The Annals of Probability
2022, Vol. 50, No. 3, 984–1012
https://doi.org/10.1214/21-AOP1552

NORMAL FLUCTUATION IN QUANTUM ERGODICITY FOR WIGNER MATRICES

BY GIORGIO CIPOLLONI[1,a], LÁSZLÓ ERDŐS[2,b] AND DOMINIK SCHRÖDER[3,c]

[1]*Department of Physics, Princeton University,* [a]*gc4233@princeton.edu*
[2]*IST Austria,* [b]*lerdos@ist.ac.at*
[3]*ETH Institute for Theoretical Studies, ETH Zurich,* [c]*dschoreder@ethz.ch*

We consider the quadratic form of a general high-rank deterministic matrix on the eigenvectors of an $N \times N$ Wigner matrix and prove that it has Gaussian fluctuation for each bulk eigenvector in the large N limit. The proof is a combination of the energy method for the Dyson Brownian motion inspired by Marcinek and Yau (2021) and our recent multiresolvent local laws (*Comm. Math. Phys.* **388** (2021) 1005–1048).

1. Introduction. Quantum unique ergodicity (QUE) in a disordered or chaotic quantum system asserts that the eigenvectors of the Hamilton operator tend to become uniformly distributed in the phase space; see [2, 16, 27, 29, 30, 33] for the seminal results and [15] for more recent references. We study a particularly strong form of this phenomenon for Wigner random matrices, the simplest prototype of a fully chaotic Hamiltonian. These are $N \times N$ random Hermitian matrices $W = W^*$ with centred, independent, identically distributed *(i.i.d.)* entries up to the symmetry constraint $w_{ab} = \overline{w_{ba}}$. Let $\{\boldsymbol{u}_i\}_{i=1}^N$ be an orthonormal eigenbasis of W corresponding to the eigenvalues $\boldsymbol{\lambda} = (\lambda_i)_{i=1}^N$ listed in increasing order. Recently, we showed [15] that for any deterministic matrix A with $\|A\| \le 1$, the eigenvector overlaps $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle$ converge to $\langle A\rangle := \frac{1}{N}\operatorname{Tr} A$, the normalized trace of A, in the large N limit. More generally, we proved that

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

holds with very high probability. We note that the bound (1) is optimal for high-rank deterministic matrices A and is coined as the *Eigenstate Thermalization Hypothesis* by Deutsch [18] and Srednicki [31]; see also [17], equation (20).

The main result of the current paper, Theorem 2.2, asserts that $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle$ has a Gaussian fluctuation on scale $N^{-1/2}$, more precisely,

$$\sqrt{N}\big[\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle - \langle A\rangle\big] \tag{2}$$

converges to a normal distribution for any Hermitian observables $A = A^*$ of high rank and for any eigenvectors $\boldsymbol{u}_i$ whose eigenvalue belongs to the bulk of the spectrum.

For Gaussian ensembles and A being a projection onto macroscopically many coordinates, (2) can be proven by using the special invariance property of the eigenvectors (see [28], Theorem 2.4). Our result concerns general Wigner matrices, and it has two main features: it concerns individual eigenvectors, *and* it is valid for general high rank observables. We now explain related previous results which all addressed only one of these features. First, Gaussianity of (2), after a small averaging in the index i, has recently been established in [13],

Received March 2021; revised September 2021.
MSC2020 subject classifications. Primary 60B20; secondary 15B52.
Key words and phrases. Local law, Dyson Brownian motion, stochastic eigenstate equation, eigenvector moment flow.

Theorem 2.3, using resolvent methods. Second, fluctuations involving individual eigenvectors in the bulk spectrum for general Wigner matrices can only be accessed by the Dyson Brownian motion approach which has only been developed for finite rank observables [3, 11, 26]. We now explain the background of these concepts.

1.1. *Dyson Brownian motion for eigenvectors.* For the simplest rank one case, $A = |\boldsymbol{q}\rangle\langle\boldsymbol{q}|$ with some deterministic unit vector $\boldsymbol{q}$, Bourgade and Yau [11] showed that the squared normalised overlaps $N|\langle \boldsymbol{u}_i, \boldsymbol{q}\rangle|^2$ converge in distribution to the square of a standard Gaussian variable as $N \to \infty$ (see also [22, 32] for the same result without DBM but under four moment matching condition in the bulk). Similar results have been obtained for deformed Wigner matrices [3], for sparse matrices [10], and for Lévy matrices [1]. Note that both the scaling and the limit distribution for the rank one case are different from (2). The basic intuition is that the coordinates of $\boldsymbol{u}_i$ are roughly independent; thus the sum in $\langle \boldsymbol{u}_i, \boldsymbol{q}\rangle = \sum_a \boldsymbol{u}_i(a)\boldsymbol{q}(a)$ obeys a central limit theorem (CLT) on scale $N^{-1/2}$. In fact, [11] also considers the joint distribution of finitely many eigenvectors tested against one fixed vector $\boldsymbol{q}$ and the joint distribution of a single eigenvector with finitely many test vectors $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_K$ for any fixed K, independent of N. Very recently, Marcinek and Yau [26] have established that the overlaps of finitely many eigenvectors *and* finitely many orthogonal test vectors are also asymptotically independent (squared) normal. Their method is very general and also applies to a large class of other random matrix ensembles, such as sparse or Lévy matrices.

The fundamental method behind all results involving individual eigenvectors for general Wigner matrices is the Dyson Brownian motion (DBM) for eigenvectors, also called the *stochastic eigenstate equation*, generated by a simple matrix Brownian motion for $W$and introduced by Bourgade and Yau in [11]. We briefly summarize the key steps in [11] in order to highlight the new ideas we needed to prove the Gaussianity of (2).

For each fixed n, the evolution of the joint nth order moments of the overlaps $N|\langle \boldsymbol{u}_i, \boldsymbol{q}\rangle|^2$ for different i's and fixed $\boldsymbol{q}$ is described by a system of parabolic evolution equations, called the *eigenvector moment flow*. Interpreting each such overlap as a particle sitting at location i in the discrete one dimensional index space $[N] = \{1, 2, \ldots, N\}$, the moment flow naturally corresponds to a Markovian jump process of n particles. It turns out that the rate of a jump from site i to site j is proportional with $N^{-1}(\lambda_i - \lambda_j)^{-2}$. Different $\boldsymbol{q}$'s can be incorporated by appropriately assigning *colours* to the particles. By the fast local equilibration property of the DBM, the moments of $N|\langle \boldsymbol{u}_i, \boldsymbol{q}\rangle|^2$ quickly become essentially independent of the index i, at least, for indices corresponding to nearby eigenvalues λ_i; hence, they can be

computed by locally averaging over i. For example, in the simplest $n = 1$ case, we have

$$f_i := \mathbf{E}\big[N\big|\langle \boldsymbol{u}_i, \boldsymbol{q}\rangle\big|^2|\boldsymbol{\lambda}\big] \approx f_{i'} = \mathbf{E}\big[N\big|\langle \boldsymbol{u}_{i'}, \boldsymbol{q}\rangle\big|^2|\boldsymbol{\lambda}\big], \quad |i - i'| \ll N, \tag{3}$$

already after a very short time $t \gg |i - i'|/N$. Here, we consider the conditional expectation of the eigenvectors, given that the eigenvalues are fixed. Since the global equilibrium of the DBM is the constant function $f_i = 1$, equilibration directly implies *smoothing* or *regularisation* in the dependence on the indices i.

On the other hand, by spectral theorem

$$\langle \boldsymbol{q}, \Im G(\lambda_i + \mathrm{i}\eta)\boldsymbol{q}\rangle = \frac{1}{N}\sum_{i'=1}^{N} \frac{\eta}{(\lambda_i - \lambda_{i'})^2 + \eta^2} N\big|\langle \boldsymbol{u}_{i'}, \boldsymbol{q}\rangle\big|^2, \tag{4}$$

where $G = G(z) = (W - z)^{-1}$ is the resolvent at a spectral parameter $z \in \mathbf{C} \setminus \mathbf{R}$. Using that the eigenvalues $\lambda_{i'}$ are *rigid*, that is, they are very close the corresponding quantiles $\gamma_{i'}$ of the

Wigner semicircle density (see (21) later), the i'-summation in (4) is a regularised averaging over indices $|i'-i| \lesssim N\eta$. Performing the i' summation in (4) by using (3), we obtain

$$\mathbf{E}\big[N|\langle \boldsymbol{u}_i, \boldsymbol{q}\rangle|^2|\boldsymbol{\lambda}\big] \approx \frac{1}{\Im m_{sc}(\gamma_i)}\mathbf{E}\big[\langle \boldsymbol{q}, \Im G(\gamma_i + \mathrm{i}\eta)\boldsymbol{q}\rangle|\boldsymbol{\lambda}\big]$$

for times $t \gg \eta$ where m_{sc} is the Stieltjes transform of the Wigner semicircle law. Choosing η slightly above the local eigenvalue spacing, $\eta = N^{-1+\epsilon}$ in the bulk of the spectrum, we have $\langle \boldsymbol{q}, \Im G(\gamma_i + \mathrm{i}\eta)\boldsymbol{q}\rangle \approx \Im m_{sc}(\gamma_i)$ not only in expectation but even in high probability by the *isotropic local law* for Wigner matrices [23]. Combining these inputs, we obtain $\mathbf{E}\, N|\langle \boldsymbol{u}_i, \boldsymbol{q}\rangle|^2 \approx 1$ along the DBM after a short time $t \gg N^{-1+\epsilon}$. A similar argument holds for higher moments. Finally, the small Gaussian component added by the DBM can be removed by standard perturbation methods, by the so called *Green function comparison* theorems.

1.2. *Dyson Brownian motion for general overlaps.* Given the method to handle $N|\langle \boldsymbol{u}_i, \boldsymbol{q}\rangle|^2$ described above, the Gaussianity of overlaps $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle$ with a general high rank matrix A can be approached in two natural ways. We now explain both of them to justify our choice. The first approach is to write $A = \sum_{k=1}^N a_k|\boldsymbol{q}_k\rangle\langle \boldsymbol{q}_k|$ in spectral decomposition with $|a_k| \lesssim 1$ and an orthonormal set $\{\boldsymbol{q}_k\}_{k=1}^N$ to have

$$\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle = \sum_{k=1}^N a_k\big|\langle \boldsymbol{u}_i, \boldsymbol{q}_k\rangle\big|^2. \tag{5}$$

If all overlaps $|\langle \boldsymbol{u}_i, \boldsymbol{q}_k\rangle|^2$, $k = 1, 2, \ldots, N$ were independent, then the central limit theorem applied to the summation in (5) would prove the normality of $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle$. This requires that the number of nonzero summands in (5), the rank of A, also goes to infinity as N increases. Hence, via the spectral decomposition of A, the Gaussianity of $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle$ appears rather an effect of the approximate independence of the overlaps $|\langle \boldsymbol{u}_i, \boldsymbol{q}_k\rangle|^2$ for different k's than their actual limit distribution. The analysis of the eigenvector moment flow [11, 26] yields this independence for finitely many k's, but it is not well suited for tracking overlaps $|\langle \boldsymbol{u}_i, \boldsymbol{q}_k\rangle|^2$ with a very large number of $\boldsymbol{q}_k$ vectors simultaneously. Hence, we discarded this approach.

The second natural approach is to generalise the eigenvector moment flow to moments of $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle$; this has been first achieved in [12]. Such flow naturally involves off-diagonal overlaps $\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle$ as well. Therefore, we need to describe conditional moments of the form $\mathbf{E}[\prod_{r=1}^n \langle \boldsymbol{u}_{i_r}, A\boldsymbol{u}_{j_r}\rangle|\boldsymbol{\lambda}]$ with different collections of *index pairs* $(i_1, j_1), (i_2, j_2), \ldots, (i_n, j_n)$ with the constraint that every index appears even number of times. Thus, the relevant moments can naturally be represented in an n-dimensional subset Λ^n of $[N]^{2n}$ (see Section 4.1). Moreover, in [12], equation (2.15), a certain symmetrised linear combination of n-order moments, the *perfect matching observable* was found that satisfies a closed equation along the Dyson Brownian motion; see (20) and (23). Moments of diagonal overlaps $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle$ can then be recovered from the perfect matching observable by setting all indices equal. Off-diagonal overlaps $\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle$, in general, cannot be recovered (except in the $n = 2$ case using an additional antisymmetric ("fermionic") version of the perfect matching observable [4]).

The main obstacle along this second approach is the lack of the analogue of (4) for general overlaps $\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle$. Consider the $n = 2$ case. A (regularized) local averaging in *one* index yields

$$\frac{1}{N}\sum_{i'=1}^N \frac{\eta}{(\lambda_i - \lambda_{i'})^2 + \eta^2} N\big|\langle \boldsymbol{u}_{i'}, A\boldsymbol{u}_j\rangle\big|^2 = \big\langle \boldsymbol{u}_j, A\Im G(\gamma_i + \mathrm{i}\eta)A\boldsymbol{u}_j\big\rangle, \tag{6}$$

which still involves an eigenvector $\boldsymbol{u}_j$, hence is not accessible solely by resolvent methods. Note that, for $A = |\boldsymbol{q}\rangle\langle \boldsymbol{q}|$, the overlap $\langle \boldsymbol{u}_{i'}, A\boldsymbol{u}_j\rangle$ factorizes, and the averaging in i' can be

done independently of j. For general A we can handle an averaging in *both* indices, that is, we will use that

$$\frac{1}{N^2}\sum_{i',j'=1}^{N}\frac{\eta}{(\lambda_i-\lambda_{i'})^2+\eta^2}\frac{\eta}{(\lambda_j-\lambda_{j'})^2+\eta^2}N\big|\langle \boldsymbol{u}_{i'}, A\boldsymbol{u}_{j'}\rangle\big|^2 = \big\langle A\Im G(\gamma_i+\mathrm{i}\eta)A\Im G(\gamma_j+\mathrm{i}\eta)\big\rangle. \tag{7}$$

The normalised trace in the right-hand side is accessible by resolvent methods using the recent multi-G local law proven in [15], Proposition 3.4. However, the generator of the eigenvector moment flow (24) involves the *sum* of averaging operators, as in (6), in all coordinate directions and not their *product*, as needed in (7). Higher moments ($n > 2$) require averaging in more than two indices simultaneously that is not apparently available in the generator. To remedy this situation, we now review how the equilibration (smoothing) property of the parabolic equation for the perfect matching observable can be manifested.

1.3. *Local smoothing of the eigenvector moment flow*: *An overview.* The technically simplest way to exploit the smoothing effect is via the maximum principle introduced in [11]. However, this requires that the generator is negative and itself has the necessary local averaging property to obtain a quantity computable by a local law; this is the case for eigenfunction overlaps, as in (4) but not for general overlaps $\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle$ in (6). We remark that the maximum principle was also used in [12] for more general overlaps, but only for getting an a priori bound and not for establishing their distribution. For this cruder purpose a rougher bound on (6) was sufficient that could be iteratively improved, but always by an N^{ϵ} factor off the optimal value.

A technically much more demanding way to exploit the equilibration of the eigenvector moment flow would be via homogenisation theory. In random matrix theory, homogenisation was originally introduced for the Dyson eigenvalue flow in [9, 24] by noticing that the generator is a discrete approximation of the one dimensional fractional Laplacian operator $|p| = \sqrt{-\Delta}$ with translation invariant kernel $(x-y)^{-2}$ whose heat kernel is explicitly known. Unfortunately, the eigenvector flow is more complicated, and a good approximation with a well-behaving continuous heat kernel is missing although homogenisation might also be accessible via a sequence of maximum principles as in [8].

Finally, the last and most flexible method for equilibration are the ultracontractivity estimates on the heat kernel that can be obtained by the standard Nash method from Poincaré or Sobolev inequalities for the Dirichlet form determined by the generator. In random matrix theory these ideas have been introduced in [20] for the eigenvalue gap statistics and have later been used as a priori bounds for the homogenisation theory. However, in the bulk regime they are barely not sufficiently strong to get the necessary precision for individual eigenvalues; they had to be complemented either by De Giorgi–Nash–Moser Hölder regularity estimates [20] or homogenisation [9, 24].

The recent work by Marcinek and Yau [26] remedies this shortcoming of the ultracontractivity bound by combining it with an energy method. The main motivation of [26] was to consider the joint distribution of the overlaps $|\langle \boldsymbol{u}_i, \boldsymbol{q}_k\rangle|^2$ for several eigenvectors and several test vectors simultaneously. The generator of the resulting *coloured eigenvector moment flow* lacks the positivity preserving property rendering the simple argument via maximum principle impossible. It turns out that this lack of positivity is due to a new *exchange term* in the generator that is present only because several $\boldsymbol{q}_k$'s (distinguished by colours) are considered simultaneously. However, the generator with the problematic exchange term is still positive in L^2-sense, and its Dirichlet form satisfies the usual Poincaré inequality from which ultracontractivity bounds can still be derived. The additional smallness now comes from an effective decay of the L^2-norm of the solution where local averaging like (4) can be exploited.

1.4. *Main ideas of the current paper.* The proof of Gaussianity of (2) consists of three steps:

Step 1. We use the energy method inspired by [26] together with the recent two-G local law from [15], Proposition 3.4, and more general multi-G local laws, proven in Section 5, to exploit an effective averaging mechanism to reduce the L^2-norm of the solution. In particular, to understand (7) we need a two-G local law instead of the single-G isotropic law used in (4).

Step 2. We use an $L^2 \to L^\infty$ ultracontractivity bound of the colourblind eigenvector moment flow from [26], Proposition 6.29.

Step 3. The first two steps prove the Gaussianity of the overlap (2) for Wigner matrices with a tiny Gaussian component. With a standard Green function comparison argument combined with the a priori bound (1) proven in [15] we remove this Gaussian component.

Step 2 and Step 3 are standard adaptations of existing previous results, so we focus only on explaining Step 1 We use the energy method in a very different way and for a very different purpose than [26] but for the same reason, its robustness. In the standard energy argument, if f_t satisfies the parabolic evolution equation $\partial_t f_t = \mathcal{L}_t f_t$ with a (time-dependent) generator $\mathcal{L}_t$, then

$$\frac{1}{2}\partial_t \| f_t \|_2^2 = \langle f_t, \mathcal{L}_t f_t \rangle =: -D_t(f_t) \le 0,$$

where D_t is the Dirichlet form (energy) associated to the generator $\mathcal{L}_t$. The goal is to give a good lower bound,

$$D_t(f) \ge c\| f \|_2^2 - \text{error}, \tag{8}$$

and use a Gronwall argument to conclude an effective L^2-decay along the dynamics. However, at this moment the Dirichlet form may first be replaced by a smaller one, $\widetilde{D}_t(f) \lesssim D_t(f)$, for which an effective lower bound (8) is easier to obtain. In our case, the gain comes from estimating the error term in (8) by exploiting the local averaging in all directions, as in (7), so that we could use the multi-G local law.

How to find $\widetilde{D}$? Very heuristically, the generator of the eigenvector moment flow is a discrete analogue of $|p_1| + |p_2| + \cdots + |p_n|$, that is, the *sum* of $|p|$-operators along all the n coordinate directions in the n-dimensional space Λ^n. However, the necessary averaging in (7) is rather the product of these one dimensional operators. Normally, sums of first order differential operators cannot be compared with their product since they scale differently with the length. But our operators have a short-range regularization on the scale η, that is, they rather correspond to $\eta^{-1}[1 - e^{-\eta|p|}]$ than just $|p|$ (see [25], Theorem 7.12). Therefore, we will prove the discrete analogue of the operator inequality

$$\frac{1}{\eta}\prod_{r=1}^{n}(1 - e^{-\eta|p_r|}) \le C(n)\sum_{r=1}^{n}\frac{1}{\eta}[1 - e^{-\eta|p_r|}] \tag{9}$$

on $\mathbf{R}^n$, and their quadratic forms will be the two Dirichlet forms $\widetilde{D}$ and D. Since the generator of $\widetilde{D}$ now averages in all directions, these averages yield traces of products $\Im GA\Im GA \ldots \Im GA$ for which we have a good local law; hence, the corresponding error in (8) is smaller than its naive a priori bound using only (1). This crucial gain provides the additional smallness to overcome the general fact that ultracontractivity bounds alone are barely not sufficient to gain sufficiently precise information on individual eigenvalues and eigenvectors in the bulk.

The actual proof requires several technical steps, such as: (i) localising the dynamics by considering a short-range approximation and treating the long-range part as a perturbation;

(ii) finite speed of propagation for the short range dynamics; (iii) cutoff the initial data in a smooth way so that cutoff and time evolution almost commute. Since these steps have appeared in the literature earlier, we will not reprove them here; we just refer to [26] where they have been adapted to the eigenvector moment flow. We will give full details only for Step 1.

Parallel with but independently of the current work, Benigni and Lopatto [6] have proved the CLT for $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle$ for the observable A, projecting onto a deterministic set of orthonormal vectors $A = \sum_{\alpha\in I} |q_\alpha\rangle\langle q_\alpha|$, with $N^\epsilon \le |I| \le N^{1-\epsilon}$, for some small fixed $\epsilon > 0$. Their low-rank assumption is complementary to our condition $\langle \mathring{A}^2\rangle \ge c$ for this class of projection operators; moreover, their result also covered the edge regime. The low-rank assumption allowed them to operate with the eigenvector moment flow from [4, 12]. However, their control can handle overlaps with at most $N^{1-\epsilon}$ vectors q_α simultaneously. It seems that this approach has a natural limitation preventing it from using it for high rank observables, for example, for $|I| \sim N$. In contrast, we consider overlaps $\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle$ directly without relying on the spectral decomposition of A.

Notation and conventions. We introduce some notations we use throughout the paper. For integers $k \in \mathbf{N}$, we use the notation $[k] := \{1, \dots, k\}$. For positive quantities f, g, we write $f \lesssim g$ and $f \sim g$ if $f \le Cg$ or $cg \le f \le Cg$, respectively, for some constants $c, C > 0$ which depend only on the constants appearing in (10). We denote vectors by bold-faced lower case 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 normalized trace of A. Moreover, for vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{C}^N$ we define

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

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

2. Main results. Let W be an $N \times N$ real symmetric or complex Hermitian Wigner matrix. We formulate the following assumptions on W.

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

$$\mathbf{E}|\chi_{\mathrm{d}}|^p + \mathbf{E}|\chi_{\mathrm{od}}|^p \le C_p. \tag{10}$$

Let $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_N$ be its eigenvalues in increasing order, and denote by $\boldsymbol{u}_1, \dots, \boldsymbol{u}_N$ the corresponding orthonormal eigenvectors. For any $N \times N$ matrix A, we denote by $\mathring{A} := A - \langle A\rangle$ the traceless part of A. We now state our main result.

THEOREM 2.2 (Central limit theorem in the QUE). *Let W be a real symmetric ($\beta = 1$) or complex Hermitian ($\beta = 2$) Wigner matrix satisfying Assumptions (2.1). Fix small $\delta, \delta' > 0$, and let $A = A^*$ be a deterministic $N \times N$ matrix with $\|A\| \lesssim 1$ and $\langle \mathring{A}^2\rangle \ge \delta'$. In the real*

symmetric case we also assume that $A \in \mathbf{R}^{N\times N}$ is real. Then, for any $i \in [\delta N, (1-\delta)N]$, it holds

$$\sqrt{\frac{\beta N}{2\langle \mathring{A}^2\rangle}}\big[\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle - \langle A\rangle\big] \Rightarrow \mathcal{N} \quad \text{as } N\to\infty \tag{11}$$

in the sense of moments, with $\mathcal{N}$ being a standard real Gaussian random variable. The speed of convergence is explicit; see (27).

3. Perfect matching observables. For definiteness, we present the proof for the real symmetric case; the analysis for the complex Hermitian case is completely analogous and so omitted. We only mention that the main difference between the two symmetry classes is that the perfect matching observables $f_{\boldsymbol{\lambda},t}$ in (20) are defined slightly differently (see [12], equation (A.3)), but the current proof can be easily adapted to this case.

Consider the matrix flow

$$\mathrm{d}W_t = \frac{\mathrm{d}\widetilde{B}_t}{\sqrt{N}}, \qquad W_0 = W, \tag{12}$$

with $\widetilde{B}_t$ being a standard real symmetric Brownian motion (see, e.g., [11], Definition 2.1). We denote the resolvent of W_t by $G = G_t(z) := (W_t - z)^{-1}$, for $z \in \mathbf{C}\setminus\mathbf{R}$. It is well known (see, e.g., [7, 21, 23]) that as $N\to\infty$ the resolvent $(W-z)^{-1}$ becomes approximately deterministic; its deterministic approximation is given by the unique solution of the scalar quadratic equation

$$-\frac{1}{m(z)} = z + m(z), \qquad \Im m(z)\Im z > 0. \tag{13}$$

In particular, $m(z) = m_{\mathrm{sc}}(z)$, $m_{\mathrm{sc}}(z)$ being the Stieltjes transform of the semicircular law $\rho_{\mathrm{sc}}(x) := (2\pi)^{-1}\sqrt{(4-x^2)_+}$. The deterministic approximation of $G_t(z)$ is given by $m_t(z)$, with m_t the solution of

$$\partial_t m_t(z) = -m_t \partial_z m_t(z), \qquad m_0 = m. \tag{14}$$

From now on, by $\rho_t = \rho_t(z)$ we denote $\rho_t(z) := \pi^{-1}\Im m_t(z)$, for any $t \geq 0$. In fact, starting from the standard semicircle $\rho_0 = \rho_{\mathrm{sc}}$, the density $\rho_t(x + \mathrm{i}0)$ is just a rescaling of ρ_0 by a factor $2+t$.

By [11], Definition 2.2, it follows that the eigenvectors $\boldsymbol{u}_1(t), \dots, \boldsymbol{u}_N(t)$ of W_t, corresponding to the eigenvalues $\lambda_1(t) \leq \lambda_2(t) \leq \cdots \leq \lambda_N(t)$, are a solution of the following system of SDE (dropping the time dependence),

$$\mathrm{d}\lambda_i = \frac{\mathrm{d}B_{ii}}{\sqrt{N}} + \frac{1}{N}\sum_{j\neq i}\frac{1}{\lambda_i - \lambda_j}\,\mathrm{d}t \tag{15}$$

$$\mathrm{d}\boldsymbol{u}_k = \frac{1}{\sqrt{N}}\sum_{j\neq i}\frac{\mathrm{d}B_{ij}}{\lambda_i - \lambda_j}\boldsymbol{u}_j - \frac{1}{2N}\sum_{j\neq i}\frac{\boldsymbol{u}_i}{(\lambda_i - \lambda_j)^2}\,\mathrm{d}t, \tag{16}$$

with $\{B_{ij}\}_{i,j\in[N]}$ being a standard real symmetric Brownian motions; see [11], Theorem 2.3, for the existence and uniqueness of the strong solution of (15)–(16).

By (16) it follows that the flow for the diagonal overlaps $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle$ naturally depends also on the off-diagonal overlap $\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle$; hence, our analysis will concern not only diagonal overlaps but also off-diagonal ones. Since $\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle - \langle A\rangle = \langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle$ and $\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle = \langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_j\rangle$ for $i \neq j$, without loss of generality we may assume for the rest of the paper, that A is traceless, $\langle A\rangle = 0$, that is, $A = \mathring{A}$. For traceless A we introduce the short-hand notation

$$p_{ij} = p_{ij}(t) := \big\langle \boldsymbol{u}_i(t), A\boldsymbol{u}_j(t)\big\rangle, \quad i,j \in [N]. \tag{17}$$

We are now ready to write the flow for monomials of p_{ii}, p_{ij} (see [12], Theorem 2.6, for the derivation of the flow). For any fixed n, we will only need to consider monomials of the form $\prod_{k=1}^{n} p_{i_k j_k}$ where each index appears even number of times; it turns out that the linear combinations of such monomials with a fixed degree n are invariant under the flow.

To encode general monomials, we use a particle picture (introduced in [11] and developed in [12, 26]) where each particle on the set of integers $[N]$ corresponds to two occurrences of an index i in the monomial product. We use the same notation as in [12], and we define $\boldsymbol{\eta} : [N] \to \mathbf{N}$, where $\eta_j := \boldsymbol{\eta}(j)$ is interpreted as the number of particles at the site j, and $n(\boldsymbol{\eta}) := \sum_j \eta_j = n$ denotes the total number of particles that are conserved under the flow. The space of n-particle configurations is denoted by Ω^n. Moreover, for any index pair $i \neq j \in [N]$, we define $\boldsymbol{\eta}^{ij}$ to be the configuration obtained moving a particle from the site i to the site j; if there is no particle in i, then we define $\boldsymbol{\eta}^{ij} = \boldsymbol{\eta}$. For any configuration $\boldsymbol{\eta}$, consider the set of vertices

$$\mathcal{V}_{\boldsymbol{\eta}} := \big\{(i,a) : 1 \le i \le n, 1 \le a \le 2\eta_i\big\}, \tag{18}$$

and let $\mathcal{G}_{\boldsymbol{\eta}}$ be the set of perfect matchings on $\mathcal{V}_{\boldsymbol{\eta}}$. Note that every particle configuration η gives rise to two vertices in $\mathcal{V}_{\boldsymbol{\eta}}$; thus, the elements of $\mathcal{V}_{\boldsymbol{\eta}}$ represent the indices in the product $\prod_{k=1}^{n} p_{i_k j_k}$.

There is no closed equation for individual products $\prod_{k=1}^{n} p_{i_k j_k}$, but there is one for a certain symmetrized linear combination; see [12], equation (2.15). Therefore, for any perfect matching $G \in \mathcal{G}_{\boldsymbol{\eta}}$, we define

$$P(G) := \prod_{e \in \mathcal{E}(G)} p(e), \quad p(e) := p_{i_1 i_2}, \tag{19}$$

where $e = \{(i_1, a_1), (i_2, a_2)\} \in \mathcal{V}_{\boldsymbol{\eta}}$, and $\mathcal{E}(G)$ denotes the edges of G. For example, for $n = 2$ and for the configuration $\boldsymbol{\eta}$ defined by $\boldsymbol{\eta}(i) = \boldsymbol{\eta}(j) = 1$ with some $i \neq j$ and zero otherwise, we have three perfect matchings corresponding to $p_{ii}p_{jj}$ and twice p_{ij}^2. For $n = 3$ and $\boldsymbol{\eta}$ defined by $\boldsymbol{\eta}(i) = \boldsymbol{\eta}(j) = \boldsymbol{\eta}(k) = 1$, we have 15 perfect matchings; $p_{ii}p_{jj}p_{kk}$, two copies $p_{ij}^2 p_{kk}$, $p_{ik}^2 p_{jj}$, $p_{jk}^2 p_{ii}$ each, and eight copies of $p_{ij}p_{jk}p_{ki}$.

We are now ready to define the *perfect matching observable* for any given configuration $\boldsymbol{\eta}$,

$$f_{\boldsymbol{\lambda},t}(\boldsymbol{\eta}) := \frac{N^{n/2}}{[2\langle A^2\rangle]^{n/2}} \frac{1}{(n-1)!!} \frac{1}{\mathcal{M}(\boldsymbol{\eta})} \mathbf{E}\Big[\sum_{G \in \mathcal{G}_{\boldsymbol{\eta}}} P(G) \Big| \boldsymbol{\lambda}\Big], \quad \mathcal{M}(\boldsymbol{\eta}) := \prod_{i=1}^{N} (2\eta_i - 1)!!, \tag{20}$$

with n being the number of particles in the configuration $\boldsymbol{\eta}$. Here, we took the conditioning on the entire flow of eigenvalues, $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}(t)\}_{t \in [0,T]}$ for some fixed $T > 0$. The observable $f_{\boldsymbol{\lambda},t}$ satisfies a parabolic partial differential equation; see (23) below.

REMARK 3.1. For any $k \in \mathbf{N}$, the double factorial $k!!$ is defined by $k!! = k(k-2)!!$, $1!! = 0!! = (-1)!! = 1$. We remark that, in [12, 26], the authors use a different convention for the double factorial, that is, in these papers $k!! = (k-1)(k-2)!!$.

Note that f in (20) is defined slightly differently compared to the definition in [12], equation (2.15), where the authors do not have the additional $(N/(2\langle A^2\rangle)^{n/2}[(n-1)!!]^{-1}$ factor. Our normalisation factor is dictated by the principle that, for traceless A, we expect $\sqrt{N}p_{ii} = \sqrt{N}[\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle]$ to be approximately a centred normal random variable with variance $2\langle A^2\rangle$. In particular, the nth moment of $(N/2\langle A^2\rangle)^{1/2}p_{ii}$ for even n is close to $(n-1)!!$. Therefore, if $\boldsymbol{\eta}$ is a configuration with n particles all sitting at the same site i, that is, $\boldsymbol{\eta}(i) = n$ and zero otherwise, then $\mathcal{M}(\boldsymbol{\eta}) = (2n-1)!!$ is the number of perfect matchings, and, therefore, we expect $f_{\boldsymbol{\lambda},t}(\boldsymbol{\eta}) \approx 1$. Note that using the a priori bound $|p_{ij}| \le N^{-1/2+\xi}$, for any

$\xi > 0$, proven in [15], Theorem 2.2, we have $|f_{\boldsymbol{\lambda},t}| \lesssim N^{\xi}$ with very high probability, while the analogous quantity $f_{\boldsymbol{\lambda},t}$, defined in [12], equation (2.15), has an a priori bound of order $N^{-n/2+\xi}$.

We always assume that the entire eigenvalue trajectory $\{\boldsymbol{\lambda}(t)\}_{t\in[0,T]}$ satisfies the usual rigidity estimate (see, e.g., [19] or [21], Theorem 7.6). More precisely, for any fixed $\xi > 0$, we define

$$\Omega = \Omega_{\xi} := \Big\{ \sup_{0\le t\le T} \max_{i\in[N]} N^{2/3}\widehat{i}^{1/3} |\lambda_i(t) - \gamma_i(t)| \le N^{\xi} \Big\}, \tag{21}$$

where $\widehat{i} := i \wedge (N+1-i)$, then we have

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

for any (small) $\xi > 0$ and (large) $D > 0$. Here, $\gamma_i(t)$ are the classical eigenvalue locations (*quantiles*) defined by

$$\int_{-\infty}^{\gamma_i(t)} \rho_t(x)\,\mathrm{d}x = \frac{i}{N}, \quad i \in [N], \tag{22}$$

where $\rho_t(x) = \frac{2}{(2+t)^2\pi}\sqrt{((2+t)^2 - x^2)_+}$ is the semicircle law corresponding to W_t. Note that $|\gamma_i(t) - \gamma_i(s)| \lesssim |t-s|$ in the bulk, for any $t, s \ge 0$, as a consequence of the smoothness of $t \to \rho_t$ in the bulk.

By [12], Theorem 2.6, we have that

$$\partial_t f_{\boldsymbol{\lambda},t} = \mathcal{B}(t) f_{\boldsymbol{\lambda},t}, \tag{23}$$

$$\mathcal{B}(t) f_{\boldsymbol{\lambda},t} = \sum_{i\ne j} c_{ij}(t) 2\eta_i(1+2\eta_j)\big(f_{\boldsymbol{\lambda},t}(\boldsymbol{\eta}^{kl}) - f_{\boldsymbol{\lambda},t}(\boldsymbol{\eta})\big), \tag{24}$$

where

$$c_{ij}(t) := \frac{1}{N(\lambda_i(t) - \lambda_j(t))^2}. \tag{25}$$

Note that c_{ij} depends on $\{\boldsymbol{\lambda}(t)\}_{t\in[0,T]}$, for some $T > 0$, but we omit this fact from the notation. We note that this flow was originally derived for special observables given in [12], equation (2.6), but the same derivation immediately holds for arbitrary A (see [12], Remark 2.8).

The main technical ingredient that will be used in the proof of Theorem 2.2 is the following proposition, whose proof is postponed to Section 4.

PROPOSITION 3.2. *For any $n \in \mathbf{N}$, there exists $c(n) > 0$ such that, for any $\epsilon > 0$ and for any $T \ge N^{-1+\epsilon}$, it holds*

$$\sup_{\boldsymbol{\eta}} \big| f_T(\boldsymbol{\eta}) - \mathbf{1}(n \text{ even}) \big| \lesssim N^{-c(n)} \tag{26}$$

with very high probability, where the supremum is taken over configurations $\boldsymbol{\eta}$ such that $\sum_i \eta_i = n$ and $\eta_i = 0$ for $i \notin [\delta N, (1-\delta)N]$, with $\delta > 0$ from Theorem 2.2. The implicit constant in (26) depends on n, ϵ, δ.

PROOF OF THEOREM 2.2. We fix $i \in [\delta N, (1-\delta)N]$, and we choose $\boldsymbol{\eta}$ to be the configuration $\boldsymbol{\eta}$ with $\eta_i = n$ and all other $\eta_j = 0$. Then, all the terms $P(G)$ are equal to p_{ii}^n in the definition of f; see (20). Then, using (26) for this particular $\boldsymbol{\eta}$, we conclude that

$$\mathbf{E}\bigg[\sqrt{\frac{N}{2\langle A^2\rangle}}\langle \boldsymbol{u}_i(T), A\boldsymbol{u}_i(T)\rangle\bigg]^n = \mathbf{1}(n \text{ even})(n-1)!! + \mathcal{O}\big(N^{-c(n)}\big) \tag{27}$$

for any $i \in [\delta N, (1-\delta)N]$ and $T \gg N^{-1}$, where we used that $\|f_T\|_\infty \le N^{n/2}$ deterministically on the complement of the high probability set on which (26) holds. With (27) we have proved that Theorem 2.2 holds for Wigner matrices with a small Gaussian component. For the general case, Theorem 2.2 follows from (27) and a standard application of the Green function comparison theorem (GFT), relating the eigenvectors/eigenvalues of W_T to those of W; see the Appendix where we recall the argument for completeness. □

4. DBM analysis. In this section we focus on the analysis of the eigenvector moment flow (23)–(24). Since in our proof we use some results proven in [26], we start giving an equivalent representation of (20) which is the same used in [26] without distinguishing the several colours.

4.1. *Equivalent representation of the flow.* Fix $n \in \mathbf{N}$; then, in the remainder of this section we will consider configurations $\boldsymbol{\eta} \in \Omega^n$, that is, such that $\sum_j \eta_j = n$. Following [26] (but without the extra complication involving colours), we now give an equivalent representation of the flow (23)–(24) which will be defined on the $2n$-dimensional lattice $[N]^{2n}$ instead of configurations of n particles. Let $\boldsymbol{x} \in [N]^{2n}$, and define

$$n_i(\boldsymbol{x}) := \big|\{a \in [2n] : x_a = i\}\big| \tag{28}$$

for all $i \in \mathbf{N}$. We define the configuration space

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

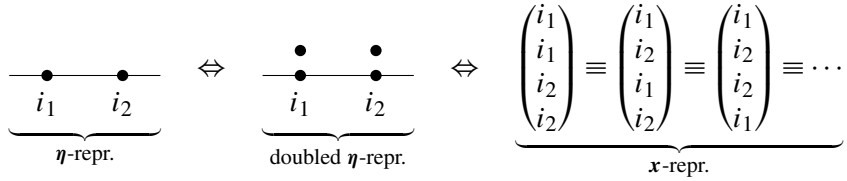
Note that Λ^n is an n-dimensional subset of the $2n$ dimensional lattice $[N]^{2n}$ in the sense that Λ^n is a finite union of n-dimensional sublattices of $[N]^{2n}$. From now on we will only consider configurations $\boldsymbol{x} \in \Lambda^n$. In particular, in this representation to each particle is associated a label $a \in [2n]$, that is, there is a particle at a site $i \in [N]$ iff there exists $a \in [2n]$ such that $x_a = i$. Additionally, by the definition of Λ^n it follows that the number of particles at a site $i \in [N]$ is always even.

REMARK 4.1. Note that in [26] the authors consider $\boldsymbol{x}$ to be an n-dimensional vector that lives in the $n/2$-dimensional subset Λ^n. For notational simplicity, in the current paper we assume that $\boldsymbol{x}$ is a $2n$-dimensional vector and that Λ^n is n-dimensional.

The natural correspondence between the two representations is given by

$$\boldsymbol{\eta} \leftrightarrow \boldsymbol{x}, \qquad \eta_i = \frac{n_i(\boldsymbol{x})}{2}. \tag{29}$$

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. As an example, the configuration with single (or doubled) particles in $i_1 \neq i_2$ corresponds to six $\boldsymbol{x} \in \Lambda^2$, as in



Let $\phi\colon \Lambda^n \to \Omega^n$, $\phi(\boldsymbol{x}) = \boldsymbol{\eta}$ denote the map that projects the $\boldsymbol{x}$-configuration space to the $\boldsymbol{\eta}$-configuration space using (29). This map naturally pulls back functions f of $\boldsymbol{\eta}$ to functions of $\boldsymbol{x}$,

$$\big(\phi^* f\big)(\boldsymbol{x}) = f\big(\phi(\boldsymbol{x})\big).$$

We will always consider functions g on $[N]^{2n}$ that are push-forward of some function f on Ω^n, $g = f \circ \phi$; that is, they correspond to functions on the configurations

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

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

$$g_{\boldsymbol{\lambda},t}(\boldsymbol{x}) := f_{\boldsymbol{\lambda},t}\big(\phi(\boldsymbol{x})\big), \tag{30}$$

where $f_{\boldsymbol{\lambda},t}$ was defined in (20). In the following we will often use the notation $g_t(\boldsymbol{x}) = g_{\boldsymbol{\lambda},t}(\boldsymbol{x})$, dropping the dependence of $g_t(\boldsymbol{x})$ on the eigenvalues.

The flow (23)–(24) can be written in the $\boldsymbol{x}$-representation as follows:

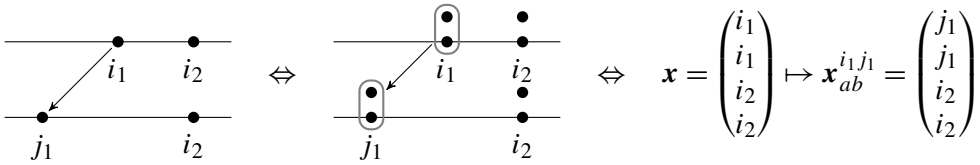
$$\partial_t g_t(\boldsymbol{x}) = \mathcal{L}(t) g_t(\boldsymbol{x}) \tag{31}$$

$$\mathcal{L}(t) := \sum_{j \neq i} \mathcal{L}_{ij}(t), \qquad \mathcal{L}_{ij}(t) g(\boldsymbol{x}) := c_{ij}(t) \frac{n_j(\boldsymbol{x}) + 1}{n_i(\boldsymbol{x}) - 1} \sum_{a \neq b \in [2n]} \big(g(\boldsymbol{x}^{ij}_{ab}) - g(\boldsymbol{x})\big), \tag{32}$$

where

$$\boldsymbol{x}^{ij}_{ab} := \boldsymbol{x} + \delta_{x_a i} \delta_{x_b i} (j - i)(\boldsymbol{e}_a + \boldsymbol{e}_b) \tag{33}$$

with $\boldsymbol{e}_a(c) = \delta_{ac}$, $a, c \in [2n]$. Clearly, this flow preserves the equivariance of g; that is, it is a map on functions defined on Λ^n. The jump operator $\boldsymbol{x}^{ij}_{ab}$, defined in (33), changes x_a, x_b from i to j if $x_a = x_b = i$ and otherwise leaves $\boldsymbol{x}$ unchanged. In the particle picture $\boldsymbol{\eta}$ this corresponds in moving one particle from the site i (if there is any) to the site j; see the following example for $n = 2$ (with $i = i_1$, $j = j_1$ and $a = 1, b = 2$):



Define the measure

$$\pi(\boldsymbol{x}) := \prod_{i=1}^{N} \big((n_i(\boldsymbol{x}) - 1)!!\big)^2 \tag{34}$$

on Λ^n and the corresponding $L^2(\Lambda^n) = L^2(\Lambda^n, \pi)$ space equipped with the scalar product

$$\langle f, g \rangle_{\Lambda^n} = \langle f, g \rangle_{\Lambda^n, \pi} := \sum_{\boldsymbol{x} \in \Lambda^n} \pi(\boldsymbol{x}) \bar{f}(\boldsymbol{x}) g(\boldsymbol{x}). \tag{35}$$

We will often drop the dependence on the measure π in the scalar product. We also define the following norm on $L^p(\Lambda^n)$:

$$\|f\|_p := \Big(\sum_{\boldsymbol{x} \in \Lambda^n} \pi(\boldsymbol{x}) |f(\boldsymbol{x})|^p \Big)^{1/p}. \tag{36}$$

The measure $\pi(\boldsymbol{x})$ clearly satisfies

$$1 \leq \pi(\boldsymbol{x}) \leq (2n - 1)!!, \tag{37}$$

uniformly in $\boldsymbol{x} \in \Lambda^n$. A direct calculation in [26], Appendix A.2, shows that the operator $\mathcal{L} = \mathcal{L}(t)$ is symmetric with respect to the measure π, and it is a negative operator on the space $L^2(\Lambda^n)$ with Dirichlet form

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

We will often omit the time dependence of the generator $\mathcal{L}(t)$. We denote by $\mathcal{U}(s,t)$ the semigroup associated to $\mathcal{L}$ from (32); that is, for any $0 \le s \le t$, it holds

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

4.2. *Short-range approximation*. Before proceeding we introduce a localised version of (31)–(32). Choose an (N-dependent) parameter $1 \ll K \le \sqrt{N}$, and define the *averaging operator* as a simple multiplication operator by a "smooth" cut-off function,

$$\text{(38)} \quad \mathrm{Av}(K, \boldsymbol{y})h(\boldsymbol{x}) := \mathrm{Av}(\boldsymbol{x}; K, \boldsymbol{y})h(\boldsymbol{x}), \qquad \mathrm{Av}(\boldsymbol{x}; K, \boldsymbol{y}) := \frac{1}{K}\sum_{j=K}^{2K-1} \mathbf{1}\big(\|\boldsymbol{x} - \boldsymbol{y}\|_1 < j\big),$$

with $\|\boldsymbol{x} - \boldsymbol{y}\|_1 := \sum_{a=1}^{2n} |x_a - y_a|$. While it was denoted and called averaging operator in [12, 26], it is rather a *localization*, that is, a multiplication by a "smooth" cutoff function $\boldsymbol{x} \to \mathrm{Av}(\boldsymbol{x}; K, \boldsymbol{y})$ which is centered at $\boldsymbol{y}$ and has a soft range of size K. The parameters $K, \boldsymbol{y}$ are considered fixed and often omitted from the notation.

Now, we define a short-range version of the dynamics (31). Fix an integer ℓ with $1 \ll \ell \ll K$, and define the short-range coefficients

$$\text{(39)} \qquad c_{ij}^{\mathcal{S}}(t) := \begin{cases} c_{ij}(t) & \text{if } i, j \in \mathcal{J} \text{ and } |i-j| \le \ell, \\ 0 & \text{otherwise,} \end{cases}$$

where $c_{ij}(t)$ is defined in (25). Here,

$$\text{(40)} \qquad \mathcal{J} = \mathcal{J}_\delta := \big\{i \in [N] : \gamma_i(0) \in \mathcal{I}_\delta\big\}, \qquad \mathcal{I}_\delta := (-2+\delta, 2-\delta)$$

with $\delta > 0$ from Theorem 2.2 so that $\mathcal{I}_\delta$ lies entirely in the bulk spectrum.

We define $h_t(\boldsymbol{x})$ as the time evolution of a localized initial data g_0 by the short-range dynamics,

$$\text{(41)} \qquad \begin{aligned} h_0(\boldsymbol{x}; \ell, K, \boldsymbol{y}) &= h_0(\boldsymbol{x}; K, \boldsymbol{y}) := \mathrm{Av}(\boldsymbol{x}; K, \boldsymbol{y})\big(g_0(\boldsymbol{x}) - \mathbf{1}(n \text{ even})\big), \\ \partial_t h_t(\boldsymbol{x}; \ell, K, \boldsymbol{y}) &= \mathcal{S}(t) h_t(\boldsymbol{x}; \ell, K, \boldsymbol{y}), \end{aligned}$$

where

$$\text{(42)} \qquad \mathcal{S}(t) := \sum_{j \ne i} \mathcal{S}_{ij}(t), \qquad \mathcal{S}_{ij}(t)h(\boldsymbol{x}) := c_{ij}^{\mathcal{S}}(t)\frac{n_j(\boldsymbol{x})+1}{n_i(\boldsymbol{x})-1} \sum_{a \ne b \in [2n]} \big(h(\boldsymbol{x}_{ab}^{ij}) - h(\boldsymbol{x})\big).$$

Here, we used the notation $h(\boldsymbol{x}) = h(\boldsymbol{x}; \ell, K, \boldsymbol{y})$ to indicate all relevant parameters: ℓ indicates the short range of the dynamics, $\boldsymbol{y}$ is the centre, and K is the range of the cut-off in the initial condition, and we always choose $\ell \ll K$. In (41) we already subtracted $\mathbf{1}(n \text{ even})$ since in our application the initial condition $g_0(\boldsymbol{x})$ after some local averaging will be close to $\mathbf{1}(n \text{ even})$; hence, after longer time we expect that h_t tends to zero since the dynamics has a smoothing effect, and it is an L^1 contraction.

4.3. *L^2-bound*. Define the distance on Λ^n as

$$\text{(43)} \qquad d(\boldsymbol{x}, \boldsymbol{y}) := \sup_{a \in [2n]} \big|\mathcal{J} \cap \big[\min(x_a, y_a), \max(x_a, y_a)\big)\big|$$

with $\mathcal{J}$ defined in (40). Note that d is not a metric since it is degenerate, but it is still symmetric and satisfies the triangle inequality [26], equation (5.6). The key ingredient to prove the L^2-bound in (49) below is to show that the short-range dynamics (41)–(42) is close to the original dynamics (31)–(32). This will be achieved using the following finite speed of propagation estimate, proven in [10], Theorem 2.1, Lemma 2.4, [26], Proposition 5.2, (see also

[12], equation (3.15)), for $\mathcal{U}_{\mathcal{S}}(s,t) = \mathcal{U}_{\mathcal{S}}(s,t;\ell)$ which is the transition semigroup associated to the short-range generator $\mathcal{S}(t)$. For any $\boldsymbol{x} \in \Lambda^n$, define the "delta-function" on Λ^n as

$$\delta_{\boldsymbol{x}}(\boldsymbol{u}) := \begin{cases} \pi(\boldsymbol{x})^{-1} & \text{if } \boldsymbol{u} = \boldsymbol{x}, \\ 0 & \text{otherwise,} \end{cases}$$

and denote the matrix entries of $\mathcal{U}_{\mathcal{S}}(s,t)$ by $\mathcal{U}_{\mathcal{S}}(s,t)_{\boldsymbol{x}\boldsymbol{y}} := \langle \delta_{\boldsymbol{x}}, \mathcal{U}_{\mathcal{S}}(s,t)\delta_{\boldsymbol{y}} \rangle$.

PROPOSITION 4.2. *Fix any small $\epsilon > 0$ and $\ell \geq N^{\epsilon}$. Then, for any $\boldsymbol{x}, \boldsymbol{y} \in \Lambda^n$ with $d(\boldsymbol{x}, \boldsymbol{y}) > N^{\epsilon}\ell$, it holds*

$$\sup_{0 \leq s_1 \leq s_2 \leq s_1 + \ell N^{-1}} \left| \mathcal{U}_{\mathcal{S}}(s_1, s_2; \ell)_{\boldsymbol{x}\boldsymbol{y}} \right| \leq e^{-N^{\epsilon}/2} \tag{44}$$

on the very high probability event Ω.

This finite speed of propagation, together with the fact that the initial condition h_0 is localized in a K-neighbourhood of a fixed center $\boldsymbol{y}$, implies that h_t is supported in a $K + N^{\epsilon}\ell \leq 2K$ neighbourhood of $\boldsymbol{y}$ up an exponentially small tail part.

Using Proposition 4.2, by [26], Corollary 5.3, we immediately conclude the following lemma.

LEMMA 4.3. *For any times s_1, s_2 such that $0 \leq s_1 \leq s_2 \leq s_1 + \ell N^{-1}$ and for any $\boldsymbol{y} \in \Lambda^n$ supported on $\mathcal{J}$ (i.e., $\boldsymbol{y}_a \in \mathcal{J}$ for any $a \in [2n]$) for the commutator of the evolution $\mathcal{U}_{\mathcal{S}}$ and the averaging operator, we have*

$$\left\| \left[\mathcal{U}_{\mathcal{S}}(s_1, s_2; \ell), \mathrm{Av}(\boldsymbol{y}, K) \right] \right\|_{\infty,\infty} \leq C(n) \frac{N^{\epsilon}\ell}{K} \tag{45}$$

for some constant $C(n) > 0$ and for any small $\epsilon > 0$, on the very high probability event Ω.

Another straightforward application of the finite speed of propagation estimate in Proposition 4.2 is the following bound $\mathcal{U}(s_1, s_2) - \mathcal{U}_{\mathcal{S}}(s_1, s_2; \ell)$. This result was proven in [26], Proposition 5.7, for a specific f, but the same proof applies for a general function f.

LEMMA 4.4. *Let $0 \leq s_1 \leq s_2 \leq s_1 + \ell N^{-1}$, and f is a function on Λ^n; then, for any $\boldsymbol{x} \in \Lambda^n$ supported on $\mathcal{J}$, it holds*

$$\left| \left(\mathcal{U}(s_1, s_2) - \mathcal{U}_{\mathcal{S}}(s_1, s_2; \ell) \right) f(\boldsymbol{x}) \right| \lesssim N^{1+n\xi} \frac{s_2 - s_1}{\ell} \|f\|_{\infty} \tag{46}$$

for any small $\xi > 0$.

PROOF. Using Proposition 4.2, the proof of (46) is completely analogous to the proof of [26], Proposition 5.7, since the only input used in [26], Proposition 5.7, is that

$$\sum_{j : |j-i| > \ell} \frac{1}{N(\lambda_i - \lambda_j)^2} \leq \frac{N^{1+\xi}}{\ell}$$

on Ω which follows by rigidity. □

Before stating the main result of this section, we define the set $\widehat{\Omega}$ 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

$$\widehat{\Omega} = \widehat{\Omega}_{\omega,\xi} := \bigcap_{\substack{z_i:\Re z_i \in \mathcal{I}_\delta \\ |\Im z_i| \in [N^{-1+\omega},10]}} \Bigg[\bigcap_{k=3}^{n} \Big\{ \sup_{0\le t\le T} |\langle G_t(z_1)A \dots G_t(z_k)A\rangle| \le \frac{N^{\xi+(k-3)/2}}{\sqrt{\eta_*}} \Big\}$$

$$\cap \Big\{ \sup_{0\le t\le T} |\langle G_t(z_1)AG_t(z_2)A\rangle - m_t(z_1)m_t(z_2)\langle A^2\rangle| \le \frac{N^\xi}{\sqrt{N\eta^*}} \Big\} \tag{47}$$

$$\cap \Big\{ \sup_{0\le t\le T} |\langle G_t(z_1)A\rangle| \le \frac{N^\xi}{N\sqrt{|\Im z_1|}} \Big\} \Bigg],$$

where $\eta_* := \min\{|\Im z_i| \mid i \in [k]\}$. The fact that $\widehat{\Omega}$ is a very high probability set follows by [15], Theorem 2.6, for $k = 1$, by [15], equation (3.9), for $k = 2$, and by Proposition 5.1 for $k \ge 3$. In particular, since $m_t(z_1)m_t(z_2)\langle A^2\rangle$ is bounded for $k = 2$, we have

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

on the very high probability event $\widehat{\Omega}_{\omega,\xi}$ which, by spectral theorem, implies

$$\sup_{0\le t\le T} \max_{i,j\in\mathcal{I}} |\langle \boldsymbol{u}_i(t), A\boldsymbol{u}_j(t)\rangle| \le N^{-1/2+\omega} \quad \text{on } \widehat{\Omega}_{\omega,\xi}. \tag{48}$$

PROPOSITION 4.5. *For any scale satisfying $N^{-1} \ll \eta \ll T_1 \ll \ell N^{-1} \ll KN^{-1}$ and any small $\epsilon, \xi > 0$, it holds*

$$\|h_{T_1}(\cdot;\ell,K,\boldsymbol{y})\|_2 \lesssim K^{n/2}\mathcal{E}, \tag{49}$$

with

$$\mathcal{E} := N^{n\xi}\Big(\frac{N^\epsilon \ell}{K} + \frac{NT_1}{\ell} + \frac{N\eta}{\ell} + \frac{N^\epsilon}{\sqrt{N\eta}} + \frac{1}{\sqrt{K}}\Big), \tag{50}$$

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

PROOF. Before presenting the formal proof, we explain the main idea. In the sense of Dirichlet forms, we will replace the generator $\mathcal{S}(t)$ (41)–(42), which is the *sum* of one-dimensional generators, with the generator $\mathcal{A}(t)$ that corresponds to the *product* of such operators (see (52) below for its definition). Considering that c_{ij} decays proportionally with $|i-j|^{-2}$ (using rigidity in (25)), it is the kernel of the discrete approximation of the one dimensional operator $|p| = \sqrt{-\Delta}$ on $\mathbf{R}$ but lifted to the n-dimensional space Λ^n. Therefore, one may think of $\mathcal{L}(t)$, and its short-range approximation $\mathcal{S}(t)$, as a discrete analogue of $|p_1| + |p_2| + \cdots + |p_n|$, that is, the sum of $|p|$-operators along all the n coordinate directions. As explained in the Introduction, using the short distance regularisation of the underlying lattice, we really have $\eta^{-1}[1 - e^{-\eta|p|}]$ instead of $|p|$, and the operator inequality (9) holds. The left-hand side of (9) corresponds to the positive operator $(-\mathcal{A})$, and the right-hand side corresponds to $(-\mathcal{S})$. The key Lemma 4.6 below asserts that $0 \le (-\mathcal{A}) \le C(n)(-\mathcal{S})$ in the sense of quadratic forms. The main purpose of this replacement is that $\mathcal{A}$ averages independently in every direction; therefore, $\mathcal{A}$, acting on the function $g = f \circ \phi$, has the effect that it averages in *all* the $i_1, i_2, \dots$ indices in the definition of $P(G)$, (19). These averages yield traces of products $\Im GA\Im GA \dots \Im GA$ for which we have a good local law on the set $\widehat{\Omega}$.

We now explain the origin of the errors in (50). The error in the multi-G local laws give the crucial fourth error $1/\sqrt{N\eta}$ in (50). The other errors come from various approximations:

the dynamics commute with the localization up to an error of order ℓ/K by Lemma 4.3, the short-range cutoff dynamics approximates the original one up to time T_1 with an error of order NT_1/ℓ, while the removed long-range part contributes with an error or order $N\eta/\ell$ to the Dirichlet form. The last $1/\sqrt{K}$ error term is technical; we do the analysis for typical index configurations where no two indices coincide and the coinciding indices have a volume factor of order $1/\sqrt{K}$ smaller than the total volume.

Now, we start with the actual proof. All the estimates in this proof hold uniformly for $\boldsymbol{y} \in \Lambda^n$ supported on $\mathcal{J}$; hence, from now on we fix a particle configuration $\boldsymbol{y}$. To make the presentation clearer, we drop the parameters $\boldsymbol{y}, K, \ell$ and use the short-hand notations $h_t(\boldsymbol{x}) = h_t(\boldsymbol{x}; \ell, K, \boldsymbol{y})$, $\mathrm{Av} = \mathrm{Av}(K, \boldsymbol{y})$, $\mathrm{Av}(\boldsymbol{x}) = \mathrm{Av}(\boldsymbol{x}; K, \boldsymbol{y})$, etc. Moreover, for any $\boldsymbol{i}, \boldsymbol{j} \in [N]^n$ by $\sum_{\boldsymbol{i}}^*$ or $\sum_{\boldsymbol{ij}}^*$, we denote the summations over indices that are all distinct; that is, the $i_1, \dots, i_n$, in the first sum, and $i_1, \dots, i_n, j_1, \dots, j_n$, in the second sum are all different. The same convention holds for summations over $\boldsymbol{a}, \boldsymbol{b} \in [2n]^n$.

Let

$$a_{ij} = a_{ij}(t) := \frac{\eta}{N((\lambda_i(t) - \lambda_j(t))^2 + \eta^2)}, \tag{51}$$

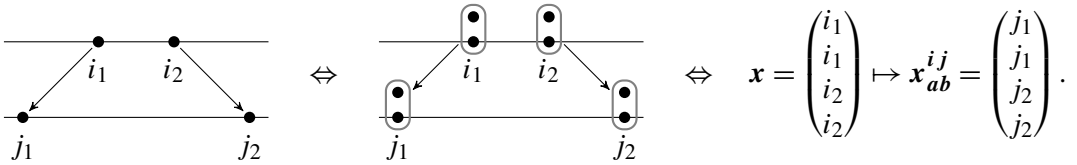
and define their short-range version $a_{ij,\mathcal{S}}$, as in (39). Define the operator $\mathcal{A} = \mathcal{A}(t)$ by

$$\mathcal{A}(t) := \sum_{\boldsymbol{i},\boldsymbol{j} \in [N]^n}^* \mathcal{A}_{\boldsymbol{ij}}(t), \qquad \mathcal{A}_{\boldsymbol{ij}}(t)h(\boldsymbol{x}) := \frac{1}{\eta}\left(\prod_{r=1}^n a^{\mathcal{S}}_{i_r,j_r}(t)\right) \sum_{\boldsymbol{a},\boldsymbol{b} \in [2n]^n}^* \big(h(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}}) - h(\boldsymbol{x})\big), \tag{52}$$

where

$$\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}} := \boldsymbol{x} + \left(\prod_{r=1}^n \delta_{x_{a_r} i_r} \delta_{x_{b_r} i_r}\right) \sum_{r=1}^n (j_r - i_r)(\boldsymbol{e}_{a_r} + \boldsymbol{e}_{b_r}). \tag{53}$$

We now explain the difference between the jump operator (53) and the one defined in (33). The operator (33) 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 $\{i_1, \dots, i_n\} \cap \{j_1, \dots, j_n\} = \varnothing$, then $\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}} \neq \boldsymbol{x}$ iff for all $r \in [n]$ it holds that $x_{a_r} = x_{b_r} = i_r$; see, for example,



Note that $\mu(\boldsymbol{x}) \equiv 1$ on Λ^n is a reversible measure for the generator $\mathcal{A}(t)$ (as a consequence of $(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}})^{\boldsymbol{ji}}_{\boldsymbol{ab}} = \boldsymbol{x}$ for any fixed $\boldsymbol{a}, \boldsymbol{b}$ and for any $\boldsymbol{x}$ such that $\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}} \neq \boldsymbol{x}$), and that $\pi(\boldsymbol{x}) \sim C(n)\mu(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Lambda^n$ (see (37)). We define the scalar product with respect to the measure $\mu(\boldsymbol{x})$ analogously to (35), and we denote it by $\langle \cdot, \cdot \rangle_{\Lambda^n,\mu}$.

We now analyse the time evolution of $\|h_t\|_2^2$,

$$\partial_t \|h_t\|_2^2 = 2\langle h_t, \mathcal{S}(t)h_t \rangle_{\Lambda^n}. \tag{54}$$

The main ingredient to give an upper bound on (54) is the following lemma, whose proof is postponed at the end of this section.

LEMMA 4.6. *Let $\mathcal{S}(t)$, $\mathcal{A}(t)$ be the generators defined in (42) and (52), respectively. Then there, exists a constant $C(n) > 0$, which depends only on n, such that*

$$\langle h, \mathcal{S}(t)h \rangle_{\Lambda^n,\pi} \le C(n)\langle h, \mathcal{A}(t)h \rangle_{\Lambda^n,\mu} \le 0 \tag{55}$$

for any $h \in L^2(\Lambda^n)$, on the very high probability set Ω.

From now on, by $C(n)$ we denote a constant that depends only on n and that may change from line to line.

Next, combining (54)–(55) 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

$$\begin{aligned}\partial_t \|h_t\|_2^2 &\le C(n)\langle h_t, \mathcal{A}(t) h_t\rangle_{\Lambda^n,\mu} \\ &= \frac{C(n)}{2\eta} \sum_{\boldsymbol{x}\in\Lambda^n} \sideset{}{^*}\sum_{\boldsymbol{i},\boldsymbol{j}\in[N]^n} \left(\prod_{r=1}^n a^{\mathcal{S}}_{i_r j_r}(t)\right) \sideset{}{^*}\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n} \overline{h_t}(\boldsymbol{x})\big(h_t(\boldsymbol{x}_{\boldsymbol{ab}}^{\boldsymbol{ij}}) - h_t(\boldsymbol{x})\big)\Psi(\boldsymbol{x}),\end{aligned} \tag{56}$$

where for any fixed $\boldsymbol{i}, \boldsymbol{a}, \boldsymbol{b}$ we defined

$$\Psi(\boldsymbol{x}) = \Psi_{\boldsymbol{i},\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x}) := \left(\prod_{r=1}^n \delta_{x_{a_r} i_r}\delta_{x_{b_r} i_r}\right).$$

Define

$$\Gamma := \{\boldsymbol{x} \in \Lambda^n : d(\boldsymbol{x}, \boldsymbol{y}) \le 3K\} \subset \Lambda^n, \tag{57}$$

and note that, by the finite speed of propagation estimate in Proposition 4.2 and the support of $h_0(\boldsymbol{x})$, the function $h_t(\boldsymbol{x})$ is supported on Γ up to an exponentially small error term (see [26], equations (5.76)–(5.77), for a more detailed calculation). For simplicity, for the rest of the proof we treat $h_t(\boldsymbol{x})$ as if it were supported on Γ, neglecting the exponentially small error term of size $\pi(\Lambda^n \setminus \Gamma)e^{-N^\epsilon} \le N^{2n}e^{-N^\epsilon}$. Since the dynamics is a linear contraction in L^∞, this small error term remains small throughout the whole evolution.

Now, we consider the term with $|h_t(\boldsymbol{x})|^2$ in (56) (here, we use the notation $\Psi(\boldsymbol{x}) = \Psi_{\boldsymbol{i},\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x})$),

$$\begin{aligned}&-\sum_{\boldsymbol{x}\in\Gamma} |h_t(\boldsymbol{x})|^2 \sideset{}{^*}\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n} \sideset{}{^*}\sum_{\boldsymbol{i},\boldsymbol{j}} \Psi(\boldsymbol{x}) \left(\prod_{r=1}^n a^{\mathcal{S}}_{i_r j_r}(t)\right) \\ &\quad= -\sum_{\boldsymbol{x}\in\Gamma} |h_t(\boldsymbol{x})|^2 \sideset{}{^*}\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n} \sideset{}{^*}\sum_{\boldsymbol{i}} \Psi(\boldsymbol{x}) \prod_{r=1}^n \left(\sum_{j_r} a^{\mathcal{S}}_{i_r j_r}(t) + \mathcal{O}\left(\frac{N^\xi}{N\eta}\right)\right) \\ &\quad= -\sum_{\boldsymbol{x}\in\Gamma} |h_t(\boldsymbol{x})|^2 \sideset{}{^*}\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n} \sideset{}{^*}\sum_{\boldsymbol{i}} \Psi(\boldsymbol{x}) \prod_{r=1}^n \left(\sum_{j_r} a_{i_r j_r}(t) + \mathcal{O}\left(\frac{N^\xi}{N\eta} + \frac{N^{1+\xi}\eta}{\ell}\right)\right) \\ &\quad\le -C(n) \sum_{\boldsymbol{x}\in\Gamma} |h_t(\boldsymbol{x})|^2 \sideset{}{^*}\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n} \sideset{}{^*}\sum_{\boldsymbol{i}} \Psi(\boldsymbol{x})\end{aligned} \tag{58}$$

on the very high probability event Ω_ξ, where the error term in the second line comes from adding back the finitely many excluded summands $j_r \in \{i_1, \dots, i_n\}$ and $j_r \in \{j_1, \dots, j_{r-1}, j_{r+1}, \dots j_n\}$. The new error in the third line comes from removing the short-range restriction from $a^{\mathcal{S}}_{i_r j_r}$, that is, adding back the regimes $|j_r - i_r| > \ell$ using

$$\sum_{j_r : |j_r - i_r| > \ell} a_{i_r j_r}(t) \le \frac{N\eta}{\ell}. \tag{59}$$

Finally, to go from the third to the fourth line in (58) we used the local law

$$\sum_{j_r} a_{i_r j_r}(t) = \big\langle \Im G_t(\lambda_{i_r} + \mathrm{i}\eta)\big\rangle = \Im m_t(\lambda_{i_r} + \mathrm{i}\eta) + \mathcal{O}\big(N^\xi (N\eta)^{-1}\big) \tag{60}$$

with very high probability on the event Ω_ξ, and that $\Im m_t(\lambda_{i_r} + \mathrm{i}\eta) \sim 1$ in the bulk whenever $\eta \ge N^{-1+\xi}$.

We now bound the last line in (58) in terms of $-\|h_t\|^2$ plus a small error term by removing the restriction from the $\boldsymbol{i}$-summation,

$$\begin{aligned}-\sum_{\boldsymbol{x}\in\Gamma}|h_t(\boldsymbol{x})|^2\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n}^{*}\sum_{\boldsymbol{i}}^{*}\Psi(\boldsymbol{x}) &= -\sum_{\boldsymbol{x}\in\Gamma}|h_t(\boldsymbol{x})|^2\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n}^{*}\sum_{\boldsymbol{i}}\Psi(\boldsymbol{x})\\ &\quad+\sum_{\boldsymbol{x}\in\Gamma}|h_t(\boldsymbol{x})|^2\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n}^{*}\left(\sum_{\boldsymbol{i}}-\sum_{\boldsymbol{i}}^{*}\right)\Psi(\boldsymbol{x})\\ &\le -C(n)\|h_t\|_2^2+C(n)N^{\xi}K^{n-1}.\end{aligned}\tag{61}$$

To estimate the first term in the right-hand side, we used that $\sum_{\boldsymbol{ab}}^{*}\sum_{\boldsymbol{i}}\Psi_{\boldsymbol{i},\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x})\ge 1$, $1\ge C(n)\pi(\boldsymbol{x})$ for all $\boldsymbol{x}\in\Gamma$, and that we can add back the regime $\Lambda^n\setminus\Gamma$ at the price of a negligible $N^n e^{-N^{\epsilon}}$ error term by finite speed of propagation. For the second term in the right-hand side of (61), we estimated $\|h_t\|_\infty\le N^{\xi}$ as a consequence of $\|h_0\|_\infty\le N^{\xi}$ and the fact that the evolution is an L^∞-contraction. Finally we used the fact that

$$\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n}^{*}\left(\sum_{\boldsymbol{i}}-\sum_{\boldsymbol{i}}^{*}\right)\Psi_{\boldsymbol{i},\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x})\neq 0,$$

only if there exist $a,b,c,d\in[2n]$, all distinct, such that $\boldsymbol{x}_a=\boldsymbol{x}_b=\boldsymbol{x}_c=\boldsymbol{x}_d$. The volume of this one codimensional subset of Γ is $C(n)K^{n-1}$, that is, by factor K^{-1} smaller than the volume of Γ which is of order K^n.

Finally, combining (58) and (61), we conclude the estimate for the term containing $|h_t(\boldsymbol{x})|^2$ in (56),

$$-\sum_{\boldsymbol{x}\in\Gamma}|h_t(\boldsymbol{x})|^2\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n}^{*}\sum_{\boldsymbol{i},\boldsymbol{j}}^{*}\Psi(\boldsymbol{x})\left(\prod_{i=1}^{r}a^{\mathcal{S}}_{i_r j_r}(t)\right)\le -C_1(n)\|h_t\|_2^2+C(n)N^{\xi}K^{n-1}.\tag{62}$$

Then, using (56) together with (62), we conclude that

$$\begin{aligned}\partial_t\|h_t\|_2^2 &\le -\frac{C_1(n)}{\eta}\|h_t\|_2^2+\frac{C(n)N^{\xi}K^{n-1}}{\eta}\\ &\quad+\frac{C_2(n)}{\eta}\sum_{\boldsymbol{x}\in\Gamma}|h_t(\boldsymbol{x})|\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n}^{*}\sum_{\boldsymbol{i}}^{*}\Psi(\boldsymbol{x})\left|\sum_{\boldsymbol{j}}^{*}\left(\prod_{r=1}^{n}a^{\mathcal{S}}_{i_r j_r}(t)\right)h_t\big(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}}\big)\right|\end{aligned}\tag{63}$$

for some constants $C_1(n),C_2(n)>0$, on the event Ω_ξ with $\xi>0$ arbitrarily small. In order to conclude the bound of $\partial_t\|h_t\|_2^2$, we are now left with the estimate of the last line in (63).

In the remainder of the proof, we will show that

$$\begin{aligned}&\frac{C_2(n)}{\eta}\sum_{\boldsymbol{x}\in\Gamma}|h_t(\boldsymbol{x})|\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n}^{*}\sum_{\boldsymbol{i}}^{*}\Psi(\boldsymbol{x})\left|\sum_{\boldsymbol{j}}^{*}\left(\prod_{r=1}^{n}a^{\mathcal{S}}_{i_r j_r}(t)\right)h_t\big(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}}\big)\right|\\ &\quad\le\frac{C_1(n)}{2\eta}\|h_t\|_2^2+\frac{C_3(n)}{\eta}\mathcal{E}^2K^n\end{aligned}\tag{64}$$

with $\mathcal{E}$ defined in (50) and $C_1(n)$ being the constant from the first line of (63). Note that, using (64), we readily conclude the proof of (49) by

$$\partial_t\|h_t\|_2^2\le-\frac{C_1(n)}{2\eta}\|h_t\|_2^2+\frac{C_3(n)}{\eta}\mathcal{E}^2K^n,\tag{65}$$

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

We now conclude the proof of (49) proving the bound in (64). We start with the analysis of

$$\sum_{\boldsymbol{j}\in[N]^n}^{*}\Big(\prod_{r=1}^{n} a^{\mathcal{S}}_{i_r j_r}(t)\Big)h_t(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}}) \tag{66}$$

for any fixed $\boldsymbol{x}\in\Gamma$, $\boldsymbol{i}\in[N]^n$, $\boldsymbol{a},\boldsymbol{b}\in[2n]^n$ with all distinct coordinates such that $\Psi(\boldsymbol{x})\neq 0$. It will be very important that the configuration $\phi(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}})$ contains exactly one particle at every index j_r; that is, we have

$$\prod_{l=1}^{N}(n_l(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}})-1)!!=1. \tag{67}$$

Similarly to [26], equations (5.89)–(5.91), equations (5.95)–(5.97), using that the function $f(\boldsymbol{x})\equiv\mathbf{1}(n \text{ even})$ is in the kernel of $\mathcal{S}(t)$, for any fixed $\boldsymbol{x}\in\Gamma$, and for any fixed $\boldsymbol{i}$, $\boldsymbol{a}$, $\boldsymbol{b}$, we conclude that

$$\begin{aligned}
&h_t(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}})\\
&\quad=\mathcal{U}_{\mathcal{S}}(0,t)\big((\mathrm{Av}\,g_0)(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}})-(\mathrm{Av}\,\mathbf{1}(n \text{ even}))(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}})\big)\\
&\quad=\mathrm{Av}(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}})\big(\mathcal{U}_{\mathcal{S}}(0,t)g_0(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}})-\mathbf{1}(n \text{ even})\big)+\mathcal{O}\Big(\frac{N^{\epsilon+n\xi}\ell}{K}\Big)\\
&\quad=\Big(\mathrm{Av}(\boldsymbol{x})+\mathcal{O}\Big(\frac{\ell}{K}\Big)\Big)\Big(\mathcal{U}(0,t)g_0(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}})-\mathbf{1}(n \text{ even})+\mathcal{O}\Big(\frac{N^{1+n\xi}t}{\ell}\Big)\Big)\\
&\qquad+\mathcal{O}\Big(\frac{N^{\epsilon+n\xi}\ell}{K}\Big)\\
&\quad=\mathrm{Av}(\boldsymbol{x})\big(g_t(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}})\big)-\mathbf{1}(n \text{ even})\big)+\mathcal{O}\Big(\frac{N^{\epsilon+n\xi}\ell}{K}+\frac{N^{1+n\xi}t}{\ell}\Big),
\end{aligned} \tag{68}$$

where the error terms are uniform in $\boldsymbol{x}\in\Gamma$. Note that to go from the first to the second line in (68) we used Lemma 4.3, to go from the second to the third line we used Lemma 4.4 together with the a priori bound $\|g_t\|_\infty\le N^{n\xi}$ for any $0\le t\le T$ on the very high probability event $\widehat{\Omega}_{\omega,\xi}$, and that

$$|\mathrm{Av}(\boldsymbol{x})-\mathrm{Av}(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}})|\le\frac{1}{K}\|\boldsymbol{x}-\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}}\|_1\le\frac{2n\ell}{K},$$

where $\|\mathbf{x}\|_1=\sum_{c=1}^{2n}|x_c|$. To go from the third to the fourth line in (68), we used that $|\mathrm{Av}(\boldsymbol{x})|\le 1$ and again that $\|g_t\|_\infty\le N^{n\xi}$. Then, from (68), we conclude that

$$\begin{aligned}
\sum_{\boldsymbol{j}}^{*}\Big(\prod_{r=1}^{n}a^{\mathcal{S}}_{i_r j_r}(t)\Big)h_t(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}})&=\mathrm{Av}(\boldsymbol{x})\sum_{\boldsymbol{j}}^{*}\Big(\prod_{r=1}^{n}a^{\mathcal{S}}_{i_r j_r}(t)\Big)\big(g_t(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{a},\boldsymbol{b}})-\mathbf{1}(n \text{ even})\big)\\
&\quad+\mathcal{O}\Big(\frac{N^{\epsilon+n\xi}\ell}{K}+\frac{N^{1+n\xi}T_1}{\ell}\Big).
\end{aligned} \tag{69}$$

From now on, we will omit the $\mathrm{Av}(\boldsymbol{x})$ prefactor in (69), since $|\mathrm{Av}(\boldsymbol{x})|\le 1$.

Using the definition of g_t from (30) and (20), for any $\boldsymbol{x} \in \Gamma$ such that $\Psi(\boldsymbol{x}) \neq 0$, and for any fixed $\boldsymbol{i}, \boldsymbol{a}, \boldsymbol{b}$, dropping the t-dependence of the eigenvalues $\lambda_i = \lambda_i(t)$, we have

$$\begin{aligned}
&\sum_{\boldsymbol{j}}^{*}\Big(\prod_{r=1}^{n} a^{S}_{i_r j_r}(t)\Big)\big(g_t(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}}) - \mathbf{1}(n \text{ even})\big) \\
&\quad = \sum_{\boldsymbol{j}}^{*}\Big(\prod_{r=1}^{n} a_{i_r j_r}(t)\Big)\Big(\frac{N^{n/2}}{\langle A^2\rangle^{n/2} 2^{n/2}(n-1)!!}\sum_{G\in\mathcal{G}_{\boldsymbol{\eta j}}} P(G) - \mathbf{1}(n \text{ even})\Big) \\
&\quad\quad + \mathcal{O}\Big(\frac{N^{1+\xi}\eta}{\ell}\Big) \qquad (70)\\
&\quad = \sum_{\boldsymbol{j}}\Big(\prod_{r=1}^{n} a_{i_r j_r}(t)\Big)\Big(\frac{N^{n/2}}{\langle A^2\rangle^{n/2} 2^{n/2}(n-1)!!}\sum_{G\in\mathcal{G}_{\boldsymbol{\eta j}}} P(G) - \mathbf{1}(n \text{ even})\Big) \\
&\quad\quad + \mathcal{O}\Big(\frac{N^{\xi}}{N\eta} + \frac{N^{1+\xi}\eta}{\ell}\Big).
\end{aligned}$$

Note that in (70) we used the notation $\boldsymbol{\eta}^{\boldsymbol{j}} := \phi(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}})$ to denote the particle configuration which has exactly one particle at each site $\{j_1, \dots, j_n\}$. Note that in the last line of (70) we do not exclude the possibility that two indices j may assume the same value, since the sum is unrestricted. In the second and third lines of (70), we simply omitted the conditional expectation $\mathbf{E}[\cdots|\boldsymbol{\lambda}]$ to shorten the formulas. Since all subsequent estimates hold with high probability, the conditional expectation does not play a role. When going from the first to the second line of (70), we removed the short-range restriction, as in (59), by adding back the summations over the regimes $|j_r - i_r| > \ell$, and we also used (67) since the coordinates of $\boldsymbol{j}$ are all distinct, and so that $\mathcal{M}(\boldsymbol{\eta}^{\boldsymbol{j}}) = 1$ in the definition of g_t in (30) and (20). Additionally, the error term in the third line of (70) comes from adding back the missing j_r-summations; in this bound we used the a priori bound $|P(G)| \le N^{\xi - n/2}$ on the very high probability event $\widehat{\Omega}_{\omega,\xi}$ and (60).

We now use the definition of $P(G)$ in (19) on the right-hand side of (70). Since every particle is doubled, we may rewrite the sum over perfect matchings as

$$\sum_{G\in\mathcal{G}_{\boldsymbol{\eta j}}} P(G) = \sum_{G\in \mathrm{Gr}_2[n]}\ \prod_{(v_1\cdots v_k)\in \mathrm{Cyc}(G)} (2k-2)!!\, p_{j_{v_1} j_{v_2}} \cdots p_{j_{v_k} j_{v_1}}, \qquad (71)$$

where $\mathrm{Gr}_2[n]$ denotes the set of 2-regular multigraphs (possibly with loop-edges) on $[n]$ and $\mathrm{Cyc}(G)$ denoting the collection of cycles in any such graph $G \in \mathrm{Gr}_2[n]$. The combinatorial factor $(2k-2)!!$ is due to the fact that, for each cycle in G, there are $(2k-2)!!$ equivalent perfect matchings giving the very same cyclic monomial. For example, for $n = 2$ there are two 2-regular multi-graphs, (11), (22) and (12), (12), and thus $\sum_{G\in\mathcal{G}_{\boldsymbol{\eta j}}} P(G) = 2p^2_{j_1 j_2} + p_{j_1 j_1} p_{j_2 j_2}$. Similarly, for $n = 3$, there are the graphs

$$\begin{aligned}
&\{(11),(22),(33)\}, \qquad \{(12),(12),(33)\}, \qquad \{(13),(13),(22)\},\\
&\{(23),(23),(11)\}, \qquad \{(12),(23),(13)\}
\end{aligned}$$

yielding

$$\begin{aligned}
\sum_{G\in\mathcal{G}_{\boldsymbol{\eta j}}} P(G) &= p_{j_1 j_1} p_{j_2 j_2} p_{j_3 j_3} + 2p_{j_1 j_1} p^2_{j_2 j_3} + 2p_{j_2 j_2} p^2_{j_1 j_3} \\
&\quad + 2p_{j_3 j_3} p^2_{j_1 j_2} + 8p_{j_1 j_2} p_{j_2 j_3} p_{j_1 j_3}.
\end{aligned}$$

For each graph $G \in \mathrm{Gr}_2[n]$, we may use the spectral theorem to perform the $\boldsymbol{j}$ summation as

$$\sum_{j_{v_1},\dots,j_{v_k}} \Big(\prod_{r\in[k]} a_{i_{v_r} j_{v_r}}(t) \Big) p_{j_{v_1} j_{v_2}} \cdots p_{j_{v_k} j_{v_1}} = N^{1-k} F_k(v_1,\dots,v_k) \tag{72}$$

with

$$F_k(v_1,\dots,v_k) := \big\langle \Im G_t(\lambda_{i_{v_1}} + \mathrm{i}\eta) A \cdots \Im G_t(\lambda_{i_{v_k}} + \mathrm{i}\eta) A \big\rangle.$$

Since each vertex appears in exactly one cycle, we can use (72) to perform the summation for the indices corresponding to any cycle separately and obtain

$$\sum_{\boldsymbol{j}} \Big(\prod_{r=1}^{n} a_{i_r j_r}(t) \Big) \sum_{G\in\mathcal{G}_{\boldsymbol{\eta j}}} P(G) = \sum_{E\in\mathrm{Gr}_2[n]} \prod_{(v_1\cdots v_k)\in\mathrm{Cyc}(E)} (2k-2)!! N^{1-k} F_k(v_1,\dots,v_k). \tag{73}$$

We note that, from (47) for each $k \geq 1$, we have the estimate

$$F_k(v_1,\dots,v_k) = \mathbf{1}(k=2)\langle A^2\rangle \Im m(z_{i_{v_1}}) \Im m(z_{i_{v_k}}) + \mathcal{O}\Big(N^{\xi} \frac{N^{k/2-1}}{\sqrt{N\eta}} \Big) \tag{74}$$

on the high-probability set $\widehat{\Omega}$. By using (74) within (73) and using the fact that there are $\mathbf{1}(n \text{ even})(n-1)!!$ graphs in $\mathrm{Gr}_2[n]$ all of which cycles have length two, it follows that

$$(73) = \mathbf{1}(n \text{ even})(n-1)!! 2^{n/2} N^{-n/2} \langle A^2\rangle^{n/2} \prod_{r\in[n]} \Im m(z_{i_r}) + \mathcal{O}\Big(N^{\xi} \frac{N^{-n/2}}{\sqrt{N\eta}} \Big), \tag{75}$$

and from (70) we conclude

$$\Psi(\boldsymbol{x}) \sum_{\boldsymbol{j}}^{*} \Big(\prod_{r=1}^{n} a^{\mathcal{S}}_{i_r j_r}(t) \Big) \big(g_t(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}}) - \mathbf{1}(n \text{ even}) \big) = \Psi(\boldsymbol{x}) \mathcal{O}\Big(\frac{N^{\xi}}{\sqrt{N\eta}} + \frac{N^{\xi}}{N\eta} + \frac{N^{1+\xi}\eta}{\ell} \Big). \tag{76}$$

Combining (69) and (76), we get that

$$\Psi(\boldsymbol{x}) \Big| \sum_{\boldsymbol{j}}^{*} \Big(\prod_{r=1}^{n} a^{\mathcal{S}}_{i_r j_r}(t) \Big) h_t(\boldsymbol{x}^{\boldsymbol{ij}}_{\boldsymbol{ab}}) \Big| \leq C(n) \Psi(\boldsymbol{x}) \mathcal{E}, \tag{77}$$

and finally, by (77), we conclude that

$$\text{l.h.s. (64)} \leq \frac{C_1(n)}{2\eta} \|h_t\|_2^2 + \frac{C_3(n)}{\eta} \mathcal{E}^2 K^n, \tag{78}$$

where we used that, for any fixed $\boldsymbol{x} \in \Lambda^n$, we have

$$\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n}^{*} \sum_{\boldsymbol{i}}^{*} \Psi_{\boldsymbol{i},\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x}) \leq C(n),$$

and that

$$\Big| \sum_{\boldsymbol{x}\in\Gamma} h_t(\boldsymbol{x}) \mathcal{E} \Big| \leq \frac{C_1(n)}{2} \sum_{\boldsymbol{x}\in\Gamma} \pi(\boldsymbol{x}) |h_t(\boldsymbol{x})|^2 + C_3(n) \mathcal{E}^2 K^n,$$

by the Schwarz inequality, the bound $1 \leq \pi(\boldsymbol{x})$ from (37), and $\sum_{\boldsymbol{x}\in\Gamma} \pi(\boldsymbol{x}) \leq C(n) K^n$. Note that, by balancing between the two terms in the Schwarz inequality, we could achieve the

same constant $C_1(n)$ with an additional 1/2 factor in front of the $\|h_t\|_2^2$ term as in the leading term in (63) with a minus sign. This concludes the proof of the bound in (64). □

PROOF OF LEMMA 4.6. All along the proof $C(n) > 0$ is a constant that depends only on n and that may change from line to line.

We consider

$$\begin{aligned}\langle h, \mathcal{S}(t)h\rangle_{\Lambda^n,\pi} &= -\frac{1}{2}\sum_{\boldsymbol{x}\in\Lambda^n}\pi(\boldsymbol{x})\sum_{j\neq i}c_{ij}^{\mathcal{S}}(t)\frac{n_j(\boldsymbol{x})+1}{n_i(\boldsymbol{x})-1}\sum_{a\neq b\in[2n]}\big|h\big(\boldsymbol{x}_{ab}^{ij}\big)-h(\boldsymbol{x})\big|^2\\ &\le -\frac{C(n)}{\eta}\sum_{\boldsymbol{x}\in\Lambda^n}\sum_{j\neq i}a_{ij}^{\mathcal{S}}(t)\sum_{a\neq b\in[2n]}\big|h\big(\boldsymbol{x}_{ab}^{ij}\big)-h(\boldsymbol{x})\big|^2\end{aligned}\tag{79}$$

and

$$\langle h, \mathcal{A}(t)h\rangle_{\Lambda^n,\mu} = -\frac{1}{2\eta}\sum_{\boldsymbol{x}\in\Lambda^n}\sum_{\boldsymbol{i},\boldsymbol{j}}^{*}\Bigg(\prod_{r=1}^{n}a_{i_r j_r}^{\mathcal{S}}(t)\Bigg)\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n}^{*}\big|h\big(\boldsymbol{x}_{\boldsymbol{ab}}^{\boldsymbol{ij}}\big)-h(\boldsymbol{x})\big|^2.\tag{80}$$

Note that in (79) we used that $a_{ij}^{\mathcal{S}}(t) \le \eta c_{ij}^{\mathcal{S}}(t)$ to compare the kernels, that $\pi(\boldsymbol{x}) \ge 1$ uniformly in $\boldsymbol{x} \in \Lambda^n$, and finally that $n_j(\boldsymbol{x}) + 1 \ge 1$, $1 \le n_i(\boldsymbol{x}) - 1 \le n$ for $\boldsymbol{x}$ and i such that $h(\boldsymbol{x}_{ab}^{ij}) \ne h(\boldsymbol{x})$.

We start with the bound

$$\begin{aligned}&\sum_{\boldsymbol{x}\in\Lambda^n}\sum_{\boldsymbol{i},\boldsymbol{j}\in[N]^n}^{*}\Bigg(\prod_{r=1}^{n}a_{i_r j_r}^{\mathcal{S}}(t)\mathbf{1}\big(n_{i_r}(\boldsymbol{x})>0\big)\Bigg)\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n}^{*}\big|h\big(\boldsymbol{x}_{\boldsymbol{ab}}^{\boldsymbol{ij}}\big)-h(\boldsymbol{x})\big|^2\\ &\quad\le C(n)\sum_{\boldsymbol{x}\in\Lambda^n}\sum_{\boldsymbol{i},\boldsymbol{j}}^{*}\Bigg(\prod_{r=1}^{n}a_{i_r j_r}^{\mathcal{S}}(t)\mathbf{1}\big(n_{i_r}(\boldsymbol{x})>0\big)\Bigg)\sum_{l=1}^{n}\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n}^{*}\big|h\big((\boldsymbol{y}_{l-1})_{a_l b_l}^{i_l j_l}\big)-h(\boldsymbol{y}_{l-1})\big|^2,\end{aligned}\tag{81}$$

where we recursively defined $\boldsymbol{y}_0 = \boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2 \ldots, \boldsymbol{y}_n = \boldsymbol{x}_{\boldsymbol{ab}}^{\boldsymbol{ij}}$ by performing the jumps $i_1 \to j_1$, $i_2 \to j_2$, etc., one by one (assuming that the choice of (a_l, b_l) allows it), otherwise $\boldsymbol{y}_l = \boldsymbol{y}_{l-1}$,

$$\boldsymbol{y}_0 = \boldsymbol{y}_0(\boldsymbol{x}) := \boldsymbol{x}, \qquad \boldsymbol{y}_l = \boldsymbol{y}_l(\boldsymbol{x}) := (\boldsymbol{y}_{l-1})_{a_l b_l}^{i_l j_l}.\tag{82}$$

In the first line of (81), we could add the indicator $\mathbf{1}(n_{i_r}(\boldsymbol{x}) > 0)$ since in case $n_{i_r}(\boldsymbol{x}) = 0$ for some r it holds that $\boldsymbol{x}_{\boldsymbol{ab}}^{\boldsymbol{ij}} = \boldsymbol{x}$. Note that, to go from the first to the second line of (81), we wrote a telescopic sum

$$h\big(\boldsymbol{x}_{\boldsymbol{ab}}^{\boldsymbol{ij}}\big)-h(\boldsymbol{x}) = \sum_{l=1}^{n}\big[h\big((\boldsymbol{y}_{l-1})_{a_l b_l}^{i_l j_l}\big)-h(\boldsymbol{y}_{l-1})\big]$$

and used Schwarz inequality.

Next, we consider

$$
\begin{aligned}
&\sum_{l=1}^{n}\sum_{\boldsymbol{x}\in\Lambda^n}\sum_{\boldsymbol{i},\boldsymbol{j}}^{*}\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n}^{*}\Bigg(\prod_{r=1}^{n}a^{\mathcal{S}}_{i_r j_r}(t)\mathbf{1}\big(n_{i_r}(\boldsymbol{x})>0\big)\Bigg)\big|h\big((\boldsymbol{y}_{l-1})^{i_l j_l}_{a_l b_l}\big)-h(\boldsymbol{y}_{l-1})\big|^2\\
&\quad=\sum_{l=1}^{n}\sum_{\boldsymbol{w}\in\Lambda^n}\sum_{\boldsymbol{i},\boldsymbol{j}}^{*}\sum_{\boldsymbol{a},\boldsymbol{b}\in[2n]^n}^{*}\Bigg(\prod_{r=1}^{n}a^{\mathcal{S}}_{i_r j_r}(t)\mathbf{1}\big(n_{i_r}(\boldsymbol{z}_{l-1})>0\big)\Bigg)\big|h\big(\boldsymbol{w}^{i_l j_l}_{a_l b_l}\big)-h(\boldsymbol{w})\big|^2\\
&\quad\le C(n)\sum_{\boldsymbol{w}\in\Lambda^n}\sum_{l=1}^{n}\sum_{i_l\neq j_l}a^{\mathcal{S}}_{i_l j_l}(t)\sum_{a_l\neq b_l\in[2n]}\big|h\big(\boldsymbol{w}^{i_l j_l}_{a_l b_l}\big)-h(\boldsymbol{w})\big|^2\\
&\qquad\times\Bigg(\prod_{r\neq l}\sum_{i_r,j_r}a_{i_r j_r}\big[\mathbf{1}\big(n_{i_r}(\boldsymbol{w})>0\big)+\mathbf{1}\big(n_{j_r}(\boldsymbol{w})>0\big)\big]\Bigg)\\
&\quad\le C(n)\sum_{\boldsymbol{w}\in\Lambda^n}\sum_{l=1}^{n}\sum_{i_l\neq j_l}a^{\mathcal{S}}_{i_l j_l}(t)\sum_{a_l\neq b_l\in[2n]}\big|h\big(\boldsymbol{w}^{i_l j_l}_{a_l b_l}\big)-h(\boldsymbol{w})\big|^2\\
&\quad\le C(n)\sum_{\boldsymbol{w}\in\Lambda^n}\sum_{i\neq j}a^{\mathcal{S}}_{ij}(t)\sum_{a\neq b\in[2n]}\big|h\big(\boldsymbol{w}^{ij}_{ab}\big)-h(\boldsymbol{w})\big|^2.
\end{aligned}
\tag{83}
$$

Note that, to go from the first to the second line, we did the change of variables $\boldsymbol{w}=\boldsymbol{y}_{l-1}(\boldsymbol{x})$, we used that $(\boldsymbol{x}^{i_l j_l}_{a_l b_l})^{j_l i_l}_{a_l b_l}=\boldsymbol{x}$ for any $\boldsymbol{x}\in\Lambda^n$ such that $\prod_r\mathbf{1}(n_{i_r}(\boldsymbol{x})>0)$, and we defined $\boldsymbol{z}_{l-1}=((\boldsymbol{w}^{j_{l-1}i_{l-1}}_{a_{l-1}b_{l-1}})\dots)^{j_1 i_1}_{a_1 b_1}$. Moreover, to go from the second to the third line in (83) we used that

$$
\begin{aligned}
\prod_{r\in[n]\setminus\{l\}}\mathbf{1}\big(n_{i_r}(\boldsymbol{z}_{l-1})>0\big)\le C(n)&\Bigg(\prod_{r=1}^{l-1}\big[\mathbf{1}\big(n_{j_r}(\boldsymbol{w})>0\big)+\mathbf{1}\big(n_{i_r}(\boldsymbol{w})>0\big)\big]\Bigg)\\
&\times\Bigg(\prod_{r=l+1}^{n}\mathbf{1}\big(n_{i_r}(\boldsymbol{w})>0\big)\Bigg)
\end{aligned}
\tag{84}
$$

for $i_1,\dots,i_n$, $j_1,\dots,j_n$ all distinct, which follows by $n_{i_r}(\boldsymbol{z}_{l-1})=n_{i_r}(\boldsymbol{w})$ if $r\ge l+1$ and

$$
\mathbf{1}\big(n_{i_r}(\boldsymbol{z}_{l-1})>0\big)\le\mathbf{1}\big(n_{i_r}(\boldsymbol{w})>0\big)+\mathbf{1}\big(n_{j_r}(\boldsymbol{w})>0\big)
$$

for $r\le l-1$. In the penultimate inequality in (83), we also used that

$$
\prod_{r\neq l}\sum_{i_r j_r}a_{i_r j_r}\big[\mathbf{1}\big(n_{i_r}(\boldsymbol{w})>0\big)+\mathbf{1}\big(n_{j_r}(\boldsymbol{w})>0\big)\big]\le C(n)
\tag{85}
$$

on the very high probability event $\widehat{\Omega}$. Combining (79)–(80), (81), and (83), we finally conclude (55). □

4.4. *Proof of Proposition 3.2.* Fix $1\ll NT_1\ll\ell_1\ll K$ and $1\ll NT_2\ll\ell_2\ll K$, with $T_1\le T_2/2$. Define the *lattice generator* $\mathcal{W}(t)$ by

$$
\mathcal{W}(t):=\sum_{i\neq j\in[N]}\mathcal{W}_{ij}(t),\mathcal{W}_{ij}(t):=c^{\mathcal{W}}_{ij}(t)\frac{n_j(\boldsymbol{x})+1}{n_i(\boldsymbol{x})-1}\sum_{a\neq b\in[2n]}\big(h\big(\boldsymbol{x}^{ij}_{ab}\big)-h(\boldsymbol{x})\big)
\tag{86}
$$

with

$$
c^{\mathcal{W}}_{ij}(t):=\begin{cases}c_{ij}(t) & \text{if } i,j\in\mathcal{J}\text{ and }1\le|i-j|\le\ell_2,\\ \dfrac{N}{|i-j|^2} & \text{otherwise.}\end{cases}
\tag{87}
$$

Denote by $\mathcal{U}_{\mathcal{W}}(s,t)$ the semigroup associated to the generator $\mathcal{W}(t)$. Note that $\mathcal{W}(t)$ is the original generator of the Dyson eigenvector flow $\mathcal{L}$ from (32) on short scales and in the interval $\mathcal{J}$ well inside the bulk, while on large scales it has an equidistant jump rate. In [26] this replacement made up for the missing rigidity (regularity) control of the eigenvalues outside of a local interval $\mathcal{J}$; in our case its role is just to handle the somewhat different scaling of the eigenvalues near the edges. We follow the setup of [26] for convenience.

On the event Ω_ξ, the coefficients $c_{ij}^{\mathcal{W}}(t)$ satisfy [26], Assumption 6.8, with a rate $v = N^{1-\xi}$, for any arbitrary small $\xi > 0$; hence, all the results in [26], Section 6, apply to the generator $\mathcal{W}(t)$. Most importantly, the Dirichlet form of $\mathcal{W}(t)$ satisfies a Poincaré inequality and, consequently, we have an $L^2 \to L^\infty$ ultracontractive decay bound for the corresponding semigroup. Their scaling properties confirm the intuition that $\mathcal{W}(t)$ is a discrete analogue of the $|p| = \sqrt{-\Delta}$ operator in $\mathbf{R}^{2n}$. In the continuous setting, standard Sobolev inequality combined with the Nash method implies that

$$\|e^{-t|p|} f\|_{L^\infty(\mathbf{R}^{2n})} \le \frac{C(n)}{t^{n/2}} \|f\|_{L^2(\mathbf{R}^{2n})} \tag{88}$$

holds for any L^2 function on $\mathbf{R}^{2n}$. The same decay holds for the semigroup generated by $\mathcal{W}(t)$ by [26], Proposition 6.29, (recall that [26] uses n to denote the dimension of the space of $\boldsymbol{x}$'s, we use $2n$). We remark that the proofs in [26], Section 6, are designed for the more involved *coloured* dynamics; here, we need only its simpler *colourblind* version which immediately follows from the coloured version by ignoring the colors. In particular, in our case the *exchange operator* $\mathcal{E}_{ij}$ is identically zero. While a direct proof of the colourblind version is possible and it would require less combinatorial complexity, for brevity, we directly use the results of [26], Section 6.

For each $\boldsymbol{y}$ supported on $\mathcal{J}$, let $q_t(\boldsymbol{x}) = q_t(\boldsymbol{x}; \boldsymbol{y})$ be the solution of

$$\begin{cases} q_0(\boldsymbol{x}) = \mathrm{Av}(\boldsymbol{x}; K, \boldsymbol{y})\big(g_0(\boldsymbol{x}) - \mathbf{1}(n \text{ even})\big), & \\ \partial_t q_t(\boldsymbol{x}) = \mathcal{S}(t) q_t(\boldsymbol{x}) & \text{for } 0 < t \le T_1, \\ \partial_t q_t(\boldsymbol{x}) = \mathcal{W}(t) q_t(\boldsymbol{x}) & \text{for } T_1 < t \le T_2 \end{cases} \tag{89}$$

with $\mathcal{S}(t)$ being the short-range generator on a scale $\ell = \ell_1$ from (42). Note that $q_t = h_t$, for any $0 \le t \le T_1$, with h_t being the solution of (41).

By Proposition 4.5, choosing $\eta = N^{-\epsilon} T_1$, we have

$$\sup_{\boldsymbol{y}: y_a \in \mathcal{J}} \|q_{T_1}(\cdot; \boldsymbol{y})\|_2 \lesssim N^\epsilon K^{n/2} \left(\frac{\ell_1}{K} + \frac{N T_1}{\ell_1} + \frac{1}{\sqrt{N T_1}} + \frac{1}{\sqrt{K}} \right) \tag{90}$$

for any arbitrary small $\epsilon > 0$, where the supremum is over all the $\boldsymbol{y}$ supported on $\mathcal{J}$. We recall that, by the finite speed of propagation estimate in Proposition 4.2, together with [26], equation (7.12), the function q_t is supported on the subset of $\Gamma \subset \Lambda^n$ such that $d(\boldsymbol{x}, \boldsymbol{y}) \le 3K$ for any $\boldsymbol{x} \in \Gamma$ (modulo a negligible exponentially small error term). Then, using the ultracontractivity bound for the dynamics of $\mathcal{W}(t)$ from [26], Proposition 6.29, with $v = N^{1-\xi}$, we get that

$$\begin{aligned} \sup_{\boldsymbol{y}} \|q_{T_2}(\cdot; \boldsymbol{y})\|_\infty &= \sup_{\boldsymbol{y}} \|\mathcal{U}_{\mathcal{W}}(T_1, T_2) q_{T_1}(\cdot; \boldsymbol{y})\|_\infty \\ &\le \sup_{\boldsymbol{y}} \|(1 - \Pi)\mathcal{U}_{\mathcal{W}}(T_1, T_2) q_{T_1}(\cdot; \boldsymbol{y})\|_\infty \\ &\quad + \sup_{\boldsymbol{y}} \|\Pi \mathcal{U}_{\mathcal{W}}(T_1, T_2) q_{T_1}(\cdot; \boldsymbol{y})\|_\infty \\ &\lesssim \frac{N^{n\xi}}{[N(T_2 - T_1)]^{n/2}} \sup_{\boldsymbol{y}} \|q_{T_1}(\cdot; \boldsymbol{y})\|_2 + \frac{K^n}{N^{n(1-\xi)}}, \end{aligned} \tag{91}$$

where Π is the orthogonal projection into the kernel of $\mathcal{L}$, $\ker(\mathcal{L}) = \bigcap_{i\neq j} \ker(\mathcal{L}_{ij})$, defined in [26], Lemma 4.17. Note that in (91) we used that by [26], Corollary 4.20, it holds

$$\|\Pi q_{T_2}\|_\infty \le C(n) N^{-n} \|q_{T_2}\|_1 \le K^n N^{-n+n\xi},$$

since $\|q_{T_2}\|_\infty \le N^{n\xi}$ on the very high probability set $\widehat{\Omega}$. We remark that in [26], Proposition 6.29, $\mathcal{U}_{\mathcal{W}}$ is replaced by $\mathcal{U}$, but this does not play any role since the only assumption on $\mathcal{L}_{ij}$, used in [26], Section 6, is that $c_{ij}(t) \ge N^{1-\xi}|i-j|^{-2}$ (see [26], Definition 6.8). Combining (90)–(91), we conclude

$$\sup_{\boldsymbol{y}} \|q_{T_2}(\cdot;\boldsymbol{y})\|_\infty \lesssim N^{2\epsilon}\Big(\frac{K}{NT_2}\Big)^{n/2}\Big(\frac{\ell_1}{K} + \frac{NT_1}{\ell_1} + \frac{1}{\sqrt{NT_1}} + \frac{1}{\sqrt{K}}\Big), \tag{92}$$

where we used that $T_1 \le T_2/2$.

Now, we compare the solution q_t from (89) with the original dynamics g_t from (31). This is done, after several steps, using [26], Proposition 7.2, with $F_t(\boldsymbol{y};\boldsymbol{y})$ replaced by $\mathbf{1}(n \text{ even})$, asserting that

$$\sup_{\boldsymbol{y}} \big|q_{T_2}(\boldsymbol{y};\boldsymbol{y}) - \big(g_{T_2}(\boldsymbol{y}) - \mathbf{1}(n \text{ even})\big)\big| \lesssim N^{\epsilon}\Big(\frac{\ell_1}{K} + \frac{NT_1}{\ell_1} + \frac{\ell_2}{K} + \frac{NT_2}{\ell_2}\Big). \tag{93}$$

In particular, the only thing used about $F_t(\boldsymbol{y};\boldsymbol{y})$ in the proof of [26], Proposition 7.2, is that F_t is in the kernel of all $\mathcal{L}_{ij}$, and this is clearly the case for $\mathbf{1}(n \text{ even})$ as well. The origins of the error terms in (93) are as follows. The smooth cutoff given by the Av localising operator in the initial condition (89) commutes with the time evolution generated by $\mathcal{S}$ up an error of order ℓ_1/K; see Lemma 4.3. The difference between the original dynamics and the short-range dynamics in the time interval $t \in [0, T_1]$ yields the error NT_1/ℓ_1; see Lemma 4.4. Similar errors hold for the approximation of the original dynamics by the time evolution generated by $\mathcal{W}$ on the time interval $t \in [T_1, T_2]$, giving rise to the errors ℓ_2/K and $N(T_2 - T_1)/\ell_2 \le NT_2/\ell_2$.

Combining (92)–(93), we conclude that

$$\begin{aligned}
&\sup_{\boldsymbol{y}} \big|g_{T_2}(\boldsymbol{y}) - \mathbf{1}(n \text{ even})\big| \\
&\lesssim N^{2\epsilon}\Big(\frac{\ell_1}{K} + \frac{NT_1}{\ell_1} + \frac{\ell_2}{K} + \frac{NT_2}{\ell_2} + \Big(\frac{K}{NT_2}\Big)^{n/2}\Big(\frac{\ell_1}{K} + \frac{NT_1}{\ell_1} + \frac{1}{\sqrt{NT_1}} + \frac{1}{\sqrt{K}}\Big)\Big) \\
&\lesssim N^{-c/(20n)},
\end{aligned} \tag{94}$$

on the with very high probability event $\Omega_\xi \cap \widehat{\Omega}_{\xi,\epsilon}$ with choosing a very small ξ. In the last step we optimised the error terms in the second line of (94) with the choice of

$$K = N^c, \qquad T_2 = N^{-1-c/(10n)}K, \qquad \ell_2 = \sqrt{NKT_2}, \qquad \ell_1 = \sqrt{NKT_1}, \qquad T_1 = \frac{\sqrt{K}}{N}$$

with some small fixed $0 < c \le 1/2$. Finally, using that

$$\sup_{\boldsymbol{y}:y_a\in\mathcal{J}} \big|g_{T_2}(\boldsymbol{y}) - \mathbf{1}(n \text{ even})\big| = \sup_{\boldsymbol{\eta}} \big|f_{T_2}(\boldsymbol{\eta}) - \mathbf{1}(n \text{ even})\big|$$

by (30), where the supremum in the right-hand side is taken over configurations $\boldsymbol{\eta}$ such that $\eta_i = 0$ for $i \in [\delta N, (1-\delta)N]^c$ and $\sum_i \eta_i = n$. The bound in (94) concludes the proof of Proposition 3.2.

5. Local law bounds. In this section we prove the local laws needed to estimate the probability of the event $\widehat{\Omega}$ in (47). We recall [21] that the resolvent $G = (W - z)^{-1}$ of the Wigner matrix W is approximately equal,

$$G_{ab} = \delta_{ab} m + \mathcal{O}\Big(\frac{N^\xi}{\sqrt{N\Im z}}\Big), \qquad \langle G\rangle = m + \mathcal{O}\Big(\frac{N^\xi}{N\Im z}\Big) \tag{95}$$

to the Stieltjes transform $m = m_{\mathrm{sc}}(z)$ of the semicircular distribution $\rho_{\mathrm{sc}} = \sqrt{4 - x^2}/2\pi$ which solves the equation

$$-\frac{1}{m} = m + z. \tag{96}$$

PROPOSITION 5.1. *Let $k \geq 3$ and $z_1, \ldots, z_k \in \mathbf{C} \setminus \mathbf{R}$ with $N \min_i(\rho_i \eta_i) \geq N^\epsilon$ for some $\epsilon > 0$ with $\eta_i := |\Im z_i|$ and $\rho_i := \rho(z_i)$, $\rho(z) := |\Im m(z)|/\pi$. Then, for arbitrary traceless matrices $A_1, \ldots, A_k$ with $\|A_i\| \lesssim 1$, we have*

$$\big|\langle G_1 A_1 \ldots G_k A_k\rangle\big| \lesssim N^{\xi + (k-3)/2} \sqrt{\frac{\rho^*}{\eta_*}} \tag{97}$$

with very high probability for any $\xi > 0$, where $\rho^* := \max_i \rho_i$ and $\eta_* := \min \eta_i$.

PROOF. Using $WG - zG = I$ and (96), we write

$$G = m - m\underline{WG} + m\langle G - m\rangle G, \tag{98}$$

where

$$\underline{WG} = WG + \langle G\rangle G$$

denotes a renormalization of WG. More generally, for functions $f(W)$ we define

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

with $\partial_{\widetilde{W}}$ denoting the directional derivative in direction $\widetilde{W}$ and $\widetilde{W}$ being an independent GUE-matrix with expectation $\widetilde{\mathbf{E}}$. We now use (98) and (95) for $G_1 = G(z_1)$ and $m_1 = m(z_1)$ to obtain

$$\Big(1 - \mathcal{O}\Big(\frac{N^\xi}{N\eta_*}\Big)\Big)\Big\langle \prod_{i=1}^k (G_i A_i)\Big\rangle = m_1 \Big\langle A_1 \prod_{i=2}^k (G_i A_i)\Big\rangle - m_1 \Big\langle \underline{WG_1} A_1 \prod_{i=2}^k (G_i A_i)\Big\rangle. \tag{99}$$

Together with

$$\begin{aligned}\Big\langle \underline{W \prod_{i=1}^k (G_i A_i)}\Big\rangle &= \Big\langle W \prod_{i=1}^k (G_i A_i)\Big\rangle + \sum_{j=1}^k \widetilde{\mathbf{E}}\Big\langle \widetilde{W}\Big[\prod_{i=1}^{j-1} (G_i A_i)\Big] G_j \widetilde{W} \prod_{i=j}^k (G_i A_i)\Big\rangle \\ &= \Big\langle \underline{WG_1} A_1 \prod_{i=2}^k (G_i A_i)\Big\rangle + \sum_{j=2}^k \Big\langle \Big[\prod_{i=1}^{j-1} (G_i A_i)\Big] G_j\Big\rangle \Big\langle \prod_{i=j}^k (G_i A_i)\Big\rangle,\end{aligned}$$

we thus have

$$\begin{aligned}\Big(1 - \mathcal{O}\Big(\frac{N^\xi}{N\eta_*}\Big)\Big)\Big\langle \prod_{i=1}^k (G_i A_i)\Big\rangle &= m_1 \Big\langle A_1 \prod_{i=2}^k (G_i A_i)\Big\rangle - m_1 \Big\langle \underline{W \prod_{i=1}^k (G_i A_i)}\Big\rangle \\ &\quad + \sum_{j=2}^k \Big\langle \Big[\prod_{i=1}^{j-1} (G_i A_i)\Big] G_j\Big\rangle \Big\langle \prod_{i=j}^k (G_i A_i)\Big\rangle.\end{aligned} \tag{100}$$

We now apply the inequality [15], equation (5.35),

$$|\langle XY\rangle| \le [\langle X^*X(YY^*)^{1/2}\rangle\langle (Y^*Y)^{1/2}\rangle]^{1/2}$$

for arbitrary matrices X, Y to $X = \prod_{i=1}^{j-1}(G_iA_i)$, $Y = G_j$ to obtain

$$\left|\left\langle \prod_{i=1}^{j-1}(G_iA_i)G_j\right\rangle\right| \le \eta_1^{-1/2}\big(\langle A_{j-1}^*G_{j-1}^*\cdots A_1^*\Im G_1A_1\cdots G_{j-1}A_{j-1}|G_j|\rangle\langle |G_j|\rangle\big)^{1/2}$$

from $G^*G = (\Im G)/\eta$. By spectral decomposition we may further estimate with very high probability, for any $\xi > 0$,

$$\begin{aligned}&|\langle A_{j-1}^*G_{j-1}^*\cdots A_1^*\Im G_1A_1\cdots G_{j-1}A_{j-1}|G_j|\rangle|\\ &\quad= \left|\frac{1}{N}\sum_{\boldsymbol{a}}\frac{\langle \boldsymbol{u}_{a_j}, A_{j-1}^*\boldsymbol{u}_{a_{1-j}}\rangle\cdots\langle \boldsymbol{u}_{a_{-2}}, A_1^*\boldsymbol{u}_{a_1}\rangle\langle \boldsymbol{u}_{a_1}, A_1\boldsymbol{u}_{a_2}\rangle\cdots\langle \boldsymbol{u}_{a_{j-1}}, A_{j-1}\boldsymbol{u}_{a_j}\rangle}{[(\lambda_{a_{j-1}}-z_{j-1})(\lambda_{a_{1-j}}-\overline{z_{j-1}})\cdots(\lambda_{a_2}-z_2)(\lambda_{a_{-2}}-\overline{z_2})]|\lambda_{a_j}-z_j|}\Im\frac{1}{\lambda_{a_1}-z_1}\right|\\ &\quad\lesssim N^{\xi+j-2}\rho(z_1)\end{aligned}$$

from the overlap bound $|\langle \boldsymbol{u}_a, A\boldsymbol{u}_b\rangle| \lesssim N^{\xi-1/2}$, and where $\sum_{\boldsymbol{a}}$ is the summation over the $2j-2$ indices $a_1, a_{\pm 2}, \ldots, a_{\pm(j-1)}, a_j$ and conclude

$$\left|\left\langle \prod_{i=1}^{j-1}(G_iA_i)G_j\right\rangle\right| \lesssim N^{\xi+j/2-1}\frac{\sqrt{\rho_1}}{\sqrt{\eta_1}}. \tag{101}$$

Similarly, we also have

$$\left|\left\langle \prod_{i=j+1}^{k}(G_iA_i)\right\rangle\right| \lesssim N^{\xi+(k-j)/2-1},$$

and the claim follows from (100) and the bound

$$\left|\left\langle \underline{W\prod_{i=1}^{k}(G_iA_i)}\right\rangle\right| \lesssim N^{\xi}N^{(k-3)/2}\frac{\sqrt{\rho^*}}{\sqrt{\eta_*}} \tag{102}$$

on the underlined term in [15], Theorem 4.1, Remark 4.3. □

APPENDIX: GREEN FUNCTION COMPARISON

Here, we briefly recall the standard Green function comparison method for eigenvector statistics. The only novelty is that, in addition to the standard entrywise local law, $|G_{ab}(z)| \lesssim N^{\zeta+\xi}$ for $\Im z \sim N^{-1-\zeta}$, we also need an analogous a priori bound for $(GAG)_{ab}$ that exploits the fact that A is traceless; see (113) later. Consider the Ornstein–Uhlenbeck flow

$$\mathrm{d}\widehat{W}_t = -\frac{1}{2}\widehat{W}_t\,\mathrm{d}t + \frac{\mathrm{d}\widehat{B}_t}{\sqrt{N}}, \qquad \widehat{W}_0 = W \tag{103}$$

with $\widehat{B}_t$ a real symmetric Brownian motion. The OU-flow (103) has the effect of adding a small Gaussian component to W so that, for any fixed T, we can decompose

$$\widehat{W}_T \stackrel{\mathrm{d}}{=} \sqrt{1-cT}\,\widetilde{W} + \sqrt{cT}\,U \tag{104}$$

with $c = c(T) > 0$ a constant very close to one as long as $T \ll 1$, and U, $\widetilde{W}$ being independent GOE/Wigner matrices. Now, let W_t be the solution of the flow (12) with initial condition $W_0 = \sqrt{1-cT}\,\widetilde{W}$ so that

$$W_{cT} \stackrel{\mathrm{d}}{=} \widehat{W}_T. \tag{105}$$

LEMMA A.1. *Let $\widehat{W}_t$ be the solution of (103), and let $\widehat{\boldsymbol{u}}_i(t)$ be its eigenvectors. Then, for any smooth test function θ of, at most, polynomial growth and any fixed $\epsilon \in (0, 1/2)$, there exists an $\omega = \omega(\theta, \epsilon) > 0$ such that, for any $i \in [\delta N, (1-\delta)N]$ (with $\delta > 0$ from Theorem 2.2) and $t = N^{-1+\epsilon}$, it holds that*

$$\mathbf{E}\,\theta\Big(\sqrt{\frac{N}{2\langle A^2\rangle}}\langle \widehat{\boldsymbol{u}}_i(t), A\widehat{\boldsymbol{u}}_i(t)\rangle\Big) = \mathbf{E}\,\theta\Big(\sqrt{\frac{N}{2\langle A^2\rangle}}\langle \widehat{\boldsymbol{u}}_i(0), A\widehat{\boldsymbol{u}}_i(0)\rangle\Big) + \mathcal{O}(N^{-\omega}). \tag{106}$$

With $T = N^{-1+\epsilon}$ and $\theta(x) = x^n$ for some integer $n \in \mathbf{N}$, it now follows that

$$\begin{aligned}\mathbf{E}\Big[\sqrt{\frac{N}{2\langle A^2\rangle}}\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle\Big]^n &= \mathbf{E}\Big[\sqrt{\frac{N}{2\langle A^2\rangle}}\langle \widehat{\boldsymbol{u}}_i(T), A\widehat{\boldsymbol{u}}_i(T)\rangle\Big]^n + \mathcal{O}(N^{-c}) \\ &= \mathbf{E}\Big[\sqrt{\frac{N}{2\langle A^2\rangle}}\langle \boldsymbol{u}_i(cT), A\boldsymbol{u}_i(cT)\rangle\Big]^n + \mathcal{O}(N^{-c}) \\ &= \mathbf{1}(n \text{ even})(n-1)!! + \mathcal{O}(N^{-c})\end{aligned} \tag{107}$$

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, concluding the proof of Theorem 2.2. Note that in (107) we used Lemma A.1 in the first, (105) in the second, and (27) 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 A.1. The proof of Lemma A.1 follows from comparing expectations of products of resolvents $G(z)$ at scales slightly below the eigenvalue spacing, that is, for $\Im z \sim N^{-1-\zeta}$. Green function comparison for eigenvectors has been presented in [23] in details and has been used in [11, 12, 26]. Since this is a standard argument, we only give an outline. Let W_t be the solution of (103), with $W_0 = W$, where W is a Wigner matrix satisfying Assumption 2.1. Here, we dropped the hat compared to the notation used in (103) to make the presentation clearer; that is, we use W_t instead of $\widehat{W}_t$, $\boldsymbol{u}_i(t)$ instead of $\widehat{\boldsymbol{u}}_i(t)$, etc. From now on, by $G_t = G_t(z)$ we denote the resolvent of W_t. Note that along the flow (103) the first two moments of W are preserved.

Due to level repulsion, as in [22], Lemma 5.2, to understand $\sqrt{N}\langle \boldsymbol{u}_i, A\boldsymbol{u}_i\rangle$ it is sufficient to understand functions of $\sqrt{N}\langle \Im G(z)A\rangle$ with $\Im z$ slightly below N^{-1}, that is, the local eigenvalue spacing. In order to prove (106), it is enough to show that

$$\sup_{E\in(-2+\delta,2-\delta)} \big|\mathbf{E}\,\theta\big(\sqrt{N}\langle \Im G_t(z)A\rangle\big) - \mathbf{E}\,\theta\big(\sqrt{N}\langle \Im G_0(z)A\rangle\big)\big| \lesssim N^{-\omega} \tag{108}$$

for $z = E + \mathrm{i}\eta$ for some $\zeta > 0, \omega > 0$ and all $\eta \geq N^{-1-\zeta}$, cf. [3], Section 4, and [11], Appendix A. Define

$$R_t := \theta\big(\sqrt{N}\langle \Im G_t(z)A\rangle\big); \tag{109}$$

then, by Itô's formula we have

$$\mathbf{E}\,\frac{\mathrm{d}R_t}{\mathrm{d}t} = \mathbf{E}\Big[-\frac{1}{2}\sum_{\alpha} w_\alpha(t)\partial_\alpha R_t + \frac{1}{2}\sum_{\alpha,\beta}\kappa_t(\alpha,\beta)\partial_\alpha\partial_\beta R_t\Big], \tag{110}$$

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,

$$\kappa_t(\alpha_1, \ldots, \alpha_l) := \kappa\big(w_{\alpha_1}(t), \ldots, w_{\alpha_l}(t)\big) \tag{111}$$

denotes the joint cumulant of $w_{\alpha_1}(t), \dots, w_{\alpha_l}(t)$, with $l \in \mathbf{N}$. Note that by (10) it follows that $|\kappa_t(\alpha_1, \dots, \alpha_l)| \lesssim N^{-l/2}$ uniformly in $t \geq 0$. Performing a cumulant expansion in (110) (see [14], equation (25), for more details), we are left with

$$\mathbf{E}\frac{\mathrm{d}R_t}{\mathrm{d}t} = \sum_{l=3}^{R} \sum_{\alpha_1,\dots,\alpha_l} \kappa_t(\alpha_1, \dots, \alpha_l)\, \mathbf{E}[\partial_{\alpha_1} \cdots \partial_{\alpha_l} R_t] + \Omega(R), \tag{112}$$

where $\Omega(R)$ is an error term, easily seen to be negligible as every additional derivative gains a further factor of $N^{-1/2}$. In order to estimate (112), we use $|(G_t)_{ab}| \leq N^{\zeta}$ and

$$\begin{aligned} \big|\big(G_t(z_1)AG_t(z_2)\big)_{ab}\big| &= \Big|\sum_{ij} \frac{\boldsymbol{u}_i(a)\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle \boldsymbol{u}_j(b)}{(\lambda_i - z_1)(\lambda_j - z_2)}\Big| \\ &\lesssim N^{1/2+\xi}\Big(\frac{1}{N}\sum_i \frac{1}{|\lambda_i - z_1|}\Big)\Big(\frac{1}{N}\sum_i \frac{1}{|\lambda_i - z_2|}\Big) \lesssim N^{1/2+\xi+2\zeta} \end{aligned} \tag{113}$$

which holds with very high probability for $z_1, z_2 \in \{z, \overline{z}\}$. In (113) we used the eigenvector delocalisation $\|\boldsymbol{u}_i\|_\infty \lesssim N^{-1/2+\xi}$ (see [21], or [5] for the optimal bound) and the optimal a priori bound $|\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle| \lesssim N^{-1/2+\xi}$ for traceless A by [15], Theorem 2.2 (note that this step crucially uses that $\langle A\rangle = 0$, the analogous bound for a general A would be larger by a factor $\sqrt{N}$). We claim that, for any $l \geq 0$, it holds that

$$\big|\partial_{\alpha_1} \dots \partial_{\alpha_l} \sqrt{N}\langle \Im G_t A\rangle\big| \leq N^{(l+3)(\zeta+\xi)}, \tag{114}$$

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

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

we are then able to estimate $|\mathbf{E}\,\mathrm{d}R_t/\mathrm{d}t|$ by the chain rule to finally obtain (108).

The bound (114) for $l = 0$ follows immediately from the a priori bound $\sqrt{N}|\langle G_t A\rangle| \lesssim N^{\xi+\zeta}$. For $l = 1$, the first derivative yields $|\partial_{ab}\sqrt{N}\langle G_t A\rangle| = N^{-1/2}|(G_t A G_t)_{ba}|$, and each additional derivative creates a single factor of G_t, and (114) follows from estimating each factor entrywise by $|(G_t)_{ab}| \lesssim N^{\xi+\zeta}$ and (113). □

Acknowledgments. L.E. would like to thank Zhigang Bao for many illuminating discussions in an early stage of this research. The authors are also grateful to Paul Bourgade for his comments on the manuscript and the anonymous referee for several useful suggestions.

REFERENCES

[1] AGGARWAL, A., LOPATTO, P. and MARCINEK, J. (2021). Eigenvector statistics of Lévy matrices. *Ann. Probab.* **49** 1778–1846. MR4260468 https://doi.org/10.1214/20-aop1493

[2] ANANTHARAMAN, N. and SABRI, M. (2019). Quantum ergodicity on graphs: From spectral to spatial delocalization. *Ann. of Math.* (2) **189** 753–835. MR3961083 https://doi.org/10.4007/annals.2019.189.3.3

[3] BENIGNI, L. (2020). Eigenvectors distribution and quantum unique ergodicity for deformed Wigner matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **56** 2822–2867. MR4164858 https://doi.org/10.1214/20-AIHP1060

[4] BENIGNI, L. (2021). Fermionic eigenvector moment flow. *Probab. Theory Related Fields* **179** 733–775. MR4242625 https://doi.org/10.1007/s00440-020-01018-0

[5] BENIGNI, L. and LOPATTO, P. (2020). Optimal delocalization for generalized Wigner matrices. Preprint. Available at arXiv:2007.09585.

[6] BENIGNI, L. and LOPATTO, P. (2021). Fluctuations in local quantum unique ergodicity for generalized Wigner matrices. Preprint. Available at arXiv:2103.12013.

[7] BLOEMENDAL, A., ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2014). Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.* **19** no. 33, 53 pp. MR3183577 https://doi.org/10.1214/ejp.v19-3054

[8] BOURGADE, P. (2021). Extreme gaps between eigenvalues of Wigner matrices. *J. Eur. Math. Soc.* **23**. https://doi.org/10.4171/JEMS/1141

[9] BOURGADE, P., ERDŐS, L., YAU, H.-T. and YIN, J. (2016). Fixed energy universality for generalized Wigner matrices. *Comm. Pure Appl. Math.* **69** 1815–1881. MR3541852 https://doi.org/10.1002/cpa.21624

[10] BOURGADE, P., HUANG, J. and YAU, H.-T. (2017). Eigenvector statistics of sparse random matrices. *Electron. J. Probab.* **22** Paper No. 64, 38 pp. MR3690289 https://doi.org/10.1214/17-EJP81

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

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

[13] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2020). Functional central limit theorems for Wigner matrices. Preprint. Available at arXiv:2012.13218.

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

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

[16] COLIN DE VERDIÈRE, Y. (1985). Ergodicité et fonctions propres du laplacien. *Comm. Math. Phys.* **102** 497–502. MR0818831

[17] D'ALESSIO, L., KAFRI, Y., POLKOVNIKOV, A. and RIGOL, M. (2016). From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics. *Adv. Phys.* **65** 239–362. Available at arXiv:1509.06411. https://doi.org/10.1080/00018732.2016.1198134

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

[19] ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2013). The local semicircle law for a general class of random matrices. *Electron. J. Probab.* **18** no. 59, 58 pp. MR3068390 https://doi.org/10.1214/EJP.v18-2473

[20] ERDŐS, L. and YAU, H.-T. (2015). Gap universality of generalized Wigner and β-ensembles. *J. Eur. Math. Soc.* (*JEMS*) **17** 1927–2036. MR3372074 https://doi.org/10.4171/JEMS/548

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

[22] KNOWLES, A. and YIN, J. (2013). Eigenvector distribution of Wigner matrices. *Probab. Theory Related Fields* **155** 543–582. MR3034787 https://doi.org/10.1007/s00440-011-0407-y

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

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

[25] LIEB, E. H. and LOSS, M. (2001). *Analysis*, 2nd ed. *Graduate Studies in Mathematics* **14**. Amer. Math. Soc., Providence, RI. MR1817225 https://doi.org/10.1090/gsm/014

[26] MARCINEK, J. and YAU, H.-T. (2020). High dimensional normality of noisy eigenvectors. Preprint. Available at arXiv:2005.08425.

[27] MARKLOF, J. and RUDNICK, Z. (2000). Quantum unique ergodicity for parabolic maps. *Geom. Funct. Anal.* **10** 1554–1578. MR1810753 https://doi.org/10.1007/PL00001661

[28] O'ROURKE, S., VU, V. and WANG, K. (2016). Eigenvectors of random matrices: A survey. *J. Combin. Theory Ser. A* **144** 361–442. MR3534074 https://doi.org/10.1016/j.jcta.2016.06.008

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

[30] ŠNIREL'MAN, A. I. (1974). Ergodic properties of eigenfunctions. *Uspekhi Mat. Nauk* **29** 181–182. MR0402834

[31] SREDNICKI, M. (1994). Chaos and quantum thermalization. *Phys. Rev. E* **50** 888–901. https://doi.org/10.1103/PhysRevE.50.888

[32] TAO, T. and VU, V. (2012). Random matrices: Universal properties of eigenvectors. *Random Matrices Theory Appl.* **1** 1150001, 27 pp. MR2930379 https://doi.org/10.1142/S2010326311500018

[33] ZELDITCH, S. (1987). Uniform distribution of eigenfunctions on compact hyperbolic surfaces. *Duke Math. J.* **55** 919–941. MR0916129 https://doi.org/10.1215/S0012-7094-87-05546-3