*Forum of Mathematics, Sigma* (2022), **Vol. 10**:e96 1–43
doi:10.1017/fms.2022.86

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

# Rank-uniform local law for Wigner matrices

Giorgio Cipolloni[1], László Erdős[2] and Dominik Schröder[3]

[1]Princeton Center for Theoretical Science, Princeton University, Princeton, NJ 08544, USA; E-mail: gc4233@princeton.edu.
[2]IST Austria, Am Campus 1, Klosterneuburg, 3400, Austria; E-mail: lerdos@ist.ac.at.
[3]Institute for Theoretical Studies, ETH Zurich, Clausiusstr. 47, Zürich, 8092, Switzerland; E-mail: dschroeder@ethz.ch.

**Received:** 10 March 2022; **Revised:** 07 September 2022; **Accepted:** 24 September 2022

**2020 Mathematics Subject Classification:** *Primary* – 60B20; *Secondary* – 15B52

**Abstract**

We prove a general local law for Wigner matrices that optimally handles observables of arbitrary rank and thus unifies the well-known averaged and isotropic local laws. As an application, we prove a central limit theorem in quantum unique ergodicity (QUE): that is, 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 generalise our previous result [17] as valid for test matrices $A$ of large rank as well as the result of Benigni and Lopatto [7] as valid for specific small-rank observables.

## Contents

**1 Introduction** **1**
**2 Main results** **5**
**3 Proof of Theorem 2.2** **8**
3.1 Proof of averaged estimate given by equation (3.8) in Proposition 3.2 . . . . . 10
3.2 Proof of the isotropic estimate given by equation (3.9) in Proposition 3.2 . . . . . 20
3.3 Reduction inequalities and bootstrap . . . . . 23
**4 Stochastic eigenstate equation and proof of Theorem 2.8** **28**
4.1 Perfect matching observables . . . . . 29
4.2 DBM analysis . . . . . 31
4.2.1 Short-range approximation . . . . . 33
4.2.2 $L^2$-bound . . . . . 33
4.2.3 Proof of Proposition 4.1 . . . . . 38
**A Proof of Theorem 2.2 in the large $d$ regime** **38**
**B Green function comparison** **39**

## 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 [53] 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*, $\rho_{\mathrm{sc}}(x) = \frac{1}{2\pi}\sqrt{4 - x^2}$, which is the $N \to \infty$ limit of the empirical density of the eigenvalues of $W$ under the standard normalisation $\mathbf{E}\,|w_{ab}|^2 = 1/N$. 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 $G(z) \to m(z) \cdot I$ holds in a natural appropriate topology, namely when tested against deterministic $N \times N$ matrices $A$: that is, in the form $\langle G(z)A\rangle \to m(z)\langle A\rangle$, where $\langle\cdot\rangle := \frac{1}{N}\operatorname{Tr}(\cdot)$ denotes the normalised trace. It is essential that the test matrix $A$ is deterministic; no analogous limit can hold if $A$ is random and strongly correlated with $W$: for example, if $A$ is a spectral projection of $W$.

The first optimal local law for Wigner matrices was proven for $A = I$ in [27]; see also [13, 32, 50, 51], extended later to more general matrices $A$ in the form that[1]

$$\left|\langle (G(z) - m(z))A\rangle\right| \le \frac{N^{\xi}\|A\|}{N\eta}, \qquad \eta := |\Im z| \tag{1.1}$$

holds with a 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 $N^{\xi}$. This is a natural byproduct of our method yielding very high probability estimates under the customary moment condition; see equation (2.2) later.[2] The estimate given by equation (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 $A$: for example, for $A = I$, the trace $\langle G - m\rangle$ is approximately complex Gaussian with standard deviation [33]

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

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

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

has been proven; see [28] for special coordinate vectors and later [38] for general vectors $\boldsymbol{x}, \boldsymbol{y}$, as well as [26, 34, 36, 40] for more general ensembles. Note that a direct application of equation (1.1) to $A = \boldsymbol{y}\boldsymbol{x}^*$ would give a bound of order $1/\eta$ instead of the optimal $1/\sqrt{N\eta}$ in equation (1.2), which is an unacceptable overestimate in the most interesting small $\eta$-regime. More generally, the average local law given by equation (1.1) performs badly when $A$ has effectively small rank: that is, if only a few eigenvalues of $A$ are comparable with the norm $\|A\|$ and most other eigenvalues are much smaller or even zero.

[1] Traditional 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, for example, [26].

[2] We remark that the $N^{\xi}$ tolerance factor can be improved to logarithmic factors under slightly different conditions; see, for example, [13, 31, 32].

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

Quite recently, we found that the average local law given by equation (1.1) is also suboptimal for another class of test matrices $A$, namely traceless matrices. In [15], we proved that

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

for any deterministic matrix $A$ with $\langle A\rangle = 0$: that is, 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 [16].

While the mechanism behind the suboptimality of equation (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 that unifies all three local laws in equations (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

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

holds with very high probability, where $\mathring{A} := A - \langle A\rangle$ is the traceless part of $A$. It is straightforward to check that equation (1.4) implies equations (1.1), (1.2) and (1.3); moreover, it optimally interpolates between full-rank and rank-one matrices $A$; hence we call equation (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 given by equations (1.1) and (1.2) are optimal only for essentially full-rank and essentially finite-rank observables, respectively. The proof of the optimality of equation (1.4) follows from identifying the scale of the Gaussian fluctuation of its left-hand side. Its standard deviation for traceless $A$ is

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

this relation was established for matrices with bounded norm $\|A\| \lesssim 1$ in [16, 42].

The key observation that traceless $A$ substantially improves the error term in equation (1.3) compared with equation (1.1) was the conceptually new input behind our recent proof of the *Eigenstate Thermalisation Hypothesis* in [15] followed by the proof of the normal fluctuation in the quantum unique ergodicity for Wigner matrices in [17]. Both results concern the behaviour of the *eigenvector overlaps*: that is, quantities of the form $\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle$, where $\{\boldsymbol{u}_i\}_{i=1}^N$ are the normalised eigenvectors of $W$. The former result stated that

$$\big|\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_j\rangle\big| = \big|\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle - \delta_{ij}\langle A\rangle\big| \le \frac{N^{\xi}\|\mathring{A}\|}{\sqrt{N}} \tag{1.6}$$

holds with very high probability for any $i, j$ and for any fixed $\xi > 0$. The latter result established the optimality of equation (1.6) for $i = j$ by showing that $\sqrt{N}\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle$ is asymptotically Gaussian when the corresponding eigenvalue lies in the bulk of the spectrum. The variance of $\sqrt{N}\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle$ was shown to be $\langle |\mathring{A}|^2\rangle$ in [17], but we needed to assume that $\langle |\mathring{A}|^2\rangle \ge c\|\mathring{A}\|^2$ with some fixed positive constant $c$: that is, 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 normalised overlaps $[N/\langle |\mathring{A}|^2\rangle]^{1/2}\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle$ 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 equation (1.4) over equation (1.3). We will also need a multi-resolvent version of this improvement since off-diagonal overlaps

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

$\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle$ are not accessible via single-resolvent local laws; in fact, $|\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle|^2$ 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 equation (1.6) (see Theorem 2.6)

$$\left|\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle - \delta_{ij}\langle A\rangle\right| \le \frac{N^{\xi}\langle |\mathring{A}|^2\rangle^{1/2}}{\sqrt{N}} \tag{1.7}$$

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 [17] was posted on the arXiv, Benigni and Lopatto in an independent work [7] proved the standard Gaussian fluctuation of $[N/|S|]^{1/2}\big[\sum_{a\in S}|u_i(a)|^2 - |S|/N\big]$ whenever $1 \ll |S| \ll N$: that is, they considered $\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle$ 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 $\langle |\mathring{A}|^2\rangle \ge c\|\mathring{A}\|^2$ from [17] 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 [7] and [17] are very different albeit they both rely on *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 $\langle q_\alpha, \boldsymbol{u}_i\rangle$ with deterministic unit vectors $q_\alpha$) in the spirit of the previous quantum ergodicity results by Bourgade and Yau [10] operating with the so-called *eigenvector moment flow* from [10] 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 $\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle$ and its off-diagonal version $\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_j\rangle$ as one unit without translating it into a sum of rank-one projections $\langle \boldsymbol{u}_i, q_\alpha\rangle\langle q_\alpha, \boldsymbol{u}_j\rangle$ 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 [12] (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 [43], and we heavily relied on their work in our proof [17].

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 in equation (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 [17]; hence we will not give all details but 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 a 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, for example, [15, 19], 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 that have typically been organised and estimated by a graphical Feynman diagrammatic scheme. A reasonably manageable power counting handles all diagrams for the purpose of proving equations (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'*)

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

was extracted at the cusp [25], (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 $1/|z_1 - z_2|$ [14] and (iii) more recently when the gain of order $\sqrt{\eta}$ due to the traceless $A$ in equation (1.3) was obtained in [15].

Extracting $\langle |\mathring{A}|^2 \rangle^{1/2}$ 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 [18] 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 [15]. In the current paper, we extend this technique to obtain the optimal control in terms of $\langle |\mathring{A}|^2 \rangle^{1/2}$ instead of $\|\mathring{A}\|$ for single resolvent local laws. However, the master inequalities in this paper are different from the ones in [18]; 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 lowercase Roman letters $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{C}^N$, for some $N \in \mathbf{N}$. Vector and matrix norms, $\|\boldsymbol{x}\|$ and $\|A\|$, indicate the usual Euclidean norm and the corresponding induced matrix norm. For any $N \times N$ matrix $A$, we use the notation $\langle A \rangle := N^{-1} \operatorname{Tr} A$ to denote the normalised trace of $A$. Moreover, for vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{C}^N$ and matrices $A \in \mathbf{C}^{N \times N}$, we define

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \sum_{i=1}^{N} \overline{x}_i y_i, \qquad A_{\boldsymbol{xy}} := \langle \boldsymbol{x}, A\boldsymbol{y} \rangle.$$

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, for example, [24]): given two families of non-negative random variables

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

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

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

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 that 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 $c, C > 0$ that depend only on the constants appearing in the moment condition; see equation (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*

$$w_{ab} \overset{\mathrm{d}}{=} N^{-1/2} \times \begin{cases} \chi_{\mathrm{od}}, & a \neq b \\ \chi_{\mathrm{d}}, & a = b \end{cases} \tag{2.1}$$

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

*for some real random variable $\chi_{\mathrm{d}}$ and some real/complex random variable $\chi_{\mathrm{od}}$ of mean $\mathbf{E}\,\chi_{\mathrm{d}} = \mathbf{E}\,\chi_{\mathrm{od}} = 0$ and variances $\mathbf{E}|\chi_{\mathrm{od}}|^2 = 1$, $\mathbf{E}\,\chi_{\mathrm{od}}^2 = 0$, $\mathbf{E}\,\chi_{\mathrm{d}}^2 = 1$ in the complex, and $\mathbf{E}|\chi_{\mathrm{od}}|^2 = \mathbf{E}\,\chi_{\mathrm{od}}^2 = 1$, $\mathbf{E}\,\chi_{\mathrm{d}}^2 = 2$ in the real case.*[3] *We furthermore assume that for every $n \geq 3$,*

$$\mathbf{E}|\chi_{\mathrm{d}}|^n + \mathbf{E}|\chi_{\mathrm{od}}|^n \leq C_n \tag{2.2}$$

*for some constant $C_n$; in particular, all higher-order cumulants $\kappa_n^{\mathrm{d}}, \kappa_n^{\mathrm{od}}$ of $\chi_{\mathrm{d}}, \chi_{\mathrm{od}}$ 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

$$-\frac{1}{m} = m + z, \qquad \Im m \Im z > 0. \tag{2.3}$$

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

We first recall the classical local law for Wigner matrices in both its tracial and isotropic forms [27, 29, 34, 38]:

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

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

*uniformly in any deterministic vectors $\boldsymbol{x}, \boldsymbol{y}$ 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 an 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 $z_1, \ldots, z_k \in \mathbf{C}$ with $N\eta\rho \geq N^{\epsilon}$, for $\eta := \min_i |\Im z_i|$, $\rho := \max_i |\Im m(z_i)|$, $d := \min_i \operatorname{dist}(z_i, [-2, 2])$, and let $A_1, \ldots, A_k$ 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 scale*[4]

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

**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 $A_k = N\boldsymbol{y}\boldsymbol{x}^* - \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ in equation (2.5).

[3] We assumed that $\sigma := \mathbf{E}\,\chi_{\mathrm{od}}^2 = 0$, $\mathbf{E}\,\chi_{\mathrm{d}}^2 = 1$ in the complex case and that $\mathbf{E}\,\chi_{\mathrm{d}}^2 = 2$ 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.

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

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

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

We now compare Theorem 2.2 to the previous result [18, Theorem 2.5], where an error term $N^{-1}\eta^{-k/2}\prod_i \|A_i\|$ was proven for equation (2.5). For clarity, we focus on the really interesting $d < 10$ regime.

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

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

is strictly better than the one in [18, 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, equation (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 $\eta$. We recall [46] that the *numerical rank* of $A$ is defined as $r(A) := N\langle |A|^2\rangle / \|A\|^2 \leq \operatorname{rank}(A)$ and say that $A$ is $\alpha$-mesoscopic for some $\alpha \in [0, 1]$ if $r(A) = N^\alpha$. If for some $k \geq 2$ all $A_i$ are $\alpha$-mesoscopic, then Theorem 2.2 improves upon [18, 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 $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$ denote the eigenvalues and $\{\boldsymbol{u}_i\}_{i=1}^N$ the corresponding orthonormal eigenvectors of $W$. Already the single-resolvent isotropic local law given by equation (2.4) implies the *eigenvector delocalisation*: that is, that $\|\boldsymbol{u}_i\|_\infty \prec N^{-1/2}$. More generally,[5] $|\langle \boldsymbol{x}, \boldsymbol{u}_i\rangle| \prec N^{-1/2}\|\boldsymbol{x}\|$: that is, eigenvectors behave as completely random unit vectors in the sense of considering their rank-1 projections onto any deterministic vector $\boldsymbol{x}$. This concept can be greatly extended to arbitrary deterministic observable matrix $A$, leading to the following results motivated both by thermalisation ideas from physics [21, 22, 23, 30] and by *quantum (unique) ergodicity (QUE)* in mathematics [2, 3, 4, 5, 20, 41, 44, 47, 48, 49, 54, 55].

**Theorem 2.6** (Eigenstate thermalisation 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*

$$\big|\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle - \delta_{ij}\langle A\rangle\big| \prec \frac{\langle |\mathring{A}|^2\rangle^{1/2}}{N^{1/2}}, \tag{2.8}$$

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

**Remark 2.7.**

1. The result given by equation (2.8) was established in [15] 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 $A = \boldsymbol{x}\boldsymbol{x}^*$, the bound given by equation (2.8) immediately implies the complete delocalisation of eigenvectors in the form $|\langle \boldsymbol{x}, \boldsymbol{u}_i\rangle| \prec N^{-1/2}\|\boldsymbol{x}\|$.

[5] Under stronger decay conditions on the distribution of $\chi_{\mathrm{d}}, \chi_{\mathrm{od}}$, even the optimal bound $\|\boldsymbol{u}_i\|_\infty \leq C\sqrt{\log N / N}$ for the bulk and $\|\boldsymbol{u}_i\|_\infty \leq C \log N / \sqrt{N}$ for the edge eigenvectors has been proven [52]; see also [45] for a comprehensive summary of related results. Very recently, even the optimal constant $C$ has been identified [8].

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

Theorem 2.6 directly follows from the bound

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

that is obtained by the spectral decomposition of both resolvents and the well-known eigenvalue rigidity, with some explicit $\delta$-dependent constants $C_\delta$ and $\epsilon = \epsilon(\delta) > 0$ (see [15, Lemma 1] for more details). The right-hand side can be directly estimated using equation (2.5); and finally, choosing $\eta = N^{-1+\xi}$ for any small $\xi > 0$ gives equation (2.8) and thus proves Theorem 2.6.

The next question is to establish a central limit theorem for the diagonal overlap in equation (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*

$$\sqrt{\frac{\beta N}{2\langle \mathring{A}^2\rangle}}\big[\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle - \langle A\rangle\big] \Rightarrow \mathcal{N}, \qquad \text{as} \quad N\to\infty \tag{2.9}$$

*with $\mathcal{N}$ being a standard real Gaussian random variable. Moreover, for any moment, the speed of convergence is explicit (see equation (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: that is, that $\mathring{A}$ has effective rank $\gg 1$. Indeed, the statement in equation (2.9) does not hold for finite-rank $A$s: for example, if $A = \mathring{A} = |\mathbf{e}_x\rangle\langle \boldsymbol{e}_x| - |\boldsymbol{e}_y\rangle\langle \mathbf{e}_y|$, for some $x \neq y \in [N]$, then $\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle = |\boldsymbol{u}_i(x)|^2 - |\boldsymbol{u}_i(y)|^2$, which is the difference of two asymptotically independent $\chi^2$-distributed random variables (e.g., see [10, Theorem 1.2]). More generally, the joint distribution of finitely many eigenvectors overlaps has been identified in [1, 10, 11, 43] for various related ensembles.

## 3. Proof of Theorem 2.2

In this section, we prove Theorem 2.2 in the critical $d < 10$ regime. The $d \ge 10$ 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 $\psi_K^{\mathrm{av/iso}} \ge 1$, of the form

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

$$\Psi_K^{\mathrm{av}} = \Psi_K^{\mathrm{av}}(\boldsymbol{A}, \boldsymbol{z}) := \frac{N^{(3-K)/2}\eta^{1/2}}{\rho^{1/2}\prod_i \langle |A_i|^2\rangle^{1/2}} |\langle [G_1A_1\cdots G_KA_K - m_1\cdots m_KA_1\cdots A_K]\rangle| \prec \psi_K^{\mathrm{av}}, \qquad K \ge 1, \tag{3.2}$$

$$\Psi_K^{\mathrm{iso}} = \Psi_K^{\mathrm{iso}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{A}, \boldsymbol{z}) := \frac{N^{(1-K)/2}\eta^{1/2}\rho^{-1/2}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|\prod_i \langle |A_i|^2\rangle^{1/2}} \times |\langle \boldsymbol{x}, [G_1A_1\cdots G_{K+1} - m_1\cdots m_{K+1}A_1\cdots A_K]\boldsymbol{y}\rangle| \prec \psi_K^{\mathrm{iso}} \tag{3.3}$$

for certain indices $K \ge 0$ have been established uniformly[6] in deterministic traceless matrices $\boldsymbol{A} = (A_1, \ldots, A_K)$, deterministic vectors $\boldsymbol{x}, \boldsymbol{y}$ and spectral parameters $\boldsymbol{z} = (z_1, \ldots, z_K)$ with $N\eta\rho \ge N^\epsilon$. We stress that we do *not* assume the estimates to be uniform in $K$. Note that $\psi_0^{\mathrm{av}}$ is defined somewhat differently from $\psi_K^{\mathrm{av}}$, $K \ge 1$, but the definition of $\psi_K^{\mathrm{iso}}$ is the same for all $K \ge 0$. For intuition, the reader

[6] In applications, the domain of uniformity may slightly change from $N\eta\rho \ge N^\epsilon$ to $N\eta\rho \ge \ell N^\epsilon$ with some fixed integer $\ell$, but for simplicity we ignore this subtlety.

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

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 $\psi_0^{\mathrm{av/iso}} = 1$.

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

**Lemma 3.1.** (i) *Assuming equations (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*

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

*where $m[z_1, \dots, z_l]$ stands for the lth divided difference of the function $m(z)$ from equation (2.3), explicitly*

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

(ii) *Assuming for some $k \geq 1$ the estimates given by equations (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*

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

*where*

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

*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 [18, Lemma 3.2]. □

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

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

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

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

*with the definition*

$$\Phi_k := \sum_{k_1+k_2+\cdots\le k} \Big(1+\frac{\psi^{\mathrm{iso}}_{2k_1}}{\sqrt{N\eta\rho}}\Big)^{1/2}\Big(1+\frac{\psi^{\mathrm{iso}}_{2k_2}}{\sqrt{N\eta\rho}}\Big)^{1/2}\prod_{i\ge 3}\Big(1+\psi^{\mathrm{iso}}_{k_i}\sqrt{\frac{\rho}{N\eta}}\Big), \tag{3.10}$$

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

This hierarchy has the structure that each $\Psi^{\mathrm{av/iso}}_k$ is estimated partly by $\psi$s with an 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 $\Psi$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 $\psi^{\mathrm{av/iso}}_l$ terms with $k+1 \le l \le 2k$ in equations (3.8) and (3.9) by $\psi$s with indices smaller than or equal to $k$.

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

$$\Psi^{\mathrm{av}}_{2k} \lesssim \sqrt{\frac{N\eta}{\rho}} + \begin{cases} \sqrt{\frac{\rho}{N\eta}}(\psi^{\mathrm{av}}_k)^2 & k \text{ even}, \\ \psi^{\mathrm{av}}_{k-1}+\psi^{\mathrm{av}}_{k+1}+\sqrt{\frac{\rho}{N\eta}}\psi^{\mathrm{av}}_{k-1}\psi^{\mathrm{av}}_{k+1} & k \text{ odd}, \end{cases} \tag{3.11}$$

*and for even $k$ also that*

$$\Psi^{\mathrm{iso}}_{k+j} \lesssim \sqrt{\frac{N\eta}{\rho}} + \Big(\frac{N\eta}{\rho}\Big)^{1/4}(\psi^{\mathrm{av}}_{2j})^{1/2} + \psi^{\mathrm{iso}}_k + \Big(\frac{\rho}{N\eta}\Big)^{1/4}(\psi^{\mathrm{av}}_{2j})^{1/2}\psi^{\mathrm{iso}}_k. \tag{3.12}$$

The rest of the present section is structured as follows: in Section 3.1, we prove equation (3.8), and in Section 3.2, we prove equation (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 $k \ge 2$, it holds that*

$$\Big|\Big\langle \prod_{i=1}^k B_i \Big\rangle\Big| \le \prod_{i=1}^k \langle |B_i|^k\rangle^{1/k} \le N^{k/2-1}\prod_{i=1}^k \langle |B_i|^2\rangle^{1/2} \tag{3.13}$$

*and*

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

In the sequel, we often drop the indices from $G, A$; hence we write $(GA)^k$ for $G_1A_1\dots G_kA_k$ 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 $A$ letters are always traceless.

### *3.1. Proof of averaged estimate given by equation (3.8) in Proposition 3.2*

We now identify the leading contribution of $\langle (GA)^k - m^kA^k\rangle$. For any matrix-valued function $f(W)$, we define the *second moment renormalisation*, denoted by underlining, as

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

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 $\mathbf{E}\, W f(W) = \widetilde{\mathbf{E}}\widetilde{W}(\partial_{\widetilde{W}} f)(W)$ whenever $W$ is a Gaussian random matrix of zero mean and $\widetilde{W}$ is an independent copy of $W$. In particular, it holds that $\mathbf{E}\, \underline{W f(W)} = 0$ whenever $W$ is a GUE-matrix, while $\mathbf{E}\, \underline{W f(W)}$ is small but nonzero for GOE or non-Gaussian matrices. By concentration and universality, we expect that to leading order $W f(W)$ 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

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

In our application, the function $f(W)$ is always 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

$$\begin{aligned}(\partial_{\widetilde{W}} G)(W) &= \lim_{\epsilon \to 0} \frac{(W + \epsilon \widetilde{W} - z)^{-1} - (W - z)^{-1}}{\epsilon} \\ &= -\lim_{\epsilon \to 0} (W + \epsilon \widetilde{W} - z)^{-1} \widetilde{W} (W - z)^{-1} = -G(W) \widetilde{W} G(W), \end{aligned} \tag{3.16}$$

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

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

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

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

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

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

and therefore, for instance,

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

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

**Lemma 3.5.** *We have*

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

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

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

*for* $k \geq 3$.

*Proof.* We start with the expansion

$$\begin{aligned}\Big(1+\mathcal{O}_{\prec}\Big(\frac{\psi_0^{\mathrm{av}}}{N\eta}\Big)\Big)\langle G_1A_1\cdots G_kA_k\rangle &= m_1\langle G_2\cdots G_kA_kA_1\rangle - m_1\langle \underline{WG_1}A_1G_2\cdots G_kA_k\rangle\\ &= m_1\langle G_2\cdots G_kA_kA_1\rangle + m_1\sum_{j=2}^{k}\langle G_1\cdots G_j\rangle\langle G_j\cdots G_kA_k\rangle\\ &\quad - m_1\langle \underline{WG_1A_1G_2\cdots G_kA_k}\rangle,\end{aligned} \tag{3.19}$$

due to

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

where for $k = 1$ the first two terms in the right-hand side of equation (3.19) are not present. In the second step, we extended the underline renormalisation to the entire product $\underline{WG_1A_1G_2\cdots G_kA_k}$ at the expense of generating additional terms collected in the summation; this identity can be obtained directly from the definition given by equation (3.15). Note that in the first line of equation (3.19), we moved the term coming from $m_1\langle G_1 - m_1\rangle G_1$ of equation (3.20) to the left-hand side, causing the error $\mathcal{O}_{\prec}(\psi_0^{\mathrm{av}}/(N\eta))$. For $k \geq 2$, using Lemmas 3.1 and 3.4, we estimate the second term in the second line of equation (3.19) by

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

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

$$m_1\langle G_2A_2A_1\rangle = m_1\langle G_2\rangle\langle A_2A_1\rangle + m_1\langle G_2(A_2A_1)^\circ\rangle = m_1\langle A_1A_2\rangle\Big(m_2+\mathcal{O}_{\prec}\Big(\frac{\psi_0^{\mathrm{av}}}{N\eta}\Big)\Big) + \mathcal{O}_{\prec}\Big(\frac{\psi_1^{\mathrm{av}}\rho^{1/2}}{(N\eta)^{1/2}}\Big), \tag{3.22}$$

where we used Lemma 3.4 to estimate

$$\langle|(A_2A_1)^\circ|^2\rangle^{1/2} = \Big(\langle|A_2A_1|^2\rangle - |\langle A_2A_1\rangle|^2\Big)^{1/2} \leq N^{1/2}. \tag{3.23}$$

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

In case $k \geq 3$, we estimate

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

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 equation (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 equations (3.21) and (3.24) and recalling $\psi_j^{\mathrm{av/iso}} \geq 1 \gtrsim \sqrt{\rho/(N\eta)}$ for all $j$, we obtain equation (3.17) with the definition of $\mathcal{E}_k$ from equation (3.18). □

Lemma 3.5 reduces understanding the local law to the underlined term in equation (3.19) since $\mathcal{E}_k^{\mathrm{av}}$ will be treated as an error term. For the underlined term, we use a cumulant expansion when calculating the high moment $\mathbf{E}|\langle (GA)^k - m^k A^k\rangle|^p$ for any fixed integer $p$. Here we will again make a notational simplification, ignoring different indices in $G$, $A$ and $m$; 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 $G = G(\overline{z_i})$ for half of the factors.

We set $\partial_{ab} := \partial/\partial w_{ab}$ as the derivative with respect to the $(a, b)$-entry of $W$: that is, 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 [35] and later revived in [33, 39]):

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

Technically, we use a truncated version of the expansion above; see, for example, [26, 33]. We thus compute[7]

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

[7]The 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 $GAG \cdots$ resolvent chains.

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

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 equation (3.27) used the fact that by definition of the underline renormalisation the $\partial_{ba}$-derivative in the first line may not act on its own $(GA)^k$.

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

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

from Lemma 3.1, estimating the deterministic leading term of $\langle (GA)^{2k}G\rangle$ by $|m^{(2)}m^{2k-1}\langle A^{2k}\rangle| \le N^{k-1}\rho/\eta$ as in equation (3.25). The first prefactor in the right-hand side of equation (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 equation (2.5).

For the second term of equation (3.27), we estimate

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

recalling that $G = G^t$ since $W$ is real symmetric.[8]

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

$$\mathcal{I}_k^{\mathrm{d}} := \left\{\boldsymbol{l} \in \mathbf{N}_0^n \,\middle|\, n \le R, \sum_i l_i = k\right\}, \quad R := (3+4k)p. \tag{3.29}$$

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

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

where for the multiset $J$, we define $\sum J := \sum_{\boldsymbol{j}\in J}|\boldsymbol{j}|$ and set

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

[8] We 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 equation (3.27) would not be present.

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

Here, for the multiset $J \subset \mathcal{I}_k^{\mathrm{d}}$, we defined its cardinality by $|J|$ and set $\sum J := \sum_{\boldsymbol{j}\in J} |\boldsymbol{j}|$. Along the product rule, the multi-index $\boldsymbol{l}$ encodes how the first factor $([(GA)^k]_{aa}$ in equation (3.30) is differentiated, while each element $\boldsymbol{j} \in J$ 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 equation (3.27), we similarly define the appropriate index set that is needed to encode the product rule[9]

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

Note that in addition to the multi-index $\boldsymbol{l}$ 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 $\partial_{ab}$ or $\partial_{ba}$ 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 equation (3.27) produces diagonal factors already after one derivative. The restriction in equation (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 $\partial_{ab}$ hits an off-diagonal term, then either one $aa$ and one $bb$ diagonal are created or none. Similarly, when an additional $\partial_{ab}$ hits a diagonal $aa$ term, then one diagonal $aa$ remains, along with a new off-diagonal $ab$. In any case, the difference between the $aa$ and $bb$ diagonals is unchanged.

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

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

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

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

Note that equation (3.33) is an overestimate: not all indices $(\boldsymbol{j}, \boldsymbol{\beta})$ indicated in equation (3.34) can actually occur after the Leibniz rule.

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

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

[9] In the definition of $\mathcal{I}_k^{\mathrm{od}}$ the indices $ab, ba, aa, bb$ should be understood symbolically, merely indicating the diagonal or off-diagonal character of the term. However, in equation (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 $a_1, a_2 = 1, \ldots, N$ in equation (3.34) and substituting $a_{(\alpha_i)_1}, a_{(\alpha_i)_2}$ for $\alpha_i$; however, this would be an excessive pedantry.

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

By combining Lemma 3.5 and equations (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

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

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

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

$$((GA)^k G)_{ab} \prec N^{k/2-1/2}\Big(\|A\boldsymbol{e}_a\| \wedge \|A\boldsymbol{e}_b\| + \psi_k^{\mathrm{iso}}\sqrt{\frac{\rho}{\eta}}\Big) \lesssim N^{k/2}\Big(1+\psi_k^{\mathrm{iso}}\sqrt{\frac{\rho}{N\eta}}\Big), \tag{3.37}$$

$$((GA)^k)_{ab} \prec N^{k/2-1/2}\|A\boldsymbol{e}_b\|\Big(1+\psi_{k-1}^{\mathrm{iso}}\sqrt{\frac{\rho}{N\eta}}\Big) \lesssim N^{k/2}\Big(1+\psi_{k-1}^{\mathrm{iso}}\sqrt{\frac{\rho}{N\eta}}\Big) \tag{3.38}$$

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

$$\sum_b \|A\boldsymbol{e}_b\|^2 = N\langle |A|^2\rangle \le N, \quad \sum_b \|A\boldsymbol{e}_b\| \le \sqrt{N}\sqrt{\sum_b \|A\boldsymbol{e}_b\|^2} \le N. \tag{3.39}$$

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

$$\Xi_k^{\mathrm{d/od}} \prec \Big(N^{k/2-1}\frac{\Omega_k}{\sqrt{N}}\Big)^{1+|J|} N^{1+\mathbf{1}(\mathrm{od})+(|J|-|\boldsymbol{l}|-\sum J)/2}, \tag{3.40}$$

where we defined

$$\Omega_k := \sum_{k_1+k_2+\cdots\le k}\prod_{i\ge 1}\Big(1+\psi_{k_i}^{\mathrm{iso}}\sqrt{\frac{\rho}{N\eta}}\Big). \tag{3.41}$$

Note that $\Omega_k \le \Phi_k$ just by choosing $k_1 = k_2 = 0$ in the definition of $\Phi_k$, equation (3.10), and thus $\Omega_k/\sqrt{N} \lesssim \Phi_k\sqrt{\rho/(N\eta)}$ since $1 \lesssim \rho/\eta$. Hence equation (3.35) follows trivially from equation (3.40) for $\Xi_k^{\mathrm{d}}$ and $\Xi_k^{\mathrm{od}}$ whenever $|\boldsymbol{l}| + \sum J \ge 2 + |J|$ or $|\boldsymbol{l}| + \sum J \ge 4 + |J|$, respectively: that is, when the exponent of $N$ in equation (3.40) is nonpositive.

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 $|\boldsymbol{l}| + \sum J - |J|$ that captures by how many factors of $N^{1/2}$ the naive estimate given by equation (3.40) exceeds the target in equation (3.35) when all $\Phi$s and $\psi$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 $\boldsymbol{\alpha}$ or in one of the tuples $\boldsymbol{\beta}$. Within case **O2**, we distinguish according to the length of $|\boldsymbol{l}|$ and $|J|$ as follows:

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

**D1** $|\boldsymbol{l}| + \sum J = |J| + 1$
**O1** $|\boldsymbol{l}| + \sum J = |J| + 3$
  **Ola** $ab \vee ba \in \boldsymbol{\alpha} \cup \bigcup_{(\boldsymbol{j},\boldsymbol{\beta})\in J} \boldsymbol{\beta}$
  **Olb** $J \in \{\{(\boldsymbol{j},(aa,bb))\},\{(\boldsymbol{j},(bb,aa))\}\}$ and $\boldsymbol{\alpha} \in \{(aa,bb),(bb,aa)\}$: that is, $\sum J = |\boldsymbol{l}| = 2$ and $|J| = 1$
**O2** $|\boldsymbol{l}| + \sum J = |J| + 2$
  **O2a** $|\boldsymbol{l}| = 1$,
  **O2b** $|\boldsymbol{l}| = 2$, $|J| \geq 2$,
  **O2c** $|\boldsymbol{l}| = 2$, $|J| = 1$, $l_1 \geq 1$,
  **O2d** $|\boldsymbol{l}| = 2$, $|J| = 1$, $l_1 = 0$.
**O3** $|\boldsymbol{l}| + \sum J = |J| + 1$

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

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

$$\begin{aligned}\Xi_k^{\mathrm{d}} &= N^{-3(1+|J|)/2}\Big|\sum_a [(GA)^k]_{aa}[(GA)^k G]_{aa}^{|J|}\Big| \\ &\prec \frac{(N^{k/2-1})^{|J|+1}}{N^{|J|/2+1}}\Omega_k^{|J|-1}\Omega_{k-1}\sum_a \|A\boldsymbol{e}_a\|\Big(\frac{\|A\boldsymbol{e}_a\|}{N^{1/2}} + \sqrt{\frac{\rho}{N\eta}}\psi_k^{\mathrm{iso}}\Big) \\ &\lesssim \Big(N^{k/2-1}\sqrt{\frac{\rho}{N\eta}}\Big)^{1+|J|}\Phi_k^{|J|-1}\Phi_{k-1}\psi_k^{\mathrm{iso}},\end{aligned} \tag{3.42}$$

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

For the off-diagonal 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 equation (3.37).
**I2** Averaging over $a$ in $|((GA)^k)_{ab}|$ gains a factor of $\sqrt{\rho/(N\eta)}$ compared to equation (3.38),

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

$$\begin{aligned}\frac{1}{N}\sum_a |[(GA)^k G]_{ab}| &\leq \frac{\sqrt{[(G^*A)^k G^* G (AG)^k]_{bb}}}{\sqrt{N}} \prec N^{k/2}\sqrt{\frac{\rho}{N\eta}}\Big(1 + \psi_{2k}^{\mathrm{iso}}\sqrt{\frac{1}{N\eta\rho}}\Big)^{1/2} \\ \frac{1}{N}\sum_a |[(GA)^k]_{ab}| &\leq \frac{\sqrt{[(AG^*)^k (GA)^k]_{bb}}}{\sqrt{N}} \prec N^{k/2-1/2}\|A\boldsymbol{e}_b\|\sqrt{\frac{\rho}{N\eta}}\Big(1 + \psi_{2(k-1)}^{\mathrm{iso}}\sqrt{\frac{1}{N\eta\rho}}\Big)^{1/2} \\ \frac{1}{N}\sum_a |[(GA)^k]_{ab}|^2 &= \frac{[(AG^*)^k (GA)^k]_{bb}}{N} \prec N^{k-1}\|A\boldsymbol{e}_b\|^2\frac{\rho}{N\eta}\Big(1 + \psi_{2(k-1)}^{\mathrm{iso}}\sqrt{\frac{1}{N\eta\rho}}\Big),\end{aligned} \tag{3.43}$$

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

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 equation (3.43).

Now we collect these gains over the naive bound given in equation (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 equation (3.40) that can be used to compensate the positive $N$-powers.

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

$$\Xi_k^{\mathrm{od}} \prec N^{-3(1+|J|)/2}(N^{k/2}\Omega_k)^{|J|-2}\sum_{ab}\big|[(GA)^k]_{ba}\big|\big|[(GA)^kG]_{ab}\big|\big|[(GA)^kG]_{ab}\big|. \tag{3.44}$$

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

$$\begin{aligned}&\sum_{ab}\big|[(GA)^k]_{ba}\big|\big|[(GA)^kG]_{ab}\big|\big|[(GA)^kG]_{ab}\big| \\ &\quad\lesssim N^{k-1}\sum_{ab}\big|[(GA)^k]_{ba}\big|\Big(\|A\boldsymbol{e}_a\|\|A\boldsymbol{e}_b\| + (\psi_k^{\mathrm{iso}})^2\frac{\rho}{\eta}\Big) \prec \Big(N^{k/2}\sqrt{\frac{\rho}{\eta}}\Big)^3\Phi_{k-1}(\psi_k^{\mathrm{iso}})^2,\end{aligned} \tag{3.45}$$

where in the second step, we performed a Schwarz inequality for the double $a, b$ summation and used the last bound in equations (3.43), (3.39) and $1 \lesssim \psi_k^{\mathrm{iso}}\sqrt{\rho/\eta}$. Thus, we conclude

$$\Xi_k^{\mathrm{od}} \prec \Big(N^{k/2-1}\sqrt{\frac{\rho}{N\eta}}\Big)^{|J|+1}\Phi_k^{|J|-2}\Phi_{k-1}(\psi_k^{\mathrm{iso}})^2. \tag{3.46}$$

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

$$\begin{aligned}\Xi_k^{\mathrm{od}} &\prec N^{-3(1+|J|)/2-1/2}(N^{k/2}\Omega_k)^{|J|-1}\sum_{ab}\big|[(GA)^k]_{ab}\big|\big|[(GA)^{j_1}G]_{\beta_1}\big|\big|[(GA)^{j_2}G]_{\beta_2}\big| \\ &\prec N^{-3(1+|J|)/2-1/2}(N^{k/2}\Omega_k)^{|J|-1}N^{k/2-1/2}\Omega_{j_2}\sum_{ab}\big|[(GA)^k]_{ab}\big|\Big(\|A\boldsymbol{e}_a\| + \|A\boldsymbol{e}_b\| + \psi_{j_1}^{\mathrm{iso}}\sqrt{\frac{\rho}{\eta}}\Big) \\ &\lesssim \Big(N^{k/2-1}\frac{\Omega_k}{\sqrt{N}}\Big)^{|J|-1}\Big(N^{k/2-1}\sqrt{\frac{\rho}{N\eta}}\Big)^2\Phi_{k-1}\psi_{j_1}^{\mathrm{iso}}\Omega_{j_2}\end{aligned} \tag{3.47}$$

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 equation (3.37) for $|[(GA)^{j_1}G]_{\beta_1}|$ and the second inequality in equation (3.37) for $|[(GA)^{j_2}G]_{\beta_2}|$. Finally, in the last step, we performed an $a, b$-Schwarz inequality, using the last bound in equations (3.43) and (3.39).

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

$$\begin{aligned}\Xi_k^{\mathrm{od}} &\prec N^{-3(1+|J|)/2-1/2}(N^{k/2}\Omega_k)^{|J|-2}\sum_{ab}|[(GA)^{l_1}G]_{\alpha_1}||[(GA)^{l_2}]_{\alpha_2}||[(GA)^kG]_{ab}|^2\\ &\prec N^{-3(1+|J|)/2-1/2}(N^{k/2}\Omega_k)^{|J|-2}N^{3k/2-3/2}\Omega_{k-1}\sum_{ab}(\|Ae_a\|+\|Ae_b\|)\Big(\|Ae_a\|\|Ae_b\|+\frac{\rho}{\eta}(\psi_k^{\mathrm{iso}})^2\Big)\\ &\lesssim \Big(N^{k/2-1}\frac{\Omega_k}{\sqrt{N}}\Big)^{|J|-2}N^{k/2-3/2}\Big(N^{k/2-1}\sqrt{\frac{\rho}{N\eta}}\Big)^2\Omega_{k-1}(\psi_k^{\mathrm{iso}})^2, \end{aligned}\tag{3.48}$$

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

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

$$\begin{aligned}\Xi_k^{\mathrm{od}} &\lesssim N^{-7/2}\sum_{ab}|[(GA)^{l_1}G]_{\alpha_1}||[(GA)^{l_2}]_{\alpha_2}||[(GA)^kG]_{ab}|\\ &\prec N^{k-5}\Omega_{l_2-1}\sum_{ab}\Big(\|Ae_b\|+\|Ae_a\|+\sqrt{\frac{\rho}{\eta}}\psi_{l_1}^{\mathrm{iso}}\Big)(\|Ae_a\|+\|Ae_b\|)\Big(\|Ae_a\|\wedge\|Ae_b\|+\sqrt{\frac{\rho}{\eta}}\psi_k^{\mathrm{iso}}\Big)\\ &\lesssim \Big(N^{k/2-1}\sqrt{\frac{\rho}{N\eta}}\Big)^2\Omega_{l_2-1}\psi_{l_1}^{\mathrm{iso}}\psi_k^{\mathrm{iso}}\end{aligned}\tag{3.49}$$

by equation (3.39).

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

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

and estimate

$$\begin{aligned}\Xi_k^{\mathrm{od}} &\lesssim N^{-7/2}\Big|\sum_{ab}G_{aa}[(GA)^k]_{bb}[(GA)^kG]_{ab}\Big| + N^{-7/2}\sum_{ab}\big|G_{ab}[(GA)^k]_{ab}[(GA)^kG]_{ab}\big|\\ &\prec N^{-7/2}\Big|\sum_b[(GA)^k]_{bb}[(GA)^kG]_{\mathbf{1}b}\Big| + N^{-7/2}\sqrt{\frac{\rho}{N\eta}}\sum_{ab}\big|[(GA)^k]_{bb}[(GA)^kG]_{ab}\big|\\ &\prec \sqrt{\frac{\rho}{\eta}}N^{k-4}\Omega_{k-1}\sum_b\|Ae_b\|\Big(\|Ae_b\|+\sqrt{\frac{\rho}{\eta}}\psi_k^{\mathrm{iso}}\Big) \lesssim (N^{k/2-1}\sqrt{\frac{\rho}{N\eta}}\Big)^2\Omega_{k-1}\psi_k^{\mathrm{iso}}\end{aligned}\tag{3.50}$$

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

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

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

$$\begin{aligned}&N^{-4}\Big|\sum_{ab}[(GA)^{k_1}G]_{aa}[(GA)^{k_2}]_{bb}[(GA)^{k_3}G]_{aa}[(GA)^{k_4}G]_{bb}\Big|\\ &\prec N^{k-5}\Omega_{k-1}\Omega_{k_3}\sum_{ab}\|Ae_b\|\Big(\|Ae_b\|+\psi_{k_4}^{\mathrm{iso}}\sqrt{\frac{\rho}{\eta}}\Big) \leq N^{k-3}\sqrt{\frac{\rho}{\eta}}\Omega_{k-1}\Big(1+\psi_{k_3}^{\mathrm{iso}}\sqrt{\frac{\rho}{N\eta}}\Big)(\psi_{k_4}^{\mathrm{iso}}+1)\\ &\lesssim \Big(N^{k/2-1}\sqrt{\frac{\rho}{N\eta}}\Big)^2\Omega_{k-1}\sum_{j=0}^{k}\psi_j^{\mathrm{iso}}\Omega_{k-j}\end{aligned}\tag{3.51}$$

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

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 equations (3.37) and (3.38) for the $k_2, k_4$ factors and the first inequality of equation (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 equation (3.37). □

### 3.2. *Proof of the isotropic estimate given by equation (3.9) in Proposition 3.2*

First we state the isotropic version of Lemma 3.5:

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

$$\left\langle \boldsymbol{x}, [(GA)^k G - m^{k+1}A^k]\boldsymbol{y} \right\rangle \left(1 + \mathcal{O}_{\prec}\left(\frac{\psi_0^{\mathrm{av}}}{N\eta}\right)\right) = -m\langle \boldsymbol{x}, \underline{W(GA)^k G}\boldsymbol{y}\rangle + \mathcal{O}_{\prec}\left(\mathcal{E}_k^{\mathrm{iso}}\right), \tag{3.52}$$

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

$$\mathcal{E}_k^{\mathrm{iso}} := N^{k/2}\sqrt{\frac{\rho}{N\eta}}\left(\psi_{k-1}^{\mathrm{iso}} + \frac{1}{N\eta}\sum_{j=1}^{k}\left(\psi_j^{\mathrm{av}} + \psi_{k-j}^{\mathrm{iso}} + \sqrt{\frac{\rho}{N\eta}}\psi_j^{\mathrm{av}}\psi_{k-j}^{\mathrm{iso}}\right)\right). \tag{3.53}$$

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

$$\begin{aligned}\left(1 + \mathcal{O}_{\prec}\left(\frac{\psi_0^{\mathrm{av}}}{N\eta}\right)\right)\langle \boldsymbol{x}, (GA)^k G\boldsymbol{y}\rangle &= m\langle \boldsymbol{x}, (AG)^k\boldsymbol{y}\rangle - m\langle \boldsymbol{x}, \underline{WG}(AG)^k\boldsymbol{y}\rangle \\ &= m^{k+1}\langle \boldsymbol{x}, A^k\boldsymbol{y}\rangle - m\langle \boldsymbol{x}, \underline{WG(AG)^k}\boldsymbol{y}\rangle \\ &\quad + m_1\sum_{j=1}^{k}\langle ((GA)^j G\rangle\langle \boldsymbol{x}, (GA)^{k-j}G\boldsymbol{y}\rangle \\ &\quad + \mathcal{O}_{\prec}\left(N^{(k-1)/2}\sqrt{\frac{\rho}{N\eta}}\|A\boldsymbol{x}\|\psi_{k-1}^{\mathrm{iso}}\right),\end{aligned} \tag{3.54}$$

where we used the definition in equation (3.3) for the first term and the definition in equation (3.15). An estimate analogous to equation (3.21) handles the sum and is incorporated in equation (3.53). This concludes the proof together with Lemma 3.1 and $\|A\boldsymbol{x}\| \leq \|A\| \leq N^{1/2}$. □

Exactly as in equation (3.27), we perform a cumulant expansion

$$\begin{aligned}&\mathbf{E}\langle \boldsymbol{x}, \underline{W(GA)^k G\boldsymbol{y}}\rangle\langle \boldsymbol{x}, [(GA)^k G - m^{k+1}A^k]\boldsymbol{y}\rangle^{p-1} \\ &= \mathbf{E}\sum_{ab}\frac{x_a[(GA)^k G]_{b\boldsymbol{y}}}{N}(\partial_{ab} + \partial_{ba})[(GA)^k G - m^{k+1}A^k]_{\boldsymbol{xy}}^{p-1} \\ &\quad + \mathbf{E}\sum_{ab}\frac{x_a\partial_{ab}[(GA)^k G]_{b\boldsymbol{y}}}{N}[(GA)^k G - m^{k+1}A^k]_{\boldsymbol{xy}}^{p-1} \\ &\quad + \sum_{j\geq 2}\frac{\kappa_{j+1}^{\mathrm{d}}}{j!N^{(j+1)/2}}\mathbf{E}\sum_a \partial_{aa}^j\left(x_a[(GA)^k G]_{a\boldsymbol{y}}[(GA)^k G - m^{k+1}A^k]_{\boldsymbol{xy}}^{p-1}\right) \\ &\quad + \sum_{j\geq 2}\frac{\kappa_{j+1}^{\mathrm{od}}}{j!N^{(j+1)/2}}\mathbf{E}\sum_{a\neq b}(\partial_{ab} + \partial_{ba})^j\left(x_a[(GA)^k G]_{b\boldsymbol{y}}[(GA)^k G - m^{k+1}A^k]_{\boldsymbol{xy}}^{p-1}\right),\end{aligned} \tag{3.55}$$

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

recalling Assumption 1 for the diagonal and off-diagonal cumulants. In fact, the formula in equation (3.55) is identical to equation (3.27) for $k+1$ instead of $k$ if the last $A = A_{k+1}$ in the product $(GA)^{k+1} = G_1A_1G_2A_2\dots G_{k+1}A_{k+1}$ is chosen specifically $A_{k+1} = \boldsymbol{yx}^*$.

For the first line of equation (3.55), after performing the derivative, we can also perform the summations and estimate the resulting isotropic resolvent chains by using the last inequality of equation (3.37) as well as Lemma 3.1 to obtain

$$\begin{aligned}
&\sum_{ab}\frac{x_a[(GA)^kG]_{by}}{N}(\partial_{ab}+\partial_{ba})[(GA)^kG-m^{k+1}A^k]_{\boldsymbol{xy}}^{p-1}\\
&=\sum_{j=0}^{2k}\frac{[(GA)^jG]_{\boldsymbol{xx}}[(GA)^kG(GA)^{k-j}G]_{\boldsymbol{yy}}+[(GA)^jG(GA)^kG]_{\boldsymbol{xy}}[(GA)^{k-j}G]_{\boldsymbol{xy}}}{N}\\
&\quad\times[(GA)^kG-m^{k+1}A^k]_{\boldsymbol{xy}}^{p-2}\\
&\prec\left(N^{k/2}\sqrt{\frac{\rho}{N\eta}}\right)^2\left(1+\sum_{j=0}^{2k}\frac{\psi_j^{\mathrm{iso}}}{\sqrt{N\eta\rho}}\left(1+\sqrt{\frac{\rho}{N\eta}}\psi_{2k-j}^{\mathrm{iso}}\right)\right)\left|[(GA)^kG-m^{k+1}A^k]_{\boldsymbol{xy}}\right|^{p-2}.
\end{aligned}\tag{3.56}$$

For the second line of equation (3.55), we estimate

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

For the third and fourth lines of equation (3.55), we distribute the derivatives according to the product rule to estimate (with the absolute value inside the summation to address both diagonal and off-diagonal terms)

$$\begin{aligned}
&\sum_{j\ge2}\frac{1}{N^{(j+1)/2}}\sum_{a,b}\left|(\partial_{ab}+\partial_{ba})^j\left(x_a[(GA)^kG]_{by}[(GA)^kG-m^{k+1}A^k]_{\boldsymbol{xy}}^{p-1}\right)\right|\\
&\qquad\le\sum_{\substack{\sum\boldsymbol{j}\ge2\\1\le|\boldsymbol{j}|\le p}}\Lambda_k(\boldsymbol{j})\left|[(GA)^kG-m^{k+1}A^k]_{\boldsymbol{xy}}\right|^{p-|\boldsymbol{j}|}
\end{aligned}\tag{3.57}$$

where

$$\Lambda_k(\boldsymbol{j}):=N^{(n-\sum\boldsymbol{j})/2}\sum_{ab}\left|\left((\partial_{ab}+\partial_{ba})^{j_0}\frac{x_a[(GA)^kG]_{by}}{\sqrt{N}}\right)\prod_{i=1}^{n}\left((\partial_{ab}+\partial_{ba})^{j_i}\frac{[(GA)^kG]_{\boldsymbol{xy}}}{\sqrt{N}}\right)\right|\tag{3.58}$$

and the summation in equation (3.57) is performed over all $\boldsymbol{j}=(j_0,\dots,j_n)\in\mathbf{N}_0^n$ with $j_0\ge0$, $j_1,\dots,j_n\ge1$ and $|\boldsymbol{j}|=n+1$. Recall that $\sum\boldsymbol{j}=j_0+j_1+j_2+\dots+j_n$.

**Lemma 3.8.** *For any admissible $\boldsymbol{j}$ in the summation of equation (3.57), it holds that*

$$\Lambda_k(\boldsymbol{j})\prec\left(N^{k/2}\sqrt{\frac{\rho}{N\eta}}\Phi_k\right)^{|\boldsymbol{j}|}.\tag{3.59}$$

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

By combining Lemmas 3.7 and 3.8 and equations (3.56), (3.57) and (3.58), we obtain

$$\big|\langle \boldsymbol{x}, [(GA)^k - m^{k+1}A^k]\boldsymbol{y}\rangle\big| \prec \mathcal{E}_k^{\mathrm{iso}} + N^{k/2}\sqrt{\frac{\rho}{N\eta}}\Big(\Phi_k + \frac{1}{N\eta}\sum_{j=0}^{2k}\sqrt{\psi_j^{\mathrm{iso}}\psi_{2k-j}^{\mathrm{iso}}} + \frac{\psi_k^{\mathrm{iso}}}{N\eta}\Big), \tag{3.60}$$

concluding the proof of equation (3.9).

*Proof of Lemma 3.8.* We recall the notations $\Omega_k, \Phi_k$ from equations (3.10) and (3.41). For a naive bound, we estimate all but the first factor trivially in equation (3.58) with

$$\left|(\partial_{ab} + \partial_{ba})^{j_i}\frac{[(GA)^k G]_{\boldsymbol{xy}}}{\sqrt{N}}\right| \prec \frac{N^{k/2}}{N^{1/2}}\Omega_k. \tag{3.61}$$

Note that the estimate is independent of the number of derivatives. For the first factor in equation (3.58), we estimate, after performing the derivatives, all but the last $[(GA)^{k_i}G]$-factor (involving $\boldsymbol{y}$) trivially by equation (3.37) as

$$\left|(\partial_{ab} + \partial_{ba})^{j_0}\frac{x_a[(GA)^k G]_{b\boldsymbol{y}}}{\sqrt{N}}\right| \prec \sum_{j=0}^{k} N^{(k-j)/2}\Omega_{k-j}|x_a|\frac{|[(GA)^j G]_{a\boldsymbol{y}}| + |[(GA)^j G]_{b\boldsymbol{y}}|}{\sqrt{N}}. \tag{3.62}$$

By combining equations (3.61) and (3.62) and the Schwarz-inequality

$$\begin{aligned}\sum_{ab}|x_a|\frac{|[(GA)^j G]_{a\boldsymbol{y}}| + |[(GA)^j G]_{b\boldsymbol{y}}|}{\sqrt{N}} &\le \sqrt{N}\|\boldsymbol{x}\|\sqrt{[(G^*A)^j G^* G(AG)^j]_{\boldsymbol{yy}}} \\ &\prec N^{j/2+1}\sqrt{\frac{\rho}{N\eta}}\Big(1 + \frac{\psi_{2j}^{\mathrm{iso}}}{\sqrt{N\eta\rho}}\Big)^{1/2},\end{aligned} \tag{3.63}$$

we conclude

$$\Lambda_k(\boldsymbol{j}) \prec N^{(n-\sum \boldsymbol{j})/2+1}N^{k/2}\sqrt{\frac{\rho}{N\eta}}\Phi_k\Big(N^{k/2}\frac{1}{\sqrt{N}}\Omega_k\Big)^{|\boldsymbol{j}|-1}, \tag{3.64}$$

which implies equation (3.59) in the case when $\sum \boldsymbol{j} \ge n+2$ using that $\Omega_k \le \Phi_k$ and $\rho/\eta \gtrsim 1$. It thus only remains to consider the cases $\sum \boldsymbol{j} = n$ and $\sum \boldsymbol{j} = n+1$.

If $\sum \boldsymbol{j} = n$, then $n \ge 2$ and $j_0 = 0$, $j_1 = j_2 = \cdots = 1$. By estimating the $j_2, j_3, \ldots$ factors in equation (3.58) using equation (3.61), we then bound

$$\begin{aligned}\Lambda_k(\boldsymbol{j}) &\prec \Big(N^{k/2}\frac{\Omega_k}{\sqrt{N}}\Big)^{|\boldsymbol{j}|-2}\sum_{ab}\frac{|x_a|\big|[(GA)^k G]_{b\boldsymbol{y}}\big|}{\sqrt{N}}\sum_{j=0}^{k}\frac{\big|[(GA)^j G]_{\boldsymbol{x}a}\big|\big|[(GA)^{k-j}G]_{b\boldsymbol{y}}\big|}{\sqrt{N}} \\ &\lesssim \Big(N^{k/2}\frac{\Omega_k}{\sqrt{N}}\Big)^{|\boldsymbol{j}|-2}\frac{\sqrt{[(G^*A)^k G^* G(AG)^k]_{\boldsymbol{yy}}}}{\sqrt{N}} \\ &\quad\times\sum_{j=0}^{k}\frac{\sqrt{[(G^*A)^j G^* G(AG)^j]_{\boldsymbol{yy}}\,[(G^*A)^{k-j}G^* G(AG)^{k-j}]_{\boldsymbol{xx}}}}{\sqrt{N}} \\ &\prec \Big(N^{k/2}\frac{\Omega_k}{\sqrt{N}}\Big)^{|\boldsymbol{j}|-2}\Big(N^{k/2}\sqrt{\frac{\rho}{N\eta}}\Big)^2\sqrt{\frac{\rho}{\eta}}\Phi_k\sum_{j=0}^{k}\Big(1 + \frac{\psi_{2j}^{\mathrm{iso}}}{\sqrt{N\eta\rho}}\Big)^{1/2}\Big(1 + \frac{\psi_{2(k-j)}^{\mathrm{iso}}}{\sqrt{N\eta\rho}}\Big)^{1/2} \lesssim \Big(N^{k/2}\sqrt{\frac{\rho}{N\eta}}\Phi_k\Big)^{|\boldsymbol{j}|}\end{aligned} \tag{3.65}$$

using $|\boldsymbol{j}| \ge 3$ and $\Omega_k \le \Phi_k$, $1 \lesssim \rho/\eta$ in the last step.

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

Finally, if $\sum \boldsymbol{j} = n+1$, then $n \geq 1$ by admissibility and either $j_0 = 0$ or $j_1 = 1$. In the first case, we estimate the $j_2, j_3, \ldots$ factors in equation (3.58) using equation (3.61) and all but the first $[(GA)^j G]_{\boldsymbol{x}\cdot}$ in the $j_1$-factor after differentiation trivially to obtain

$$\begin{aligned}\Lambda_k(\boldsymbol{j}) &\prec N^{-1/2}\Big(N^{k/2}\frac{\Omega_k}{\sqrt{N}}\Big)^{|\boldsymbol{j}|-2}\sum_{ab}\frac{|x_a|\big|[(GA)^k G]_{\boldsymbol{by}}\big|}{\sqrt{N}}\sum_{j=0}^{k}N^{(k-j)/2}\Omega_{k-j}\frac{\big|[(GA)^j G]_{\boldsymbol{x}a}\big|+\big|[(GA)^j G]_{\boldsymbol{x}b}\big|}{\sqrt{N}}\\ &\prec \Big(N^{k/2}\frac{\Omega_k}{\sqrt{N}}\Big)^{|\boldsymbol{j}|-2}\Big(N^{k/2}\sqrt{\frac{\rho}{N\eta}}\Phi_k\Big)^2, \end{aligned} \tag{3.66}$$

again using a Schwarz inequality. Finally, in the $j_1 = 1$ case, we estimate two $j_0$-factor using equation (3.62), the $j_2, j_3, \ldots$ factors trivially and to bound

$$\begin{aligned}\Lambda_k(\boldsymbol{j}) &\prec N^{-1/2}\Big(N^{k/2}\frac{\Omega_k}{\sqrt{N}}\Big)^{|\boldsymbol{j}|-2}\\ &\quad\times\sum_{ab}\sum_{j,l=0}^{k}N^{(k-l)/2}\Omega_{k-l}|x_a|\frac{|[(GA)^l G]_{a\boldsymbol{y}}|+|[(GA)^l G]_{b\boldsymbol{y}}|}{\sqrt{N}}\frac{\big|[(GA)^j G]_{\boldsymbol{x}a}\big|\big|[(GA)^{k-j} G]_{b\boldsymbol{y}}\big|}{\sqrt{N}}\\ &\prec \Big(N^{k/2}\frac{\Omega_k}{\sqrt{N}}\Big)^{|\boldsymbol{j}|-2}\Big(N^{k/2}\sqrt{\frac{\rho}{N\eta}}\Phi_k\Big)^2, \end{aligned} \tag{3.67}$$

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

### *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 $\psi_k^{\mathrm{av/iso}} \lesssim 1$ for any $k \geq 0$.

*Proof of Lemma 3.3.* The proof of this proposition is very similar to [18, 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

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

$$\lesssim \sqrt{\frac{N\eta}{\rho}} + \frac{N^{(3-2k)/2+1}\eta^{1/2}}{\rho^{1/2}\langle|A|^2\rangle^k}\left(N^{k/2-1}\langle|A|^2\rangle^{k/2} + \frac{\rho^{1/2}\langle|A|^2\rangle^{k/2}}{N^{(3-k)/2}\eta^{1/2}}\psi_k^{\mathrm{av}}\right)^2$$
$$\lesssim \sqrt{\frac{N\eta}{\rho}} + \sqrt{\frac{\rho}{N\eta}}(\psi_k^{\mathrm{av}})^2.$$

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

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

$$X \prec A + \frac{x}{B} + x^{1-\alpha}C^{\alpha} \tag{3.68}$$

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 $\delta, \alpha$ and $D$ are $N$-independent positive constants; other quantities may depend on $N$), then we also know that $X \prec x$ implies

$$X \prec A + C. \tag{3.69}$$

The proof is simply to iterate equation (3.68) finitely many times (depending only on $\delta, \alpha$ and $D$). The fact that $\Psi_k^{\mathrm{av/iso}} \prec N^D$ 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 equation (2.4), we know that $\psi_0^{\mathrm{av}} = \psi_0^{\mathrm{iso}} = 1$. Using the master inequalities in Proposition 3.2 and the reduction bounds from Lemma 3.3, in the first step, we will show that $\Psi_k^{\mathrm{av/iso}} \prec \rho^{-k/4}$ 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 $\Psi_k^{\mathrm{av/iso}} \prec 1$. 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 $\Psi_k^{\mathrm{av/iso}}$ than the trivial $\Psi_k^{\mathrm{av/iso}} \leq N^{k/2}\eta^{-k-1}$ 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 $\Psi_k^{\mathrm{av/iso}} \prec \rho^{-k/4}$ proceeds by a step-two induction: we first prove that $\Psi_k^{\mathrm{av,iso}} \prec \rho^{-k/4}$ for $k = 1, 2$ and then show that if $\Psi_n^{\mathrm{av/iso}} \prec \rho^{-n/4}$ holds for all $n \leq k-2$, for some $k \geq 4$, then it also holds for $\Psi_{k-1}^{\mathrm{av/iso}}$ and $\Psi_k^{\mathrm{av/iso}}$.

Using equations (3.8)–(3.9), we have

$$\begin{aligned}\Psi_1^{\mathrm{av}} &\prec 1 + \frac{\sqrt{\psi_2^{\mathrm{iso}}}}{(N\eta\rho)^{1/4}} + \frac{\sqrt{\psi_2^{\mathrm{av}}}}{(N\eta\rho)^{1/4}} + (\psi_1^{\mathrm{iso}})^{2/3} + \psi_1^{\mathrm{iso}}\sqrt{\frac{\rho}{N\eta}} + \frac{\psi_1^{\mathrm{av}}}{\sqrt{N\eta}} \\ \Psi_1^{\mathrm{iso}} &\prec 1 + \frac{\sqrt{\psi_2^{\mathrm{iso}}}}{(N\eta\rho)^{1/4}} + \frac{\psi_1^{\mathrm{av}}}{N\eta} + \psi_1^{\mathrm{iso}}\sqrt{\frac{\rho}{N\eta}} + \frac{\psi_1^{\mathrm{iso}}}{N\eta}\end{aligned} \tag{3.70}$$

for $k = 1$, using

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

Similarly, for $k = 2$, estimating explicitly

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

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

$$\begin{aligned}
\Psi_2^{\mathrm{av}} &\prec 1 + \psi_1^{\mathrm{av}} + \frac{\psi_2^{\mathrm{iso}}}{(N\eta\rho)^{1/12}} + \frac{\sqrt{\psi_4^{\mathrm{iso}}}}{(N\eta\rho)^{1/4}} + \frac{\sqrt{\psi_4^{\mathrm{av}}}}{(N\eta\rho)^{1/4}} + (\psi_2^{\mathrm{iso}})^{2/3} + \sqrt{\psi_1^{\mathrm{iso}}\psi_2^{\mathrm{iso}}} \\
&\quad + \frac{(\psi_2^{\mathrm{iso}})^{3/4}(\psi_1^{\mathrm{iso}})^{1/2}}{(N\eta\rho)^{1/8}} + \frac{\rho^{1/3}(\psi_2^{\mathrm{iso}})^{2/3}(\psi_1^{\mathrm{iso}})^{1/3}}{(N\eta\rho)^{1/6}} + \frac{\sqrt{\rho}(\psi_1^{\mathrm{av}})^2}{(N\eta)^{3/2}} + (\psi_1^{\mathrm{iso}})^2\frac{\rho}{N\eta} \\
&\quad + (\psi_1^{\mathrm{iso}})^{3/2}\sqrt{\frac{\rho}{N\eta}} + \frac{\rho^{1/2}\psi_1^{\mathrm{iso}}(\psi_2^{\mathrm{iso}})^{1/2}}{(N\eta\rho)^{1/4}} + \frac{\psi_2^{\mathrm{av}}}{\sqrt{N\eta}}, \\
\Psi_2^{\mathrm{iso}} &\prec 1 + \psi_1^{\mathrm{iso}} + \frac{\psi_1^{\mathrm{av}}}{N\eta} + \frac{\psi_2^{\mathrm{iso}} + \psi_2^{\mathrm{av}}}{(N\eta\rho)^{1/2}} + \frac{\sqrt{\psi_4^{\mathrm{iso}}}}{(N\eta\rho)^{1/4}} + \frac{\sqrt{\psi_1^{\mathrm{iso}}\psi_3^{\mathrm{iso}}}}{N\eta} + \frac{\psi_1^{\mathrm{av}}\psi_1^{\mathrm{iso}}}{(N\eta)^{3/2}} + (\psi_1^{\mathrm{iso}})^2\frac{\rho}{N\eta} + \frac{\psi_2^{\mathrm{iso}}}{N\eta}.
\end{aligned} \tag{3.71}$$

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

By equations (3.70)–(3.71), using iteration for the sum $\Psi_1^{\mathrm{av}} + \Psi_1^{\mathrm{iso}}$, we readily conclude

$$\Psi_1^{\mathrm{av}} + \Psi_1^{\mathrm{iso}} \prec 1 + \frac{\sqrt{\psi_2^{\mathrm{iso}}}}{(N\eta\rho)^{1/4}} + \frac{\sqrt{\psi_2^{\mathrm{av}}}}{(N\eta\rho)^{1/4}}. \tag{3.72}$$

Note that since equation (3.72) holds uniformly in the hidden parameters $A, z, \boldsymbol{x}, \boldsymbol{y}$ in $\Psi_1^{\mathrm{av/iso}}$, this bound serves as an upper bound on $\psi_1^{\mathrm{av}} + \psi_1^{\mathrm{iso}}$ (in the sequel, we will frequently use an already proven upper bound on $\Psi_k$ as an effective upper bound on $\psi_k$ in the next steps without explicitly mentioning it). Next, using this upper bound together with an iteration for $\Psi_2^{\mathrm{av}} + \Psi_2^{\mathrm{iso}}$, we have from equation (3.71)

$$\Psi_2^{\mathrm{av}} + \Psi_2^{\mathrm{iso}} \prec 1 + \frac{\sqrt{\psi_4^{\mathrm{iso}}}}{(N\eta\rho)^{1/4}} + \frac{\sqrt{\psi_4^{\mathrm{av}}}}{(N\eta\rho)^{1/4}} + \frac{\sqrt{\psi_1^{\mathrm{iso}}\psi_3^{\mathrm{iso}}}}{N\eta}, \tag{3.73}$$

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

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

$$\begin{aligned}
\Psi_4^{\mathrm{av}} &\prec \sqrt{\frac{N\eta}{\rho}} + \sqrt{\frac{\rho}{N\eta}}(\psi_2^{\mathrm{av}})^2 \\
\Psi_4^{\mathrm{iso}} &\prec \sqrt{\frac{N\eta}{\rho}} + \psi_2^{\mathrm{av}} + \psi_2^{\mathrm{iso}} + \sqrt{\frac{\rho}{N\eta}}\psi_2^{\mathrm{av}}\psi_2^{\mathrm{iso}} \\
\Psi_3^{\mathrm{iso}} &\prec \sqrt{\frac{N\eta}{\rho}} + \left(\frac{N\eta}{\rho}\right)^{1/4}\sqrt{\psi_2^{\mathrm{av}}} + \psi_2^{\mathrm{iso}} + \left(\frac{\rho}{N\eta}\right)^{1/4}\psi_2^{\mathrm{iso}}\sqrt{\psi_2^{\mathrm{av}}},
\end{aligned} \tag{3.74}$$

where the first inequality was already inserted into the right-hand side of equation (3.12) to get the second inequality in equation (3.74).

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

Then, inserting equations (3.74) and (3.72) into equation (3.73) and using iteration, we conclude

$$\Psi_2^{\mathrm{av}} + \Psi_2^{\mathrm{iso}} \prec \frac{1}{\sqrt{\rho}} + \frac{\sqrt{\psi_2^{\mathrm{iso}}} + \sqrt{\psi_2^{\mathrm{av}}}}{(N\eta\rho)^{1/4}} + \frac{\psi_2^{\mathrm{av}} + \psi_2^{\mathrm{iso}}}{(N\eta)^{1/2}}, \tag{3.75}$$

which together with equation (3.72) implies

$$\Psi_1^{\mathrm{iso}} + \Psi_1^{\mathrm{av}} \prec \rho^{-1/4}, \qquad \Psi_2^{\mathrm{iso}} + \Psi_2^{\mathrm{av}} \prec \rho^{-1/2}. \tag{3.76}$$

We now proceed with a step-two induction on $k$. The initial step of the induction is equation (3.76). Fix an even $k \geq 4$, and assume that $\Psi_n^{\mathrm{av/iso}} \prec \rho^{-n/4}$ holds for any $n \leq k-2$. In this case, by substituting this bound for $\psi_n^{\mathrm{av/iso}}$ whenever possible, for any even $l \leq k$, we have the following upper bounds on $\Phi_l$ and $\Phi_{l-1}$:

$$\begin{aligned}
\Phi_l \lesssim \mathfrak{E}_l &:= \frac{1}{\rho^{l/4}} + \sum_{j=0}^{l/2-1} \frac{\sqrt{\psi_{2l-2j}^{\mathrm{iso}}}}{(N\eta\rho)^{1/4}} \left(1 + \frac{1}{\rho^{j/4}(N\eta\rho)^{1/4}}\right) + \frac{\sqrt{\psi_l^{\mathrm{iso}}}}{(N\eta\rho)^{1/4}} \left(1 + \frac{1}{\rho^{l/8}(N\eta\rho)^{1/4}}\right) \mathbf{1}(l \leq k-2) \\
&\quad + \left[\psi_{k-1}^{\mathrm{iso}} \sqrt{\frac{\rho}{N\eta}} + \frac{\psi_k^{\mathrm{iso}}}{(N\eta\rho)^{1/2}}\right] \mathbf{1}(l = k), \\
\Phi_{l-1} \lesssim \mathfrak{E}_{l-1} &:= \frac{1}{\rho^{(l-1)/4}} + \sum_{j=1}^{l/2-1} \frac{\sqrt{\psi_{2l-2j}^{\mathrm{iso}}}}{(N\eta\rho)^{1/4}} \left(1 + \frac{1}{\rho^{(j-1)/4}(N\eta\rho)^{1/4}}\right) + \psi_{k-1}^{\mathrm{iso}} \sqrt{\frac{\rho}{N\eta}} \mathbf{1}(l = k).
\end{aligned} \tag{3.77}$$

We now plug equation (3.77) into equations (3.8) and (3.9) and, again using the bound $\Psi_n^{\mathrm{av/iso}} \prec \rho^{-n/4}$, $n \leq k-2$, whenever possible, get

$$\begin{aligned}
\Psi_{k-1}^{\mathrm{av}} &\prec \frac{1}{\rho^{(k-1)/4}} + \mathfrak{E}_{k-1} + \frac{\sqrt{\psi_{2k-2}^{\mathrm{av}}}}{(N\eta\rho)^{1/4}} + (\psi_{k-1}^{\mathrm{iso}})^{2/3} \mathfrak{E}_{k-2}^{1/3} + \sum_{j=1}^{k-2} \frac{1}{\rho^{j/8}} \sqrt{\mathfrak{E}_{k-1-j}(\psi_{k-1}^{\mathrm{iso}} + \mathfrak{E}_{k-2})} \\
&\quad + \frac{\psi_{k-1}^{\mathrm{av}}}{N\eta\rho^{1/4}} + \frac{\psi_{k-1}^{\mathrm{av}}}{\sqrt{N\eta}}, \\
\Psi_{k-1}^{\mathrm{iso}} &\prec \frac{1}{\rho^{(k-1)/4}} + \mathfrak{E}_{k-1} + \frac{\psi_{k-1}^{\mathrm{av}} + \psi_{k-1}^{\mathrm{iso}}}{N\eta} + \frac{1}{N\eta} \sum_{j=0}^{k-2} \frac{1}{\rho^{j/8}} \sqrt{\psi_{2k-2-j}^{\mathrm{iso}}} + \frac{\psi_{k-1}^{\mathrm{iso}}}{N\eta},
\end{aligned} \tag{3.78}$$

and

$$\begin{aligned}
\Psi_k^{\mathrm{av}} &\prec \frac{1}{\rho^{k/4}} + \mathfrak{E}_k + \frac{\sqrt{\psi_{2k}^{\mathrm{av}}}}{(N\eta\rho)^{1/4}} + \psi_{k-1}^{\mathrm{av}} + (\psi_k^{\mathrm{iso}})^{2/3} \mathfrak{E}_{k-1}^{1/3} + \sum_{j=1}^{k-2} \frac{1}{\rho^{j/8}} \sqrt{\mathfrak{E}_{k-j}(\psi_k^{\mathrm{iso}} + \mathfrak{E}_{k-1})} \\
&\quad + \sqrt{\psi_{k-1}^{\mathrm{iso}} \mathfrak{E}_1 (\psi_k^{\mathrm{iso}} + \mathfrak{E}_{k-1})} + \frac{\psi_k^{\mathrm{av}}}{\sqrt{N\eta}}, \\
\Psi_k^{\mathrm{iso}} &\prec \frac{1}{\rho^{k/4}} + \mathfrak{E}_k + \psi_{k-1}^{\mathrm{iso}} + \frac{\psi_k^{\mathrm{av}} + \psi_k^{\mathrm{iso}}}{N\eta} + \frac{\psi_{k-1}^{\mathrm{av}}}{N\eta} + \frac{1}{N\eta} \sum_{j=0}^{k-2} \frac{1}{\rho^{j/8}} \sqrt{\psi_{2k-j}^{\mathrm{iso}}} + \frac{\sqrt{\psi_{k-1}^{\mathrm{iso}} \psi_{k+1}^{\mathrm{iso}}}}{N\eta} + \frac{\psi_k^{\mathrm{iso}}}{N\eta}.
\end{aligned} \tag{3.79}$$

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

By iteration for $\Psi^{\mathrm{av}}_{k-1}+\Psi^{\mathrm{iso}}_{k-1}$ from equation (3.78), we thus get

$$\Psi^{\mathrm{av}}_{k-1}+\Psi^{\mathrm{iso}}_{k-1}\prec\frac{1}{\rho^{(k-1)/4}}+\mathfrak{E}_{k-1}+\frac{\sqrt{\psi^{\mathrm{av}}_{2k-2}}}{(N\eta\rho)^{1/4}}+\sum_{j=1}^{k-2}\frac{1}{\rho^{j/8}}\sqrt{\mathfrak{E}_{k-1-j}(\psi^{\mathrm{iso}}_{k-1}+\mathfrak{E}_{k-2})}+\frac{1}{N\eta}\sum_{j=1}^{k-2}\frac{1}{\rho^{j/8}}\sqrt{\psi^{\mathrm{iso}}_{2k-j}},\tag{3.80}$$

where we used that $\mathfrak{E}_{k-2}\le\mathfrak{E}_{k-1}$. Then using iteration for $\Psi^{\mathrm{av}}_{k}+\Psi^{\mathrm{iso}}_{k}$ from equation (3.79), we have

$$\begin{aligned}\Psi^{\mathrm{av}}_{k}+\Psi^{\mathrm{iso}}_{k}\prec{}&\frac{1}{\rho^{k/4}}+\mathfrak{E}_{k}+\frac{\sqrt{\psi^{\mathrm{av}}_{2k}}}{(N\eta\rho)^{1/4}}+\sum_{j=1}^{k-2}\frac{1}{\rho^{j/8}}\sqrt{\mathfrak{E}_{k-j}(\psi^{\mathrm{iso}}_{k}+\mathfrak{E}_{k-1})}+\frac{1}{N\eta}\sum_{j=0}^{k-2}\frac{1}{\rho^{j/8}}\sqrt{\psi^{\mathrm{iso}}_{2k-j}}\\&+\sqrt{\psi^{\mathrm{iso}}_{k-1}\mathfrak{E}_{1}(\psi^{\mathrm{iso}}_{k}+\mathfrak{E}_{k-1})}+\psi^{\mathrm{av}}_{k-1}+\psi^{\mathrm{iso}}_{k-1}+\frac{\sqrt{\psi^{\mathrm{iso}}_{k-1}\psi^{\mathrm{iso}}_{k+1}}}{N\eta},\end{aligned}\tag{3.81}$$

where we used that $\mathfrak{E}_{k-1}\le\mathfrak{E}_{k}$.

We will now use the reduction inequalities from Lemma 3.3 in the following form:

$$\Psi^{\mathrm{av}}_{2j}\prec\begin{cases}\sqrt{\frac{N\eta}{\rho}}+\sqrt{\frac{\rho}{N\eta}}(\psi^{\mathrm{av}}_{k})^2 & j=k,\\ \sqrt{\frac{N\eta}{\rho}}+\psi^{\mathrm{av}}_{k}+\frac{1}{\rho^{(k-2)/4}}+\frac{1}{(N\eta\rho)^{1/2}\rho^{(k-6)/4}}\psi^{\mathrm{av}}_{k} & j=k-1,\\ \sqrt{\frac{N\eta}{\rho}}+\frac{1}{\rho^{(j+1)/4}}+\frac{1}{(N\eta\rho)^{1/2}\rho^{(j-2)/2}} & j\le k-2,\end{cases}\tag{3.82}$$

and

$$\Psi^{\mathrm{iso}}_{l+(l-2j)}\prec\sqrt{\frac{N\eta}{\rho}}\left(1+\left(\frac{\rho}{N\eta}\right)^{1/4}(\psi^{\mathrm{av}}_{2(l-2j)})^{1/2}\right)\left(1+\sqrt{\frac{\rho}{N\eta}}\psi^{\mathrm{iso}}_{l}\right)\lesssim\frac{(N\eta\rho)^{1/2}}{\rho^{(l-j)/2}}\tag{3.83}$$

for any $j\le l/2$, where $l\le k-2$ is even. In the last step, we also used the last line of equation (3.82) to estimate $\psi^{\mathrm{av}}_{2(l-2j)}$. Then by plugging equation (3.83) into equation (3.77), we readily conclude that

$$\mathfrak{E}_{r}\lesssim\frac{1}{\rho^{r/4}}+\frac{\sqrt{\psi^{\mathrm{iso}}_{2k}}}{(N\eta\rho)^{1/4}}\mathbf{1}(r=k)\tag{3.84}$$

for any $r\le k$.

Plugging equation (3.84) into equations (3.80) and (3.81) and using iteration, we conclude

$$\begin{aligned}\Psi^{\mathrm{av}}_{k-1}+\Psi^{\mathrm{iso}}_{k-1}\prec{}&\frac{1}{\rho^{(k-1)/4}}+\frac{\sqrt{\psi^{\mathrm{av}}_{2k-2}}}{(N\eta\rho)^{1/4}}+\frac{1}{N\eta}\sum_{j=1}^{k-2}\frac{1}{\rho^{j/8}}\sqrt{\psi^{\mathrm{iso}}_{2k-j}}\\ \Psi^{\mathrm{av}}_{k}+\Psi^{\mathrm{iso}}_{k}\prec{}&\frac{1}{\rho^{k/4}}+\frac{\sqrt{\psi^{\mathrm{av}}_{2k}}+\sqrt{\psi^{\mathrm{iso}}_{2k}}}{(N\eta\rho)^{1/4}}+\frac{1}{N\eta}\sum_{j=0}^{k-2}\frac{1}{\rho^{j/8}}\sqrt{\psi^{\mathrm{iso}}_{2k-j}}+\frac{1}{\rho^{1/8}}\sqrt{\psi^{\mathrm{iso}}_{k-1}\psi^{\mathrm{iso}}_{k}}+\psi^{\mathrm{av}}_{k-1}+\psi^{\mathrm{iso}}_{k-1}\\&+\frac{\sqrt{\psi^{\mathrm{iso}}_{k-1}\psi^{\mathrm{iso}}_{k+1}}}{N\eta}.\end{aligned}\tag{3.85}$$

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

We will now additionally use that by equation (3.12) for any $r \in \{k-1, k\}$, we have

$$\begin{aligned}\Psi_{k+r}^{\mathrm{iso}} &\prec \sqrt{\frac{N\eta}{\rho}}\left(1+\left(\frac{\rho}{N\eta}\right)^{1/4}(\psi_{2r}^{\mathrm{av}})^{1/2}\right)\left(1+\sqrt{\frac{\rho}{N\eta}}\psi_k^{\mathrm{iso}}\right)\\ &\lesssim \left(\frac{N\eta}{\rho}\right)^{1/2}\left(1+\sqrt{\frac{\rho}{N\eta}}\psi_k^{\mathrm{iso}}\right)\times\begin{cases}1+\sqrt{\frac{\rho}{N\eta}}\psi_k^{\mathrm{av}} & r=k,\\ 1+\left(\frac{\rho}{N\eta}\right)^{1/4}\sqrt{\psi_k^{\mathrm{av}}}+\frac{1}{\rho^{(k-2)/8}}+\frac{1}{(N\eta\rho)^{1/4}\rho^{(k-6)/8}}\sqrt{\psi_k^{\mathrm{av}}} & r=k-1,\end{cases}\end{aligned} \tag{3.86}$$

and that

$$\Psi_{2k-j}^{\mathrm{iso}} = \Psi_{k+(k-j)}^{\mathrm{iso}} \prec (N\eta\rho)^{1/2}\frac{1}{\rho^{(2k-j)/4}}$$

for any $2 \le j \le k-1$.

Plugging these bounds, together with equation (3.82) for $j = k-1$ and $j = k$, into equation (3.85), and using iteration first for $\Psi_{k-1}^{\mathrm{av}} + \Psi_{k-1}^{\mathrm{iso}}$ and then for $\Psi_k^{\mathrm{av}} + \Psi_k^{\mathrm{iso}}$, we conclude that $\Psi_{k-1}^{\mathrm{av/iso}} \prec \rho^{-(k-1)/4}$ and that $\Psi_k^{\mathrm{av/iso}} \prec \rho^{-k/4}$. This completes the step-two induction and hence the first and pivotal step of the proof.

In the second step, we improve the bounds $\Psi_k^{\mathrm{av/iso}} \prec \rho^{-k/4}$ to $\Psi_k^{\mathrm{av/iso}} \prec 1$ for all $k$. Recall that the definition of stochastic domination given by equation (1.8) involved an arbitrary small but fixed exponent $\xi$. Now we fix this $\xi$, a large exponent $D$, and fix an upper threshold $K$ for the indices. Our goal is to prove that

$$\max_{k\le K}\sup_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{A},\boldsymbol{z}} \mathbf{P}\left[\Psi_k^{\mathrm{av/iso}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{A},\boldsymbol{z}) > N^{\xi}\right] \le N^{-D}, \tag{3.87}$$

where the supremum over all indicated parameters are meant in the sense described below equation (3.3).

Now we distinguish two cases in the supremum over the collection of spectral parameters $\boldsymbol{z}$ in equation (3.87). In the regime where $\rho = \rho(\boldsymbol{z}) \ge N^{-\xi/K}$, the bounds $\Psi_k^{\mathrm{av/iso}} \prec \rho^{-k/4}$, already proven for all $k$, imply equation (3.87). Hence we can work in the opposite regime where $\rho < N^{-\xi/K}$, and from now on, we restrict the supremum in equation (3.87) to this parameter regime. By plugging this bound into the master inequalities in Proposition 3.2 and noticing that $\Phi_k \le 1 + \rho^{-k/4}(N\eta\rho)^{-1/4}$, we directly conclude that

$$\Psi_k^{\mathrm{av/iso}} \prec 1 + \rho^{-k/4}[\rho^{1/4} + (N\eta\rho)^{-1/12}] \le 1 + \rho^{-k/4}[N^{-\xi/4K} + (N\eta\rho)^{-1/12}] \tag{3.88}$$

for any $k \ge 0$. Now we can use this improved inequality by again plugging it into the master inequalities to achieve

$$\Psi_k^{\mathrm{av/iso}} \prec 1 + \rho^{-k/4}[N^{-\xi/4K} + (N\eta\rho)^{-1/12}]^2 \tag{3.89}$$

and so on. Recalling the assumption that $N\eta\rho \ge N^{\epsilon}$ and recalling that $\rho \gtrsim \eta^{1/2} \ge N^{-1/3}$, we need to iterate this process finitely many times (depending on $k, \xi, K, \epsilon$) to also achieve $\Psi_k^{\mathrm{av/iso}} \prec 1$ in the second regime. This concludes the proof of the theorem. □

## 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 [17, Theorem 2.2], which is presented in [17, Sections 3 and 4]. For this reason, we only explain the differences and refer

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

to [17] for a fully detailed proof. We mention that the proof in [17] relies heavily on the theory of the stochastic eigenstate equation initiated in [10] and then further developed in [12, 43].

Similarly to [17, 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: that is, we consider the Dyson Brownian motion (DBM) with initial condition $W$ and show that the overlaps of the eigenvectors have Gaussian fluctuations after a time $t$ slightly bigger than $N^{-1}$. With a separate argument 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

$$\mathrm{d}W_t = \frac{\mathrm{d}\widetilde{B}_t}{\sqrt{N}}, \qquad W_0 = W, \tag{4.1}$$

where $\widetilde{B}_t$ is a standard real symmetric matrix Brownian motion (see, for example, [10, 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 $N \to \infty$, 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

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

with initial condition $m(z) = m_{\mathrm{sc}}(z)$ 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 $\rho_t(x+\mathrm{i}0)$ is a rescaling of $\rho_0 = \rho_{\mathrm{sc}}$ by a factor $1+t$. In fact, $W_t$ is a Wigner matrix itself, with a normalisation $\mathbf{E}\,|(W_t)_{ab}|^2 = N^{-1}(1+t)$ with a Gaussian component.

Denote by $\lambda_1(t) \le \lambda_2(t) \le \cdots \le \lambda_N(t)$ the eigenvalues of $W_t$, and let $\{\boldsymbol{u}_i(t)\}_{i\in[N]}$ be the corresponding eigenvectors. Then it is known [10, Theorem 2.3] that $\lambda_i = \lambda_i(t)$, $\boldsymbol{u}_i = \boldsymbol{u}_i(t)$ are the unique strong solutions of the following system of stochastic differential equations:

$$\mathrm{d}\lambda_i = \frac{\mathrm{d}B_{ii}}{\sqrt{N}} + \frac{1}{N}\sum_{j\neq i}\frac{1}{\lambda_i - \lambda_j}\,\mathrm{d}t \tag{4.3}$$

$$\mathrm{d}\boldsymbol{u}_i = \frac{1}{\sqrt{N}}\sum_{j\neq i}\frac{\mathrm{d}B_{ij}}{\lambda_i - \lambda_j}\boldsymbol{u}_j - \frac{1}{2N}\sum_{j\neq i}\frac{\boldsymbol{u}_i}{(\lambda_i - \lambda_j)^2}\,\mathrm{d}t, \tag{4.4}$$

where $B_t = (B_{ij})_{i,j\in[N]}$ is a standard real symmetric matrix Brownian motion (see, for example, [10, Definition 2.1]).

Note that the flow for the diagonal overlaps $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle$, by equation (4.4), naturally also depends on the off-diagonal overlap $\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle$. 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 [12], Bourgade, Yau and Yin proved that the *perfect matching observable* $f_{\lambda,t}$, which is presented in equation (4.6) below, satisfies a parabolic PDE (see equation (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, $\langle A\rangle = 0$: that is, $A = \mathring{A}$. We introduce the shorthand notation for the *eigenvector overlaps*

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

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 [10] and heavily used in [12, 43]. Each particle configuration is encoded by a function $\boldsymbol{\eta} : [N] \to \mathbf{N}_0$, 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 $\Omega^n$. Moreover, for any index pair $i \neq j \in [N]$, we define $\boldsymbol{\eta}^{ij}$ to be the configuration obtained moving a particle from site $i$ to site $j$; if there is no particle in $i$, then $\boldsymbol{\eta}^{ij} := \boldsymbol{\eta}$.

We now define the *perfect matching observable* (introduced in [12]) for any given configuration $\boldsymbol{\eta}$:

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

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

$$\mathcal{V}_{\boldsymbol{\eta}} := \{(i,a) : 1 \le i \le n, 1 \le a \le 2\eta_i\},$$

and

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

where $e = \{(i_1,a_1),(i_2,a_2)\} \in \mathcal{V}_{\boldsymbol{\eta}}^2$, and $\mathcal{E}(G)$ denotes the edges of $G$. Note that in equation (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 formalise it, we define

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

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

$$\int_{-\infty}^{\gamma_i(t)} \rho_t(x)\,\mathrm{d}x = \frac{i}{N}, \qquad i \in [N], \tag{4.9}$$

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 any $t, s \ge 0$.

The well-known rigidity estimate (see, for example, [24, Theorem 7.6] or [29]) asserts that

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

for any (small) $\xi > 0$ and (large) $D > 0$. This was proven for any fixed $t$: for example, in [24, Theorem 7.6] or [29], the extension to all $t$ follows by a grid argument together with the fact that $\boldsymbol{\lambda}(t)$ is stochastically $1/2$-Hölder in $t$, which follows by Weyl's inequality

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

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

By [12, Theorem 2.6], we know that the perfect matching observable $f_{\lambda,t}$ is a solution of the following parabolic discrete PDE

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

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

where

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

Note that the number of particles $n = n(\boldsymbol{\eta})$ is preserved under the flow of equation (4.10). The eigenvalue trajectories are fixed in this proof; hence we will often omit $\boldsymbol{\lambda}$ from the notation: for example, 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 (compare to [17, 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*

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

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

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 equation (2.9) follows immediately from equation (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 [17, Proof of Theorem 2.2] for a more detailed proof. □

### *4.2. DBM analysis*

Since the current DBM analysis of equation (4.10) heavily relies on [17, Section 4], before starting it, we introduce an equivalent representation of equation (4.6) used in [17] (which itself is based on the particles representation from [43]).

Fix $n \in \mathbf{N}$, and consider configurations $\boldsymbol{\eta} \in \Omega^n$: that is, such that $\sum_j \eta_j = n$. We now give an equivalent representation of equations (4.10) and (4.11) that is defined on the $2n$-dimensional lattice $[N]^{2n}$ instead of configurations of $n$ particles (see [17, Section 4.1] for a more detailed description). Let $\boldsymbol{x} \in [N]^{2n}$, and define the configuration space

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

where

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

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

The correspondence between these two representations is given by

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

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

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

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

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

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

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

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

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

where

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

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

We now define the scalar product and the natural measure on $\Lambda^n$:

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

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

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

By [43, Appendix A.2], it follows that the operator $\mathcal{L} = \mathcal{L}(t)$ is symmetric with respect to the measure $\pi$, and it is a negative operator on $L^2(\Lambda^n)$ with Dirichlet form

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

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

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

#### 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 equation (4.26) below) of $g_t$ that will be exponentially small, evaluated on $\boldsymbol{x}$s that are not fully supported in the bulk.

Recall the definition of the quantiles $\gamma_i(0)$ from equation (4.9). Then we define the sets

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

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:

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

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

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

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

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

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

where

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

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

#### 4.2.2. $L^2$-bound

By standard finite speed propagation estimates (see [17, 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 $\Lambda^n$. Then for any $\boldsymbol{x} \in \Lambda^n$ supported on $\mathcal{J}$, it holds*

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

*for any small $\xi > 0$. The implicit constant in equation (4.13) depends on $n$, $\epsilon$, $\delta$.*

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 we will prove a bound in the $L^2$-sense that will be enhanced to an $L^\infty$ bound by standard parabolic regularity arguments.

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

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

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

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

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

with $\eta_k = \eta_k(t)$ defined by $N\eta_k\rho(\gamma_{i_k}(t)+\mathrm{i}N^{-2/3}) = N^{\omega}$. 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 equation (4.29) for $k = 2$, we have

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

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

$$|\langle \boldsymbol{u}_i(t), A\boldsymbol{u}_j(t)\rangle|^2 \le \frac{N^{2\omega}\langle A^2\rangle}{N[\rho(\gamma_i(t)+\mathrm{i}N^{-2/3})\wedge\rho(\gamma_j(t)+\mathrm{i}N^{-2/3})]} \qquad \text{on } \widehat{\Omega}_{\omega,\xi}\cap\widetilde{\Omega}_{\xi} \tag{4.31}$$

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

**Remark 4.3.** The set $\widehat{\Omega}$ defined in equation (4.29) is slightly different from its analogue[10] in [17, Eq. (4.20)]. First, all the error terms now explicitly depend on $\langle A^2\rangle$, whilst in [17, Eq. (4.20)], we just bounded the error terms using the operator norm of $A$ (which was smaller than 1 in [17, Eq. (4.20)]). Second, we have a slightly weaker bound (compared to [17, 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

[10]The definition of $\widehat{\Omega}$ in the published version of [17, Equation (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 [17]; Remark 4.3 refers to the corrected version.

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

optimally. As a consequence of this slightly worse bound close to the edges, we get the overlap bound in equation (4.31) instead of the optimal bound [17, Equation (4.21)]; however, this difference will not cause any change in the result. We remark that the bound in equation (4.31) is optimal for bulk indices.

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

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

*with*

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

*uniformly for particle configuration* $\boldsymbol{y} \in \Lambda^n$ *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 [17, Proposition 4.5]; hence we will only explain the main differences. The reader should consult [17] for a fully detailed proof. The key idea is to replace the operator $\mathcal{S}(t)$ in equations (4.26) and (4.27) by the following operator

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

where

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

and $a^{\mathcal{S}}_{ij}$ are their short-range version defined as in equation (4.25) and

$$\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}} := \boldsymbol{x} + \left(\prod_{r=1}^{n} \delta_{x_{a_r} i_r}\delta_{x_{b_r} i_r}\right) \sum_{r=1}^{n} (j_r - i_r)(\boldsymbol{e}_{a_r} + \boldsymbol{e}_{b_r}). \tag{4.36}$$

We remark that $\boldsymbol{x}^{ij}_{ab}$ from equation (4.20) changes two entries of $\boldsymbol{x}$ per time; instead, $\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}}$ changes all the coordinates of $\boldsymbol{x}$ at the same time: that is, let $\boldsymbol{i} := (i_1, \dots, i_n), \boldsymbol{j} := (j_1, \dots, j_n) \in [N]^n$, with $\{i_1, \dots, i_n\} \cap \{j_1, \dots, j_n\} = \emptyset$; then $\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}} \neq \boldsymbol{x}$ iff for all $r \in [n]$, it holds that $x_{a_r} = x_{b_r} = i_r$. This means $\mathcal{S}(t)$ 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 [17]). *Let* $\mathcal{S}(t)$, $\mathcal{A}(t)$ *be the generators defined in equations (4.27) and (4.34), respectively, and let* $\mu$ *denote the uniform measure on* $\Lambda^n$ *for which* $\mathcal{A}(t)$ *is reversible. Then there exists a constant* $C(n) > 0$ *such that*

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

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

Next, combining

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

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

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

The star over $\sum$ means summation over two $n$-tuples of fully distinct indices. Then proceeding as in the proof of [17, Proposition 4.5], we conclude that

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

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

We point out that to go from equation (4.39) to equation (4.40), we proceed exactly as in the proof of [17, Proposition 4.5] (with the additional $\langle A^2 \rangle^{k/2}$, $\langle A^2 \rangle^{n/2}$ factors in [17, Equation (4.47)] and [17, Equation (4.48)], respectively) except for the estimate in [17, Equation (4.43)]. The error terms in this estimate used that $|P(G)| \le N^{n\xi - n/2}$ uniformly in the spectrum, a fact that we cannot establish near the edges as a consequence of the weaker bound in equation (4.31). We now explain how we can still prove [17, Equation (4.43)] in the current case. The main mechanism is that the strong bound $|P(G)| \le N^{n\xi - n/2} \langle A^2 \rangle^{n/2}$ 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 equation (4.31) still provides a nontrivial bound of order $N^{-1/3}$ for $|\langle \boldsymbol{u}_i, A \boldsymbol{u}_j \rangle|$ since $\rho(\gamma_i(t) + \mathrm{i}N^{-2/3}) \gtrsim N^{-1/3}$ uniformly in $0 \le t \le T$.

We start with removing the short-range cutoff from the kernel $a_{ij}^{\mathcal{S}}(t)$ in the left-hand side of [17, Equation (4.43)]:

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

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

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 equation (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 $|j_r - i_r| \le \ell$ for any $r \in \{2, \dots, n\}$. The

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

proof in the other cases is completely analogous and so omitted. Additionally, to make our presentation easier, we assume that $n = 2$:

$$\begin{aligned}&\sum_{\substack{|j_1-i_1|>\ell,|j_2-i_2|\le\ell,\\ j_1\neq j_2}} a_{i_1j_2}(t)a_{i_1j_2}(t)\left(\frac{N}{2\langle A^2\rangle}\sum_{G\in\mathcal{G}_{\eta^{\boldsymbol{ij}}}} P(G)\right)\\ &=\left(\sum_{\substack{cN\ge|j_1-i_1|>\ell,|j_2-i_2|\le\ell,\\ j_1\neq j_2}} + \sum_{\substack{|j_1-i_1|>cN,|j_2-i_2|\le\ell,\\ j_1\neq j_2}}\right)a_{i_1j_2}(t)a_{i_1j_2}(t)\left(\frac{N}{2\langle A^2\rangle}\sum_{G\in\mathcal{G}_{\eta^{\boldsymbol{ij}}}} P(G)\right).\end{aligned} \tag{4.42}$$

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

$$\begin{aligned}\sum_{\substack{|j_1-i_1|>cN,|j_2-i_2|\le\ell,\\ j_1\neq j_2}} a_{i_1j_1}(t)a_{i_2j_2}(t)\left(\frac{N}{2\langle A^2\rangle}\sum_{G\in\mathcal{G}_{\eta^{\boldsymbol{ij}}}} P(G)\right) &\lesssim \frac{N^{1+\xi}\eta}{N^{2/3}}\sum_{|j_2-i_2|\le\ell} a_{i_2j_2}(t)\\ &\lesssim \frac{N^{1+\xi}\eta}{N^{2/3}} \le \frac{N\eta}{\ell},\end{aligned} \tag{4.43}$$

where we used that $a_{i_1j_1}(t) \lesssim \eta N^{-1}$, $\ell \ll K \ll \sqrt{N}$ and that

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

by equation (4.31). We point out that to go from the first to the second line of equation (4.43), we also used that $\sum_{j_2} a_{i_2j_2}(t) \lesssim 1$ on $\widehat{\Omega}$. This concludes the proof that the last line of equation (4.41) is bounded by $N^{1+n\xi}\eta\ell^{-1}$. We thus conclude that

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

Proceeding in a similar way – that is, 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 equation (4.44):

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

Finally, by equations (4.44) and (4.45), we conclude

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

which is exactly the same as [17, Equation (4.43)]. Given equation (4.46), the remaining part of the proof of this proposition is completely analogous to the proof of [17, Proposition 4.5]; the only difference is that now in [17, Eq. (4.48)], using that $|m_t(z_i)| \lesssim 1$ uniformly in $0 \le t \le T$, we will have an additional error term

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

coming from the deterministic term in equation (4.29) (the mixed terms when we use the error term in equation (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

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

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

#### 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 $L^\infty$-bound and hence proving Proposition 4.1 is completely analogous to the proof of [17, Proposition 3.2] presented in [17, Section 4.4] and so omitted.

## A. Proof of Theorem 2.2 in the large $d$ regime

The $d \ge 10$ regime is much simpler mainly because the trivial norm bound $\|G(z)\| \le 1/d$ 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 analogous[11] to [18, Appendix B]; hence we will be very brief.

We now assume that equation (2.5) has been proven up to some $k-1$ in the $d \ge 10$ regime. Using equation (3.19) and estimating all resolvent chains in the right-hand side of equation (3.19) by the induction hypotheses (after splitting $A_kA_1 = \langle A_kA_1\rangle + (A_kA_1)^\circ$), using the analogue of Lemma 3.1 to estimate $\langle (GA)^{j-1}G\rangle$ in terms of the induction hypothesis, we easily obtain

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

in place of Lemma 3.5. In estimating the leading terms in equation (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 $\langle m^kA^k\rangle$ under the normalisation $\langle |A|^2\rangle = 1$, and the small factor $1/Nd^2$ represents the smallness of the negligible

[11] We point out that the $N$-scaling here is naturally different from that in [18, Appendix B] simply due to the fact that here we chose the normalisation $\langle |A_i|^2\rangle = 1$ instead of $\|A_i\| = 1$.

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

error term. We now follow the argument in Section 3 starting from equation (3.26). For the Gaussian term in equation (3.28), we simply bound

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

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

## B. Green function comparison

The Green function comparison argument is very similar to the one presented in [17, Appendix A]; hence we only explain the minor differences.

Consider the Ornstein-Uhlenbeck flow

$$\mathrm{d}\widehat{W}_t = -\frac{1}{2}\widehat{W}_t \, \mathrm{d}t + \frac{\mathrm{d}\widehat{B}_t}{\sqrt{N}}, \qquad \widehat{W}_0 = W, \tag{B.1}$$

with $\widehat{B}_t$ a real symmetric Brownian motion. Along the OU-flow in equation (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

$$\widehat{W}_T \stackrel{\mathrm{d}}{=} \sqrt{1 - cT}\widetilde{W} + \sqrt{cT}U, \tag{B.2}$$

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

$$W_{cT} \stackrel{\mathrm{d}}{=} \widehat{W}_T \,. \tag{B.3}$$

**Lemma B.1.** *Let $\widehat{W}_t$ be the solution of equation (B.1), and let $\widehat{\boldsymbol{u}}_i(t)$ be its eigenvectors. Then for any smooth test function $\theta$ 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*

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

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

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

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

for some small $c = c(n, \epsilon) > 0$, with $\boldsymbol{u}_i, \widehat{\boldsymbol{u}}_i(t), \boldsymbol{u}_i(t)$ being the eigenvectors of $W, \widehat{W}_t, W_t$, respectively. This concludes the proof of Theorem 2.8. Note that in equation (B.5), we used Lemma B.1 in the first step, equation (B.3) in the second step and equation (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 $\widetilde{W}_0 = \widetilde{W}$.

*Proof of Lemma B.1.* The proof of this lemma is very similar to the proof of [17, 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 [17, Appendix A], but this is automatic. Second, since the current overlap bound in equation (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 [17, Appendix A] in some detail.

As a consequence of the repulsion of the eigenvalues (level repulsion), as in [37, Lemma 5.2], to understand the overlap $\sqrt{N}\langle A^2 \rangle^{-1/2} \langle \boldsymbol{u}_i, A\boldsymbol{u}_i \rangle$, 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 $N^{-1}$: that is, the local eigenvalue spacing. In particular, to prove equation (B.4), it is enough to show that

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

for $t = N^{-1+\epsilon}$, $z = E + \mathrm{i}\eta$ for some $\zeta > 0, \omega > 0$ and all $\eta \geq N^{-1-\zeta}$; compare to [6, Section 4] and [10, Appendix A].

To prove this, we define

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

and then use Itô's formula:

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

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,

$$\kappa_t(\alpha_1, \ldots, \alpha_l) := \kappa(w_{\alpha_1}(t), \ldots, w_{\alpha_l}(t)) \tag{B.9}$$

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

By cumulant expansion, we get

$$\mathbf{E}\, \frac{\mathrm{d}R_t}{\mathrm{d}t} = \sum_{l=3}^{R} \sum_{\alpha_1, \ldots, \alpha_l} \kappa_t(\alpha_1, \ldots, \alpha_l)\, \mathbf{E}[\partial_{\alpha_1} \cdots \partial_{\alpha_l} R_t] + \Omega(R), \tag{B.10}$$

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 equation (B.10), we realise that $\partial_{ab}$-derivatives of $\langle \Im GA \rangle$ result in factors of the form $(GAG)_{ab}$, $(GAG)_{aa}$. For such factors, we use that

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

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

where we used that $\|\boldsymbol{u}_i\|_\infty \lesssim N^{-1/2+\xi}$, $|\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle| \le N^{-1/3+\xi}$, for any $\xi > 0$, uniformly in the spectrum by [29] and Theorem 2.6, respectively. We remark that in [17, Equation (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 $|\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle|$ for indices close to the edge (however, in [17, Equation (A.11)], we did not have $\langle A^2\rangle^{1/2}$). While our estimate on $(GAG)_{ab}$ is now weaker by a factor $N^{1/6}$, this is still sufficient to complete the Green function comparison argument.

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

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

and so, together with

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

by equation (B.10), we conclude equation (B.6). □

**Author contributions.** All the authors contributed equally.

**Conflict of Interest.** The authors have no conflicts of interest to declare.

**Ethical standards.** The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

**Funding statement.** L.E. acknowledges support by ERC Advanced Grant 'RMTBeyond' No. 101020331. D.S. acknowledges the support of Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zürich Foundation.

## References

[1] A. Aggarwal, P. Lopatto, and J. Marcinek, 'Eigenvector statistics of Lévy matrices', *Ann. Probab.* **49** (2021), 1778–1846, MR4260468.
[2] N. Anantharaman and E. Le Masson, 'Quantum ergodicity on large regular graphs', *Duke Math. J.* **164** (2015), 723–765, MR3322309.
[3] N. Anantharaman and M. Sabri, 'Quantum ergodicity on graphs: from spectral to spatial delocalization', *Ann. of Math. (2)* **189** (2019), 753–835, MR3961083.
[4] R. Bauerschmidt, J. Huang, and H.-T. Yau, 'Local Kesten-McKay law for random regular graphs', *Comm. Math. Phys.* **369** (2019), 523–636, MR3962004.
[5] R. Bauerschmidt, A. Knowles, and H.-T. Yau, 'Local semicircle law for random regular graphs', *Comm. Pure Appl. Math.* **70** (2017), 1898–1960, MR3688032.
[6] L. Benigni, 'Eigenvectors distribution and quantum unique ergodicity for deformed Wigner matrices', *Ann. Inst. Henri Poincaré Probab. Stat.* **56** (2020), 2822–2867, MR4164858.
[7] L. Benigni and P. Lopatto, 'Fluctuations in local quantum unique ergodicity for generalized Wigner matrices', *Comm. Math. Phys.* **391** (2022), 401–454, MR4397177.
[8] L. Benigni and P. Lopatto, 'Optimal delocalization for generalized Wigner matrices', *Adv. Math.* **396** (2022), Paper No. 108109, MR4370471.
[9] L. Benigni, 'Fermionic eigenvector moment flow', *Probab. Theory Related Fields* **179** (2021), 733–775, MR4242625.
[10] P. Bourgade and H.-T. Yau, 'The eigenvector moment flow and local quantum unique ergodicity', *Comm. Math. Phys.* **350** (2017), 231–278, MR3606475.
[11] P. Bourgade, J. Huang, and H.-T. Yau, 'Eigenvector statistics of sparse random matrices', *Electron. J. Probab.* **22** (2017), Paper No. 64, 38, MR3690289.

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

[12] P. 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** (2020), 1526–1596, MR4156609.
[13] C. Cacciapuoti, A. Maltsev, and B. Schlein, 'Bounds for the Stieltjes transform and the density of states of Wigner matrices', *Probab. Theory Related Fields* **163** (2015), 1–59, MR3405612.
[14] G. 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.
[15] G. Cipolloni, L. Erdős, and D. Schröder, 'Eigenstate thermalization hypothesis for Wigner matrices', *Comm. Math. Phys.* **388** (2021), 1005–1048, MR4334253.
[16] G. Cipolloni, L. Erdős, and D. Schröder, 'Functional central limit theorems for Wigner matrices', accepted for publication in *Ann. Appl. Probab* (2020), arXiv:2012.13218.
[17] G. Cipolloni, L. Erdős, and D. Schröder, 'Normal fluctuation in quantum ergodicity for Wigner matrices', *Ann. Probab.* **50** (2022), 984–1012, MR4413210.
[18] G. Cipolloni, L. Erdős, and D. Schröder, 'Optimal multi-resolvent local laws for Wigner matrices,' *Electron. J. Probab*. **27** (2022), MR4479913.
[19] G. Cipolloni, L. Erdős, and D. Schröder, 'Thermalisation for Wigner matrices', *J. Funct. Anal*. **282** (2022), Paper No. 109394, 37, MR4372147.
[20] Y. Colin de Verdière, 'Ergodicité et fonctions propres du laplacien', *Comm. Math. Phys.* **102** (1985), 497–502, MR818831.
[21] L. D'Alessio, Y. Kafri, A. Polkovnikov, and M. Rigol, 'From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics', *Advances in Physics* **65** (2016), 239–362 .
[22] J. Deutsch, 'Eigenstate thermalization hypothesis', *Rep. Prog. Phys*. **81** (2018), 082001, PMID29862983.
[23] B. Eckhardt, S. Fishman, J. Keating, O. Agam, J. Main, and K. Müller, 'Approach to ergodicity in quantum wave functions', *Physical Review E* **52** (1995), 5893–5903, PMID9964105.
[24] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, 'The local semicircle law for a general class of random matrices', *Electron. J. Probab*. **18**(59) (2013), 58, MR3068390.
[25] L. Erdős, T. Krüger, and D. Schröder, 'Cusp universality for random matrices I: local law and the complex Hermitian case', *Comm. Math. Phys.* **378** (2020), 1203–1278, MR4134946.
[26] L. Erdős, T. Krüger, and D. Schröder, 'Random matrices with slow correlation decay', *Forum Math. Sigma* **7** (2019), e8, 89, MR3941370.
[27] L. Erdős, B. Schlein, and H.-T. Yau, 'Local semicircle law and complete delocalization for Wigner random matrices', *Comm. Math. Phys.* **287** (2009), 641–655, MR2481753.
[28] L. Erdős, H.-T. Yau, and J. Yin, 'Bulk universality for generalized Wigner matrices', *Probab. Theory Related Fields* **154** (2012), 341–407, MR2981427.
[29] L. Erdős, H.-T. Yau, and J. Yin, 'Rigidity of eigenvalues of generalized Wigner matrices', *Adv. Math.* **229** (2012), 1435–1515, MR2871147.
[30] M. Feingold and A. Peres, 'Distribution of matrix elements of chaotic systems', *Physical Review A* **34** (1986), 591–595, PMID9897286.
[31] F. Gëttse, A. A. Naumov, and A. N. Tikhomirov, 'A local semicircle law under moment conditions: the Stieltjes transfer, rigidity and delocalization', *Teor. Veroyatn. Primen.* **62** (2017), 72–103, MR3633466.
[32] F. Götze, A. Naumov, and A. Tikhomirov, 'Local semicircle law under fourth moment condition', *J. Theoret. Probab.* **33** (2020), 1327–1362, MR4125959.
[33] Y. He and A. Knowles, 'Mesoscopic eigenvalue statistics of Wigner matrices', *Ann. Appl. Probab.* **27** (2017), 1510–1550, MR3678478.
[34] Y. He, A. Knowles, and R. Rosenthal, 'Isotropic self-consistent equations for mean-field random matrices', *Probab. Theory Related Fields* **171** (2018), 203–249, MR3800833.
[35] A. M. Khorunzhy, B. A. Khoruzhenko, and L. A. Pastur, 'Asymptotic properties of large random matrices with independent entries', *J. Math. Phys.* **37** (1996), 5033–5060, MR1411619.
[36] A. Knowles and J. Yin, 'Anisotropic local laws for random matrices', *Probab. Theory Related Fields* **169** (2017), 257–352, MR3704770.
[37] A. Knowles and J. Yin, 'Eigenvector distribution of Wigner matrices', *Probab. Theory Related Fields* **155** (2013), 543–582, MR3034787.
[38] A. Knowles and J. Yin, 'The isotropic semicircle law and deformation of Wigner matrices', *Comm. Pure Appl. Math.* **66** (2013), 1663–1750, MR3103909.
[39] J. O. Lee and K. Schnelli, 'Edge universality for deformed Wigner matrices', *Rev. Math. Phys*. **27** (2015), 1550018, 94, MR3405746.
[40] J. O. Lee and K. Schnelli, 'Local deformed semicircle law and complete delocalization for Wigner matrices with random potential', *J. Math. Phys*. **54** (2013), 103504, 62, MR3134604.
[41] W. Z. Luo and P. Sarnak, 'Quantum ergodicity of eigenfunctions on $\mathrm{PSL}_2(\mathbf{Z}) \setminus \mathbf{H}^2$', *Inst. Hautes Études Sci. Publ. Math*., 207–237 (1995), MR1361757.
[42] A. Lytova, 'On non-Gaussian limiting laws for certain statistics of Wigner matrices', *Zh. Mat. Fiz. Anal. Geom*. **9** (2013), 536–581, 611, 615, MR3155024.

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press

[43] J. Marcinek, 'High dimensional normality of noisy eigenvectors', Ph.D. thesis, Harvard University (ProQuest LLC, Ann Arbor, MI, 2020), 145, MR4272266.
[44] J. Marklof and Z. Rudnick, 'Quantum unique ergodicity for parabolic maps', *Geom. Funct. Anal.* **10** (2000), 1554–1578, MR1810753.
[45] S. O'Rourke, V. Vu, and K. Wang, 'Eigenvectors of random matrices: a survey', *J. Combin. Theory Ser. A* **144** (2016), 361–442, MR3534074.
[46] M. Rudelson and R. Vershynin, 'Sampling from large matrices: an approach through geometric functional analysis', *J. ACM* **54** (2007), Art. 21, 19, MR2351844.
[47] Z. Rudnick and P. Sarnak, 'The behaviour of eigenstates of arithmetic hyperbolic manifolds', *Comm. Math. Phys.* **161** (1994), 195–213, MR1266075.
[48] A. I. Snirelman, 'Ergodic properties of eigenfunctions', *Uspehi Mat. Nauk* **29** (1974), 181–182, MR0402834.
[49] K. Soundararajan, 'Quantum unique ergodicity for $\mathrm{SL}_2(\mathbb{Z}) \setminus \mathbb{H}$', *Ann. of Math. (2)* **172** (2010), 1529–1538, MR2680500.
[50] T. Tao and V. Vu, 'Random matrices: universality of local eigenvalue statistics', *Acta Math.* **206** (2011), 127–204, MR2784665.
[51] T. Tao and V. Vu, 'Random matrices: universality of local eigenvalue statistics up to the edge', *Comm. Math. Phys.* **298** (2010), 549–572, MR2669449.
[52] V. Vu and K. Wang, 'Random weighted projections, random quadratic forms and random eigenvectors', *Random Structures Algorithms* **47** (2015), 792–821, MR3418916.
[53] E. P. Wigner, 'Characteristic vectors of bordered matrices with infinite dimensions', *Ann. of Math. (2)* **62** (1955), 548–564, MR77805.
[54] S. Zelditch, 'Recent developments in mathematical quantum chaos', in *Current Developments in Mathematics, 2009*, 115–204 (Int. Press, Somerville, MA, 2010), MR2757360.
[55] S. Zelditch, 'Uniform distribution of eigenfunctions on compact hyperbolic surfaces', *Duke Math. J.* **55** (1987), 919–941, MR916129.

https://doi.org/10.1017/fms.2022.86 Published online by Cambridge University Press