \left| E_\kappa^2 \right| - (2p - |E_\kappa|) = 2 \left| E_\kappa^2 \right| + \left| E_\kappa^{\geq 3} \right| - 2p, \tag{93}$$

and, trivially, $n \leq 2p$. This concludes the proof of (P8) in the mutually exclusive cases $j \in \mathfrak{a}$ and $j \in \mathfrak{t}$ (recall that $\mathfrak{a} \cap \mathfrak{t} = \emptyset$ by assumption).

For the claim (92a) on the number of G-edges in (P3) note that the number of $\Im G$'s remains invariant under the derivative actions. For (92b) note that each derivative acting on some G increases the number of G's by one, while each of the $2p - |E_\kappa|$ derivatives acting on some W leaves the number of G's invariant. Thus we conclude that the total number of G's is

$$2lp + \sum_{e \in E_\kappa} d_g(e) - |E_\kappa| - (2p - |E_\kappa|) = \sum_{e \in E_\kappa} d_g(e) + 2(l-1)p$$

and (92b) follows.   □

Remark 7. Proposition 3 holds true verbatim also under the alternative definition of the renormalisation outlined in Remark 5 in case no G is transposed. Also the proof of the proposition remains unchanged except for the proof of Property (P5). For the alternative renormalisation also for degree two edges when expanding $\underline{WG \cdots} = \sum_{ab} \Delta^{ab} G \cdots \partial_{ba}$ the derivative $\sigma \partial_{ab}$ may act on its own trace. However, since no G is transposed this action will necessarily result in $\Delta^{ab} G \cdots G \Delta^{ab}$ and therefore no loops are created.

Using Proposition 3, in order to conclude Theorem 5, it remains to estimate $\mathrm{Val}(\Gamma)$ for each $\Gamma \in \mathcal{G}_p$ as follows. We note that the following Proposition is valid for any av-/iso-graphs $\Gamma \in \mathcal{G}$ from Definition 4, i.e. graphs satisfying the properties (P1)-(P8) and $(\mathrm{P}^{\mathrm{av}}9)$-$(\mathrm{P}^{\mathrm{av}}10)$/$(\mathrm{P}^{\mathrm{iso}}9)$-$(\mathrm{P}^{\mathrm{iso}}10)$ above rather than only for the specific families of graphs $\mathcal{G}_p^{\mathrm{av}}, \mathcal{G}_p^{\mathrm{iso}}$ arising in the cumulant expansion.

Proposition 4. (Value estimate) *For each av-graph $\Gamma \in \mathcal{G}$ for some parameters $a, t, l, p, i \in \mathbf{N}$ we have the bound*

$$|\mathrm{Val}(\Gamma)| \prec \begin{cases} \rho^{2(b+1)p} N^{2bp} K^{-2bp}, & b = l \\ \Lambda_+^{2ap} \Pi_+^{2tp} \rho^{2ip \vee 2(b+1)p} N^{p(a+t+2b)} K^{-p(1+2b)}, & b < l, \end{cases}, \quad b := l - a - t \tag{94}$$

with K as in (75), while for each iso-graph Γ for some parameters $a, t, l, p, i \in \mathbf{N}$ we have the bound

$$|\mathrm{Val}(\Gamma)| \prec \Lambda_+^{2ap} \Pi_+^{2tp} \rho^{2ip \vee 2(b+1)p} N^{p(a+t+2b)} K^{-p(1+2b)}, \quad b := l - a - t - 1. \tag{95}$$

Proof. (Theorem 5) Theorem 5 follows immediately by combining Propositions 3 and 4 under the simplifying assumptions made at the beginning of Sect. 5. Their removal is a routine technicality whose details we present in Appendix A of the arXiv:2012.13215 version of the present paper. Following the proof Proposition 4 it is evident that both Λ_+^{2ap} and Π_+^{2tp} can be replaced by the product of individual $\Lambda_+^{B_k}, \Pi_+^{B_k}$ for $k \in \mathfrak{a} \cup \mathfrak{t}$, as claimed in Theorem 5.

Finally, regarding the replacement of ρ^i by $\prod_{k \in \mathfrak{i}} \rho(z_k)$ in the bounds of Theorem 5, it is easy to see that during the cumulant expansion the number of $\Im G(z_k)$ is preserved and each gives rise to a factor $\rho(z_k)$ in Proposition 4, hence the factor ρ^{2ip} may be replaced by the factor $\prod_{k \in \mathfrak{i}} \rho(z_k)^{2p}$. Similarly, for the replacement of Λ_+^a by $\prod_{k \in \mathfrak{a}} \Lambda_+^{B_k}$ we note that each B_k appears exactly $2p$ times also after the cumulant expansions, and therefore each $\Lambda_+^{B_k}$ can only appear in at most the $2p$-th power on the rhs. of (94)–(95). $\qquad\square$

5.2. Estimating graph values: Proof of Proposition 4.

The proof of Proposition 4 goes in three major steps formulated in Lemmata 2, 3 and 4 which we first state and then use to conclude the proof of Proposition 4.

First, we express the value $\text{Val}(\Gamma) = \text{Val}(\Gamma_{\text{red}})$ as the value of the *reduced graph* Γ_{red} obtained from Γ by collapsing all degree-2 vertices $V_{\mathfrak{i}} \cup V_\kappa^2$. Thus, in graph-theoretic terms, Γ_{red} is the minimal (with the least number of edges) graph having $\Gamma \setminus E_\kappa^2$ as a *subdivision*. We claim that each summation index a_v for $v \in V_{\mathfrak{i}} \cup V_\kappa^2$ appears in exactly two $\mathcal{G}$-factors and no κ-matrices, and thus the summation can be written as a matrix product after (potentially) transposing one of the two $\mathcal{G}$'s in the cases of two incoming or two outgoing edges, e.g. $\sum_{a_v} (GB)_{xa_v} G_{ya_v} = (GBG^t)_{xy}$. Indeed, the index a_v appears only in exactly two G-edges since $d_g(v) = 2$, cf. Definition 2. Moreover, due to (P1) no κ-edge is adjacent to $V_{\mathfrak{i}}$ while for $v \in V_\kappa^2$ the corresponding κ-edge (uv) or (vu) due to (78) and (88) is given by $\kappa^{1,1}$ or $\kappa^{2,0}$ which are constant-1, and constant σ-matrices, and thus effectively the index a_v does not appear in any $\kappa^{(vu)/(uv)}$ matrix.

In the reduction process the value of Γ effectively reduces to a summation over vertices of degree at least 3, traces of G-cycles and entries of G-chains and $E_\kappa^{\geq 3}$-matrices, represented by Γ_{red}. Here we use the terminology that a *G-cycle* is a cycle of G-edges on $V_\kappa^2 \cup V_{\mathfrak{i}}$ vertices, irrespective of the edge orientation, and that a *G-chain* is a chain of G-edges with internal $V_\kappa^2 \cup V_{\mathfrak{i}}$-vertices and external $V_\kappa^{\geq 3} \cup V_e$-vertices, again irrespective of the edge orientation. Note that the reduction completely collapses each E_g-cycles on $V_{\mathfrak{i}} \cup V_\kappa^2$-vertices into a single vertex with a loop edge. The sets of these single vertices and loop edges are denoted by V_{cyc} and $E_g^{\text{red,cyc}}$. Therefore the edge set of reduced graph Γ_{red} is naturally partitioned into $V(\Gamma_{\text{red}}) := V_\kappa^{\geq 3} \dot\cup V_e \dot\cup V_{\text{cyc}}$ and its edge set is $E(\Gamma_{\text{red}}) := E_g^{\text{red}} \dot\cup E_\kappa^{\geq 3}$.

The graph reduction by partial resummations corresponds to generalising the definition of value to

$$\text{Val}(\Gamma_{\text{red}}) := N^{-|E_\kappa^2|+|V_{\text{cyc}}|} \sum_{\substack{a_v \in [N] \\ v \in V_\kappa^{\geq 3}}} \left[\prod_{(uv) \in E_\kappa^{\geq 3}} \left(N^{-d_g(u)/2} \kappa_{a_u a_v}^{(uv)} \right) \right]$$

$$\times \left(\prod_{v \in V_{\text{cyc}}} \langle \mathcal{G}^{(vv)} \rangle \right) \left(\prod_{(uv) \in E_g^{\text{red}} \setminus E_g^{\text{red,cyc}}} \mathcal{G}_{x_u x_v}^{(uv)} \right). \tag{96}$$

where we defined

$$\mathcal{G}^{(v_1 v_k)} := (\mathcal{G}^{(v_1 v_2)})^{(t)} \cdots (\mathcal{G}^{(v_{k-1} v_k)})^{(t)} \tag{97}$$

as a matrix product of (possibly transposed, depending on the in- and out-degrees) of $\mathcal{G}^{(v_1 v_2)}, \ldots, \mathcal{G}^{(v_{k-1} v_k)}$, whenever $d_g(v_2) = \cdots = d_g(v_{k-1}) = 2$. For each edge $e \in E_g^{\mathrm{red}}$ we record the number of $\Im G$'s, the total number of G-edges and the number of summed up $V_o^{t/\mathrm{0tr}}$-vertices in the corresponding chains and cycles by $i(e), l(e), t(e), a(e)$, respectively and set $o(e) := a(e) + t(e)$. The letter o refers to the counting of vertices with the asymptotic orthogonality effect. Note that for cycles all $V_o^{t/\mathrm{0tr}}$-vertices in the cycle contribute towards $t(e), a(e)$ while for chains the first and last vertex necessarily are in $V_\kappa^{\geq 3} \cup V_e$ and hence, by definition cannot be $V_o^{t/\mathrm{0tr}}$-vertices. Thus the parameters a, t, i, l satisfy the relations

$$1 \leq l(e), \qquad 0 \leq i(e) \leq l(e),$$

$$0 \leq a(e) + t(e) = o(e) \leq \begin{cases} l(e), & e \in E_g^{\mathrm{red,cyc}}, \\ l(e) - 1, & e \in E_g^{\mathrm{red}} \setminus E_g^{\mathrm{red,cyc}}. \end{cases} \tag{98}$$

We denote the set of $v \in V_{\mathrm{cyc}}$ with $o((vv)) = k$ by $V_{\mathrm{cyc}}^{o=k}$ which are of cardinality $\left| V_{\mathrm{cyc}}^{o=k} \right| = n_{\mathrm{cyc}}^{o=k}$, c.f. (P6).

Lemma 2. *For each av-/iso-graph $\Gamma \in \mathcal{G}$ with parameters a, t, l, i, p and the selected vertex sets $V_o^{\mathrm{0tr}}, V_o^t$, let $\Gamma_{\mathrm{red}} = (V_\kappa^{\geq 3} \cup V_e \cup V_{\mathrm{cyc}}, E_g^{\mathrm{red}} \cup E_\kappa^{\geq 3})$ denote its reduction. The reduced graph then satisfies*

$$\left| E_g^{\mathrm{red}} \right| = |E_g| - |V_i| - \left| V_\kappa^2 \right| + |V_{\mathrm{cyc}}| = |E_g| - |V_i| - 2 \left| E_\kappa^2 \right| + |V_{\mathrm{cyc}}| \tag{99}$$

and

$$\mathrm{Val}(\Gamma_{\mathrm{red}}) = \mathrm{Val}(\Gamma).$$

Moreover, we have

$$\sum_{e \in E_g^{\mathrm{red}}} t(e) = \left| V_o^t \right|, \quad \sum_{e \in E_g^{\mathrm{red}}} a(e) = \left| V_o^{\mathrm{0tr}} \right|, \quad \sum_{e \in E_g^{\mathrm{red}}} l(e) = \left| E_g \right|, \quad \sum_{e \in E_g^{\mathrm{red}}} i(e) = 2ip. \tag{100}$$

Second, we estimate the value of each graph by bounding the size of each of the reduced G-edges entrywise and the summations trivially.

Lemma 3. *For each av-/iso-graph $\Gamma \in \mathcal{G}$ with the selected vertex sets $V_o^{\mathrm{0tr}}, V_o^t$ we have $|\mathrm{Val}(\Gamma_{\mathrm{red}})| \prec \mathrm{I}_2\text{-Est}(\Gamma)$ with*

$$\mathrm{I}_2\text{-Est}(\Gamma) := \Lambda_+^{\left| V_o^{\mathrm{0tr}} \right|} \Pi_+^{\left| V_o^t \right|} \rho^{2ip \vee (|V_i| + 2|E_\kappa^2| - |V_o|)} N^{|V_i| + \left| E_\kappa^2 \right| + \left| E_\kappa^3 \right|/2 - |V_o|/2 - \delta^{\geq 4}}$$
$$\times K^{|V_o| - |V_i| - 2 \left| E_\kappa^2 \right| + \left| V_{\mathrm{cyc}}^{o=0} \right| + \left| V_{\mathrm{cyc}}^{o=1} \right|/2}, \tag{101}$$

where

$$\delta^{\geq 4} := \sum_{e \in E_\kappa} \left(\frac{d_g(e)}{2} - 2 \right)_+$$

Finally, in the third step we improve upon the entrywise estimate as by estimating summations corresponding to some $V_\kappa^{\geq 3}$-vertices more effectively, using a Schwarz inequality followed by the Ward identity $GG^* = \Im G/\eta$.

Lemma 4. *For each av-graph $\Gamma \in \mathcal{G}$ with the selected vertex sets $V_{\mathrm{o}}^{\mathrm{0tr}}$, $V_{\mathrm{o}}^{\mathrm{0tr}}$ we have* $|\mathrm{Val}(\Gamma_{\mathrm{red}})| \prec \mathrm{I}_3\text{-}\mathrm{Est}(\Gamma)$ *with*

$$\mathrm{I}_3\text{-}\mathrm{Est}(\Gamma) := \Lambda_+^{\left|V_{\mathrm{o}}^{\mathrm{0tr}}\right|}\Pi_+^{\left|V_{\mathrm{o}}^t\right|}\rho^{2ip\vee(|V_{\mathrm{i}}|+2|E_\kappa^2|+|E_\kappa^3|-|V_{\mathrm{o}}|)}N^{|V_{\mathrm{i}}|+|E_\kappa^2|+|E_\kappa^3|/2-|V_{\mathrm{o}}|/2-\delta^{\geq 4}}$$
$$\times K^{|V_{\mathrm{o}}|-|V_{\mathrm{i}}|-2|E_\kappa^2|-|E_\kappa^3|/2+\left|V_{\mathrm{cyc}}^{o=0}\right|+\left|V_{\mathrm{cyc}}^{o=1}\right|/2} \tag{102a}$$

and for each iso-graph $\Gamma \in \mathcal{G}$ we have $|\mathrm{Val}(\Gamma_{\mathrm{red}})| \prec \mathrm{I}_3\text{-}\mathrm{Est}(\Gamma)$ with

$$\mathrm{I}_3\text{-}\mathrm{Est}(\Gamma) := \Lambda_+^{\left|V_{\mathrm{o}}^{\mathrm{0tr}}\right|}\Pi_+^{\left|V_{\mathrm{o}}^t\right|}\rho^{2ip\vee(|V_{\mathrm{i}}|+2|E_\kappa^2|+|E_\kappa^3|-|V_{\mathrm{o}}|)}N^{|V_{\mathrm{i}}|+|E_\kappa^2|+|E_\kappa^3|/2-|V_{\mathrm{o}}|/2-\delta^{\geq 4}}$$
$$\times K^{|V_{\mathrm{o}}|-|V_{\mathrm{i}}|-2|E_\kappa^2|-|E_\kappa^3|/2+\left|V_{\mathrm{cyc}}^{o=0}\right|+\left|V_{\mathrm{cyc}}^{o=1}\right|/2-\left(p-|E_\kappa^2|+|V_{\mathrm{cyc}}|-\delta^{\geq 4}\right)_+} \tag{102b}$$

Before proving Lemmata 2–4 we conclude the proof of Proposition 4.

Proof. (Proposition 4) The proof of Proposition 4 distinguishes several cases. For the averaged bound we consider the two cases $a = t = 0$ and $a+t = o > 0$, $|V_{\mathrm{i}} \cap V_{\mathrm{o}}| = 2(o-1)p$ separately, while for the isotropic bound we consider the cases $o \geq 0$, $|V_{\mathrm{i}} \cap V_{\mathrm{o}}| = 2op$ and $o > 0$, $|V_{\mathrm{i}} \cap V_{\mathrm{o}}| = 2(o-1)p$ separately.

We first consider the $o = 0$ case of the averaged bound where we obtain from Lemma 2, (102a) with $V_{\mathrm{o}} = \emptyset$ from (P8), $\left|V_{\mathrm{cyc}}\right| \leq \left|E_\kappa^2\right|$ from (P6) and $|V_{\mathrm{i}}| = 2p(l-1)$ from (P^{av}10) that

$$|\mathrm{Val}(\Gamma)| \prec \rho^{|V_{\mathrm{i}}|+2|E_\kappa^2|+|E_\kappa^3|}N^{|V_{\mathrm{i}}|+|E_\kappa^2|+|E_\kappa^3|/2}K^{-|V_{\mathrm{i}}|-|E_\kappa^2|-|E_\kappa^3|/2}$$
$$= \rho^{2(l-1)p+2|E_\kappa^2|+|E_\kappa^3|}N^{2lp}K^{-2p(l-1)}N^{-2p+|E_\kappa^2|+|E_\kappa^3|/2}K^{-|E_\kappa^2|-|E_\kappa^3|/2}$$
$$\lesssim \rho^{2p(l+1)}N^{2lp}K^{-2lp},$$

where in the last step we used $K \lesssim N\rho^2$ due to $\eta = \min_k \eta_k \lesssim \max_k \rho_k = \rho$ and $\left|E_\kappa^2\right| + \left|E_\kappa^3\right|/2 \leq |E_\kappa| \leq 2p$ from (P2) and (P4) .

Next, we consider the $|V_{\mathrm{o}} \cap V_{\mathrm{i}}| = 2op$ case of isotropic bound, where we obtain from Lemma 2, (102b), and $|V_{\mathrm{i}}| = 2p(o+b-1)$ from (P^{iso}10) that

$$|\mathrm{Val}(\Gamma)| \prec \Lambda_+^{2ap}\Pi_+^{2tp}\rho^{2ip\vee(2p(b-1)+2|E_\kappa^2|+|E_\kappa^3|)}N^{p(o+2b)}N^{|E_\kappa^2|+|E_\kappa^3|/2-2p-\delta^{\geq 4}}$$
$$\times K^{p(1-2b)-|E_\kappa^2|-|E_\kappa^3|/2+\delta^{\geq 4}}$$
$$\lesssim \Lambda_+^{2ap}\Pi_+^{2tp}\rho^{2ip\vee 2p(b+1)}N^{p(o+2b)}K^{-p(1+2b)}$$

again using $K \lesssim N\rho^2$ and $\left|E_\kappa^2\right| + \left|E_\kappa^3\right|/2 \leq |E_\kappa| \leq 2p$.

Next, we consider the $|V_{\mathrm{o}} \cap V_{\mathrm{i}}| = 2(o-1)p$, $o > 0$ case of both the averaged bound, and the isotropic bound where we similarly obtain (from estimating $(\ldots)_+ \geq 0$ for the iso-graphs)

$|\mathrm{Val}(\Gamma)|$

$$\prec \Lambda_+^{\left|V_o^{0tr}\right|} \Pi_+^{\left|V_o^t\right|} \rho^{i'} N^{|V_i|+\left|E_\kappa^2\right|+\left|E_\kappa^3\right|/2-|V_o|/2} K^{|V_o|-|V_i|-2\left|E_\kappa^2\right|-\left|E_\kappa^3\right|/2+\left|V_{cyc}^{o=0}\right|+\left|V_{cyc}^{o=1}\right|/2}$$

$$= \Lambda_+^{\left|V_o^{0tr}\right|} \Pi_+^{\left|V_o^t\right|} \rho^{i'} N^{p(o+2b-1)+\left|E_\kappa^2\right|+\left|E_\kappa^3\right|/2-|V_o \cap V_\kappa|/2}$$

$$\times K^{|V_o \cap V_\kappa|-2pb-2\left|E_\kappa^2\right|-\left|E_\kappa^3\right|/2+\left|V_{cyc}^{o=0}\right|+\left|V_{cyc}^{o=1}\right|/2}$$

$$\le \Lambda_+^{2pa} \Pi_+^{2pt} \rho^{i'} N^{p(o+2b)} K^{-(2b+1)p} \left(\frac{K}{N}\right)^{p+|V_o \cap V_\kappa|/2-\left|E_\kappa^2\right|-\left|E_\kappa^3\right|/2}$$

$$\lesssim \Lambda_+^{2pa} \Pi_+^{2pt} \rho^{2ip\vee 2(b+1)p} N^{p(o+2b)} K^{-(2b+1)p}, \tag{103}$$

with

$$i' = 2ip \vee (|V_i| + 2\left|E_\kappa^2\right| + \left|E_\kappa^3\right| - |V_o|) = 2ip \vee \left(2bp + 2\left|E_\kappa^2\right| + \left|E_\kappa^3\right| - |V_o \cap V_\kappa|\right).$$

Here we used (P3) and (P^{av}10)/(P^{iso}10) and $V_o \subset V_i \cup V_\kappa^2$ (since by definition V_o are degree-2 vertices, while $V_e = \emptyset$ due to (P^{av}10) in the averaged case and $d_g(v) = 1$, $v \in V_e$ due to (P^{iso}10) in the isotropic case and $V_\kappa^{\ge 3}$-vertices have degree at least 3 by (P4)) in the equality. Furthermore, we used (P6) in the first inequality, and (P8) in the second inequality, and $K/N \lesssim \rho^2$ and (P8) in the final step. $\qquad\square$

5.2.1. Graph reduction: Proof of Lemma 2

Since for $d_g((uv)) = 2$ we have $\kappa_{ab}^{(uv)} = 1$ or $\kappa_{ab}^{(uv)} = \sigma$ for all a, b due to (78) (using Assumption (A-i)) it is possible to write (with potential transpositions) the summation over a_v for $v \in V_i \cup V_\kappa^2$ as matrix products which are then associated with edges of the reduced graph Γ_{red}. In this way G-chains $(v_1 v_2), \ldots, (v_{k-1} v_k) \in E_g$ with $v_2, \ldots, v_{k-1} \in V_\kappa^2 \cup V_i$ and $v_1, v_k \notin V_\kappa^2 \cup V_i$ are reduced to the edge $(v_1 v_k) \in E_g^{red}$, and G-cycles $(v_1 v_2), \ldots, (v_k v_1) \in E_g$ with $v_1, \ldots, v_k \in V_\kappa^2 \cup V_i$ are reduced to isolated loops which we represent by the vertex $v_1 \in V_{cyc}$ and the loop-edge $(v_1 v_1) \in E_g^{red,cyc} \subset E_g^{red}$. For each cycle of length k we arbitrarily pick one of the k possible reductions since they are all equivalent.

The first relation in (99) follows trivially since for each of the carried out summations corresponding to $V_\kappa^2 \cup V_i$ the number of G-edges is reduced by one with the exception that for cycles the last index is kept in V_{cyc}. The second relation in (99) is a direct consequence of (P1). Next, the claim (100) follows from (P3) and by noting that the definition of $a(e), t(e)$ is consistent with the counting of $t/0tr$-vertices in Γ. This concludes the proof of Lemma 2.

5.2.2. Entrywise bound: Proof of Lemma 3

For edges in the reduced graph we use the bound from the following lemma. Note that $o(e) \le l(e)$ for cycles e and $o(e) \le l(e) - 1$ for chains e and therefore the exponents of K below are guaranteed to be non-positive.

Lemma 5. *For $e \in E_g^{red,cyc}$ we have the averaged bound*

$$\left|\langle \mathcal{G}^e\rangle\right| \prec \Lambda_+^{a(e)} \Pi_+^{t(e)} \rho^{i(e)\vee(l(e)-o(e)+\mathbf{1}[0<o(e)<l(e)])} N^{l(e)-\frac{o(e)}{2}-1}$$

$$\times K^{o(e)-l(e)+\mathbf{1}(o(e)=0)+\frac{\mathbf{1}(o(e)=1)}{2}} \tag{104a}$$

and for $e \in E_g^{\mathrm{red}} \setminus E_g^{\mathrm{red,cyc}}$ the isotropic bound

$$\left| \langle \boldsymbol{v}, \mathcal{G}^e \boldsymbol{w} \rangle \right| \prec \|\boldsymbol{v}\| \, \|\boldsymbol{w}\| \, \Lambda_+^{a(e)} \Pi_+^{t(e)} \rho^{i(e) \vee (l(e)-o(e)-\mathbf{1}[o(e)=l(e)-1])} N^{l(e)-\frac{o(e)}{2}-1}$$
$$\times \, K^{o(e)-l(e)+1} \tag{104b}$$

for any two deterministic vectors $\boldsymbol{v}$, $\boldsymbol{w}$. Moreover, the same bounds hold true if within the chain $\mathcal{G}$ absolute values of resolvents $|G(z)|$ appear in addition, to $(\Im G)^{(t)}$, $(G^*)^{(t)}$, $(G)^{(t)}$.

Remark 8. The estimates (104a)–(104b) are designed to take advantage of the asymptotic orthogonality vertices. Indeed, using that a posteriori we will show that $\Lambda_+ + \Pi_+ \prec 1$ in the bulk, $\rho \sim 1$, both inequalities essentially depend on the number of orthogonality-vertices as $(K/\sqrt{N})^o \sim (\sqrt{N}\eta)^o$ (ignoring some K factors in (104a) for $o = 0, 1$). Therefore as long as $\eta \ll N^{-1/2}$ the orthogonality helps and our bounds do exploit this effect. However, for $\eta \gg N^{-1/2}$ it is better to use (104a)–(104b) by simply ignoring the asymptotic orthogonality, i.e. choosing $\mathfrak{a} = \mathfrak{t} = \emptyset$.

Using Lemma 5, the proof of which we defer to the the end of the subsection, we now conclude the proof of Lemma 3. From Lemma 2 we obtain $\mathrm{Val}(\Gamma) = \mathrm{Val}(\Gamma_{\mathrm{red}})$ with $\mathrm{Val}(\Gamma_{\mathrm{red}})$ as in (96). By estimating $\left| \kappa_{ab}^{(uv)} \right| \lesssim 1$ and $\mathcal{G}$ via Lemma 5 we obtain from (100),

$$\prod_{(uv) \in E_\kappa^{\geq 3}} \left(\sum_{a_u, a_v} N^{-d_g((uv))/2} \right) = \prod_{k \geq 3} \left(N^{2-k/2} \right)^{|E_\kappa^k|} = N^{|E_\kappa^3|/2 - \delta^{\geq 4}},$$

and $\left| V_{\mathrm{cyc}} \right| = \left| E_g^{\mathrm{red,cyc}} \right|$ that

$$|\mathrm{Val}(\Gamma)| \prec \Lambda_+^{\left| V_o^{0\mathrm{tr}} \right|} \Pi_+^{\left| V_o^t \right|} \rho^{i''} N^{-\left| E_\kappa^2 \right| + \left| V_{\mathrm{cyc}} \right| + \left| E_\kappa^3 \right|/2 + \left| E_g \right| - \left| V_o \right|/2 - \left| E_g^{\mathrm{red}} \right| - \delta^{\geq 4}}$$
$$\times \, K^{\left| V_o \right| - \left| E_g \right| + \left| E_g^{\mathrm{red}} \right| - \left| V_{\mathrm{cyc}} \right| + \left| V_{\mathrm{cyc}}^{o=0} \right| + \left| V_{\mathrm{cyc}}^{o=1} \right|/2}$$
$$\lesssim \Lambda_+^{\left| V_o^{0\mathrm{tr}} \right|} \Pi_+^{\left| V_o^t \right|} \rho^{2ip \vee (|V_i| + 2\left| E_\kappa^2 \right| - |V_o|)} N^{\left| E_\kappa^2 \right| + \left| E_\kappa^3 \right|/2 + |V_i| - |V_o|/2 - \delta^{\geq 4}}$$
$$\times \, K^{|V_o| - |V_i| - 2\left| E_\kappa^2 \right| + \left| V_{\mathrm{cyc}}^{o=0} \right| + \left| V_{\mathrm{cyc}}^{o=1} \right|/2}, \tag{105}$$

where we used (99) in the second step. Here we counted the factors of ρ as

$$i'' := \sum_{e \in E_g^{\mathrm{red}}} i(e) \vee \begin{cases} l(e) - o(e) + \mathbf{1}(0 < o(e) < l(e)), & e \in E_g^{\mathrm{red,cyc}}, \\ l(e) - o(e) - \mathbf{1}(o(e) = l(e) - 1), & e \in E_g^{\mathrm{red}} \setminus E_g^{\mathrm{red,cyc}}, \end{cases}$$
$$\geq \sum_{e \in E_g^{\mathrm{red}}} i(e) \vee [l(e) - o(e) - \mathbf{1}(e \notin E_g^{\mathrm{red,cyc}})]$$
$$\geq 2ip \vee \left(\left| E_g \right| - |V_o| - \left| E_g^{\mathrm{red}} \right| + \left| V_{\mathrm{cyc}} \right| \right) = 2ip \vee \left(|V_i| + 2\left| E_\kappa^2 \right| - |V_o| \right) \tag{106}$$

due to (99), completing the proof of Lemma 3.

Proof. (Lemma 5) We actually prove a slightly more general bound which allows for chains $\mathcal{G}^e = G_1 B_1 \cdots G_l B_l$ with

$$G_k \in \left\{ G(z_k), G(z_k)^*, \Im G(z_k), |G(z_k)|, (G(z_k))^t, (G(z_k)^*)^t, (\Im G(z_k))^t, |G(z_k)|^t \right\},$$

i.e. including factors of the form $|G| = \sqrt{G^*G} = \sqrt{GG^*}$. Within the proof we will repeatedly use (98) which implies $o(e) \le l(e)$ for $e \in E_g^{\mathrm{red,cyc}}$ and $o(e) \le l(e) - 1$ for $e \in E_g^{\mathrm{red}} \setminus E_g^{\mathrm{red,cyc}}$. We prove (104a) by distinguishing several cases depending on the parameters $o(e)$ and $l(e)$ and a new parameter $c(e)$ counting the number of *alternating chains* associated with e defined as follows. For any $e \in E_g^{\mathrm{red}}$ we consider the original chain or cycle in Γ that was reduced to e. The alternating chains associated with e are the maximal subchains of these original chain/cycle with internal vertices from V_o and at least one V_o-vertex. For example, if $e \in E_g^{\mathrm{red}}$ was the reduction of the cycle $\langle (GA)(\Im GA)(BG^*)(G)(AG^*)(\Im G)^t \rangle$ then the alternating chains associated with e are $(GA)(\Im GA)$ and $(G)(AG^*)(\Im G)^t$. By maximality, $o(e)$, the number of V_o-vertices in the original chain/cycle that has been reduced to e is equal to the total number of V_o vertices in the alternating chains associated with e. In particular, $c(e) \le o(e)$.

Averaged bound for $o(e) = 0$ In the case without alternating chains, i.e. for $o(e) = 0$ we simply split off any G-factor by Cauchy-Schwarz and obtain

$$|\langle G_1 B_1 G_2 B \cdots G_l B_l \rangle| \le \sqrt{\langle G_1 |B_1|^2 G_1^* \rangle \langle G_2 B_2 \cdots G_l |B_l|^2 G_l^* \cdots B_2^* G_2^* \rangle}$$
$$\prec \frac{\rho}{\eta^{l-1}} \le \rho^l N^{l-1} K^{1-l}.$$

Here, and frequently in the remaining proof we use the norm bounds $\|G\| \lesssim 1/\eta$, $\|B_k\| \lesssim 1$, and the *Ward identity* $G(z)G(z)^* = \Im G(z)/\Im z$.

Averaged bound for $a(e) = l(e)$ For $\mathcal{G}^e = G_1 B_1 G_2 B_2 \cdots G_l B_l$ we use spectral decomposition to write

$$\langle \mathcal{G}^e \rangle = N^{-1} \sum_{a_1 \ldots a_l} \langle \boldsymbol{u}_{a_1}^{(1)}, B_1 \boldsymbol{u}_{a_2}^{(2)} \rangle \cdots \langle \boldsymbol{u}_{a_l}^{(l)}, B_l \boldsymbol{u}_{a_1}^{(1)} \rangle p_{a_1}^{(1)} \cdots p_{a_l}^{(l)},$$

where $p_a^{(k)} = (\lambda_a - z_k)^{-1}, (\lambda_a - \overline{z_k})^{-1}, \Im(\lambda_a - z_k)^{-1}, |\lambda_a - z_k|^{-1}$ depending on whether $G_k = G, G^*, \Im G, |G|$, and $\boldsymbol{u}_a^{(k)} \in \{\boldsymbol{u}_a, \overline{\boldsymbol{u}_a}\}$, depending on whether G_k is transposed or not. By additional averaging using the analogue of (40), Cauchy-Schwarz and the high-probability bounds

$$\sum_a \frac{1}{|\lambda_a - z|} \lesssim N \log N, \qquad \sum_a \left| \Im \frac{1}{\lambda_a - z} \right| \lesssim \rho(z) N, \tag{107}$$

from rigidity (34) it follows that

$$
|\langle \mathcal{G}^e \rangle| \lesssim \frac{1}{N} \sum_{\substack{a_k \in [N] \\ k \in [l]}} \left| p_{a_1}^{(1)} \right| \cdots \left| p_{a_l}^{(l)} \right| \frac{1}{L^l} \sum_{\substack{|b_k - a_k| \leq L \\ k \in [l]}} \left| \langle \boldsymbol{u}_{b_1}^{(1)}, B_1 \boldsymbol{u}_{b_2}^{(2)} \rangle \right| \cdots \left| \langle \boldsymbol{u}_{b_l}^{(l)}, B_l \boldsymbol{u}_{b_1}^{(1)} \rangle \right|
$$

$$
\lesssim \frac{1}{N} \sum_{\substack{a_k \in [N] \\ k \in [l]}} \left| p_{a_1}^{(1)} \right| \cdots \left| p_{a_l}^{(l)} \right| \sqrt{\frac{1}{L^2} \sum_{|a_1 - b_1| \leq L} \sum_{|a_2 - b_2| < L} \left| \langle \boldsymbol{u}_{b_1}^{(1)}, B_1 \boldsymbol{u}_{b_2}^{(2)} \rangle \right|^2} \cdots
$$

$$
\times \sqrt{\frac{1}{L^2} \sum_{|a_l - b_l| \leq L} \sum_{|a_1 - b_1| < L} \left| \langle \boldsymbol{u}_{b_l}^{(l)}, B_l \boldsymbol{u}_{b_1}^{(1)} \rangle \right|^2}
$$

$$
\prec \Lambda_+^a \Pi_+^t \rho^i N^{l/2 - 1}, \tag{108}
$$

where $\log N$ factors have been incorporated into the $\prec$ notation in the ultimate inequality.

Averaged bound for $o(e) = 1$ By cyclicity we may assume $\mathcal{G}^e = G_1 B_1 G_2 B_l \cdots G_l B_l$ is such that the index between $G_1 B_1$ and $G_2 B_2$ is the asymptotic orthogonality index and estimate

$$
|\langle G_1 B_1 G_2 B_2 \cdots G_l B_l \rangle| \leq \sqrt{\langle G_1 B_1 G_2 G_2^* B_2^* G_1^* \rangle \langle B_2 G_3 \cdots G_l |B_l|^2 G_l^* \cdots G_3^* B_2^* \rangle}
$$

$$
\prec \eta^{-1} \Lambda_+ \rho N^{l-5/2} \rho^{l-2} K^{5/2-l} \leq \Lambda_+ \rho^l N^{l-3/2} K^{3/2-l},
$$

from the $o(e) = 0$ and $o(e) = l(e)$ cases, and using the Ward identity.

Averaged bound for $2 \leq o(e) < l(e)$ and $c(e) = 1$ For this case we may assume by cyclicity that

$$
\mathcal{G}^e = G_1 B_1 \cdots G_o B_o G_{o+1} B_{o+1} \cdots G_l B_l
$$

such that the summations between G_1 and G_o correspond to orthogonality indices. Here we make use of the inequality

$$
|\langle XYZ \rangle| \leq \left[\langle X^* X (YY^*)^{1/2} \rangle \langle ZZ^* (Y^*Y)^{1/2} \rangle \right]^{1/2} \tag{109}
$$

for arbitrary matrices X, Y, Z which follows from singular value decomposition of $Y = USV^*$ and Cauchy-Schwarz in the form

$$
|\langle XYZ \rangle|^2 = \left| \langle XU\sqrt{S}\sqrt{S}V^*Z \rangle \right|^2
$$

$$
\leq \langle XUSU^*X^* \rangle \langle Z^*VSV^*Z \rangle = \langle X^*X(YY^*)^{1/2} \rangle \langle ZZ^*(Y^*Y)^{1/2} \rangle.
$$

By (109) with $X = G_1 B_1, Y = G_2, Z = B_2 G_3 \cdots B_o G_{o+1} B_{o+1} \cdots G_l B_l$ we obtain

$$
|\langle G_1 B_1 \cdots G_o B_o G_{o+1} \cdots G_l B_l \rangle|
$$

$$
\leq \sqrt{\langle B_1^* G_1^* G_1 B_1 |G_2| \rangle \langle |G_2|^{1/2} B_2 G_3 \cdots G_l |B_l|^2 G_l^* \cdots G_3^* B_2^* |G_2|^{1/2} \rangle}
$$

$$
\lesssim \frac{1}{\eta^{l-o}} \left[\langle B_1^* \Im G_1 B_1 |G_2| \rangle \langle |G_2| B_2 G_3 \cdots B_o \Im G_{o+1} B_o^* \cdots G_3^* B_2^* \rangle \right]^{1/2}
$$

$$
\prec \Lambda_+^a \Pi_+^t \rho^{l-o+1+i_{2\ldots o}} N^{l-o/2-1} K^{o-l},
$$

where $i_{2\ldots o}$ is the number of $\Im G$'s among $G_2, \ldots G_o$, and we used the previously considered $o(e) = l(e)$ case in the last step.

Averaged bound for $2 \le o(e) < l(e)$ and $c(e) \ge 2$ For at least two alternating chains, $c(e) \ge 2$, we may write by cyclicity $\langle \mathcal{G}^e \rangle = \langle \mathcal{G}^{e_1} \cdots \mathcal{G}^{e_{c(e)}} \rangle$ for

$$\mathcal{G}^{e_j} = G_{j,1} B_{j,1} G_{j,2} \cdots B_{j,o_j} G_{j,o_j+1} B_{j,o_j+1} \cdots G_{j,l_j} B_{j,l_j},$$

for some $1 \le o_j \le l_j - 1$ such for each $\mathcal{G}^{e_j}$ the first o_j internal summation indices are orthogonality indices. By Cauchy-Schwarz it follows that

$$
\begin{aligned}
\left| \langle \mathcal{G}^e \rangle \right| &\le \frac{1}{N} \sqrt{ \prod_{j \in [c(e)]} \operatorname{Tr} \mathcal{G}^{e_j} (\mathcal{G}^{e_j})^* } \\
&\lesssim \frac{1}{N} \prod_{j \in [c(e)]} \frac{1}{\eta^{l_j - o_j}} \sqrt{ \operatorname{Tr} \Im G_{j,1} B_{j,1} \cdots G_{j,o_j} B_{j,o_j} \Im G_{j,o_j+1} B_{j,o_j}^* G_{j,o_j}^* \cdots B_{j,1}^* } \\
&\prec \frac{1}{N} \prod_{j \in [c(e)]} \frac{N^{o_j/2} \rho^{i_j+1}}{\eta^{l_j - o_j}} \Lambda_+^{a_j} \Pi_+^{t_j} \\
&\le \Lambda_+^{\sum_j a_j} \Pi_+^{\sum_j t_j} \rho^{\sum_j (l_j - o_j + i_j + 1)} N^{\sum_j (l_j - o_j/2) - 1} K^{\sum_j (o_j - l_j)}, \quad\quad (110)
\end{aligned}
$$

where i_j denotes the number of $\Im G$'s among $G_{j,2}, \ldots, G_{j,o_j}$, and we used the previously discussed $o(e) = l(e)$ case in the third inequality. This concludes the proof of (104a).

Isotropic bound for $o(e) = l(e) - 1$ The claimed bound is trivial if $l(e) = 1$ (and hence $o(e) = 0$). Otherwise for $l(e) \ge 2$ we estimate

$$
\begin{aligned}
&\left| \langle \boldsymbol{v}, \mathcal{G}^e \boldsymbol{w} \rangle \right| \\
&= \left| \langle \boldsymbol{v}, G_1 B_1 G_2 B_2 \cdots B_{l-1} G_l, \boldsymbol{w} \rangle \right| \\
&\lesssim \sum_{\substack{a_k \in [N] \\ k \in [l]}} \left| p_{a_1}^{(1)} \right| \cdots \left| p_{a_l}^{(l)} \right| \left| \langle \boldsymbol{v}, \boldsymbol{u}_{a_1}^{(1)} \rangle \right| \left| \langle \boldsymbol{u}_{a_1}^{(1)}, B_1 \boldsymbol{u}_{a_2}^{(2)} \rangle \right| \cdots \left| \langle \boldsymbol{u}_{a_{l-1}}^{(l-1)}, B_{l-1} \boldsymbol{u}_{a_l}^{(l)} \rangle \right| \left| \langle \boldsymbol{u}_{a_l}^{(l)}, \boldsymbol{w} \rangle \right| \\
&\prec \frac{1}{N} \sum_{\substack{a_k \in [N] \\ k \in [l]}} \left| p_{a_1}^{(1)} \right| \cdots \left| p_{a_l}^{(l)} \right| \frac{1}{L^l} \sum_{\substack{|b_k - a_k| \le L \\ k \in [l]}} \left| \langle \boldsymbol{u}_{b_1}^{(1)}, B_1 \boldsymbol{u}_{b_2}^{(2)} \rangle \right| \cdots \left| \langle \boldsymbol{u}_{b_{l-1}}^{(l-1)}, B_{l-1} \boldsymbol{u}_{b_l}^{(l)} \rangle \right| \\
&\prec \Lambda_+^a \Pi_+^t \rho^i N^{l/2 - 1/2},
\end{aligned}
$$

using delocalisation $|\langle \boldsymbol{u}_a, \boldsymbol{v} \rangle| + |\langle \overline{\boldsymbol{u}_a}, \boldsymbol{v} \rangle| \prec N^{-1/2}$ for any deterministic $\boldsymbol{v}$ with $\|\boldsymbol{v}\| \lesssim 1$, by the isotropic law in (12), in the second inequality.

Isotropic bound for $o(e) \le l(e) - 2$ We decompose $\mathcal{G}^e = \mathcal{G}^{e_1} \cdots \mathcal{G}^{e_k}$ such that each of $\mathcal{G}^{e_2}, \ldots, \mathcal{G}^{e_{k-1}}$ begins with a new alternating chain followed (potentially) by further G's, $\mathcal{G}^{e_1}$ either begins with an alternating chain, or is a chain without orthogonality indices, and $\mathcal{G}^{e_k}$ is either an alternating chain or a chain without orthogonality indices. For example, by brackets denoting the decomposition, we would separate

$$\langle \boldsymbol{v}, (G B_1 G^* B_2 (\Im G)^t B_3)(G B_4 G^t B_5)(G^* B_6) \boldsymbol{w} \rangle$$

if the indices associated with B_1, B_2, B_4 are orthogonality indices, and estimate

$$
\begin{aligned}
&\left| \langle \boldsymbol{v}, \mathcal{G}^e \boldsymbol{w} \rangle \right| \\
&\le \left[\langle \boldsymbol{v}, \mathcal{G}^{e_1} (\mathcal{G}^{e_1})^* \boldsymbol{v} \rangle (\operatorname{Tr} \mathcal{G}^{e_2} (\mathcal{G}^{e_2})^*) \cdots (\operatorname{Tr} \mathcal{G}^{e_{k-1}} (\mathcal{G}^{e_{k-1}})^*) \langle \boldsymbol{w}, (\mathcal{G}^{e_k})^* \mathcal{G}^{e_k} \boldsymbol{w} \rangle \right]^{1/2}.
\end{aligned}
$$

For the two isotropic factors of length l_j with o_j orthogonality indices and i_j many $\Im G$'s we claim that

$$\left| \langle \boldsymbol{v}, \mathcal{G}^{e_j} (\mathcal{G}^{e_j})^* \boldsymbol{v} \rangle \right| \prec \frac{N^{2l_j - o_j - 1} \rho^{2i_j \vee 2(l_j - o_j)} \Lambda_+^{2a_j} \Pi_+^{2t_j}}{K^{2(l_j - o_j) - 1}} \tag{111}$$

which follows from

$$
\begin{aligned}
&\left| \langle \boldsymbol{v}, G_1 B_1 \cdots G_o B_o G_{o+1} B_{o+1} \cdots G_l G_l^* \cdots B_{o+1}^* G_{o+1}^* B_o^* G_o \cdots B_1^* G_1^* \boldsymbol{v} \rangle \right| \\
&\qquad \lesssim \frac{\left| \langle \boldsymbol{v}, G_1 B_1 \cdots G_o B_o \Im G_{o+1} B_o^* G_o \cdots B_1^* G_1^* \boldsymbol{v} \rangle \right|}{\eta^{2(l-o)-1}} \\
&\qquad \prec \frac{N^{2l-o-1} \rho^{2i_{1\cdots o} + 2(l-o)} \Lambda_+^{2a} \Pi_+^{2t}}{K^{2(l-o)-1}},
\end{aligned}
$$

where $i_{1\cdots o}$ is the number of $\Im G$'s among $G_1, \ldots, G_o$. For the tracial factors we have, as in (110), that

$$\operatorname{Tr} \mathcal{G}^{e_j} (\mathcal{G}^{e_j})^* \prec \frac{N^{2l_j - o_j} \rho^{2i_j \vee 2(l_j - o_j)} \Lambda_+^{2a_j} \Pi_+^{2t_j}}{K^{2(l_j - o_j)}}. \tag{112}$$

By combining (111)–(112) we obtain

$$
\begin{aligned}
\left| \langle \boldsymbol{v}, \mathcal{G}^e \boldsymbol{w} \rangle \right| &\prec \frac{K}{N} \prod_{j \in [k]} \frac{N^{l_j - o_j/2} \rho^{i_j \vee (l_j - o_j)} \Lambda_+^{a_j} \Pi_+^{t_j}}{K^{l_j - o_j}} \\
&= \Lambda_+^a \Pi_+^t \rho^{i \vee (l-o)} N^{l - o/2 - 1} K^{o - l + 1},
\end{aligned}
$$

completing the proof of (104b) also in this case. □

5.2.3. Improved degree three estimate: Proof of Lemma 4 The proof of Lemma 4 consists of identifying improvements over the estimate given in Lemma 3 that relied solely on entrywise bounds for each individual $\mathcal{G}$-factor. In order to quantify the improvement we distinguish the two different entrywise bounds in Lemma 3 as

$$|\mathrm{Val}(\Gamma_{\mathrm{red}})| \prec \mathrm{I}_2^{\mathrm{i}}\text{-Est}(\Gamma) \wedge \mathrm{I}_2^0\text{-Est}(\Gamma), \tag{113}$$

where $\mathrm{I}_2^{\mathrm{i}}$-Est, I_2^0-Est are defined as in (101) but with $\mathrm{I}_2^{\mathrm{i}}$-Est having ρ-exponent $2ip$, and I_2^0-Est having ρ-exponent $|V_{\mathrm{i}}| + 2 \left| E_\kappa^2 \right| - |V_o|$. Note that $\rho \lesssim 1$ and therefore the maximum in the exponent of ρ in (101) corresponds to the minimum of $\mathrm{I}_2^{\mathrm{i}}$-Est, I_2^0-Est.

Within the reduced graphs we call a subset $E_{\mathrm{Ward}} \subset E_g^{\mathrm{red}} \backslash (E_g^{\mathrm{red,cyc}} \cup \left\{ (vv) \big| v \in V_\kappa^{\geq 3} \right\})$ *Wardable* if each subgraph $\Gamma' \subset (V_{\mathrm{e}} \cup V_\kappa^{\geq 3}, E_{\mathrm{Ward}})$ satisfies $\min \left\{ d_g^{\Gamma'}(v) \big| v \in V_\kappa^{\geq 3} \right\} \leq 2$. The contribution of these Wardable edges will be estimated better than their trivial entrywise bound to obtain I_3-Est. We start with a simple alternative characterization of Wardable subsets (see [27, Lemma 4.5] and [31,41]).

Lemma 6. *A subset E_{Ward} is Wardable if and only if there exists an ordering $V_\kappa^{\geq 3} = \{v_1, v_2, \ldots\}$ such that the sequence of graphs $\Gamma_0 := (V_{\mathrm{e}} \cup V_\kappa^{\geq 3}, E_{\mathrm{Ward}})$, $\Gamma_k := \Gamma_{k-1} \backslash \{v_k\}$ satisfies $d_g^{\Gamma_{k-1}}(v_k) \leq 2$ for each $k \geq 1$, where it is understood that Γ_k is obtained from Γ_{k-1} by removing v_k and all adjacent edges.*

Proof. Suppose that E_{Ward} is Wardable. Then by definition there exists v_1 with $d_g^{\Gamma_0}(v_1) \leq 2$ and we obtain Γ_1 which in turn contains some vertex v_2 with $d_g^{\Gamma_1}(v_2) \leq 2$. Continuing inductively yields the desired ordering.

For the reverse implication let $v_1, v_2, \ldots$ be the given ordering and let Γ' be arbitrary. Set $k_{\min} := \min \left\{ k \middle| v_k \in \Gamma' \right\}$ so that $\Gamma' \subset \Gamma_{k_{\min}-1}$ and consequently $d_g^{\Gamma'}(v_{k_{\min}}) \leq d_g^{\Gamma_{k_{\min}-1}}(v_{k_{\min}}) \leq 2$.   □

Lemma 4 follows immediately from combining the following two statements (where for the iso-graphs we simply estimate $\rho^{|E_{\text{Ward}}|} \leq \rho^{|E_\kappa^3|}$ in the definition of I$_3$-Est(Γ) below):

(S1) For each av-graph Γ the reduced graph Γ_{red} admits a Wardable set E_{Ward} of size

$$|E_{\text{Ward}}| \geq \left| E_\kappa^3 \right|. \tag{114a}$$

and for each iso-graph Γ satisfying the reduced graph Γ_{red} admits a Wardable set of size

$$|E_{\text{Ward}}| \geq \left| E_\kappa^3 \right| + \left(2p - 2 (\left| E_\kappa^2 \right| - \left| V_{\text{cyc}} \right|) - \sum_{e \in E_\kappa} (d_g(e) - 4)_+ \right)_+. \tag{114b}$$

(S2) For any av- or iso-graph $\Gamma \in \mathcal{G}$ and a given Wardable set E_{Ward} we have the improved estimates

$$|\text{Val}(\Gamma_{\text{red}})| \prec \text{I}_3^{\text{i}}\text{-Est}(\Gamma) \wedge \text{I}_3^0\text{-Est}(\Gamma)$$

with

$$\text{I}_3^{\text{i}}\text{-Est}(\Gamma) := K^{-|E_{\text{Ward}}|/2} \, \text{I}_2^{\text{i}}\text{-Est}(\Gamma),$$
$$\text{I}_3^0\text{-Est}(\Gamma) := \rho^{|E_{\text{Ward}}|} K^{-|E_{\text{Ward}}|/2} \, \text{I}_2^0\text{-Est}(\Gamma).$$

Proof. (Step (S1)) We start with two inequalities that will be proven later. Denoting the number of E_g^{red}-edges between two subsets of vertices $V', V'' \subset V$ by $e_g(V', V'')$, we claim that for av-/iso graphs Γ we have

$$e_g(V_\kappa^3, V_e) + e_g(V_\kappa^3, V_\kappa^{\geq 3}) \geq 3 \left| E_\kappa^3 \right|, \tag{115a}$$

while for iso-graphs Γ we also have

$$e_g(V_\kappa^{\geq 3}, V_e) + e_g(V_\kappa^{\geq 3}, V_\kappa^{\geq 3}) \geq \sum_{e \in E_\kappa^{\geq 3}} d_g(e) + 2p - 2 (\left| E_\kappa^2 \right| - \left| V_{\text{cyc}} \right|). \tag{115b}$$

Armed with these inequalities, we first construct *candidate sets* of edges within $E_g^{\text{red}} \setminus E_g^{\text{red,cyc}}$ which are not necessarily Wardable, and then iteratively remove certain edges to make the sets Wardable. For the proof of $|E_{\text{Ward}}| \geq \left| E_\kappa^3 \right|$ for both av- and iso-graphs we start with the candidate set consisting of all G-edges adjacent to V_κ^3-vertices. The size of this set is $e_g(V_\kappa^3, V_e) + e_g(V_\kappa^3, V_\kappa^{\geq 3})$. We remove at most one edge adjacent to any $v \in V_\kappa^3$, so that the at most two remaining edges are not loops. After doing so in arbitrary order for all V_κ^3-vertices we obtain an edge set which is Wardable by

construction. Since the total number of removed edges is at most $\left|V_\kappa^3\right| = 2\left|E_\kappa^3\right|$, we immediately obtain (114a), and (114b) in case $(\ldots)_+ = 0$ from (115a).

For the proof of (114b) in case $(\ldots)_+ > 0$ we consider a larger candidate set of size $e_g(V_\kappa^{\geq 3}, V_e) + e_g(V_\kappa^{\geq 3}, V_\kappa^{\geq 3})$ that consists of all edges adjacent to $V_\kappa^{\geq 3}$-vertices. Going through all $V_\kappa^{\geq 3}$-vertices in arbitrary order we remove at most $k - 2$ edges for each vertex $v \in V_\kappa^k$, so that the at most two remaining edges are not loops; this yields again a Wardable set. Since $\left|V_\kappa^k\right| = 2\left|E_\kappa^k\right|$, the total number of removed edges is at most

$$\sum_{k \geq 3} \sum_{e \in E_\kappa^k} 2(k-2) = \sum_{e \in E_\kappa^{\geq 3}} (2d_g(e) - 4) = \sum_{e \in E_\kappa^{\geq 3}} d_g(e) + \sum_{e \in E_\kappa} (d_g(e) - 4)_+ - \left|E_\kappa^3\right|,$$

which, together with (115b) yields (114b). This completes the proof of (S1) modulo (115) that we prove now. $\qquad\square$

Proof. (Eq. (115)) The bound (115a) follows from

$$6\left|E_\kappa^3\right| = 2 \sum_{(uv) \in E_\kappa^3} d_g((uv)) = \sum_{v \in V_\kappa^3} d_g(v)$$
$$= 2e_g(V_\kappa^3, V_\kappa^3) + e_g(V_\kappa^3, V_\kappa^{\geq 4} \cup V_e) \leq 2e_g(V_\kappa^3, V_\kappa^{\geq 3} \cup V_e).$$

For the bound (115b) we note that the set $E_g^{\text{red}} \setminus E_g^{\text{red,cyc}}$ can be partitioned into edges within $V_\kappa^{\geq 3}$, edges within V_e and edges between these two sets, and thus from (P3)–($P^{\text{iso}}10$) and (99) we obtain

$$e_g(V_\kappa^{\geq 3}, V_e) + e_g(V_\kappa^{\geq 3}, V_\kappa^{\geq 3})$$
$$= \left|E_g^{\text{red}} \setminus E_g^{\text{red,cyc}}\right| - e_g(V_e, V_e) = \left|E_g^{\text{red}}\right| - \left|V_{\text{cyc}}\right| - e_g(V_e, V_e)$$
$$= \sum_{e \in E_\kappa^{\geq 3}} d_g(e) + 2p - e_g(V_e, V_e).$$

Furthermore, by ($P^{\text{iso}}9$) each V_e-V_e edge corresponds to at least one V_κ^2-vertex, while by (P5) each cycle $E_g^{\text{red,cyc}}$ corresponds to at least two V_κ^2-vertices in Γ (which are in particular not part of any chain), whence

$$e_g(V_e, V_e) \leq \left|V_\kappa^2\right| - 2\left|V_{\text{cyc}}\right| = 2(\left|E_\kappa^2\right| - \left|V_{\text{cyc}}\right|)$$

and the claim follows. $\qquad\square$

Proof. (Claim 5.2.3) We recall from the proof of Lemma 3 that (101) is the minimum of two different estimates given in (113). Estimating each $\mathcal{G}^e$ for $e \in E_g^{\text{red}}$ by Lemma 5 with a ρ-exponent of $i(e)$ in (96) yields the first bound $|\text{Val}(\Gamma_{\text{red}})| \prec I_2^{\frac{1}{2}}\text{-Est}(\Gamma)$. Similarly, estimating each $\mathcal{G}^e$ by Lemma 5 with a ρ-exponent of $l(e) - o(e) - \mathbf{1}(e \in E_g^{\text{red}} \setminus E_g^{\text{red,cyc}})$ yields the second bound $|\text{Val}(\Gamma_{\text{red}})| \prec I_2^0\text{-Est}(\Gamma)$, cf. the first inequality in (106). In order to prove 5.2.3 for a given Wardable set E_{Ward} we estimate $\mathcal{G}^e$ for $e \in E_g^{\text{red,cyc}} \cup (E_g^{\text{red}} \setminus (E_g^{\text{red,cyc}} \cup E_{\text{Ward}}))$ exactly as in Lemma 3 and remove the corresponding edges from the graph, leaving only E_{Ward}-edges. In order to conclude the proof it remains to establish

an additional gain of $K^{-1/2}$ (compared to the first bound) and $\rho K^{-1/2}$ (compared to the second bound) per E_{Ward}-edge e compared to the entrywise estimates.

Let $v_1, v_2, \dots$ denote the ordering of $V_\kappa^{\geq 3}$ guaranteed to exist by Lemma 6. By definition of E_{Ward} at most two Wardable edges are adjacent to v_1 and whence the part of the value depending on a_{v_1} can be estimated by either

$$\sum_{a_{v_1}} \left| \mathcal{G}_{\boldsymbol{x}_w a_{v_1}}^{(wv)} \right| \leq N^{1/2} \sqrt{[\mathcal{G}^{(wv_1)}(\mathcal{G}^{(wv_1)})^*]_{\boldsymbol{x}_w \boldsymbol{x}_x}} \tag{116}$$

or

$$\sum_{a_{v_1}} \left| \mathcal{G}_{\boldsymbol{x}_w a_{v_1}}^{(wv_1)} \right| \left| \mathcal{G}_{a_{v_1} \boldsymbol{x}_y}^{(v_1 y)} \right| \leq \sqrt{[\mathcal{G}^{(wv_1)}(\mathcal{G}^{(wv_1)})^*]_{\boldsymbol{x}_w \boldsymbol{x}_w}} \sqrt{[(\mathcal{G}^{(v_1 y)})^* \mathcal{G}^{(yv_1)}]_{\boldsymbol{x}_y \boldsymbol{x}_y}} \tag{117}$$

using Cauchy-Schwarz for some $w, y \in V_\kappa^{\geq 3} \cup V_e$. In case of I_2^i-Est the entrywise estimate on the lhs. of (116)–(117) used in the proof of Lemma 3 is at least

$$\Lambda_+^a \Pi_+^t \rho^i N^{l-o/2} K^{o-l+1} \quad \text{and} \quad \Lambda_+^{a+a'} \Pi_+^{t+t'} \rho^{i+i'} N^{l+l'-o/2-o'/2-1} K^{o+o'-l-l'+2}$$

with $i = i((wv_1))$, $l = l((wv_1))$, $a = a((wv_1))$, $t = t((wv_1))$, $o = t + a$ and $i' = i((v_1 y))$, $l' = l((v_1 y))$, $a' = a((v_1 y))$, $t' = t((v_1 y))$, $o' = t' + a'$ while applying Lemma 5 to the rhs. yields

$$\Lambda_+^a \Pi_+^t \rho^i N^{l-o/2} K^{o-l+1/2} \quad \text{and} \quad \Lambda_+^{a+a'} \Pi_+^{t+t'} \rho^{i+i'} N^{l+l'-o/2-o'/2-1} K^{o+o'-l-l'+1},$$

demonstrating the gains of at least $K^{-1/2}$ and $(K^{-1/2})^2$, respectively. Similarly, the I_2^0-Est-estimate on the lhs. of (116)–(117) is at least

$$\Lambda_+^a \Pi_+^t \rho^{l-o-1} N^{l-o/2} K^{o-l+1}$$
$$\text{and}$$
$$\Lambda_+^{o+o'} \rho^{l+l'-o-o'-2} N^{l+l'-o/2-o'/2-1} K^{o+o'-l-l'+2}$$

while, in comparison, when applying Lemma 5 to the rhs. of (116)–(117), we obtain bounds of

$$\Lambda_+^a \Pi_+^t \rho^{l-o} N^{l-o/2} K^{o-l+1/2} \quad \text{and} \quad \Lambda_+^{o+o'} \rho^{l+l'-o-o'} N^{l+l'-o/2-o'/2-1} K^{o+o'-l-l'+1},$$

demonstrating exactly the claimed gain of $\rho K^{-1/2}$ per edge. Here, for example, we counted that $\mathcal{G}^{(wv_1)}(\mathcal{G}^{(wv_1)})^*$ contains $2l$ factors of G and $2o$ orthogonality indices satisfying $2o \leq 2l - 2 < 2l - 1$.

The proof now follows by induction since by Lemma 6 after the removal of v_1, the next vertex v_2 has degree at most 2 etc. and (116)–(117) can be used to establish the gain of $(\rho) K^{-1/2}$ iteratively for each $e \in E_{\text{Ward}}$. $\square$

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Anantharaman, N.: Quantum ergodicity on regular graphs. Commun. Math. Phys. **353**, 633–690 (2017)
2. Anantharaman, N., Le Masson, E.: Quantum ergodicity on large regular graphs. Duke Math. J. **164**, 723–765 (2015)
3. Anantharaman, N., Sabri, M.: Quantum ergodicity on graphs: from spectral to spatial delocalization. Ann. Math. **2**(189), 753–835 (2019)
4. Bao, Z., Erdős, L., Schnelli, K.: Equipartition principle for Wigner matrices. Forum Math. Sigma **9**(e44), 21 (2021)
5. Barnett, A.H.: Asymptotic rate of quantum ergodicity in chaotic Euclidean billiards. Commun. Pure Appl. Math. **59**, 1457–1488 (2006)
6. Bauerschmidt, R., Huang, J., Yau, H.-T.: Local Kesten–McKay law for random regular graphs. Commun. Math. Phys. **369**, 523–636 (2019)
7. Bauerschmidt, R., Knowles, A., Yau, H.-T.: Local semicircle law for random regular graphs. Commun. Pure Appl. Math. **70**, 1898–1960 (2017)
8. Benigni, L.: Eigenvectors distribution and quantum unique ergodicity for deformed Wigner matrices. Ann. Inst. Henri Poincareé Probab. Stat. **56**, 2822–2867 (2020)
9. Benigni, L.: Fermionic eigenvector moment flow, preprint (2019). arXiv:1908.10855
10. Benigni, L., Lopatto, P.: Optimal Delocalization for Generalized Wigner Matrices, preprint (2020). arXiv:2007.09585
11. Bloemendal, A., Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Isotropic local laws for sample covariance and generalized Wigner matrices. Electron. J. Probab. **19**(33), 53 (2014)
12. Bourgade, P., Yau, H.-T.: The eigenvector moment flow and local quantum unique ergodicity. Commun. Math. Phys. **350**, 231–278 (2017)
13. Bourgade, P., Yau, H.-T., Yin, J.: Random band matrices in the delocalized phase I: quantum unique ergodicity and universality. Commun. Pure Appl. Math. **73**, 1526–1596 (2020)
14. Boutet de Monvel, A., Khorunzhy, A.: Asymptotic distribution of smoothed eigenvalue density. II. Wigner random matrices. Random Oper. Stoch. Equ. **7**, 149–168 (1999)
15. Brooks, S., Lindenstrauss, E.: Joint quasimodes, positive entropy, and quantum unique ergodicity. Invent. Math. **198**, 219–259 (2014)
16. Brooks, S., Lindenstrauss, E.: Non-localization of eigenfunctions on large regular graphs. Israel J. Math. **193**, 1–14 (2013)
17. Cipolloni, G., Erdős, L., Schröder, D.: Central limit theorem for linear eigenvalue statistics of non-Hermitian random matrices. Commun. Pure Appl. Math. (to appear). arXiv:1912.04100
18. Cipolloni, G., Erdős, L., Schröder, D.: Functional Central Limit Theorems for Wigner Matrices, preprint (2020). arXiv:2012.13218
19. Colin de Verdière, Y.: Ergodicité et fonctions propres du laplacien. Commun. Math. Phys. **102**, 497–502 (1985)
20. D'Alessio, L., Kafri, Y., Polkovnikov, A., Rigol, M.: From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics. Adv. Phys. **65**, 239–362 (2016)
21. Deutsch, J.: Eigenstate thermalization hypothesis. Rep. Prog. Phys. **81**, 082001 (2018)
22. Deutsch, J.: Quantum statistical mechanics in a closed system. Phys. Rev. A **43**, 2046–2049 (1991)
23. Eckhardt, B., Fishman, S., Keating, J., Agam, O., Main, J., Müller, K.: Approach to ergodicity in quantum wave functions. Phys. Rev. E **52**, 5893–5903 (1995)
24. Erdős, L., Knowles, A.: The Altshuler–Shklovskii formulas for random band matrices I: the unimodular case. Commun. Math. Phys. **333**, 1365–1416 (2015)
25. Erdős, L., Knowles, A., Yau, H.-T.: Averaging fluctuations in resolvents of random band matrices. Ann. Henri Poincareé **14**, 1837–1926 (2013)

26. Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: The local semicircle law for a general class of random matrices. Electron. J. Probab. **18**(59), 58 (2013)
27. Erdős, L., Krüger, T., Schröder, D.: Cusp universality for random matrices I: local law and the complex Hermitian case. Commun. Math. Phys. **378**, 1203–1278 (2020)
28. Erdős, L., Krüger, L., Schröder, D.: Random matrices with slow correlation decay. Forum Math. Sigma **7**, e8, 89 (2019)
29. Erdős, L., Schlein, B., Yau, H.-T.: Local semicircle law and complete delocalization for Wigner random matrices. Commun. Math. Phys. **287**, 641–655 (2009)
30. Erdős, L., Yau, H.-T., Yin, J.: Rigidity of eigenvalues of generalized Wigner matrices. Adv. Math. **229**, 1435–1515 (2012)
31. Erdős, P., Hajnal, A.: On chromatic number of graphs and set-systems. Acta Math. Acad. Sci. Hungar. **17**, 61–99 (1966)
32. Feingold, M., Peres, A.: Distribution of matrix elements of chaotic systems. Phys. Rev. A **34**, 591–595 (1986)
33. Hachem, W., Hardy, A., Najim, J.: Large complex correlated Wishart matrices: the Pearcey kernel and expansion at the hard edge. Electron. J. Probab. **21**(1), 36 (2016)
34. He, Y., Knowles, A.: Mesoscopic eigenvalue statistics of Wigner matrices. Ann. Appl. Probab. **27**, 1510–1550 (2017)
35. Hezari, H., Rivière, G.: Quantitative equidistribution properties of toral eigenfunctions. J. Spectr. Theory **7**, 471–485 (2017)
36. Holowinsky, R., Soundararajan, K.: Mass equidistribution for Hecke eigenforms. Ann. Math. **2**(172), 1517–1528 (2010)
37. Knowles, A., Yin, J.: Eigenvector distribution of Wigner matrices. Probab. Theory Relat. Fields **155**, 543–582 (2013)
38. Knowles, A., Yin, J.: The isotropic semicircle law and deformation of Wigner matrices. Commun. Pure Appl. Math. **66**, 1663–1750 (2013)
39. Kurlberg, P., Rudnick, Z.: On the distribution of matrix elements for the quantum cat map. Ann. Math. **2**(161), 489–507 (2005)
40. Le Masson, E., Sahlsten, T.: Quantum ergodicity and Benjamini–Schramm convergence of hyperbolic surfaces. Duke Math. J. **166**, 3425–3460 (2017)
41. Lick, D.R., White, A.T.: k-degenerate graphs. Can. J. Math. **22**, 1082–1096 (1970)
42. Lindenstrauss, E.: Invariant measures and arithmetic quantum unique ergodicity. Ann. Math. **2**(163), 165–219 (2006)
43. Luo, W.Z., Sarnak, P.: Quantum ergodicity of eigenfunctions on $\mathrm{PSL}_2(\mathbf{Z})\backslash\mathbf{H}^2$. Inst. Hautes Études Sci. Publ. Math. **207–237** (1995)
44. Marcinek, J., Yau, H.-T.: High dimensional normality of noisy eigenvectors. Commun. Math. Phys. (to appear). arXiv:2005.08425
45. Marklof, J., Rudnick, Z.: Quantum unique ergodicity for parabolic maps. Geom. Funct. Anal. **10**, 1554–1578 (2000)
46. Rudnick, Z., Sarnak, P.: The behaviour of eigenstates of arithmetic hyperbolic manifolds. Commun. Math. Phys. **161**, 195–213 (1994)
47. Schubert, R.: On the rate of quantum ergodicity for quantised maps. Ann. Henri Poincaré **9**, 1455–1477 (2008)
48. Schubert, R.: Upper bounds on the rate of quantum ergodicity. Ann. Henri Poincaré **7**, 1085–1098 (2006)
49. Šnirel'man, A.I.: Ergodic properties of eigenfunctions. Uspehi Mat. Nauk **29**, 181–182 (1974)
50. Soundararajan, K.: Quantum unique ergodicity for $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$. Ann. Math. **2**(172), 1529–1538 (2010)
51. Srednicki, M.: Chaos and quantum thermalization. Phys. Rev. E **50**, 888–901 (1994)
52. Tao, H.-T., Vu, H.-T.: Random matrices: universal properties of eigenvectors, Random Matrices Theory Appl. **1**, 1150001, 27 (2012)
53. Wigner, E.P.: Characteristic vectors of bordered matrices with infinite dimensions. Ann. Math. **2**(62), 548–564 (1955)
54. Zelditch, S.: "Recent developments in mathematical quantum chaos", Current developments in mathematics, 2009, pp. 115–204. Int. Press, Somerville, MA (2010)
55. Zelditch, S.: On the rate of quantum ergodicity. I. Upper bounds. Commun. Math. Phys. **160**, 81–92 (1994)
56. Zelditch, S.: Uniform distribution of eigenfunctions on compact hyperbolic surfaces. Duke Math. J. **55**, 919–941 (1987)