

RANK-UNIFORM LOCAL LAW FOR WIGNER MATRICES

GIORGIO CIPOLLONI

Princeton Center for Theoretical Science, Princeton University, Princeton, NJ 08544, USA

LÁSZLÓ ERDŐS#

IST Austria, Am Campus 1, 3400 Klosterneuburg, Austria

DOMINIK SCHRÖDER*

Institute for Theoretical Studies, ETH Zurich, Clausiusstr. 47, 8092 Zurich, Switzerland

ABSTRACT. We prove a general local law for Wigner matrices which optimally handles observables of arbitrary rank and thus it unifies the well-known averaged and isotropic local laws. As an application, we prove a central limit theorem in quantum unique ergodicity (QUE), i.e. we show that the quadratic forms of a general deterministic matrix A on the bulk eigenvectors of a Wigner matrix have approximately Gaussian fluctuation. For the bulk spectrum, we thus generalize our previous result [18] valid for test matrices A of large rank as well as the result of Benigni and Lopatto [10] valid for specific small rank observables.

1. INTRODUCTION

Wigner random matrices are $N \times N$ random Hermitian matrices $W = W^*$ with centred, independent, identically distributed (*i.i.d.*) entries up to the symmetry constraint $w_{ab} = \overline{w_{ba}}$. Originally introduced by E. Wigner [55] to study spectral gaps of large atomic nuclei, Wigner matrices have become the most studied random matrix ensemble since they represent the simplest example of a fully chaotic quantum Hamiltonian beyond the explicitly computable Gaussian case.

A key conceptual feature of Wigner matrices, as well as a fundamental technical tool to study them, is the fact that their resolvent $G(z) := (W - z)^{-1}$, with a spectral parameter z away from the real axis, becomes asymptotically deterministic in the large N limit. The limit is the scalar matrix $m(z) \cdot I$, where $m(z) = \frac{1}{2}(-z + \sqrt{z^2 - 4})$ is the Stieltjes transform of the *Wigner semicircular density*, $\rho_{\text{sc}}(x) = \frac{1}{2\pi}\sqrt{4 - x^2}$, which is the $N \rightarrow \infty$ limit of the empirical density of the eigenvalues of W under the standard normalization $\mathbf{E} |w_{ab}|^2 = 1/N$. The *local law on optimal scale* asserts that this limit holds even when z is very close to the real axis as long as $|\Im z| \gg 1/N$. Noticing that the imaginary part of the Stieltjes transform resolves the spectral measure on a scale comparable with $|\Im z|$, this condition is necessary for a deterministic limit to hold since on scales of order $1/N$, comparable with the typical eigenvalue spacing, the resolvent is genuinely fluctuating.

The limit $G(z) \rightarrow m(z) \cdot I$ holds in a natural appropriate topology, namely when tested against deterministic $N \times N$ matrices A , i.e. in the form $\langle G(z)A \rangle \rightarrow m(z)\langle A \rangle$, where $\langle \cdot \rangle := \frac{1}{N} \text{Tr}(\cdot)$ denotes the normalized trace. It is essential that the test matrix A is deterministic, no analogous limit can hold if A were random and strongly correlated with W , e.g. if A were a spectral projection of W .

E-mail addresses: gc4233@princeton.edu, lerdos@ist.ac.at, dschroeder@ethz.ch

Date: September 8, 2023.

2010 Mathematics Subject Classification. 60B20, 15B52.

Key words and phrases. Local law, quantum unique ergodicity, eigenvector overlap.

Supported by ERC Advanced Grant “RMTBeyond” No. 101020331.

* Supported by Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zürich Foundation.

The first optimal local law for Wigner matrices was proven for $A = I$ in [29], see also [52, 53, 14, 34] extended later to more general matrices A in the form that¹

$$|\langle (G(z) - m(z))A \rangle| \leq \frac{N^\xi \|A\|}{N\eta}, \quad \eta := |\Im z|, \quad (1.1)$$

holds with very high probability for any fixed $\xi > 0$ if N is sufficiently large. By *optimality* in this paper we always mean up to a tolerance factor N^ξ , this is a natural byproduct of our method yielding very high probability estimates under the customary moment condition, see (2.2) later². The estimate (1.1) is called the *average local law* and it controls the error in terms of the standard Euclidean matrix norm $\|A\|$ of A . It holds for arbitrary deterministic matrices A and it is also optimal in this generality with respect to the dependence on A : e.g. for $A = I$ the trace $\langle G - m \rangle$ is approximately complex Gaussian with standard deviation [35]

$$\sqrt{\mathbf{E}|\langle G - m \rangle|^2} \approx \frac{|m'(z)| |\Im m(z)|}{N\eta |m(z)|^2} \sim \frac{1}{N\eta}, \quad \eta = |\Im z| = N^{-\alpha}, \quad \alpha \in [0, 1),$$

but (1.1) is far from being optimal when applied to matrices with small rank. Rank one matrices, $A = \mathbf{y}\mathbf{x}^*$, are especially important since they give the asymptotic behaviour of resolvent matrix elements $G_{\mathbf{x}\mathbf{y}} := \langle \mathbf{x}, G\mathbf{y} \rangle$. For such special test matrices, a separate *isotropic local law* of the optimal form

$$|\langle \mathbf{x}, (G(z) - m(z))\mathbf{y} \rangle| \leq \frac{N^\xi \rho^{1/2} \|\mathbf{x}\| \|\mathbf{y}\|}{\sqrt{N\eta}}, \quad \eta = |\Im z|, \quad \rho := |\Im m(z)|, \quad (1.2)$$

has been proven; see [30] for special coordinate vectors and later [40] for general vectors \mathbf{x}, \mathbf{y} , as well as [42, 38, 36, 28, 36] for more general ensembles. Note that a direct application of (1.1) to $A = \mathbf{y}\mathbf{x}^*$ would give a bound of order $1/\eta$ instead of the optimal $1/\sqrt{N\eta}$ in (1.2) which is an unacceptable overestimate in the most interesting small η -regime. More generally, the average local law (1.1) performs badly when A has effectively small rank, i.e. if only a few eigenvalues of A are comparable with the norm $\|A\|$ and most other eigenvalues are much smaller or even zero.

Quite recently we found that the average local law (1.1) is also suboptimal for another class of test matrices A , namely for traceless matrices. In [16] we proved that

$$|\langle (G(z) - m(z))A \rangle| = |\langle G(z)A \rangle| \leq \frac{N^\xi \|A\|}{N\sqrt{\eta}}, \quad \eta = |\Im z|, \quad (1.3)$$

for any deterministic matrix A with $\langle A \rangle = 0$, i.e. traceless observables yield an additional $\sqrt{\eta}$ improvement in the error. The optimality of this bound for general traceless A was demonstrated by identifying the nontrivial Gaussian fluctuation of $N\sqrt{\eta}\langle G(z)A \rangle$ in [17].

While the mechanism behind the suboptimality of (1.1) for small rank and traceless A is very different, their common core is that estimating the size of A simply by the Euclidean norm is too crude for several important classes of A . In this paper we present a local law which unifies all three local laws (1.1), (1.2) and (1.3) by identifying the appropriate way to measure the size of A . Our main result (Theorem 2.2, $k = 1$ case) shows that

$$|\langle (G(z) - m(z))A \rangle| \leq \frac{N^\xi}{N\eta} |\langle A \rangle| + \frac{N^\xi \rho^{1/2} \langle \hat{A}^2 \rangle^{1/2}}{N\sqrt{\eta}}, \quad \eta = |\Im z|, \quad \rho = |\Im m(z)|, \quad (1.4)$$

holds with very high probability, where $\hat{A} := A - \langle A \rangle$ is the traceless part of A . It is straightforward to check that (1.4) implies (1.1), (1.2) and (1.3), moreover, it optimally interpolates between full rank and rank-one matrices A , hence we call (1.4) the *rank-uniform local law* for Wigner matrices. Note that an optimal local law for matrices of intermediate rank was previously unknown; indeed the local laws (1.1)–(1.2) are optimal only for essentially full rank and essentially finite rank observables, respectively. The proof of the optimality of (1.4) follows from identifying the scale of the Gaussian fluctuation of its left hand side. Its standard deviation for traceless A is

$$\sqrt{\mathbf{E}|\langle GA \rangle|^2} \approx \frac{|m| \sqrt{\Im m} \langle AA^* \rangle^{1/2}}{N\sqrt{\eta}} \sim \frac{\rho^{1/2} \langle AA^* \rangle^{1/2}}{N\sqrt{\eta}}; \quad (1.5)$$

this relation was established for matrices with bounded norm $\|A\| \lesssim 1$ in [44, 17].

¹Traditional local laws for Wigner matrices did not consider a general test matrix A ; this concept appeared later in connection with more general random matrix ensembles, see e.g. [28].

²We remark that the N^ξ tolerance factor can be improved to logarithmic factors under slightly different conditions, see e.g. [14, 33, 34].

The key observation that traceless A substantially improves the error term (1.3) compared with (1.1) was the conceptually new input behind our recent proof of the *Eigenstate Thermalisation Hypothesis* in [16] followed by the proof of the normal fluctuation in the quantum unique ergodicity for Wigner matrices in [19]. Both results concern the behavior of the *eigenvector overlaps*, i.e. quantities of the form $\langle \mathbf{u}_i, A\mathbf{u}_j \rangle$, where $\{\mathbf{u}_i\}_{i=1}^N$ are the normalized eigenvectors of W . The former result stated that

$$|\langle \mathbf{u}_i, \mathring{A}\mathbf{u}_j \rangle| = |\langle \mathbf{u}_i, A\mathbf{u}_j \rangle - \delta_{ij}\langle A \rangle| \leq \frac{N^\xi \|\mathring{A}\|}{\sqrt{N}} \quad (1.6)$$

holds with very high probability for any i, j and for any fixed $\xi > 0$. The latter result established the optimality of (1.6) for $i = j$ by showing that $\sqrt{N}\langle \mathbf{u}_i, \mathring{A}\mathbf{u}_i \rangle$ is asymptotically Gaussian when the corresponding eigenvalue lies in the bulk of the spectrum. The variance of $\sqrt{N}\langle \mathbf{u}_i, \mathring{A}\mathbf{u}_i \rangle$ was shown to be $\langle |\mathring{A}|^2 \rangle$ in [19] but we needed to assume that $\langle |\mathring{A}|^2 \rangle \geq c\|\mathring{A}\|^2$ with some fixed positive constant c , i.e. that the rank of \mathring{A} was essentially macroscopic.

As the second main result of the current paper, we now remove this unnatural condition and show the standard Gaussianity of the normalized overlaps $[N/\langle |\mathring{A}|^2 \rangle]^{1/2}\langle \mathbf{u}_i, \mathring{A}\mathbf{u}_i \rangle$ for bulk indices under the optimal and natural condition that $\langle |\mathring{A}|^2 \rangle \gg N^{-1}\|\mathring{A}\|^2$, which essentially ensures that \mathring{A} is not of finite rank. This improvement is possible thanks to improving the dependence of the error terms in the local laws from $\|\mathring{A}\|$ to $\langle |\mathring{A}|^2 \rangle^{1/2}$ similarly to the improvement in (1.4) over (1.3). We will also need a multi-resolvent version of this improvement since off-diagonal overlaps $\langle \mathbf{u}_i, A\mathbf{u}_j \rangle$ are not accessible via single-resolvent local laws; in fact $|\langle \mathbf{u}_i, A\mathbf{u}_j \rangle|^2$ is intimately related to $\langle \Im G(z) A \Im G(z') A^* \rangle$ with two different spectral parameters z, z' , analysed in Theorem 2.2. As a corollary we will show the following improvement of (1.6) (see Theorem 2.6)

$$|\langle \mathbf{u}_i, A\mathbf{u}_j \rangle - \delta_{ij}\langle A \rangle| \leq \frac{N^\xi \langle |\mathring{A}|^2 \rangle^{1/2}}{\sqrt{N}} \quad (1.7)$$

for the bulk indices. The analysis at the edge is deferred to later work.

Gaussian fluctuation of diagonal overlaps with a special low rank observable has been proven earlier. Right after [19] was posted on the arXiv, Benigni and Lopatto in an independent work [7] proved the standard Gaussian fluctuation of $[N/|S|]^{1/2}[\sum_{a \in S} |u_i(a)|^2 - |S|/N]$ whenever $1 \ll |S| \ll N$, i.e. they considered $\langle \mathbf{u}_i, \mathring{A}\mathbf{u}_i \rangle$ for the special case when the matrix A is the projection on coordinates from the set S . Their result also holds at the edge. The condition $|S| \ll N$ requires A to have small rank, hence it is complementary to our old condition $\langle |\mathring{A}|^2 \rangle \geq c\|\mathring{A}\|^2$ from [19] for projection operators. The natural condition $|S| \gg 1$ is the special case of our new improved condition $\langle |\mathring{A}|^2 \rangle \gg N^{-1}\|\mathring{A}\|^2$. In particular, our new result covers [7] as a special case in the bulk and it gives a uniform treatment of all observables in full generality.

The methods of [19] and [7] are very different albeit they both rely on the *Dyson Brownian motion (DBM)*, complemented by fairly standard *Green function comparison (GFT)* techniques. Benigni and Lopatto focused on the joint Gaussianity of the individual eigenvector entries $u_i(a)$ (or more generally, linear functionals $\langle q_\alpha, \mathbf{u}_i \rangle$ with deterministic unit vectors q_α) in the spirit of the previous quantum ergodicity results by Bourgade and Yau [11] operating with the so-called *eigenvector moment flow* from [11] complemented by its "fermionic" version by Benigni [9]. This approach becomes less effective when more entries need to be controlled simultaneously and it seems to have a natural limitation at $|S| \ll N$.

Our method viewed the eigenvector overlap $\langle \mathbf{u}_i, \mathring{A}\mathbf{u}_i \rangle$ and its offdiagonal version $\langle \mathbf{u}_i, \mathring{A}\mathbf{u}_j \rangle$ as one unit without translating it into a sum of rank one projections $\langle \mathbf{u}_i, q_\alpha \rangle \langle q_\alpha, \mathbf{u}_j \rangle$ via the spectral decomposition of \mathring{A} . The corresponding flow for overlaps with arbitrary A , called the *stochastic eigenstate equation*, was introduced by Bourgade, Yau and Yin in [13] (even though they applied it to the special case when A is a projection, their formalism is general). The analysis of this new flow is more involved than the eigenvector moment flow since it operates on a geometrically more complicated higher dimensional space. However, the substantial part of this analysis has been done by Marcinek and Yau [45] and we heavily relied on their work in our proof [19].

We close this introduction by commenting on our methods. The main novelty of the current paper is the proof of the rank-uniform local laws involving the Hilbert-Schmidt norm $\langle |\mathring{A}|^2 \rangle^{1/2}$ instead of the Euclidean matrix norm $\|\mathring{A}\|$. This is done in Section 3 and it will directly imply the improved overlap estimate (1.7). Once this estimate is available, both the DBM and the GFT parts of the proof in the current paper are essentially the same as in [19], hence we will not give all details, we only point out the differences. While this can be done very concisely for the GFT in Appendix B, for the DBM part we need to recall large part of the necessary setup in Section 4 for the convenience of the reader.

As to our main result, the general scheme to prove single resolvent local laws has been well established and traditionally it consisted of two parts: (i) the derivation of an approximate self-consistent equation that $G - m$ satisfies and

(ii) estimating the key fluctuation term in this equation. The proofs of the multi-resolvent local laws follow the same scheme, but the self-consistent equation is considerably more complicated and its stability is more delicate, see e.g. [16, 21] where general multi-resolvent local laws were proven. The main complication lies in part (ii) where a high moment estimate is needed for the fluctuation term. The corresponding cumulant expansion results in many terms which have typically been organized and estimated by a graphical Feynman diagrammatic scheme. A reasonably manageable power counting handles all diagrams for the purpose of proving (1.1) and (1.2). However, in the multi-resolvent setup or if we aim at some improvement, the diagrammatic approach becomes very involved since the right number of additional improvement factors needs to be gained from every single graph. This was the case many times before: (i) when a small factor (so-called “sigma-cell”) was extracted at the cusp [27], (ii) when we proved that the correlation between the resolvents of the Hermitization of an i.i.d. random matrix shifted by two different spectral parameters z_1, z_2 decays in $1/|z_1 - z_2|$ [15], and (iii) more recently when the gain of order $\sqrt{\eta}$ due to the traceless A in (1.3) was obtained in [16].

Extracting $\langle |\hat{A}|^2 \rangle^{1/2}$ instead of $\|A\|$, especially in the multi-resolvent case, seems even more involved in this way since estimating A simply by its norm appears everywhere in any diagrammatic expansion. However, very recently in [20] we introduced a new method of a *system of master inequalities* that circumvents the full diagrammatic expansion. The power of this method was demonstrated by fully extracting the maximal $\sqrt{\eta}$ -gain from traceless A even in the multi-resolvent setup; the same result seemed out of reach with the diagrammatic method used for the single-resolvent setup in [16]. In the current paper we extend this technique to obtain the optimal control in terms of $\langle |\hat{A}|^2 \rangle^{1/2}$ instead of $\|\hat{A}\|$ for single resolvent local laws. However, the master inequalities in this paper are different from the ones in [20]; in fact they are much tighter, since the effect we extract now is much more delicate. We also obtain a similar optimal control for the multi-resolvent local laws needed to prove the Gaussianity of the bulk eigenvector overlaps under the optimal condition on A .

Notations and conventions. We denote vectors by bold-faced lower case Roman letters $\mathbf{x}, \mathbf{y} \in \mathbf{C}^N$, for some $N \in \mathbf{N}$. Vector and matrix norms, $\|\mathbf{x}\|$ and $\|A\|$, indicate the usual Euclidean norm and the corresponding induced matrix norm. For any $N \times N$ matrix A we use the notation $\langle A \rangle := N^{-1} \text{Tr } A$ to denote the normalized trace of A . Moreover, for vectors $\mathbf{x}, \mathbf{y} \in \mathbf{C}^N$ and matrices $A \in \mathbf{C}^{N \times N}$ we define

$$\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^N \bar{x}_i y_i, \quad A_{\mathbf{x}\mathbf{y}} := \langle \mathbf{x}, A\mathbf{y} \rangle.$$

We will use the concept of “with very high probability” meaning that for any fixed $D > 0$ the probability of an N -dependent event is bigger than $1 - N^{-D}$ if $N \geq N_0(D)$. We introduce the notion of *stochastic domination* (see e.g. [26]): given two families of non-negative random variables

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

indexed by N (and possibly some parameter u in some parameter space $U^{(N)}$), we say that X is stochastically dominated by Y , if for all $\xi, D > 0$ we have

$$\sup_{u \in U^{(N)}} \mathbf{P} \left[X^{(N)}(u) > N^\xi Y^{(N)}(u) \right] \leq N^{-D} \quad (1.8)$$

for large enough $N \geq N_0(\xi, D)$. In this case we use the notation $X \prec Y$ or $X = \mathcal{O}_\prec(Y)$. We also use the convention that $\xi > 0$ denotes an arbitrary small constant which is independent of N .

Finally, 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 the moment condition, see (2.2) later.

2. MAIN RESULTS

Assumption 1. We say that $W = W^* \in \mathbf{C}^{N \times N}$ is a real symmetric/complex hermitian Wigner-matrix if the entries $(w_{ab})_{a \leq b}$ in the upper triangular part are independent and satisfy

$$w_{ab} \stackrel{d}{=} N^{-1/2} \times \begin{cases} \chi_{\text{od}}, & a \neq b \\ \chi_{\text{d}}, & a = b, \end{cases} \quad (2.1)$$

for some real random variable χ_d and some real/complex random variable χ_{od} of mean $\mathbf{E} \chi_d = \mathbf{E} \chi_{od} = 0$ and variances $\mathbf{E} |\chi_{od}|^2 = 1$, $\mathbf{E} \chi_{od}^2 = 0$, $\mathbf{E} \chi_d^2 = 1$ in the complex, and $\mathbf{E} |\chi_{od}|^2 = \mathbf{E} \chi_{od}^2 = 1$, $\mathbf{E} \chi_d^2 = 2$ in the real case³. We furthermore assume that for every $n \geq 3$

$$\mathbf{E} |\chi_d|^n + \mathbf{E} |\chi_{od}|^n \leq C_n \quad (2.2)$$

for some constant C_n , in particular all higher order cumulants $\kappa_n^d, \kappa_n^{od}$ of χ_d, χ_{od} are finite for any n .

Our results hold for both symmetry classes, but for definiteness we prove the main results in the real case, the changes for the complex case being minimal.

For a spectral parameter $z \in \mathbf{C}$ with $\eta := |\Im z| \gg N^{-1}$ the resolvent $G = G(z) = (W - z)^{-1}$ of a $N \times N$ Wigner matrix W is well approximated by a constant multiple $m \cdot I$ of the identity matrix, where $m = m(z)$ is the Stieltjes transform of the semicircular distribution $\sqrt{4 - x^2}/(2\pi)$ and satisfies the equation

$$-\frac{1}{m} = m + z, \quad \Im m \Im z > 0. \quad (2.3)$$

We set $\rho(z) := |\Im m(z)|$, which approximates the density of eigenvalues near $\Re z$ in a window of size η .

We first recall the classical local law for Wigner matrices, both in its tracial and isotropic form [29, 31, 40, 36]:

Theorem 2.1. *Fix any $\epsilon > 0$, then it holds that*

$$|\langle G - m \rangle| \prec \frac{1}{N\eta}, \quad |\langle \mathbf{x}, (G - m)\mathbf{y} \rangle| \prec \|\mathbf{x}\| \|\mathbf{y}\| \left(\sqrt{\frac{\rho}{N\eta}} + \frac{1}{N\eta} \right) \quad (2.4)$$

uniformly in any deterministic vectors \mathbf{x}, \mathbf{y} and spectral parameter z with $\eta = |\Im z| \geq N^{-1+\epsilon}$ and $\Re z \in \mathbf{R}$, where $\rho = |\Im m(z)|$.

Our main result is the following optimal multi-resolvent local law with Hilbert-Schmidt norm error terms. Compared to Theorem 2.1 we formulate the bound only in averaged sense since, due to the Hilbert-Schmidt norm in the error term, the isotropic bound is a special case with one of the traceless matrices being a centred rank-one matrix, see Corollary 2.4.

Theorem 2.2 (Averaged multi-resolvent local law). *Fix $\epsilon > 0$, let $k \geq 1$ and consider $z_1, \dots, z_k \in \mathbf{C}$ with $N\eta\rho \geq N^\epsilon$, for $\eta := \min_i |\Im z_i|$, $\rho := \max_i |\Im m(z_i)|$, $d := \min_i \text{dist}(z_i, [-2, 2])$, and let A_1, \dots, A_k be deterministic traceless matrices, $\langle A_i \rangle = 0$. Set $G_i := G(z_i)$ and $m_i := m(z_i)$ for all $i \leq k$. Then we have the local law on optimal scale⁴*

$$|\langle G_1 A_1 \cdots G_k A_k - m_1 \cdots m_k A_1 \cdots A_k \rangle| \prec N^{k/2-1} \prod_{i=1}^k \langle |A_i|^2 \rangle^{1/2} \times \begin{cases} \sqrt{\frac{\rho}{N\eta}}, & d < 10 \\ \frac{1}{\sqrt{N}d^{k+1}}, & d \geq 10. \end{cases} \quad (2.5)$$

Remark 2.3. *We also obtain generalisations of Theorem 2.2 where each G may be replaced by a product of G 's and $|G|$'s, see Lemma 3.1 later.*

Due to the Hilbert-Schmidt sense of the error term we obtain an isotropic variant of Theorem 2.2 as an immediate corollary by choosing $A_k = N\mathbf{y}\mathbf{x}^* - \langle \mathbf{x}, \mathbf{y} \rangle$ in (2.5).

Corollary 2.4 (Isotropic local law). *Under the setup and conditions of Theorem 2.2, for any vectors \mathbf{x}, \mathbf{y} it holds that*

$$|\langle \mathbf{x}, (G_1 A_1 \cdots A_{k-1} G_k - m_1 \cdots m_k A_1 \cdots A_{k-1}) \mathbf{y} \rangle| \prec \|\mathbf{x}\| \|\mathbf{y}\| N^{\frac{k-1}{2}} \prod_{i=1}^{k-1} \langle |A_i|^2 \rangle^{1/2} \times \begin{cases} \sqrt{\frac{\rho}{N\eta}}, & d < 10 \\ \frac{1}{\sqrt{N}d^{k+1}}, & d \geq 10. \end{cases} \quad (2.6)$$

We now compare Theorem 2.2 to the previous result [20, Theorem 2.5] where an error term $N^{-1}\eta^{-k/2} \prod_i \|A_i\|$ was proven for (2.5). For clarity we focus on the really interesting $d < 10$ regime.

Remark 2.5. *For $k = 1$ our new estimate for traceless A :*

$$|\langle (G - m)A \rangle| = |\langle GA \rangle| \prec \frac{\sqrt{\rho}}{N\sqrt{\eta}} \langle |A|^2 \rangle^{1/2}, \quad (2.7)$$

³We assumed that $\sigma := \mathbf{E} \chi_{od}^2 = 0$, $\mathbf{E} \chi_d^2 = 1$ in the complex case, and that $\mathbf{E} \chi_d^2 = 2$ in the real case only for notational simplicity. All the results presented below hold under the more general assumption $|\sigma| < 1$ and general variance for diagonal entries. The necessary modifications in the proofs are straightforward and will be omitted.

⁴The constant 10 is arbitrary and can be replaced by any positive constant.

is strictly better than the one in [20, Theorem 2.5], since $\langle |A|^2 \rangle \leq \|A\|^2$ always holds, but $\langle |A|^2 \rangle$ can be much smaller than $\|A\|^2$ for small rank A . In addition, (2.7) features an additional factor $\sqrt{\rho} \lesssim 1$ that is considerably smaller than 1 near the spectral edges.

For larger $k \geq 2$ the relationship depends on the relative size of the Hilbert-Schmidt and operator norm of the A_i 's, as well as on the size of η . We recall [48] that the numerical rank of A is defined as $r(A) := N \langle |A|^2 \rangle / \|A\|^2 \leq \text{rank}(A)$ and say that A is α -mesoscopic for some $\alpha \in [0, 1]$ if $r(A) = N^\alpha$. If for some $k \geq 2$ all A_i are α -mesoscopic, then Theorem 2.2 improves upon [20, Theorem 2.5] whenever $\eta \ll N^{(1-\alpha k)/(k-1)}$.

Local laws on optimal scales can give certain information on eigenvectors as well. Let $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$ denote the eigenvalues and $\{\mathbf{u}_i\}_{i=1}^N$ the corresponding orthonormal eigenvectors of W . Already the single-resolvent isotropic local law (2.4) implies the *eigenvector delocalization*, i.e. that $\|\mathbf{u}_i\|_\infty \prec N^{-1/2}$. More generally⁵ $|\langle \mathbf{x}, \mathbf{u}_i \rangle| \prec N^{-1/2} \|\mathbf{x}\|$, i.e. eigenvectors behave as completely random unit vectors in the sense of considering their rank-1 projections onto any deterministic vector \mathbf{x} . This concept can be greatly extended to arbitrary deterministic observable matrix A leading to the following results motivated both by thermalisation ideas from physics [24, 32, 25, 23] as well as by *Quantum (Unique) Ergodicity (QUE)* in mathematics [50, 57, 22, 43, 49, 56, 51, 46, 3, 2, 4, 5].

Theorem 2.6 (Eigenstate Thermalization Hypothesis). *Let W be a Wigner matrix satisfying Assumption 1 and let $\delta > 0$. Then for any deterministic matrix A and any bulk indices $i, j \in [\delta N, (1 - \delta)N]$ it holds that*

$$|\langle \mathbf{u}_i, A\mathbf{u}_j \rangle - \delta_{ij} \langle A \rangle| \prec \frac{\langle \mathring{A}^2 \rangle^{1/2}}{N^{1/2}}, \quad (2.8)$$

where $\mathring{A} := A - \langle A \rangle$ is the traceless part of A .

Remark 2.7.

- (1) The result (2.8) was established in [16] with $\langle \mathring{A}\mathring{A}^* \rangle^{1/2}$ replaced by $\|\mathring{A}\|$ uniformly in the spectrum (i.e. also for edge indices).
- (2) For rank-1 matrices $A = \mathbf{x}\mathbf{x}^*$ the bound (2.8) immediately implies the complete delocalisation of eigenvectors in the form $|\langle \mathbf{x}, \mathbf{u}_i \rangle| \prec N^{-1/2} \|\mathbf{x}\|$.

Theorem 2.6 directly follows from the bound

$$\max_{i, j \in [\delta N, (1-\delta)N]} N \left| \langle \mathbf{u}_i, \mathring{A}\mathbf{u}_j \rangle \right|^2 \leq C_\delta (N\eta)^2 \max_{E, E' \in [-2+\epsilon, 2-\epsilon]} \langle \Im G(E + i\eta) \mathring{A} \Im G(E' + i\eta) \mathring{A}^* \rangle$$

that is obtained by the spectral decomposition of both resolvents and the well-known eigenvalue rigidity, with some explicit δ -dependent constants C_δ and $\epsilon = \epsilon(\delta) > 0$ (see [16, Lemma 1] for more details). The right hand side can be directly estimated using (2.5) and finally choosing $\eta = N^{-1+\xi}$ for any small $\xi > 0$ gives (2.8) and thus proves Theorem 2.6.

The next question is to establish a central limit theorem for the diagonal overlap in (2.8).

Theorem 2.8 (Central Limit Theorem in the QUE). *Let W be a real symmetric ($\beta = 1$) or complex Hermitian ($\beta = 2$) Wigner matrix satisfying Assumption 1. Fix small $\delta, \delta' > 0$ and let $A = A^*$ be a deterministic $N \times N$ matrix with $N^{-1+\delta'} \|\mathring{A}\|^2 \lesssim \langle \mathring{A}^2 \rangle \lesssim 1$. In the real symmetric case we also assume that $A \in \mathbf{R}^{N \times N}$ is real. Then for any bulk index $i \in [\delta N, (1 - \delta)N]$ we have a central limit theorem*

$$\sqrt{\frac{\beta N}{2 \langle \mathring{A}^2 \rangle}} [\langle \mathbf{u}_i, A\mathbf{u}_i \rangle - \langle A \rangle] \Rightarrow \mathcal{N}, \quad \text{as } N \rightarrow \infty \quad (2.9)$$

with \mathcal{N} being a standard real Gaussian random variable. Moreover, for any moment the speed of convergence is explicit (see (B.5)).

We require that $\langle \mathring{A}^2 \rangle \gtrsim N^{-1+\delta'} \|\mathring{A}\|^2$ in order to ensure that the spectral distribution of \mathring{A} is not concentrated to a finite number eigenvalues, i.e. that \mathring{A} has effective rank $\gg 1$. Indeed, the statement in (2.9) does not hold for finite rank A 's, e.g. if $A = \mathring{A} = |\mathbf{e}_x\rangle\langle \mathbf{e}_x| - |\mathbf{e}_y\rangle\langle \mathbf{e}_y|$, for some $x \neq y \in [N]$, then $\langle \mathbf{u}_i, \mathring{A}\mathbf{u}_i \rangle = |\mathbf{u}_i(x)|^2 - |\mathbf{u}_i(y)|^2$, which is the difference of two asymptotically independent χ^2 -distributed random variables (e.g. see [11, Theorem 1.2]). More generally, the joint distribution of finitely many eigenvector overlaps has been identified in [11, 45, 1, 12] for various related ensembles.

⁵Under stronger decay conditions on the distribution of χ_d, χ_{od} even the optimal bound $\|\mathbf{u}_i\|_\infty \leq C \sqrt{\log N/N}$ for the bulk and $\|\mathbf{u}_i\|_\infty \leq C \log N/\sqrt{N}$ for the edge eigenvectors has been proven [54], see also [47] for a comprehensive summary of related results. Very recently even the optimal constant C has been identified [8].

3. PROOF OF THEOREM 2.2

Within this section we prove Theorem 2.2 in the critical $d < 10$ regime. The $d \geq 10$ regime is handled similarly but the estimates are much simpler; the necessary modifications are outlined in Appendix A.

In the subsequent proof we will often assume that a priori bounds, with some control parameters $\psi_K^{\text{av/iso}} \geq 1$, of the form

$$\Psi_0^{\text{av}} = \Psi_0^{\text{av}}(z_1) := N\eta |\langle G_1 - m_1 \rangle| \prec \psi_0^{\text{av}} \quad (3.1)$$

$$\Psi_K^{\text{av}} = \Psi_K^{\text{av}}(\mathbf{A}, \mathbf{z}) := \frac{N^{(3-K)/2} \eta^{1/2}}{\rho^{1/2} \prod_i \langle |A_i|^2 \rangle^{1/2}} |\langle [G_1 A_1 \cdots G_K A_K - m_1 \cdots m_K A_1 \cdots A_K] \rangle| \prec \psi_K^{\text{av}}, \quad K \geq 1, \quad (3.2)$$

$$\Psi_K^{\text{iso}} = \Psi_K^{\text{iso}}(\mathbf{x}, \mathbf{y}, \mathbf{A}, \mathbf{z}) := \frac{N^{(1-K)/2} \eta^{1/2} \rho^{-1/2}}{\|\mathbf{x}\| \|\mathbf{y}\| \prod_i \langle |A_i|^2 \rangle^{1/2}} |\langle \mathbf{x}, [G_1 A_1 \cdots G_{K+1} - m_1 \cdots m_{K+1} A_1 \cdots A_K] \mathbf{y} \rangle| \prec \psi_K^{\text{iso}} \quad (3.3)$$

for certain indices $K \geq 0$ have been established uniformly in deterministic traceless matrices $\mathbf{A} = (A_1, \dots, A_K)$, deterministic vectors \mathbf{x}, \mathbf{y} , and spectral parameters $\mathbf{z} = (z_1, \dots, z_K)$ with⁶ $N\eta\rho \geq N^\epsilon$. We stress that we do *not* assume the estimates to be uniform in K . Note that ψ_0^{av} is defined somewhat differently from ψ_K^{av} , $K \geq 1$, but the definition of ψ_K^{iso} is the same for all $K \geq 0$. For intuition, the reader should think of the control parameters as essentially order one quantities, in fact our main goal will be to prove this fact. Note that by Theorem 2.1 we may set $\psi_0^{\text{av/iso}} = 1$.

As a first step we observe that (3.1), (3.2) and (3.3) immediately imply estimates on more general averaged resolvent chains and isotropic variants.

Lemma 3.1. (i) Assuming (3.1) and (3.3) for $K = 0$ holds uniformly in z_1 , then for any z_1, \dots, z_l with $N\eta\rho \geq N^\epsilon$ it holds that

$$\begin{aligned} |\langle G_1 G_2 \cdots G_l - m[z_1, \dots, z_l] \rangle| &\prec \frac{\psi_0^{\text{av}}}{N\eta^l}, \\ |\langle \mathbf{x}, (G_1 G_2 \cdots G_l - m[z_1, \dots, z_l]) \mathbf{y} \rangle| &\prec \frac{\|\mathbf{x}\| \|\mathbf{y}\| \psi_0^{\text{iso}}}{\eta^{l-1}} \sqrt{\frac{\rho}{N\eta}} \end{aligned} \quad (3.4)$$

where $m[z_1, \dots, z_l]$ stands for the l -th divided difference of the function $m(z)$ from (2.3), explicitly

$$m[z_1, \dots, z_l] = \int_{-2}^2 \frac{\sqrt{4-x^2}}{2\pi} \prod_{i=1}^l \frac{1}{x-z_i} dx. \quad (3.5)$$

(ii) Assuming for some $k \geq 1$ the estimates (3.2) and (3.3) for $K = k$ have been established uniformly, then for $\mathcal{G}_j := G_{j,1} \cdots G_{j,l_j}$ with $G_{j,i} \in \{G(z_{j,i}), |G(z_{j,i})|\}$, traceless matrices A_i and $\eta := \min_{j,i} |\Im z_{j,i}|$, $\rho := \max_{j,i} \rho(z_{j,i})$ it holds that

$$\begin{aligned} \left| \langle \mathcal{G}_1 A_1 \cdots \mathcal{G}_k A_k - m^{(1)} \cdots m^{(k)} A_1 \cdots A_k \rangle \right| &\prec \psi_k^{\text{av}} N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \prod_j \frac{\langle |A_j|^2 \rangle^{1/2}}{\eta^{l_j-1}}, \\ \left| \langle \mathbf{x}, [\mathcal{G}_1 A_1 \cdots A_k \mathcal{G}_{k+1} - m^{(1)} \cdots m^{(k+1)} A_1 \cdots A_k] \mathbf{y} \rangle \right| &\prec \psi_k^{\text{iso}} \|\mathbf{x}\| \|\mathbf{y}\| N^{k/2} \sqrt{\frac{\rho}{N\eta}} \prod_j \frac{\langle |A_j|^2 \rangle^{1/2}}{\eta^{l_j-1}}, \end{aligned} \quad (3.6)$$

where

$$m^{(j)} := \int_{-2}^2 \frac{\sqrt{4-x^2}}{2\pi} \prod_i g_{j,i}(x) dx \quad (3.7)$$

and $g_{j,i}(x) = (x - z_{j,i})^{-1}$ or $g_{j,i}(x) = |x - z_{j,i}|^{-1}$, depending on whether $G_{j,i} = G(z_{j,i})$ or $G_{j,i} = |G(z_{j,i})|$.

Proof. Analogous to [20, Lemma 3.2]. \square

The main result of this section is the following hierarchy of master inequalities.

⁶In some estimates the domain of uniformity for the spectral parameters may shrink a bit. The pedantic way to track this effect is to define the concept of (ϵ, ℓ) -uniformity (see [20, Definition 3.1]) meaning that an estimate holds uniformly for $N\eta\rho \geq \ell N^\epsilon$ with some $\ell \in \mathbf{N}$. We keep ϵ fixed but ℓ may increase by one in some steps. However, this happens only finitely many times and it is inconsequential to our main argument, hence we entirely omit tracking the ℓ -dependence.

Proposition 3.2 (Hierarchy of master inequalities). *Fix $k \geq 1$, and assume that (3.2) and (3.3) have been established uniformly in \mathbf{A} , and z with $N\eta\rho \geq N^\epsilon$, for all $K \leq 2k$. Then⁷ it holds that*

$$\begin{aligned} \Psi_k^{\text{av}} &\prec \Phi_k + \left(\frac{\psi_{2k}^{\text{av}}}{\sqrt{N\eta\rho}}\right)^{1/2} + \psi_{k-1}^{\text{av}} + \frac{\psi_k^{\text{av}}}{\sqrt{N\eta}} + (\psi_k^{\text{iso}})^{2/3}\Phi_{k-1}^{1/3} + \sum_{j=1}^{k-1} \sqrt{\psi_j^{\text{iso}}\Omega_{k-j}(\psi_k^{\text{iso}} + \Phi_{k-1})} \\ &\quad + \frac{1}{N\eta} \sum_{j=1}^{k-1} \psi_j^{\text{av}} \left(1 + \psi_{k-j}^{\text{av}} \sqrt{\frac{\rho}{N\eta}}\right) \end{aligned} \quad (3.8)$$

$$\Psi_k^{\text{iso}} \prec \Phi_k + \psi_{k-1}^{\text{iso}} + \frac{1}{N\eta} \left[\sum_{j=1}^k \psi_j^{\text{av}} \left(1 + \sqrt{\frac{\rho}{N\eta}} \psi_{k-j}^{\text{iso}}\right) + \sum_{j=0}^{2k} \sqrt{\psi_j^{\text{iso}}\psi_{2k-j}^{\text{iso}}} + \psi_k^{\text{iso}} \right] \quad (3.9)$$

with the definitions

$$\Omega_k := \sum_{k_1+k_2+\dots\leq k} \prod_{i\geq 1} \left(1 + \psi_{k_i}^{\text{iso}} \sqrt{\frac{\rho}{N\eta}}\right) \leq \Phi_k := \sum_{k_1+k_2+\dots\leq k} \prod_{i=1}^2 \left(1 + \frac{\psi_{2k_i}^{\text{iso}}}{\sqrt{N\eta\rho}}\right)^{1/2} \prod_{i\geq 3} \left(1 + \psi_{k_i}^{\text{iso}} \sqrt{\frac{\rho}{N\eta}}\right), \quad (3.10)$$

where the second sum is taken over an arbitrary number of non-negative integers k_i , with $k_i \geq 1$ for $i \geq 3$, under the condition that their sum does not exceed k (in the case of only one non-zero k_1 the second factor and product in (3.10) are understood to be one and $\Phi_0 = 1$).

This hierarchy has the structure that each $\Psi_k^{\text{av/iso}}$ is estimated partly by ψ 's with index higher than k , which potentially is uncontrollable even if the coefficient of the higher order terms is small (recall that $1/(N\eta)$, and $1/(N\eta\rho)$ are small quantities). Thus the hierarchy must be complemented by another set of inequalities that estimate higher indexed Ψ 's with smaller indexed ones even at the expense of a large constant. The success of this scheme eventually depends on the relative size of these small and large constants, so it is very delicate. We prove the following reduction inequalities to estimate the $\psi_l^{\text{av/iso}}$ terms with $k+1 \leq l \leq 2k$ in (3.8)–(3.9) by ψ 's with indices smaller or equal than k .

Lemma 3.3 (Reduction lemma). *Fix $1 \leq j \leq k$ and assume that (3.2)–(3.3) have been established uniformly for $K \leq 2k$. Then it holds that*

$$\Psi_{2k}^{\text{av}} \lesssim \sqrt{\frac{N\eta}{\rho}} + \begin{cases} \sqrt{\frac{\rho}{N\eta}} (\psi_k^{\text{av}})^2 & k \text{ even,} \\ \psi_{k-1}^{\text{av}} + \psi_{k+1}^{\text{av}} + \sqrt{\frac{\rho}{N\eta}} \psi_{k-1}^{\text{av}} \psi_{k+1}^{\text{av}} & k \text{ odd,} \end{cases} \quad (3.11)$$

and for even k also that

$$\Psi_{k+j}^{\text{iso}} \lesssim \sqrt{\frac{N\eta}{\rho}} + \left(\frac{N\eta}{\rho}\right)^{1/4} (\psi_{2j}^{\text{av}})^{1/2} + \psi_k^{\text{iso}} + \left(\frac{\rho}{N\eta}\right)^{1/4} (\psi_{2j}^{\text{av}})^{1/2} \psi_k^{\text{iso}}. \quad (3.12)$$

The rest of the present section is structured as follows: In Section 3.1 we prove (3.8) while in Section 3.2 we prove (3.9). Then in Section 3.3 we prove Lemma 3.3 and conclude the proof of Theorem 2.2. Before starting the main proof we collect some trivial estimates between Hilbert-Schmidt and operator norms using matrix Hölder inequalities.

Lemma 3.4. *For $N \times N$ matrices B_1, \dots, B_k and $k \geq 2$ it holds that*

$$\left| \left\langle \prod_{i=1}^k B_i \right\rangle \right| \leq \prod_{i=1}^k \langle |B_i|^k \rangle^{1/k} \leq N^{k/2-1} \prod_{i=1}^k \langle |B_i|^2 \rangle^{1/2}, \quad (3.13)$$

and

$$\|B\| = \sqrt{\lambda_{\max}(|B|^2)} \leq N^{1/2} \langle |B|^2 \rangle^{1/2}. \quad (3.14)$$

In the sequel we often drop the indices from G, A , hence we write $(GA)^k$ for $G_1 A_1 \dots G_k A_k$, and assume without loss of generality that $A_i = A_i^*$ and $\langle A_i^2 \rangle = 1$. We also introduce the convention in this paper that matrices denoted by capital A letter are always traceless.

⁷Following⁶, we omit tracking the spectral domains in the main text. We only mention here that the pedantic formulation of Proposition 3.2 would assert that if (3.2)–(3.3) hold (ϵ, ℓ) -uniformly, for some $\ell \in \mathbf{N}$, then the conclusions in (3.8)–(3.9) hold $(\epsilon, \ell + 1)$ -uniformly.

3.1. **Proof of averaged estimate (3.8) in Proposition 3.2.** We now identify the leading contribution of $\langle (GA)^k - m^k A^k \rangle$. For any matrix-valued function $f(W)$ we define the *second moment renormalisation*, denoted by underline, as

$$\underline{Wf(W)} := Wf(W) - \widetilde{\mathbf{E}}_{\text{GUE}} \widetilde{W} (\partial_{\widetilde{W}} f)(W) \quad (3.15)$$

in terms of the directional derivative $\partial_{\widetilde{W}}$ in the direction of an independent GUE-matrix \widetilde{W} . The motivation for the second moment renormalisation is that by Gaussian integration by parts it holds that $\mathbf{E} Wf(W) = \widetilde{\mathbf{E}} \widetilde{W} (\partial_{\widetilde{W}} f)(W)$ whenever W is a Gaussian random matrix of zero mean, and \widetilde{W} is an independent copy of W . In particular it holds that $\mathbf{E} \underline{Wf(W)} = 0$ whenever W is a GUE-matrix, while $\mathbf{E} \underline{Wf(W)}$ is small but non-zero for GOE or non-Gaussian matrices. By concentration and universality we expect that to leading order $Wf(W)$ may be approximated by $\widetilde{\mathbf{E}} \widetilde{W} (\partial_{\widetilde{W}} f)(W)$. Here the directional derivative $\partial_{\widetilde{W}} f$ should be understood as

$$(\partial_{\widetilde{W}} f)(W) := \lim_{\epsilon \rightarrow 0} \frac{f(W + \epsilon \widetilde{W}) - f(W)}{\epsilon}.$$

In our application the function $f(W)$ always is a (product of) matrix resolvents $G(z) = (W - z)^{-1}$ and possibly deterministic matrices A_i . This time we view the resolvent as a function of W , $G(W) = (W - z)^{-1}$ for any fixed z . By the resolvent identity it follows that

$$(\partial_{\widetilde{W}} G)(W) = \lim_{\epsilon \rightarrow 0} \frac{(W + \epsilon \widetilde{W} - z)^{-1} - (W - z)^{-1}}{\epsilon} = - \lim_{\epsilon \rightarrow 0} (W + \epsilon \widetilde{W} - z)^{-1} \widetilde{W} (W - z)^{-1} = -G(W) \widetilde{W} G(W), \quad (3.16)$$

while the expectation of a product of GUE-matrices acts as an averaged trace in the sense

$$\widetilde{\mathbf{E}}_{\text{GUE}} \widetilde{W} A \widetilde{W} = \frac{1}{N} \sum_{ab} \Delta^{ab} A \Delta^{ba} = \langle A \rangle I,$$

where I denotes the identity matrix and $(\Delta^{ab})_{cd} := \delta_{ac} \delta_{bd}$. Therefore, for instance, we have the identities

$$\underline{WG} = WG + \langle G \rangle G, \quad \underline{WGAG} = WGAG + \langle G \rangle GAG + \langle GAG \rangle G = \underline{WGAG} + \langle GAG \rangle G.$$

Finally, we want to comment on the choice of renormalising with respect to an independent GUE rather than GOE matrix. In fact this is purely a matter of convenience and we could equally have chosen the GOE-renormalisation. Indeed, we have

$$\widetilde{\mathbf{E}}_{\text{GOE}} \widetilde{W} A \widetilde{W} = \langle A \rangle I + \frac{A^t}{N}$$

and therefore, for instance,

$$\underline{WG}_{\text{GOE}} = \underline{WG}_{\text{GUE}} + \frac{G^t G}{N}$$

which is a negligible difference. Our formulas below will be slightly simpler with our choice in (3.15) even though now $\underline{EWf(W)}$ is not exactly zero even for $W \sim \text{GOE}$.

Lemma 3.5. *We have*

$$\left\langle \prod_{i=1}^k (G_i A_i) - \prod_{i=1}^k m_i A_i \right\rangle \left(1 + \mathcal{O}_{\prec} \left(\frac{\psi_0^{\text{av}}}{N\eta} \right) \right) = -m_1 \langle \underline{WG_1 A_1 \cdots G_k A_k} \rangle + \mathcal{O}_{\prec}(\mathcal{E}_k^{\text{av}}), \quad (3.17)$$

where $\mathcal{E}_1^{\text{av}} = 0$ and

$$\begin{aligned} \mathcal{E}_2^{\text{av}} &:= \sqrt{\frac{\rho}{N\eta}} \left(\psi_1^{\text{av}} + \frac{\psi_0^{\text{av}}}{\sqrt{N\eta\rho}} + \frac{1}{N\eta} \sqrt{\frac{\rho}{N\eta}} (\psi_1^{\text{av}})^2 \right), \\ \mathcal{E}_k^{\text{av}} &:= N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \left(\psi_{k-1}^{\text{av}} + \frac{1}{N\eta} \sum_{j=1}^{k-1} \psi_j^{\text{av}} \left(1 + \psi_{k-j}^{\text{av}} \sqrt{\frac{\rho}{N\eta}} \right) \right) \end{aligned} \quad (3.18)$$

for $k \geq 3$.

Proof. We start with the expansion

$$\begin{aligned}
\left(1 + \mathcal{O}_{\prec}\left(\frac{\psi_0^{\text{av}}}{N\eta}\right)\right) \langle G_1 A_1 \cdots G_k A_k \rangle &= m_1 \langle G_2 \cdots G_k A_k A_1 \rangle - m_1 \langle \underline{W G_1 A_1 G_2 \cdots G_k A_k} \rangle \\
&= m_1 \langle G_2 \cdots G_k A_k A_1 \rangle + m_1 \sum_{j=2}^k \langle G_1 \cdots G_j \rangle \langle G_j \cdots G_k A_k \rangle \\
&\quad - m_1 \langle \underline{W G_1 A_1 G_2 \cdots G_k A_k} \rangle,
\end{aligned} \tag{3.19}$$

due to

$$G = m - m \underline{W G} + m \langle G - m \rangle G, \tag{3.20}$$

where for $k = 1$ the first two terms in the right hand side of (3.19) are not present. In the second step we extended the underline renormalization to the entire product $\underline{W G_1 A_1 G_2 \cdots G_k A_k}$ at the expense of generating additional terms collected in the summation; this identity can directly be obtained from the definition (3.15). Note that in the first line of (3.19) we moved the term coming from $m_1 \langle G_1 - m_1 \rangle G_1$ of (3.20) to the left hand side causing the error $\mathcal{O}_{\prec}(\psi_0^{\text{av}}/(N\eta))$. For $k \geq 2$, using Lemmas 3.1 and 3.4 we estimate the second term in the second line of (3.19) by

$$\begin{aligned}
&|\langle G_1 \cdots G_j \rangle \langle G_j \cdots G_k A_k \rangle| \\
&\prec \left(\frac{\rho}{\eta} |\langle A_1 \cdots A_{j-1} \rangle| + \frac{\psi_{j-1}^{\text{av}} \rho^{1/2} N^{j/2-2}}{\eta^{3/2}} \right) \left(|\langle A_j \cdots A_k \rangle| + \frac{\psi_{k-j+1}^{\text{av}} \rho^{1/2} N^{(k-j)/2-1}}{\eta^{1/2}} \right) \\
&\lesssim \frac{N^{k/2-1} \rho}{N\eta} \left(1 + \frac{\psi_{j-1}^{\text{av}}}{\sqrt{N\eta\rho}} \right) \left(1 + \psi_{k-j+1}^{\text{av}} \sqrt{\frac{\rho}{N\eta}} \right).
\end{aligned} \tag{3.21}$$

For the first term in the second line of (3.19) we distinguish the cases $k = 2$ and $k \geq 3$. In the former we write

$$m_1 \langle G_2 A_2 A_1 \rangle = m_1 \langle G_2 \rangle \langle A_2 A_1 \rangle + m_1 \langle G_2 (A_2 A_1)^\circ \rangle = m_1 \langle A_1 A_2 \rangle \left(m_2 + \mathcal{O}_{\prec}\left(\frac{\psi_0^{\text{av}}}{N\eta}\right) \right) + \mathcal{O}_{\prec}\left(\frac{\psi_1^{\text{av}} \rho^{1/2}}{(N\eta)^{1/2}}\right), \tag{3.22}$$

where we used Lemma 3.4 to estimate

$$\left(|(A_2 A_1)^\circ|^2 \right)^{1/2} = \left(|A_2 A_1|^2 - |(A_2 A_1)|^2 \right)^{1/2} \leq N^{1/2} \tag{3.23}$$

In case $k \geq 3$ we estimate

$$\begin{aligned}
m_1 \langle G_2 \cdots G_k A_k A_1 \rangle &= m_1 \langle G_2 \cdots G_k (A_k A_1)^\circ \rangle + m_1 \langle G_2 \cdots G_k \rangle \langle A_k A_1 \rangle \\
&= m_1 \cdots m_k \langle A_2 \cdots A_{k-1} (A_k A_1)^\circ \rangle \\
&\quad + \mathcal{O}_{\prec}\left(N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \left[\psi_{k-1}^{\text{av}} + \sqrt{\frac{\rho}{N\eta}} \left(1 + \frac{\psi_{k-2}^{\text{av}}}{\sqrt{N\eta\rho}} \right) \right] \right).
\end{aligned} \tag{3.24}$$

Note that the leading deterministic term of $\langle G_2 \cdots G_k \rangle$ was simply estimated as

$$\left| m[z_2, z_k] m_3 \cdots m_{k-1} \left\langle \prod_{i=2}^{k-1} A_i \right\rangle \right| \lesssim \frac{\rho}{\eta} N^{k/2-2}. \tag{3.25}$$

From (3.24) we write $\langle A_2 \cdots A_{k-1} (A_k A_1)^\circ \rangle = \langle A_1 \cdots A_k \rangle - \langle A_1 A_k \rangle \langle A_2 \cdots A_{k-1} \rangle$ where the second term can simply be estimated as $|\langle A_1 A_k \rangle \langle A_2 \cdots A_{k-1} \rangle| \leq N^{k/2-2}$, due to Lemma 3.4, and included in the error term. Collecting all other error terms from (3.21) and (3.24) and recalling $\psi_j^{\text{av}/\text{iso}} \geq 1 \gtrsim \sqrt{\rho/(N\eta)}$ for all j , we obtain (3.17) with the definition of \mathcal{E}_k from (3.18). \square

Lemma 3.5 reduces understanding the local law to the underline term in (3.19) since $\mathcal{E}_k^{\text{av}}$ will be treated as an error term. For the underline term we use a cumulant expansion when calculating the high moment $\mathbf{E}|(GA)^k - m^k A^k|^p$ for any fixed integer p . Here we will make again a notational simplification ignoring different indices in G , A and m , and, in particular we may write

$$\left| \left\langle \prod_{i=1}^k (G_i A_i) - \prod_{i=1}^k m_i A_i \right\rangle \right|^p = \langle (GA)^k - m^k A^k \rangle^p \tag{3.26}$$

by choosing $G = G(\bar{z}_i)$ for half of the factors.

We set $\partial_{ab} := \partial/\partial w_{ab}$ as the derivative with respect to the (a, b) -entry of W , i.e. we consider w_{ab} and w_{ba} as independent variables in the following cumulant expansion (such expansion was first used in the random matrix context in [37] and later revived in [35, 41])

$$\mathbf{E} w_{ab} f(W) = \sum_{j=1}^{\infty} \frac{1}{j! N^{(j+1)/2}} \begin{cases} \kappa_{j+1}^{\text{od}} \mathbf{E} (\partial_{ab} + \partial_{ba})^j f(W), & a \neq b, \\ \kappa_{j+1}^{\text{d}} \mathbf{E} \partial_{aa}^j f(W), & a = b. \end{cases}$$

Technically we use a truncated version of the expansion above, see e.g. [28, 35, 28]. We thus compute⁸

$$\begin{aligned} & \mathbf{E} \langle \underline{W(GA)^k} \rangle \langle (GA)^k - m^k A^k \rangle^{p-1} \\ &= \frac{1}{N} \mathbf{E} \sum_{ab} \frac{[(GA)^k]_{ba}}{N} (\partial_{ab} + \partial_{ba}) \langle (GA)^k - m^k A^k \rangle^{p-1} + \mathbf{E} \sum_{ab} \frac{\partial_{ab} [((GA)^k)_{ba}]}{N^2} \langle (GA)^k - m^k A^k \rangle^{p-1} \\ &+ \sum_{j=2}^R \frac{\kappa_{j+1}^{\text{d}}}{j! N^{(j+3)/2}} \mathbf{E} \sum_a \partial_{aa}^j \left([(GA)^k]_{aa} \langle (GA)^k - m^k A^k \rangle^{p-1} \right) \\ &+ \sum_{j=2}^R \frac{\kappa_{j+1}^{\text{od}}}{j! N^{(j+3)/2}} \mathbf{E} \sum_{a \neq b} (\partial_{ab} + \partial_{ba})^j \left([(GA)^k]_{ba} \langle (GA)^k - m^k A^k \rangle^{p-1} \right) + \mathcal{O}_{\prec}((N^{k/2-3/2})^p) \end{aligned} \quad (3.27)$$

recalling Assumption 1 for the diagonal and off-diagonal cumulants. The summation runs over all indices $a, b \in [N]$. The second cumulant calculation in (3.27) used the fact that by definition of the underline renormalisation the ∂_{ba} -derivative in the first line may not act on its own $(GA)^k$.

For the first term of (3.27) we use $\partial_{ab} \langle (GA)^k \rangle = -k N^{-1} ((GA)^k G)_{ba}$ due to (3.16) with $\widetilde{W} = \Delta^{ab}$ so that using $G^t = G$ we can perform the summation and obtain

$$\begin{aligned} & \left| \frac{1}{N} \sum_{ab} \frac{[(GA)^k]_{ba}}{N} (\partial_{ab} + \partial_{ba}) \langle (GA)^k - m^k A^k \rangle^{p-1} \right| \\ & \lesssim \left| \frac{\langle (GA)^{2k} G \rangle}{N^2} \right| |\langle (GA)^k - m^k A^k \rangle|^{p-2} \prec \left(N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \right)^2 \left(1 + \frac{\psi_{2k}^{\text{av}}}{\sqrt{N\eta\rho}} \right) |\langle (GA)^k - m^k A^k \rangle|^{p-2} \end{aligned} \quad (3.28)$$

from Lemma 3.1 with estimating the deterministic leading term of $\langle (GA)^{2k} G \rangle$ by $|m^{(2)} m^{2k-1} \langle A^{2k} \rangle| \leq N^{k-1} \rho/\eta$ as in (3.25). The first prefactor in the right hand side of (3.28) is already written as the square of the target size $N^{k/2-1} \sqrt{\rho/(N\eta)}$ for $\langle (GA)^k - m^k A^k \rangle$, see (2.5).

For the second term of (3.27) we estimate

$$\begin{aligned} \sum_{ab} \frac{\partial_{ab} [((GA)^k)_{ba}]}{N^2} &= -\frac{1}{N^2} \sum_{j=0}^{k-1} \sum_{ab} ((GA)^j G)_{ba} ((GA)^{k-j})_{ba} = -\frac{1}{N} \sum_{j=0}^{k-1} \langle (GA)^j G (A^t G)^{k-j} \rangle \\ &= \mathcal{O}_{\prec} \left(N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \left(\sqrt{\frac{\rho}{N\eta}} + \frac{\sqrt{\rho} \psi_k^{\text{av}}}{N\eta} \right) \right), \end{aligned}$$

recalling that $G = G^t$ since W is real symmetric⁹.

For the second line of (3.27) we define the set of multi-indices $\mathbf{l} = (l_1, l_2, \dots, l_n)$ with arbitrary length n , denoted by $|\mathbf{l}| := n$, and total size $k = \sum_i l_i$ as

$$\mathcal{I}_k^{\text{d}} := \left\{ \mathbf{l} \in \mathbf{N}_0^n \mid n \leq R, \sum_i l_i = k \right\}, \quad R := (3 + 4k)p. \quad (3.29)$$

⁸The truncation error of the cumulant expansion after $R = (3 + 4k)p$ terms can be estimated trivially by the single- G local law for resolvent entries, and by norm for entries of $GAG \dots$ resolvent chains.

⁹We recall that we present the proof for the slightly more involved real symmetric case. In the complex Hermitian case the second term on the right hand side of (3.27) would not be present.

Note that the set \mathcal{I}_k^d is a finite set with cardinality depending only on k, p . We distribute the derivatives according to the product rule to estimate

$$\begin{aligned} & \left| \sum_{j \geq 2} \frac{\kappa_{j+1}^d}{j! N^{(j+3)/2}} \sum_a \partial_{aa}^j \left([(GA)^k]_{aa} \langle (GA)^k - m^k A^k \rangle^{p-1} \right) \right| \\ & \leq \sum_{\substack{\mathbf{l} \in \mathcal{I}_k^d, J \subset \mathcal{I}_k^d \\ l_{|l|} \geq 1, |l| + \sum J \geq 3}} \Xi_k^d(\mathbf{l}, J) |\langle (GA)^k - m^k A^k \rangle|^{p-1-|J|}, \end{aligned} \quad (3.30)$$

where for the multi-set J we define $\sum J := \sum_{j \in J} |j|$ and set

$$\Xi_k^d := \frac{N^{-\frac{|l| + \sum J}{2}}}{N^{1+|J|}} \left| \sum_a [(GA)^{l_1} G]_{aa} \cdots [(GA)^{l_{|l|-1}} G]_{aa} [(GA)^{l_{|l|}}]_{aa} \prod_{j \in J} [(GA)^{j_1} G]_{aa} \cdots [(GA)^{j_{|j|}} G]_{aa} \right|. \quad (3.31)$$

Here for the multi-set $J \subset \mathcal{I}_k^d$ we defined its cardinality by $|J|$ and we set $\sum J := \sum_{j \in J} |j|$. Along the product rule, the multi-index \mathbf{l} encodes how the first factor $[(GA)^k]_{aa}$ in (3.30) is differentiated, while each element $j \in J$ is a multi-index that encodes how another factor $\langle (GA)^k - m^k A^k \rangle$ is differentiated. Note that $|J|$ is the number of such factors affected by derivatives, the remaining $p-1-|J|$ factors are untouched.

For the third line of (3.27) we similarly define the appropriate index set that is needed to encode the product rule¹⁰

$$\mathcal{I}_k^{\text{od}} := \left\{ (\mathbf{l}, \boldsymbol{\alpha}) \in \mathbf{N}_0^{|\mathbf{l}|} \times \{ab, ba, aa, bb\}^{|\mathbf{l}|} \mid |\mathbf{l}| \leq R, \sum_i l_i = k, |\{i \mid \alpha_i = aa\}| = |\{i \mid \alpha_i = bb\}| \right\}. \quad (3.32)$$

Note that in addition to the multi-index \mathbf{l} encoding the distribution of the derivatives after the Leibniz rule similarly to the previous diagonal case, the second element $\boldsymbol{\alpha}$ of the new type of indices also keeps track of whether after the differentiations the corresponding factor is evaluated at ab, ba, aa or bb . While a single ∂_{ab} or ∂_{ba} acting on $\langle (GA)^k - m^k A^k \rangle$ results in an off-diagonal term of the form $[(GA)^k G]_{ab}$ or $[(GA)^k G]_{ba}$, a second derivative also produces diagonal terms. The derivative action on the first factor $[(GA)^k]_{ba}$ in the third line of (3.27) produces diagonal factors already after one derivative. The restriction in (3.31) that the number of aa and bb -type diagonal elements must coincide comes from a simple counting of diagonal indices along derivatives: when an additional ∂_{ab} hits an off-diagonal term, then either one aa and one bb diagonal is created or none. Similarly, when an additional ∂_{ab} hits a diagonal aa term, then one diagonal aa remains, along with a new off-diagonal ab . In any case the difference of the aa and bb diagonals is unchanged.

Armed with this notation, similarly to (3.30) we estimate

$$\begin{aligned} & \left| \sum_{j \geq 2} \frac{\kappa_{j+1}^{\text{od}}}{j! N^{(j+3)/2}} \sum_{a,b} (\partial_{ab} + \partial_{ba})^j \left([(GA)^k]_{ba} \langle (GA)^k - m^k A^k \rangle^{p-1} \right) \right| \\ & \leq \sum_{\substack{(\mathbf{l}, \boldsymbol{\alpha}) \in \mathcal{I}_k^{\text{od}}, J \subset \mathcal{I}_k^{\text{od}} \\ l_{|l|} \geq 1, |l| + \sum J \geq 3}} \Xi_k^{\text{od}}((\mathbf{l}, \boldsymbol{\alpha}), J) |\langle (GA)^k - m^k A^k \rangle|^{p-1-|J|}, \end{aligned} \quad (3.33)$$

where for the multi-set $J \subset \mathcal{I}_k^{\text{od}}$ we define $\sum J := \sum_{(j, \beta) \in J} |j|$ and set

$$\Xi_k^{\text{od}} := \frac{N^{-\frac{|l| + \sum J}{2}}}{N^{1+|J|}} \left| \sum_{ab} [(GA)^{l_1} G]_{\alpha_1} \cdots [(GA)^{l_{|l|}}]_{\alpha_{|l|}} \prod_{(j, \beta) \in J} [(GA)^{j_1} G]_{\beta_1} \cdots [(GA)^{j_{|j|}} G]_{\beta_{|j|}} \right|. \quad (3.34)$$

Note that (3.33) is an overestimate: not all indices (j, β) indicated in (3.34) can actually occur after the Leibniz rule.

¹⁰In the definition of $\mathcal{I}_k^{\text{od}}$ the indices ab, ba, aa, bb should be understood symbolically, merely indicating the diagonal or off-diagonal character of the term. However, in the formula (3.34) below the concrete summation indices a, b are substituted for the symbolic expressions. Alternatively, we could have avoided this slight abuse of notation by defining $\alpha_i \in \{(1, 1), (1, 2), (2, 1), (2, 2)\}$, sum over $a_1, a_2 = 1, \dots, N$ in (3.34) and substitute $a_{(\alpha_i)_1}, a_{(\alpha_i)_2}$ for α_i , however this would be an excessive pedantry.

Lemma 3.6. *For any $k \geq 1$ it holds that*

$$\Xi_k^{\text{d}} + \Xi_k^{\text{od}} \prec \left(N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \left(\Phi_k + (\psi_k^{\text{iso}})^{2/3} \Phi_{k-1}^{1/3} + \sum_{j=1}^{k-1} \sqrt{(\Phi_{k-1} + \psi_k^{\text{iso}}) \psi_j^{\text{iso}} \Omega_{k-j}} \right) \right)^{1+|J|} \quad (3.35)$$

By combining Lemma 3.5 and (3.27), (3.28), (3.30) and (3.33) with Lemma 3.6 and using a simple Hölder inequality, we obtain, for any fixed $\xi > 0$, that

$$\begin{aligned} & \left(\mathbf{E} | \langle (GA)^k - m^k A^k \rangle |^p \right)^{1/p} \\ & \lesssim N^\xi N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \left(\Phi_k + \left(\frac{\psi_{2k}^{\text{av}}}{\sqrt{N\eta\rho}} \right)^{1/2} + \psi_{k-1}^{\text{av}} + \frac{\psi_k^{\text{av}}}{\sqrt{N\eta}} + (\psi_k^{\text{iso}})^{2/3} \Phi_{k-1}^{1/3} \right. \\ & \quad \left. + \sum_{j=1}^{k-1} \sqrt{(\Phi_{k-1} + \psi_k^{\text{iso}}) \psi_j^{\text{iso}} \Omega_{k-j}} + \frac{1}{N\eta} \sum_{j=1}^{k-1} \psi_j^{\text{av}} \left(1 + \psi_{k-j}^{\text{av}} \sqrt{\frac{\rho}{N\eta}} \right) \right), \end{aligned} \quad (3.36)$$

where we used the Ξ_k^{d} term to add back the $a = b$ part of the summation in (3.33) compared to (3.27). By taking p large enough, ξ arbitrarily small, and using the definition of \prec and the fact that the bound (3.36) holds uniformly in the spectral parameters and the deterministic matrices, we conclude the proof of (3.8).

Proof of Lemma 3.6. The proof repeatedly uses (3.3) in the form

$$((GA)^k G)_{ab} \prec N^{k/2-1/2} \left(\|Ae_a\| \wedge \|Ae_b\| + \psi_k^{\text{iso}} \sqrt{\frac{\rho}{\eta}} \right) \lesssim N^{k/2} \left(1 + \psi_k^{\text{iso}} \sqrt{\frac{\rho}{N\eta}} \right), \quad (3.37)$$

$$((GA)^k)_{ab} \prec N^{k/2-1/2} \|Ae_b\| \left(1 + \psi_{k-1}^{\text{iso}} \sqrt{\frac{\rho}{N\eta}} \right) \lesssim N^{k/2} \left(1 + \psi_{k-1}^{\text{iso}} \sqrt{\frac{\rho}{N\eta}} \right) \quad (3.38)$$

with e_b being the b -th coordinate vector, where we estimated the deterministic leading term $m^k (A^k)_{ab}$ by $|(A^k)_{ab}| \leq \|A\|^{k-1} \|Ae_b\| \leq N^{(k-1)/2} \|Ae_b\|$ using (3.14). Recalling the normalization $\langle |A|^2 \rangle = 1$, the best available bound on $\|Ae_b\|$ is $\|Ae_b\| \leq N^{1/2}$, however this can be substantially improved under a summation over the index b :

$$\sum_b \|Ae_b\|^2 = N \langle |A|^2 \rangle \leq N, \quad \sum_b \|Ae_b\| \leq \sqrt{N} \sqrt{\sum_b \|Ae_b\|^2} \leq N. \quad (3.39)$$

Using (3.37) and (3.38) for each entry of (3.31) and (3.34), we obtain the following *naive (or a priori) estimates* on $\Xi_k^{\text{d/od}}$

$$\Xi_k^{\text{d/od}} \prec \left(N^{k/2-1} \frac{\Omega_k}{\sqrt{N}} \right)^{1+|J|} N^{1+\mathbf{1}(\text{od})+(|J|-|l|-\sum J)/2} \quad (3.40)$$

where we recall the definition of Ω_k from (3.10). Using $\Omega_k/\sqrt{N} \lesssim \Phi_k \sqrt{\rho/(N\eta)}$ due to $1 \lesssim \rho/\eta$ the claim (3.35) follows trivially from (3.40) for Ξ_k^{d} and Ξ_k^{od} whenever $|l| + \sum J \geq 2 + |J|$ or $|l| + \sum J \geq 4 + |J|$, respectively, i.e. when the exponent of N in (3.40) is non-positive.

In the rest of the proof we consider the remaining diagonal **D1** and off-diagonal cases **O1–O3** that we will define below. The cases are organised according to the quantity $|l| + \sum J - |J|$ which captures by how many factors of $N^{1/2}$ the naive estimate (3.40) exceeds the target (3.35) when all Φ 's and ψ 's are set to be order one. Within case **O1** we further differentiate whether an off-diagonal index pair ab or ba appears at least once in the tuple α or in one of the tuples β . Within case **O2** we distinguish according to the length of $|l|$ and $|J|$ as follows:

D1 $|l| + \sum J = |J| + 1$

O1 $|l| + \sum J = |J| + 3$

O1a $ab \vee ba \in \alpha \cup \bigcup_{(j,\beta) \in J} \beta$

O1b $J \in \{ \{(j, (aa, bb))\}, \{(j, (bb, aa))\} \}$ and $\alpha \in \{ (aa, bb), (bb, aa) \}$, i.e. $\sum J = |l| = 2$ and $|J| = 1$

O2 $|l| + \sum J = |J| + 2$

O2a $|l| = 1,$

O2b $|l| = 2, |J| \geq 2,$

O2c $|l| = 2, |J| = 1, l_1 \geq 1,$

O2d $|l| = 2, |J| = 1, l_1 = 0.$

O3 $|\mathbf{l}| + \sum J = |J| + 1$

The list of four cases above is exhaustive since $\sum J + |\mathbf{l}| \geq |J| + 1$ by definition, and the subcases of **O2** are obviously exhaustive. Within case **O1** either some off-diagonal element appears in α or some β (hence we are in case **O1a**), or the number of elements in α and all β is even, c.f. the constraint on the number of diagonal elements in (3.32). The latter case is only possible if $|J| = 1$, $|\mathbf{l}| = \sum J = 2$ which is case **O1b** (note that $|\mathbf{l}| \geq 2$ implies $|J| \leq 1$, and $|J| = 0$ is impossible as it would imply $|\mathbf{l}| = 3$, the number of elements in α , is odd).

Now we give the estimates for each case separately. For case **D1**, using the restriction in the summation in (3.33) to get $3 \leq |\mathbf{l}| + \sum J = 1 + |J|$, we estimate

$$\begin{aligned} \Xi_k^{\text{d}} &= N^{-3(1+|J|)/2} \left| \sum_a [(GA)^k]_{aa} [(GA)^k G]_{aa}^{|J|} \right| \\ &< \frac{(N^{k/2-1})^{|J|+1}}{N^{|J|/2+1}} \Omega_k^{|J|-1} \Omega_{k-1} \sum_a \|Ae_a\| \left(\frac{\|Ae_a\|}{N^{1/2}} + \sqrt{\frac{\rho}{N\eta}} \psi_k^{\text{iso}} \right) \\ &\lesssim \left(N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \right)^{1+|J|} \Phi_k^{|J|-1} \Phi_{k-1} \psi_k^{\text{iso}} \end{aligned} \quad (3.41)$$

where we used the first inequalities of (3.37) and (3.38) for the $(GA)^k$ and one of the $(GA)^k G$ factors, and the second inequality of (3.37) for the remaining factors, and in the last step we used (3.39) and $\psi_k^{\text{iso}} \sqrt{\rho/\eta} \gtrsim 1$. Finally we use Young's inequality $\Phi_k^{|J|-1} \Phi_{k-1} \psi_k^{\text{iso}} \leq \Phi_k^{|J|+1} + (\Phi_{k-1} \psi_k^{\text{iso}})^{(|J|+1)/2}$. This confirms (3.35) in case **D1**.

For the offdiagonal cases we will use the following so-called *Ward-improvements*:

- I1** Averaging over a or b in $|((GA)^k G)_{ab}|$ gains a factor of $\sqrt{\rho/(N\eta)}$ compared to (3.37),
- I2** Averaging over a in $|((GA)^k)_{ab}|$ gains a factor of $\sqrt{\rho/(N\eta)}$ compared to (3.38),

at the expense of replacing a factor of $(1 + \psi_k^{\text{iso}} \sqrt{\rho/(N\eta)})$ in the definition of Ω_k by a factor of $(1 + \psi_{2k}^{\text{iso}}/\sqrt{N\eta\rho})^{1/2}$. These latter replacements necessitate changing Ω_k to the larger Φ_k as a main control parameter in the estimates after Ward improvements. Indeed, **I1** and **I2** follow directly from (3.6) of Lemma 3.1 and $|m^{(2)}| \lesssim \rho/\eta$, more precisely

$$\begin{aligned} \frac{1}{N} \sum_a |[(GA)^k G]_{ab}| &\leq \frac{\sqrt{[(G^*A)^k G^* G (AG)^k]_{bb}}}{\sqrt{N}} < N^{k/2} \sqrt{\frac{\rho}{N\eta}} \left(1 + \psi_{2k}^{\text{iso}} \sqrt{\frac{1}{N\eta\rho}}\right)^{1/2} \\ \frac{1}{N} \sum_a |[(GA)^k]_{ab}| &\leq \frac{\sqrt{[(AG^*)^k (GA)^k]_{bb}}}{\sqrt{N}} < N^{k/2-1/2} \|Ae_b\| \sqrt{\frac{\rho}{N\eta}} \left(1 + \psi_{2(k-1)}^{\text{iso}} \sqrt{\frac{1}{N\eta\rho}}\right)^{1/2} \\ \frac{1}{N} \sum_a |[(GA)^k]_{ab}|^2 &= \frac{[(AG^*)^k (GA)^k]_{bb}}{N} < N^{k-1} \|Ae_b\|^2 \frac{\rho}{N\eta} \left(1 + \psi_{2(k-1)}^{\text{iso}} \sqrt{\frac{1}{N\eta\rho}}\right), \end{aligned} \quad (3.42)$$

where the first step in each case followed from a Schwarz inequality and summing up the indices explicitly. This improvement is essentially equivalent to using the *Ward-identity* $GG^* = \Im G/\eta$ in (3.42).

Now we collect these gains over the naive bound given in (3.40) for each case. Note that whenever a factor $\sqrt{\rho/(N\eta)}$ is gained, the additional $1/\sqrt{N}$ is freed up along the second inequality in (3.40) which can be used to compensate the positive N -powers.

For case **O3** we have $|J| \geq 2$ and estimate all but the first two (j, β) factors in (3.34) trivially using the last inequality in (3.37) to obtain

$$\Xi_k^{\text{od}} < N^{-3(1+|J|)/2} (N^{k/2} \Omega_k)^{|J|-2} \sum_{ab} |[(GA)^k]_{ba}| |[(GA)^k G]_{ab}| |[(GA)^k G]_{ab}|. \quad (3.43)$$

For the last two factors we use first inequality in (3.37) and then estimate as

$$\begin{aligned} &\sum_{ab} |[(GA)^k]_{ba}| |[(GA)^k G]_{ab}| |[(GA)^k G]_{ab}| \\ &\lesssim N^{k-1} \sum_{ab} |[(GA)^k]_{ba}| \left(\|Ae_a\| \|Ae_b\| + (\psi_k^{\text{iso}})^2 \frac{\rho}{\eta} \right) < \left(N^{k/2} \sqrt{\frac{\rho}{\eta}} \right)^3 \Phi_{k-1} (\psi_k^{\text{iso}})^2, \end{aligned} \quad (3.44)$$

where in the second step we performed a Schwarz inequality for the double a, b summation and used the last bound in (3.42), (3.39) and $1 \lesssim \psi_k^{\text{iso}} \sqrt{\rho/\eta}$. Thus, we conclude

$$\Xi_k^{\text{od}} \prec \left(N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \right)^{|J|+1} \Phi_k^{|J|-2} \Phi_{k-1} (\psi_k^{\text{iso}})^2. \quad (3.45)$$

In case **O2a** there exists some j with $|j| = 2$ (recall that $\sum J = |J| + 1$). By estimating the remaining J -terms trivially by (3.37), we obtain

$$\begin{aligned} \Xi_k^{\text{od}} &\prec N^{-3(1+|J|)/2-1/2} (N^{k/2} \Omega_k)^{|J|-1} \sum_{ab} |[(GA)^k]_{ab}| |[(GA)^{j_1} G]_{\beta_1}| |[(GA)^{j_2} G]_{\beta_2}| \\ &\prec N^{-3(1+|J|)/2-1/2} (N^{k/2} \Omega_k)^{|J|-1} N^{k/2-1/2} \Omega_{j_2} \sum_{ab} |[(GA)^k]_{ab}| \left(\|Ae_a\| + \|Ae_b\| + \psi_{j_1}^{\text{iso}} \sqrt{\frac{\rho}{\eta}} \right) \\ &\lesssim \left(N^{k/2-1} \frac{\Omega_k}{\sqrt{N}} \right)^{|J|-1} \left(N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \right)^2 \Phi_{k-1} \psi_{j_1}^{\text{iso}} \Omega_{j_2} \end{aligned} \quad (3.46)$$

for some $j_1 + j_2 = k$ and double indices $\beta_1, \beta_2 \in \{aa, bb, ab, ba\}$. Here in the second step we assumed without loss of generality $j_1 \geq 1$ (the case $j_2 \geq 1$ being completely analogous) and used the first inequality in (3.37) for $|[(GA)^{j_1} G]_{\beta_1}|$ and the second inequality in (3.37) for $|[(GA)^{j_2} G]_{\beta_2}|$. Finally, in the last step we performed an a, b -Schwarz inequality, used the last bound in (3.42) and (3.39).

In case **O2b** we have $|j| = 1$ for all j since $\sum J + |l| = |J| + 2$ implies $\sum J = |J|$, and we estimate all but two J -factors trivially by the last inequality in (3.37), the other two J -factors (which are necessarily offdiagonal) by the first inequality in (3.37), the l_1 -factor by the last inequality in (3.37) and the l_2 factor by the first inequality in (3.38) (note that $l_2 \geq 1$) to obtain

$$\begin{aligned} \Xi_k^{\text{od}} &\prec N^{-3(1+|J|)/2-1/2} (N^{k/2} \Omega_k)^{|J|-2} \sum_{ab} |[(GA)^{l_1} G]_{\alpha_1}| |[(GA)^{l_2}]_{\alpha_2}| |[(GA)^k G]_{ab}|^2 \\ &\prec N^{-3(1+|J|)/2-1/2} (N^{k/2} \Omega_k)^{|J|-2} N^{3k/2-3/2} \Omega_{k-1} \sum_{ab} (\|Ae_a\| + \|Ae_b\|) \left(\|Ae_a\| \|Ae_b\| + \frac{\rho}{\eta} (\psi_k^{\text{iso}})^2 \right) \\ &\lesssim \left(N^{k/2-1} \frac{\Omega_k}{\sqrt{N}} \right)^{|J|-2} N^{k/2-3/2} \left(N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \right)^2 \Omega_{k-1} (\psi_k^{\text{iso}})^2, \end{aligned} \quad (3.47)$$

where the last step used (3.39) and $\psi_k^{\text{iso}} \sqrt{\rho/\eta} \gtrsim 1$.

In case **O2c** we use the first inequalities of (3.37) and (3.38) for the l_1, l_2 -terms (since $l_1, l_2 \geq 1$) and the first inequality of (3.37) for the $(GA)^k G$ factor to obtain

$$\begin{aligned} \Xi_k^{\text{od}} &\lesssim N^{-7/2} \sum_{ab} |[(GA)^{l_1} G]_{\alpha_1}| |[(GA)^{l_2}]_{\alpha_2}| |[(GA)^k G]_{ab}| \\ &\prec N^{k-5} \Omega_{l_2-1} \sum_{ab} \left(\|Ae_b\| + \|Ae_a\| + \sqrt{\frac{\rho}{\eta}} \psi_{l_1}^{\text{iso}} \right) (\|Ae_a\| + \|Ae_b\|) \left(\|Ae_a\| \wedge \|Ae_b\| + \sqrt{\frac{\rho}{\eta}} \psi_k^{\text{iso}} \right) \\ &\lesssim \left(N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \right)^2 \Omega_{l_2-1} \psi_{l_1}^{\text{iso}} \psi_k^{\text{iso}} \end{aligned} \quad (3.48)$$

by (3.39).

In case **O2d** we write the single- G diagonal as $G_{aa} = m + \mathcal{O}_{\prec} \left(\sqrt{\rho/(N\eta)} \right)$ and use *isotropic resummation* for the leading m term into the $\mathbf{1} = (1, 1, \dots)$ vector of norm $\|\mathbf{1}\| = \sqrt{N}$, i.e.

$$\sum_a G_{aa} [(GA)^k G]_{ab} = m [(GA)^k G]_{\mathbf{1}b} + \mathcal{O}_{\prec} \left(\sqrt{\frac{\rho}{N\eta}} \right) \sum_a |[(GA)^k G]_{ab}|,$$

and estimate

$$\begin{aligned}
\Xi_k^{\text{od}} &\lesssim N^{-7/2} \left| \sum_{ab} G_{aa} [(GA)^k]_{bb} [(GA)^k G]_{ab} \right| + N^{-7/2} \sum_{ab} |G_{ab} [(GA)^k]_{ab} [(GA)^k G]_{ab}| \\
&\prec N^{-7/2} \left| \sum_b [(GA)^k]_{bb} [(GA)^k G]_{1b} \right| + N^{-7/2} \sqrt{\frac{\rho}{N\eta}} \sum_{ab} |[(GA)^k]_{bb} [(GA)^k G]_{ab}| \\
&\prec \sqrt{\frac{\rho}{\eta}} N^{k-4} \Omega_{k-1} \sum_b \|Ae_b\| \left(\|Ae_b\| + \sqrt{\frac{\rho}{\eta}} \psi_k^{\text{iso}} \right) \lesssim \left(N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \right)^2 \Omega_{k-1} \psi_k^{\text{iso}}
\end{aligned} \tag{3.49}$$

using the first inequalities of (3.37) and (3.38).

In case **O1a** we use either **I1** or **I2** depending on whether the off-diagonal matrix is of the form $(GA)^l G$ or $(GA)^l$ to gain one factor of $\sqrt{\rho/(N\eta)}$ in either case and conclude (3.35).

Finally we consider case **O1b** where there is no off-diagonal element to perform Ward-improvement, but for which, using (3.39), we estimate

$$\begin{aligned}
&N^{-4} \left| \sum_{ab} [(GA)^{k_1} G]_{aa} [(GA)^{k_2}]_{bb} [(GA)^{k_3} G]_{aa} [(GA)^{k_4} G]_{bb} \right| \\
&\prec N^{k-5} \Omega_{k-1} \Omega_{k_3} \sum_{ab} \|Ae_b\| \left(\|Ae_b\| + \psi_{k_4}^{\text{iso}} \sqrt{\frac{\rho}{\eta}} \right) \leq N^{k-3} \sqrt{\frac{\rho}{\eta}} \Omega_{k-1} \left(1 + \psi_{k_3}^{\text{iso}} \sqrt{\frac{\rho}{N\eta}} \right) (\psi_{k_4}^{\text{iso}} + 1) \\
&\lesssim \left(N^{k/2-1} \sqrt{\frac{\rho}{N\eta}} \right)^2 \Omega_{k-1} \sum_{j=0}^k \psi_j^{\text{iso}} \Omega_{k-j}
\end{aligned} \tag{3.50}$$

for any exponents with $k_1 + k_2 = k_3 + k_4 = k$. Here in case $k_4 > 0$ we used the second inequalities of (3.37) and (3.38) for the k_2, k_4 factors and the first inequality of (3.37) for the k_1, k_3 factors. The case $k_4 = 0$ is handled similarly, with the same result, by estimating $[(GA)^{k_3} G]_{aa}$ instead of $[(GA)^{k_4} G]_{bb}$ using the first inequality of (3.37). \square

3.2. Proof of isotropic estimate (3.9) in Proposition 3.2. First we state the isotropic version of Lemma 3.5:

Lemma 3.7. *For any deterministic unit vectors \mathbf{x}, \mathbf{y} and $k \geq 0$ we have*

$$\langle \mathbf{x}, [(GA)^k G - m^{k+1} A^k] \mathbf{y} \rangle \left(1 + \mathcal{O}_{\prec} \left(\frac{\psi_0^{\text{av}}}{N\eta} \right) \right) = -m \langle \mathbf{x}, \underline{W(GA)^k G} \mathbf{y} \rangle + \mathcal{O}_{\prec}(\mathcal{E}_k^{\text{iso}}), \tag{3.51}$$

where $\mathcal{E}_0^{\text{iso}} = 0$ and for $k \geq 1$

$$\mathcal{E}_k^{\text{iso}} := N^{k/2} \sqrt{\frac{\rho}{N\eta}} \left(\psi_{k-1}^{\text{iso}} + \frac{1}{N\eta} \sum_{j=1}^k \left(\psi_j^{\text{av}} + \psi_{k-j}^{\text{iso}} + \sqrt{\frac{\rho}{N\eta}} \psi_j^{\text{av}} \psi_{k-j}^{\text{iso}} \right) \right). \tag{3.52}$$

Proof. From (3.20) applied to the first factor $G = G_1$, similarly to (3.19), we obtain

$$\begin{aligned}
&\left(1 + \mathcal{O}_{\prec} \left(\frac{\psi_0^{\text{av}}}{N\eta} \right) \right) \langle \mathbf{x}, (GA)^k G \mathbf{y} \rangle = m \langle \mathbf{x}, (AG)^k \mathbf{y} \rangle - m \langle \mathbf{x}, \underline{W(GA)^k G} \mathbf{y} \rangle \\
&= m^{k+1} \langle \mathbf{x}, A^k \mathbf{y} \rangle - m \langle \mathbf{x}, \underline{W(GA)^k G} \mathbf{y} \rangle \\
&\quad + m_1 \sum_{j=1}^k \langle (GA)^j G \mathbf{x}, (GA)^{k-j} G \mathbf{y} \rangle \\
&\quad + \mathcal{O}_{\prec} \left(N^{(k-1)/2} \sqrt{\frac{\rho}{N\eta}} \|A\mathbf{x}\| \psi_{k-1}^{\text{iso}} \right),
\end{aligned} \tag{3.53}$$

where we used the definition (3.3) for the first term and the definition (3.15). An estimate analogous to (3.21) handles the sum and is incorporated in (3.52). This concludes the proof together with Lemma 3.1 and $\|A\mathbf{x}\| \leq \|A\| \leq N^{1/2}$. \square

Exactly as in (3.27) we perform a cumulant expansion

$$\begin{aligned}
& \mathbf{E}\langle \mathbf{x}, \underline{W(GA)^k G \mathbf{y}} \rangle \langle \mathbf{x}, [(GA)^k G - m^{k+1} A^k] \mathbf{y} \rangle^{p-1} \\
&= \mathbf{E} \sum_{ab} \frac{x_a [(GA)^k G]_{b\mathbf{y}}}{N} (\partial_{ab} + \partial_{ba}) [(GA)^k G - m^{k+1} A^k]_{\mathbf{x}\mathbf{y}}^{p-1} \\
&\quad + \mathbf{E} \sum_{ab} \frac{x_a \partial_{ab} [(GA)^k G]_{b\mathbf{y}}}{N} [(GA)^k G - m^{k+1} A^k]_{\mathbf{x}\mathbf{y}}^{p-1} \\
&\quad + \sum_{j \geq 2} \frac{\kappa_{j+1}^d}{j! N^{(j+1)/2}} \mathbf{E} \sum_a \partial_{aa}^j \left(x_a [(GA)^k G]_{a\mathbf{y}} [(GA)^k G - m^{k+1} A^k]_{\mathbf{x}\mathbf{y}}^{p-1} \right) \\
&\quad + \sum_{j \geq 2} \frac{\kappa_{j+1}^{\text{od}}}{j! N^{(j+1)/2}} \mathbf{E} \sum_{a \neq b} (\partial_{ab} + \partial_{ba})^j \left(x_a [(GA)^k G]_{b\mathbf{y}} [(GA)^k G - m^{k+1} A^k]_{\mathbf{x}\mathbf{y}}^{p-1} \right),
\end{aligned} \tag{3.54}$$

recalling Assumption 1 for the diagonal and off-diagonal cumulants. In fact, the formula (3.54) is identical to (3.27) for $k+1$ instead of k if the last $A = A_{k+1}$ in the product $(GA)^{k+1} = G_1 A_1 G_2 A_2 \dots G_{k+1} A_{k+1}$ is chosen specifically $A_{k+1} = \mathbf{y} \mathbf{x}^*$.

For the first line of (3.54), after performing the derivative, we can also perform the summations and estimate the resulting isotropic resolvent chains by using the last inequality of (3.37) as well as Lemma 3.1 to obtain

$$\begin{aligned}
& \sum_{ab} \frac{x_a [(GA)^k G]_{b\mathbf{y}}}{N} (\partial_{ab} + \partial_{ba}) [(GA)^k G - m^{k+1} A^k]_{\mathbf{x}\mathbf{y}}^{p-1} \\
&= \sum_{j=0}^{2k} \frac{[(GA)^j G]_{\mathbf{x}\mathbf{x}} [(GA)^k G (GA)^{k-j} G]_{\mathbf{y}\mathbf{y}} + [(GA)^j G (GA)^k G]_{\mathbf{x}\mathbf{y}} [(GA)^{k-j} G]_{\mathbf{x}\mathbf{y}}}{N} [(GA)^k G - m^{k+1} A^k]_{\mathbf{x}\mathbf{y}}^{p-2} \\
&\prec \left(N^{k/2} \sqrt{\frac{\rho}{N\eta}} \right)^2 \left(1 + \sum_{j=0}^{2k} \frac{\psi_j^{\text{iso}}}{\sqrt{N\eta\rho}} \left(1 + \sqrt{\frac{\rho}{N\eta}} \psi_{2k-j}^{\text{iso}} \right) \right) |[(GA)^k G - m^{k+1} A^k]_{\mathbf{x}\mathbf{y}}|^{p-2}.
\end{aligned} \tag{3.55}$$

For the second line of (3.54) we estimate

$$\begin{aligned}
\sum_{ab} \frac{x_a \partial_{ab} [(GA)^k G]_{b\mathbf{y}}}{N} &= - \sum_{j=0}^k \sum_{ab} \frac{x_a [(GA)^j G]_{ba} [(GA)^{k-j} G]_{b\mathbf{y}}}{N} \\
&= - \sum_{j=0}^k \frac{[(GA^t)^j G (GA)^{k-j} G]_{\mathbf{x}\mathbf{y}}}{N} = \mathcal{O}_{\prec} \left(N^{k/2} \frac{\rho}{N\eta} \left(1 + \frac{\psi_k^{\text{iso}}}{\sqrt{N\eta\rho}} \right) \right).
\end{aligned}$$

For the third and fourth line of (3.54) we distribute the derivatives according to the product rule to estimate (with absolute value inside the summation to address both diagonal and off-diagonal terms)

$$\begin{aligned}
& \sum_{j \geq 2} \frac{1}{N^{(j+1)/2}} \sum_{a,b} \left| (\partial_{ab} + \partial_{ba})^j \left(x_a [(GA)^k G]_{b\mathbf{y}} [(GA)^k G - m^{k+1} A^k]_{\mathbf{x}\mathbf{y}}^{p-1} \right) \right| \\
&\leq \sum_{\substack{\sum j \geq 2 \\ 1 \leq |j| \leq p}} \Lambda_k(\mathbf{j}) |[(GA)^k G - m^{k+1} A^k]_{\mathbf{x}\mathbf{y}}|^{p-|\mathbf{j}|}
\end{aligned} \tag{3.56}$$

where

$$\Lambda_k(\mathbf{j}) := N^{(n-\sum j)/2} \sum_{ab} \left| \left((\partial_{ab} + \partial_{ba})^{j_0} \frac{x_a [(GA)^k G]_{b\mathbf{y}}}{\sqrt{N}} \right) \prod_{i=1}^n \left((\partial_{ab} + \partial_{ba})^{j_i} \frac{[(GA)^k G]_{\mathbf{x}\mathbf{y}}}{\sqrt{N}} \right) \right| \tag{3.57}$$

and the summation in (3.56) is performed over all $\mathbf{j} = (j_0, \dots, j_n) \in \mathbf{N}_0^n$ with $j_0 \geq 0, j_1, \dots, j_n \geq 1$ and $|\mathbf{j}| = n+1$. Recall that $\sum \mathbf{j} = j_0 + j_1 + j_2 + \dots + j_n$.

Lemma 3.8. *For any admissible \mathbf{j} in the summation (3.56) it holds that*

$$\Lambda_k(\mathbf{j}) \prec \left(N^{k/2} \sqrt{\frac{\rho}{N\eta}} \Phi_k \right)^{|\mathbf{j}|}. \quad (3.58)$$

By combining Lemmas 3.7 and 3.8 and (3.55) and (3.56) and (3.57) we obtain

$$|\langle \mathbf{x}, [(GA)^k - m^{k+1} A^k] \mathbf{y} \rangle| \prec \mathcal{E}_k^{\text{iso}} + N^{k/2} \sqrt{\frac{\rho}{N\eta}} \left(\Phi_k + \frac{1}{N\eta} \sum_{j=0}^{2k} \sqrt{\psi_j^{\text{iso}} \psi_{2k-j}^{\text{iso}}} + \frac{\psi_k^{\text{iso}}}{N\eta} \right), \quad (3.59)$$

concluding the proof of (3.9).

Proof of Lemma 3.8. We recall the notations Ω_k, Φ_k from (3.10). For a naive bound we estimate all but the first factor trivially in (3.57) with

$$\left| (\partial_{ab} + \partial_{ba})^{j_i} \frac{[(GA)^k G]_{\mathbf{x}\mathbf{y}}}{\sqrt{N}} \right| \prec \frac{N^{k/2}}{N^{1/2}} \Omega_k. \quad (3.60)$$

Note that the estimate is independent of the number of derivatives. For the first factor in (3.57) we estimate, after performing the derivatives, all but the last $[(GA)^{k_i} G]$ -factor (involving \mathbf{y}) trivially by (3.37) as

$$\left| (\partial_{ab} + \partial_{ba})^{j_0} \frac{x_a [(GA)^k G]_{b\mathbf{y}}}{\sqrt{N}} \right| \prec \sum_{j=0}^k N^{(k-j)/2} \Omega_{k-j} |x_a| \frac{|[(GA)^j G]_{a\mathbf{y}}| + |[(GA)^j G]_{b\mathbf{y}}|}{\sqrt{N}}. \quad (3.61)$$

By combining (3.60) and (3.61) and the Schwarz-inequality

$$\begin{aligned} \sum_{ab} |x_a| \frac{|[(GA)^j G]_{a\mathbf{y}}| + |[(GA)^j G]_{b\mathbf{y}}|}{\sqrt{N}} &\leq \sqrt{N} \|\mathbf{x}\| \sqrt{|[(G^*A)^j G^* G (AG)^j]_{\mathbf{y}\mathbf{y}}} \\ &\prec N^{j/2+1} \sqrt{\frac{\rho}{N\eta}} \left(1 + \frac{\psi_{2j}^{\text{iso}}}{\sqrt{N\eta\rho}} \right)^{1/2} \end{aligned} \quad (3.62)$$

we conclude

$$\Lambda_k(\mathbf{j}) \prec N^{(n-\sum \mathbf{j})/2+1} N^{k/2} \sqrt{\frac{\rho}{N\eta}} \Phi_k \left(N^{k/2} \frac{1}{\sqrt{N}} \Omega_k \right)^{|\mathbf{j}|-1}, \quad (3.63)$$

which implies (3.58) in the case when $\sum \mathbf{j} \geq n+2$ using that $\Omega_k \leq \Phi_k$ and $\rho/\eta \gtrsim 1$. It thus only remains to consider the cases $\sum \mathbf{j} = n$ and $\sum \mathbf{j} = n+1$.

If $\sum \mathbf{j} = n$, then $n \geq 2$ and $j_0 = 0, j_1 = j_2 = \dots = 1$. By estimating the j_2, j_3, \dots factors in (3.57) using (3.60) we then bound

$$\begin{aligned} \Lambda_k(\mathbf{j}) &\prec \left(N^{k/2} \frac{\Omega_k}{\sqrt{N}} \right)^{|\mathbf{j}|-2} \sum_{ab} \frac{|x_a| |[(GA)^k G]_{b\mathbf{y}}|}{\sqrt{N}} \sum_{j=0}^k \frac{|[(GA)^j G]_{\mathbf{x}a}| |[(GA)^{k-j} G]_{b\mathbf{y}}|}{\sqrt{N}} \\ &\lesssim \left(N^{k/2} \frac{\Omega_k}{\sqrt{N}} \right)^{|\mathbf{j}|-2} \frac{\sqrt{|[(G^*A)^k G^* G (AG)^k]_{\mathbf{y}\mathbf{y}}}}{\sqrt{N}} \sum_{j=0}^k \frac{\sqrt{|[(G^*A)^j G^* G (AG)^j]_{\mathbf{y}\mathbf{y}}| |[(G^*A)^{k-j} G^* G (AG)^{k-j}]_{\mathbf{x}\mathbf{x}}}}{\sqrt{N}} \\ &\prec \left(N^{k/2} \frac{\Omega_k}{\sqrt{N}} \right)^{|\mathbf{j}|-2} \left(N^{k/2} \sqrt{\frac{\rho}{N\eta}} \right)^2 \sqrt{\frac{\rho}{\eta}} \Phi_k \sum_{j=0}^k \left(1 + \frac{\psi_{2j}^{\text{iso}}}{\sqrt{N\eta\rho}} \right)^{1/2} \left(1 + \frac{\psi_{2(k-j)}^{\text{iso}}}{\sqrt{N\eta\rho}} \right)^{1/2} \lesssim \left(N^{k/2} \sqrt{\frac{\rho}{N\eta}} \Phi_k \right)^{|\mathbf{j}|} \end{aligned} \quad (3.64)$$

using $|\mathbf{j}| \geq 3$ and $\Omega_k \leq \Phi_k, 1 \lesssim \rho/\eta$ in the last step.

Finally, if $\sum \mathbf{j} = n+1$, then $n \geq 1$ by admissibility and either $j_0 = 0$ or $j_1 = 1$. In the first case we estimate the j_2, j_3, \dots factors in (3.57) using (3.60), and all but the first $[(GA)^j G]_{\mathbf{x}}$ in the j_1 -factor after differentiation trivially to obtain

$$\begin{aligned} \Lambda_k(\mathbf{j}) &\prec N^{-1/2} \left(N^{k/2} \frac{\Omega_k}{\sqrt{N}} \right)^{|\mathbf{j}|-2} \sum_{ab} \frac{|x_a| |[(GA)^k G]_{b\mathbf{y}}|}{\sqrt{N}} \sum_{j=0}^k N^{(k-j)/2} \Omega_{k-j} \frac{|[(GA)^j G]_{\mathbf{x}a}| + |[(GA)^j G]_{\mathbf{x}b}|}{\sqrt{N}} \\ &\prec \left(N^{k/2} \frac{\Omega_k}{\sqrt{N}} \right)^{|\mathbf{j}|-2} \left(N^{k/2} \sqrt{\frac{\rho}{N\eta}} \Phi_k \right)^2, \end{aligned} \quad (3.65)$$

again using a Schwarz inequality. Finally, in the $j_1 = 1$ case we estimate two j_0 -factor using (3.61), the j_2, j_3, \dots factors trivially, and to bound

$$\begin{aligned} \Lambda_k(\mathbf{j}) &\prec N^{-1/2} \left(N^{k/2} \frac{\Omega_k}{\sqrt{N}} \right)^{|\mathbf{j}|-2} \\ &\quad \times \sum_{ab} \sum_{j,l=0}^k N^{(k-l)/2} \Omega_{k-l}|x_a| \frac{|[(GA)^l G]_{ay}| + |[(GA)^l G]_{by}|}{\sqrt{N}} \frac{|[(GA)^j G]_{xa}| |[(GA)^{k-j} G]_{by}|}{\sqrt{N}} \\ &\prec \left(N^{k/2} \frac{\Omega_k}{\sqrt{N}} \right)^{|\mathbf{j}|-2} \left(N^{k/2} \sqrt{\frac{\rho}{N\eta}} \Phi_k \right)^2, \end{aligned} \quad (3.66)$$

where we used the trivial bound for the $|[(GA)^j G]_{xa}|$ in order to estimate the remaining terms by a Schwarz inequality. This completes the proof of the lemma. \square

3.3. Reduction inequalities and bootstrap. In this section we prove the reduction inequalities in Lemma 3.3 and conclude the proof of our main result Theorem 2.2 showing that $\psi_k^{\text{av}/\text{iso}} \lesssim 1$ for any $k \geq 0$.

Proof of Lemma 3.3. The proof of this proposition is very similar to [20, Lemma 3.6], we thus present only the proof in the averaged case. Additionally, we only prove the case when k is even, if k is odd the proof is completely analogous.

Define $T = T_k := A(GA)^{k/2-1}$, write $(GA)^{2k} = GTGTGTGT$ and use the spectral theorem for these four intermediate resolvents. Then, using that $|m_i| \lesssim 1$ and that $|\langle A^k \rangle| \lesssim N^{k/2-1} \langle |A|^2 \rangle^{k/2}$, after a Schwarz inequality in the third line, we conclude that

$$\begin{aligned} \Psi_{2k}^{\text{av}} &= \frac{N^{(3-2k)/2} \eta^{1/2}}{\rho^{1/2} \langle |A|^2 \rangle^k} |\langle (GA)^{2k} - m_1 \dots m_{2k} A^{2k} \rangle| \\ &\lesssim \sqrt{\frac{N\eta}{\rho}} + \frac{N^{(3-2k)/2} \eta^{1/2}}{N \rho^{1/2} \langle |A|^2 \rangle^k} \left| \sum_{ijml} \frac{\langle \mathbf{u}_i, T \mathbf{u}_j \rangle \langle \mathbf{u}_j, T \mathbf{u}_m \rangle \langle \mathbf{u}_m, T \mathbf{u}_l \rangle \langle \mathbf{u}_l, T \mathbf{u}_i \rangle}{(\lambda_i - z_1)(\lambda_j - z_{k/2+1})(\lambda_m - z_{k+1})(\lambda_l - z_{3k/2+1})} \right| \\ &\lesssim \sqrt{\frac{N\eta}{\rho}} + \frac{N^{(3-2k)/2+1} \eta^{1/2}}{\rho^{1/2} \langle |A|^2 \rangle^k} \langle |G|A(GA)^{k/2-1} |G|A(G^*A)^{k/2-1} \rangle \langle |G|A(GA)^{k/2-1} |G|A(G^*A)^{k/2-1} \rangle \\ &\lesssim \sqrt{\frac{N\eta}{\rho}} + \frac{N^{(3-2k)/2+1} \eta^{1/2}}{\rho^{1/2} \langle |A|^2 \rangle^k} \left(N^{k/2-1} \langle |A|^2 \rangle^{k/2} + \frac{\rho^{1/2} \langle |A|^2 \rangle^{k/2}}{N^{(3-k)/2} \eta^{1/2}} \psi_k^{\text{av}} \right)^2 \\ &\lesssim \sqrt{\frac{N\eta}{\rho}} + \sqrt{\frac{\rho}{N\eta}} (\psi_k^{\text{av}})^2. \end{aligned}$$

We remark that to bound $\langle |G|A(GA)^{k/2-1} |G|A(G^*A)^{k/2-1} \rangle$ in terms of ψ_k^{av} we used (ii) of Lemma 3.1 together with $G^*(z) = G(\bar{z})$. \square

We are now ready to conclude the proof of our main result.

Proof of Theorem 2.2. The proof repeatedly uses a simple argument called *iteration*. By this we mean the following observation: whenever we know that $X \prec x$ implies

$$X \prec A + \frac{x}{B} + x^{1-\alpha} C^\alpha, \quad (3.67)$$

for some constants $B \geq N^\delta$, $A, C > 0$, and exponent $0 < \alpha < 1$, and we know that $X \prec N^D$ initially (here δ, α and D are N -independent positive constants, other quantities may depend on N) then we also know that $X \prec x$ implies

$$X \prec A + C. \quad (3.68)$$

The proof is simply to iterate (3.67) finitely many times (depending only on δ, α and D). The fact that $\Psi_k^{\text{av}/\text{iso}} \prec N^D$ follows by a simple norm bound on the resolvents and A , so the condition $X \prec N^D$ is always satisfied in our applications.

By the standard single resolvent local laws in (2.4) we know that $\psi_0^{\text{av}} = \psi_0^{\text{iso}} = 1$. Using the master inequalities in Proposition 3.2 and the reduction bounds from Lemma 3.3, in the first step we will show that $\Psi_k^{\text{av}/\text{iso}} \prec \rho^{-1}$ for any $k \geq 1$ as an a priori bound. Then, in the second step, we feed this bound into the tandem of the master inequalities and

the reduction bounds to improve the estimate to $\Psi_k^{\text{av/iso}} \prec 1$. The first step is the critical stage of the proof, here we need to show that our bounds are sufficiently strong to close the hierarchy of our estimates to yield a better bound on $\Psi_k^{\text{av/iso}}$ than the trivial $\Psi_k^{\text{av/iso}} \leq N^{k/2}\eta^{-k-1}$ estimate obtained by using the norm bounds $\|G\| \leq \eta^{-1}$ and $\|A\| \leq N^{1/2}$. Once some improvement is achieved, it can be relatively easily iterated.

The proof of $\Psi_k^{\text{av/iso}} \prec \rho^{-1}$ proceeds by a step-two induction, we first prove that $\Psi_k^{\text{av/iso}} \prec \rho^{-1}$ for $k = 1, 2$ and then show that if $\Psi_n^{\text{av/iso}} \prec \rho^{-1}$ holds for all $n \leq k - 2$, for some $k \geq 4$, then it also holds for $\Psi_{k-1}^{\text{av/iso}}$ and $\Psi_k^{\text{av/iso}}$.

Using (3.8)–(3.9) we have

$$\begin{aligned}\Psi_1^{\text{av}} &\prec 1 + \frac{\sqrt{\psi_2^{\text{iso}}}}{(N\eta\rho)^{1/4}} + \frac{\sqrt{\psi_2^{\text{av}}}}{(N\eta\rho)^{1/4}} + (\psi_1^{\text{iso}})^{2/3} + \psi_1^{\text{iso}} \sqrt{\frac{\rho}{N\eta}} + \frac{\psi_1^{\text{av}}}{\sqrt{N\eta}} \\ \Psi_1^{\text{iso}} &\prec 1 + \frac{\sqrt{\psi_2^{\text{iso}}}}{(N\eta\rho)^{1/4}} + \frac{\psi_1^{\text{av}}}{N\eta} + \psi_1^{\text{iso}} \sqrt{\frac{\rho}{N\eta}} + \frac{\psi_1^{\text{iso}}}{N\eta}\end{aligned}\tag{3.69}$$

for $k = 1$, using

$$\Phi_1 \lesssim 1 + \psi_1^{\text{iso}} \sqrt{\frac{\rho}{N\eta}} + \frac{\sqrt{\psi_2^{\text{iso}}}}{(N\eta\rho)^{1/4}}.$$

Similarly, for $k = 2$, using that $\Omega_1 \leq \Phi_1$, and estimating explicitly

$$\Phi_2 \lesssim 1 + (\psi_1^{\text{iso}})^2 \frac{\rho}{N\eta} + \frac{\psi_2^{\text{iso}}}{(N\eta\rho)^{1/2}} + \frac{(\psi_4^{\text{iso}})^{1/2}}{(N\eta\rho)^{1/4}}$$

by Schwarz inequalities and plugging it into (3.8)–(3.9) we have

$$\begin{aligned}\Psi_2^{\text{av}} &\prec 1 + \psi_1^{\text{av}} + \frac{\psi_2^{\text{iso}}}{(N\eta\rho)^{1/12}} + \frac{\sqrt{\psi_4^{\text{iso}}}}{(N\eta\rho)^{1/4}} + \frac{\sqrt{\psi_4^{\text{av}}}}{(N\eta\rho)^{1/4}} + (\psi_2^{\text{iso}})^{2/3} + \sqrt{\psi_1^{\text{iso}}\psi_2^{\text{iso}}} \\ &\quad + \frac{(\psi_2^{\text{iso}})^{3/4}(\psi_1^{\text{iso}})^{1/2}}{(N\eta\rho)^{1/8}} + \frac{\rho^{1/3}(\psi_2^{\text{iso}})^{2/3}(\psi_1^{\text{iso}})^{1/3}}{(N\eta\rho)^{1/6}} + \frac{\sqrt{\rho}(\psi_1^{\text{av}})^2}{(N\eta)^{3/2}} + (\psi_1^{\text{iso}})^2 \frac{\rho}{N\eta} \\ &\quad + (\psi_1^{\text{iso}})^{3/2} \sqrt{\frac{\rho}{N\eta}} + \frac{\rho^{1/2}\psi_1^{\text{iso}}(\psi_2^{\text{iso}})^{1/2}}{(N\eta\rho)^{1/4}} + \frac{\psi_2^{\text{av}}}{\sqrt{N\eta}}, \\ \Psi_2^{\text{iso}} &\prec 1 + \psi_1^{\text{iso}} + \frac{\psi_1^{\text{av}}}{N\eta} + \frac{\psi_2^{\text{iso}} + \psi_2^{\text{av}}}{(N\eta\rho)^{1/2}} + \frac{\sqrt{\psi_4^{\text{iso}}}}{(N\eta\rho)^{1/4}} + \frac{\sqrt{\psi_1^{\text{iso}}\psi_3^{\text{iso}}}}{N\eta} + \frac{\psi_1^{\text{av}}\psi_1^{\text{iso}}}{(N\eta)^{3/2}} + (\psi_1^{\text{iso}})^2 \frac{\rho}{N\eta} + \frac{\psi_2^{\text{iso}}}{N\eta}.\end{aligned}\tag{3.70}$$

In these estimates we frequently used that $\psi_k^{\text{av/iso}} \geq 1$, $\rho \lesssim 1$, $\rho/N\eta \leq 1$ and $N\eta\rho \geq 1$ to simplify the formulas.

By (3.69)–(3.70), using iteration for the sum $\Psi_1^{\text{av}} + \Psi_1^{\text{iso}}$, we readily conclude

$$\Psi_1^{\text{av}} + \Psi_1^{\text{iso}} \prec 1 + \frac{\sqrt{\psi_2^{\text{iso}}}}{(N\eta\rho)^{1/4}} + \frac{\sqrt{\psi_2^{\text{av}}}}{(N\eta\rho)^{1/4}}.\tag{3.71}$$

Note that since (3.71) holds uniformly in the hidden parameters $A, z, \mathbf{x}, \mathbf{y}$ in $\Psi_1^{\text{av/iso}}$, this bound serves as an upper bound on $\psi_1^{\text{av}} + \psi_1^{\text{iso}}$ (in the sequel, we will frequently use an already proven upper bound on Ψ_k as an effective upper bound on ψ_k in the next steps without explicitly mentioning it). Next, using this upper bound together with an iteration for $\Psi_2^{\text{av}} + \Psi_2^{\text{iso}}$ we have from (3.70)

$$\Psi_2^{\text{av}} + \Psi_2^{\text{iso}} \prec 1 + \frac{\sqrt{\psi_4^{\text{iso}}}}{(N\eta\rho)^{1/4}} + \frac{\sqrt{\psi_4^{\text{av}}}}{(N\eta\rho)^{1/4}} + \frac{\sqrt{\psi_1^{\text{iso}}\psi_3^{\text{iso}}}}{N\eta},\tag{3.72}$$

again after several simplifications by Young's inequality and the basic inequalities $\psi_k^{\text{av/iso}} \geq 1$, $\rho \lesssim 1$ and $N\eta\rho \geq 1$.

We now apply the reduction inequalities from Lemma 3.3 in the form

$$\begin{aligned}\Psi_4^{\text{av}} &\prec \sqrt{\frac{N\eta}{\rho}} + \sqrt{\frac{\rho}{N\eta}} (\psi_2^{\text{av}})^2 \\ \Psi_4^{\text{iso}} &\prec \sqrt{\frac{N\eta}{\rho}} + \psi_2^{\text{av}} + \psi_2^{\text{iso}} + \sqrt{\frac{\rho}{N\eta}} \psi_2^{\text{av}} \psi_2^{\text{iso}} \\ \Psi_3^{\text{iso}} &\prec \sqrt{\frac{N\eta}{\rho}} + \left(\frac{N\eta}{\rho}\right)^{1/4} \sqrt{\psi_2^{\text{av}}} + \psi_2^{\text{iso}} + \left(\frac{\rho}{N\eta}\right)^{1/4} \psi_2^{\text{iso}} \sqrt{\psi_2^{\text{av}}},\end{aligned}\tag{3.73}$$

where the first inequality was inserted already into the right hand side of (3.12) to get the second inequality in (3.73).

Then, inserting (3.73) and (3.71) into (3.72) and using iteration, we conclude

$$\Psi_2^{\text{av}} + \Psi_2^{\text{iso}} \prec \frac{1}{\sqrt{\rho}} + \frac{\sqrt{\psi_2^{\text{iso}}} + \sqrt{\psi_2^{\text{av}}}}{(N\eta\rho)^{1/4}} + \frac{\psi_2^{\text{av}} + \psi_2^{\text{iso}}}{(N\eta)^{1/2}}\tag{3.74}$$

which together with (3.71) implies

$$\Psi_1^{\text{iso}} + \Psi_1^{\text{av}} \prec \rho^{-1/4}, \quad \Psi_2^{\text{iso}} + \Psi_2^{\text{av}} \prec \rho^{-1/2}.\tag{3.75}$$

We now proceed with a step-two induction on k . The initial step of the induction is (3.75). Fix an even $k \geq 4$ and assume that

$$\Psi_n^{\text{av/iso}} \prec \frac{1}{\rho} \quad \text{for } n \leq k-2.\tag{3.76}$$

First of all we notice that using the reduction inequality (3.11) for $k/2$ we obtain (assuming that k is divisible by four)

$$\Psi_k^{\text{av}} \prec \sqrt{\frac{N\eta}{\rho}} + \sqrt{\frac{\rho}{N\eta}} (\psi_{k/2}^{\text{av}})^2 \prec \sqrt{\frac{N\eta}{\rho}},\tag{3.77}$$

where in the last inequality we used (3.76). If k is not divisible by four we conclude the same bound using the second inequality in (3.11) instead of the first one. Next, using again the reduction inequality (3.11) and the induction hypothesis (3.76) we obtain

$$\Psi_{2n}^{\text{av}} \prec \sqrt{\frac{N\eta}{\rho}} \quad \text{for } n \leq k.\tag{3.78}$$

We point out that for $n = k$ we used

$$\Psi_{2k}^{\text{av}} \prec \sqrt{\frac{N\eta}{\rho}} + \sqrt{\frac{\rho}{N\eta}} (\psi_k^{\text{av}})^2 \prec \sqrt{\frac{N\eta}{\rho}},$$

where in the last inequality we used (3.77), and a similar bound for $n = k-1$. Similarly, in the isotropic case, using (3.12), we obtain

$$\Psi_{k+j}^{\text{iso}} \prec \sqrt{\frac{N\eta}{\rho}} + \left(\frac{N\eta}{\rho}\right)^{1/4} \sqrt{\psi_{2j}^{\text{av}}} + \psi_k^{\text{iso}} + \left(\frac{\rho}{N\eta}\right)^{1/4} \sqrt{\psi_{2j}^{\text{av}} \psi_k^{\text{iso}}} \prec \sqrt{\frac{N\eta}{\rho}} \quad \text{for } j \leq k,\tag{3.79}$$

where in the last inequality we used (3.78) and that by (3.12) it follows

$$\psi_k^{\text{iso}} = \psi_{(k-2)+2}^{\text{iso}} \prec \sqrt{\frac{N\eta}{\rho}} + \left(\frac{N\eta}{\rho}\right)^{1/4} \sqrt{\psi_4^{\text{av}}} + \psi_{k-2}^{\text{iso}} + \left(\frac{\rho}{N\eta}\right)^{1/4} \sqrt{\psi_4^{\text{av}} \psi_{k-2}^{\text{iso}}} \prec \sqrt{\frac{N\eta}{\rho}}.\tag{3.80}$$

We point out that similarly, writing $k-1 = (k-2) + 1$, we also conclude that $\psi_{k-1}^{\text{iso}} \prec \sqrt{N\eta/\rho}$. Then, by using (3.78) and the induction hypothesis (3.76) in the definition of Φ_n in (3.10), we readily conclude that

$$\Phi_n \prec \frac{1}{\rho} \quad \text{for } n \leq k.\tag{3.81}$$

Furthermore, we notice that, by using (3.76) and (3.80) (as well as the similar bound for ψ_{k-1}^{iso} stated below it) in the definition of Ω_n in (3.10), we also have

$$\Omega_n \prec 1 \quad \text{for } n \leq k.\tag{3.82}$$

We are now ready to consider the master inequalities for $\Psi_{k-1}^{\text{av/iso}}$ and $\Psi_k^{\text{av/iso}}$. Using *iteration* as in (3.67)–(3.68), together with the bounds (3.78), (3.79), (3.81)–(3.82), by (3.8)–(3.9), we obtain (recall that $N\eta\rho \gg 1$ and $\rho \lesssim 1$)

$$\begin{aligned}\Psi_{k-1}^{\text{av}} &\prec \frac{1}{\rho} + (\psi_{k-1}^{\text{iso}})^{2/3} \rho^{-1/3} + \rho^{-1/2} \sqrt{\psi_{k-1}^{\text{iso}}} \\ \Psi_{k-1}^{\text{iso}} &\prec \frac{1}{\rho} + \frac{\psi_{k-1}^{\text{av}}}{N\eta},\end{aligned}\tag{3.83}$$

and

$$\begin{aligned}\Psi_k^{\text{av}} &\prec \frac{1}{\rho} + (\psi_k^{\text{iso}})^{2/3} \rho^{-1/3} + \rho^{-1/2} (\sqrt{\psi_k^{\text{iso}}} + \sqrt{\psi_{k-1}^{\text{iso}}}) + \sqrt{\psi_{k-1}^{\text{iso}} \psi_k^{\text{iso}}} + \frac{\psi_{k-1}^{\text{av}}}{N\eta} \\ \Psi_k^{\text{iso}} &\prec \frac{1}{\rho} + \frac{\psi_k^{\text{av}} + \psi_{k-1}^{\text{av}}}{N\eta}.\end{aligned}\tag{3.84}$$

Then, using *iteration* in (3.83) for $\Psi_{k-1}^{\text{av}} + \Psi_{k-1}^{\text{iso}}$, we immediately conclude that $\Psi_{k-1}^{\text{av/iso}} \prec \rho^{-1}$. Finally, plugging this information into (3.84), and using iteration once again for $\Psi_k^{\text{av}} + \Psi_k^{\text{iso}}$ we conclude that $\Psi_k^{\text{av/iso}} \prec \rho^{-1}$ as well. This completes the step-two induction hence the first and the pivotal step of the proof.

In the second step we improve $\Psi_k^{\text{av/iso}} \prec \rho^{-1}$ to $\Psi_k^{\text{av/iso}} \prec 1$ for all k . By plugging the bound $\Psi_k^{\text{av/iso}} \prec \rho^{-1}$ into the master inequalities in Proposition 3.2 and noticing that $\Phi_k \leq 1 + \rho^{-1}(N\eta\rho)^{-1/4}$, we directly conclude that

$$\Psi_k^{\text{av/iso}} \prec 1 + \frac{\rho^{-1}}{(N\eta\rho)^{1/12}}\tag{3.85}$$

for any $k \geq 0$. We point that that the exponent 1/12 comes from the fifth term in the first line of (3.8). Now we can use this improved inequality by plugging it again in the master inequalities to achieve

$$\Psi_k^{\text{av/iso}} \prec 1 + \rho^{-1} \left(\frac{1}{(N\eta\rho)^{1/12}} \right)^2,\tag{3.86}$$

and so on. Recalling the assumption that $N\eta\rho \geq N^\epsilon$, we need to iterate this process finitely many times (depending on k, ξ, K, ϵ) to achieve $\Psi_k^{\text{av/iso}} \prec 1$ also in the second regime. This concludes the proof of the theorem. \square

4. STOCHASTIC EIGENSTATE EQUATION AND PROOF OF THEOREM 2.8

Armed with the new local law (Theorem 2.2) and its direct corollary on the eigenvector overlaps (Theorem 2.6), the rest of the proof of Theorem 2.8 is very similar to the proof of [19, Theorem 2.2], which is presented in [19, Sections 3-4]. For this reason we only explain the differences and refer to [19] for a fully detailed proof. We mention that the proof in [19] heavily relies on the theory of the stochastic eigenstate equation initiated in [11] and then further developed in [13, 45].

Similarly to [19, Sections 3-4], we present the proof only in the real case (the complex case is completely analogous and so omitted). We will prove Theorem 2.8 dynamically, i.e. we consider the Dyson Brownian motion (DBM) with initial condition W and we will show that the overlaps of the eigenvectors have Gaussian fluctuations after a time t slightly bigger than N^{-1} . With a separate argument then in Appendix B we show that the (small) Gaussian component added along the DBM flow can be removed at the price of a negligible error.

More precisely, we consider the matrix flow

$$dW_t = \frac{d\tilde{B}_t}{\sqrt{N}}, \quad W_0 = W,\tag{4.1}$$

where \tilde{B}_t is a standard real symmetric matrix Brownian motion (see e.g. [11, Definition 2.1]). We denote the resolvent of W_t by $G = G_t(z) := (W_t - z)^{-1}$, for $z \in \mathbf{C} \setminus \mathbf{R}$. It is well known that in the limit $N \rightarrow \infty$ the resolvent $G_t(z) := (W_t - z)^{-1}$, for $z \in \mathbf{C} \setminus \mathbf{R}$, becomes approximately deterministic, and that its deterministic approximation is given by the scalar matrix $m_t \cdot I$. The function $m_t = m_t(z)$ is the unique solution of the complex Burgers equation

$$\partial_t m_t(z) = -m_t \partial_z m_t(z), \quad m_0(z) = m(z),\tag{4.2}$$

with initial condition $m(z) = m_{\text{sc}}(z)$ being the Stieltjes transform of the semicircular law. Denote $\rho_t = \rho_t(z) := \pi^{-1} \Im m_t(z)$, then it is easy to see that $\rho_t(x + i0)$ is a rescaling of $\rho_0 = \rho_{\text{sc}}$ by a factor $1 + t$. In fact, W_t is a Wigner matrix itself, with a normalization $\mathbf{E} |(W_t)_{ab}|^2 = N^{-1}(1 + t)$ with a Gaussian component.

Denote by $\lambda_1(t) \leq \lambda_2(t) \leq \dots \leq \lambda_N(t)$ the eigenvalues of W_t , and let $\{\mathbf{u}_i(t)\}_{i \in [N]}$ be the corresponding eigenvectors. Then, it is known [11, Theorem 2.3] that $\lambda_i = \lambda_i(t)$, $\mathbf{u}_i = \mathbf{u}_i(t)$ are the unique strong solutions of the following system of stochastic differential equations:

$$d\lambda_i = \frac{dB_{ii}}{\sqrt{N}} + \frac{1}{N} \sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} dt \quad (4.3)$$

$$d\mathbf{u}_i = \frac{1}{\sqrt{N}} \sum_{j \neq i} \frac{dB_{ij}}{\lambda_i - \lambda_j} \mathbf{u}_j - \frac{1}{2N} \sum_{j \neq i} \frac{\mathbf{u}_i}{(\lambda_i - \lambda_j)^2} dt, \quad (4.4)$$

where $B_t = (B_{ij})_{i,j \in [N]}$ is a standard real symmetric matrix Brownian motion (see e.g. [11, Definition 2.1]).

Note that the flow for the diagonal overlaps $\langle \mathbf{u}_i, A\mathbf{u}_i \rangle$, by (4.4), naturally also depends on the off-diagonal overlap $\langle \mathbf{u}_i, A\mathbf{u}_j \rangle$. Hence, even if we are only interested in diagonal overlaps, our analysis must also handle off-diagonal overlaps. In particular, this implies that there is no closed differential equation for only diagonal or only off-diagonal overlaps. However, in [13] Bourgade, Yau, and Yin proved that the *perfect matching observable* $f_{\lambda,t}$, which is presented in (4.6) below, satisfies a parabolic PDE (see (4.10) below). We now describe how the observable $f_{\lambda,t}$ is constructed.

4.1. Perfect matching observables. Without loss of generality for the rest of the paper we assume that A is traceless, $\langle A \rangle = 0$, i.e. $A = \dot{A}$. We introduce the short-hand notation for the *eigenvector overlaps*

$$p_{ij} = p_{ij}(t) := \langle \mathbf{u}_i(t), A\mathbf{u}_j(t) \rangle, \quad i, j \in [N]. \quad (4.5)$$

To compute the moments, we will consider monomials of eigenvector overlaps of the form $\prod_k p_{i_k j_k}$ where each index occurs an even number of times. We start by introducing a particle picture and a certain graph that encode such monomials: each particle on the set of integers $[N]$ corresponds to two occurrences of an index i in the monomial product. This particle picture was introduced in [11] and heavily used also in [13, 45]. Each particle configuration is encoded by a function $\boldsymbol{\eta} : [N] \rightarrow \mathbf{N}_0$, where $\eta_j := \boldsymbol{\eta}(j)$ denotes the number of particles at the site j , and $n(\boldsymbol{\eta}) := \sum_j \eta_j = n$ is the total number of particles. We denote the space of n -particle configurations by Ω^n . Moreover, for any index pair $i \neq j \in [N]$, we define $\boldsymbol{\eta}^{ij}$ to be the configuration obtained moving a particle from the site i to the site j , if there is no particle in i then $\boldsymbol{\eta}^{ij} := \boldsymbol{\eta}$.

We now define the *perfect matching observable* (introduced in [13]) for any given configuration $\boldsymbol{\eta}$:

$$f_{\lambda,t}(\boldsymbol{\eta}) := \frac{N^{n/2}}{[2\langle A^2 \rangle]^{n/2}} \frac{1}{(n-1)!!} \frac{1}{\mathcal{M}(\boldsymbol{\eta})} \mathbf{E} \left[\sum_{G \in \mathcal{G}_n} P(G) \Big| \boldsymbol{\lambda} \right], \quad \mathcal{M}(\boldsymbol{\eta}) := \prod_{i=1}^N (2\eta_i - 1)!!, \quad (4.6)$$

with n being the number of particles in the configuration $\boldsymbol{\eta}$. Here \mathcal{G}_n denotes the set of perfect matchings on the complete graph with vertex set

$$\mathcal{V}_n := \{(i, a) : 1 \leq i \leq n, 1 \leq a \leq 2\eta_i\},$$

and

$$P(G) := \prod_{e \in \mathcal{E}(G)} p(e), \quad p(e) := p_{i_1 i_2}, \quad (4.7)$$

where $e = \{(i_1, a_1), (i_2, a_2)\} \in \mathcal{V}_n^2$, and $\mathcal{E}(G)$ denotes the edges of G . Note that in (4.6) we took the conditioning on the entire flow of eigenvalues, $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}(t)\}_{t \in [0, T]}$ for some fixed $T > 0$. From now on we will always assume that $T \ll 1$ (even if not stated explicitly).

We always assume that the entire eigenvalue trajectory $\{\boldsymbol{\lambda}(t)\}_{t \in [0, T]}$ satisfies the usual rigidity estimate asserting that the eigenvalues are very close to the deterministic quantiles of the semicircle law with very high probability. To formalize it, we define

$$\tilde{\Omega} = \tilde{\Omega}_\xi := \left\{ \sup_{0 \leq t \leq T} \max_{i \in [N]} N^{2/3} i^{1/3} |\lambda_i(t) - \gamma_i(t)| \leq N^\xi \right\} \quad (4.8)$$

for any $\xi > 0$, where $\widehat{i} := i \wedge (N + 1 - i)$. Here $\gamma_i(t)$ denote the *quantiles* of ρ_t , defined by

$$\int_{-\infty}^{\gamma_i(t)} \rho_t(x) \, dx = \frac{i}{N}, \quad i \in [N], \quad (4.9)$$

where $\rho_t(x) = \frac{1}{2(1+t)\pi} \sqrt{(4(1+t)^2 - x^2)_+}$ is the semicircle law corresponding to W_t . Note that $|\gamma_i(t) - \gamma_i(s)| \lesssim |t - s|$ for any bulk index i and for any $t, s \geq 0$.

The well known rigidity estimate (see e.g. [26, Theorem 7.6] or [31]) asserts that

$$\mathbf{P}(\widetilde{\Omega}_\xi) \geq 1 - C(\xi, D)N^{-D}$$

for any (small) $\xi > 0$ and (large) $D > 0$. This was proven for any fixed t e.g. in [26, Theorem 7.6] or [31], the extension to all t follows by a grid argument together with the fact that $\lambda(t)$ is stochastically $1/2$ -Hölder in t , which follows by Weyl's inequality

$$\|\lambda(t) - \lambda(s)\|_\infty \lesssim \|W_t - W_s\| \stackrel{d}{=} \|W + \sqrt{s}U_1 + \sqrt{t-s}U_2 - W - \sqrt{s}U_1\| \lesssim \sqrt{t-s},$$

with $s \leq t$ and U_1, U_2 being independent GUE/GOE matrices, which are also independent of W .

By [13, Theorem 2.6] we know that the perfect matching observable $f_{\lambda,t}$ is a solution of the following parabolic discrete PDE

$$\partial_t f_{\lambda,t} = \mathcal{B}(t)f_{\lambda,t}, \quad (4.10)$$

$$\mathcal{B}(t)f_{\lambda,t} = \sum_{i \neq j} c_{ij}(t) 2\eta_i(1 + 2\eta_j)(f_{\lambda,t}(\boldsymbol{\eta}^{kl}) - f_{\lambda,t}(\boldsymbol{\eta})). \quad (4.11)$$

where

$$c_{ij}(t) := \frac{1}{N(\lambda_i(t) - \lambda_j(t))^2}. \quad (4.12)$$

Note that the number of particles $n = n(\boldsymbol{\eta})$ is preserved under the flow (4.10). The eigenvalue trajectories are fixed in this proof, hence we will often omit λ from the notation, e.g. we will use $f_t = f_{\lambda,t}$, and so on.

The main technical input in the proof of Theorem 2.8 is the following result (cf. [19, Proposition 3.2]):

Proposition 4.1. *For any $n \in \mathbf{N}$ there exists $c(n) > 0$ such that for any $\epsilon > 0$, and for any $T \geq N^{-1+\epsilon}$ it holds*

$$\sup_{\boldsymbol{\eta}} |f_T(\boldsymbol{\eta}) - \mathbf{1}(n \text{ even})| \lesssim N^{-c(n)}, \quad (4.13)$$

with very high probability, where the supremum is taken over configurations $\boldsymbol{\eta} \in \Omega^n$ supported in the bulk, i.e. such that $\eta_i = 0$ for $i \notin [\delta N, (1 - \delta)N]$, with $\delta > 0$ from Theorem 2.8. The implicit constant in (4.13) depends on n, ϵ, δ .

We are now ready to prove Theorem 2.8.

Proof of Theorem 2.8. Fix $i \in [\delta N, (1 - \delta)N]$, then the convergence in (2.9) follows immediately from (4.13) choosing $\boldsymbol{\eta}$ to be the configuration with $\eta_i = n$ and all other $\eta_j = 0$, together with a standard application of the Green function comparison theorem (GFT), relating the eigenvectors/eigenvalues of W_T to those of W ; see Appendix B where we recall the GFT argument for completeness. We defer the interested reader to [19, Proof of Theorem 2.2] for a more detailed proof. □

4.2. DBM analysis. Since the current DBM analysis of (4.10) heavily relies on [19, Section 4], before starting it we introduce an equivalent representation of (4.6) used in [19] (which itself is based on the particles representation from [45]).

Fix $n \in \mathbf{N}$, and consider configurations $\boldsymbol{\eta} \in \Omega^n$, i.e. such that $\sum_j \eta_j = n$. We now give an equivalent representation of (4.10)–(4.11) which is defined on the $2n$ -dimensional lattice $[N]^{2n}$ instead of configurations of n particles (see [19, Section 4.1] for a more detailed description). Let $\boldsymbol{x} \in [N]^{2n}$ and define the configuration space

$$\Lambda^n := \{\boldsymbol{x} \in [N]^{2n} : n_i(\boldsymbol{x}) \text{ is even for every } i \in [N]\}, \quad (4.14)$$

where

$$n_i(\boldsymbol{x}) := |\{a \in [2n] : x_a = i\}| \quad (4.15)$$

for all $i \in \mathbf{N}$.

The correspondence between these two representations is given by

$$\boldsymbol{\eta} \leftrightarrow \boldsymbol{x} \quad \eta_i = \frac{n_i(\boldsymbol{x})}{2}. \quad (4.16)$$

Note that \boldsymbol{x} uniquely determines $\boldsymbol{\eta}$, but $\boldsymbol{\eta}$ determines only the coordinates of \boldsymbol{x} as a multi-set and not its ordering. Let $\phi: \Lambda^n \rightarrow \Omega^n$, $\phi(\boldsymbol{x}) = \boldsymbol{\eta}$ be the projection from the \boldsymbol{x} -configuration space to the $\boldsymbol{\eta}$ -configuration space using (4.16). We will then always consider functions g on $[N]^{2n}$ that are push-forwards of some function f on Ω^n , $g = f \circ \phi$, i.e. they correspond to functions on the configurations

$$f(\boldsymbol{\eta}) = f(\phi(\boldsymbol{x})) = g(\boldsymbol{x}).$$

In particular g is supported on Λ^n and it is equivariant under permutation of the arguments, i.e. it depends on \boldsymbol{x} only as a multiset. We thus consider the observable

$$g_t(\boldsymbol{x}) = g_{\boldsymbol{\lambda},t}(\boldsymbol{x}) := f_{\boldsymbol{\lambda},t}(\phi(\boldsymbol{x})) \quad (4.17)$$

where $f_{\boldsymbol{\lambda},t}$ was defined in (4.6).

Using the \boldsymbol{x} -representation space, we can now write the flow (4.10)–(4.11) as follows:

$$\partial_t g_t(\boldsymbol{x}) = \mathcal{L}(t)g_t(\boldsymbol{x}) \quad (4.18)$$

$$\mathcal{L}(t) := \sum_{j \neq i} \mathcal{L}_{ij}(t), \quad \mathcal{L}_{ij}(t)g(\boldsymbol{x}) := c_{ij}(t) \frac{n_j(\boldsymbol{x}) + 1}{n_i(\boldsymbol{x}) - 1} \sum_{a \neq b \in [2n]} (g(\boldsymbol{x}_{ab}^{ij}) - g(\boldsymbol{x})), \quad (4.19)$$

where

$$\boldsymbol{x}_{ab}^{ij} := \boldsymbol{x} + \delta_{x_a i} \delta_{x_b i} (j - i) (\boldsymbol{e}_a + \boldsymbol{e}_b), \quad (4.20)$$

with $\boldsymbol{e}_a(c) = \delta_{ac}$, $a, c \in [2n]$. This flow is map on functions defined on $\Lambda^n \subset [N]^{2n}$ and it preserves equivariance.

We now define the scalar product and the natural measure on Λ^n :

$$\langle f, g \rangle_{\Lambda^n} = \langle f, g \rangle_{\Lambda^n, \pi} := \sum_{\boldsymbol{x} \in \Lambda^n} \pi(\boldsymbol{x}) \bar{f}(\boldsymbol{x}) g(\boldsymbol{x}), \quad \pi(\boldsymbol{x}) := \prod_{i=1}^N ((n_i(\boldsymbol{x}) - 1)!)^2, \quad (4.21)$$

as well as the norm on $L^p(\Lambda^n)$:

$$\|f\|_p = \|f\|_{L^p(\Lambda^n, \pi)} := \left(\sum_{\boldsymbol{x} \in \Lambda^n} \pi(\boldsymbol{x}) |f(\boldsymbol{x})|^p \right)^{1/p}. \quad (4.22)$$

By [45, Appendix A.2] it follows that the operator $\mathcal{L} = \mathcal{L}(t)$ is symmetric with respect to the measure π and it is a negative operator on $L^2(\Lambda^n)$ with Dirichlet form

$$D(g) = \langle g, (-\mathcal{L})g \rangle_{\Lambda^n} = \frac{1}{2} \sum_{\boldsymbol{x} \in \Lambda^n} \pi(\boldsymbol{x}) \sum_{i \neq j} c_{ij}(t) \frac{n_j(\boldsymbol{x}) + 1}{n_i(\boldsymbol{x}) - 1} \sum_{a \neq b \in [2n]} |g(\boldsymbol{x}_{ab}^{ij}) - g(\boldsymbol{x})|^2.$$

Let $\mathcal{U}(s, t)$ be the semigroup associated to \mathcal{L} , i.e. for any $0 \leq s \leq t$ it holds

$$\partial_t \mathcal{U}(s, t) = \mathcal{L}(t)\mathcal{U}(s, t), \quad \mathcal{U}(s, s) = I.$$

4.2.1. Short range approximation. Most of our DBM analysis will be completely local, hence we will introduce a short range approximation h_t (see its definition in (4.26) below) of g_t that will be exponentially small evaluated on \boldsymbol{x} 's which are not fully supported in the bulk.

Recall the definition of the quantiles $\gamma_i(0)$ from (4.9), then we define the sets

$$\mathcal{J} = \mathcal{J}_\delta := \{i \in [N] : \gamma_i(0) \in \mathcal{I}_\delta\}, \quad \mathcal{I}_\delta := (-2 + \delta, 2 - \delta), \quad (4.23)$$

which correspond to indices and spectral range in the bulk, respectively. From now on we fix a point $\boldsymbol{y} \in \mathcal{J}$, and an N -dependent parameter K such that $1 \ll K \leq \sqrt{N}$. Next, we define the *averaging operator* as a simple multiplication operator by a “smooth” cut-off function:

$$\text{Av}(K, \boldsymbol{y})h(\boldsymbol{x}) := \text{Av}(\boldsymbol{x}; K, \boldsymbol{y})h(\boldsymbol{x}), \quad \text{Av}(\boldsymbol{x}; K, \boldsymbol{y}) := \frac{1}{K} \sum_{j=K}^{2K-1} \mathbf{1}(\|\boldsymbol{x} - \boldsymbol{y}\|_1 < j), \quad (4.24)$$

with $\|\mathbf{x} - \mathbf{y}\|_1 := \sum_{a=1}^{2n} |x_a - y_a|$. Additionally, fix an integer ℓ with $1 \ll \ell \ll K$, and define the short range coefficients

$$c_{ij}^S(t) := \begin{cases} c_{ij}(t) & \text{if } i, j \in \mathcal{J} \text{ and } |i - j| \leq \ell \\ 0 & \text{otherwise,} \end{cases} \quad (4.25)$$

where $c_{ij}(t)$ is defined in (4.12). The parameter ℓ is the length of the short range interaction.

The short range approximation $h_t = h_t(\mathbf{x})$ of g_t is defined as the unique solution of the parabolic equation

$$\begin{aligned} \partial_t h_t(\mathbf{x}; \ell, K, \mathbf{y}) &= \mathcal{S}(t)h_t(\mathbf{x}; \ell, K, \mathbf{y}) \\ h_0(\mathbf{x}; \ell, K, \mathbf{y}) &= h_0(\mathbf{x}; K, \mathbf{y}) := \text{Av}(\mathbf{x}; K, \mathbf{y})(g_0(\mathbf{x}) - \mathbf{1}(n \text{ even})), \end{aligned} \quad (4.26)$$

where

$$\mathcal{S}(t) := \sum_{j \neq i} \mathcal{S}_{ij}(t), \quad \mathcal{S}_{ij}(t)h(\mathbf{x}) := c_{ij}^S(t) \frac{n_j(\mathbf{x}) + 1}{n_i(\mathbf{x}) - 1} \sum_{a \neq b \in [2n]} (h(\mathbf{x}_{ab}^{ij}) - h(\mathbf{x})). \quad (4.27)$$

Since K, \mathbf{y} and ℓ are fixed for the rest of this section we will often omit them from the notation. We conclude this section defining the transition semigroup $\mathcal{U}_S(s, t) = \mathcal{U}_S(s, t; \ell)$ associated to the short range generator $\mathcal{S}(t)$.

4.2.2. *L²-bound.* By standard finite speed propagation estimates (see [19, Proposition 4.2, Lemmata 4.3–4.4]), we conclude that

Lemma 4.2. *Let $0 \leq s_1 \leq s_2 \leq s_1 + \ell N^{-1}$, and f be a function on Λ^n , then for any $\mathbf{x} \in \Lambda^n$ supported on \mathcal{J} it holds*

$$\left| (\mathcal{U}(s_1, s_2) - \mathcal{U}_S(s_1, s_2; \ell))f(\mathbf{x}) \right| \lesssim N^{1+n\xi} \frac{s_2 - s_1}{\ell} \|f\|_\infty, \quad (4.28)$$

for any small $\xi > 0$. The implicit constant in (4.13) depends on n, ϵ, δ .

In particular, this lemma shows that the observable g_t and its short-range approximation h_t are close to each other up to times $t \ll \ell/N$, hence to prove Proposition 4.1 will be enough to estimate h_t . First in Proposition 4.4 below will prove a bound in L^2 -sense that will be enhanced to an L^∞ bound by standard parabolic regularity arguments.

Define the event $\widehat{\Omega}$ on which the local laws for certain products of resolvents and traceless matrices A hold, i.e. for a small $\omega > 2\xi > 0$ we define

$$\begin{aligned} \widehat{\Omega} &= \widehat{\Omega}_{\omega, \xi} \\ &:= \bigcap_{\substack{z_i: \Re z_i \in [-3, 3], \\ |\Im z_i| \in [N^{-1+\omega}, 10]}} \left[\bigcap_{k=2}^n \left\{ \sup_{0 \leq t \leq T} (\rho_t^*)^{-1/2} \left| \langle G_t(z_1)A \dots G_t(z_k)A \rangle - \langle A^k \rangle \prod_{i=1}^k m_t(z_i) \right| \leq \frac{N^{\xi+k/2-1} \langle A^2 \rangle^{k/2}}{\sqrt{N} \eta_*} \right\} \right. \\ &\quad \left. \cap \left\{ \sup_{0 \leq t \leq T} (\rho_{1,t})^{-1/2} |\langle G_t(z_1)A \rangle| \leq \frac{N^\xi \langle A^2 \rangle^{1/2}}{N \sqrt{|\Im z_1|}} \right\} \right], \end{aligned} \quad (4.29)$$

where $\eta_* := \min\{|\Im z_i| \mid i \in [k]\}$, $\rho_{i,t} := |\Im m_t(z_i)|$, and $\rho_t^* := \max_i \rho_{i,t}$. Theorem 2.2 shows that $\widehat{\Omega}$ is a very high probability event, by using standard grid argument for the spectral parameters and stochastic continuity in the time parameter. Note that by rigidity (4.8) and the spectral theorem we have (recall the definition of $\gamma_i(0)$ from (4.9)):

$$\begin{aligned} &(\rho_t^*)^{-1} (\Im G_t(\gamma_{i_1}(t) + i\eta_1)A \Im G_t(\gamma_{i_2}(t) + i\eta_2)A) \\ &= \frac{1}{N \rho_t^*} \sum_{i,j=1}^N \frac{\eta^2 |\langle \mathbf{u}_i(t), A \mathbf{u}_j(t) \rangle|^2}{((\lambda_i(t) - \gamma_{i_1}(t))^2 + \eta_1^2)((\lambda_i(t) - \gamma_{i_2}(t))^2 + \eta_2^2)} \\ &\geq \frac{|\langle \mathbf{u}_{i_1}(t), A \mathbf{u}_{i_2}(t) \rangle|^2}{N \eta_1 \eta_2 \rho_t^*} \\ &= \frac{N [\rho(\gamma_{i_1}(t) + iN^{-2/3}) \wedge \rho(\gamma_{i_2}(t) + iN^{-2/3})] \cdot |\langle \mathbf{u}_{i_1}(t), A \mathbf{u}_{i_2}(t) \rangle|^2}{N^2 \eta_1 \eta_2 \rho_t^* [\rho(\gamma_{i_1}(t) + iN^{-2/3}) \wedge \rho(\gamma_{i_2}(t) + iN^{-2/3})]} \\ &= N^{1-2\omega} [\rho(\gamma_{i_1}(t) + iN^{-2/3}) \wedge \rho(\gamma_{i_2}(t) + iN^{-2/3})] \cdot |\langle \mathbf{u}_{i_1}(t), A \mathbf{u}_{i_2}(t) \rangle|^2 \end{aligned} \quad (4.30)$$

with $\eta_k = \eta_k(t)$ defined by $N\eta_k\rho(\gamma_{i_k}(t) + iN^{-2/3}) = N^\omega$. In particular, since $|\Im m_t(z_1)\Im m_t(z_2)| \lesssim \rho(z_1)\rho(z_2)$, by the first line of (4.29) for $k = 2$ we have

$$\sup_{0 \leq t \leq T} \sup_{z_1, z_2} (\rho_t^*)^{-1} \langle \Im G_t(z_1) A \Im G_t(z_2) A \rangle \lesssim \langle A^2 \rangle,$$

on $\widehat{\Omega}_{\omega, \xi}$, which by (4.30), choosing $z_k = \gamma_{i_k}(t) + i\eta$, implies

$$|\langle \mathbf{u}_i(t), A \mathbf{u}_j(t) \rangle|^2 \leq \frac{N^{2\omega} \langle A^2 \rangle}{N[\rho(\gamma_i(t) + iN^{-2/3}) \wedge \rho(\gamma_j(t) + iN^{-2/3})]} \quad \text{on } \widehat{\Omega}_{\omega, \xi} \cap \widetilde{\Omega}_\xi, \quad (4.31)$$

simultaneously for all $i, j \in [N]$ and $0 \leq t \leq T$. We recall that the quantiles $\gamma_i(t)$ are defined in (4.9).

Remark 4.3. The set $\widehat{\Omega}$ defined in (4.29) is slightly different from its analogue¹¹ in [19, Eq. (4.20)]. First, all the error terms now explicitly depend on $\langle A^2 \rangle$, whilst in [19, Eq. (4.20)] we just bounded the error terms using the operator norm of A (which was smaller than 1 in [19, Eq. (4.20)]). Second, we now have a slightly weaker bound (compared to [19, Eq. (4.20)]) for $\langle \Im G_t(z_1) A \Im G_t(z_2) A \rangle - \Im m_t(z_1) \Im m_t(z_2) \langle A^2 \rangle$, since we now do not carry the dependence on the $\rho_{i,t}$'s optimally; as a consequence of this slightly worse bound close to the edges we get the overlap bound (4.31), instead of the optimal bound [19, Eq. (4.21)], however this difference will not cause any change in the result. We remark that the bound (4.31) is optimal for bulk indices.

Proposition 4.4. For any parameters satisfying $N^{-1} \ll \eta \ll T_1 \ll \ell N^{-1} \ll KN^{-1}$, and any small $\epsilon, \xi > 0$ it holds

$$\|h_{T_1}(\cdot; \ell, K, \mathbf{y})\|_2 \lesssim K^{n/2} \mathcal{E}, \quad (4.32)$$

with

$$\mathcal{E} := N^{n\xi} \left(\frac{N^\epsilon \ell}{K} + \frac{NT_1}{\ell} + \frac{N\eta}{\ell} + \frac{N^\epsilon}{\sqrt{N\eta}} + \frac{1}{\sqrt{K}} \right), \quad (4.33)$$

uniformly for particle configuration $\mathbf{y} \in \Lambda^n$ supported on \mathcal{J} and eigenvalue trajectory $\boldsymbol{\lambda}$ in the high probability event $\widetilde{\Omega}_\xi \cap \widehat{\Omega}_{\omega, \xi}$.

Proof. This proof is very similar to that of [19, Proposition 4.5], hence we will only explain the main differences. The reader should consult with [19] for a fully detailed proof. The key idea is to replace the operator $\mathcal{S}(t)$ in (4.26)–(4.27), by the following operator

$$\mathcal{A}(t) := \sum_{\mathbf{i}, \mathbf{j} \in [N]^n}^* \mathcal{A}_{\mathbf{i}\mathbf{j}}(t), \quad \mathcal{A}_{\mathbf{i}\mathbf{j}}(t) h(\mathbf{x}) := \frac{1}{\eta} \left(\prod_{r=1}^n a_{i_r, j_r}^{\mathcal{S}}(t) \right) \sum_{\mathbf{a}, \mathbf{b} \in [2n]^n}^* (h(\mathbf{x}_{\mathbf{a}\mathbf{b}}^{ij}) - h(\mathbf{x})), \quad (4.34)$$

where

$$a_{ij} = a_{ij}(t) := \frac{\eta}{N((\lambda_i(t) - \lambda_j(t))^2 + \eta^2)}, \quad (4.35)$$

and $a_{ij}^{\mathcal{S}}$ are their short range version defined as in (4.25), and

$$\mathbf{x}_{\mathbf{a}\mathbf{b}}^{ij} := \mathbf{x} + \left(\prod_{r=1}^n \delta_{x_{a_r}, i_r} \delta_{x_{b_r}, j_r} \right) \sum_{r=1}^n (j_r - i_r) (\mathbf{e}_{a_r} + \mathbf{e}_{b_r}). \quad (4.36)$$

We remark that $\mathbf{x}_{\mathbf{a}\mathbf{b}}^{ij}$ from (4.20) changes two entries of \mathbf{x} per time, instead $\mathbf{x}_{\mathbf{a}\mathbf{b}}^{ij}$ changes all the coordinates of \mathbf{x} at the same time, i.e. let $\mathbf{i} := (i_1, \dots, i_n), \mathbf{j} := (j_1, \dots, j_n) \in [N]^n$, with $\{i_1, \dots, i_n\} \cap \{j_1, \dots, j_n\} = \emptyset$, then $\mathbf{x}_{\mathbf{a}\mathbf{b}}^{ij} \neq \mathbf{x}$ iff for all $r \in [n]$ it holds that $x_{a_r} = x_{b_r} = i_r$. This means that $\mathcal{S}(t)$ makes a jump only in one direction at a time, while $\mathcal{A}(t)$ jumps in all directions simultaneously. Technically, the replacement of $\mathcal{S}(t)$ by $\mathcal{A}(t)$ is done on the level of Dirichlet forms:

Lemma 4.5 (Lemma 4.6 of [19]). Let $\mathcal{S}(t), \mathcal{A}(t)$ be the generators defined in (4.27) and (4.34), respectively, and let μ denote the uniform measure on Λ^n for which $\mathcal{A}(t)$ is reversible. Then there exists a constant $C(n) > 0$ such that

$$\langle h, \mathcal{S}(t)h \rangle_{\Lambda^n, \pi} \leq C(n) \langle h, \mathcal{A}(t)h \rangle_{\Lambda^n, \mu} \leq 0, \quad (4.37)$$

for any $h \in L^2(\Lambda^n)$, on the very high probability set $\widetilde{\Omega}_\xi \cap \widehat{\Omega}_{\omega, \xi}$.

¹¹The definition of $\widehat{\Omega}$ in the published version of [19, Eq. (4.20)] contained a small error; the constraints were formally restricted only to spectral parameters in the bulk, even though the necessary bounds were directly available at the edge as well. This slightly imprecise formulation is corrected in the latest arXiv version of [19]; Remark 4.3 refers to the corrected version.

Next, combining

$$\partial_t \|h_t\|_2^2 = 2\langle h_t, \mathcal{S}(t)h_t \rangle_{\Lambda^n}, \quad (4.38)$$

which follows from (4.26), with (4.37), and using that $\mathbf{x}_{ab}^{ij} = \mathbf{x}$ unless $\mathbf{x}_{a_r} = \mathbf{x}_{b_r} = i_r$ for all $r \in [n]$, we conclude that

$$\begin{aligned} \partial_t \|h_t\|_2^2 &\leq C(n)\langle h_t, \mathcal{A}(t)h_t \rangle_{\Lambda^n, \mu} \\ &= \frac{C(n)}{2\eta} \sum_{\mathbf{x} \in \Lambda^n} \sum_{\mathbf{i}, \mathbf{j} \in [N]^n} \left(\prod_{r=1}^n a_{i_r j_r}^S(t) \right) \sum_{\mathbf{a}, \mathbf{b} \in [2n]^n} \bar{h}_t(\mathbf{x}) (h_t(\mathbf{x}_{ab}^{ij}) - h_t(\mathbf{x})) \left(\prod_{r=1}^n \delta_{x_{a_r} i_r} \delta_{x_{b_r} i_r} \right). \end{aligned} \quad (4.39)$$

The star over \sum means summation over two n -tuples of fully distinct indices. Then, proceeding as in the proof of [19, Proposition 4.5], we conclude that

$$\partial_t \|h_t\|_2^2 \leq -\frac{C_1(n)}{2\eta} \|h_t\|_2^2 + \frac{C_3(n)}{\eta} \mathcal{E}^2 K^n, \quad (4.40)$$

which implies $\|h_{T_1}\|_2^2 \leq C(n)\mathcal{E}^2 K^n$, by a simple Gronwall inequality, using that $T_1 \gg \eta$.

We point out that to go from (4.39) to (4.40) we proceed exactly as in the proof of [19, Proposition 4.5] (with the additional $\langle A^2 \rangle^{k/2}$, $\langle A^2 \rangle^{n/2}$ factors in [19, Eq. (4.47)] and [19, Eq. (4.48)], respectively) except for the estimate in [19, Eq. (4.43)]. The error terms in this estimate used that $|P(G)| \leq N^{n\xi - n/2}$ uniformly in the spectrum, a fact that we cannot establish near the edges as a consequence of the weaker bound (4.31). We now explain how we can still prove [19, Eq. (4.43)] in the current case. The main mechanism is that the strong bound $|P(G)| \leq N^{n\xi - n/2} \langle A^2 \rangle^{n/2}$ holds for bulk indices and when an edge index j is involved together with a bulk index i , then the kernel $a_{ij} \lesssim \eta/N$ is very small which balances the weaker estimate on the overlap. Note that (4.31) still provides a nontrivial bound of order $N^{-1/3}$ for $|\langle \mathbf{u}_i, \mathbf{A}\mathbf{u}_j \rangle|$ since $\rho(\gamma_i(t) + iN^{-2/3}) \gtrsim N^{-1/3}$ uniformly in $0 \leq t \leq T$.

We start with removing the short range cutoff from the kernel $a_{ij}^S(t)$ in the left hand side of [19, Eq. (4.43)]:

$$\begin{aligned} &\sum_j^* \left(\prod_{r=1}^n a_{i_r j_r}^S(t) \right) (g_t(\mathbf{x}_{ab}^{ij}) - \mathbf{1}(n \text{ even})) \\ &= \sum_j^* \left(\prod_{r=1}^n a_{i_r j_r}(t) \right) \left(\frac{N^{n/2}}{\langle A^2 \rangle^{n/2} 2^{n/2} (n-1)!!} \sum_{G \in \mathcal{G}_{\eta j}} P(G) - \mathbf{1}(n \text{ even}) \right) \\ &\quad - \sum_j^{**} \left(\prod_{r=1}^n a_{i_r j_r}(t) \right) \left(\frac{N^{n/2}}{\langle A^2 \rangle^{n/2} 2^{n/2} (n-1)!!} \sum_{G \in \mathcal{G}_{\eta j}} P(G) - \mathbf{1}(n \text{ even}) \right). \end{aligned} \quad (4.41)$$

Here \sum_j^{**} denotes the sum over distinct j_1, \dots, j_n such that at least one $|i_r - j_r|$ is bigger than ℓ .

Here the indices i_1, \dots, i_n are fixed and such that $i_l \in [\delta N, (1-\delta)N]$, for any $l \in [n]$. We will now show that the second line in (4.41) is estimated by $N^{1+n\xi}\eta\ell^{-1}$. This is clear for the terms containing $\mathbf{1}(n \text{ even})$, hence we now show that this bound is also valid for the terms containing $P(G)$. We present this bound only for the case when $|j_1 - i_1| > \ell$ and $|j_r - i_r| \leq \ell$ for any $r \in \{2, \dots, n\}$. The proof in the other cases is completely analogous and so omitted. Additionally, to make our presentation easier we assume that $n = 2$:

$$\begin{aligned} &\sum_{\substack{|j_1 - i_1| > \ell, |j_2 - i_2| \leq \ell, \\ j_1 \neq j_2}} a_{i_1 j_2}(t) a_{i_1 j_2}(t) \left(\frac{N}{2\langle A^2 \rangle} \sum_{G \in \mathcal{G}_{\eta j}} P(G) \right) \\ &= \left(\sum_{\substack{cN \geq |j_1 - i_1| > \ell, |j_2 - i_2| \leq \ell, \\ j_1 \neq j_2}} + \sum_{\substack{|j_1 - i_1| > cN, |j_2 - i_2| \leq \ell, \\ j_1 \neq j_2}} \right) a_{i_1 j_2}(t) a_{i_1 j_2}(t) \left(\frac{N}{2\langle A^2 \rangle} \sum_{G \in \mathcal{G}_{\eta j}} P(G) \right). \end{aligned} \quad (4.42)$$

Here $c \leq \delta/2$ is a small fixed constant so that j_1 is still a bulk index if $|i_1 - j_1| \leq cN$. The fact that the first summation in the second line of (4.42) is bounded by $N^{1+n\xi}\eta\ell^{-1}$ follows from (4.31), i.e. that $|\langle \mathbf{u}_i, \mathbf{A}\mathbf{u}_j \rangle| \leq N^{-1/2+\omega} \langle A^2 \rangle^{1/2}$, with very high probability, for any bulk indices i, j , in particular the bound $|P(G)| \leq N^{n\xi - n/2} \langle A^2 \rangle^{n/2}$ holds for this term.

For the second summation we have that

$$\begin{aligned} \sum_{\substack{|j_1 - i_1| > cN, |j_2 - i_2| \leq \ell, \\ j_1 \neq j_2}} a_{i_1 j_1}(t) a_{i_2 j_2}(t) \left(\frac{N}{2\langle A^2 \rangle} \sum_{G \in \mathcal{G}_{\eta^j}} P(G) \right) &\lesssim \frac{N^{1+\xi}\eta}{N^{2/3}} \sum_{|j_2 - i_2| \leq \ell} a_{i_2 j_2}(t) \\ &\lesssim \frac{N^{1+\xi}\eta}{N^{2/3}} \leq \frac{N\eta}{\ell}, \end{aligned} \quad (4.43)$$

where we used that $a_{i_1 j_1}(t) \lesssim \eta N^{-1}$, $\ell \ll K \ll \sqrt{N}$, and that

$$|P(G)| = |\langle \mathbf{u}_{j_1}, A\mathbf{u}_{j_1} \rangle \langle \mathbf{u}_{j_2}, A\mathbf{u}_{j_2} \rangle + 2|\langle \mathbf{u}_{j_1}, A\mathbf{u}_{j_2} \rangle|^2| \lesssim \frac{N^\xi}{N^{2/3}} \langle A^2 \rangle$$

by (4.31). We point out that to go from the first to the second line of (4.43) we also used that $\sum_{j_2} a_{i_2 j_2}(t) \lesssim 1$ on $\widehat{\Omega}$. This concludes the proof that the last line of (4.41) is bounded by $N^{1+n\xi}\eta\ell^{-1}$. We thus conclude that

$$\begin{aligned} &\sum_{\mathbf{j}}^* \left(\prod_{r=1}^n a_{i_r j_r}^S(t) \right) (g_t(\mathbf{x}_{\mathbf{ab}}^{i\mathbf{j}}) - \mathbf{1}(n \text{ even})) \\ &= \sum_{\mathbf{j}}^* \left(\prod_{r=1}^n a_{i_r j_r}(t) \right) \left(\frac{N^{n/2}}{\langle A^2 \rangle^{n/2} 2^{n/2} (n-1)!!} \sum_{G \in \mathcal{G}_{\eta^j}} P(G) - \mathbf{1}(n \text{ even}) \right) + \mathcal{O}\left(\frac{N^{1+n\xi}\eta}{\ell}\right). \end{aligned} \quad (4.44)$$

Proceeding in a similar way, i.e. splitting bulk and edge regimes and using the corresponding bounds for the overlaps, we then add back the missing indices in the summation in the second line of (4.44):

$$\begin{aligned} &\sum_{\mathbf{j}}^* \left(\prod_{r=1}^n a_{i_r j_r}(t) \right) \left(\frac{N^{n/2}}{\langle A^2 \rangle^{n/2} 2^{n/2} (n-1)!!} \sum_{G \in \mathcal{G}_{\eta^j}} P(G) - \mathbf{1}(n \text{ even}) \right) \\ &= \sum_{\mathbf{j}} \left(\prod_{r=1}^n a_{i_r j_r}(t) \right) \left(\frac{N^{n/2}}{\langle A^2 \rangle^{n/2} 2^{n/2} (n-1)!!} \sum_{G \in \mathcal{G}_{\eta^j}} P(G) - \mathbf{1}(n \text{ even}) \right) + \mathcal{O}\left(\frac{N^{n\xi}}{N\eta}\right). \end{aligned} \quad (4.45)$$

Finally, by (4.44)–(4.45), we conclude

$$\begin{aligned} &\sum_{\mathbf{j}}^* \left(\prod_{r=1}^n a_{i_r j_r}^S(t) \right) (g_t(\mathbf{x}_{\mathbf{ab}}^{i\mathbf{j}}) - \mathbf{1}(n \text{ even})) \\ &= \sum_{\mathbf{j}} \left(\prod_{r=1}^n a_{i_r j_r}(t) \right) \left(\frac{N^{n/2}}{\langle A^2 \rangle^{n/2} 2^{n/2} (n-1)!!} \sum_{G \in \mathcal{G}_{\eta^j}} P(G) - \mathbf{1}(n \text{ even}) \right) + \mathcal{O}\left(\frac{N^{n\xi}}{N\eta} + \frac{N^{1+n\xi}\eta}{\ell}\right), \end{aligned} \quad (4.46)$$

which is exactly the same of [19, Eq. (4.43)]. Given (4.46), the remaining part of the proof of this proposition is completely analogous to the proof of [19, Proposition 4.5], the only difference is that now in [19, Eq. (4.48)], using that $|m_t(z_i)| \lesssim 1$ uniformly in $0 \leq t \leq T$, we will have an additional error term

$$\frac{N^{n/2}}{\langle A^2 \rangle^{n/2}} \sum_{r=1}^n \sum_{k_1 + \dots + k_r = n}^* \prod_{i=1}^r N^{1-k_i} \langle A^{k_i} \rangle \lesssim \frac{N^{n/2}}{\langle A^2 \rangle^{n/2}} \sum_{r=1}^n \sum_{k_1 + \dots + k_r = n}^* \prod_{i=1}^r N^{-k_i/2} N^{-\delta'(k_i/2-1)} \langle A^2 \rangle^{k_i/2} \lesssim N^{-\delta'}$$

coming from the deterministic term in (4.29) (the mixed terms when we use the error term in (4.29) for some terms and the leading term for the remaining terms are estimated in the same way). We remark that in the first inequality we used that

$$\langle A^{k_i} \rangle \leq \|A\|^{k_i-2} \langle A^2 \rangle \lesssim (N^{1-\delta'})^{(k_i-2)/2} \langle A^2 \rangle^{k_i/2}$$

by our assumption $\langle A^2 \rangle \gtrsim N^{-1+\delta'} \|A\|^2$ from Theorem 2.8. Here $\sum_{k_1 + \dots + k_r = n}^*$ denotes the summation over all $k_1, \dots, k_r \geq 2$ such that there exists at least one r_0 such that $k_{r_0} \geq 3$. \square

4.2.3. *Proof of Proposition 4.1.* Given the finite speed of propagation estimates in Lemma 4.2 and the L^2 -bound on h_t from Proposition 4.4 as an input, enhancing this bound to an L^∞ -bound and hence proving Proposition 4.1 is completely analogous to the proof of [19, Proposition 3.2] presented in [19, Section 4.4] and so omitted.

APPENDIX A. PROOF OF THEOREM 2.2 IN THE LARGE d REGIME

The $d \geq 10$ regime is much simpler mainly because the trivial norm bound $\|G(z)\| \leq 1/d$ on every resolvent is affordable. In particular, no system of master inequalities and their meticulously bootstrapped analysis are necessary; a simple induction on k is sufficient. We remark that the argument using these drastic simplifications is completely analogous¹² to [20, Appendix B], hence we will be very brief.

We now assume that (2.5) has been proven up to some $k - 1$ in the $d \geq 10$ regime. Using (3.19) and estimating all resolvent chains in the right hand side of (3.19) by the induction hypotheses (after splitting $A_k A_1 = \langle A_k A_1 \rangle + (A_k A_1)^\circ$), using the analogue of Lemma 3.1 to estimate $\langle (GA)^{j-1} G \rangle$ in terms of the induction hypothesis, we easily obtain

$$\langle (GA)^k - m^k A^k \rangle \left(1 + \mathcal{O}_\prec \left(\frac{1}{Nd^2} \right) \right) = -m \langle W(GA)^k \rangle + \mathcal{O}_\prec \left(\frac{N^{k/2-1}}{d^k} \frac{1}{Nd^2} \right) \quad (\text{A.1})$$

in place of Lemma 3.5. In estimating the leading terms in (3.19) we used that $|m[z_1, z_k] - m(z_1)m(z_k)| \lesssim d^{-4}$. Note that $N^{k/2-1}/d^k$ is the natural size of the leading deterministic term $\langle m^k A^k \rangle$ under the normalization $\langle |A|^2 \rangle = 1$ and the small factor $1/Nd^2$ represents the smallness of the negligible error term. We now follow the argument in Section 3 starting from (3.26). For the Gaussian term (3.28) we simply bound

$$\left| m \frac{\langle (GA)^{2k} G \rangle}{N^2} \right| \prec \frac{N^{k-3}}{d^{2k+2}} = \left(\frac{N^{k/2-1}}{d^k} \frac{1}{\sqrt{Nd}} \right)^2 \quad (\text{A.2})$$

indicating a gain of order $1/(\sqrt{Nd})$ over the natural size of the leading term in (A.1); this gives the main error term in (2.5). The modifications to the non-Gaussian terms (3.27), i.e. the estimates of (3.30) and (3.33) are similarly straightforward and left to the reader. This completes the proof in the remaining $d \geq 10$ regime.

APPENDIX B. GREEN FUNCTION COMPARISON

The Green function comparison argument is very similar to the one presented in [19, Appendix A], hence we only explain the minor differences.

Consider the Ornstein-Uhlenbeck flow

$$d\widehat{W}_t = -\frac{1}{2}\widehat{W}_t dt + \frac{d\widehat{B}_t}{\sqrt{N}}, \quad \widehat{W}_0 = W, \quad (\text{B.1})$$

with \widehat{B}_t a real symmetric Brownian motion. Along the OU-flow (B.1) the moments of the entries of \widehat{W}_t remain constant, additionally, this flow adds a small Gaussian component to W , so that for any fixed T we have

$$\widehat{W}_T \stackrel{d}{=} \sqrt{1 - cT} \widetilde{W} + \sqrt{cT} U, \quad (\text{B.2})$$

with $c = c(T) > 0$ a constant very close to one as long as $T \ll 1$, and U, \widetilde{W} being independent GOE/Wigner matrices. Now consider the solution of the flow (4.1) W_t with initial condition $W_0 = \sqrt{1 - cT} \widetilde{W}$, so that

$$W_{cT} \stackrel{d}{=} \widehat{W}_T. \quad (\text{B.3})$$

Lemma B.1. *Let \widehat{W}_t be the solution of (B.1), and let $\widehat{u}_i(t)$ be its eigenvectors. Then for any smooth test function θ of at most polynomial growth, and any fixed $\epsilon \in (0, 1/2)$ there exists an $\omega = \omega(\theta, \epsilon) > 0$ such that for any bulk index $i \in [\delta N, (1 - \delta)N]$ (with $\delta > 0$ from Theorem 2.8) and $t = N^{-1+\epsilon}$ it holds that*

$$\mathbf{E} \theta \left(\sqrt{\frac{N}{2\langle A^2 \rangle}} \langle \widehat{u}_i(t), A \widehat{u}_i(t) \rangle \right) = \mathbf{E} \theta \left(\sqrt{\frac{N}{2\langle A^2 \rangle}} \langle \widehat{u}_i(0), A \widehat{u}_i(0) \rangle \right) + \mathcal{O}(N^{-\omega}). \quad (\text{B.4})$$

¹²We point out that the N -scaling here is naturally different from that in [20, Appendix B] simply due to the fact that here we chose the normalization $\langle |A_i|^2 \rangle = 1$ instead of $\|A_i\| = 1$.

We now show how to conclude Theorem 2.8 using the GFT result from Lemma B.1. Choose $T = N^{-1+\epsilon}$ and $\theta(x) = x^n$ for some integer $n \in \mathbf{N}$, then we have

$$\begin{aligned} \mathbf{E} \left[\sqrt{\frac{N}{2\langle A^2 \rangle}} \langle \mathbf{u}_i, A\mathbf{u}_i \rangle \right]^n &= \mathbf{E} \left[\sqrt{\frac{N}{2\langle A^2 \rangle}} \langle \widehat{\mathbf{u}}_i(T), A\widehat{\mathbf{u}}_i(T) \rangle \right]^n + \mathcal{O}(N^{-c}) \\ &= \mathbf{E} \left[\sqrt{\frac{N}{2\langle A^2 \rangle}} \langle \mathbf{u}_i(cT), A\mathbf{u}_i(cT) \rangle \right]^n + \mathcal{O}(N^{-c}) \\ &= \mathbf{1}(n \text{ even})(n-1)!! + \mathcal{O}(N^{-c}), \end{aligned} \quad (\text{B.5})$$

for some small $c = c(n, \epsilon) > 0$, with $\mathbf{u}_i, \widehat{\mathbf{u}}_i(t), \mathbf{u}_i(t)$ being the eigenvectors of W, \widehat{W}_t, W_t , respectively. This concludes the proof of Theorem 2.8. Note that in (B.5) we used Lemma B.1 in the first step, (B.3) in the second step and (4.13) for $\boldsymbol{\eta}$ such that $\eta_i = n$ and $\eta_j = 0$ for any $j \neq i$ in the third step, using that in distribution the eigenvectors of W_{cT} are equal to those of $\widetilde{W}_{cT/(1-cT)}$ with \widetilde{W}_t being the solution to the DBM flow with initial condition $\widetilde{W}_0 = \widehat{W}$.

Proof of Lemma B.1. The proof of this lemma is very similar to the proof of [19, Appendix A]. The differences come from the somewhat different local law. First, we now systematically carry the factor $\langle A^2 \rangle$ instead of $\|A\|^2 = 1$ as in [19, Appendix A], but this is automatic. Second, since the current overlap bound (4.31) is somewhat weaker near the edge, we need to check that for resolvents with spectral parameters in the bulk this will make no essential difference. This is the main purpose of repeating the standard proof from [19, Appendix A] in some details.

As a consequence of the repulsion of the eigenvalues (level repulsion), as in [39, Lemma 5.2], to understand the overlap $\sqrt{N}\langle A^2 \rangle^{-1/2} \langle \mathbf{u}_i, A\mathbf{u}_i \rangle$ it is enough to understand functions of $\sqrt{N}\langle A^2 \rangle^{-1/2} \langle \Im G(z)A \rangle$ with $\Im z$ slightly below N^{-1} , i.e. the local eigenvalue spacing. In particular, to prove (B.4) it is enough to show that

$$\sup_{E \in (-2+\delta, 2-\delta)} \left| \mathbf{E} \theta(\sqrt{N}\langle A^2 \rangle^{-1/2} \langle \Im G_t(z)A \rangle) - \mathbf{E} \theta(\sqrt{N}\langle A^2 \rangle^{-1/2} \langle \Im G_0(z)A \rangle) \right| \lesssim N^{-\omega}, \quad (\text{B.6})$$

for $t = N^{-1+\epsilon}$, $z = E + i\eta$ for some $\zeta > 0, \omega > 0$ and all $\eta \geq N^{-1-\zeta}$, c.f. [6, Section 4] and [11, Appendix A].

To prove this we define

$$R_t := \theta(\sqrt{N}\langle A^2 \rangle^{-1/2} \langle \Im G_t(z)A \rangle), \quad (\text{B.7})$$

and then use Itô's formula:

$$\mathbf{E} \frac{dR_t}{dt} = \mathbf{E} \left[-\frac{1}{2} \sum_{\alpha} w_{\alpha}(t) \partial_{\alpha} R_t + \frac{1}{2} \sum_{\alpha, \beta} \kappa_t(\alpha, \beta) \partial_{\alpha} \partial_{\beta} R_t \right], \quad (\text{B.8})$$

where $\alpha, \beta \in [N]^2$ are double indices, $w_{\alpha}(t)$ are the entries of W_t , and $\partial_{\alpha} := \partial_{w_{\alpha}}$. Here

$$\kappa_t(\alpha_1, \dots, \alpha_l) := \kappa(w_{\alpha_1}(t), \dots, w_{\alpha_l}(t)) \quad (\text{B.9})$$

denotes the joint cumulant of $w_{\alpha_1}(t), \dots, w_{\alpha_l}(t)$, with $l \in \mathbf{N}$. Note that by (2.2) it follows that $|\kappa_t(\alpha_1, \dots, \alpha_l)| \lesssim N^{-l/2}$ uniformly in $t \geq 0$.

By cumulant expansion we get

$$\mathbf{E} \frac{dR_t}{dt} = \sum_{l=3}^R \sum_{\alpha_1, \dots, \alpha_l} \kappa_t(\alpha_1, \dots, \alpha_l) \mathbf{E}[\partial_{\alpha_1} \cdots \partial_{\alpha_l} R_t] + \Omega(R), \quad (\text{B.10})$$

where $\Omega(R)$ is an error term, easily seen to be negligible as every additional derivative gains a further factor of $N^{-1/2}$. Then to estimate (B.10) we realize that ∂_{ab} -derivatives of $\langle \Im GA \rangle$ result in factors of the form $(GAG)_{ab}, (GAG)_{aa}$. For

such factors we use that

$$\begin{aligned} |(G_t(z_1)AG_t(z_2))_{ab}| &= \left| \sum_{ij} \frac{\mathbf{u}_i(a)\langle \mathbf{u}_i, A\mathbf{u}_j \rangle \mathbf{u}_j(b)}{(\lambda_i - z_1)(\lambda_j - z_2)} \right| \\ &\lesssim N^{2/3+\xi} \langle A^2 \rangle^{1/2} \left(\frac{1}{N} \sum_i \frac{1}{|\lambda_i - z_1|} \right) \left(\frac{1}{N} \sum_i \frac{1}{|\lambda_i - z_2|} \right) \\ &\lesssim N^{2/3+\xi+2\zeta} \langle A^2 \rangle^{1/2}, \end{aligned} \tag{B.11}$$

where we used that $\|\mathbf{u}_i\|_\infty \lesssim N^{-1/2+\xi}$, $|\langle \mathbf{u}_i, A\mathbf{u}_j \rangle| \leq N^{-1/3+\xi}$, for any $\xi > 0$, uniformly in the spectrum by [31], and Theorem 2.6, respectively. We remark that in [19, Eq. (A.11)] we could bound $(G_t(z_1)AG_t(z_2))_{ab}$ by $N^{1/2+\xi+2\zeta}$ as a consequence of the better bound on $|\langle \mathbf{u}_i, A\mathbf{u}_j \rangle|$ for indices close to the edge (however in [19, Eq. (A.11)] we did not have $\langle A^2 \rangle^{1/2}$). While our estimate on $(GAG)_{ab}$ is now by a factor $N^{1/6}$ weaker, this is still sufficient to complete the Green function comparison argument.

Indeed, using (B.11) and that $|(G_t)_{ab}| \leq N^\zeta$, for any $\zeta > 0$, we conclude that

$$\left| \partial_{\alpha_1} \dots \partial_{\alpha_l} \frac{\sqrt{N}}{\langle A^2 \rangle} \langle \Im G_t A \rangle \right| \leq N^{1/3+(l+3)(\zeta+\xi)}, \tag{B.12}$$

and so, together with

$$\sum_{\alpha_1, \dots, \alpha_l} |\kappa_t(\alpha_1, \dots, \alpha_l)| \lesssim N^{2-l/2},$$

by (B.10), we conclude (B.6). \square

REFERENCES

- ¹A. Aggarwal, P. Lopatto, and J. Marcinek, *Eigenvector statistics of Lévy matrices*, *Ann. Probab.* **49**, 1778–1846 (2021), [MR4260468](#).
- ²N. Anantharaman and E. Le Masson, *Quantum ergodicity on large regular graphs*, *Duke Math. J.* **164**, 723–765 (2015), [MR3322309](#).
- ³N. Anantharaman and M. Sabri, *Quantum ergodicity on graphs: from spectral to spatial delocalization*, *Ann. of Math. (2)* **189**, 753–835 (2019), [MR3961083](#).
- ⁴R. Bauerschmidt, J. Huang, and H.-T. Yau, *Local Kesten-McKay law for random regular graphs*, *Comm. Math. Phys.* **369**, 523–636 (2019), [MR3962004](#).
- ⁵R. Bauerschmidt, A. Knowles, and H.-T. Yau, *Local semicircle law for random regular graphs*, *Comm. Pure Appl. Math.* **70**, 1898–1960 (2017), [MR3688032](#).
- ⁶L. Benigni, *Eigenvectors distribution and quantum unique ergodicity for deformed Wigner matrices*, *Ann. Inst. Henri Poincaré Probab. Stat.* **56**, 2822–2867 (2020), [MR4164858](#).
- ⁷L. Benigni and P. Lopatto, *Fluctuations in local quantum unique ergodicity for generalized Wigner matrices*, *Comm. Math. Phys.* **391**, 401–454 (2022), [MR4397177](#).
- ⁸L. Benigni and P. Lopatto, *Optimal delocalization for generalized Wigner matrices*, *Adv. Math.* **396**, Paper No. 108109 (2022), [MR4370471](#).
- ⁹L. Benigni, *Fermionic eigenvector moment flow*, *Probab. Theory Related Fields* **179**, 733–775 (2021), [MR4242625](#).
- ¹⁰L. Benigni and P. Lopatto, *Fluctuations in local quantum unique ergodicity for generalized Wigner matrices*, *Commun. Math. Phys.* (2021), [arXiv:2103.12013](#).
- ¹¹P. Bourgade and H.-T. Yau, *The eigenvector moment flow and local quantum unique ergodicity*, *Comm. Math. Phys.* **350**, 231–278 (2017), [MR3606475](#).
- ¹²P. Bourgade, J. Huang, and H.-T. Yau, *Eigenvector statistics of sparse random matrices*, *Electron. J. Probab.* **22**, Paper No. 64, 38 (2017), [MR3690289](#).
- ¹³P. Bourgade, H.-T. Yau, and J. Yin, *Random band matrices in the delocalized phase I: Quantum unique ergodicity and universality*, *Comm. Pure Appl. Math.* **73**, 1526–1596 (2020), [MR4156609](#).
- ¹⁴C. Cacciapuoti, A. Maltsev, and B. Schlein, *Bounds for the Stieltjes transform and the density of states of Wigner matrices*, *Probab. Theory Related Fields* **163**, 1–59 (2015), [MR3405612](#).

- ¹⁵G. Cipolloni, L. Erdős, and D. Schröder, *Central limit theorem for linear eigenvalue statistics of non-Hermitian random matrices*, *Comm. Pure Appl. Math.* (2019), [arXiv:1912.04100](#).
- ¹⁶G. Cipolloni, L. Erdős, and D. Schröder, *Eigenstate thermalization hypothesis for Wigner matrices*, *Comm. Math. Phys.* **388**, 1005–1048 (2021), [MR4334253](#).
- ¹⁷G. Cipolloni, L. Erdős, and D. Schröder, *Functional central limit theorems for Wigner matrices*, Accepted for publication in *Ann. Appl. Probab.* (2020), [arXiv:2012.13218](#).
- ¹⁸G. Cipolloni, L. Erdős, and D. Schröder, *Normal fluctuation in quantum ergodicity for Wigner matrices*, Accepted for publication in *Ann. Probab.* (2021), [arXiv:2103.06730](#).
- ¹⁹G. Cipolloni, L. Erdős, and D. Schröder, *Normal fluctuation in quantum ergodicity for Wigner matrices*, *Ann. Probab.* **50**, 984–1012 (2022), [MR4413210](#).
- ²⁰G. Cipolloni, L. Erdős, and D. Schröder, *Optimal multi-resolvent local laws for Wigner matrices*, *Electron. J. Probab.* **27**, – (2022), [MR4479913](#).
- ²¹G. Cipolloni, L. Erdős, and D. Schröder, *Thermalisation for Wigner matrices*, *J. Funct. Anal.* **282**, Paper No. 109394, 37 (2022), [MR4372147](#).
- ²²Y. Colin de Verdière, *Ergodicité et fonctions propres du laplacien*, *Comm. Math. Phys.* **102**, 497–502 (1985), [MR818831](#).
- ²³L. D’Alessio, Y. Kafri, A. Polkovnikov, and M. Rigol, *From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics*, *Advances in Physics* **65**, 239–362 (2016).
- ²⁴J. Deutsch, *Eigenstate thermalization hypothesis*, *Rep. Prog. Phys.* **81**, 082001 (2018), [PMID29862983](#).
- ²⁵B. Eckhardt, S. Fishman, J. Keating, O. Agam, J. Main, and K. Müller, *Approach to ergodicity in quantum wave functions*, *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics* **52**, 5893–5903 (1995), [PMID9964105](#).
- ²⁶L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *The local semicircle law for a general class of random matrices*, *Electron. J. Probab.* **18**, no. 59, 58 (2013), [MR3068390](#).
- ²⁷L. Erdős, T. Krüger, and D. Schröder, *Cusp universality for random matrices I: local law and the complex Hermitian case*, *Comm. Math. Phys.* **378**, 1203–1278 (2020), [MR4134946](#).
- ²⁸L. Erdős, T. Krüger, and D. Schröder, *Random matrices with slow correlation decay*, *Forum Math. Sigma* **7**, e8, 89 (2019), [MR3941370](#).
- ²⁹L. Erdős, B. Schlein, and H.-T. Yau, *Local semicircle law and complete delocalization for Wigner random matrices*, *Comm. Math. Phys.* **287**, 641–655 (2009), [MR2481753](#).
- ³⁰L. Erdős, H.-T. Yau, and J. Yin, *Bulk universality for generalized Wigner matrices*, *Probab. Theory Related Fields* **154**, 341–407 (2012), [MR2981427](#).
- ³¹L. Erdős, H.-T. Yau, and J. Yin, *Rigidity of eigenvalues of generalized Wigner matrices*, *Adv. Math.* **229**, 1435–1515 (2012), [MR2871147](#).
- ³²M. Feingold and A. Peres, *Distribution of matrix elements of chaotic systems*, *Physical review. A, General physics* **34**, 591–595 (1986), [PMID9897286](#).
- ³³F. Götze, A. A. Naumov, and A. N. Tikhomirov, *A local semicircle law under moment conditions: the Stieltjes transfer, rigidity and delocalization*, *Teor. Veroyatn. Primen.* **62**, 72–103 (2017), [MR3633466](#).
- ³⁴F. Götze, A. Naumov, and A. Tikhomirov, *Local semicircle law under fourth moment condition*, *J. Theoret. Probab.* **33**, 1327–1362 (2020), [MR4125959](#).
- ³⁵Y. He and A. Knowles, *Mesoscopic eigenvalue statistics of Wigner matrices*, *Ann. Appl. Probab.* **27**, 1510–1550 (2017), [MR3678478](#).
- ³⁶Y. He, A. Knowles, and R. Rosenthal, *Isotropic self-consistent equations for mean-field random matrices*, *Probab. Theory Related Fields* **171**, 203–249 (2018), [MR3800833](#).
- ³⁷A. M. Khorunzhy, B. A. Khoruzhenko, and L. A. Pastur, *Asymptotic properties of large random matrices with independent entries*, *J. Math. Phys.* **37**, 5033–5060 (1996), [MR1411619](#).
- ³⁸A. Knowles and J. Yin, *Anisotropic local laws for random matrices*, *Probab. Theory Related Fields* **169**, 257–352 (2017), [MR3704770](#).
- ³⁹A. Knowles and J. Yin, *Eigenvector distribution of Wigner matrices*, *Probab. Theory Related Fields* **155**, 543–582 (2013), [MR3034787](#).
- ⁴⁰A. Knowles and J. Yin, *The isotropic semicircle law and deformation of Wigner matrices*, *Comm. Pure Appl. Math.* **66**, 1663–1750 (2013), [MR3103909](#).

- ⁴¹J. O. Lee and K. Schnelli, *Edge universality for deformed Wigner matrices*, *Rev. Math. Phys.* **27**, 1550018, 94 (2015), [MR3405746](#).
- ⁴²J. O. Lee and K. Schnelli, *Local deformed semicircle law and complete delocalization for Wigner matrices with random potential*, *J. Math. Phys.* **54**, 103504, 62 (2013), [MR3134604](#).
- ⁴³W. Z. Luo and P. Sarnak, *Quantum ergodicity of eigenfunctions on $\mathrm{PSL}_2(\mathbf{Z}) \backslash \mathbf{H}^2$* , *Inst. Hautes Études Sci. Publ. Math.*, 207–237 (1995), [MR1361757](#).
- ⁴⁴A. Lytova, *On non-Gaussian limiting laws for certain statistics of Wigner matrices*, *Zh. Mat. Fiz. Anal. Geom.* **9**, 536–581, 611, 615 (2013), [MR3155024](#).
- ⁴⁵J. Marcinek, *High Dimensional Normality of Noisy Eigenvectors*, Thesis (Ph.D.)–Harvard University (ProQuest LLC, Ann Arbor, MI, 2020), p. 145, [MR4272266](#).
- ⁴⁶J. Marklof and Z. Rudnick, *Quantum unique ergodicity for parabolic maps*, *Geom. Funct. Anal.* **10**, 1554–1578 (2000), [MR1810753](#).
- ⁴⁷S. O’Rourke, V. Vu, and K. Wang, *Eigenvectors of random matrices: a survey*, *J. Combin. Theory Ser. A* **144**, 361–442 (2016), [MR3534074](#).
- ⁴⁸M. Rudelson and R. Vershynin, *Sampling from large matrices: an approach through geometric functional analysis*, *J. ACM* **54**, Art. 21, 19 (2007), [MR2351844](#).
- ⁴⁹Z. Rudnick and P. Sarnak, *The behaviour of eigenstates of arithmetic hyperbolic manifolds*, *Comm. Math. Phys.* **161**, 195–213 (1994), [MR1266075](#).
- ⁵⁰A. I. Snirelman, *Ergodic properties of eigenfunctions*, *Uspehi Mat. Nauk* **29**, 181–182 (1974), [MR0402834](#).
- ⁵¹K. Soundararajan, *Quantum unique ergodicity for $\mathrm{SL}_2(\mathbb{Z}) \backslash \mathbb{H}$* , *Ann. of Math. (2)* **172**, 1529–1538 (2010), [MR2680500](#).
- ⁵²T. Tao and V. Vu, *Random matrices: universality of local eigenvalue statistics*, *Acta Math.* **206**, 127–204 (2011), [MR2784665](#).
- ⁵³T. Tao and V. Vu, *Random matrices: universality of local eigenvalue statistics up to the edge*, *Comm. Math. Phys.* **298**, 549–572 (2010), [MR2669449](#).
- ⁵⁴V. Vu and K. Wang, *Random weighted projections, random quadratic forms and random eigenvectors*, *Random Structures Algorithms* **47**, 792–821 (2015), [MR3418916](#).
- ⁵⁵E. P. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, *Ann. of Math. (2)* **62**, 548–564 (1955), [MR77805](#).
- ⁵⁶S. Zelditch, “Recent developments in mathematical quantum chaos”, *Current developments in mathematics, 2009* (Int. Press, Somerville, MA, 2010), pp. 115–204, [MR2757360](#).
- ⁵⁷S. Zelditch, *Uniform distribution of eigenfunctions on compact hyperbolic surfaces*, *Duke Math. J.* **55**, 919–941 (1987), [MR916129](#).