Random Band Matrices in the Delocalized Phase I: Quantum Unique Ergodicity and Universality

PAUL BOURGADE
Courant Institute

HORNG-TZER YAU
Harvard University

AND

JUN YIN
University of California, Los Angeles

Contents

1 Introduction

1.1 Random Matrices Beyond Mean Field

In Wigner's vision, random matrices play the role of a mean-field model for large quantum systems of high complexity. His paradigm has been confirmed with significant progress in understanding the universal behavior of many random graph and random matrix models. However, regarding his core thesis that random matrices can be used to model non-mean-field systems, our understanding is much more limited. Even for one of the simplest non mean-field models, the random Schrödinger operator, there is no result concerning the existence of the delocalized regime in which random matrix statistics are expected to hold.

A slightly more tractable model is the *random band matrix* characterized by the property that H_{ij} becomes negligible if $\mathrm{dist}(i, j)$ exceeds a parameter W, called the *bandwidth*. In general, i, j are lattice points in $\mathbb{Z}^d$, but in this article we consider only the case $d = 1$. Based on numerics, it was conjectured [9, 10] that

Communications on Pure and Applied Mathematics, Vol. LXXIII, 1526–1596 (2020)
© 2020 Wiley Periodicals, Inc.

the eigenvectors of band matrices satisfy a localization-delocalization transition, in the bulk of the spectrum, with a corresponding sharp transition for the eigenvalue distribution [18]:

(i) for $W \gg \sqrt{N}$, delocalization and Gaussian orthogonal ensemble (GOE) spectral statistics hold;

(ii) for $W \ll \sqrt{N}$, eigenstates are localized and the eigenvalues converge to a Poisson point process.

This transition was also supported by heuristic arguments [36] and a nonrigorous supersymmetry method [19].

There have been many partial results concerning localization-delocalization for band matrices. For general distribution of the matrix entries, localization of eigenvectors was first shown for $W \ll N^{1/8}$ [28], and improved to $W \ll N^{1/7}$ for Gaussian entries [26]. Delocalization was proved in some averaged sense, for $W \gg N^{6/7}$ in [14], $W \gg N^{4/5}$ in [16], $W \gg N^{7/9}$ in [20]. The Green's function was controlled down to the scale $\operatorname{Im} z \gg W^{-1}$ in [17], implying a lower bound of order W for the localization length of all eigenvectors. We also mention that at the edge of the spectrum, the transition for one-dimensional band matrices (with critical exponent $N^{5/6}$) was understood in [33], thanks to the method of moments.

When the entries of band matrices are Gaussian with some specific covariance profile, one can apply supersymmetry techniques (see [13,34] for overviews). With this method, for $d = 3$, precise estimates on the density of states [12] were first obtained. Then, random matrix local spectral statistics were proved for $W = \Omega(N)$ [30], and delocalization was obtained for all eigenvectors when $W \gg N^{6/7}$ and the first four moments of the matrix entries match the Gaussian ones [2] (these results assume complex entries and hold in a part of the bulk). Still with the supersymmetry technique, a transition around $N^{1/2}$ was proved in [29, 31], concerning moments of characteristics polynomials.

1.2 Mean Field Reduction and Quantum Unique Ergodicity

The main difficulties in analyzing spectral properties of band matrices with general entries are twofold.

(i) There is currently no effective diagrammatical method to estimate the Green's function when $\operatorname{Im} z \ll W^{-1}$, while delocalization of eigenvectors requires estimates up to $\operatorname{Im} z \gg N^{-1}$.

(ii) For the universality of local spectral statistics, the comparison method used for mean-field models does not apply to band matrices since the majority of matrix elements (effectively) vanish.

In an earlier paper [4], we proposed a *mean-field reduction* method to prove universality of local spectral statistics for band matrices with $W = \Omega(N)$. This method relies on a notion much stronger than delocalization, the probabilistic quantum unique ergodicity (QUE). Historically, QUE was introduced by Rudnick and

Sarnak [27], asserting that for negatively curved compact Riemannian manifolds, *all* high energy Laplacian eigenfunctions become completely flat. Quantum ergodicity, essentially an averaged version of QUE, had previously been proved for more general manifolds [11, 32, 38]. For d-regular graphs, the eigenvectors of the discrete Laplacian also satisfy quantum ergodicity, under certain assumptions on the injectivity radius and spectral gap of the adjacency matrices [1].

A probabilistic version of QUE was proposed and proved for Wigner matrices in [7]. To state it, let H be a size N random matrix with eigenvectors $\boldsymbol{\psi}_j$ associated to eigenvalues λ_j. Then, there exists $\varepsilon > 0$ such that for any deterministic $1 \leq j \leq N$ and $I \subset [\![1, N]\!]$ and for any $\delta > 0$ we have

$$\mathbb{P}\left(\Big| \sum_{i \in I} |\psi_j(i)|^2 - \frac{|I|}{N} \Big| \geq \delta \right) \leq N^{-\varepsilon}/\delta^2. \tag{1.1}$$

To explain the mean-field reduction, we block-decompose a band matrix H and its eigenvectors:

$$H = \begin{pmatrix} A & B^* \\ B & D \end{pmatrix}, \quad \boldsymbol{\psi}_j = \begin{pmatrix} \mathbf{w}_j \\ \mathbf{p}_j \end{pmatrix}, \tag{1.2}$$

where A is a $W \times W$ Wigner matrix. From the eigenvector equation $H\boldsymbol{\psi}_j = \lambda_j \boldsymbol{\psi}_j$,

$$Q_{\lambda_j} \mathbf{w}_j = \lambda_j \mathbf{w}_j \quad \text{where } Q_e = A - B^* \frac{1}{D - e} B. \tag{1.3}$$

Thus $\mathbf{w}_j$ is an eigenvector to Q_e with eigenvalue λ_j when $e = \lambda_j$. The basic observation from the earlier paper [4] can be summarized as follows. Suppose that the probabilistic QUE for the eigenvectors of H holds. Then the eigenvalues of H near a fixed energy E can be reconstructed from the eigenvalues of Q_e near the origin with e near E. Thus if we can prove the spectral universality for Q_e, the same statement holds for H. On the other hand, to establish QUE for the band matrix H, assume first that it holds for the $W \times W$ operator Q_e. If we can substitute e by λ_j, then the eigenvector $\boldsymbol{\psi}_j$ is flat in the first W coordinates. Clearly, we can stitch together the flatnesses of $\boldsymbol{\psi}_j$ in sufficiently many windows of size W to establish the global flatness of $\boldsymbol{\psi}_j$ provided that the error in each window is sufficiently small.

To summarize, the mean-field reduction method reduces the universality and QUE for the band matrix H to those of Q_e. Thanks to the recent progress on these topics [5, 22, 23], the inputs to prove these properties require precise estimates on the Green's function $(Q_e - z)^{-1}$ only for $\operatorname{Im} z \sim N^{-\varepsilon}$. For probabilistic QUE, we also need to establish the error probability in the sense of "very high probability." In the following, we start with a discussion on the Green's function $(Q_e - z)^{-1}$.

1.3 Generalized Green's Functions

It is clear that, if we estimate the Green's function $(Q_e - z)^{-1}$ directly, some bound on the matrix $(D - e)^{-1}$ appearing in Q_e will be needed. Since e is real, estimating $(D - e)^{-1}$ is clearly a much harder problem than estimating the original

Green's function $(H - z')^{-1}$. Fortunately, we only need this estimate with $\operatorname{Im} z \sim N^{-\varepsilon}$. Clearly, one can interpret $(Q_e - z)^{-1}$ as the $W \times W$ corner of the *generalized Green's function*

$$G(z, w) = \left(H - \begin{pmatrix} z\,\mathrm{I}_W & 0 \\ 0 & w\,\mathrm{I}_{N-W} \end{pmatrix} \right)^{-1} \tag{1.4}$$

when $w = e$. In [4], we use a somehow involved induction argument and an uncertainty principle to estimate $G(z, e)$ for $W = \Omega(N)$. In this work, we provide accurate estimates, Theorem 4.5, on $G(z, e)$ for $\operatorname{Im} z \sim N^{-\varepsilon}$ when $W \gg N^{3/4}$. Our method is to derive a self-consistent equation for the (off-diagonal) entries of the generalized Green's function (a similar equation for the standard Green's function was called the T equation [15]). Notice that Ward's identity, which is instrumental in many random matrix estimations, is not valid for generalized Green's functions. More precisely, Ward's identity asserts that for any Green's function of a Hermitian operator H,

$$\sum_j |G_{ij}(z)|^2 \leq (\operatorname{Im} z)^{-1} \operatorname{Im} G_{ii}. \tag{1.5}$$

For the generalized Green's function $G(z, w)$, the last property fails. Our strategy is to establish an estimate on $\sum_j |G_{ij}(z)|^2$ by successively decreasing the imaginary part of w and using repeatedly the self-consistent T equation in each step. Besides overcoming this difficulty, we also devise a new diagrammatic expansion in deriving the T equation. Finally, we remark that the main condition $W \gg N^{3/4}$ is mainly used in estimating $G(z, e)$. Besides extending the region of validity from $W = \Omega(N)$ to $W \gg N^{3/4}$, our current approach allows the estimate on $G(z, e)$ to be completely independent from all other arguments in this work (e.g., the mean-field reduction). The proof of Theorem 4.5 will be delayed to parts 2 and 3 of this series.

1.4 Probabilistic QUE with High Probability

The proof of the quantum unique ergodicity (1.1) for Q_e in [4] relies on two different tools.

(i) A priori estimates on the Green's function $(Q_e - z)^{-1}$ (for large Im z) provide flatness of eigenvectors on average. This a priori information is necessary to obtain the following.

(ii) The eigenvector moment flow from [7] is a random walk in a dynamic random environment whose relaxation means flatness of individual eigenvectors (quantum unique ergodicity).

We have just outlined our new estimates on the Green's function $(Q_e - z)^{-1}$ for $W \gg N^{3/4}$. The main new technique developed in this work concerns (ii): Theorem 2.5 states that

Averaged quantum unique ergodicity implies a high probability quantum unique ergodicity, after adding a small GOE component.

Compared to (1.1), this new result is a strong probabilistic QUE, as it first allows much more general observables of eigenvectors and is valid with probability $1 - N^{-D}$ for any $D > 0$. Therefore all bulk eigenvectors are now simultaneously flat. The proof of Theorem 2.5 relies on a remarkable combinatorial identity: the perfect matching observables defined in (2.15) satisfy the eigenvector moment flow parabolic equation; see Theorem 2.6.

Thanks to this new strong version of QUE, the eigenvectors of Q_e are flat for all e in a discrete subset of size N^C for any C fixed. Thus to establish flatness of $\boldsymbol{\psi}_j$ on the first W coordinates, we only need to compare eigenvectors of Q_e and Q_{λ_j} for $|e - \lambda_j| \leq N^{-C}$ with C a large constant. An eigenvector perturbation formula is enough to compute the difference between these eigenvectors, with sufficient a priori estimates given by a weak uncertainty principle as developed in [4].

Therefore, our work presents an improvement from $W = \Omega(N)$ [4] to $W \gg N^{3/4}$ thanks to new results both on (i) and (ii). As discussed in Remark 4.7, our hypothesis $W \gg N^{3/4}$ for delocalization comes from the generalized Green's function estimates (ii). Heuristics for the transition at bandwidth $N^{1/2}$ are given in the same remark.

1.5 The Model and Results

All results in this paper apply to both real and complex band matrices. For the definiteness of notation, we consider only the real symmetric case, and we use the convention that all eigenvectors are real. In the following definition, $\mathbb{Z}_N$ denotes the set of residues mod N so that our matrices are assumed to have periodic boundary condition.

DEFINITION 1.1 (Band matrix H_N with bandwidth W_N). Let H_N be a $N \times N$ matrix with real centered entries $(H_{ij}, i, j \in \mathbb{Z}_N)$ which are independent up to the condition $H_{ij} = H_{ji}$. We say that H_N is band matrix with bandwidth $W = W_N$ if

$$s_{ij} := \mathbb{E}|H_{ij}|^2 = f(i-j) \tag{1.6}$$

for some $f : \mathbb{Z}_N \to \mathbb{R}$ satisfying $\sum_{x \in \mathbb{Z}_N} f(x) = 1$, and there exist a small positive constant c_s and a large constant C_s such that

$$c_s\, W^{-1} \cdot 1_{|x| \leq W} \leq f(x) \leq C_s\, W^{-1} \cdot 1_{|x| \leq C_s W}, \quad x \in \mathbb{Z}_N, \tag{1.7}$$

where $|\cdot|$ is the periodic distance on $\mathbb{Z}_N$.

The method in this paper also allows us to treat cases with progressive decay of the variance away from the diagonal (e.g., $f(x) \leq C_s W^{-1} \cdot 1_{|x| \leq C_s W}$ instead of $f(x) \leq C_s W^{-1} \cdot 1_{|x| \leq W}$), or variants with exponentially small mass away from the bandwidth. We work under the hypothesis (1.7) for simplicity.

For technical reasons we assume the following condition on the fourth moment of the matrix entries: there is $\varepsilon_m > 0$ (here the subscript m indicates the moment condition) such that for $|i-j| \le W$,

$$\min_{|i-j|\le W}\bigl(\mathbb{E}\xi_{ij}^4 - \bigl(\mathbb{E}\xi_{ij}^3\bigr)^2 - 1\bigr) \ge N^{-\varepsilon_m}, \tag{1.8}$$

where $\xi_{ij} := H_{ij}(s_{ij})^{-1/2}$ is the normalized random variable with mean 0 and variance 1. It is well-known that for any real random variable ξ with mean 0 and variance 1, $\mathbb{E}\xi^4 - (\mathbb{E}\xi^3)^2 - 1 \ge 0$, and the equality holds if and only if ξ is a Bernoulli random variable (lemma 28 of [35]). Therefore, one simply has $\varepsilon_m = 0$ when the ξ_{ij}'s ($|i-j| \le W$) all have the same law, different from the Bernoulli distribution. In the more general setting (1.8), all our results are restricted to $0 \le \varepsilon_m < 1/2$ because of the following condition (1.11).

We also assume that for some $\delta_d > 0$ (subscript d stands for "decay") we have

$$\sup_{N,i,j} \mathbb{E}\bigl(e^{\delta_d W H_{ij}^2}\bigr) < \infty. \tag{1.9}$$

This tail condition can be weakened to a finite high moment condition. We assume (1.9) mainly for the convenience of presentation. The constants in the following theorems depend on the fixed parameters c_s, C_s, ε_m, and δ_d, in (1.7), (1.8), and (1.9), but we will only keep track of the dependence on ε_m.

Denote the eigenvalues of H by $\lambda_1 \le \cdots \le \lambda_N$, and let $(\boldsymbol{\psi}_k)_{k=1}^N$ be the corresponding L^2-normalized eigenvector, i.e., $H\boldsymbol{\psi}_k = \lambda_k \boldsymbol{\psi}_k$. Thanks to the condition $\sum f(x) = 1$, it is known that the empirical spectral measure $\frac{1}{N}\sum_{k=1}^N \delta_{\lambda_k}$ converges almost surely to the Wigner semicircle law with density

$$\rho_{\mathrm{sc}}(x) = \frac{1}{2\pi}\sqrt{(4-x^2)_+}\,.$$

The concept of localization/delocalization can be defined in many ways. For definiteness, we use the L^∞ norm. For any small constant $c > 0$ and $\tau > 0$, one expects that

$$\begin{aligned}&\mathbb{P}\bigl(N^{-\tau} \le \min(N, W^2)\|\boldsymbol{\psi}_k\|_\infty^2 \le N^\tau \text{ for all } k \in [\![cN, (1-c)N]\!]\bigr)\\ &\quad = 1 - \mathrm{o}(1),\end{aligned} \tag{1.10}$$

meaning that a localization-delocalization transition occurs at $\log_N W = 1/2$, where $\log_N W = \log W / \log N$. Our first result proves (1.10) in the delocalization regime $\log_N W > 3/4$.

THEOREM 1.2 (Delocalization for $\log_N W > 3/4$). *Let $(H_N)_{N\ge 1}$ be band matrices with bandwidth W_N satisfying the conditions* (1.8) *and* (1.9). *Recall that $\varepsilon_m > 0$ is defined in* (1.8). *Suppose that for some constant $a > 0$,*

$$\log_N W \ge \max\biggl(\frac{3}{4}, \frac{1}{2} + \varepsilon_m\biggr) + a. \tag{1.11}$$

For any (small) constants $\kappa, \tau > 0$ and (large) $D > 0$, there exists N_0 such that for all $N \geq N_0$ we have

$$\mathbb{P}\big(\|\boldsymbol{\psi}_k\|_\infty^2 \leq N^{-1+\tau} \text{ for all } k \in [\![\kappa N, (1-\kappa)N]\!]\big) \geq 1 - N^{-D}. \tag{1.12}$$

The above delocalization holds together with a local semicircle law down to the optimal scale.

THEOREM 1.3 (Local semicircle law for $\log_N W > 3/4$). *Under the same assumptions as Theorem* 1.2, *there exists $\varepsilon > 0$ such that for any (small) $\kappa, \tau > 0$ and (large) $D > 0$ there exists N_0 such that for any $E_1, E_2 \in [-2+\kappa, 2-\kappa]$ and any $N \geq N_0$ we have*

$$\begin{aligned}\mathbb{P}\bigg(\bigg|\#\{\lambda_k \in [E_1, E_2]\} - N\int_{E_1}^{E_2} \mathrm{d}\rho_{\mathrm{sc}}\bigg| < N^\tau + |E_1 - E_2|N^{1-\varepsilon}\bigg)& \\ \geq 1 - N^{-D}.&\end{aligned} \tag{1.13}$$

In the following fixed energy universality statement, we denote by $\rho_H^{(k)}$ the k-point correlation function (understood in the sense of distributions) for the spectral measure of an $N \times N$ random matrix H.

THEOREM 1.4 (Universality for $\log_N W > 3/4$). *Under the same assumptions as Theorem* 1.2, *for any $\kappa > 0$, any integer k, and any smooth test function $O \in \mathscr{C}^\infty(\mathbb{R}^k)$ with compact support, there are constants $c, C > 0$ such that for any $|E| \leq 2 - \kappa$ we have*

$$\begin{aligned}\bigg|\int_{\mathbb{R}^k} O(\mathbf{a})\rho_H^{(k)}\bigg(E + \frac{\mathbf{a}}{N\rho_{\mathrm{sc}}(E)}\bigg)\mathrm{d}\mathbf{a} - \int_{\mathbb{R}^k} O(\mathbf{a})\rho_{\mathrm{GOE}}^{(k)}\bigg(E + \frac{\mathbf{a}}{N\rho_{\mathrm{sc}}(E)}\bigg)\mathrm{d}\mathbf{a}\bigg|& \\ \leq CN^{-c}.&\end{aligned} \tag{1.14}$$

For the proof of Theorems 1.2, 1.3 and 1.4, the first step is to show that delocalization, the local semicircle law, eigenvalue universality, and quantum unique ergodicity hold under the following additional assumption: H is a Gaussian divisible band matrix; i.e., there exists independent band matrices H_1 and H_2 with the same width W and $c > 0$ such that H_1 satisfies (1.8) and (1.9), and

$$\begin{aligned}&H = H_1 + H_2 \\ &\qquad \text{where } (H_2)_{ij} = 1_{|i-j| \leq W} \cdot (1 + 1_{ij})^{1/2} \cdot \mathscr{N}(0,\ c\ W^{-1}N^{-\varepsilon_m}).\end{aligned} \tag{1.15}$$

Remember that ε_m is defined in (1.8). Here, c is a small enough constant depending only on δ_d from (1.9).

THEOREM 1.5. *Assume that H is a band matrix of type* (1.15), *with bandwidth W_N satisfying* (1.11).

(i) *The eigenvectors are delocalized as in* (1.12).
(ii) *The eigenvalues satisfy the local semicircle law as in* (1.13).
(iii) *Fixed energy universality holds as in* (1.14).

(iv) *For any (small) $\tau, \kappa > 0$, and (large) $D > 0$, there exists $N_0 > 0$ such that for any $N \geq N_0$ we have*

$$\mathbb{P}\left(\left|\frac{N}{W}\sum_{\alpha=\ell}^{\ell+W}|\psi_j(\alpha)|^2 - 1\right| < N^{-\frac{3}{2}a+\tau} \text{ for all } 1 \leq j, \ell \leq N \text{ such that } |\lambda_j| \leq 2-\kappa\big) \geq 1 - N^{-D}\right),$$

where $a > 0$ was given in (1.11) *and all indices are defined modulo N.*

1.6 Organization of the Paper

This work is essentially divided in two parts.

The first part (Sections 2 and 3) concerns quantum unique ergodicity for mean-field blocks, and improves on the estimate (1.1): Theorem 2.5 gives flatness of the eigenvectors with overwhelming probability (*overwhelming probability* refers to the arbitrary large choice of $D > 0$, for example in Theorem 2.5), and with optimal fluctuations scale for the L^2 mass of eigenvectors on subsets of $[\![1, N]\!]$. This result is the main technical novelty of our work.

The first aspect of the proof is algebraic (Section 2). A new function of the eigenvector overlaps is defined in equation (2.15), and it follows the eigenvector moment flow dynamics; see Theorem 2.6. These dynamics of *perfect matching observables* generalize an earlier observation from [7]. In this previous work, the eigenvector evolution was related to a random walk in a dynamic random environment, after dimension reduction through projection on a given fixed direction. Projections can now occur on an arbitrary number of directions; see Remark 2.8. The proof of Theorem 2.6 is combinatorial and given at the end of Section 2.

The second aspect of the proof of Theorem 2.5 is analytic (Section 3). As proved by a sequence of maximum principles and approximations with short range dynamics, the eigenvector moment flow reaches equilibrium after some time depending on the initial condition. This allows us to identify the scale of the perfect matching observables. Our proof is more involved than the Hölder regularity of the eigenvector moment flow in [7], because our observables are more general: in [7], the scale of observables was a priori known and the dynamics were used to identify the distribution of fluctuations.

The second part of the paper (Sections 4 and 5) applies the strong form of quantum unique ergodicity to delocalization for random band matrices. First, Theorem 1.5 is proved by the mean-field reduction technique from [4], then it is extended to more general band matrices by a moment matching argument.

The proof of Theorem 1.5 (Section 4) is sketched in Section 4.2. Section 4.4 contains the first important input for the proof: the resolvent estimates for $(Q_e - z)^{-1}$. As explained after (1.4), these estimates from Theorem 4.5 amount to an averaged form of QUE for the eigenvectors of Q_e. From this a priori estimate, quantum unique ergodicity is deduced for the Gaussian divisible version of Q_e (Section 4.5). To access flatness of eigenvectors of our original eigenvectors $\boldsymbol{\psi}_k$, we need to patch QUE estimates for eigenvectors of Q_e when $e = \lambda_k$. By a net argument in e, with mesh size N^{-C} (C is fixed and arbitrarily large because Theorem 2.5 holds with overwhelming probability), we only need to control eigenvector shifts under tiny perturbations in e. This is the role of another input for the proof of Theorem 1.5, the weak uncertainty principle. It is inspired by a more difficult result from [4], and proved in Section 4.6. We refer to (4.58) for eigenvectors bounds thanks to the weak uncertainty principle. Section 4.7 concludes the proof of Theorem 1.5.

In Section 5, delocalization, the local semicircle law, and universality (Theorems 1.2, 1.3, and 1.4) are obtained beyond the Gaussian divisible ensemble. The proof relies on moment matching, exhibiting a matrix $\widetilde{H}$ of type (1.15) whose first four moments of the entries match those of H. This idea appeared in [35] for the purpose of universality for Wigner matrices and required some a priori information on delocalization and local semicircle law. In our work, such information is only available for $\widetilde{H}$, by Theorem 1.5. It is extended to H thanks to an implementation of the moment matching strategy at the level of the Green's functions [17], and a self-consistent method to obtain these estimates by continuously interpolating from $\widetilde{H}$ to H [21].

Finally, although this work focuses on symmetric matrices, the method applies to the Hermitian class. The only substantial difference is the algebraic part of QUE for mean-field models: the perfect matching observables are defined in a different way for real and complex matrices, as explained in the Appendix.

2 Quantum Unique Ergodicity for Deformed Matrices

This and the next section are self-sufficient. In these sections, the size of the matrices is denoted by n. The main result (Theorem 2.5) will then be applied to mean-field blocks of type Q_e from (1.3) (or more precisely its generalization Q_e^g; see (4.10)), i.e., for $n = W$.

2.1 Eigenvector Dynamics

In this subsection, we first recall the stochastic differential equation for the eigenvectors under the Dyson Brownian motion, as stated in [7, sec. 2].

The matrix Brownian motion dynamics are defined as follows, either at the matrix, eigenvalues, or eigenvectors level (remember we only consider the symmetric case, the Hermitian one being detailed in the Appendix). Let B be an $n \times n$ matrix such that B_{ij} $(i < j)$ and $B_{ii}/\sqrt{2}$ are independent standard Brownian motions, and $B_{ij} = B_{ji}$. We abbreviate $Z(t) = B(t)/\sqrt{n}$. The $n \times n$ symmetric Dyson Brownian motion K with initial value $K(0) = V$ is defined as

$$K(t) = V + Z(t). \tag{2.1}$$

Let $\boldsymbol{\lambda}_0 \in \Sigma_n = \{\lambda_1 < \cdots < \lambda_n\}$, $u_0 \in \mathrm{O}(n)$. The symmetric Dyson Brownian motion/vector flow with initial condition $\boldsymbol{\lambda}_0 = (\lambda_1, \dots, \lambda_n)$, $u_0 = (u_1, \dots, u_n)$, is defined through the dynamics

$$\mathrm{d}\lambda_k = \frac{\mathrm{d}B_{kk}}{\sqrt{n}} + \left(\frac{1}{n} \sum_{\ell \neq k} \frac{1}{\lambda_k - \lambda_\ell}\right) \mathrm{d}t, \tag{2.2}$$

$$\mathrm{d}u_k = \frac{1}{\sqrt{n}} \sum_{\ell \neq k} \frac{\mathrm{d}B_{k\ell}}{\lambda_k - \lambda_\ell} u_\ell - \frac{1}{2n} \sum_{\ell \neq k} \frac{\mathrm{d}t}{(\lambda_k - \lambda_\ell)^2} u_k. \tag{2.3}$$

With a slight abuse of notation, we will write $\boldsymbol{\lambda}_t$ either for $(\lambda_1(t), \dots, \lambda_n(t))$ or for the $n \times n$ diagonal matrix with entries $\lambda_1(t), \dots, \lambda_n(t)$.

The link between the previously defined matrix and spectral dynamics is given as follows. See [7] for a proof, with the main ideas being due to McKean [24] for the existence and uniqueness of solutions, and Bru [8] for the eigenvector dynamics in the Wishart case.

THEOREM 2.1. *The following statements about the Dyson Brownian motion and eigenvalue/vector flow hold.*

(a) *Existence and strong uniqueness hold for the system of stochastic differential equations* (2.2)–(2.3). *Let $(\boldsymbol{\lambda}_t, u_t)_{t \geq 0}$ be the solution. Almost surely, for any $t \geq 0$ we have $\boldsymbol{\lambda}_t \in \Sigma_n$ and $u_t \in \mathrm{O}(n)$.*

(b) *Let $(K(t))_{t \geq 0}$ be a symmetric Dyson Brownian motion with initial condition $K(0) = u_0 \boldsymbol{\lambda}_0 u_0^*$, $\boldsymbol{\lambda}_0 \in \Sigma_n$. Then the processes $(K(t))_{t \geq 0}$ and $(u_t \boldsymbol{\lambda}_t u_t^*)_{t \geq 0}$ have the same distribution.*

(c) *Existence and strong uniqueness hold for* (2.2). *For any $T > 0$, let $\nu_T^{K(0)}$ be the distribution of $(\boldsymbol{\lambda}_t)_{0 \leq t \leq T}$ with initial value the spectrum of a matrix $K(0)$. For $0 \leq T \leq T_0$ and any given continuous trajectory $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_t)_{0 \leq t \leq T_0} \subset \Sigma_n$, existence and strong uniqueness holds for* (2.3) *on* $[0, T]$. *Let $\mu_T^{K(0), \boldsymbol{\lambda}}$ be the distribution of $(u_t)_{0 \leq t \leq T}$ with the initial matrix $K(0)$ and the path $\boldsymbol{\lambda}$ given.*

Let F be continuous bounded, from the set of continuous paths (*on* $[0, T]$) *on $n \times n$ symmetric matrices to $\mathbb{R}$. Then for any initial matrix $K(0)$ we have*

$$\mathbb{E}^{K(0)}\big(F((K(t))_{0 \leq t \leq T})\big) = \int \mathrm{d}\nu_T^{K(0)}(\boldsymbol{\lambda}) \int \mathrm{d}\mu_T^{K(0), \boldsymbol{\lambda}}(u) F((u_t \boldsymbol{\lambda}_t u_t^*)_{0 \leq t \leq T}).$$

Following [7], we introduce the notations (the dependence in t will often be omitted for $c_{k\ell}$, $1 \leq k < \ell \leq n$)

$$c_{k\ell}(t) = \frac{1}{n(\lambda_k(t) - \lambda_\ell(t))^2}, \tag{2.4}$$

$$u_k \partial_{u_\ell} = \sum_{\alpha=1}^{n} u_k(\alpha) \partial_{u_\ell(\alpha)}, \tag{2.5}$$

$$X_{k\ell}^{(s)} = u_k \partial_{u_\ell} - u_\ell \partial_{u_k},$$

We then have the following generator for the eigenvector dynamics. For a proof, see [7].

LEMMA 2.2. *For the diffusion* (2.3) *the generator acting on smooth functions* $f:\mathbb{R}^{n^2}\to\mathbb{R}$ *is*

$$\mathrm{L}_t^{(s)}=\sum_{1\le k<\ell\le n}c_{k\ell}(t)\big(X_{k\ell}^{(s)}\big)^2.$$

The above lemma means $\mathrm{d}\mathbb{E}(g(u_t))/\mathrm{d}t=\mathbb{E}(\mathrm{L}_t^{(s)}g(u_t))$ for the stochastic differential equation (2.3).

2.2 Main Result

Let I be a deterministic subset of $[\![1,n]\!]$. We denote the eigenvector overlaps as

$$\begin{aligned}p_{ij}&=\sum_{\alpha\in I}u_i(\alpha)u_j(\alpha),\quad i\neq j\in[\![1,n]\!],\\ p_{ii}&=\sum_{\alpha\in I}u_i(\alpha)^2-C_0,\quad i\in[\![1,n]\!],\end{aligned}\tag{2.6}$$

where C_0 is an arbitrary but fixed constant independent of i. We will eventually choose $C_0=|I|/n$ so that the diagonal overlaps are properly normalized, but many results in this section do not depend on the actual value of C_0. Moreover, these overlaps are functions of t ($\mathbf{u}$ satisfies the dynamics (2.3)) but this dependence is omitted in the notation.

Remember the notation (2.1) and denote

$$G(t,z)=\frac{1}{K(t)-z}.$$

For a matrix H, we abbreviate the Stieltjes transform as

$$m_H(z)=\frac{1}{n}\operatorname{Tr}\frac{1}{H-z}.$$

Assumption 2.3 (Notations and conditions for relaxation flow). Fix a small number $\mathfrak{a}>0$. A matrix V is said to be bounded if the norm of V is bounded; i.e., there is a constant $C_1>0$ such that

$$\|V\|:=\|V\|_{\mathrm{op}}\le n^{C_1}.\tag{2.7}$$

A deterministic matrix V is called $(\boldsymbol{\eta}_*,\,\boldsymbol{\eta}^*,r)$-regular at E_0 if η_*, η^* and r satisfy

$$n^{-1+\mathfrak{a}}\le\eta_*,\quad \eta_*n^{\mathfrak{a}}\le r\le n^{-\mathfrak{a}}\eta^*,\quad \eta^*n^{\mathfrak{a}}\le 1\tag{2.8}$$

and there exists C_2 such that the imaginary part of the Stieltjes transform of V is bounded from above and below by

$$C_2^{-1}\le\Im(m_V(z))\le C_2,\quad m_V(z):=\frac{1}{n}\operatorname{Tr}(V-z)^{-1},\tag{2.9}$$

uniformly for any

$$z \in \{E + \mathrm{i}\eta : E \in [E_0 - r, E_0 + r],\ \eta_* \le \eta \le \eta^*\}.$$

Our main result requires not only the above hypothesis about the Stieltjes transform but also the following estimates on individual diagonal resolvent entries.

Assumption 2.4. The following holds uniformly in

$$z \in \{E + \mathrm{i}\eta : E \in [E_0 - r, E_0 + r], \eta_* < \eta < \eta^*\}.$$

(i) Diagonal entries all have the same order:

$$\operatorname{Im} G(0, z)_{ii} \le \frac{2}{n} \operatorname{Im} \operatorname{Tr} G(0, z). \tag{2.10}$$

(ii) There exists a constant $0 < \mathfrak{c} < 1$ such that the averages over I and $[\![1, n]\!]$ coincide up to $n^{-\mathfrak{c}}$:

$$\left| \frac{1}{|I|} \sum_{i \in I} G(0, z)_{ii} - \frac{1}{n} \operatorname{Tr} G(0, z) \right| \le n^{-\mathfrak{c}}. \tag{2.11}$$

In the remainder of this article, to simplify the exposition we also assume that the deterministic set I from (2.6) satisfies

$$|I| \ge cn \tag{2.12}$$

for some small fixed constant c. This is enough for our purpose, as $|I| \sim n/2$ in the next sections. We define, for any $r > 0$ and $0 < \kappa < 1$,

$$I_\kappa^r(E) := \mathscr{I}_{E,(1-\kappa)r}, \quad \mathscr{I}_{E,r} = (E - r,\ E + r). \tag{2.13}$$

The main result of this section is the following, where we choose $C_0 = |I|/N$ in (2.6).

THEOREM 2.5 (Quantum unique ergodicity for deformed matrices). *Remember the notation* (2.6) *for the centered partial overlaps, and take $C_0 = |I|/n$ and assume* (2.12). *Under Assumption* 2.3 *and Assumption* 2.4, *the following statement holds. For any* (*small*) $\kappa, \varepsilon > 0$, *(large) $D > 0$, and $i, j \in [\![1, n]\!]$ for any t_0, t_1 such that $n^{\mathfrak{a}}\eta_* \le t_0 \le t_1 \le n^{-\mathfrak{a}} r$, we have*

$$\begin{aligned} &\mathbb{P}\left(\exists t_0 < t < t_1 : 1_{\lambda_i(t),\lambda_j(t) \in I_r^\kappa(E_0)}(|p_{ii}| + |p_{ij}|) \ge n^{\varepsilon}\left(\frac{1}{n^{\mathfrak{c}}} + \frac{1}{\sqrt{n t_0}}\right)\right)_0 \\ &\quad \le n^{-D} \end{aligned} \tag{2.14}$$

for large enough N. Here, the constant $\mathfrak{c}$ *is from* (2.11). *In other words, the errors consist of the initial error $n^{-\mathfrak{c}}$ and the dynamical error $(nt_0)^{-1/2}$.*

2.3 Perfect Matching Observables

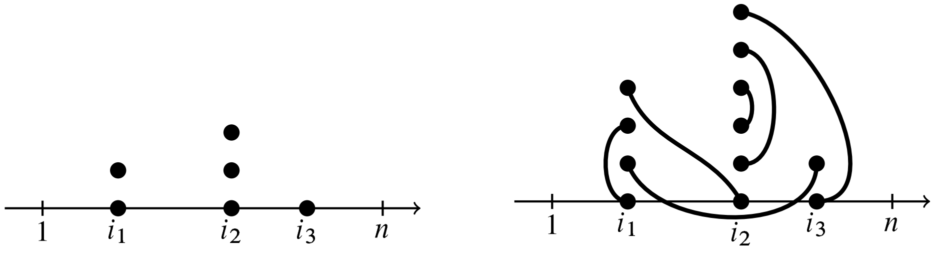
We will need the following notations.

First, as in [7], we define $\boldsymbol{\eta} : [\![1,n]\!] \to \mathbb{N}$ where $\eta_j := \boldsymbol{\eta}(j)$ is interpreted as the number of particles at the site j. Thus $\boldsymbol{\eta}$ denotes the configuration space of particles. We denote $\mathcal{N}(\boldsymbol{\eta}) = \sum_j \eta_j = d$ the total number of particles. Define $\boldsymbol{\eta}^{i,j}$ to be the configuration obtained by moving one particle from i to j. If there is no particle at i then $\boldsymbol{\eta}^{i,j} = \boldsymbol{\eta}$. Notice that there is a direction and the particle is moved from i to j.

Second, for any given configuration $\boldsymbol{\eta}$, consider the set of vertices

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

Let $\mathcal{G}_{\boldsymbol{\eta}}$ be the set of perfect matchings of the complete graph on $\mathcal{V}_{\boldsymbol{\eta}}$; i.e., this is the set of graphs G with vertices $V_{\boldsymbol{\eta}}$ and edges $\mathcal{E}(G) \subset \{\{v_1, v_2\} : v_1 \in \mathcal{V}_{\boldsymbol{\eta}}, v_2 \in \mathcal{V}_{\boldsymbol{\eta}}, v_1 \neq v_2\}$ being a partition of $\mathcal{V}_{\boldsymbol{\eta}}$.



(A) A configuration $\boldsymbol{\eta}$ with $\mathcal{N}(\boldsymbol{\eta}) = 6$, $\eta_{i_1} = 2$, $\eta_{i_2} = 3$, $\eta_{i_3} = 1$.

(B) A perfect matching $G \in \mathcal{G}_{\boldsymbol{\eta}}$. Here, $P(G) = p_{i_1 i_1} p_{i_1 i_2} p^2_{i_2 i_2} p_{i_2 i_3} p_{i_3 i_1}$.

Third, for any given edge $e = \{(i_1, a_1), (i_2, a_2)\}$, we define $p(e) = p_{i_1, i_2}$, $P(G) = \prod_{e \in \mathcal{E}(G)} p(e)$, and

$$f^{(s)}_{\boldsymbol{\lambda},t}(\boldsymbol{\eta}) = \frac{1}{\mathcal{M}(\boldsymbol{\eta})} \mathbb{E}\Big(\sum_{G \in \mathcal{G}_{\boldsymbol{\eta}}} P(G) \mid \boldsymbol{\lambda} \Big), \quad \mathcal{M}(\boldsymbol{\eta}) = \prod_{i=1}^{n} (2\eta_i)!!, \tag{2.15}$$

where $(2m)!! = \prod_{k \leq 2m, k \text{ odd}} k$ is the number of perfect matchings of the complete graph on $2m$ vertices. Remarkably, the above function f satisfies a parabolic partial differential equation.

THEOREM 2.6 (Perfect matching observables for the eigenvector moment flow: symmetric case). *Suppose that u is the solution to the symmetric eigenvector dynamics* (2.3) *and $f^{(s)}_{\boldsymbol{\lambda},t}(\boldsymbol{\eta})$ is given by* (2.15). *Then $f^{(s)}_{\boldsymbol{\lambda},t}$ satisfies the equation*

$$\partial_t f^{(s)}_{\boldsymbol{\lambda},t} = \mathscr{B}^{(s)}(t) f^{(s)}_{\boldsymbol{\lambda},t}, \tag{2.16}$$

$$\mathscr{B}^{(s)}(t) f(\boldsymbol{\eta}) = \sum_{k \neq \ell} c_{k\ell}(t) 2\eta_k (1 + 2\eta_\ell) \big(f(\boldsymbol{\eta}^{k\ell}) - f(\boldsymbol{\eta}) \big). \tag{2.17}$$

Remark 2.7. An important property of the eigenvector moment flow is the reversibility with respect to a simple explicit equilibrium measure:

$$\pi(\boldsymbol{\eta}) = \prod_{p=1}^{n} \phi(\eta_p), \ \phi(k) = \prod_{i=1}^{k}\left(1 - \frac{1}{2i}\right). \tag{2.18}$$

For any function f on the configuration space, the Dirichlet form is given by

$$\sum_{\boldsymbol{\eta}} \pi(\boldsymbol{\eta}) f(\boldsymbol{\eta}) \mathscr{B}(t) f(\boldsymbol{\eta}) = \sum_{\boldsymbol{\eta}} \pi(\boldsymbol{\eta}) \sum_{i \neq j} c_{ij} \eta_i (1 + 2\eta_j)\big(f(\boldsymbol{\eta}^{ij}) - f(\boldsymbol{\eta})\big)^2.$$

Remark 2.8. The above theorem is independent of our choice of C_0 and of the canonical basis and, more remarkably, the projection vectors don't have to be orthogonal. More precisely, let $(\mathbf{q}_\alpha)_{\alpha \in I}$ be any family of fixed vectors. Define

$$\begin{aligned} p_{ij} &= \sum_{\alpha \in I} \langle u_i, \mathbf{q}_\alpha \rangle \langle u_j, \mathbf{q}_\alpha \rangle, \quad i \neq j \in [\![1, n]\!], \\ p_{ii} &= \sum_{\alpha \in I} \langle u_i, \mathbf{q}_\alpha \rangle^2 - C_0, \quad\quad i \in [\![1, n]\!], \end{aligned}$$

and $f_{t,\boldsymbol{\lambda}}$ accordingly. Then (2.16) holds. In particular, Theorem 2.6 generalizes [7, theorem 3.1(i)] by just choosing $|I| = 1$.

2.4 Proof of Theorem 2.6

To start the proof of Theorem 2.6, let

$$g(\boldsymbol{\eta}) = \frac{1}{\mathscr{M}(\boldsymbol{\eta})} \sum_{G \in \mathscr{G}_{\boldsymbol{\eta}}} P(G) \tag{2.19}$$

and let $1 \leq k < \ell \leq n$ be fixed for the rest of this subsection. We abbreviate $X = X_{k\ell}^{(s)}$. Using Lemma 2.2, we only need to prove

$$X^2 g(\boldsymbol{\eta}) = 2\eta_k(1 + 2\eta_\ell)(g(\boldsymbol{\eta}^{k\ell}) - g(\boldsymbol{\eta})) + 2\eta_\ell(1 + 2\eta_k)(g(\boldsymbol{\eta}^{\ell k}) - g(\boldsymbol{\eta})). \tag{2.20}$$

We therefore want to calculate $X^2 P(G)$ for any $G \in \mathscr{G}_{\boldsymbol{\eta}}$. For that purpose, we first need the following definition.

DEFINITION 2.9. Let $\boldsymbol{\eta}$ and $k < \ell$ be fixed. The following notations will be useful for calculating $X^2 P(G)$.

(i) $\mathscr{V}_i \subset \mathscr{V}_{\boldsymbol{\eta}}$ is the set of vertices of type (i, a), $1 \leq a \leq 2\eta_i$.

(ii) For any two vertices $v, w \in \mathscr{V}_k \cup \mathscr{V}_\ell$, we denote

$$\varepsilon(v, w) = \begin{cases} 1 & \text{if } v, w \text{ are in the same } \mathscr{V}_i,\ i = k \text{ or } \ell \\ -1 & \text{if } v, w \text{ are in different } \mathscr{V}_i\text{'s}. \end{cases}$$

(iii) Let $G \in \mathscr{G}_{\boldsymbol{\eta}}$ and $v, w \in \mathscr{V}_k \cup \mathscr{V}_\ell$.

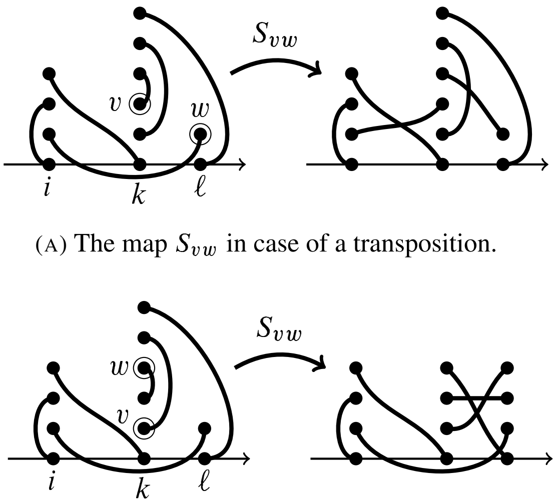
Assume $v \in \mathscr{V}_k$ and $w \in \mathscr{V}_\ell$. Then we define $S_{wv}G = S_{vw}G \in \mathscr{G}_{\boldsymbol{\eta}}$ as the perfect matching obtained by transposition of v and w. More precisely, let τ_{vw} be the permutation of $\mathscr{V}_{\boldsymbol{\eta}}$ transposing v and w. Then

$$\mathscr{E}(S_{vw}G) = \{\{\tau_{v,w}(v_1), \tau_{v,w}(v_2)\} : \{v_1, v_2\} \in \mathscr{E}(G)\}.$$

Assume $v = (k, a)$ and $w = (k, b)$ $(a < b)$ are both in $\mathscr{V}_k$. Then we define $S_{wv}G = S_{vw}G \in \mathscr{G}_{\boldsymbol{\eta}^{k\ell}}$ as the perfect matching obtained by a jump of v and w to ℓ. More precisely, let $j_{vw} = j_{wv}$ be the following bijection from $\mathscr{V}_{\boldsymbol{\eta}}$ to $\mathscr{V}_{\boldsymbol{\eta}^{k\ell}}$: $j_{vw}(v) = (\ell, 2\eta_\ell{+}1)$, $j_{vw}(w) = (\ell, 2\eta_\ell{+}2)$, $j_{vw}((k, c)) = (k, c{-}2)$ if $b < c$, $j_{vw}((k, c)) = (k, c - 1)$ if $a < c < b$ and $j_{vw}(v_1) = v_1$ in all other cases. Then

$$\mathscr{E}(S_{vw}G) = \{\{j_{v,w}(v_1), j_{v,w}(v_2)\} : \{v_1, v_2\} \in \mathscr{E}(G)\}.$$

A similar definition applies if both v and w are in $\mathscr{V}_\ell$, the jump now being towards k.



(A) The map S_{vw} in case of a transposition.

(B) The map S_{vw} in case of a jump.

In this proof, for any set A we denote $A^2_* = \{(a, b) \in A^2 : a \neq b\}$. The following result is the key step in our proof of Theorem 2.6.

LEMMA 2.10. *For any $G \in \mathscr{G}_{\boldsymbol{\eta}}$, we have*

$$X^2 P(G) = \sum_{(v,w)\in(\mathscr{V}_k\cup\mathscr{V}_\ell)^2_*} \varepsilon(v, w) P(S_{vw}G) - (2\eta_k + 2\eta_\ell) P(G). \tag{2.21}$$

We postpone the proof of the above lemma and first finish the proof of Theorem 2.6. Let

$$h(\boldsymbol{\eta}) = \sum_{G\in\mathscr{G}_{\boldsymbol{\eta}}} P(G).$$

Note that if $v \in \mathscr{V}_k$ and $w \in \mathscr{V}_\ell$, S_{vw} is a permutation of $\mathscr{G}_{\boldsymbol{\eta}}$. Moreover, if v and w are both in $\mathscr{V}_k$, S_{vw} is a bijection from $\mathscr{G}_{\boldsymbol{\eta}}$ to $\mathscr{G}_{\boldsymbol{\eta}^{k\ell}}$. The summation of (2.21) over all $G \in \mathscr{G}_{\boldsymbol{\eta}}$ therefore gives

$$\begin{aligned} X^2 h(\boldsymbol{\eta}) = & \sum_{(v,w)\in(\mathscr{V}_k)^2_*} \sum_{G\in\mathscr{G}_{\boldsymbol{\eta}}} P(S_{vw}G) + \sum_{(v,w)\in(\mathscr{V}_\ell)^2_*} \sum_{G\in\mathscr{G}_{\boldsymbol{\eta}}} P(S_{vw}G) \\ & - 2 \sum_{(v,w)\in\mathscr{V}_k\times\mathscr{V}_\ell} \sum_{G\in\mathscr{G}_{\boldsymbol{\eta}}} P(S_{vw}G) - 2(\eta_k + \eta_\ell) h(\boldsymbol{\eta}) \\ = & \sum_{(v,w)\in(\mathscr{V}_k)^2_*} h(\boldsymbol{\eta}^{k\ell}) + \sum_{(v,w)\in(\mathscr{V}_\ell)^2_*} h(\boldsymbol{\eta}^{\ell k}) \\ & - 2 \sum_{(v,w)\in\mathscr{V}_k\times\mathscr{V}_\ell} h(\boldsymbol{\eta}) - 2(\eta_k + \eta_\ell) h(\boldsymbol{\eta}) \\ X^2 h(\boldsymbol{\eta}) = & \; 2\eta_k(2\eta_k - 1) h(\boldsymbol{\eta}^{k\ell}) + 2\eta_\ell(2\eta_\ell - 1) h(\boldsymbol{\eta}^{\ell k}) \\ & - (2\eta_k(2\eta_\ell + 1) + 2\eta_\ell(2\eta_k + 1)) h(\boldsymbol{\eta}). \end{aligned}$$

The above equation implies (2.20) after renormalization by $\mathscr{M}(\boldsymbol{\eta})$. This concludes the proof of Theorem 2.6.

PROOF OF LEMMA 2.10. Let $G \in \mathscr{G}_{\boldsymbol{\eta}}$ and $1 \le k < \ell \le n$ be fixed. The Leibniz rule applies: for any smooth functions $f, g((u_i(\alpha))_{1\le i,\alpha\le n}) : \mathbb{R}^{n^2} \to \mathbb{R}$ we have $X(fg) = fX(g) + gX(f)$, so that

$$\begin{aligned} X^2 P(G) = & \sum_{(e_1,e_2)\in\mathscr{E}(G)^2_*} Xp(e_1)Xp(e_2) \prod_{e\in\mathscr{E}(G)\setminus\{e_1,e_2\}} p(e) \\ & + \sum_{e_1\in\mathscr{E}(G)} X^2 p(e_1) \prod_{e\in\mathscr{E}(G)\setminus\{e_1\}} p(e). \end{aligned} \tag{2.22}$$

The above sums will be decomposed depending on the following edge group (single, double or transverse):

$$\mathscr{E}_s = \mathscr{E}(G) \cap \{\{v,w\} : v \in \mathscr{V}_k \cup \mathscr{V}_\ell, w \notin \mathscr{V}_k \cup \mathscr{V}_\ell\}, \tag{2.23}$$

$$\mathscr{E}_d = \mathscr{E}(G) \cap \{\{v,w\} : (v,w) \in \mathscr{V}_k^2 \cup \mathscr{V}_\ell^2\}, \tag{2.24}$$

$$\mathscr{E}_t = \mathscr{E}(G) \cap \{\{v,w\} : v \in \mathscr{V}_k, w \in \mathscr{V}_\ell\}. \tag{2.25}$$

For any $v \in \mathscr{V}_{\boldsymbol{\eta}}$, let e_v be the edge containing v, and v' be the vertex such that $e_v = \{v, v'\}$. We denote

$$\begin{aligned} \mathscr{V}_s &= \{v \in \mathscr{V}_k \cup \mathscr{V}_\ell : \{v,v'\} \in \mathscr{E}_s\}, \\ \mathscr{V}_d &= \{v \in \mathscr{V}_k \cup \mathscr{V}_\ell : \{v,v'\} \in \mathscr{E}_d\}, \\ \mathscr{V}_t &= \{v \in \mathscr{V}_k \cup \mathscr{V}_\ell : \{v,v'\} \in \mathscr{E}_t\}. \end{aligned}$$

Our calculations will be based on the following basic facts: if $e \notin \mathscr{E}_s \cup \mathscr{E}_d \cup \mathscr{E}_t$, then $X_{k\ell} p(e) = 0$, and

$$Xp_{ki} = -p_{\ell i}, \tag{2.26}$$

$$Xp_{k\ell} = p_{kk} - p_{\ell\ell}, \tag{2.27}$$

$$Xp_{\ell\ell} = 2p_{k\ell}. \tag{2.28}$$

From (2.22) we have

$$X^2 P(G) = \text{(I)} + \text{(II)} + \text{(III)} + \text{(IV)} + \text{(V)} + \text{(VI)} + \text{(VII)} + \text{(VIII)} + \text{(IX)}$$

where all terms are defined and calculated below. First,

$$\begin{aligned}\text{(I)} &:= \sum_{(e_1,e_2)\in(\mathscr{E}_s)^2_*} Xp(e_1)Xp(e_2) \prod_{e\in\mathscr{E}(G)\setminus\{e_1,e_2\}} p(e)\\ &= \sum_{(v,w)\in(\mathscr{V}_s)^2_*} Xp_{\{w,w'\}}Xp_{\{v,v'\}} \prod_{e\in\mathscr{E}(G)\setminus\{e_v,e_w\}} p(e).\end{aligned}$$

From (2.26), $Xp_{\{v,v'\}}Xp_{\{w,w'\}} = -p_{\{w,v'\}}p_{\{v,w'\}}$ if v and w are in distinct $\mathscr{V}_i$'s, and $Xp_{\{v,v'\}}Xp_{\{w,w'\}} = p_{\{j_{v,w}(v),v'\}}p_{\{j_{v,w}(w),w'\}}$ if they are both in the same $\mathscr{V}_i$. In all cases, we have proved

$$\text{(I)} = \sum_{(v,w)\in(\mathscr{V}_s)^2_*} \varepsilon(v,w)P(S_{vw}G). \tag{2.29}$$

We now consider

$$\begin{aligned}\text{(II)} &:= \sum_{(e_1,e_2)\in\mathscr{E}_s\times\mathscr{E}_d\cup\mathscr{E}_d\times\mathscr{E}_s} Xp(e_1)Xp(e_2) \prod_{e\in\mathscr{E}(G)\setminus\{e_1,e_2\}} p(e)\\ &= \sum_{(v,w)\in\mathscr{V}_s\times\mathscr{V}_d} Xp_{\{v,v'\}}Xp_{\{w,w'\}} \prod_{e\in\mathscr{E}(G)\setminus\{e_v,e_w\}} p(e).\end{aligned}$$

For the second equality, note that vertices on a double edge need to be weighted by a factor of $1/2$. From (2.28) and (2.26), $Xp_{\{v,v'\}}Xp_{\{w,w'\}} = -2p_{\{w,v'\}}Xp_{\{v,w'\}}$ if v and w are in distinct $\mathscr{V}_i$'s, and $2p_{\{j_{vw}(v),v'\}}p_{\{j_{vw}(w),w'\}}$ if they are in the same $\mathscr{V}_i$. We therefore have

$$\text{(II)} = \sum_{(v,w)\in\mathscr{V}_s\times\mathscr{V}_d\cup\mathscr{V}_d\times\mathscr{V}_s} \varepsilon(v,w)P(S_{vw}G). \tag{2.30}$$

For the contribution of

$$\begin{aligned}\text{(III)} &:= \sum_{(e_1,e_2)\in(\mathscr{E}_d)^2_*} Xp(e_1)Xp(e_2) \prod_{e\in\mathscr{E}(G)\setminus\{e_1,e_2\}} p(e)\\ &= \frac{1}{4} \sum_{(v,w)\in(\mathscr{V}_d)^2_*:w\neq v'} Xp_{\{v,v'\}}Xp_{\{w,w'\}} \prod_{e\in\mathscr{E}(G)\setminus\{e_v,e_w\}} p(e)\end{aligned}$$

from (2.28) we have $Xp_{\{v,v'\}}Xp_{\{w,w'\}} = -4p_{\{w,v'\}}Xp_{\{v,w'\}}$ if v and w are in distinct $\mathscr{V}_i$'s, and $2p_{\{j_{vw}(v),v'\}}p_{\{j_{vw}(w),w'\}}$ if they are in the same $\mathscr{V}_i$. We therefore proved

$$\text{(III)} = \sum_{(v,w)\in(\mathscr{V}_d)^2_*} \varepsilon(v,w)P(S_{vw}G) - \sum_{v\in\mathscr{V}_d} P(S_{vv'}G). \tag{2.31}$$

We now calculate

$$\begin{aligned}\text{(IV)} &:= \sum_{e_1\in\mathscr{E}_s} X^2 p(e_1) \prod_{e\in\mathscr{E}(G)\backslash\{e_1\}} p(e)\\ &= \sum_{v\in\mathscr{V}_s} X^2 p_{\{v,v'\}} \prod_{e\in\mathscr{E}(G)\backslash\{e_v\}} p(e) = -\sum_{v\in\mathscr{V}_s} P(G)\end{aligned} \tag{2.32}$$

where we used (2.26) twice to obtain $X^2p_{\{v,v'\}} = -p_{\{v,v'\}}$.

For the term

$$\text{(V)} := \sum_{e_1\in\mathscr{E}_d} X^2 p(e_1) \prod_{e\in\mathscr{E}(G)\backslash\{e_1\}} p(e) = \frac{1}{2}\sum_{v\in\mathscr{V}_d} X^2 p_{\{v,v'\}} \prod_{e\in\mathscr{E}(G)\backslash\{e_1\}} p(e),$$

note that we have $X^2p_{\{v,v'\}} = 2p_{kk}-2p_{\ell\ell}$ if $v\in\mathscr{V}_\ell$, and $2p_{\ell\ell}-2p_{kk}$ otherwise. This yields

$$\text{(V)} = \sum_{v\in\mathscr{V}_d} (P(S_{v,v'}(G)) - P(G)). \tag{2.33}$$

We now consider cases where transverse edges appear:

$$\begin{aligned}\text{(VI)} &:= \sum_{(e_1,e_2)\in\mathscr{E}_s\times\mathscr{E}_t\cup\mathscr{E}_t\times\mathscr{E}_s} Xp(e_1)Xp(e_2) \prod_{e\in\mathscr{E}(G)\backslash\{e_1,e_2\}} p(e)\\ &= 2\sum_{v\in\mathscr{V}_s,\{w,w'\}\in\mathscr{E}_t} Xp_{\{v,v'\}}Xp_{\{w,w'\}} \prod_{e\in\mathscr{E}(G)\backslash\{e_v,e_w\}} p(e).\end{aligned} \tag{2.34}$$

Up to transposing w and w', we can assume that v and w are in the same $\mathscr{V}_i$. With (2.26) and (2.27), a calculation gives $Xp_{\{v,v'\}}Xp_{\{w,w'\}} = p_{j_{vw}(v)v'}p_{j_{vw}(w)w'} - p_{\tau_{vw'}(v)v'}p_{\tau_{vw'}(w')w}$. This yields

$$\text{(VI)} := \sum_{(v,w)\in\mathscr{V}_s\times\mathscr{V}_t\cup\mathscr{V}_t\times\mathscr{V}_s} \varepsilon(v,w)P(S_{vw}(G)). \tag{2.35}$$

We also have

$$\begin{aligned}\text{(VII)} &:= \sum_{(e_1,e_2)\in\mathscr{E}_d\times\mathscr{E}_t\cup\mathscr{E}_t\times\mathscr{E}_d} Xp(e_1)Xp(e_2) \prod_{e\in\mathscr{E}(G)\backslash\{e_1,e_2\}} p(e)\\ &= \sum_{v\in\mathscr{V}_d,\{w,w'\}\in\mathscr{E}_t} Xp_{\{v,v'\}}Xp_{\{w,w'\}} \prod_{e\in\mathscr{E}(G)\backslash\{e_v,e_w\}} p(e).\end{aligned}$$

We can assume v and w are in the same $\mathscr{V}_i$. Then (2.27) and (2.28) give

$$Xp_{\{v,v'\}}Xp_{\{w,w'\}} = 2(p_{j_{vw}(v)v'}p_{j_{vw}(w)w'} - p_{\tau_{vw'}(v)v'}p_{\tau_{vw'}(w')w}),$$

so that

$$\text{(VII)} := \sum_{(v,w)\in\mathscr{V}_d\times\mathscr{V}_t\cup\mathscr{V}_t\times\mathscr{V}_d} \varepsilon(v,w)P(S_{vw}(G)). \tag{2.36}$$

For two transverse edges, we have

$$\begin{aligned}\text{(VIII)} &:= \sum_{(e_1,e_2)\in(\mathscr{E}_t)^2_*} Xp(e_1)Xp(e_2) \prod_{e\in\mathscr{E}(G)\backslash\{e_1,e_2\}} p(e)\\ &= \frac{1}{4}\sum_{(v,w)\in(\mathscr{V}_t)^2_*, w\neq v'} Xp_{\{v,v'\}}Xp_{\{w,w'\}} \prod_{e\in\mathscr{E}(G)\backslash\{e_v,e_w\}} p(e).\end{aligned}$$

Without loss of generality, assume v and w are in the same $\mathscr{V}_i$. Equation (2.27) yields

$$\begin{aligned}Xp_{\{v,v'\}}Xp_{\{w,w'\}} = {}& p_{\{j_{v,w}(v),v'\}}p_{\{j_{v,w}(w),w'\}} + p_{\{j_{v',w'}(v'),v\}}p_{\{j_{v',w'}(w'),w\}}\\ &- p_{\{\tau_{v,w'}(v),v'\}}p_{\{\tau_{v,w'}(w'),w\}} - p_{\{\tau_{v',w}(v),v\}}p_{\{\tau_{v',w}(w),w'\}}.\end{aligned}$$

We therefore have

$$\text{(VIII)} = \sum_{(v,w)\in(\mathscr{V}_t)^2_*} \varepsilon(v,w)P(S_{vw}(G)) + \sum_{v\in\mathscr{V}_t} P(G). \tag{2.37}$$

Finally, from (2.27) we have $X^2 p_{k\ell} = -4p_{k\ell}$, so that

$$\text{(IX)} := \sum_{e_1\in\mathscr{E}_t} X^2 p(e_1) \prod_{e\in\mathscr{E}(G)\backslash\{e_1\}} p(e) = -2\sum_{v\in\mathscr{V}_t} P(G) \tag{2.38}$$

By summing all the equations (2.29), (2.30), (2.31), (2.32), (2.33), (2.35), (2.36), (2.37), and (2.38), the right-hand sides of (2.21) and (2.22) exactly coincide, concluding the proof of Lemma 2.10. □

3 Analysis of the Eigenvector Moment Flow

Before getting into the details of the proof of Theorem 2.5, i.e., relaxation for the eigenvector moment flow (2.17), we note substantial differences with the setting and proof in [7]. The dynamics equation (2.17) already appeared in [7], but the observables associated with equation (2.17) are now much more general (see Remark 2.8), and their natural scale (i.e., the order of the sizes of these observables) is not known a priori.

Indeed, in [7], the order of magnitude of $f_t(\boldsymbol{\eta})$ was a priori known: $f_t(\boldsymbol{\eta}) = \mathbb{E}(|\sqrt{n}\langle \mathbf{q}, u_k\rangle|^d \mid \boldsymbol{\lambda}) \leq n^{\varepsilon}$ thanks to the local law. The eigenvector moment flow was used in [7] to find fluctuations around this scale.

On the contrary, in the current paper, the eigenvector moment flow (2.17) allows us to find the natural scale for a wider class of observables. For $|I| \sim cn$, local laws only give the trivial estimate $|p_{ii}| \leq 1$ for example, although the dynamics yield Theorem 2.5, i.e., $|p_{ii}| \leq n^{-1/2+\varepsilon}$ for t approaching 1.

This differences about observables and scales require the following notable novelties in the proof of Theorem 2.5:

(i) The decomposition between long-range and short-range dynamics is now more intricate. In particular, our bound on the long-range contribution improves in inductive steps (see Lemma 3.5 to be compared with [7, lemma 6.1]).

(ii) The maximum principle, Proposition 3.7, also gives stronger results once it is used inductively, on space-time embedded domains, while the analogue [7, theorem 7.4] only required one time step.

In summary, the error terms in the finite speed of propagation and the maximum principle estimates depend on the size of $f_t(\boldsymbol{\eta})$. In this paper, the a priori bound on $f_t(\boldsymbol{\eta})$ is far from its real size. Hence we need to bootstrap our estimates in a suitable way in order to get a sharp estimate at the end of the proof.

We now introduce a few notations that will be useful in the statement and proof of the following lemma 3.1 and in following this section. For a fixed and arbitrarily small $\omega > 0$, we define the control parameter

$$\psi = n^{\omega}$$

with $\omega \leq \mathfrak{a}/100$, and the following time and spectral domains:

$$\mathcal{T}_{\omega}(\eta_*, \eta^*, r) = \{t : \eta_* \psi \leq t \leq \psi^{-1} r\}, \tag{3.1}$$

3.1 A Priori Estimates

For $K(t)$ in (2.1), we denote the initial matrix $V = U_0 \Lambda_0 U_0^*$, where $\Lambda_0 = \operatorname{diag}\{\lambda_1(0), \dots, \lambda_n(0)\}$, and U_0 is the orthogonal matrix of its eigenvectors. Let $m_{\mathrm{fc},t}$ be the Stieltjes transform of the free convolution between the empirical spectral measure of V and the Gaussian orthogonal ensemble Z_t. Then $m_{\mathrm{fc},t}$ solves the equation

$$\begin{aligned} m_{\mathrm{fc},t}^{(n)}(z) &= m_V\big(z + t m_{\mathrm{fc},t}^{(n)}(z)\big) \\ &= \frac{1}{n} \sum_{i=1}^{n} g_i(t,z),\, g_i(t,z) := \frac{1}{\lambda_i(0) - z - t m_{\mathrm{fc},t}(z)}. \end{aligned} \tag{3.2}$$

Here $m_{\mathrm{fc},t}^{(n)}(z)$ is the Stieltjes transform of a measure with density denoted $\rho_{\mathrm{fc},t}^{(n)}$. For notational convenience we will suppress the superscript and use the notations $m_{\mathrm{fc},t}(z)$, $\rho_{\mathrm{fc},t}$.

The typical location $\gamma_i(s)$ of the i^{th} eigenvalue $\lambda_i(s)$ is defined through $\int_{-\infty}^{\gamma_i(s)} \mathrm{d}\rho_{\mathrm{fc},s} = \frac{i}{n}$. We also recall the following stability property of the typical locations; see [23, lemma 3.4]: for any $0 < q_1 < q_2 < 1$ and $\omega > 0$, for large enough n we have, for all $s, t \in \mathcal{T}_{\omega}(\eta_*, \eta^*, r)$,

$$\{i : \gamma_i(s) \in \mathcal{I}_{E_0, q_1 r}\} \subset \{i : \gamma_i(t) \in \mathcal{I}_{E_0, q_2 r}\}. \tag{3.3}$$

LEMMA 3.1 (Delocalization for deformed matrices). *Let $\tau > 0$ and let $\mathbf{u}_{i,t}$ denote the normalized eigenvector of $K(t)$ in* (2.1)*, whose eigenvalues are $\lambda_{i,t}$, $1 \le i \le n$. We assume that $t \in \mathscr{T}_\omega(\eta_*, \eta^*, r)$, V is (η_*, η^*, r)-regular at E_0 and bounded as in* (2.7)*, and that there exists $C > 0$ such that for any $D > 0$, for large enough n we have*

$$\mathbb{P}(\exists E \in \mathscr{I}_{E_0,r}, \eta_* < \eta < rn^{-\omega} : \operatorname{Im} G(0, E + \mathrm{i}\eta)_{ii} \ge C) \le n^{-D} \tag{3.4}$$

for any $1 \le i \le n$. Here $G(0, z)$ is the Green function of the initial matrix V. Then for any $\kappa, \tau, D > 0$, provided that n is sufficiently large we have

$$\mathbb{P}\big(\mathbb{1}_{|\lambda_{k,t}-E_0|\le(1-\kappa)r}\|u_{k,t}\|_\infty^2 \ge n^{-1+\tau}\big) \le n^{-D}$$

uniformly in $1 \le k \le n$.

Remark 3.2. This lemma is essentially a restatement of [5, theorem 2.1], which holds in the domain $\{z = E + \mathrm{i}\eta : E \in I_\kappa^r(E_0), \psi^4/n \le \eta \le 1 - \kappa r\}$ under the $(\eta_*, 1, r)$-regularity for V; see [5, assumption 1.3].

In Lemma 3.1 the assumption is weaker: we only have (η_*, η^*, r)-regularity for V. A simple inspection of the proof of [5, theorem 2.1] shows that its conclusion remains, in the restricted domain $\psi^4/n \le \eta \le rn^{-\omega}$ (which will be sufficient for our purpose) under this (η_*, η^*, r)-regularity assumption.

PROOF. We bound the eigenvector coordinates by the diagonal entries of the resolvent through

$$|u_{k,t}(i)|^2 \le n^{-1+\tau} \operatorname{Im} G(t, \lambda_{k,t} + \mathrm{i}n^{-1+\tau})_{ii}. \tag{3.5}$$

If $|\lambda_{k,t} - E_0| \le (1-\kappa)r$, denoting $z = \lambda_{k,t} + \mathrm{i}n^{-1+\tau}$ we have

$$z \in \left\{E + \mathrm{i}\eta : |E - E_0| < (1-\kappa)r, \frac{\psi^4}{n} \le \eta \le r\eta^*\psi^{-1}\right\}. \tag{3.6}$$

The local law from [5, theorem 2.1] with the domain adjustment from Remark 3.2 states that $U_0 \operatorname{diag}\{g_1(t,z), g_2(t,z), \dots, g_n(t,z)\}U_0^*$ is a good approximation for $G(t,z)$; i.e., for any $\eta_* \ll t \ll r$ and any unit vector $\mathbf{q}$, uniformly for any z as in (3.6), the following holds with overwhelming probability:

$$\left|\langle \mathbf{q}, G(t,z)\mathbf{q}\rangle - \sum_{i=1}^n \langle \mathbf{u}_i(0), \mathbf{q}\rangle^2 g_i(t,z)\right| \le \frac{\psi^2}{\sqrt{n\eta}} \operatorname{Im}\left(\sum_{i=1}^n \langle \mathbf{u}_i(0), \mathbf{q}\rangle^2 g_i(t,z)\right). \tag{3.7}$$

Clearly, we can restate the last result as

$$\begin{aligned} &|\langle \mathbf{q}, G(t,z)\mathbf{q}\rangle - \langle \mathbf{q}, G(0, z + tm_{\mathrm{fc},t}(z))\mathbf{q}\rangle| \\ &\quad \le \frac{\psi^2}{\sqrt{n\eta}} \operatorname{Im}\langle \mathbf{q}, G(0, z + tm_{\mathrm{fc},t}(z))\mathbf{q}\rangle, \end{aligned} \tag{3.8}$$

where $G(0, z)$ is the Green's function of V. Since $\psi^2/\sqrt{n\eta} \le 1$, we have

$$|\operatorname{Im} G(t,z)_{ii} - \operatorname{Im} G(0, z + tm_{\mathrm{fc},t}(z))_{ii}| \le \operatorname{Im} G(0, z + tm_{\mathrm{fc},t}(z))_{ii}. \tag{3.9}$$

From [5, proposition 2.2], for some fixed constant $C > 0$ we have

$$C^{-1} < \operatorname{Im} m_{\mathrm{fc},t}(z) < C \quad \text{and} \quad |\operatorname{Re} m_{\mathrm{fc},t}(z)| < C \log n,$$

so that $\operatorname{Re}(z + t m_{\mathrm{fc},t}(z)) \in \mathscr{I}_{E_0,r}$ and $\eta_* < \operatorname{Im}(z + t m_{\mathrm{fc},t}(z)) < n^{-\omega} r$. With (3.4), we deduce that

$$\operatorname{Im} G(0, z + t m_{\mathrm{fc},t}(z))_{ii} \le C \tag{3.10}$$

with overwhelming probability. Equations (3.5), (3.9), and (3.10) conclude the proof. □

Similarly to [5,7], we split the operator $\mathscr{B}(t)$ from (2.17) into a short-range part and a long-range part through a short-range parameter ℓ : $\mathscr{B}(t) = \mathscr{S}(t) + \mathscr{L}(t)$, with

$$(\mathscr{S} f_t)(\boldsymbol{\eta}) = \sum_{0<|j-k|\le\ell} c_{jk}(t) 2\eta_j (1 + 2\eta_k)(f_t(\boldsymbol{\eta}^{jk}) - f_t(\boldsymbol{\eta})),$$

$$(\mathscr{L} f_t)(\boldsymbol{\eta}) = \sum_{|j-k|>\ell} c_{jk}(t) 2\eta_j (1 + 2\eta_k)(f_t(\boldsymbol{\eta}^{jk}) - f_t(\boldsymbol{\eta})).$$

Notice that $\mathscr{S}$ and $\mathscr{L}$ are also reversible with respect to the measure π from (2.18). We denote by $\mathrm{U}_{\mathscr{B}}(s,t)$, $\mathrm{U}_{\mathscr{S}}(s,t)$, and $\mathrm{U}_{\mathscr{L}}(s,t)$ the semigroup associated with $\mathscr{B}$, $\mathscr{S}$, and $\mathscr{L}$, respectively, from time s to t, i.e.,

$$\partial_t \mathrm{U}_{\mathscr{B}}(s,t) = \mathscr{B}(t) \mathrm{U}_{\mathscr{B}}(s,t).$$

For a fixed $\kappa > 0$, consider the following “distance" on n-particle configurations:

$$d(\boldsymbol{\eta}, \boldsymbol{\xi}) = \max_{1\le\alpha\le d} \#\{i \in [\![1,n]\!] : \gamma_i(t_0) \in I_\kappa^r(E_0),\ i \in [\![x_\alpha, y_\alpha]\!] \cup [\![y_\alpha, x_\alpha]\!]\}, \tag{3.11}$$

where $\boldsymbol{\eta}$: $1 \le x_1 \le x_2 \le \cdots \le x_d \le n$ and $\boldsymbol{\xi}$: $1 \le y_1 \le y_2 \le \cdots \le y_d \le n$, and an initial time t_0 defined in the next lemma. Note that we use the notation d for $\widetilde{d}$ defined in [5, equation (3.10)].

LEMMA 3.3. *Assume the initial estimates* (2.7), (2.8), (2.9), *and* (2.11) *hold. We fix times* t_0, t_1, *and the range parameter* ℓ *such that*

$$\psi \eta_* \le t_0 \le t_1 \le \frac{\ell}{n\psi} \le \frac{r}{\psi^{10}}. \tag{3.12}$$

The matrix Brownian motion $(K(s))_{0\le s\le t_1}$ *defined in Section* (2.1)*) induces a measure on the space of eigenvalues and eigenvectors* $(\boldsymbol{\lambda}(s), \boldsymbol{u}(s))$ *for* $0 \le s \le t_1$ *such that, for any* $\kappa > 0$, *the following event* A *holds with overwhelming probability*:

(i) *The eigenvalue rigidity estimate holds*: $\sup_{t_0\le s\le t_1} |m_s(z) - m_{\mathrm{fc},s}(z)| \le \psi(n\eta)^{-1}$ *uniformly in* $z \in \mathscr{D}_\kappa$ *and* $\sup_{t_0\le s\le t_1} |\lambda_i(s) - \gamma_i(s)| \le \psi n^{-1}$ *uniformly for indices* i *such that* $\gamma_i(s) \in I_r^\kappa(E_0)$.

(ii) *When we condition on the trajectory $\boldsymbol{\lambda} \in A$, with overwhelming probability, the following holds*:

$$\sup_{t_0 \le s \le t_1} |G(s,z)_{ii} - G(0, z + s m_{\mathrm{fc},s}(z))_{ii}| \le C \frac{\psi^2}{\sqrt{n\eta}}, \tag{3.13}$$

$$\sup_{t_0 \le s \le t_1} \left| \frac{1}{|I|} \sum_{i \in I} G(s,z)_{ii} - \frac{1}{n} \operatorname{Tr} G(s,z) \right| \le \frac{C\psi^2}{n^{\mathfrak{c}}} + \frac{C\psi^4}{n\eta}, \tag{3.14}$$

uniformly in $z \in \mathscr{D}_\kappa$, where $\mathfrak{b}, \mathfrak{c}$ *are defined in Assumption* 2.4.

(iii) *Finite speed of propagation holds*: *for any d there exists $C_d, c_d > 0$ such that uniformly, for any function h on the space of d particle configurations and particle configuration $\boldsymbol{\xi}$, which is away from the support of h in the sense that $d(\boldsymbol{\eta}, \boldsymbol{\xi}) \ge \psi\ell$, we have for any $\boldsymbol{\eta}$ in the support of h that*

$$\sup_{t_0 \le s' \le s \le t_1} \mathrm{U}_{\mathscr{S}}(s', s) h(\boldsymbol{\xi}) \le C_d \|h\|_\infty n^d e^{-c_d \psi}. \tag{3.15}$$

PROOF. Statement (i) was proved in [23, theorem 3.3 and 3.5], and (3.13) and (iii) were given in theorem 2.1 and lemma 3.4 of [5], respectively. For the proof of (3.14), we decompose

$$\begin{aligned} &\frac{1}{|I|} \sum_{i \in I} G(s,z)_{ii} - \frac{1}{n} \operatorname{Tr} G(s,z) \\ &\quad = \frac{1}{|I|} \sum_{i \in I} \big(G(s,z)_{ii} - G(0, z + s m_{\mathrm{fc},s}(z))_{ii} \big) \\ &\qquad + \left(\frac{1}{|I|} \sum_{i \in I} G(0, z + s m_{\mathrm{fc},s}(z))_{ii} - \frac{1}{n} \sum_{1 \le i \le n} G(0, z + s m_{\mathrm{fc},s}(z))_{ii} \right) \\ &\qquad - \frac{1}{n} \sum_{1 \le i \le n} \big(G(s,z)_{ii} - G(0, z + s m_{\mathrm{fc},s}(z))_{ii} \big) \end{aligned}$$

The second sum is exactly the left-hand side of (2.11), so it is bounded by $n^{-\mathfrak{c}}$. The third sum is just the difference between Stieltjes transforms, and it was proved in [23, theorem 3.3] is of order at most $n^{\varepsilon}/(n\eta)$ thanks to (3.8). Notice that we have used $m_{\mathrm{fc},s}(z) = G(0, z + s m_{\mathrm{fc},s}(z))$ by definition.

The first sum of the last displayed equation is of the same type as the third one except that the average is over not all entries but a macroscopic fraction of them. The proof in [23, theorem 3.3], based on a fluctuation averaging lemma, can be replicated to yield that

$$\left| \frac{1}{|I|} \sum_{i \in I} \big(G(s,z)_{ii} - G(0, z + s m_{\mathrm{fc},s}(z))_{ii} \big) \right| \le \frac{n^{\varepsilon}}{n\eta}$$

with overwhelming probability. This completes the proof of Lemma 3.3. □

Remark 3.4. The following is an elementary consequence of the above rigidity estimate (i) together with (3.3). For any $t_0 \leq s \leq t_1$ and interval $I \subset I^r_\kappa(E_0)$ with $|I| \geq \psi^4/n$, we have

$$C^{-1}|I|n \leq \#\{i : \gamma_i(s) \in I\} + \#\{i : \lambda_i(s) \in I\} \leq C|I|n. \tag{3.16}$$

3.2 Approximation with Short-Range Dynamics

We introduce the notation

$$\mathrm{S}_I^{(u,v)} = \sup_{\boldsymbol{\eta}\subset I, u\leq s\leq v} f_s(\boldsymbol{\eta})$$

for the following lemma. For $i \in \mathbb{Z}$ and $J \subset \mathbb{Z}$, let $d(i, J) = \inf_{j\in J} |i - j|$. Finally, from here we assume that the number of particles of the eigenvector moment flow is even, i.e.,

$$d = 2m.$$

LEMMA 3.5. *Under the assumptions of Lemma* 3.3, *consider* $\boldsymbol{\lambda} \in A$, *with* A *defined in the same lemma. Consider the perfect matching observables* f_u *from* (A.3). *Then, for large enough* n, *for any intervals* $J_{\text{in}} \subset \{i : \gamma_i(t_0) \in I^r_{2\kappa}(E)\}$ *and* $J_{\text{out}} = \{i : d(i, J_{\text{in}}) \leq \psi\ell\}$, *any* d*-particle configuration* $\boldsymbol{\xi}$ *supported on* J_{in}, *and any* $t_0 < u < v < t_1$ *we have*

$$\begin{aligned}&|((\mathrm{U}_{\mathscr{B}}(u,v) - \mathrm{U}_{\mathscr{S}}(u,v))f_u)(\boldsymbol{\xi})|\\ &\quad\leq \psi^4\frac{n|u-v|}{\ell}\left(\mathrm{S}^{(u,v)}_{J_{\text{out}}} + \frac{1}{n^{\mathfrak{c}}}\big(\mathrm{S}^{(u,v)}_{J_{\text{out}}}\big)^{\frac{d-1}{d}} + \frac{1}{\ell}\big(\mathrm{S}^{(u,v)}_{J_{\text{out}}}\big)^{\frac{d-2}{d}}\right).\end{aligned}$$

PROOF. We first define, similarly to [5, 7], the following flattening operators on the space of functions of configurations with d points:

$$(\mathrm{Flat}_a(f))(\boldsymbol{\eta}) = \begin{cases} f(\boldsymbol{\eta}) & \text{if } \boldsymbol{\eta} \subset \{i : d(i, J_{\text{in}}) \leq a\},\\ 0 & \text{otherwise,}\end{cases}$$

By Duhamel's formula,

$$((\mathrm{U}_{\mathscr{S}}(u,v) - \mathrm{U}_{\mathscr{B}}(u,v))f_u)(\boldsymbol{\xi}) = \int_u^v \mathrm{U}_{\mathscr{S}}(s,v)\mathscr{L}(s)f_s(\boldsymbol{\xi})\mathrm{d}s.$$

Notice that $d(\mathrm{supp}(\mathscr{L}(s)f_s - \mathrm{Flat}_{\psi\ell}(\mathscr{L}(s)f_s)), \boldsymbol{\xi}) \geq \psi\ell$. Therefore by the finite speed of propagation (3.15) in Lemma 3.3 of $\mathrm{U}_{\mathscr{S}}$, we have

$$\begin{aligned}|(\mathrm{U}_{\mathscr{S}}(s,v)\mathscr{L}(s)f_s)(\boldsymbol{\xi})| &= |\mathrm{U}_{\mathscr{S}}(s,v)\mathrm{Flat}_{\psi\ell}(\mathscr{L}(s)f_s)(\boldsymbol{\xi})| + \mathrm{O}(e^{-c\psi/2})\\ &\leq \max_{\widetilde{\boldsymbol{\eta}}} |\mathrm{Flat}_{\psi\ell}(\mathscr{L}(s)f_s)(\widetilde{\boldsymbol{\eta}})| + \mathrm{O}(e^{-c\psi/2}).\end{aligned} \tag{3.17}$$

where in the last inequality, we used that $\mathrm{U}_{\mathscr{S}}$ is a contraction in L^∞.

Let $\widetilde{\boldsymbol{\eta}}$ be a configuration $\{(i_1, j_1), \dots, (i_d, j_d)\}$ with support in J_{out}. In view of (3.17), we only need to prove that

$$|(\mathscr{L}(s)f_s)(\widetilde{\boldsymbol{\eta}})| \leq \psi^4\frac{n}{\ell}\left(\mathrm{S}^{(u,v)}_{J_{\text{out}}} + \frac{1}{n^{\mathfrak{c}}}\big(\mathrm{S}^{(u,v)}_{J_{\text{out}}}\big)^{\frac{d-1}{d}} + \frac{1}{\ell}\big(\mathrm{S}^{(u,v)}_{J_{\text{out}}}\big)^{\frac{d-2}{d}}\right). \tag{3.18}$$

We have

$$\mathscr{L}(s) f_s(\widetilde{\boldsymbol{\eta}}) \le \left| \sum_{|j-k|\ge \ell} \frac{f_s(\widetilde{\boldsymbol{\eta}}^{jk})}{n(\lambda_j-\lambda_k)^2} \right| + |f_s(\widetilde{\boldsymbol{\eta}})| \sum_{1\le p\le d, |i_p-k|\ge \ell} \frac{1}{n(\lambda_{i_p}-\lambda_k)^2}.$$

Notice that $i_p \in J_{\mathrm{out}}$, and thus $\lambda_{i_p}(s) \in I_k^r(E_0)$. We denote $\eta_q = 2^q \ell/n$. From the local law and a dyadic decomposition we have

$$\sum_{k:|i_p-k|\ge \ell} \frac{1}{n(\lambda_{i_p}-\lambda_k)^2} \le \sum_{q=1}^{\lceil \log_2 n/\ell \rceil} \frac{1}{\eta_q} \sum_k \frac{\eta_q}{n((\lambda_{i_p}-\lambda_k)^2+\eta_q^2)} \le \frac{n}{\ell},$$

so that the second term on the right-hand side of (3.18) is bounded by the right-hand side of (3.17), as desired.

More subtle bounds are required for

$$\begin{aligned}
&\sum_{|j-k|\ge \ell} \frac{f_s(\widetilde{\boldsymbol{\eta}}^{jk})}{n(\lambda_j-\lambda_k)^2} \\
&\quad = \sum_{|j-k|\ge \ell, \widetilde{\eta}_k=0} \frac{f_s(\widetilde{\boldsymbol{\eta}}^{jk})}{n(\lambda_j-\lambda_k)^2} + \mathrm{O}\left(\frac{n}{\ell^2}\right) \sup_{u\le s\le v, \boldsymbol{\eta}\subset J_{\mathrm{out}}} |f_s(\boldsymbol{\eta})|
\end{aligned}$$

where we used that $\widetilde{\boldsymbol{\eta}}^{jk} \subset J_{\mathrm{out}}$ if $\widetilde{\eta}_k \neq 0$, and $1/(n(\lambda_j-\lambda_k)^2) \le 1/(n(\ell/n)^2)$ for $|j-k| \ge \ell$ by rigidity; see Lemma 3.3(i). For fixed p, we therefore want to bound

$$\sum_{|i_p-k|\ge \ell, \widetilde{\eta}_k=0} \sum_{G\in \mathscr{G}_{\widetilde{\boldsymbol{\eta}} i_p k}} \frac{\mathbb{E}(P(G) \mid \boldsymbol{\lambda})}{n(\lambda_{i_p}-\lambda_k)^2} = \text{(I)} + \text{(II)}$$

where (I) corresponds to perfect matchings such that $\{(k,1),(k,2)\}$ is not an edge, and (II) corresponds to perfect matchings with an edge of type $\{(k,1),(k,2)\}$. More precisely,

$$\begin{aligned}
\text{(I)} &= \sum_{1\le q_1,q_2\le d} \mathbb{E}\left(P^{(q_1,q_2)}(p(e)_{e\in \mathscr{E}_{\widetilde{\boldsymbol{\eta}}}}) \sum_{|k-i_p|>\ell, \widetilde{\eta}_k=0} \frac{p_{i_{q_1}k} p_{i_{q_2}k}}{n(\lambda_{i_p}-\lambda_k)^2} \;\Big|\; \boldsymbol{\lambda} \right), \\
\text{(II)} &= \mathbb{E}\left(P^{(p)}((p(e)_{e\in \mathscr{E}_{\widetilde{\boldsymbol{\eta}}}})) \sum_{|k-i_p|>\ell, \widetilde{\eta}_k=0} \frac{p_{kk}}{n(\lambda_{i_p}-\lambda_k)^2} \;\Big|\; \boldsymbol{\lambda} \right),
\end{aligned}$$

with $\mathscr{E}_{\widetilde{\boldsymbol{\eta}}}$ the set of all possible edges between between vertices from $\mathscr{V}_{\widetilde{\boldsymbol{\eta}}}$, $P^{(p,q)}$ is a finite sum of monic monomials of degree $d-2$, and $P^{(p)}$ is a finite sum of monic monomials of degree $d-1$.

To bound (I), we simply write

$$\sum_{|k-i_p|>\ell, \widetilde{\eta}_k=0} \frac{p_{i_{q_1}k} p_{i_{q_2}k}}{n(\lambda_{i_p}-\lambda_k)^2} = \mathrm{O}\left(\frac{1}{n(\ell/n)^2} \sum_k (p_{i_{q_1}k}^2 + p_{i_{q_2}k}^2) \right) = \mathrm{O}\left(n^{\varepsilon} \frac{|I|}{\ell^2} \right),$$

where we slightly changed the meaning of p_{kk} (only in the equation above and the equation below, $p_{kk} = \sum_{\alpha \in I} u_k(\alpha)^2$, i.e., $C_0 = 0$ in (2.6)) and used the elementary identity

$$\sum_k p_{ik}^2 = \sum_{\alpha \in I} u_i(\alpha)^2 = \mathrm{O}\left(n^{\varepsilon} \frac{|I|}{n}\right). \tag{3.19}$$

The above second equality follows from Lemma 3.1. Moreover, with Lemma 3.6, we have

$$\begin{aligned} \mathbb{E}\big(|P^{(q_1,q_2)}(p(e)_{e \in \mathscr{E}_{\tilde{\boldsymbol{\eta}}}})| \;\big|\; \boldsymbol{\lambda}\big) &= \mathrm{O}\Big(\sup_{u \le s \le v, \boldsymbol{\eta} \subset J_{\text{out}}} |f_s(\boldsymbol{\eta})| \mathbb{1}_{\mathcal{N}(\boldsymbol{\eta}) = d-2}\Big) \\ &= \mathrm{O}\Big(\big(\mathrm{S}_{J_{\text{out}}}^{(u,v)}\big)^{\frac{d-2}{d}}\Big), \end{aligned}$$

where we used Hölder's inequality and Lemma 3.6. This concludes our estimate for (I).

The term (II) is more complicated to bound. For fixed p and s, let $E_1 = \gamma_{i_p - \ell}$, $E_1^- = \gamma_{i_p - \ell - n^{\varepsilon}}$, $E_1^+ = \gamma_{i_p - \ell + n^{\varepsilon}}$, $E_2 = \gamma_{i_p + \ell}$, $E_2^- = \gamma_{i_p + \ell - n^{\varepsilon}}$, and $E_2^+ = \gamma_{i_p + \ell + n^{\varepsilon}}$. We also define the contour Γ as the rectangle with vertices $E_1 \pm \mathrm{i}\frac{\ell}{n}$, $E_2 \pm \mathrm{i}\frac{\ell}{n}$. Let

$$\begin{aligned} f(z) &= \sum_{k : \gamma_k \notin [E_1^-, E_2^+]} \frac{p_{kk}}{n(z - \lambda_k)}, \\ g(z) &= \sum_{k : \gamma_k \notin [E_1^-, E_1^+] \cup [E_2^-, E_2^+]} \frac{p_{kk}}{n(z - \lambda_k)}. \end{aligned} \tag{3.20}$$

We now assume $|z - \lambda_{i_p}| \le n^{-\varepsilon} \frac{\ell}{n}$. By Cauchy's formula, we have

$$f(z) = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma} \frac{f(\xi)}{\xi - z} \, \mathrm{d}\xi = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma} \frac{g(\xi)}{\xi - z} \, \mathrm{d}\xi,$$

where for the second equality we used that, for any λ_k (and z) inside Γ we have

$$\int_{\Gamma} \frac{\mathrm{d}\xi}{(\xi - \lambda_k)(\xi - z)} = 0,$$

from a residue calculus. Define

$$\Gamma_{\text{int}} = \{z = E + \mathrm{i}\eta : E = E_1 \text{ or } E_2, |\eta| < n^{\varepsilon}/n\} \quad \text{and} \quad \Gamma_{\text{ext}} = \Gamma / \Gamma_{\text{int}}.$$

We first bound the contribution due to small η: we have

$$\left| \int_{\Gamma_{\text{int}}} \frac{g(\xi)}{\xi - z} \, \mathrm{d}\xi \right| \le \frac{n}{\ell} \int_{\Gamma_{\text{int}}} \sum_{\substack{k < i_p - \ell - n^{\varepsilon}, \\ i_p + \ell + n^{\varepsilon} < k, \\ i_p - \ell + n^{\varepsilon} < k < i_p + \ell - n^{\varepsilon}}} \frac{|p_{kk}|}{n |\lambda_k - \xi|}.$$

We simply bound $|p_{kk}|$ by 1 and obtain that the corresponding integral is at most

$$\left|\int_{\Gamma_{\mathrm{int}}} \frac{g(\xi)}{\xi - z}\,\mathrm{d}\xi\right| \leq \frac{n}{\ell}\frac{n^{\varepsilon}}{n}\sum_{k\geq\ell}\frac{1}{n(k/n)} = \mathrm{O}\left(\frac{n^{\varepsilon}}{\ell}\right).$$

We now bound the contribution from Γ_{ext}. On this domain, we can afford extending the definition of g to the full sum $1 \leq k \leq n$, up to an error of order

$$n^{\varepsilon}\frac{n}{\ell}\int_{\Gamma_{\mathrm{Ext}}} \frac{|p_{kk}|}{n|\xi - E_1|} \leq \frac{n^{\varepsilon}}{\ell}.$$

We therefore proved

$$\begin{aligned} f(z) &\leq \frac{n}{\ell}\int_{\Gamma_{\mathrm{Ext}}}\left|\frac{1}{n}\sum_{1\leq k\leq n}\frac{p_{kk}}{\xi - \lambda_k}\right||\mathrm{d}\xi| + \frac{n^{\varepsilon}}{\ell} \\ &\leq \frac{n}{\ell}\int_{\Gamma_{\mathrm{Ext}}}\left(\frac{\psi^4}{n\operatorname{Im}\xi} + \frac{\psi}{n^{\mathrm{c}}}\right)|\mathrm{d}\xi| + \frac{n^{\varepsilon}}{\ell} = \mathrm{O}(n^{\varepsilon})\left(\frac{1}{\ell} + \frac{1}{n^{\mathrm{c}}}\right), \end{aligned} \tag{3.21}$$

where we used (3.14) in the second inequality. We conclude that $\partial_z f(\lambda_{i_p}) = \mathrm{O}(n^{\varepsilon})\frac{n}{\ell}(\frac{1}{\ell} + \frac{1}{n^{\mathrm{c}}})$, so that

$$\text{(II)} = \mathrm{O}\left(\frac{n}{\ell^2} + \frac{n}{\ell n^{\mathrm{c}}}\right)\left(\mathrm{S}^{(u,v)}_{J_{\mathrm{out}}}\right)^{\frac{d-1}{d}}$$

where we used Hölder's inequality and Lemma 3.6. This concludes the proof of (3.18) and the lemma (note that $\frac{n}{\ell^2}(\mathrm{S}^{(u,v)}_{J_{\mathrm{out}}})^{\frac{d-1}{d}} = \mathrm{O}(\frac{n}{\ell^2}(\mathrm{S}^{(u,v)}_{J_{\mathrm{out}}})^{\frac{d-2}{d}})$). □

LEMMA 3.6. *Denote by $\boldsymbol{\eta}$ the configuration with m particles at site i, m particles at site j, and no particles elsewhere. Moreover, denote by $\boldsymbol{\eta}^{(1)}$ ($\boldsymbol{\eta}^{(2)}$ resp.) the configurations with $d = 2m$ particles on the site i (resp., site j) and no particles elsewhere. Then there exists $C_1, C_2, C > 0$ depending only on d such that for any $i < j$ and any time s we have*

$$\mathbb{E}(p_{ij}(s)^d \mid \boldsymbol{\lambda}) \leq C_1 f_{\boldsymbol{\lambda},s}(\boldsymbol{\eta}^{(1)}) + C_2 f_{\boldsymbol{\lambda},s}(\boldsymbol{\eta}^{(2)}) + C f_{\boldsymbol{\lambda},s}(\boldsymbol{\eta}).$$

PROOF. From (A.3), we have

$$f_{\boldsymbol{\lambda},s}(\boldsymbol{\eta}) = a_d\mathbb{E}\big(p_{ij}^d \mid \boldsymbol{\lambda}\big) + \sum_{\alpha+\beta+\gamma=d,\alpha<d} b_{\alpha,\beta,\gamma}\mathbb{E}\big(p_{ij}^{\alpha}p_{ii}^{\beta}p_{jj}^{\gamma} \mid \boldsymbol{\lambda}\big) \tag{3.22}$$

for some coefficients $a_d > 0$, $b_{\alpha,\beta,\gamma} \geq 0$. From Young's inequality, for any $\varepsilon > 0$ we have

$$\left|\mathbb{E}\big(p_{ij}^{\alpha}p_{ii}^{\beta}p_{jj}^{\gamma} \mid \boldsymbol{\lambda}\big)\right| \leq \frac{\alpha\varepsilon^2}{d}\mathbb{E}\big(p_{ij}^d \mid \boldsymbol{\lambda}\big) + \frac{\beta}{d\varepsilon}\mathbb{E}\big(p_{ii}^d \mid \boldsymbol{\lambda}\big) + \frac{\gamma}{d\varepsilon}\mathbb{E}\big(p_{jj}^d \mid \boldsymbol{\lambda}\big). \tag{3.23}$$

Equations (3.22) and (3.23) imply

$$\begin{aligned}&\mathbb{E}\big(p_{ij}^d \mid \boldsymbol{\lambda}\big)\\ &\quad\le \frac{f_{\boldsymbol{\lambda},s}(\boldsymbol{\eta})}{a_d}\\ &\qquad+ \sum_{\substack{\alpha+\beta+\gamma=d,\\ \alpha<d}} \frac{b_{\alpha,\beta,\gamma}}{a_d}\left(\frac{\alpha\varepsilon^2}{d}\big(p_{ij}^d \mid \boldsymbol{\lambda}\big) + \frac{\beta\mathbb{E}(p_{ii}^d \mid \boldsymbol{\lambda}) + \gamma\mathbb{E}(p_{jj}^d \mid \boldsymbol{\lambda})}{d\varepsilon}\right).\end{aligned}$$

The result follows by choosing $\varepsilon = \varepsilon(d)$ small enough. □

3.3 Maximum Principle

Iterations of the following proposition will give the main result, Theorem 2.5.

PROPOSITION 3.7. *For any eigenvalue trajectory* $(\boldsymbol{\lambda}(s))_{0\le s\le t_1} \in A$ *defined in Lemma* 3.3, *let* f *be a solution of the* d*-particle eigenvector moment flow* (2.16) *with initial matrix* $K(0)$. *For any* $C > 0$, *there exists* n_0 *such that for any* $n \ge n_0$ *the following holds. For any intervals* $J_{\text{in}} \subset \{i : \gamma_i(t_0) \in I_{3\kappa}^r(E_0)\}$, $J_{\text{out}} = \{i : d(i, J_{\text{in}}) \le nr/\psi\}$, *and* $[t, t+u] \subset [t_0, t_1]$ *with* $u > t/\psi$, *we have*

$$\begin{aligned}\mathrm{S}_{J_{\text{in}}}^{(t+\frac{u}{2},t+u)} \le \psi^3\left(\left(\frac{u}{r}\right)^{1/2} + \frac{1}{nt}\right)\mathrm{S}_{J_{\text{out}}}^{(t,t+u)} + \frac{\psi^3}{n^{\mathfrak{c}}}\big(\mathrm{S}_{J_{\text{out}}}^{(t,t+u)}\big)^{\frac{d-1}{d}}\\ + \frac{\psi^3}{nt}\big(\mathrm{S}_{J_{\text{out}}}^{(t,t+u)}\big)^{\frac{d-2}{d}} + n^{-C}.\end{aligned} \tag{3.24}$$

PROOF. For a general number of particles d, consider now the following modification of the eigenvector moment flow (2.16). We only keep the short-range dynamics (depending on the short-range parameter ℓ, chosen later) and modify the initial condition to be 0 when there is a particle far from J_{in}:

$$\begin{aligned}&\partial_s g_s = \mathscr{S}(s) g_s,\\ &g_t(\boldsymbol{\eta}) = (\operatorname{Av} f_t)(\boldsymbol{\eta}),\end{aligned} \quad t \le s \le t+u, \tag{3.25}$$

where

$$\operatorname{Av}(f) = \frac{3\psi}{nr}\sum_{\frac{1}{3}\frac{nr}{\psi} < a < \frac{2}{3}\frac{nr}{\psi}} \operatorname{Flat}_a(f).$$

We can write

$$\operatorname{Av}(f)(\boldsymbol{\eta}) = a_{\boldsymbol{\eta}} f(\boldsymbol{\eta})$$

for some coefficient $a_{\boldsymbol{\eta}} \in [0,1]$ ($a_{\boldsymbol{\eta}} = 0$ if $\boldsymbol{\eta} \not\subset J_{\text{out}}$, $a_{\boldsymbol{\eta}} = 1$ if $\boldsymbol{\eta} \subset J_{\text{in}}$). We will only use the elementary property

$$|a_{\boldsymbol{\eta}} - a_{\boldsymbol{\xi}}| \le \frac{\psi}{nr} d(\boldsymbol{\eta}, \boldsymbol{\xi}), \tag{3.26}$$

where the distance is defined in (3.11).

For any $\boldsymbol{\eta} \subset J_{\rm in}$, we have

$$\begin{aligned} &|f_s(\boldsymbol{\eta}) - g_s(\boldsymbol{\eta})| \\ &\quad\le |(\mathrm{U}_{\mathscr{B}}(t,s)f_t - \mathrm{U}_{\mathscr{S}}(t,s)f_t)(\boldsymbol{\eta})| + |\mathrm{U}_{\mathscr{S}}(t,s)(f_t - \mathrm{Av}\, f_t)(\boldsymbol{\eta})| \\ &\quad\le \psi^4 \frac{nu}{\ell}\left(\mathrm{S}^{(t,t+u)}_{J_{\rm out}} + \frac{1}{n^{\mathfrak{c}}}\big(\mathrm{S}^{(t,t+u)}_{J_{\rm out}}\big)^{\frac{d-1}{d}} + \frac{1}{\ell}\big(\mathrm{S}^{(t,t+u)}_{J_{\rm out}}\big)^{\frac{d-2}{d}}\right) + e^{-c\psi/2}, \end{aligned} \tag{3.27}$$

where we bounded the first term by Lemma 3.5, and the second term by finite speed of propagation (3.15), since $f_{t_0} - \mathrm{Av} f_{t_0}$ vanishes for any $\boldsymbol{\xi}$ such that $\boldsymbol{\xi} \subset \{i : d(i, J_{\rm in}) \le nr/3\psi\}$ (note that $\psi\ell \le nr/3\psi$).

In the following we will prove that for large enough n we have

$$\begin{aligned} &\sup_{\boldsymbol{\eta}\subset J_{\rm in}, t+\frac{u}{2}\le s\le t+u} g_s(\boldsymbol{\eta}) \\ &\quad\le \psi\left(\frac{nu}{\ell} + \frac{\ell}{nr} + \frac{\psi^2}{nt}\right)\mathrm{S}^{(t,t+u)}_{J_{\rm out}} \\ &\qquad + \frac{\psi}{n^{\mathfrak{c}}}\big(\mathrm{S}^{(t,t+u)}_{J_{\rm out}}\big)^{\frac{d-1}{d}} + \psi\left(\frac{nu}{\ell^2} + \frac{\psi^2}{nt}\right)\big(\mathrm{S}^{(t,t+u)}_{J_{\rm out}}\big)^{\frac{d-2}{d}} + n^{-C} \end{aligned} \tag{3.28}$$

by a maximum principle argument. Equations (3.27) and (3.28) together give the expected result (3.24) by choosing

$$\ell = n\psi^2(ur)^{1/2},$$

which satisfies (3.12) If the left-hand side of (3.28) is smaller than n^{-C}, there is nothing to prove. If it is greater than n^{-C} by the finite speed of propagation property (3.15) for any $t < s < t+u$, the configuration(s) $\widetilde{\boldsymbol{\eta}}$ such that

$$g_s(\widetilde{\boldsymbol{\eta}}) = \sup_{\boldsymbol{\eta}} g_s(\boldsymbol{\eta})$$

need to be supported in $\{i : d(i, J_{\rm in}) \le \frac{3}{4}\frac{nr}{\psi}\}$.

From the dynamics (3.25), for any parameter $\psi^4/n \le \eta \le \ell/n$ to be chosen, we have

$$\begin{aligned} \partial_t g_s(\widetilde{\boldsymbol{\eta}}) &= \sum_{0<|j-k|\le\ell} c_{jk} 2\widetilde{\eta}_j(1+2\widetilde{\eta}_k)\big(g_s(\widetilde{\boldsymbol{\eta}}^{jk}) - g_s(\widetilde{\boldsymbol{\eta}})\big) \\ &\le \frac{C}{n}\sum_{\substack{1\le p\le d,\\ k:0<|i_p-k|\le\ell}} \frac{g_s(\widetilde{\boldsymbol{\eta}}^{i_p k}) - g_s(\widetilde{\boldsymbol{\eta}})}{(\lambda_{i_p}-\lambda_k)^2 + \eta^2} \\ &= \frac{1}{n\eta}\sum_{\substack{1\le p\le d,\\ k:0<|i_p-k|\le\ell}} \mathrm{Im}\,\frac{g_s(\widetilde{\boldsymbol{\eta}}^{i_p k})}{z_{i_p} - \lambda_k} \\ &\quad - \frac{1}{n\eta} g_s(\widetilde{\boldsymbol{\eta}}) \sum_{\substack{1\le p\le d,\\ k:0<|i_p-k|\le\ell}} \Im \frac{1}{z_{i_p} - \lambda_k} \end{aligned} \tag{3.29}$$

where we define $z_{i_p} = \lambda_{i_p} + \mathrm{i}\eta$. For the second term in (3.29), note that

$$\sum_{\substack{1\le p\le d,\\ k:0<|i_p-k|\le\ell}} \Im\frac{1}{z_{i_p}-\lambda_k} \ge \sum_{p=1}^{d}\sum_{k:0<|i_p-k|\le\ell}\frac{\eta}{(\lambda_{i_p}-\lambda_k)^2+\eta^2} \ge \sum_{p=1}^{d}\sum_{k:|\lambda_k-\lambda_{i_p}|\le\eta}\frac{\eta}{2\eta^2}\gtrsim n,$$

where we used (3.16). For the first term in (3.29), we claim that for any fixed p we have

$$\begin{aligned}&\frac{1}{n}\sum_{k:0<|i_p-k|\le\ell}\mathrm{Im}\frac{g_s(\tilde{\boldsymbol{\eta}}^{i_pk})}{z_{i_p}-\lambda_k}\\ &\quad= \mathrm{O}(\psi)\Big(\frac{1}{n\eta}+\frac{\ell}{nr}+\frac{nu}{\ell}\Big)\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}+\mathrm{O}(\psi)\frac{1}{n^{\mathfrak{c}}}\big(\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}\big)^{\frac{d-1}{d}}\\ &\qquad+\mathrm{O}(\psi)\Big(\frac{1}{n\eta}+\frac{nu}{\ell^2}\Big)\big(\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}\big)^{\frac{d-2}{d}}.\end{aligned}\tag{3.30}$$

For this, we can bound the left-hand side of (3.30) by (3.31) + (3.32) + (3.33) where

$$\mathrm{Im}\sum_{k:0<|k-i_p|\le\ell}\frac{1}{n}\frac{(\mathrm{U}_{\mathscr{S}}(t,s)\,\mathrm{Av}\,f_t)(\tilde{\boldsymbol{\eta}}^{i_pk})-(\mathrm{Av}\,\mathrm{U}_{\mathscr{S}}(t,s)f_t)(\tilde{\boldsymbol{\eta}}^{i_pk})}{z_{i_p}-\lambda_k},\tag{3.31}$$

$$\mathrm{Im}\sum_{k:0<|i_p-k|\le\ell}\frac{1}{n}\frac{(\mathrm{Av}\,\mathrm{U}_{\mathscr{S}}(t,s)f_t)(\tilde{\boldsymbol{\eta}}^{i_pk})-(\mathrm{Av}\,\mathrm{U}_{\mathscr{B}}(t,s)f_t)(\tilde{\boldsymbol{\eta}}^{i_pk})}{z_{i_p}-\lambda_k},\tag{3.32}$$

$$\mathrm{Im}\sum_{k:0<|i_p-k|\le\ell}\frac{1}{n}\frac{(\mathrm{Av}\,\mathrm{U}_{\mathscr{B}}(t,s)f_t)(\tilde{\boldsymbol{\eta}}^{i_pk})}{z_{i_p}-\lambda_k}.\tag{3.33}$$

The term (3.31) will be controlled by finite speed of propagation; (3.32) will be controlled by Lemma 3.5, and (3.33) by the local law.

To bound (3.31), we write

$$\begin{aligned}&(\mathrm{U}_{\mathscr{S}}(t,s)\,\mathrm{Av}\,f_t)(\tilde{\boldsymbol{\eta}}^{i_pk})-(\mathrm{Av}\,\mathrm{U}_{\mathscr{S}}(t,s)f_t)(\tilde{\boldsymbol{\eta}}^{i_pk})\\ &\quad=\frac{2\psi}{nr}\sum_{\frac{nr}{2\psi}<a<\frac{nr}{\psi}}\big(\mathrm{U}_{\mathscr{S}}(t,s)\mathrm{Flat}_a f_t-\mathrm{Flat}_a\mathrm{U}_{\mathscr{S}}(t,s)f_t\big)(\tilde{\boldsymbol{\eta}}^{i_pk}).\end{aligned}$$

For fixed a, let $L_1\subset L_2$ be defined as $L_1=\{i:d(i,J_{\mathrm{in}})\le a-\psi\ell\}$, $L_2=\{i:d(i,J_{\mathrm{in}})\le a+\psi\ell\}$. We consider three cases: $\tilde{\boldsymbol{\eta}}^{i_pk}\not\subset L_2$, $\tilde{\boldsymbol{\eta}}^{i_pk}\subset L_1$, or neither.

For $\tilde{\boldsymbol{\eta}}^{i_pk}\not\subset L_2$, by our definition, $\mathrm{Flat}_a\mathrm{U}_{\mathscr{S}}(t,s)f_t(\tilde{\boldsymbol{\eta}}^{i_pk})=0$. By the finite speed of propagation (3.15), the total mass of $\mathrm{U}_{\mathscr{S}}(t,s)\mathrm{Flat}_a f_t$ outside L_2 is exponentially small. In particular, $|\mathrm{U}_{\mathscr{S}}(t,s)\mathrm{Flat}_a f_t(\tilde{\boldsymbol{\eta}}^{i_pk})|\le\exp(-c\psi/2)$.

For $\widetilde{\boldsymbol{\eta}}^{i_pk} \subset L_1$, we have

$$\begin{aligned}&\big|(\mathrm{U}_{\mathscr{S}}(t,s)\mathrm{Flat}_a f_t - \mathrm{Flat}_a \mathrm{U}_{\mathscr{S}}(t,s) f_t)(\widetilde{\boldsymbol{\eta}}^{i_pk})\big| \\ &\quad = \big|(\mathrm{U}_{\mathscr{S}}(t,s)\mathrm{Flat}_a f_t - \mathrm{U}_{\mathscr{S}}(t,s) f_t)(\widetilde{\boldsymbol{\eta}}^{i_pk})\big| \\ &\quad = \big|(\mathrm{U}_{\mathscr{S}}(t,s)(f_t - \mathrm{Flat}_a f_t))(\widetilde{\boldsymbol{\eta}}^{i_pk})\big| \le \exp(-c\psi/2).\end{aligned}$$

We used the finite speed of propagation (3.15) in the last inequality, since $f_t - \mathrm{Flat}_a f_t$ vanishes for any $\boldsymbol{\xi}$ supported in $\{i : d(i, J_{\mathrm{in}}) \le a\}$.

For the last case, we have $\widetilde{\boldsymbol{\eta}}^{i_pk} \subset L_2$, and some particle(s) of $\widetilde{\boldsymbol{\eta}}^{i_pk}$ is in L_2/L_1. There are at most $2n\psi\ell$ such a. Moreover, since $\mathrm{U}_{\mathscr{S}}$ is a contraction in L^∞, we have

$$\begin{aligned}&\big|(\mathrm{U}_{\mathscr{S}}(t,s)\mathrm{Flat}_a f_t - \mathrm{Flat}_a \mathrm{U}_{\mathscr{S}}(t,s) f_t)(\widetilde{\boldsymbol{\eta}}^{i_pk})\big| \\ &\quad \le |\mathrm{U}_{\mathscr{S}}(t,s)\mathrm{Flat}_a f_t| + \big|\mathrm{Flat}_a \mathrm{U}_{\mathscr{S}}(t,s)\mathrm{Flat}_{a+2\psi\ell} f_t(\widetilde{\boldsymbol{\eta}}^{i_p,k})\big| \\ &\qquad + \big|\mathrm{Flat}_a \mathrm{U}_{\mathscr{S}}(t,s)(f_t - \mathrm{Flat}_{a+2\psi\ell} f_t)(\widetilde{\boldsymbol{\eta}}^{i_pk})\big| \\ &\quad \le \|\mathrm{Flat}_a f_t\|_\infty + \|\mathrm{Flat}_{a+\psi\ell} f_t\|_\infty + e^{-c\psi/2}.\end{aligned}$$

We bound $\|\mathrm{Flat}_a f_t\|_\infty, \|\mathrm{Flat}_{a+2\psi\ell} f_t\|_\infty \le \mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}$. From these estimates, we have (3.31) $\le \psi^2 \frac{\ell}{nr} \mathrm{S}^{(t_0,t_0+u)}_{J_{\mathrm{out}}}$.

We now bound (3.32). For $|k - i_p| \le \ell$, $\widetilde{\boldsymbol{\eta}}^{i_pk}$ is supported in $\{i : \gamma_i(t_0) \in I^r_{2\kappa}(E)\}$, so that we can apply Lemma 3.5:

$$\begin{aligned}&\big|(\mathrm{Av}\,\mathrm{U}_{\mathscr{S}}(t,s) f_t)(\widetilde{\boldsymbol{\eta}}^{i_pk}) - (\mathrm{Av}\,\mathrm{U}_{\mathscr{B}}(t,s) f_t)(\widetilde{\boldsymbol{\eta}}^{i_pk})\big| \\ &\quad \le \big|(\mathrm{U}_{\mathscr{S}}(t,s) f_t - \mathrm{U}_{\mathscr{B}}(t,s) f_t)(\widetilde{\boldsymbol{\eta}}^{i_pk})\big| \\ &\quad \le \psi^4 \frac{nu}{\ell}\bigg(\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}} + \frac{1}{n^{\mathfrak{c}}}\big(\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}\big)^{\frac{d-1}{d}} + \frac{1}{\ell}\big(\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}\big)^{\frac{d-2}{d}}\bigg).\end{aligned}$$

As a consequence, we have

$$(3.32) \le \psi^4 \frac{nu}{\ell}\bigg(\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}} + \frac{1}{n^{\mathfrak{c}}}\big(\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}\big)^{\frac{d-1}{d}} + \frac{1}{\ell}\big(\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}\big)^{\frac{d-2}{d}}\bigg).$$

Finally, for (3.33), note that $\widetilde{\boldsymbol{\eta}}^{i_pk}$ is supported on J_{out}, so that

$$\begin{aligned}&\frac{1}{n}\,\mathrm{Im} \sum_{k:0<|i_p-k|\le\ell} \frac{(\mathrm{Av}\, f_t)(\widetilde{\boldsymbol{\eta}}^{i_pk})}{z_{i_p} - \lambda_k} \\ &\quad = \frac{1}{n}\,\mathrm{Im} \sum_{k:0<|i_p-k|\le\ell} \frac{a_{\widetilde{\boldsymbol{\eta}}} f_t(\widetilde{\boldsymbol{\eta}}^{i_pk}) + (a_{\widetilde{\boldsymbol{\eta}}^{i_pk}} - a_{\widetilde{\boldsymbol{\eta}}}) f_t(\widetilde{\boldsymbol{\eta}}^{i_pk})}{z_{i_p} - \lambda_k} \\ &\quad = \frac{a_{\widetilde{\boldsymbol{\eta}}}}{n}\,\mathrm{Im} \sum_{k:0<|i_p-k|\le\ell} \frac{f_t(\widetilde{\boldsymbol{\eta}}^{i_pk})}{z_{i_p} - \lambda_k} + \mathrm{O}\bigg(\psi \frac{\ell}{nr} \mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}\bigg),\end{aligned}$$

where we used that $|a_{\tilde{\boldsymbol{\eta}}^{i_pk}} - a_{\tilde{\boldsymbol{\eta}}}| \leq \psi d(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\eta}}^{i_pk})/(nr) \leq \psi\ell/(nr)$ from (3.26).

In the above imaginary part, the contribution of all $k \in \{i_1, \dots, i_d\}$ is of order $\frac{1}{n\eta}\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}$, so that (here k_0 is any index not in $\{i_1, \dots, i_d\}$)

$$
\begin{aligned}
&\frac{1}{n}\operatorname{Im}\sum_{k:0<|i_p-k|\leq\ell}\frac{f_t(\tilde{\boldsymbol{\eta}}^{i_pk})}{z_{i_p}-\lambda_k}\\
&\quad=\frac{1}{n}\frac{1}{\mathscr{M}(\tilde{\boldsymbol{\eta}}^{i_pk_0})}\operatorname{Im}\sum_{k:0<|i_p-k|\leq\ell}\sum_{G\in\mathscr{G}_{\tilde{\boldsymbol{\eta}}^{i_pk}}}\frac{\mathbb{E}(P(G)\mid\boldsymbol{\lambda})}{z_{i_p}-\lambda_k}+\mathrm{O}\biggl(\frac{1}{n\eta}\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}\biggr)\\
&\quad=(\mathrm{I})+(\mathrm{II})+\mathrm{O}\biggl(\frac{1}{n\eta}\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}\biggr)
\end{aligned}
$$

where (I) corresponds to perfect matchings for which $\{(k,1),(k,2)\}$ is not an edge, and (II) corresponds to perfect matchings for which $\{(k,1),(k,2)\}$ is an edge. More precisely,

$$
\begin{aligned}
(\mathrm{I})&=\operatorname{Im}\sum_{1\leq q_1,q_2\leq d}\mathbb{E}\biggl(P^{(q_1,q_2)}(p(e)_{e\in\mathscr{E}_{\tilde{\boldsymbol{\eta}}}})\sum_{k:0<|i_p-k|\leq\ell}\frac{p_{i_{q_1}k}p_{i_{q_2}k}}{n(z_{i_p}-\lambda_k)}\mid\boldsymbol{\lambda}\biggr),\\
(\mathrm{II})&=\operatorname{Im}\mathbb{E}\biggl(P^{(p)}((p(e)_{e\in\mathscr{E}_{\tilde{\boldsymbol{\eta}}}}))\sum_{k:0<|i_p-k|\leq\ell}\frac{p_{kk}}{n(z_{i_p}-\lambda_k)}\mid\boldsymbol{\lambda}\biggr),
\end{aligned}
$$

with $\mathscr{E}_{\tilde{\boldsymbol{\eta}}}$ the set of all possible edges between between vertices from $\mathscr{V}_{\tilde{\boldsymbol{\eta}}}$, $P^{(p,q)}$ is a finite sum of monic monomials of degree $n-2$, and $P^{(p)}$ is a finite sum of monic monomials of degree $n-1$.

To bound (I), we simply write

$$
\operatorname{Im}\sum_{k:0<|i_p-k|\leq\ell}\frac{p_{i_{q_1}k}p_{i_{q_2}k}}{n(z_{i_p}-\lambda_k)}=\mathrm{O}\biggl(\frac{1}{n\eta}\sum_k\bigl(p^2_{i_{q_1}k}+p^2_{i_{q_2}k}\bigr)\biggr)=\mathrm{O}\biggl(\frac{1}{n\eta}n^{\varepsilon}\frac{|I|}{n}\biggr).
$$

Here we slightly changed the meaning of p_{kk} (in both equations above and below, $p_{kk}=\sum_{\alpha\in I}u_k(\alpha)^2$, i.e., $C_0=0$ in (2.6)) and used the elementary identity (3.19). The above second equality follows from Lemma 3.1.

Moreover, with Lemma 3.6, we have

$$
\begin{aligned}
\mathbb{E}\bigl(|P^{(q_1,q_2)}(p(e)_{e\in\mathscr{E}_{\tilde{\boldsymbol{\eta}}}})|\mid\boldsymbol{\lambda}\bigr)&=\mathrm{O}\Bigl(\sup_{t\leq s\leq t+u,\boldsymbol{\eta}\subset J_{\mathrm{out}}}|f_s(\boldsymbol{\eta})|\mathbb{1}_{\mathscr{N}(\boldsymbol{\eta})=n-2}\Bigr)\\
&=\mathrm{O}\Bigl(\bigl(\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}\bigr)^{\frac{d-2}{d}}\Bigr),
\end{aligned}
$$

where we used Hölder's inequality and Lemma 3.6. This concludes our bound for (I), $\frac{1}{n\eta}\bigl(\mathrm{S}^{(t,t+u)}_{J_{\mathrm{out}}}\bigr)^{\frac{d-2}{d}}$.

More subtle bounds are required for the term (II).

$$\operatorname{Im} \sum_{k:0<|i_p-k|\leq \ell} \frac{p_{kk}}{n(z_{i_p}-\lambda_k)} = \mathrm{O}\left(\frac{1}{n\eta}+\frac{1}{n^{\mathfrak{c}}}\right) - \operatorname{Im} \sum_{k:|i_p-k|>\ell} \frac{p_{kk}}{n(z_{i_p}-\lambda_k)}$$

where we used (3.14). This last term can be bounded exactly as between (3.20) and (3.21), and we obtain

$$\operatorname{Im} \sum_{k:0<|i_p-k|\leq \ell} \frac{p_{kk}}{n(z_{i_p}-\lambda_k)} = \mathrm{O}\left(\frac{1}{n\eta}+\frac{1}{n^{\mathfrak{c}}}\right),$$

where we used that $\eta \leq \ell/n$. This concludes the proof of (3.30).

We define $h(s) = \sup_{\boldsymbol{\eta}} g_s(\boldsymbol{\eta})$. Equations (3.29) and (3.30) yield

$$\begin{aligned} h'(s) \leq \frac{C\psi}{\eta}\Biggl(\left(\frac{1}{n\eta}+\frac{\ell}{nr}+\frac{nu}{\ell}\right)\mathrm{S}_{J_{\mathrm{out}}}^{(t,t+u)} + \frac{1}{n^{\mathfrak{c}}}\left(\mathrm{S}_{J_{\mathrm{out}}}^{(t,t+u)}\right)^{\frac{d-1}{d}} \\ + \left(\frac{1}{n\eta}+\frac{nu}{\ell^2}\right)\left(\mathrm{S}_{J_{\mathrm{out}}}^{(t,t+u)}\right)^{\frac{d-2}{d}}\Biggr) - c\frac{h(s)}{\eta} \end{aligned}$$

for any $t < s < t+u$. We now choose $\eta = t/\psi^2$, so that $u/\eta > \psi$, and obtain

$$\begin{aligned} h(s) < C\psi\left(\frac{1}{n\eta}+\frac{\ell}{nr}+\frac{nu}{\ell}\right)\mathrm{S}_{J_{\mathrm{out}}}^{(t,t+u)} + C\frac{\psi}{n^{\mathfrak{c}}}\left(\mathrm{S}_{J_{\mathrm{out}}}^{(t,t+u)}\right)^{\frac{d-1}{d}} \\ + C\psi\left(\frac{1}{n\eta}+\frac{nu}{\ell^2}\right)\left(\mathrm{S}_{J_{\mathrm{out}}}^{(t,t+u)}\right)^{\frac{d-2}{d}} + n^{-C} \end{aligned}$$

for any $t+u/2 < s < t+u$, which is (3.28) and concludes the proof. □

PROOF OF THEOREM 2.5. We proceed by iterating the bound from Proposition 3.7. We are given a small ε such that $\varepsilon < \mathfrak{a}/5$ and a large $D > 0$, as in the statement of Theorem 2.5.

We first choose $d = \lfloor 5D/\varepsilon \rfloor$ and define (implicitly, for J_{i+1})

$$\begin{cases} s_0 = t_0, \\ s_{i+1} = \frac{s_i+t_1}{2}, \end{cases} \qquad \begin{cases} J_0 = \{i : \gamma_i(t_0) \in I_{3\kappa}^r(E_0)\}, \\ J_i = \{i : d(i, J_{i+1}) \leq \frac{nr}{\psi}\}. \end{cases}$$

A direct application of Proposition 3.7 together with the bounds $n^{-1+\mathfrak{a}} \leq t_0 \leq t_1 \leq n^{-\mathfrak{a}} r$ yields

$$\begin{aligned} \mathrm{S}_{J_{i+1}}^{(s_{i+1},t_1)} \leq \psi^3\left(n^{-\mathfrak{a}/2}+2^i n^{-\mathfrak{a}}\right)\mathrm{S}_{J_i}^{(s_i,t_1)} \\ + \frac{\psi^3}{n^{\mathfrak{c}}}\left(\mathrm{S}_{J_{\mathrm{out}}}^{(s_i,t_1)}\right)^{\frac{d-1}{d}} + \frac{\psi^3 2^i}{nt_0}\left(\mathrm{S}_{J_i}^{(s_i,t_1)}\right)^{\frac{d-2}{d}} + n^{-C}. \end{aligned}$$

In particular, we have

$$\mathrm{S}_{J_{i+1}}^{(s_{i+1},t_1)} \leq n^{-\varepsilon/3}\mathrm{S}_{J_i}^{(s_i,t_1)}$$

provided that

$$\left(\mathrm{S}_{J_i}^{(s_i,t_0)}\right)^{1/d} \geq \frac{n^{\varepsilon/2}2^i}{\sqrt{nt_1}} + \frac{n^{\varepsilon/2}}{n^{\mathfrak{c}}}.$$

This implies that for $k = \lfloor 4\varepsilon^{-1} \rfloor$ we have

$$\left|\left(\mathrm{S}_{J_k}^{(s_k,t_1)}\right)^{1/d}\right| \le \frac{n^{3\varepsilon/4}}{\sqrt{nt_0}} + \frac{n^{3\varepsilon/4}}{n^{\mathfrak{c}}}.$$

For each fixed i, by choosing $\boldsymbol{\eta}$ as the configuration with $d = 2m$ particles on the site i and no particles elsewhere, we have $|p_{ii}(s)|^d \le C_d f_s(\boldsymbol{\eta})$. Hence by Markov's inequality, the last displayed equation implies that

$$\begin{aligned}\mathbb{P}&\left(\exists s_k < t < t_1 : 1_{\lambda_i(t) \in I_{4\kappa}^r(E_0)} |p_{ii}| \ge n^{\varepsilon}\left(\frac{1}{n^{\mathfrak{c}}} + \frac{1}{\sqrt{nt_0}}\right)\right)\\ &\le (n^{\varepsilon})^{-d}(n^{3\varepsilon/4})^d \le n^{-D}.\end{aligned}$$

Here we used that $\{i : \gamma_i(t_0) \in I_{4\kappa}^r(E_0)\} \subset J_k$ because $k\frac{nr}{\psi} < \kappa r$ for $k = \lfloor 4\varepsilon^{-1} \rfloor$ and n large enough.

Finally, by Lemma 3.6, p_{ij}^d can be estimated in terms f_t, p_{ii}, and p_{jj}. Hence the previous estimate also holds if we replace

$$1_{\lambda_i(t) \in I_{4\kappa}^r(E_0)} |p_{ii}| \text{ by } 1_{\lambda_i(t), \lambda_j(t) \in I_{4\kappa}^r(E_0)} |p_{ij}|.$$

This concludes the proof of the theorem, up to redefining t_0 and κ by a constant factor. □

4 Mean-Field Reduction

This section proves Theorem 1.5. We actually just need to prove it when a tiny GOE regularization is added, as explained in the next paragraph.

4.1 Small Regularization

Consider matrices of type

$$(4.1) \quad H = H_1 + H_2 + N^{-A} H^{\mathrm{G}} \quad \text{where } H_{ij}^{\mathrm{G}} \overset{\text{(d)}}{=} (1 + 1_{ij})^{1/2} \cdot \mathscr{N}(0, N^{-1}),$$

where H_1 and H_2 are defined in (1.15). Our main result in this section is the following:

THEOREM 4.1. *Let $A > 10$ be any fixed constant. Assume that H is a band matrix of type* (4.1)*, with bandwidth W_N satisfying (*1.11*).*

(i) *The eigenvectors are delocalized as in* (1.12).
(ii) *The eigenvalues satisfy the local semicircle law as in* (1.13).
(iii) *Fixed energy universality holds as in* (1.14).

(iv) *For any* (*small*) $\tau, \kappa > 0$ *and* (*large*) $D > 0$, *there exists* $N_0 > 0$ *such that for any* $N \geq N_0$ *we have*

$$\mathbb{P}\left(\left|\frac{N}{W}\sum_{\alpha=\ell}^{\ell+W}|\psi_j(\alpha)|^2 - 1\right| < N^{-\frac{3}{2}a+\tau} \text{ for all } 1 \leq j, \ell \leq N \text{ such that } |\lambda_j| \leq 2-\kappa\right) \geq 1 - N^{-D}, \tag{4.2}$$

where $a > 0$ *was given in* (1.11) *and all indices are defined modulo* N.

The same results hold for all submatrices of H *of type* $H^{(k)} = (H_{ij})_{i,j\in[\![1,N]\!]\setminus\{k\}}$.

The following simple lemma shows that all properties of delocalization only need to be established for the slightly regularized matrices. It is proved by perturbative arguments.

LEMMA 4.2. *Theorem* 4.1 *implies Theorem* 1.5.

PROOF. Let $H' = H_1 + H_2$ have distribution (1.15) and $H = H' + N^{-A}H^{\mathrm{G}}$, with respective ordered eigenvalues and eigenvectors λ'_k, $\boldsymbol{\psi}'_k$, λ_k, $\boldsymbol{\psi}_k$. Let $\mathscr{A} = \{\|H^{\mathrm{G}}\|_\infty \leq N^{-A-1/2+\varepsilon}\}$. By Gaussian decay of the entries of H^{G}, for any $\varepsilon, C > 0$, for large enough N we have

$$\mathbb{P}(\mathscr{A}) \geq 1 - N^{-C}. \tag{4.3}$$

The conclusions (ii) and (iii) of Theorem 1.5 for H' therefore follow from the Hoffman-Wieland inequality:

$$\begin{aligned}\sup_k |\lambda_k - \lambda'_k|1_{\mathscr{A}} &\leq N^{1/2}\Big(\sum_k |\lambda_k - \lambda'_k|^2\Big)^{1/2}1_{\mathscr{A}} \\ &\leq N^{1/2}(\mathrm{Tr}(H'-H)^2)^{1/2}1_{\mathscr{A}} \leq N^{-A+3}.\end{aligned} \tag{4.4}$$

Moreover, the conclusion (i) of Theorem 1.5 also hold for H'. Indeed, we have $\eta^{-1}|\psi'_k(i)|^2 \leq \operatorname{Im} G'_{ii}(\lambda'_k + \mathrm{i}\eta)$ and the simple inequality

$$\|(H'-z)^{-1}\|_\infty = \|(H-z)^{-1}\|_\infty + \mathrm{O}\left(\frac{N^2}{\eta^2}\|H'-H\|_\infty\right)$$

obtained by resolvent expansion. From the local law and eigenvector delocalization for H, for any $z = E + \mathrm{i}\eta$, $\eta > N^{-1+\varepsilon}$, $E \in [-2+\kappa, 2-\kappa]$, for any $D > 0$ we have $\mathbb{P}(\|(H-z)^{-1}\|_\infty \leq N^{\varepsilon}) \geq 1 - N^{-D}$ for some $C > 0$ for large enough N. Moreover, on $\mathscr{A}$ we have $\frac{N^2}{\eta^2}\|H'-H\|_\infty \leq N^{-2}$, which concludes the proof of (i) for H'.

The proof of (iv) is more involved. We want to obtain (4.2) for H' for a given large $D > 0$. Take $A = 4D$ in (4.1) and denote $t = N^{-A}$. The perturbation formula for the $\boldsymbol{\psi}_k(s)$'s, eigenvectors of $H' + sH^{\mathrm{G}}$ associated to eigenvalues $\lambda_k(s)$'s, is

$$\frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{\psi}_k(s) = \sum_{\ell \neq k} \frac{\langle \boldsymbol{\psi}_\ell(s), H^{\mathrm{G}}\boldsymbol{\psi}_k(s)\rangle}{\lambda_k(s) - \lambda_\ell(s)}\boldsymbol{\psi}_\ell(s).$$

On $\mathscr{A}$, we therefore have

$$\|\boldsymbol{\psi}_k - \boldsymbol{\psi}'_k\|_\infty \leq N^2 \int_0^t \mathrm{d}s\left(\frac{1}{|\lambda_k(s) - \lambda_{k+1}(s)|} + \frac{1}{|\lambda_k(s) - \lambda_{k-1}(s)|}\right). \tag{4.5}$$

Consider eigenvalues $\lambda_k < \lambda_{k+1}$ for H, with $\lambda_k^{(i)} \in (\lambda_k, \lambda_{k+1})$ an eigenvalue for the minor $H^{(i)}$, with associated normalized eigenvector $\boldsymbol{\psi}_k^{(i)}$. Denote

$$\mathscr{A}_k^{(i)} = \left\{\sum_{|\alpha - i| < W} |\psi_k^{(i)}(\alpha)^2| > \frac{W}{10N}\right\}.$$

By QUE for $H^{(i)}$, for any $C > 0$, for large enough N we have

$$\mathbb{P}\Big(\bigcap_{i,k:\lambda_k \in [-2+\kappa, 2-\kappa]} \mathscr{A}_k^{(i)}\Big) \geq 1 - N^{-C}. \tag{4.6}$$

By a Schur complement as in [25, sec. 4], for any $\delta > 0$ we have

$$\begin{aligned}&\mathbb{P}\big(\{|\lambda_{k+1} - \lambda_k| < \delta\} \cap \mathscr{A}_k^{(1)} \cap \cdots \cap \mathscr{A}_k^{(N)}\big)\\ &\quad \leq N\mathbb{P}\big(\{|\langle \widetilde{H}^{(1)}, \boldsymbol{\psi}_k^{(1)}\rangle| < \delta\sqrt{N}\} \cap \mathscr{A}_k^{(1)}\big)\end{aligned}$$

where $\widetilde{H}^{(i)} = (H_{ij})_{j \neq i}$. Take $\delta = N^{-2D}$. On $\mathscr{A}_k^{(1)}$, $\langle \widetilde{H}^{(1)}, \boldsymbol{\psi}_k^{(1)}\rangle$ is a random variable with density bounded by N^2, so that

$$\mathbb{P}\big(\{|\lambda_{k+1} - \lambda_k| < \delta\} \cap \mathscr{A}_k^{(1)} \cap \cdots \cap \mathscr{A}_k^{(N)}\big) \leq N^{-2D+4}.$$

Moreover, similarly to (4.4), we have $\sup_{0 \leq s \leq t} |\lambda_k(t) - \lambda_k(s)| \mathbb{1}_{\mathscr{A}} \leq N^{-A+3}$, which together with the previous equation gives

$$\begin{aligned}&\mathbb{P}\big(\{|\lambda_{k+1}(s) - \lambda_k(s)| < \delta \text{ for some } 0 < s < t\} \cap \mathscr{A}_k^{(1)} \cap \cdots \cap \mathscr{A}_k^{(N)} \cap \mathscr{A}\big)\\ &\quad \leq N^{-2D+4} + N^{-A+3}.\end{aligned} \tag{4.7}$$

From equations (4.3), (4.5), (4.6), and (4.7), for any $C > 0$ we have, for large enough N,

$$\mathbb{P}\big(\|\boldsymbol{\psi}_k - \boldsymbol{\psi}'_k\|_\infty < N^{-A+2+2D}\big) \geq 1 - N^{-2D+4} - N^{-A+3} - N^{-C}.$$

This concludes the proof of QUE for $\boldsymbol{\psi}'_k$, knowing QUE for $\boldsymbol{\psi}_k$. □

4.2 Notations

We now explain the ideas for the proof of Theorem 4.1. We start with the following definition, which generalizes band matrices by allowing diagonal perturbations.

DEFINITION 4.3 (Definition of H_ζ^{g}). For any positive constant ζ and any $\mathrm{g} \in \mathbb{R}^N$, H_ζ and H_ζ^{g} will denote real symmetric $N \times N$ matrices satisfying the following properties.

The matrix H_ζ is centered, it has independent entries up to the symmetry condition that satisfy (1.8) and (1.9) and is of the form

$$H_\zeta = \begin{pmatrix} A_\zeta & B^* \\ B & D \end{pmatrix}, \tag{4.8}$$

where A_ζ is a $W \times W$ matrix and

$$\operatorname{Var}((H_\zeta)_{ij}) = (s_\zeta)_{ij} = s_{ij} - \frac{\zeta(1+\delta_{ij})}{W} 1_{i,j \in [\![1,W]\!]},$$

where $s_{ij} = f(i-j)$ and $\sum_{x \in \mathbb{Z}_N} f(x) = 1$.

The matrix H_ζ^{g} is defined by

$$\big(H_\zeta^{\mathrm{g}}\big)_{ij} := (H_\zeta)_{ij} - g_i\delta_{ij}, \quad H_\zeta^{\mathrm{g}} =: \begin{pmatrix} A_\zeta^{\mathrm{g}} & B^* \\ B & D^{\mathrm{g}} \end{pmatrix}, \quad \mathrm{g} = (g_1, g_2, \dots, g_N). \tag{4.9}$$

We denote the eigenvalues and eigenvectors of H_ζ^{g} by λ_k^{g} and

$$\boldsymbol{\psi}_k^{\mathrm{g}} = \begin{pmatrix} \mathbf{w}_k^{\mathrm{g}} \\ \mathbf{v}_k^{\mathrm{g}} \end{pmatrix} \quad \text{where } \mathbf{w}_k^{\mathrm{g}} \in \mathbb{R}^W.$$

In the special case $g_j = g 1_{j>W}$, we will denote H_ζ^{g} by H_ζ^{g}, and for $\zeta = 0$ we abbreviate H_ζ^{g} (resp., H_ζ^{g}) by H^{g} (resp., H^{g}).

In fact, the matrices H^{g} we consider will always be of type H^{g}, up to a translation of the basis indices mod N.

We now define some curves, illustrated in Figure 4.1. The eigenvector equation $H^{\mathrm{g}}\boldsymbol{\psi}_k^{\mathrm{g}} = \lambda_k^{\mathrm{g}}\boldsymbol{\psi}_k^{\mathrm{g}}$ immediately implies that

$$(A^{\mathrm{g}} - B^{\mathrm{g},*}(D^{\mathrm{g}} - \lambda_k^{\mathrm{g}})^{-1}B^{\mathrm{g}})\mathbf{w}_k^{\mathrm{g}} = \lambda_k^{\mathrm{g}}\mathbf{w}_k^{\mathrm{g}}.$$

Hence we will consider the eigenvector equation

$$Q_e^{\mathrm{g}}\mathbf{u}_k^{\mathrm{g}}(e) = \xi_k^{\mathrm{g}}(e)\mathbf{u}_k^{\mathrm{g}}(e), \quad Q_e^{\mathrm{g}} := (A^{\mathrm{g}} - B^{\mathrm{g},*}(D^{\mathrm{g}} - e)^{-1}B^{\mathrm{g}}), \tag{4.10}$$

where $\xi_k^{\mathrm{g}}(e)$ and $\mathbf{u}_k^{\mathrm{g}}(e)$ are eigenvalues and normalized eigenvectors. From now on, we assume that k is an index in the bulk of the spectrum for H^{g}, i.e., for some $\kappa > 0, \kappa N < k < (1-\kappa)N$.

Since the matrix elements have Gaussian components (4.1), it is easy to check that the eigenvalue flows $g \to \lambda_k^{g}$ are smooth and nonintersecting with probability 1. Assuming that the function $g \to e = \lambda_k^{g} + g$ has a regular inverse (for the existence of such an inverse, see Section 4.7), for any e close enough to λ_k, there exists a g such that $e = \lambda_k^g + g$, so that we can define

$$\mathscr{C}_k(e) = \lambda_k^g.$$

The curves $(\mathscr{C}_k(e))_{1\leq k\leq N}$are labeled in increasing order by their intersections with the diagonal $\mathscr{C}(e) = e$. We refer to [4, equation (4.16)] for a detailed discussion of the domain of $\mathscr{C}_k$.

We defined $\xi_i(e)$ $(1 \leq i \leq W)$, the eigenvalue of $Q_e = Q_e^{\mathrm{g}=0}$. A simulation of the curves $e \to \xi(e)$ is given in Figure 4.1. Since $\xi_i(e)$ is also an eigenvalue of $H^{e-\xi_i(e)}$, it is equal to $\mathscr{C}_j(e)$ for some j. We follow the convention in [4] to denote $k' \in [\![1, W]\!]$ to be the index given by the relation $\xi_{k'}(e) = \mathscr{C}_k(e)$. Here $k' = k'(e)$ depends on the energy e, and $\xi_{k'(e)}(e)$ is increasing in k. As e approaches an eigenvalue of D, one eigenvalue from $\boldsymbol{\xi}(e)$ tends to $\pm\infty$. The other eigenvalues follow the smooth curves $\mathscr{C}_k$ and the labels $k'(e)$ gets shifted by ± 1 whenever e crosses an eigenvalue of D. Since the curve $\mathscr{C}_k$ passes through (λ_k, λ_k), we have

$$\begin{gathered} H\boldsymbol{\psi}_k = \lambda_k \boldsymbol{\psi}_k, \quad \xi_{k'}(\lambda_k) = \lambda_k, \quad \boldsymbol{\psi}_k = \begin{pmatrix} \mathbf{w}_k \\ \mathbf{v}_k \end{pmatrix}, \\ Q_{\lambda_k}\mathbf{u}_{k'} = \xi_{k'}\mathbf{u}_{k'}, \quad \mathbf{u}_{k'}(\lambda_k) = \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|_2}. \end{gathered} \tag{4.11}$$

4.3 Outline of Proof of Theorem 4.1

We explain the main steps of the proof, with QUE for mean-field blocks, QUE for H from (4.1), and its application to local law, universality, and delocalization.

Step 1. *QUE for mean-field blocks* Q_e^{g}. Remember the definition of H from (4.1) and denote $\widetilde{H} = (1 + N^{-A}\frac{N+1}{N})^{-1/2}H$. Consider a parameter $\zeta = T = N^{-c}$ where $c > \varepsilon_m$ is defined in (4.22). Then, thanks to the Gaussian matrix H_2 defining H, we can write

$$\widetilde{H} = H_T + \sqrt{T}\begin{pmatrix} H_W^{\mathrm{G}} & 0 \\ 0 & 0 \end{pmatrix}$$

for some H_T as defined in (4.8), and H_W^{G} is a $W \times W$ GOE matrix. To this H_T we associate H_T^{g} from formula (4.9), and denote $V = A_T^{\mathrm{g}} - B^{\mathrm{g},*}(D^{\mathrm{g}} - e)^{-1}B^{\mathrm{g}}$. Consider the flow

$$K_T^{\mathrm{g},e}(t) = V + Z(t) \tag{4.12}$$

as in (2.1). Notice that we have the equality in distribution

$$K_T^{\mathrm{g},e}(t) \overset{(\mathrm{d})}{=} K_{T-t}^{\mathrm{g},e}(0) = A_{T-t}^{\mathrm{g}} - B^{\mathrm{g},*}(D^{\mathrm{g}} - e)^{-1}B^{\mathrm{g}}. \tag{4.13}$$

In particular, the distributions of $Q_e^{\mathrm{g}} = A^{\mathrm{g}} - B^{\mathrm{g},*}(D^{\mathrm{g}} - e)^{-1}B^{\mathrm{g}}$ from (4.10) are the same as for $K_T^{\mathrm{g},e}(T)$.

We therefore obtain QUE for the mean-field blocks Q_e^{g} by using Theorem 2.5, i.e., by interpreting this matrix ensemble as the result of the flow $K(T) = K_T^{\mathrm{g},e}(T)$. As an hypothesis for Theorem 2.5, some estimates on $V = K(0)$ are necessary and given in Section 4.4.

Step 2. *QUE for* H^{g}. To simplify the notations we set $\mathrm{g} = 0$, but QUE will be obtained similarly for any small enough g. For the proof, we combine an ε-net argument with perturbations of eigenvectors.

For this, we first need to choose good points for our net. Let $M = N^C$ with C a large constant that will be chosen in the rigorous proof. We will prove that there is another large number C' such that for each $n \in \mathbb{Z}$ fixed such that $E_n = nN^{-C'} \in [-2+\kappa, 2-\kappa]$, there is a deterministic $e_n \in [E_n, E_{n+1}]$ (i.e., the choice of e_n may depend on the law of D but is independent of the random matrix elements of D)

$$\inf_j \left|\lambda_j^D - e_n\right| \geq M^{-1}$$

with high probability, where the λ_j^D's are eigenvalues of D (recall that λ_j denotes an eigenvalue of H). In other words, the bulk eigenvalues of D will stay away from the grid points $(e_n)_{n\in\mathbb{Z}}$ by at least N^{-C}: the norm of Q_{e_n} is polynomially bounded, a hypothesis necessary to prove QUE by flow methods.

We now consider QUE for these good points $(e_n)_{n\in\mathbb{Z}}$. Let J be the $W \times W$ matrix defined by

$$(J)_{ij} = \delta_{ij} \cdot 1_{1\leq i\leq W/2}. \tag{4.14}$$

By the QUE property for mean-field blocks (see Lemma 4.8) for all n and l satisfying $|\xi_l(e_n) - e_n| \leq W^{-1}$ we have

$$\left| \|J \cdot \mathbf{u}_l(e_n)\|_2^2 - \frac{1}{2} \right| \leq \frac{N^{1/2+\tau}}{W} + \frac{N^{\frac{\varepsilon_m}{2}+\tau}}{W^{1/2}} \tag{4.15}$$

with overwhelming probability, where $\tau > 0$ is an arbitrarily small positive constant and ε_m is defined in (1.8).

For a given bulk index k, let $\widetilde{e} = \sup_n\{e_n : e_n < \lambda_k\}$. Recall that $\mathscr{C}_k(\lambda_k) = \lambda_k$ and $k' \in \llbracket 1, W \rrbracket$ is the index given by the relation $\xi_{k'}(e) = \mathscr{C}_k(e)$ for all e, as explained in Section 4.2. By the eigenvector perturbation formula for the matrix Q_e, we have

$$\frac{\mathrm{d}}{\mathrm{d}e}\mathbf{u}_{k'}(e) = \sum_{\ell\neq k} \frac{\mathbf{u}_{\ell'}(e)}{\mathscr{C}_k(e) - \mathscr{C}_\ell(e)} \left(\mathbf{u}_{\ell'}(e), B^* \frac{1}{(D-e)^2} B \ \underline{\mathrm{u}}_{k'}(e) \right). \tag{4.16}$$

Notice that we used the labeling associated with the curve $\mathscr{C}$ since $\mathscr{C}_k(e)$ is continuous, i.e., the label k, ℓ does not change as e pass through the eigenvalues of D. However, the label k' for the eigenvector depends on e.

Our goal is to approximate $\mathbf{u}_{k'}(\lambda_k)$, which is proportional to the first W components of the eigenvector $\boldsymbol{\psi}_k$ of H, by the eigenvector $\mathbf{u}_{k'}(\widetilde{e})$ which satisfies the

QUE by (4.15). Integration gives

$$\|\mathbf{u}_{k'}(\widetilde{e}) - \mathbf{u}_{k'}(\lambda_k)\| \leq \tag{4.17}$$
$$N^{-C'} \sup_{\lambda_k \leq e \leq \widetilde{e}} \sum_{\ell \neq k} \frac{1}{|\mathscr{C}_k(e) - \mathscr{C}_\ell(e)|} \left| \left(\mathbf{u}_{\ell'}(e), B^* \frac{1}{(D-e)^2} B \, \mathbf{u}_{k'}(e) \right) \right|.$$

We will show that for some $C_1 > 0$ the following two estimates hold with high probability.

(i) Level repulsion: for fixed k we have

$$\min_{e \in [\widetilde{e}, \lambda_k]} \min_{\ell : \ell \neq k} |\mathscr{C}_k(e) - \mathscr{C}_\ell(e)| \geq N^{-C_1/2}.$$

(ii) A consequence of the weak uncertainty principle from Section 4.6,

$$\sup_{\lambda_k \leq e \leq \widetilde{e}} \left| \left(\mathbf{u}_{\ell'}(e), B^* \frac{1}{(D-e)^2} B \, \mathbf{u}_{k'}(e) \right) \right| \leq N^{C_1/2}.$$

If these two bounds hold then (4.17) gives stability of the eigenvector under perturbation in e, provided that $C' \gg C_1$. Delocalization and QUE of $\mathbf{u}_{k'}(\widetilde{e})$ therefore imply the same properties for $\mathbf{u}_{k'}(\lambda_k)$.

Thus, denoting by ε_N the right-hand side of (4.15) and

$$X_n = \sum_{1 \leq i \leq W/2} |\psi_k(i + nW/2)|^2,$$

we have

$$\frac{X_1}{X_2} = 1 + \mathrm{O}(\varepsilon_N) \tag{4.18}$$

with overwhelming probability. Now we can shift the indices by $W/2$ and repeat the same argument, so that for any $1 \leq \ell < m \leq 2N/W$, we have

$$\frac{X_\ell}{X_m} = (1 + \mathrm{O}(\varepsilon_N))^{m-\ell} = 1 + \mathrm{O}\left(\frac{N}{W}\varepsilon_N\right).$$

provided that $\frac{N}{W}\varepsilon_N = \mathrm{o}(1)$. Summing over ℓ for fixed m gives, with overwhelming probability,

$$\frac{N}{W} X_m = \frac{1}{2} + \mathrm{O}\left(\frac{N}{W}\varepsilon_N\right). \tag{4.19}$$

This concludes the outline that QUE for the eigenvector $\boldsymbol{\psi}_k$ holds, when $\frac{N}{W}\varepsilon_N = \mathrm{o}(1)$.

Step 3. *Applications of QUE.* We successively outline the proofs of delocalization, universality, and local law for H from (4.1).

Delocalization for the mean-field blocks Q_e^{g} holds thanks to a priori resolvent estimates from Section 4.4, and regularization of the resolvent by Dyson Brownian motion, as in (3.13). By stability as in (4.17), this delocalization is extended to $\mathbf{u}_{k'}(\lambda_k)$. As

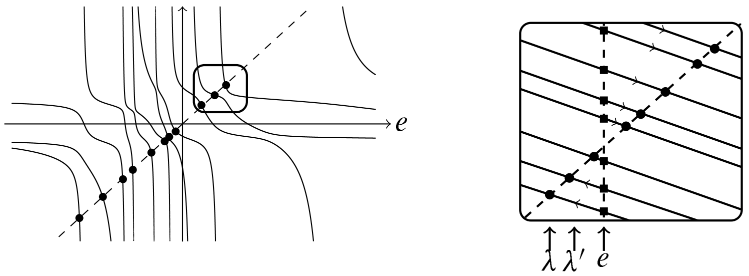
$$((u_{k'}(\lambda_k))(i))_{1\leq i\leq W} = (\psi_k(i))_{1\leq i\leq W}/\|\boldsymbol{\psi}_k\|_{\mathrm{L}^2([\![1,W]\!])},$$

delocalization for $\boldsymbol{\psi}_k$ follows from both delocalization of $\mathbf{u}_{k'}(\lambda_k)$ to the QUE estimate (4.19) about $\|\boldsymbol{\psi}_k\|_{\mathrm{L}^2([\![1,W]\!])}$.

For universality, remember that for any e, $\mathscr{C}_k(e) = \xi_{k'}(e)$ denotes the eigenvalues of Q_e, and that the intersection points of the curves $e \to \xi_k(e)$ with the diagonal $e = \xi$ are eigenvalues for H (see Figure 4.1). Thus λ_j can be determined by the spectrum $\boldsymbol{\xi}(e)$ for a fixed e, and the slope of the curves $e \to \xi_{k'}(e)$. On the one hand, $\boldsymbol{\xi}(e)$ follows GOE statistics as a consequence of [22]. On the other hand, a simple calculation yields

$$\partial_e \mathscr{C}_k(e) = 1 - \frac{1}{\sum_{i=1}^{W} |\psi_k^g(i)|^2},$$

where $\boldsymbol{\psi}^g$ is the corresponding eigenvector of H^g with g the solution to $e = \lambda_k^g + g$. From the QUE (4.19) for H^g, all slopes are equal at leading order, so that the statistics of λ_j will be given by those of ξ_k up to some trivial scaling. In the same way, the local law for H follows from a local law for Q_e by parallel projection.



(A) A simulation of eigenvalues of $Q_e = A - B^*(D-e)^{-1}B$, i.e., functions $e \mapsto \xi_j(e)$. Here $N = 12$ and $W = 3$. The λ_i's are the abscissas of the intersections with the diagonal.

(B) Framed region of Figure (a) for large N, W: the curves ξ_j are almost parallel, with slope $1 - N/W$. The eigenvalues of Q_e and H are related by a projection to the diagonal.

FIGURE 4.1. The idea of mean-field reduction, from [4]: universality of gaps between eigenvalues for fixed e implies universality on the diagonal through parallel projection. For e fixed, we label the curves by $\xi_k(e)$.

4.4 Generalized Resolvent Estimates

In this subsection, we do not need to assume (1.8).

Recall that we have added a GOE regularization of size N^{-A} in (4.1). Since N^{-A} is tiny, all resolvent estimates cited in this paper for matrices are valid after adding this small regularizing GOE. A formal proof can be obtained by the standard resolvent identity $(B-C)^{-1} = B^{-1} + B^{-1}CB^{-1} + \cdots$, which we will skip. In this section, all results will be proved without this regularization so as to simplify the notations.

Our first goal is to show that $K^{\mathrm{g},e}_{T,t}$ (4.13) is (η_*, η^*, r)-regular at $e = E_0$, in the sense of Assumption 2.3, for some range of t. The precise choice of the parameters r, T, η_*, η^* will be given in (4.22). Recall the matrix $H^{\mathrm{g}}_{T,t}$ is defined by

$$H^{\mathrm{g}}_{T,t} = \begin{pmatrix} A^{\mathrm{g}}_T + Z_t & B^* \\ B & D^{\mathrm{g}} \end{pmatrix}$$

As in (1.4), define the “generalized resolvent" of $H^{\mathrm{g}}_{T,t}$ by

$$G^{\mathrm{g}}_{T,t}(z,e) = \left(H^{\mathrm{g}}_{T,t} - \begin{pmatrix} z\mathrm{I}_W & 0 \\ 0 & e\mathrm{I}_{N-W} \end{pmatrix} \right)^{-1}.$$

The distribution of $H^{\mathrm{g}}_{T,t}$ is the same as H^{g}_{T-t} defined in (4.9), so we will also denote $G^{\mathrm{g}}_{T,t}(z,e)$ by $G^{\mathrm{g}}_{T-t}(z,e)$.

Clearly, the $W \times W$ component of $G^{\mathrm{g}}_{T,t}(z,e)$ is the resolvent $(K^{\mathrm{g},e}_{T-t} - z)^{-1}$. We will state estimates on this generalized resolvent in Theorem 4.5, an important input for our mean-field reduction method. The proof appears in the companion papers [6, 37]. On the one hand, the absence of imaginary part on most of the diagonal elements of the generalized resolvent is a major obstacle to estimate it. On the other hand, Theorem 4.5 assumes $\eta = \operatorname{Im} z$ is large (almost of order 1), which is a sufficient input to apply Theorem 2.5 and obtain quantum unique ergodicity.

Define $M^{\zeta,\mathrm{g}}_i(z,\widetilde{z})$ as the solution of the self consistent equation

$$\left(M^{\zeta,\mathrm{g}}_i\right)^{-1}(z,\widetilde{z}) = -(\widetilde{z}-z)1_{i>W} - z - g_i - \sum_j (s_\zeta)_{ij} M^{\zeta,\mathrm{g}}_j(z,\widetilde{z}), \quad z,\widetilde{z} \in \mathbb{C}^+ \cup \mathbb{R}$$

with the constraint that

$$M^{0,0}_i(\widetilde{z},\widetilde{z}) = m_{\mathrm{sc}}(\widetilde{z} + \mathrm{i}0^+),$$

the Stieltjes transform of the semicircle law. For simplicity of notations, we denote by $M^{\zeta,\mathrm{g}}(z,\widetilde{z})$ the matrix with entries

$$M^{\zeta,\mathrm{g}}_{ij} := M^{\zeta,\mathrm{g}}_i \delta_{ij}.$$

We will show that $M^{\zeta,\mathrm{g}}(z,\widetilde{z})$ is the limit of the generalized resolvent $G^{\mathrm{g}}_\zeta(z,\widetilde{z})$. For this purpose, we first collect basic properties of M in the following lemma, which is proved in [6].

LEMMA 4.4. *Assume* $|\operatorname{Re}\tilde{z}| \le 2-\kappa$ *for some* $\kappa > 0$. *There exist* $c, C > 0$ *such that the following hold*:

(i) Existence and Lipschitz continuity. *If*

$$\zeta + \|\mathrm{g}\|_\infty + |z-\tilde{z}| \le c, \tag{4.20}$$

then $M_i^{\zeta,\mathrm{g}}(z,\tilde{z})$ *exists and*

$$\max_i \bigl|M_i^{\zeta,\mathrm{g}}(z,\tilde{z}) - m_{\mathrm{sc}}(\tilde{z}+\mathrm{i}0^+)\bigr| \le C\bigl(\zeta + \|\mathrm{g}\|_\infty + |z-\tilde{z}|\bigr).$$

If, in addition, we assume $\zeta' + \|\mathrm{g}'\|_\infty + |z'-\tilde{z}| \le c$, *then*

$$\max_i \bigl|M_i^{\zeta',\mathrm{g}'}(z',\tilde{z}) - M_i^{\zeta,\mathrm{g}}(z,\tilde{z})\bigr| \le C\bigl(\|\mathrm{g}-\mathrm{g}'\|_\infty + |z'-z| + |\zeta'-\zeta|\bigr).$$

(ii) Uniqueness. *The vector* $M_i^{\zeta,\mathrm{g}}(z,\tilde{z})$ $(1 \le i \le N)$ *is unique under the constraints* (4.20) *and*

$$\max_i \bigl|M_i^{\zeta,\mathrm{g}}(z,\tilde{z}) - m_{\mathrm{sc}}(\tilde{z}+\mathrm{i}0^+)\bigr| \le c.$$

Since s_{ij} from (1.6) is a periodic function of $i-j$, by the uniqueness of the previous lemma, we have $M_i^{\zeta,0}(z,\tilde{z}) = M_{W-i}^{\zeta,0}(z,\tilde{z})$, so that

$$\sum_{i=1}^{W/2} M_i^{\zeta,0}(z,\tilde{z}) = \frac{1}{2}\sum_{i=1}^{W} M_i^{\zeta,0}(z,\tilde{z}). \tag{4.21}$$

This equation will be necessary to obtain the averaged QUE estimate for Q_e^{g} in (4.28). Our main results on the generalized resolvent of H_ζ^{g} is the following, proved in a companion paper.

THEOREM 4.5 (Generalized resolvent estimate). *Recall* η_*, η^*, *and* r *from Assumptions* 2.3 *and* 2.4. *Suppose these parameters are of the form*

$$\begin{gathered}\eta_* = N^{-\varepsilon_*}, \quad \eta^* = N^{-\varepsilon^*}, \quad r = N^{-\varepsilon_*+3\varepsilon^*},\\ T = N^{-\varepsilon_*+\varepsilon^*}, \quad 0 < \varepsilon^* \le \varepsilon_*/20,\end{gathered} \tag{4.22}$$

where T *is a new parameter used in the equation* (4.24) *below. Assume that*

$$\log_N W > \max\left(\frac{3}{4}+\varepsilon^*, \frac{1}{2}+\varepsilon_*+\varepsilon^*\right). \tag{4.23}$$

For any small $\tau, \kappa > 0$ *and large* D, *uniformly in* $|e| < 2-\kappa$, *for large enough* N *the following holds. For any deterministic* z, ζ, *and* g *satisfying*

$$|\operatorname{Re} z - e| \le r, \quad \eta_* \le \operatorname{Im} z \le \eta^*, \quad 0 \le \zeta \le T, \quad \|\mathrm{g}\|_\infty \le W^{-3/4}, \tag{4.24}$$

we have (*we denote* $\|A\|_{\max} = \max_{i,j}|A_{ij}|$)

$$\mathbb{P}\left(\bigl\|G_\zeta^{\mathrm{g}}(z,e) - M^{\zeta,\mathrm{g}}(z,e)\bigr\|_{\max} \ge N^\tau\left(\frac{N^{1/2}}{W} + \frac{1}{\sqrt{W\operatorname{Im} z}}\right)\right) \le N^{-D}. \tag{4.25}$$

The following corollary is an immediate consequence of the above generalized resolvent estimate, the deterministic Lemma 4.4, and (4.21). In the statement, we use the notation $\mathscr{I}_{e,r} = (e-r, e+r)$ as in (2.13).

COROLLARY 4.6. *We follow the assumptions and conventions of Theorem* 4.5. *Then for any $z = E + \mathrm{i}\eta$ with $E \in \mathscr{I}_{e,r}$ and $\eta_* \le \eta \le \eta^*$, any t satisfying $0 \le t \le T$ and any fixed* (*large*) *number $D > 0$ the following statements hold for N large enough:*

$$\mathbb{P}\left(\exists E \in \mathscr{I}_{e,r}\ \left|\operatorname{Im}\left(K^{\mathrm{g},e}_{T-t} - z\right)^{-1}_{kk}\right| \ge \frac{2}{W}\operatorname{Im}\operatorname{Tr}\left(K^{\mathrm{g},e}_{T-t} - z\right)^{-1}\right) \le W^{-D}, \tag{4.26}$$

$$\mathbb{P}\left(\left|\frac{1}{W}\operatorname{Tr}\left(K^{\mathrm{g},e}_{T-t} - z\right)^{-1} - m_{\mathrm{sc}}(z)\right| \ge N^{-\varepsilon^*/2}\right) \le W^{-D}, \tag{4.27}$$

$$\begin{aligned}\mathbb{P}\Bigg(\max_{E\in\mathscr{I}_{e,r}}\Bigg|\frac{1}{W}\sum_{1\le k\le W/2}\left(K^{\mathrm{g},e}_{T-t} - z\right)^{-1}_{kk} - \frac{1}{2W}\operatorname{Tr}\left(K^{\mathrm{g},e}_{T-t} - z\right)^{-1}\Bigg|& \\ \ge N^{\tau}\left(\frac{N^{1/2}}{W} + \frac{1}{\sqrt{W\operatorname{Im} z}}\right)\Bigg) \le W^{-D}&.\end{aligned} \tag{4.28}$$

In particular, $K^{\mathrm{g},e}_{T-t}$ satisfies the regularity assumptions (2.9), (2.10), *and* (2.11) *in the range* $0 \le t \le T$.

Remark 4.7. This corollary gives control of the error in QUE for mean-field blocks and therefore controls the range of W for which delocalization can be proved.

More precisely, assume $\varepsilon_* = 0$ to simplify. The error $N^{-\mathfrak{c}}$ in Assumption 2.4, which governs the error in Theorem 2.5, is of order $N^{-\mathfrak{c}} \sim \frac{N^{1/2}}{W}$, from (4.28). In order to patch this estimate to get QUE for the band matrix H, we will need $\frac{N}{W} \cdot \frac{N^{1/2}}{W} \ll 1$. This explains our condition $W \gg N^{3/4}$.

However, the error $\frac{\sqrt{N}}{W}$ in (4.28) is taken from (4.25); this error in (4.28) usually can be improved by taking into account the average of the index k. We believe that the key error term in Theorem 2.5 comes from the last term in (2.14). If we take t_0 close to 1 and replace n by W, this error is of order $W^{-1/2}$. We therefore expect that for $\frac{N}{W} \cdot \frac{1}{W^{1/2}} \ll 1$, i.e., $W \gg N^{2/3}$, the QUE for band matrices holds. If we additionally assume that these errors associated to different blocks are centered and asymptotically independent, then the total error for the QUE of the band matrix H would be $(\frac{N}{W})^{1/2} \cdot \frac{1}{W^{1/2}}$, which is much smaller than 1 when $W \gg N^{1/2}$.

4.5 Eigenvector and Eigenvalue Estimates for Mean-Field Blocks

The following lemma concerns the QUE and related properties of the $W \times W$ matrix Q^{g}_e from (4.10). It is an important building block for the proof of Theorem 4.1.

For the statement, recall the notations from Section 4.2. In particular, the matrix Q^{g}_e and its eigenvalues and eigenvectors $\xi^{\mathrm{g}}_k(e)$ and $\mathbf{u}^{\mathrm{g}}_k(e)$ are defined in (4.10).

LEMMA 4.8. *Let H satisfy the assumptions in Theorem* 4.1 *and $\kappa, \tau > 0$ be small constants. For $e \in \mathbb{R}$, $1 \leq k \leq W$, and $C > 0$, denote by χ the set*

$$\chi(k, e, C, \mathrm{g}) := \{|D^{\mathrm{g}} - e| \geq N^{-C}\} \cap \{|\xi_k^{\mathrm{g}}(e) - e| \leq W^{-1}\}. \tag{4.29}$$

Uniformly in deterministic $|e| \leq 2 - \kappa$, the following statements hold.

(i) Delocalization. *For any $C, D > 0$, for N large enough, we have*

$$\max_{\|\mathrm{g}\|_\infty \leq W^{-3/4}} \mathbb{P}(\{\|\mathbf{u}_k^{\mathrm{g}}(e)\|_\infty^2 \geq W^{-1+\tau}\} \cap \chi(k, e, C, \mathrm{g}))0 \leq N^{-D}. \tag{4.30}$$

(ii) Level repulsion. *For any $C, D > 0$, there exists $N_0 = N_0(C, D)$ such that for $N \geq N_0(C, D)$ and for any $x > 0$ we have*

$$\begin{aligned} &\max_k \max_{\|\mathrm{g}\|_\infty \leq W^{-3/4}} \mathbb{P}\left(\left\{|\xi_{k\pm 1}^{\mathrm{g}}(e) - \xi_k^{\mathrm{g}}(e)| \leq \frac{x}{W}\right\} \cap \chi(k, e, C, \mathrm{g})\right) \\ &\quad \leq W^\tau x^{2-\tau} + N^{-D}. \end{aligned} \tag{4.31}$$

(iii) QUE. *Recall that ε_m is defined in* (1.11) *and J in* (4.14). *For any $C, D > 0$ for N large enough,*

$$\begin{aligned} \max_{\|\mathrm{g}\|_\infty \leq W^{-3/4}} \mathbb{P}\Bigg(\Bigg\{\left|\|J \cdot \mathbf{u}_k^{\mathrm{g}}(e)\|_2^2 - \frac{1}{2}\right| \geq \frac{N^{1/2+\tau}}{W} + \frac{N^{\frac{\varepsilon_m}{2}+\tau}}{W^{1/2}}\Bigg\} \\ \cap \chi(k, e, C, \mathrm{g})\Bigg) \leq N^{-D}. \end{aligned} \tag{4.32}$$

(iv) Local law. *Take $\mathrm{g} = 0$. There exists $\varepsilon > 0$ that does not depend on τ such that for any $C, D > 0$ and for sufficiently large N we have*

$$\begin{aligned} \mathbb{P}\Bigg(\Bigg\{\sup_{0 \leq e' - e \leq W^{-1+\varepsilon}} \left|\#\{k : \xi_k(e) \in [e, e']\} - W \int_e^{e'} \rho_{\mathrm{sc}}(x) \mathrm{d}x\right| \geq W^\tau\Bigg\} \\ \cap \{|D - e| \geq N^{-C}\}\Bigg) \leq N^{-D}. \end{aligned} \tag{4.33}$$

Notice that for $\tau < \varepsilon$, we have $W^\tau < W \int_e^{e+W^{-1+\varepsilon}} \rho_{\mathrm{sc}}(x) \mathrm{d}x$.

Remark 4.9. The constraint $|\xi_k^{\mathrm{g}}(e) - e| \leq W^{-1}$ in (4.29) can be replaced by $|\xi_k^{\mathrm{g}}(e) - e| \leq W^{-1+\varepsilon}$ for some $\varepsilon > 0$, with little change in the proof. In the application of this lemma in our paper, we only need to use AW^{-1} for any large fixed constant A.

PROOF. Recall the operator $K_{T,t}^{\mathrm{g},e}$ in (4.13) and η_*, η^*, r, and T in (4.22). Denote the eigenvalues and eigenvectors of $K_T^{\mathrm{g},e}(t)$ by $\lambda_k^T(t)$ and $\mathbf{u}_k^T(t)$. Hence the distributions of the eigenvalue $\xi_k^{\mathrm{g}}(e)$ and eigenvector $\mathbf{u}_k^{\mathrm{g}}(e)$ of Q_e^{g} are given by

$$\xi_k^{\mathrm{g}}(e) \overset{(\mathrm{d})}{=} \lambda_{k,T}(t), \quad \mathbf{u}_k^{\mathrm{g}}(e) \overset{(\mathrm{d})}{=} \mathbf{u}_{k,T}(t).$$

By definition of $K_T^{\mathrm{g},e}(t)$, it is trivial to prove that for any $C_0 > 0$ there exists C_1 such that

$$1_{|D^{\mathrm{g}}-e|\geq N^{-C_0}} \|K_T^{\mathrm{g},e}(t)\| \leq W^{C_1} \tag{4.34}$$

holds with a very high probability for any $0 \leq t \leq T$. Corollary 4.6 and (4.34) imply that $K_T^{\mathrm{g},e}(t)$ is (η_*, η^*, r)-regular (Assumption 2.3) at $E_0 = e$ for any $0 \leq t \leq T$ (under the condition $1_{|D^{\mathrm{g}}-e|\geq N^{-C_0}}$). In addition, the conditions in Assumption 2.4 are guaranteed by (4.26) and (4.28). By Theorem 2.5, for any small $\varepsilon > 0$, with overwhelming probability we have

$$\left| \|J \cdot \mathbf{u}_k^{\mathrm{g}}(e)\|_2^2 - \frac{1}{2}\right| \leq W^{\varepsilon}\left(\frac{N^{1/2}}{W} + (WN^{-\varepsilon_*+\varepsilon^*})^{-1/2}\right),$$

where we have used T defined in (4.22). We now choose $\varepsilon_* = \varepsilon_m + \varepsilon$ and $\varepsilon^* = \varepsilon$. For small enough ε, thanks to (1.11), the constraint (4.23) is satisfied. Together with the above equation, this proves (4.32). Moreover, by (3.13), $(K_{T,t}^{\mathrm{g},e} - z)_{ii}^{-1}$ is uniformly bounded for $z \in \mathscr{D}_\kappa$ and this implies (4.30).

To prove (4.31), we need a level repulsion result from [23].

LEMMA 4.10 (Theorem 3.5 and 3.6 of [23]). *Let $\lambda_{i,t}$ denote the eigenvalues of $K(t)$ (2.1) with V (η_*, η^*, r)-regular at E_0 and bounded as in Definition 2.3. Assume that there exists $c < 1$ such that*

$$|\log \eta^*| \leq c|\log \eta_*|. \tag{4.35}$$

Let $\tau > 0$. Then for large enough N and for any $x > 0$ we have

$$\max_{t\in\mathscr{T}_\omega} \mathbb{P}\left(\left\{|\lambda_{i,t} - E_0| \leq W^{-1}\right\} \cap \left\{|\lambda_{i,t} - \lambda_{i\pm1,t}| \leq W^{-1}x\right\}\right) \leq W^{\tau} x^{2-\tau},$$

where $\mathscr{T}_\omega$ is defined in (3.1).

We now apply Lemma 4.10 to the flow (4.12). The condition (4.35) is trivially verified by the choice of $\varepsilon_*, \varepsilon^*$ in Lemma 4.5. Hence Lemma 4.10 implies the level repulsion estimate (4.31).

It remains to prove (4.33). From Lemma 3.3(i) applied to $K_T^{0,e}(t)$ at time $t = T$, we have

$$\mathbb{P}\left(\left\{\sup_{0\leq e'-e\leq W^{-1+\varepsilon}}\left|\#\{k : \xi_k(e) \in [e, e']\} - W\int_e^{e'} \rho_{\mathrm{fc},T}(x)\mathrm{d}x\right| \geq W^{\tau}\right\} \cap \{|D - e| \geq N^{-C}\}\right) \leq N^{-D}.$$

We therefore just need to prove

$$\left|\int_e^{e'} \rho_{\mathrm{fc},T}(x)\mathrm{d}x - \int_e^{e'} \rho_{\mathrm{sc}}(x)\mathrm{d}x\right| \leq W^{-1+\tau}. \tag{4.36}$$

Recall the following relation between $m_{\mathrm{fc},t}$ and V:

$$m_{\mathrm{fc},t}(z) = m_V(z + tm_{\mathrm{fc},t}(z)) = \frac{1}{W}\operatorname{Tr}(V - z - t\, m_{\mathrm{fc},t}(z))^{-1} \tag{4.37}$$

where $V = A^{\mathrm{g}}_T - B^{\mathrm{g},*}(D^{\mathrm{g}} - e)^{-1}B^{\mathrm{g}} = K^{\mathrm{g},e}_T$. For $z = E + \mathrm{i}\eta$ with $|E - e| \le r$ and $\eta_* \le \eta \le \eta^*$, (4.27) implies that

$$m_{\mathrm{fc},0}(z) - m_{\mathrm{sc}}(z) = \frac{1}{W}\operatorname{Tr}(V - z)^{-1} - m_{\mathrm{sc}}(z) = \mathrm{O}(N^{-\varepsilon^*/2})$$

holds with high probability. Similarly, by (4.27) and (4.37), for any $t \ge 0$ we have

$$\begin{aligned} &m_{\mathrm{fc},t}(z) - m_{\mathrm{sc}}(z + t\, m_{\mathrm{fc},t}(z)) \\ &\quad = \frac{1}{W}\operatorname{Tr}\bigl(V - z - t\, m_{\mathrm{fc},t}(z)\bigr)^{-1} - m_{\mathrm{sc}}(z + tm_{\mathrm{fc},t}(z)) = \mathrm{O}(N^{-\varepsilon^*/2}) \end{aligned} \tag{4.38}$$

provided that

$$\begin{aligned} \eta_* \le \operatorname{Im}(z + tm_{\mathrm{fc},t}(z)) = \eta + t \operatorname{Im} m_{\mathrm{fc},t}(z) \le \eta^*, \\ |\operatorname{Re}(z + tm_{\mathrm{fc},t}(z)) - e| \le r. \end{aligned} \tag{4.39}$$

For $t = T$ as defined in Lemma 4.5, we have

$$\eta_* \ll T \ll \eta^*, \quad T \ll r/2, \quad |E - e| \le r/2, \quad 0 \le \eta \le \eta^*/2. \tag{4.40}$$

Moreover, as proved in [23, lemma 7.2], for any $0 < \eta < \eta^*$, we have

$$c \le \operatorname{Im} m_{\mathrm{fc},T}(z) \le C', \quad |m_{\mathrm{fc},T}(z)| < C' \log N, \tag{4.41}$$

for some positive constants c, C'. Equations (4.40) and (4.41) show that the assumption (4.39) holds for $t = T$, and the proof of (4.36) is concluded by taking $\eta = 0^+$ in (4.38). □

4.6 Regularity and the Weak Uncertainty Principle

The GOE component in (4.1) implies the following regularity property and weak uncertainty principle. This lemma does not require the decomposition (1.15), i.e., the Gaussian divisibility for the band matrix elements; we state it under this assumption for simplicity.

LEMMA 4.11. *Let H be as in Theorem 4.1 for some fixed $A > 10$. Let $\boldsymbol{\phi} \in \mathbb{R}^N$ be defined by*

$$\phi_i = 1_{W \le i \le N}.$$

Recall the notatons from Definition 4.3. Then there exists a (large) constant $C_r = C_r(A)$ (the subscript r is used to indicate that the constant is related to the regularity) such that for any fixed $D > 0$

$$\begin{aligned} \max_{\|\mathrm{g}\|_\infty \le W^{0.9}/N} &\mathbb{P}\Bigl(\exists t : |t| \le 20,\ k \in \mathbb{Z}_N \textit{ such that} \\ &\qquad |\lambda^{\mathrm{g}+t\boldsymbol{\phi}}_k| \le 20,\ \|\mathbf{w}^{\mathrm{g}+t\boldsymbol{\phi}}_k\|_2^2 \le N^{-C_r}\Bigr) \le N^{-D}, \end{aligned} \tag{4.42}$$

(4.43)
$$\max_{\|g\|_\infty \le W^{0.9}/N} \mathbb{P}\biggl(\exists e : |e| \le 10,\\ B^* \frac{1}{(D^{g} - e)^2} B \ge N^{C_r} \biggl(B^* \frac{1}{(D^{g} - e)} B \biggr)^2 + N^{C_r} \biggr) \le N^{-D}.$$

The proof of this lemma follows the one for [4, proposition 3.1]. Lemma 4.11 is weaker in the sense that the error $N^{\pm C_r}$ was originally given by order 1 quantities in [4]. In addition, [4, proposition 3.1] applies to any approximate eigenvector without assuming the small GOE regularization $N^{-A}H^{G}$. On the other hand, Lemma 4.11 works for $W \ge N^{3/4+c}$ (in fact, $W \ge N^{1/2+c}$ is enough), while [4] required $W = \Omega(N)$.

PROOF OF LEMMA 4.11. We first note that 0.9 in (4.42) and (4.43) can be replaced by any fixed number less than 1 in the following arguments.

We will first prove the following form of an uncertainty principle: approximate eigenvectors for D^{g} have some weight on the first W coordinates in the sense that there exists $C > 0$ such that for any fixed $D > 0$,

$$\max_{\|g\|_\infty \le W^{0.9}/N} \mathbb{P}\Bigl(\exists \mathbf{v} \in \mathbb{R}^{N-W} \text{ with } \|\mathbf{v}\|_1 = 1,\ e \in \mathbb{R},\\ \|B^*\mathbf{v}\|_2 + \|(D^{g} - e)\mathbf{v}\|_2 \le N^{-C}\Bigr) \le N^{-D}. \tag{4.44}$$

We first consider $B^*\mathbf{v}$. Thanks to the component $N^{-A}H^{G}$ in (4.1), for any fixed $\mathbf{v}$ and $1 \le n \le W$, there is an a_0 independent of H^{G} such that $(B^*\mathbf{v})_n = a_0 + N^{-A}\xi_n \cdot \mathbf{v}$ with $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_W$ having independent Gaussian entries of variance order $1/N$. Thus there exists $C > 0$ such that for any $\|\mathbf{v}\|_2 = 1$, we have

$$\mathbb{P}\bigl(|(B^*\mathbf{v})_n| \le N^{-C}\bigr) \le 1/2$$

for all $1 \le n \le W$. Taking the intersection of these independent events, we have proved that there exist $C > 0$ and $c > 0$ such that for any $\mathbf{v}$ as above,

$$\mathbb{P}\bigl(\|B^*\mathbf{v}\|_2 \le N^{-C}\bigr) \le e^{-cW}. \tag{4.45}$$

The matrix D in H is itself a band matrix of size $N - W$ and band width W. Denote by λ_k^D the eigenvalues of D. The local law in [17, theorem 2.1] was established up to the scale $W^{-1+\tau}$ for any constant $\tau > 0$, strictly speaking for random band matrices satisfying $\sum_j s_{ij} = 1$. For D, this assumption is violated for i in a set of size at most $2W$, but [17, theorem 2.1] still holds by elementary adjustments left to the reader. This theorem implies in particular that with probability $1 - N^{-D}$

$$\max_{e \in \mathbb{R}} \frac{\#\{k,\ \lambda_k^D \in [e, e + W^{-1+\tau}]\}}{N W^{-1+\tau}} \le 10. \tag{4.46}$$

Denote by $\lambda_k^{D^{g}}$ and $\boldsymbol{\psi}_k^{D^{g}}$, $1 \le k \le N - W$, the eigenvalues and eigenvectors of D^{g}. A trivial bound on the eigenvalue perturbation gives $|\lambda_k^{D^{g}} - \lambda_k^D| = \|g\|_\infty \le$

$W^{0.9}/N$, so that $\lambda_k^{D^{\mathrm{g}}} \in [e, e+N^{-1}]$ implies $\lambda_k^D \in [e - W^{0.9}/N, e + 2W^{0.9}/N]$. Hence with high probability we have

$$\left|\{k : \lambda_k^{D^{\mathrm{g}}} \in [e, e+N^{-1}]\}\right| \le \left|\{j : \lambda_j^D \in [e - W^{0.9}/N, e + 2W^{0.9}/N]\}\right| \\ \le 10W^{0.9},$$

where we have used (4.46) and $W^{0.9}/N \ge W^{-1+\tau}$. Hence for any $D > 0$, for large enough N we have

$$\mathbb{P}\big(\exists e \in \mathbb{R},\ \left|\{k : \lambda_k^{D^{\mathrm{g}}} \in [e, e+N^{-1}]\}\right| \ge 30W^{0.9}\big) \le N^{-D}.$$

As a consequence, if we define $S_e = \operatorname{span}\{\boldsymbol{\psi}_k^{D^{\mathrm{g}}} : \lambda_k^{D^{\mathrm{g}}} \in [e, e+N^{-1}]\}$, then $\dim(S_e) = \mathrm{O}(W^{0.9})$ with high probability. For such an S_e of dimension $\mathrm{O}(W^{0.9})$, we can choose a finite set $\mathscr{N}$ in the unit sphere of S_e with $|\mathscr{N}| = N^{\mathrm{O}(W^{0.9})}$ such that for any $\mathbf{v}$ in the unit sphere there is a $p \in \mathscr{N}$ such that $|\mathbf{v} - p| \le N^{-C-1}$ with C being the constant in (4.45). Together with (4.45) and the fact that $\|B\| \le N$ holds with very high probability, we obtain that there exists $C > 0$ such that for any fixed $D > 0$,

$$\mathbb{P}\big(\exists \mathbf{v} \in S_e,\ \|B^*\mathbf{v}\|_2 \le N^{-C}\big) \le N^{-D},$$

where we have used $e^{-cW}N^{\mathrm{O}(W^{0.9})} \le e^{-c'W}$ for some $c' > 0$. Therefore there exists $C > 0$ such that for any fixed e and $D > 0$, for large enough N we have $\mathbb{P}(A_e) \le N^{-D}$ where

$$A_e = \big\{\exists \mathbf{v} \in \mathbb{R}^{N-W} \text{ with } \|\mathbf{v}\| = 1,\ \|B^*\mathbf{v}\|_2 + \|(D^{\mathrm{g}} - e)\mathbf{v}\|_2 \le N^{-C}\big\}. \tag{4.47}$$

By union bound, we also have $\mathbb{P}(\bigcap_{e \in N^{-2C}\mathbb{Z}, |e| < N^C} A_e) \le N^{-D}$. Moreover, $\|D^{\mathrm{g}}\| \le N^C$ with high probability, so that (4.44) follows easily.

We now show how (4.42) follows from (4.44). By definition $D^{\mathrm{g}+t\phi} = D^{\mathrm{g}} - t$ and $A^{\mathrm{g}+t\phi} = A^{\mathrm{g}}$, so the eigenvector equation is

$$A^{\mathrm{g}}\mathbf{w}_k^{\mathrm{g}+t\boldsymbol{\phi}} + B^*\mathbf{v}_k^{\mathrm{g}+t\boldsymbol{\phi}} = \lambda_k^{\mathrm{g}+t\boldsymbol{\phi}}\mathbf{w}_k^{\mathrm{g}+t\boldsymbol{\phi}},$$
$$B\mathbf{w}_k^{\mathrm{g}+t\boldsymbol{\phi}} + (D^{\mathrm{g}} - t)\mathbf{v}_k^{\mathrm{g}+t\boldsymbol{\phi}} = \lambda_k^{\mathrm{g}+t\boldsymbol{\phi}}\mathbf{v}_k^{\mathrm{g}+t\boldsymbol{\phi}}.$$

If $\|\mathbf{w}_k^{\mathrm{g}+t\boldsymbol{\phi}}\|_2 \le N^{-C}$ for some $C > 0$, then $\|A^{\mathrm{g}}\mathbf{w}_k^{\mathrm{g}+t\boldsymbol{\phi}}\| + \|B\mathbf{w}_k^{\mathrm{g}+t\boldsymbol{\phi}}\| \le N^{-C/2}$ with very high probability. Hence $\mathbf{v}_k^{\mathrm{g}+t\boldsymbol{\phi}}$, after normalization, realizes the condition (4.47) with $e = t + \lambda_k^{\mathrm{g}+t\boldsymbol{\phi}}$. Therefore, (4.42) follows from (4.44).

We now prove (4.43). The event in (4.43) means that for some normalized $\mathbf{v} \in \mathbb{R}^{N-W}$ and $|e| < 10$,

$$\left\|\frac{1}{(D^{\mathrm{g}} - e)}B\mathbf{v}\right\|_2 \ge N^{C_r}\left(\left\|B^*\frac{1}{(D^{\mathrm{g}} - e)}B\mathbf{v}\right\|_2 + 1\right). \tag{4.48}$$

Denoting $\widetilde{\mathbf{v}} = (D^{\mathrm{g}} - e)^{-1} B\mathbf{v} / \|(D^{\mathrm{g}} - e)^{-1} B\mathbf{v}\|_2$, it follows from (4.48) that

$$\begin{aligned}\|(D^{\mathrm{g}} - e)\widetilde{\mathbf{v}}\|_2 &= \frac{\|B\mathbf{v}\|_2}{\|(D^{\mathrm{g}} - e)^{-1} B\mathbf{v}\|_2} \\ &\leq \frac{\|B\mathbf{v}\|_2}{N^{C_r}\big(\big\|B^* \frac{1}{(D^{\mathrm{g}}-e)} B\mathbf{v}\big\|_2 + 1\big)} \leq N^{-C_r}\|B\mathbf{v}\|_2, \\ \|B^*\widetilde{\mathbf{v}}\|_2 &= \frac{\|B^*(D^{\mathrm{g}} - e)^{-1} B\mathbf{v}\|_2}{\|(D^{\mathrm{g}} - e)^{-1} B\mathbf{v}\|_2} \leq N^{-C_r}.\end{aligned}$$

Since $\|B\|_{\mathrm{op}} \leq N$ with high probability, $\widetilde{v}$ realizes the event (4.47), so that (4.44) implies (4.43). □

4.7 Proof of Theorem 4.1

We make rigorous our proof sketch from Section 4.3. We consider the full band matrix H, the proof for the minors $H^{(k)}$ being the same up to trivial adjustments.

Recall the notations from Section 4.2. There, we assumed that the map $g \to e = \lambda_k^g + g$ has a regular inverse, which enables us to define the curve $\mathscr{C}_k(e) = \lambda_k^g$. To prove this, a simple perturbation calculation yields that $\partial(\lambda_k^g + g)/\partial g = \sum_{i=1}^W |\psi_k^g(i)|^2$ By (4.42), $|\psi_k^g(i)|^2 \geq N^{-C_r}$ for all $|g| < 20$ for some constant $C_r > 0$, with high probability. Thus the invertibility is proved with high probability, and from now on we shall restrict ourselves to this case. By differentiating w.r.t. g in the identity $\mathscr{C}_k(\lambda_k^g + g) = \lambda_k^g$, we have

$$\left|\frac{\partial}{\partial e}\mathscr{C}_k(e)\right| = \left|1 - \left(\sum_{i=1}^W |\psi_k^g(i)|^2\right)^{-1}\right| \leq N^{C_r}. \tag{4.49}$$

We now complete the proof of Theorem 4.1, successively considering QUE for some small mean-field matrices, then QUE and delocalization for band matrices, the semicircle law, and universality.

Part 1A: *QUE for a small matrix.* We will prove that a slightly more general form of (1.12) holds for eigenvectors $\boldsymbol{\psi}^{\mathrm{g}}$ of H^{g} with any $\|\mathrm{g}\|_\infty \leq W^{-3/4}$. But for simplicity, we present the proof for $\mathrm{g} = 0$, and point out the modification for the general case whenever it is needed.

We will prove the following delocalization and QUE for the eigenvector $\mathbf{u}_{k'}(\lambda_k)$ of $Q_{\lambda_k}^{\mathrm{g}}$ defined in (4.10):

$$\begin{aligned}\mathbb{P}\Big(&\exists k \in [\![1, W]\!], j \in \mathbb{Z}_N : |\lambda_j| \leq 2 - \kappa,\ \xi_k(\lambda_j) = \lambda_j, \\ &\|\mathbf{u}_k(\lambda_j)\|_\infty^2 \geq W^{-1+\tau}\Big) \leq N^{-D},\end{aligned} \tag{4.50}$$

$$\begin{aligned}(4.51)\qquad &\mathbb{P}\bigg(\exists k \in [\![1, W]\!], j \in \mathbb{Z}_N : |\lambda_j| \leq 2-\kappa,\ \xi_k(\lambda_j) = \lambda_j,\\ &\bigg|\|J \cdot \mathbf{u}_k(\lambda_j)\|_2^2 - \frac{1}{2}\bigg| \geq \frac{N^{\frac{1}{2}+\tau}}{W} + \frac{N^{\frac{\varepsilon m}{2}+\tau}}{W^{1/2}}\bigg) \leq N^{-D}.\end{aligned}$$

The difference between (4.50)–(4.51) and (4.30)–(4.32) is the randomness of their arguments, i.e., e in (4.30)–(4.32) is replaced by λ_j in (4.50)–(4.51). To prove (4.50) and (4.51), our basic strategy is combining an ε-net argument with a perturbation theory of eigenvectors. Let $M = N^{2C_1+2D}$; here D is the constant appearing in (4.50) and (4.51) (not to be confused with the notation that D was also used to denote a matrix) and $C_1 = D + 6C_r$ where C_r is the constant in (4.49).

We denote $E_n = nN^{-C_1}$ and claim that for each fixed n satisfying

$$[E_n, E_{n+1}] \subset [-2+\kappa, 2-\kappa],$$

there is a deterministic e_n with $E_n < e_n < E_{n+1}$ such that

$$\mathbb{P}\big(\inf_{j,n}|\lambda_j^D - e_n| \leq M^{-1}\big) \leq N^{-D} \tag{4.52}$$

where the λ_j^D's are the eigenvalues of D. To see this, note that for any $\eta > 0$

$$\int_{E_n}^{E_{n+1}} \mathbb{E}\,\mathrm{Im}[D - E - \mathrm{i}\eta]^{-1}\,\mathrm{d}E \leq \mathbb{E}\int_{\mathbb{R}} \mathrm{Im}[D - E - \mathrm{i}\eta]^{-1}\,\mathrm{d}E \leq N.$$

Hence there is an $e_n \in [E_n, E_{n+1}]$ such that, with $\eta = M^{-1}$,

$$\mathbb{E}\,\mathrm{Im}[D - e_n - \mathrm{i}\eta]^{-1} \leq CN^{C_1+1}.$$

By the Markov inequality,

$$\mathbb{P}\big(\inf_{j}|\lambda_j^D - e_n| \leq M^{-1}\big) \leq N^{C_1+1}M^{-1}$$

and (4.52) holds, so that we can restrict our consideration to the set $|D-e| \geq N^{-C}$ for any $C \geq 2C_1 + 2D$. By Lemma 4.8, equations (4.30), (4.31), and (4.32) hold. In particular, (4.31) holds with $x = N^{-C_1/2}$. Hence for all n and l satisfying $|\xi_l(e_n) - e_n| \leq W^{-1}$ we have

$$\begin{aligned}(4.53)\qquad &\|\mathbf{u}_l(e_n)\|_\infty^2 \leq W^{-1+\tau}, \quad \bigg|\|J \cdot \mathbf{u}_l(e_n)\|_2^2 - \frac{1}{2}\bigg| \leq \frac{N^{1/2+\tau}}{W} + \frac{N^{\frac{\varepsilon m}{2}+\tau}}{W^{1/2}},\\ &|\xi_{l\pm1}(e_n) - \xi_l(e_n)| \geq N^{-C_1/2},\end{aligned}$$

with probability larger than $(1 - N^{-D} - N^{-C_1(1-\tau)}W^{2+\tau}) \geq 1 - 2N^{-D}$. Here we have used the choice $C_1 = D + 6C_r$, and τ can be an arbitrarily small number.

We define

$$m(\lambda_k) = \sup_n\{n : e_n < \lambda_k\}. \tag{4.54}$$

For simplicity we denote $\widetilde{e} = e_{m(\lambda_k)}$. Recall that $\mathscr{C}_k(\lambda_k) = \lambda_k$, and thus (4.49) and (4.42) assert that $|\partial\mathscr{C}_k(e)/\partial e| \leq N^{C_r}$ holds with high probability. Since

$e_{n+1} - e_n \le 2N^{-C_1}$, (4.54) implies $|\widetilde{e} - \lambda_k| \le 2N^{-C_1}$. Since $C_1 \ge 6C_r$ so that $N^{-C_1}N^{C_r} \ll N^{-0.8C_1}$, we have

$$\begin{aligned} |\mathscr{C}_k(\widetilde{e}) - \widetilde{e}\,| &\le |\mathscr{C}_k(\widetilde{e}) - \mathscr{C}_k(\lambda_k)| + |\lambda_k - \widetilde{e}\,| \\ &\le N^{C_r}|\lambda_k - \widetilde{e}\,| \le 2N^{-C_1}N^{C_r} \le N^{-0.8C_1} \ll W^{-1} \end{aligned} \tag{4.55}$$

with probability larger than $1 - N^{-D}$.

Recall that $k' \in [\![1, W]\!]$ is the index given by the relation $\xi_{k'}(e) = \mathscr{C}_k(e)$ for all e. Applying (4.53) with e_n set to be $\widetilde{e}$ and using $C_1 \gg D$, we obtain the level repulsion bound

$$\mathbb{P}\big(|\xi_{k'\pm 1}(\widetilde{e}) - \xi_{k'}(\widetilde{e})| \ge N^{-C_1/2}\big) \ge 1 - N^{-D}.$$

Together with the continuity argument used in (4.55), the level repulsion holds between $\mathscr{C}_k$ and $\mathscr{C}_{k\pm 1}$ in the interval $[\widetilde{e}, \lambda_k\,]$, i.e.,

$$\mathbb{P}\left(\exists\, e \in [\widetilde{e}, \lambda_k] \text{ s.t. } |\mathscr{C}_{k\pm 1}(e) - \mathscr{C}_k(e)| \le \frac{1}{2}N^{-C_1/2}\right) \le N^{-D}. \tag{4.56}$$

Integrating the perturbation formula (4.16), we get

$$\begin{aligned} \mathbf{u}_{k'}(\lambda_k) &= \mathbf{u}_{k'}(\widetilde{e}) \\ &\quad - \int_{\lambda_k}^{\widetilde{e}} \sum_{\ell \ne k} \frac{\mathrm{u}_{\ell'}(e)}{\mathscr{C}_k(e) - \mathscr{C}_\ell(e)} \left(\mathrm{u}_{\ell'}(e), B^* \frac{1}{(D-e)^2} B\, \mathbf{u}_{k'}(e)\right) \mathrm{d}e. \end{aligned} \tag{4.57}$$

Inserting (4.56) into (4.57) and using $|\widetilde{e} - \lambda_k| \le 2N^{-C_1}$, we obtain

$$\begin{aligned} &\|\mathbf{u}_{k'}(\lambda_k) - \mathbf{u}_{k'}(\widetilde{e})\|_\infty \\ &\quad \le CN^{-C_1/2} \max_{e \in [\widetilde{e}, \lambda_k]} \max_{\ell \ne k} \frac{|(\mathbf{u}_{\ell'}, B^*(D-e)^{-2}B\mathbf{u}_{k'})|}{|\mathscr{C}_\ell(e)| + 1}. \end{aligned} \tag{4.58}$$

The numerator of the last term can be bounded by using (4.43) so that

$$\begin{aligned} &|(\mathbf{u}_{\ell'}, B^*(D-e)^{-2}B\mathbf{u}_{k'})| \\ &\quad \le N^{C_r}\big(\|B^*(D-e)^{-1}B\mathbf{u}_{\ell'}\|_2^2 + 1\big)^{1/2}\big(\|B^*(D-e)^{-1}B\mathbf{u}_{k'}\|_2^2 + 1\big)^{1/2}. \end{aligned} \tag{4.59}$$

Inserting the identity $B^*(D-e)^{-1}B\mathbf{u}_{\ell'} = Q_e\mathbf{u}_{\ell'} - A\mathbf{u}_{\ell'} = \xi_{\ell'}(e)\mathbf{u}_{\ell'} - A\mathbf{u}_{\ell'}$ into the right-hand side of (4.59), we obtain

$$|(\mathbf{u}_{\ell'}, B^*(D-e)^{-2}B\mathbf{u}_{k'})| \le N^{C_r}\big(\xi_{\ell'}(e) + \|A\|_{\mathrm{op}} + 1\big)\big(\xi_{k'}(e) + \|A\|_{\mathrm{op}} + 1\big).$$

It is easy to prove that $\|A\|_{\mathrm{op}} = \mathrm{O}(N)$ with high probability. Together with the fact that $\xi_{k'}(e) \in [\lambda_j, \xi_{k'}(\widetilde{e})]$ for $e \in [\widetilde{e}, \lambda_k]$, which follows from $\partial_e \mathscr{C}_j(e) < 0$, we have proved

$$|(\mathbf{u}_{\ell'}, B^*(D-e)^{-2}B\mathbf{u}_{k'})| \le N^{C_r+2}(|\xi_{\ell'}(e)| + 1).$$

Inserting this bound into (4.58) and using that $\xi_{\ell'}(e) = \mathscr{C}_\ell(e)$ in the denominator and the choice $C_1 = D + 6C_r$, we have proved (4.50) and (4.51).

Notice that the constant C_r is associated with the uncertainty principle in (4.43) and N^{-C_1} is the grid size. We can make the grid size small by choosing large

C_1; the price to pay is that the initial data in the stochastic flow argument becomes large; i.e., the constant C_1 in (2.7) is large. However, the results we use on the stochastic flow (e.g., 4.10) are insensitive to this constant, which is the main reason we can choose C_1 large.

Part 1B: *Delocalization and QUE for the band matrices.* By definition (4.11), $\mathbf{u}_{k'}(\lambda_k) = \mathbf{w}_k/\|\mathbf{w}_k\|_2$. Equations (4.50) and (4.51) can be written in the following form: for any fixed large $D > 0$ and small $\tau > 0$,

$$\mathbb{P}\left(\exists k \in \mathbb{Z}_N : |\lambda_k| \le 2-\kappa,\ \frac{\max_{1\le i\le W} |\psi_k(i)|^2}{\sum_{i=1}^{W} |\psi_k(i)|^2} \ge W^{-1+\tau}\right) \le N^{-D},$$

$$\mathbb{P}\left(\exists k \in \mathbb{Z}_N : |\lambda_k| \le 2-\kappa,\right.$$
$$\left.\left|\frac{\sum_{i=1}^{W/2} |\psi_k(i)|^2}{\sum_{i=1}^{W} |\psi_k(i)|^2} - \frac{1}{2}\right| \ge \frac{N^{1/2+\tau}}{W} + \frac{N^{\frac{\varepsilon_m}{2}+\tau}}{W^{1/2}}\right) \le N^{-D}. \tag{4.60}$$

Clearly we can shift the indices so that

$$\mathbb{P}\left(\exists k \in \mathbb{Z}_N : |\lambda_k| \le 2-\kappa,\right.$$
$$\left.\max_{n\in\mathbb{Z}_N}\left|\frac{\sum_{i=1}^{W/2} |\psi_k(n+i)|^2}{\sum_{i=1}^{W} |\psi_k(n+i)|^2} - \frac{1}{2}\right| \ge \frac{N^{1/2+\tau}}{W} + \frac{N^{\frac{\varepsilon_m}{2}+\tau}}{W^{1/2}}\right) \le N^{-D},$$

and a similar shifted version of (4.60) holds. Since W, N, and ε_m are related by (1.11), we have $\frac{N^{1/2+\tau}}{W} + \frac{N^{\frac{\varepsilon_m}{2}+\tau}}{W^{1/2}} = \mathrm{o}(W/N)$, so that exactly as in (4.18)–(4.19) we obtain

$$\mathbb{P}\left(\exists k \in \mathbb{Z}_N, : |\lambda_k| \le 2-\kappa,\right.$$
$$\left.\max_{n\in\mathbb{Z}_N}\left|\frac{N}{W}\sum_{i=1}^{W/2} |\psi_k(i+n)|^2 - \frac{1}{2}\right| \ge \frac{N}{W}\left(\frac{N^{1/2+\tau}}{W} + \frac{N^{\frac{\varepsilon_m}{2}+\tau}}{W^{1/2}}\right)\right) \le N^{-D}.$$

when N/W is an integer. If N/W is not an integer, the delocalization estimate (4.60) can be used to lead to the same conclusion. Moreover, from (1.11) we have

$$\frac{N}{W}\cdot\frac{N^{1/2+\tau}}{W} < N^{-2a} \quad \text{and} \quad \frac{N}{W}\cdot\frac{N^{\frac{\varepsilon_m}{2}+\tau}}{W^{1/2}} < N^{-\frac{3}{2}a}$$

with $a > 0$ given in (1.11). We have thus proved the QUE part of Theorem 4.1. Finally, note that the above QUE for length interval $W/2$ obviously implies the same estimate for length W.

Finally, the proof of Theorem 4.1 just given above holds for all $\|\mathrm{g}\| \le W^{-3/4}$ since all lemmas were proved under this assumption. We have thus proved that for any fixed $\tau, D > 0$, for large enough N we have

$$\min_{\|g\|\le W^{-3/4}} \mathbb{P}\Bigg(\exists j \in \mathbb{Z}_N : |\lambda_j^g| \le 2-\kappa, \quad \frac{N}{W}\sum_{i=1}^{W} |\psi_j^g(i)|^2 = 1 + \mathrm{O}(N^{-\frac{3}{2}a+\tau})\Bigg) \ge 1 - N^{-D}. \tag{4.61}$$

Part 2: *The semicircle law.* Following the mean-field reduction method, we first prove the following lemma.

LEMMA 4.12. *Recall the definition of the constant a in* (1.11). *Under the assumption of Theorem* 4.1, *for any fixed e_0 with $|e_0| \le 2-\kappa$ we have*

$$\max_j \mathbb{P}\Bigg(|\lambda_j - e_0| \le N^{-1+\frac{a}{2}} \text{ and } \left|(\mathscr{C}_j(e_0) - e_0) - \frac{N}{W}(\lambda_j - e_0)\right| \ge W^{-1-\frac{a}{10}}\Bigg) \le N^{-D}. \tag{4.62}$$

PROOF. Recall the definition of the matrix H^g from (4.9) and the relation

$$\frac{\partial(g+\lambda_j^g)}{\partial g} = \sum_{i=1}^{W} \big|\psi_j^g(i)\big|^2.$$

Integrating this relation from g_0 to 0 with g_0 defined by the equation $g_0 + \lambda_j^{g_0} = e_0$, we have

$$\int_{g_0}^{0} \sum_{i=1}^{W} \big|\psi_j^g(i)\big|^2 \,\mathrm{d}g = \lambda_j - e_0. \tag{4.63}$$

By (4.61), for each $|g| \le W^{-3/4}$ fixed, we have

$$\sum_{i=1}^{W} \big|\psi_j^g(i)\big|^2 = W/N(1+\mathrm{O}(N^{-a}))$$

with high probability. The left side of (4.63) is equal to $-g_0 W/N(1+\mathrm{O}(N^{-a}))$ with high probability. By definition, $\mathscr{C}_j(e_0) = \lambda_j^{g_0} = e_0 - g_0$. Inserting this relation into (4.63), we have proved (4.62). □

We now prove the local semicircle law by using (4.33). For $\varepsilon > 0$ small enough, we consider $E_2 > E_1$ with $\Delta := E_2 - E_1 \le N^{-1+\varepsilon}$. Clearly, we can assume $\Delta \ge 1/N$. We apply Lemma 4.12 with the choice $e_0 = E_1$: for any $D > 0$, for large enough N we have

$$\begin{aligned}
\#\left\{k : \mathscr{C}_k(e_0) \in \left[e_0, e_0 + \Delta\frac{N}{W} - \frac{1}{W}\right]\right\} \\
\leq \#\{k : \lambda_k \in [E_1, E_2]\} \\
\leq \#\left\{k : \mathscr{C}_k(e_0) \in \left[e_0, e_0 + \Delta\frac{N}{W} + \frac{1}{W}\right]\right\}
\end{aligned}$$

with probability $1-N^{-D}$. Since $\mathscr{C}_k = \xi_{k'}$ represents the same curve, we can apply (4.33) with the choices $e = E_1$, $e' - e = \Delta N/W - 1/W$, or $\Delta N/W + 1/W$. Hence the estimate (4.33) implies the local semicircle law and we have completed the proof of Theorem 4.1.

Part 3: *Eigenvalue local statistics.* We rely on a fixed energy universality result for a matrix flow from [22] (note that the constraint $\omega_0 > 2/3$ below is probably not optimal but sufficient in our setting).

THEOREM 4.13 (Fixed energy universality for the Dyson Brownian motion [22]). *Let V be an $n \times n$ deterministic matrix and Z be a $n \times n$ standard GOE matrix. Consider $H = V + \sqrt{t_0}Z$ with $t_0 = n^{\omega_0}/n$. Assume that V is $(n^{-\delta_1}t_0, n^{-\delta_2}, n^{\delta_3}t_0)$ regular at E (see Assumption 2.3) with (c is a universal small constant)*

$$\frac{2}{3} < \omega_0 < 1, \ \delta_2 < \min\left(\frac{1-\omega_0}{4}, \delta_3, c\right).$$

Remember the notation $\rho_{\mathrm{fc},t_0}^{(n)}$ for the density corresponding to the Stieltjes transform $m_{\mathrm{fc},t_0}^{(n)}$ defined in (3.2).

For any smooth test function $O \in \mathscr{C}^\infty(\mathbb{R}^k)$ with compact support, there are constants $c, C > 0$ such that

$$\begin{aligned}
&\left|\int_{\mathbb{R}^k} O(\mathbf{a}) p_H^{(k)}\left(E + \frac{\mathbf{a}}{N\rho_{\mathrm{fc},t_0}^{(n)}(E)}\right)\mathrm{d}\mathbf{a} - \int_{\mathbb{R}^k} O(\mathbf{a}) p_{\mathrm{GOE}}^{(k)}\left(E + \frac{\mathbf{a}}{N\rho_{\mathrm{fc},t_0}^{(n)}(E)}\right)\mathrm{d}\mathbf{a}\right| \\
&\quad \leq Cn^{-c}.
\end{aligned}$$

We apply this result to the $W \times W$ matrix flow $t \to K_{T-t}^{\mathrm{g},e}$ at $t = T$ with the initial data $V = K_T^{\mathrm{g},e}$, i.e., $n = W$. By Corollary 4.6, V is (η_*, η^*, r) regular with the parameters defined in Theorem 4.5. With η_*, η^*, r, T defined in (4.22), we have the following identifications:

$$\begin{gathered}
\delta_3 = 2\varepsilon^* \log_W N, \quad \delta_1 = \varepsilon^* \log_W N, \\
\delta_2 = \varepsilon^* \log_W N, \quad 1 - \omega_0 = \log_W N(\varepsilon_* - \varepsilon^*).
\end{gathered}$$

The above theorem with $e = E$ gives (we consider the case $k = 2$ to simplify the presentation)

$$\left| \int_{\mathbb{R}^2} O(\mathbf{a}) p^{(2)}_{Q_E}\left(E + \frac{\mathbf{a}}{N\rho_{\mathrm{sc}}(E)} \right) \mathrm{d}\mathbf{a} - \int_{\mathbb{R}^2} O(\mathbf{a}) p^{(2)}_{\mathrm{GOE}}\left(E + \frac{\mathbf{a}}{N\rho_{\mathrm{sc}}(E)} \right) \mathrm{d}\mathbf{a} \right| \\ \leq CN^{-c}.$$

where we replaced $\rho^{(n)}_{\mathrm{fc},t_0}$ with ρ_{sc} by taking $\eta = 0^+$ in (4.38). We can write

$$\begin{aligned} &\int_{\mathbb{R}^2} O(\mathbf{a}) p^{(2)}_{Q_E}\left(E + \frac{\mathbf{a}}{N\rho_{\mathrm{sc}}(E)} \right) \mathrm{d}\mathbf{a} \\ &\quad = \frac{1}{2} \sum_{k' \neq j'} \mathbb{E}\, O(W\rho_{\mathrm{sc}}(E)(\xi_{k'} - E), W\rho_{\mathrm{sc}}(E)(\xi_{j'} - E)) \\ &\quad = \frac{1}{2} \sum_{k \neq j} \mathbb{E}\, O(W\rho_{\mathrm{sc}}(E)(\mathscr{C}_k(E) - E), W\rho_{\mathrm{sc}}(E)(\mathscr{C}_j(E) - E)). \end{aligned}$$

Recall that there is a shift of indices $k \to k'$ (depending on the randomness) so that $\mathscr{C}_k(E) = \xi_{k'}$. In the expression above, we have summed over all indices, and thus this shift is irrelevant for our purpose.

Applying (4.62) with $e_0 = E$, we can substitute $W\rho_{\mathrm{sc}}(E)(\mathscr{C}_k(E) - E)$ with $N\rho_{\mathrm{sc}}(E)(\lambda_k - E) + \mathrm{O}(W^{-a/10})$ in the above equation. Note that (4.62) holds only for eigenvalues in a small neighborhood of E. Since O is compactly supported, this restriction does not affect the usage of (4.62) in the last equation. Finally, the error term $\mathrm{O}(W^{-a/10})$ is negligible, which concludes the the proof.

5 A Comparison Method

In this section, we prove the theorems 1.2, 1.3, and 1.4. The basic idea follows the Green function comparison method in [17], interpolating between resolvents of two matrices H and $\widetilde{H}$. However, contrary to the setting from [17], we only have a priori estimates on the Green's function for $\widetilde{H}$, and not for H. A self-consistent Green function comparison method for band matrices was developed in [2], which only requires estimates on the Green's function of one of both matrices. Our a priori estimates on the Green's function are different so that we proceed with another self-consistent method from [21].

5.1 Elementary Facts

Recall that the resolvent of a matrix H can be written as

$$G_{ij}(z) = \sum_k \frac{\psi_k(i)\psi_k(j)}{\lambda_k - z}, \quad \operatorname{Im} G_{ii}(E + \mathrm{i}\eta) = \sum_k \frac{\eta|\psi_k(i)|^2}{(E - \lambda_k)^2 + \eta^2}, \tag{5.1}$$

where $\boldsymbol{\psi}_k$ is the k^{th} eigenvector with eigenvalue λ_k. The following lemma is a classical fact connecting the Green function with delocalization of eigenvectors and local laws.

LEMMA 5.1. *Let $(H_N)_{N\geq 1}$ be a sequence of $N \times N$ random symmetric matrices. Suppose that for any $c > 0$ there exists a constant $\kappa_c > 0 \in \mathbb{R}$ such that for any $D > 0$*

$$\inf_{j\in[\![cN,(1-c)N]\!]} \mathbb{P}(|\lambda_j| \leq 2 - \kappa_c) \geq 1 - N^{-D} \tag{5.2}$$

provided that N is large enough. Consider the following assertion: *for any small $\kappa, \tau > 0$ and $D > 0$*

$$\sup_{|E|\leq 2-\kappa, N^{-1}\leq\eta\leq 1} \max_i (\operatorname{Im} G_{ii}(z)) = \mathrm{O}(N^{\tau}), \tag{5.3}$$

$$\sup_{|E|\leq 2-\kappa, N^{-1}\leq\eta\leq 1} \max_{i,j} |G_{ij}(z)| = \mathrm{O}(N^{\tau}), \tag{5.4}$$

hold with probability larger than $1 - N^{-D}$. Then

(i) (5.3) *implies* (1.12).
(ii) (1.12) *and* (1.13) *imply* (5.4).

PROOF. For any $k \in [\![cN, (1-c)N]\!]$, by (5.2), we can assume $|\lambda_k| \leq 2 - \kappa_c$ for some $\kappa_c > 0$. Then (5.3) implies that, with high probability,

$$\eta^{-1}|\psi_k(i)|^2 \leq \operatorname{Im} G_{ii}(\lambda_k + \mathrm{i}\eta) = \mathrm{O}(N^{\tau}), \quad \eta = N^{-1},$$

which is (1.12). On the other hand, the bound (1.12) on $\psi_k(i)$ and the eigenvalue distribution estimate (1.13) inserted in (5.1) yield (5.4) by a simple dyadic decomposition. □

5.2 Proof of Theorem 1.2

By Theorem 1.5, (1.12) of Theorem 1.2 is just a corollary of the following lemma. In the remainder of this section, we will prove Lemma 5.2.

LEMMA 5.2. *If the statement* (5.3) *holds for all H in* (1.15), *then* (5.3) *holds for any H in Theorem* 1.2.

To prove Lemma 5.2, note that for each H in Theorem 1.2, there is $\widetilde{H}$ of type (1.15) such that the first four moments of the entries of H and $\widetilde{H}$ coincide. A precise statement is the following lemma, about a single random variable. The proof is easily adapted from [35, corollary 30] and is left to the reader.

LEMMA 5.3. *Let H be a band matrix satisfying the conditions in Theorem* 1.2. *Then there exists a matrix ensemble $\widetilde{H}$ of the form* (1.15) *satisfying the assumptions of Theorem* 1.5 *such that*

$$\mathbb{E}(H_{ij})^n = \mathbb{E}(\widetilde{H}_{ij})^n, \quad |i-j| \leq W, \quad n = 1, 2, 3, 4.$$

PROOF OF LEMMA 5.2. Let H be the matrix in Theorem 1.2 and $\widetilde{H}$ the one given in Lemma 5.3. Denote by $\widetilde{G}(z) = (\widetilde{H} - z)^{-1}$ the Green function of $\widetilde{H}$.

By Theorem 1.5, (1.12) and (1.13) hold for the eigenvalues and eigenvectors of $\widetilde{H}$. Together with Lemma 5.1, we get

$$\sup_{|E|\le 2-\kappa, N^{-1}\le\eta\le 1} \|\widetilde{G}(z)\|_{\max} = \mathrm{O}(N^\tau), \quad z = E + \mathrm{i}\eta, \tag{5.5}$$

with probability $1 - N^{-D}$. We now prove that the same estimate holds for G, i.e.,

$$\sup_{|E|\le 2-\kappa, N^{-1}\le\eta\le 1} \|G(z)\|_{\max} = \mathrm{O}(N^\tau). \tag{5.6}$$

We follow the self-consistent comparison method from [21]. We start with a very weak estimate $\|G(z)\|_{\max} \le \eta^{-1}$, i.e., (5.6) holds for $\eta \sim 1$. For $\eta < 1$, let $\varepsilon_0 > 0$ be a small parameter and define

$$\eta_m = N^{-m\varepsilon_0}, \quad z_m = E + \mathrm{i}\eta_m, \quad 1 \le m \le \varepsilon_0.$$

Our goal is to prove by induction that for $z = z_m$, $1 \le m \le \varepsilon_0^{-1}$, (5.6) holds, which implies (5.5). Thus it remains to prove that if (5.6) holds for $z = z'_m$, $0 \le m' \le m$, then (5.6) holds for $z = z_{m+1}$, $1 \le m+1 \le \varepsilon_0^{-1}$.

As in [21], we define the symmetric interpolation matrix H^θ by

$$(H^\theta)_{ij} = (1 - \chi_{ij}^\theta)H_{ij}^0 + \chi_{ij}^\theta H_{ij}^1, \quad H^1 = H, \quad H^0 = \widetilde{H}, \tag{5.7}$$

where for $i \le j$, χ_{ij}^θ are i.i.d. Bernoulli random variables such that $\mathbb{P}(\chi_{ij}^\theta = 1) = \theta$. Denote $G^\theta(z) = (H^\theta - z)^{-1}$. We can now recast the induction as follows: if for any (small) τ and (large) D, and $|E| < 2-\kappa$, we have

$$\max_{0\le\theta\le 1}\max_{m'\le m} \|G^\theta(z_{m'})\|_{\max} = \mathrm{O}(N^\tau), \quad z_{m'} = E + \mathrm{i}\eta_{m'},$$

then for any τ and D, and $|E| < 2-\kappa$, we have

$$\max_{0\le\theta\le 1} \|G^\theta(z_{m+1})\|_{\max} = \mathrm{O}(N^\tau), \quad z_{m+1} = E + \mathrm{i}\eta_{m+1}. \tag{5.8}$$

We know that (5.8) holds for $\theta = 0$ and all $m \le \varepsilon_0^{-1}$. Our aim is to prove (5.8) for $0 < \theta \le 1$.

From [3, lemma 10.2], we have $\|G(E + \mathrm{i}\eta/r)\|_{\max} \le r\|G(E + \mathrm{i}\eta)\|_{\max}$ for any $r > 1$. As a consequence,

$$\max_{0\le n\le m} \|G^\theta(z_n)\|_{\max} = \mathrm{O}(N^\tau) \quad \text{implies} \quad \|G^\theta(z_{m+1})\|_{\max} = \mathrm{O}(N^{\tau+\varepsilon_0}).$$

Thus it remains to show that, under the assumption (5.5), if we have

$$\max_{0\le\theta\le 1} \|G^\theta(z_{m+1})\|_{\max} = \mathrm{O}(N^{\tau+\varepsilon_0}), \tag{5.9}$$

then (5.8) also holds.

By (5.5), for any $p \in \mathbb{N}$ and for any $\tau > 0$, we have

$$\max_{kl} \mathbb{E}\big|G_{kl}^{\theta=0}(z_{m+1})\big|^{2p} = \mathrm{O}(N^{2p\tau}). \tag{5.10}$$

We will use the following lemma from [21] to extend the above bound to general $\theta \in [0,1]$:

$$\max_{\theta} \max_{kl} \mathbb{E}\left|G^{\theta}_{kl}(z_{m+1})\right|^{2p} = \mathrm{O}(N^{2p\tau}), \tag{5.11}$$

which completes the proof of (5.8) by Markov's inequality.

LEMMA 5.4. *For fixed $i, j \in Z_N$ and $\lambda \in \mathbb{R}$, we define the matrix*

$$\left(H^{\theta,\lambda}_{(ij)}\right)_{ab} = \begin{cases} \lambda & \text{if } \{a,b\} = \{i,j\}, \\ H^{\theta}_{ab} & \text{if } \{a,b\} \neq \{i,j\}. \end{cases}$$

For bounded and smooth $F \colon \mathbb{R}^{N\times N} \to \mathbb{C}$ we have

$$\partial_\theta \mathbb{E} F(H^\theta) = \sum_{i \le j} \left(\mathbb{E} F\left(H^{\theta, H^1_{ij}}_{(ij)}\right) - \mathbb{E} F\left(H^{\theta, H^0_{ij}}_{(ij)}\right)\right).$$

We now return to prove (5.11). Choose the function F as follows:

$$F(X) = F_{kl,p,z}(X) = \left|\left((X-z)^{-1}\right)_{kl}\right|^{2p}.$$

By (5.10), for any $\tau > 0$ and $p \in \mathbb{N}$, $\mathbb{E}F(H^0) = \mathrm{O}(N^{2\tau p})$. Thus, (5.11) for $H^\theta, 0 \le \theta \le 1$, follows from Gronwall's inequality and the following inequality, to be proved in the remainder of this paragraph (here and below, $z = z_{m+1}$): for any $p \ge 100$, there exists $c > 0$ such that

$$\begin{aligned} \left|\partial_\theta \mathbb{E} F(H^\theta)\right| &= \left|\sum_{i,j} \left(\mathbb{E} F\left(H^{\theta, H^1_{ij}}_{(ij)}\right) - \mathbb{E} F\left(H^{\theta, H^0_{ij}}_{(ij)}\right)\right)\right| \\ &\le N^{-c}\left(1 + \mathbb{E}F(H^\theta)\right) \end{aligned} \tag{5.12}$$

for any $0 \le \theta \le 1$. Note that the above equality is Lemma 5.4.

The matrices $H^{\theta, H^1_{ij}}_{(ij)}$ and $H^{\theta, H^0_{ij}}_{(ij)}$ are identical except for the entries (i,j) and (j,i) when $|i-j| \le W$, so we now compare them by a perturbative argument. We fix i, j and define

$$f(\lambda) = f_{ij,kl,p,z,\theta}(\lambda) := F_{kl,p,z}\left(H^{\theta,\lambda}_{(ij)}\right).$$

By definition, $f(H^\theta_{ij}) = F(H^\theta)$ with H^θ as in (5.7). The n^{th} derivative of f, $f^{(n)}$, is a sum of products of some $2p+n$ matrix entries of the resolvent and its conjugate. From (5.9), we therefore have

$$f^{(n)}(H^\theta_{ij}) = \mathrm{O}(N^{(\tau+\varepsilon_0)(2p+n)})$$

with high probability. By standard iterated resolvent identities, the same bound holds for any $y = \mathrm{O}(W^{-1/2+\tau})$:

$$f^{(n)}(y) = \mathrm{O}(N^{(\tau+\varepsilon_0)(2p+n)})$$

with overwhelming probability. Hence, by Taylor's expansion with respect to $y = 0$, for any $m \geq 1$, we have

$$\mathbb{E}F\Big(H_{(ij)}^{\theta,H_{ij}^1}\Big) - \mathbb{E}F\Big(H_{(ij)}^{\theta,H_{ij}^0}\Big) = \sum_{5\leq n\leq m} \frac{\mathbb{E}(f^{(n)}(0))}{n!}\Big(\mathbb{E}\big((H_{ij}^1)^n\big) - \mathbb{E}\big((H_{ij}^0)^n\big)\Big) \\ + \mathrm{O}\big(N^{(\tau+\varepsilon_0)(2p+m+1)}(W^{-\frac{1}{2}-\tau})^{m+1}\big)$$

where we used that the first four moments of H_{ij}^1 and H_{ij}^0 are the same. On the one hand, we choose $m = p$ so that the above error term is at most $\mathrm{O}(N^{-p/10})$ when $\tau + \varepsilon_0 < 1/100$. On the other hand, $f^{(n)}(0)$ is a sum of products of some $2p + n$ matrix entries of the resolvent and its conjugate, among which at least $2p - n$ are either $\big(H_{(ij)}^{\theta,0} - z\big)_{kl}^{-1}$ or its conjugate. With the resolvent identity, these two quantities are easily bounded by $|G_{kl}^{\theta}| + W^{-1/2+\varepsilon}$ for any $\varepsilon > 2(\tau + \varepsilon_0)$, with high probability. The remaining $2n$ resolvent entries are bounded using (5.9). Therefore, for $n \leq p$,

$$\begin{aligned} |\mathbb{E}f^{(n)}(0)| &\leq C_p N^{2n(\tau+\varepsilon_0)}\big(\mathbb{E}\big(|G_{kl}^{\theta}|^{2p-n}\big) + W^{(-1/2+\varepsilon)(2p-n)}\big) \\ &\leq C_p N^{2n(\tau+\varepsilon_0)}\big(1 + \mathbb{E}(F(H^{\theta}))\big). \end{aligned}$$

The above estimates together give

$$\big|\partial_\theta \mathbb{E}F(H^{\theta})\big| \leq C_p N W^{1-\frac{5}{2}+10(\tau+\varepsilon_0)}(1 + \mathbb{E}F(H^{\theta})) + C_p N^{1-\frac{p}{10}} W.$$

As $W \geq N^{3/4}$ and $p \geq 100$, this concludes the proof of the inequality in (5.12) (and Lemma 5.2). □

5.3 Proof of Theorem 1.3

We keep the notations from the proof of Lemma 5.2 for H and $\widetilde{H}$. On the one hand, from the local law in Theorem 1.5, for any $\kappa, \tau > 0$ there exist $\varepsilon > 0$ such that for any $z = E + N^{-1+\tau}$ with $-2 + \kappa < E < 2 - \kappa$, $\mathrm{Im}\, N^{-1}\,\mathrm{Tr}\,\widetilde{G}(z) - \mathrm{Im}\, m_{\mathrm{sc}}(z) = \mathrm{O}(N^{-\varepsilon})$. On the other hand, by repeating exactly the proof of Lemma 5.2, the estimate (5.12) also holds for

$$F(X) = \big|\mathrm{Im}\, N^{-1}\,\mathrm{Tr}(X - z)^{-1} - \mathrm{Im}\, m_{\mathrm{sc}}(z)\big|^{2p},$$

so that $\mathrm{Im}\, N^{-1}\,\mathrm{Tr}\, G(z) - \mathrm{Im}\, m_{\mathrm{sc}}(z) = \mathrm{O}(N^{-c})$ for some $c > 0$. In turn, this implies the local law for H.

5.4 Proof of Theorem 1.4

Again, we follow the notations from the proof of Lemma 5.2 for H and $\widetilde{H}$. Theorem 1.5 gives universality for $\widetilde{H}$, so that Theorem 1.4 follows by applying the Green's functions comparison theorem from [17]. The input for this theorem is the four-moment matching of the matrix entries, given by construction of $\widetilde{H}$, and resolvent bounds as proved for our band matrices in (5.6).

Appendix A Perfect Matching Observables for Hermitian Matrices

Although the main results of this paper are stated for symmetric matrices, they can be adapted to the Hermitian class. The only major modification concerns the definition of the perfect matching observables. We explain below the Hermitian counterpart of Section 2.

A.1 Eigenvector Dynamics

Let $B^{(h)}$ be a $n \times n$ matrix such that $\Re(B_{ij}^{(h)})$, $\Im(B_{ij}^{(h)})$ $(i < j)$ and $B_{ii}^{(h)}/\sqrt{2}$ are independent standard Brownian motions, and $B_{ji}^{(h)} = (B_{ij}^{(h)})^*$. The $n \times n$ Hermitian Dyson Brownian motion $K^{(h)}$ with initial value $K^{(h)}(0)$ is $K^{(h)}(t) = K^{(h)}(0) + \frac{1}{\sqrt{2n}} B^{(h)}(t)$.

Let $\boldsymbol{\lambda}_0 \in \Sigma_n$, $u_0 \in \mathrm{U}(n)$. The Hermitian Dyson Brownian motion/vector flow with initial condition $(\lambda_1, \dots, \lambda_n) = \boldsymbol{\lambda}_0$, $(u_1, \dots, u_n) = u_0$, is

$$\begin{aligned} \mathrm{d}\lambda_k &= \frac{\mathrm{d}B_{kk}^{(h)}}{\sqrt{2n}} + \left(\frac{1}{n}\sum_{\ell \neq k} \frac{1}{\lambda_k - \lambda_\ell}\right)\mathrm{d}t, \\ \mathrm{d}u_k &= \frac{1}{\sqrt{2n}} \sum_{\ell \neq k} \frac{\mathrm{d}B_{k\ell}^{(h)}}{\lambda_k - \lambda_\ell} u_\ell - \frac{1}{2n} \sum_{\ell \neq k} \frac{\mathrm{d}t}{(\lambda_k - \lambda_\ell)^2} u_k. \end{aligned} \tag{A.1}$$

With the above definitions, the strict analogue of Theorem 2.1 holds in this Hermitian setting.

In addition to (2.4) and (2.5), we define

$$u_k \partial_{\bar{u}_\ell} = \sum_{\alpha=1}^{n} u_k(\alpha) \partial_{\bar{u}_\ell(\alpha)},$$

$$X_{k\ell}^{(h)} = u_k \partial_{u_\ell} - \bar{u}_\ell \partial_{\bar{u}_k}, \quad \bar{X}_{k\ell}^{(h)} = \bar{u}_k \partial_{\bar{u}_\ell} - u_\ell \partial_{u_k}.$$

Here ∂_{u_ℓ} and $\partial_{\bar{u}_\ell}$ are defined by considering u_ℓ as a complex number; i.e., if we write $u_\ell = x + \mathrm{i}y$, then $\partial_{u_\ell} = \frac{1}{2}(\partial_x - \mathrm{i}\partial_y)$. The analogue of Lemma 2.2 for the generator is then (see [7])

$$\mathrm{L}_t^{(h)} = \frac{1}{2} \sum_{1 \leq k < \ell \leq n} c_{k\ell}(t) \Big(X_{k\ell}^{(h)} \bar{X}_{k\ell}^{(h)} + \bar{X}_{k\ell}^{(h)} X_{k\ell}^{(h)} \Big), \tag{A.2}$$

meaning that $\mathrm{d}\mathbb{E}(g(u_t))/\mathrm{d}t = \mathbb{E}(\mathrm{L}_t^{(h)} g(u_t)))$ for the stochastic differential equation (A.1).

A.2 The Observables

As in Section 2, let I be a fixed subset of $[\![1,n]\!]$, and denote the eigenvector overlaps

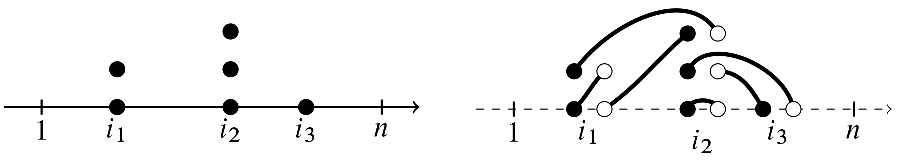
$$p_{ij} = \sum_{\alpha\in I} u_i(\alpha)\bar{u}_j(\alpha), \qquad i \neq j \in [\![1,n]\!],$$
$$p_{ii} = \sum_{\alpha\in I} u_i(\alpha)\bar{u}_i(\alpha) - C_0, \quad i \in [\![1,n]\!],$$

where C_0 is an arbitrary but fixed constant independent of i. Note that contrary to the real case, we now have $p_{ij} \neq p_{ji}$ for $i \neq j$.

With this definition, Theorem 2.5 still holds. For the proof, we keep the same definition for our configuration space as in the real case: $\boldsymbol{\eta}\colon [\![1,n]\!] \to \mathbb{N}$ where $\eta_j := \boldsymbol{\eta}(j)$ is interpreted as the number of particles at the site j. For any given configuration $\boldsymbol{\eta}$, consider the set of vertices

$$\mathscr{V}_{\boldsymbol{\eta}} = \{(i,a,\varepsilon) : 1 \le i \le n, 1 \le a \le \eta_i, \varepsilon \in \{b,w\}\}.$$

We represent vertices corresponding to $\varepsilon = b$ (resp., $\varepsilon = w$) by a black (resp., white) disk. Let $\mathscr{A}_{\boldsymbol{\eta}}$ be the graph with vertices $\mathscr{V}_{\boldsymbol{\eta}}$ and with edges all possible $\{v_1, v_2\}$ with $\varepsilon_1 \neq \varepsilon_2$, where $v_1 = (i_1, a_1, \varepsilon_1)$, $v_2 = (i_2, a_2, \varepsilon_2)$. In words, $\mathscr{A}_{\boldsymbol{\eta}}$ is the complete graph on $\mathscr{V}_{\boldsymbol{\eta}}$ except that edges between vertices of the same color are forbidden. Let $\mathscr{G}_{\boldsymbol{\eta}}$ be the set of perfect matchings of $\mathscr{A}_{\boldsymbol{\eta}}$. Let $\mathscr{E}(G)$ be the set of edges of a graph $G \in \mathscr{G}_{\boldsymbol{\eta}}$.



(A) A configuration $\boldsymbol{\eta}$ with $\mathscr{N}(\boldsymbol{\eta}) = 6$, $\eta_{i_1} = 2$, $\eta_{i_2} = 3$, $\eta_{i_3} = 1$.

(B) A perfect matching $G \in \mathscr{G}_{\boldsymbol{\eta}}$. Here, $P(G) = p_{i_1i_1}p_{i_1i_2}p_{i_2i_1}p_{i_2i_2}p_{i_2i_3}p_{i_3i_2}$.

Moreover, for any given edge $e = \{(i_1, a_1, \varepsilon_1), (i_2, a_2, \varepsilon_2)\}$, we define $p(e) = p_{i_1,i_2}$ if $\varepsilon_1 = b$, and $p(e) = p_{i_2,i_1}$ if $\varepsilon_2 = b$. Let $P(G) = \prod_{e\in\mathscr{E}(G)} p(e)$ and

$$f^{(h)}_{\boldsymbol{\lambda},t}(\boldsymbol{\eta}) = \frac{1}{\mathscr{L}(\boldsymbol{\eta})}\mathbb{E}\Big(\sum_{G\in\mathscr{G}_{\boldsymbol{\eta}}} P(G) \mid \boldsymbol{\lambda}\Big), \quad \mathscr{L}(\boldsymbol{\eta}) = \prod_{i=1}^{n} \eta_i!. \tag{A.3}$$

We have the following complex analogue of Theorem 2.6.

THEOREM A.1 (Perfect matching observables for the eigenvector moment flow: Hermitian case). *Suppose that u is the solution to the Hermitian eigenvector dynamics* (A.1)*, and $f^{(h)}_{\boldsymbol{\lambda},t}$ is given by* (A.3)*. Then it satisfies the equation*

$$\begin{aligned}
&\partial_t f^{(h)}_{\boldsymbol{\lambda},t} = \mathscr{B}^{(h)}(t) f^{(h)}_{\boldsymbol{\lambda},t}, \\
&\mathscr{B}^{(h)}(t) f(\boldsymbol{\eta}) = \sum_{k\neq\ell} c_{k\ell}(t)\eta_k(1+\eta_\ell)(f(\boldsymbol{\eta}^{k,\ell}) - f(\boldsymbol{\eta})).
\end{aligned} \tag{A.4}$$

As in the real case, the above theorem is independent of our choice of the canonical basis; see Remark 2.8. It therefore generalizes the class of observables for the eigenvector moment flow from [7, theorem 3.12(ii)].

A.3 Proof of Theorem A.1

We naturally replace the definition (2.19) with

$$g(\boldsymbol{\eta}) = \frac{1}{\mathscr{L}(\boldsymbol{\eta})} \sum_{G\in\mathscr{G}_{\boldsymbol{\eta}}} P(G),$$

and let $1 \leq k < \ell \leq n$ be fixed for the rest of this subsection. We abbreviate $X = X^{(h)}_{k\ell}$, $\bar{X} = \bar{X}^{(h)}_{k\ell}$. With (A.2) the proof reduces to

$$\begin{aligned}
\frac{1}{2}(X\bar{X} + \bar{X}X)g(\boldsymbol{\eta}) &= \eta_k(1+\eta_\ell)(g(\boldsymbol{\eta}^{k\ell}) - g(\boldsymbol{\eta})) \\
&\quad + \eta_\ell(1+\eta_k)(g(\boldsymbol{\eta}^{\ell k}) - g(\boldsymbol{\eta})).
\end{aligned} \tag{A.5}$$

To calculate $\frac{1}{2}(X\bar{X} + \bar{X}X)P(G)$ for any $G \in \mathscr{G}_{\boldsymbol{\eta}}$, we first need the following definition.

DEFINITION A.2. Let $\boldsymbol{\eta}$ and $k < \ell$ be fixed.

(i) $\mathscr{V}_i \subset \mathscr{V}_{\boldsymbol{\eta}}$ is the set of vertices of type (i, a, ε), $1 \leq a \leq \eta_i$, $\varepsilon \in \{b, w\}$.

(ii) $\mathscr{V}_i^b \subset \mathscr{V}_i$ is the set of vertices of type (i, a, b), $1 \leq a \leq \eta_i$, and similarly for $\mathscr{V}_i^w$.

(iii) For any two sets, denote $A \cdot B = (A \times B) \cup (B \times A)$. We define

$$\varepsilon(v_1, v_2) = \begin{cases} 1 & \text{if } (v_1, v_2) \in (\mathscr{V}_k^b \cdot \mathscr{V}_k^w) \cup (\mathscr{V}_\ell^b \cdot \mathscr{V}_\ell^w), \\ -1 & \text{if } (v_1, v_2) \in (\mathscr{V}_k^b \cdot \mathscr{V}_\ell^b) \cup (\mathscr{V}_k^w \cdot \mathscr{V}_\ell^w), \\ 0 & \text{otherwise.} \end{cases}$$

(iv) Let $G \in \mathscr{G}_{\boldsymbol{\eta}}$ and $(v_1, v_2) \in (\mathscr{V}_k \cup \mathscr{V}_\ell)^2_*$.

Assume $(v_1, v_2) \in (\mathscr{V}_k^b \cdot \mathscr{V}_\ell^b) \cup (\mathscr{V}_k^w \cdot \mathscr{V}_\ell^w)$. Then we define $S_{v_1 v_2} G = S_{v_2 v_1} G \in \mathscr{G}_{\boldsymbol{\eta}}$ as the perfect matching obtained by transposition of v_1 and v_2. More precisely, let $\tau_{v_1 v_2}$ be the permutation of $\mathscr{V}_{\boldsymbol{\eta}}$ transposing v_1 and v_2. Then

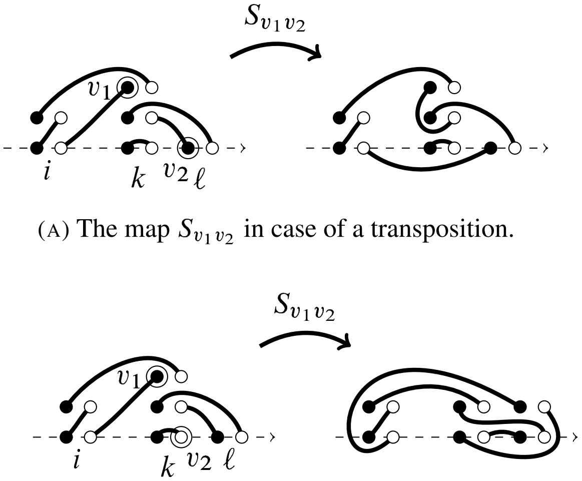
$$\mathscr{E}(S_{v_1 v_2} G) = \big\{\{\tau_{v_1,v_2}(w_1), \tau_{v_1,v_2}(w_2)\} : \{w_1, w_2\} \in \mathscr{E}(G)\big\}.$$

Assume $(v_1, v_2) \in \mathscr{V}_k^b \cdot \mathscr{V}_k^w$, and write $v_1 = (k, a_1, b)$, $v_2 = (k, a_2, w)$, for example, where $1 \le a_1, a_2 \le \eta_k$. Then we define $S_{v_1 v_2} G = S_{v_2 v_2} G \in \mathscr{G}_{\boldsymbol{\eta}^{k\ell}}$ as the perfect matching obtained by a jump of v_1 and v_2 to ℓ. More precisely, let $j_{v_1 v_2} = j_{v_2 v_1}$ be the following bijection from $\mathscr{V}_{\boldsymbol{\eta}}$ to $\mathscr{V}_{\boldsymbol{\eta}^{k\ell}}$: $j_{v_1 v_2}(v_1) = (\ell, \eta_\ell + 1, b)$, $j_{v_1 v_2}(v_2) = (\ell, \eta_\ell + 1, w)$, $j_{v_1 v_2}((k, c, b)) = (k, c - 1, b)$ if $a_1 < c$, $j_{v_1 v_2}((k, c, w)) = (k, c - 1, w)$ if $a_1 < c$, and $j_{v_1 v_2}(w_1) = w_1$ in all other cases. Then

$$\mathscr{E}(S_{v_1 v_2} G) = \big\{\{j_{v_1 v_2}(w_1), j_{v_1 v_2}(w_2)\} : \{w_1, w_2\} \in \mathscr{E}(G)\big\}.$$

A similar definition applies if $(v_1, v_2) \in \mathscr{V}_\ell^b \cdot \mathscr{V}_\ell^w$, the jump now being towards k.

Finally, if $(v_1, v_2) \notin (\mathscr{V}_k^b \cdot \mathscr{V}_\ell^b) \cup (\mathscr{V}_k^w \cdot \mathscr{V}_\ell^w) \cup (\mathscr{V}_k^b \cdot \mathscr{V}_k^w) \cup (\mathscr{V}_\ell^b \cdot \mathscr{V}_\ell^w)$, we define $S_{v_1 v_2} G = G$ (or any arbitrary function).



(A) The map $S_{v_1 v_2}$ in case of a transposition.

(B) The map $S_{v_1 v_2}$ in case of a jump.

Below is the main result for the proof of Theorem A.1.

LEMMA A.3. *For any $G \in \mathscr{G}_{\boldsymbol{\eta}}$, we have*

$$\begin{aligned}&\frac{1}{2}(X\bar{X} + \bar{X}X)P(G)\\ &\quad = \frac{1}{2} \sum_{(v_1, v_2) \in (\mathscr{V}_k \cup \mathscr{V}_\ell)_*^2} \varepsilon(v_1, v_2) P(S_{v_1 v_2} G) - (\eta_k + \eta_\ell) P(G).\end{aligned} \tag{A.6}$$

Assuming the above lemma we can complete the proof of Theorem A.1. Let

$$h(\boldsymbol{\eta}) = \sum_{G \in \mathscr{G}_{\boldsymbol{\eta}}} P(G).$$

Note that if $(v_1, v_2) \in (\mathscr{V}_k^b \cdot \mathscr{V}_\ell^b) \cup (\mathscr{V}_k^w \cdot \mathscr{V}_\ell^w)$, then $S_{v_1 v_2}$ is a permutation of $\mathscr{G}_{\boldsymbol{\eta}}$. Moreover, if $(v_1, v_2) \in \mathscr{V}_k^b \cdot \mathscr{V}_k^w$ (resp., $\mathscr{V}_\ell^b \cdot \mathscr{V}_\ell^w$) then $S_{v_1 v_2}$ is a bijection from $\mathscr{G}_{\boldsymbol{\eta}}$ to $\mathscr{G}_{\boldsymbol{\eta}^{k\ell}}$ (resp., $\mathscr{G}_{\boldsymbol{\eta}^{\ell k}}$). Summing (A.6) over all $G \in \mathscr{G}_{\boldsymbol{\eta}}$ therefore gives

$$\begin{aligned}
&\frac{1}{2}\big(X\bar{X} + \bar{X}X\big)h(\boldsymbol{\eta}) \\
&\quad = \frac{1}{2}\big(2\eta_k^2 h(\boldsymbol{\eta}^{k\ell}) + 2\eta_\ell^2 h(\boldsymbol{\eta}^{\ell k}) - (2\eta_k\eta_\ell + 2\eta_\ell\eta_k)h(\boldsymbol{\eta}) - 2(\eta_k + \eta_\ell)h(\boldsymbol{\eta})\big) \\
&\quad = \eta_k^2 h(\boldsymbol{\eta}^{k\ell}) + \eta_\ell^2 h(\boldsymbol{\eta}^{\ell k}) - (\eta_k(\eta_\ell + 1) + \eta_\ell(\eta_k + 1))h(\boldsymbol{\eta}).
\end{aligned}$$

The above equation implies (A.5) after renormalization by $\mathscr{L}(\boldsymbol{\eta})$. This concludes the proof of Theorem A.1.

PROOF OF LEMMA A.3. Let $L = \frac{1}{2}(X\bar{X} + \bar{X}X)$. We have

$$\begin{aligned}
LP(G) = &\sum_{(e_1,e_2)\in\mathscr{E}(G)_*^2} Xp(e_1)\bar{X}p(e_2) \prod_{e\in\mathscr{E}(G)\setminus\{e_1,e_2\}} p(e) \\
&+ \sum_{e_1\in\mathscr{E}(G)} Lp(e_1) \prod_{e\in\mathscr{E}(G)\setminus\{e_1\}} p(e).
\end{aligned} \tag{A.7}$$

We keep the notations (2.23), (2.24), and (2.25) for the single, double, and transverse edges. Remember that for any $v \in \mathscr{V}_{\boldsymbol{\eta}}$, e_v is the unique edge containing v, and v' is the unique vertex such that $e_v = \{v, v'\}$. We still denote

$$\begin{aligned}
\mathscr{V}_s &= \{v \in \mathscr{V}_k \cup \mathscr{V}_\ell : \{v, v'\} \in \mathscr{E}_s\}, \\
\mathscr{V}_d &= \{v \in \mathscr{V}_k \cup \mathscr{V}_\ell : \{v, v'\} \in \mathscr{E}_d\}, \\
\mathscr{V}_t &= \{v \in \mathscr{V}_k \cup \mathscr{V}_\ell : \{v, v'\} \in \mathscr{E}_t\},
\end{aligned}$$

and $\mathscr{V}_{k,s}^b$ are the single, black vertices in $\mathscr{V}_k$ (and similarly for $\mathscr{V}_{k,s}^w$, etc.). We will need the following elementary rules: if $i \neq \ell$ and $j \neq k$, $Xp_{ij} = 0$ and

$$Xp_{ik} = -p_{i\ell}, \quad Xp_{\ell j} = p_{kj}, \tag{A.8}$$

$$Xp_{\ell k} = p_{kk} - p_{\ell\ell}, \tag{A.9}$$

$$Xp_{kk} = -p_{k\ell}, \quad Xp_{\ell\ell} = p_{kk}. \tag{A.10}$$

We also obviously have $\bar{X}p = \overline{X\bar{p}}$. Equation (A.7) can be written as

$$LP(G) = \text{(I)} + \text{(II)} + \text{(III)} + \text{(IV)} + \text{(V)} + \text{(VI)} + \text{(VII)} + \text{(VIII)} + \text{(IX)}$$

where all terms are defined and calculated below. First,

$$\begin{aligned}
\text{(I)} :=& \sum_{(e_1,e_2)\in(\mathscr{E}_s)_*^2} Xp(e_1)\bar{X}p(e_2) \prod_{e\in\mathscr{E}(G)\setminus\{e_1,e_2\}} p(e) \\
=& \sum_{(v_1,v_2)\in(\mathscr{V}_s)_*^2} Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_2,v_2'\}} \prod_{e\in\mathscr{E}(G)\setminus\{e_{v_1},e_{v_2}\}} p(e).
\end{aligned}$$

From (A.8), $Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_2,v_2'\}} = -p_{\{v_2,v_1'\}}p_{\{v_1,v_2'\}}$ if $(v_1, v_2) \in (\mathscr{V}^b_{\ell,s} \times \mathscr{V}^b_{k,s}) \cup (\mathscr{V}^w_{k,s} \times \mathscr{V}^w_{\ell,s})$; $Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_2,v_2'\}} = p_{\{j_{v_1,v_2}(v_1),v_1'\}}p_{\{j_{v_1,v_2}(v_2),v_2'\}}$ if $(v_1, v_2) \in (\mathscr{V}^b_{k,s} \times \mathscr{V}^w_{k,s}) \cup (\mathscr{V}^b_{\ell,s} \times \mathscr{V}^w_{\ell,s})$. In all other cases, $Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_2,v_2'\}} = 0$. We therefore have proved

$$\begin{aligned}
(\mathrm{I}) &= \sum_{\substack{(v_1,v_2)\in(\mathscr{V}^b_{\ell,s}\times\mathscr{V}^b_{\ell,s})\cup(\mathscr{V}^w_{k,s}\times\mathscr{V}^w_{\ell,s})\\ \cup(\mathscr{V}^b_{k,s}\times\mathscr{V}^w_{k,s})\cup(\mathscr{V}^b_{\ell,s}\times\mathscr{V}^w_{\ell,s})}} \varepsilon(v_1, v_2)P(S_{v_1v_2}G) \\
&= \frac{1}{2}\sum_{(v_1,v_2)\in(\mathscr{V}_s)^2_*} \varepsilon(v_1, v_2)P(S_{v_1v_2}G).
\end{aligned} \tag{A.11}$$

We now consider

$$\begin{aligned}
(\mathrm{II}) &:= \sum_{(e_1,e_2)\in\mathscr{E}_s\cdot\mathscr{E}_d} Xp(e_1)\bar{X}p(e_2) \prod_{e\in\mathscr{E}(G)\setminus\{e_1,e_2\}} p(e) \\
&= \frac{1}{2}\sum_{(v_1,v_2)\in\mathscr{V}_s\cdot\mathscr{V}_d} Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_2,v_2'\}} \prod_{e\in\mathscr{E}(G)\setminus\{e_{v_1},e_{v_2}\}} p(e).
\end{aligned}$$

We used that vertices on a double edge need to be weighted by a factor $1/2$. From (A.8) and (A.10),

$$\begin{aligned}
&Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_2,v_2'\}} = -p_{\{v_2,v_1'\}}p_{\{v_1,v_2'\}} \\
&\text{if } (v_1, v_2) \in \big(\mathscr{V}^w_{k,d} \times \mathscr{V}^w_{\ell,s}\big) \cup \big(\mathscr{V}^b_{\ell,s} \times \mathscr{V}^b_{k,d}\big) \cup \big(\mathscr{V}^b_{\ell,d} \times \mathscr{V}^b_{k,s}\big)\cup\big(\mathscr{V}^w_{k,s} \times \mathscr{V}^w_{\ell,d}\big), \\
&Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_2,v_2'\}} = p_{\{j_{v_1v_2}(v_1),v_1'\}}p_{\{j_{v_1v_2}(v_2),v_2'\}} \\
&\text{if } (v_1, v_2) \in \big(\mathscr{V}^w_{k,d} \times \mathscr{V}^b_{k,s}\big) \cup \big(\mathscr{V}^b_{\ell,d} \times \mathscr{V}^w_{\ell,s}\big) \cup \big(\mathscr{V}^w_{k,s} \times \mathscr{V}^b_{k,d}\big) \cup \big(\mathscr{V}^b_{\ell,s} \times \mathscr{V}^w_{\ell,d}\big).
\end{aligned}$$

We therefore have

$$(\mathrm{II}) = \frac{1}{2}\sum_{(v_1,v_2)\in\mathscr{V}_s\cdot\mathscr{V}_d} \varepsilon(v_1, v_2)P(S_{v_1v_2}G). \tag{A.12}$$

Concerning

$$\begin{aligned}
(\mathrm{III}) &:= \sum_{(e_1,e_2)\in(\mathscr{E}_d)^2_*} Xp(e_1)\bar{X}p(e_2) \prod_{e\in\mathscr{E}(G)\setminus\{e_1,e_2\}} p(e) \\
&= \frac{1}{4}\sum_{(v_1,v_2)\in(\mathscr{V}_d)^2_*:v_1\neq v_2'} Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_2,v_2'\}} \prod_{e\in\mathscr{E}(G)\setminus\{e_{v_1},e_{v_2}\}} p(e),
\end{aligned}$$

using (A.10) we have $Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_2,v_2'\}} = -p_{\{v_2,v_1'\}}p_{\{v_1,v_2'\}}$ if v_1 and v_2 are in distinct $\mathscr{V}_i$'s and with the same color, and $p_{\{j_{v_1v_2}(v_1),v_1'\}}p_{\{j_{v_1v_2}(v_2),v_2'\}}$ if they are in the same $\mathscr{V}_i$ with distinct colors. All together, we always have

$$Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_2,v_2'\}} = \varepsilon(v_1, v_2)P(S_{v_1v_2}G) + \varepsilon(v_1', v_2)P(S_{v_1'v_2}G).$$

We have therefore proved

$$\begin{aligned}(\mathrm{III}) :={}& \frac{1}{4}\sum_{(v_1,v_2)\in(\mathscr{V}_d)^2_*}\big(\varepsilon(v_1,v_2)P(S_{v_1v_2}G)+\varepsilon(v_1',v_2)P(S_{v_1'v_2}G)\big)\\ &-\frac{1}{2}\sum_{v\in\mathscr{V}_d}P(S_{vv'}G)\\ ={}& \frac{1}{2}\sum_{(v_1,v_2)\in(\mathscr{V}_d)^2_*}\varepsilon(v_1,v_2)P(S_{v_1v_2}G)-\frac{1}{2}\sum_{v\in\mathscr{V}_d}P(S_{vv'}G).\end{aligned}\tag{A.13}$$

Our next estimate is a diagonal term, namely

$$\begin{aligned}(\mathrm{IV}) := \sum_{e_1\in\mathscr{E}_s}Lp(e_1)\prod_{e\in\mathscr{E}(G)\setminus\{e_1\}}p(e) &= \sum_{v\in\mathscr{V}_s}Lp_{\{v,v'\}}\prod_{e\in\mathscr{E}(G)\setminus\{e_v\}}p(e)\\ &= -\frac{1}{2}\sum_{v\in\mathscr{V}_s}P(G)\end{aligned}\tag{A.14}$$

where we used (A.8) twice to obtain $Lp_{\{v,v'\}}=-\frac{1}{2}p_{\{v,v'\}}$.

Another diagonal term is

$$(\mathrm{V}) := \sum_{e_1\in\mathscr{E}_d}Lp(e_1)\prod_{e\in\mathscr{E}(G)\setminus\{e_1\}}p(e) = \frac{1}{2}\sum_{v\in\mathscr{V}_d}Lp_{\{v,v'\}}\prod_{e\in\mathscr{E}(G)\setminus\{e_1\}}p(e).$$

Note that we have $Lp_{\{v,v'\}}=p_{kk}-p_{\ell\ell}$ if $v\in\mathscr{V}_\ell$, and $p_{\ell\ell}-p_{kk}$ otherwise. This proves

$$(\mathrm{V}) = \frac{1}{2}\sum_{v\in\mathscr{V}_d}\big(P(S_{vv'}(G))-P(G)\big).\tag{A.15}$$

We now consider cases where transverse edges appear:

$$\begin{aligned}(\mathrm{VI}) :={}& \sum_{(e_1,e_2)\in\mathscr{E}_s\times\mathscr{E}_t\cup\mathscr{E}_t\times\mathscr{E}_s}Xp(e_1)\bar{X}p(e_2)\prod_{e\in\mathscr{E}(G)\setminus\{e_1,e_2\}}p(e)\\ ={}& \sum_{\substack{v_1\in\mathscr{V}_s,\\ \{v_2,v_2'\}\in\mathscr{E}_t}}\big(Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_2,v_2'\}}+\bar{X}p_{\{v_1,v_1'\}}Xp_{\{v_2,v_2'\}}\big)\\ &\times\prod_{e\in\mathscr{E}(G)\setminus\{e_{v_1},e_{v_2}\}}p(e).\end{aligned}$$

Up to transposing v_2 and v_2', we can assume that v_1 and v_2 are in the same $\mathscr{V}_i$. With (A.8) and (A.9) a calculation gives $Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_2,v_2'\}}+\bar{X}p_{\{v_1,v_1'\}}Xp_{\{v_2,v_2'\}}=$

$p_{j_{v_1v_2}(v_1)v_1'}p_{j_{v_1v_2}(v_2)v_2'} - p_{\tau_{v_1v_2'}(v_1)v_1'}p_{\tau_{v_1v_2'}(v_2')v_2}$. This yields

$$\text{(A.16)}\qquad \begin{aligned}\text{(VI)} &= \sum_{(v_1,v_2)\in\mathscr{V}_s\times\mathscr{V}_t} \varepsilon(v_1,v_2)P(S_{v_1v_2}(G))\\ &= \frac{1}{2}\sum_{(v_1,v_2)\in\mathscr{V}_s\cdot\mathscr{V}_t} \varepsilon(v_1,v_2)P(S_{v_1v_2}(G)).\end{aligned}$$

We also consider

$$\begin{aligned}\text{(VII)} :&= \sum_{(e_1,e_2)\in\mathscr{E}_d\times\mathscr{E}_t\cup\mathscr{E}_t\times\mathscr{E}_d} Xp(e_1)\bar{X}p(e_2)\prod_{e\in\mathscr{E}(G)\setminus\{e_1,e_2\}} p(e)\\ &= \sum_{v_1\in\mathscr{V}_d,\{v_2,v_2'\}\in\mathscr{E}_t}\big(Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_1,v_2'\}} + \bar{X}p_{\{v_1,v_1'\}}Xp_{\{v_1,v_2'\}}\big)\\ &\qquad\times\prod_{e\in\mathscr{E}(G)\setminus\{e_v,e_w\}} p(e).\end{aligned}$$

Without loss of generality we can assume v_1 and v_2 are in the same $\mathscr{V}_i$. Assume they also have a different color. Then (A.9) and (A.10) give

$$\begin{aligned}&Xp_{\{v_1,v_1'\}}\bar{X}p_{\{v_1,v_2'\}} + \bar{X}p_{\{v_1,v_1'\}}Xp_{\{v_2,v_2'\}}\\ &\quad= p_{j_{v_1v_2}(v_1)v_1'}p_{j_{v_1v_2}(v_2)v_2'} - p_{\tau_{v_1v_2'}(v_1)v_1'}p_{\tau_{v_1v_2'}(v_2')v_2}.\end{aligned}$$

If v_1 and v_2 have the same color, a similar equation holds, permuting v_1 and v_1'. This implies

$$\text{(A.17)}\qquad \begin{aligned}\text{(VII)} &= \sum_{(v_1,v_2)\in\mathscr{V}_d\times\mathscr{V}_t} \varepsilon(v_1,v_2)P(S_{v_1v_2}(G))\\ &= \frac{1}{2}\sum_{(v_1,v_2)\in\mathscr{V}_d\cdot\mathscr{V}_t} \varepsilon(v_1,v_2)P(S_{v_1v_2}(G)).\end{aligned}$$

For two transverse edges, with (A.9) we first compute

$$\frac{1}{2}(Xp_{k\ell}\bar{X}p_{k\ell} + \bar{X}p_{k\ell}Xp_{k\ell}) = 0,$$

and indeed $\varepsilon(v_1,v_2) = 0$ when v_1, v_2 are the same color on the same site, or different colors on different sites. Moreover, $\frac{1}{2}(Xp_{k\ell}\bar{X}p_{\ell k} + \bar{X}p_{k\ell}Xp_{\ell k}) = \frac{1}{2}(p_{kk}^2 + p_{\ell\ell}^2 - 2p_{kk}p_{\ell\ell})$, so that in all cases we have proved

$$\text{(A.18)}\qquad \begin{aligned}\text{(VIII)} :&= \sum_{(e_1,e_2)\in(\mathscr{E}_t)^2_*} Xp(e_1)\bar{X}p(e_2)\prod_{e\in\mathscr{E}(G)\setminus\{e_1,e_2\}} p(e)\\ &= \frac{1}{2}\sum_{(v_1,v_2)\in(\mathscr{V}_t^2)_*} \varepsilon(v_1,v_2)P(S_{v_1v_2}(G).\end{aligned}$$

Finally, from (A.9) we have $Lp_{k\ell} = -p_{k\ell}$, so that

$$\text{(IX)} := \sum_{e_1 \in \mathscr{E}_t} Lp(e_1) \prod_{e \in \mathscr{E}(G) \setminus \{e_1\}} p(e) = -\frac{1}{2} \sum_{v \in \mathscr{V}_t} P(G). \tag{A.19}$$

By summation of all equations (A.11), (A.12), (A.13), (A.14), (A.15), (A.16), (A.17), (A.18), and (A.19), the right-hand sides of (A.6) and (2.22) are the same. This concludes the proof of Lemma A.3. □

Acknowledgment. The work of P.B. is partially supported by National Science Foundation Grant DMS#1513587. The work of H.-T. Y. is partially supported by National Science Foundation Grant DMS-1606305 and a Simons Investigator award. The work of J.Y. is partially supported by National Science Foundation Grant DMS#1552192.

Bibliography

[1] Anantharaman, N.; Le Masson, E. Quantum ergodicity on large regular graphs. *Duke Math. J.* **164** (2015), no. 4, 723–765. doi:10.1215/00127094-2881592

[2] Bao, Z.; Erdős, L. Delocalization for a class of random block band matrices. *Probab. Theory Related Fields* **167** (2017), no. 3-4, 673–776. doi:10.1007/s00440-015-0692-y

[3] Benaych-Georges, F.; Knowles, A. Local semicircle law for Wigner matrices. In *Advanced topics in random matrices*. *Panor. Synthèses*, 53, 1–90. Soc. Math. France, Paris, 2017.

[4] Bourgade, P.; Erdős, L.; Yau, H.-T.; Yin, J. Universality for a class of random band matrices. *Adv. Theor. Math. Phys.* **21** (2017), no. 3, 739–800. doi:10.4310/ATMP.2017.v21.n3.a5

[5] Bourgade, P.; Huang, J.; Yau, H.-T. Eigenvector statistics of sparse random matrices. *Electron. J. Probab.* **22** (2017), Paper No. 64, 38 pp. doi:10.1214/17-EJP81

[6] Bourgade, P.; Yang, F.; Yau, H.-T.; Yin, J. Random band matrices in the delocalized phase, II: generalized resolvent estimates. *J. Stat. Phys.* **174** (2019), no. 6, 1189–1221. doi:10.1007/s10955-019-02229-z

[7] Bourgade, P.; Yau, H.-T. The eigenvector moment flow and local quantum unique ergodicity. *Comm. Math. Phys.* **350** (2017), no. 1, 231–278. doi:10.1007/s00220-016-2627-6

[8] Bru, M.-F. Diffusions of perturbed principal component analysis. *J. Multivariate Anal.* **29** (1989), no. 1, 127–136. doi:10.1016/0047-259X(89)90080-8

[9] Casati, G.; Guarneri, I.; Izrailev, F.; Scharf, R. Scaling behavior of localization in quantum chaos. *Phys. Rev. Lett.* **64** (1990), no. 1. 5–8. doi:10.1103/PhysRevLett.64.5

[10] Casati, G.; Molinari, L.; Izrailev, F. Scaling properties of band random matrices. *Phys. Rev. Lett.* **64** (1990), no. 16, 1851–1854. doi:10.1103/PhysRevLett.64.1851

[11] Colin de Verdière, Y. Ergodicité et fonctions propres du laplacien. *Comm. Math. Phys.* **102** (1985), no. 3, 497–502.

[12] Disertori, M.; Pinson, H.; Spencer, T. Density of states for random band matrices. *Comm. Math. Phys.* **232** (2002), no. 1, 83–124. doi:10.1007/s00220-002-0733-0

[13] Efetov, K. *Supersymmetry in disorder and chaos*. Cambridge University Press, Cambridge, 1997.

[14] Erdős, L.; Knowles, A. Quantum diffusion and delocalization for band matrices with general distribution. *Ann. Henri Poincaré* **12** (2011), no. 7, 1227–1319. doi:10.1007/s00023-011-0104-5

[15] Erdős, L.; Knowles, A.; Yau, H.-T.; Yin, J. Delocalization and diffusion profile for random band matrices. *Comm. Math. Phys.* **323** (2013), no. 1, 367–416. doi:10.1007/s00220-013-1773-3

[16] Erdős, L.; Knowles, A.; Yau, H.-T; Yin, J. The local semicircle law for a general class of random matrices. *Electron. J. Probab.* **18** (2013), no. 59, 58 pp. doi:10.1214/EJP.v18-2473

[17] Erdős, L.; Yau, H.-T.; Yin, J. Bulk universality for generalized Wigner matrices. *Probab. Theory Related Fields* **154** (2012), no. 1-2, 341–407. doi:10.1007/s00440-011-0390-3

[18] Feingold, M.; Leitner, D. M.; Wilkinson, M. Spectral statistics in semiclassical random-matrix ensembles. *Phys. Rev. Lett.* **66** (1991), no. 8, 986–989. doi:10.1103/PhysRevLett.66.986

[19] Fyodorov, Y. V.; Mirlin, A. D. Scaling properties of localization in random band matrices: a σ-model approach. *Phys. Rev. Lett.* **67** (1991), no. 18, 2405–2409. doi:10.1103/PhysRevLett.67.2405

[20] He, Y.; Marcozzi, M. Diffusion profile for random band matrices: a short proof. *J. Stat. Phys.* **177** (2019), no. 4, 666–716. doi:10.1007/s10955-019-02385-2

[21] Knowles, A.; Yin, J. Anisotropic local laws for random matrices. *Probab. Theory Related Fields* **169** (2017), no. 1-2, 257–352. doi:10.1007/s00440-016-0730-4

[22] Landon, B.; Sosoe, P.; Yau, H.-T. Fixed energy universality of Dyson Brownian motion. *Adv. Math.* **346** (2019), 1137–1332. doi:10.1016/j.aim.2019.02.010

[23] Landon, B.; Yau, H.-T. Convergence of local statistics of Dyson Brownian motion. *Comm. Math. Phys.* **355** (2017), no. 3, 949–1000. doi:10.1007/s00220-017-2955-1

[24] McKean, H. P., Jr. *Stochastic integrals.* Probability and Mathematical Statistics, 5. Academic Press, New York–London, 1969.

[25] Nguyen, H.; Tao, T.; Vu, V. Random matrices: tail bounds for gaps between eigenvalues. *Probab. Theory Related Fields* **167** (2017), no. 3-4, 777–816. doi:10.1007/s00440-016-0693-5

[26] Peled, R.; Schenker, J.; Shamis, M.; Sodin, S. On the Wegner orbital model. *Int. Math. Res. Not. IMRN* (2019), no. 4, 1030–1058. doi:10.1093/imrn/rnx145

[27] Rudnick, Z.; Sarnak, P. The behaviour of eigenstates of arithmetic hyperbolic manifolds. *Comm. Math. Phys.* **161** (1994), no. 1, 195–213.

[28] Schenker, J. Eigenvector localization for random band matrices with power law band width. *Comm. Math. Phys.* **290** (2009), no. 3, 1065–1097. doi:10.1007/s00220-009-0798-0

[29] Shcherbina, T. On the second mixed moment of the characteristic polynomials of 1D band matrices. *Comm. Math. Phys.* **328** (2014), no. 1, 45–82. doi:10.1007/s00220-014-1947-7

[30] Shcherbina, T. Universality of the local regime for the block band matrices with a finite number of blocks. *J. Stat. Phys.* **155** (2014), no. 3, 466–499. doi:10.1007/s10955-014-0964-4

[31] Shcherbina, M.; Shcherbina, T. Characteristic polynomials for 1D random band matrices from the localization side. *Comm. Math. Phys.* **351** (2017), no. 3, 1009–1044. doi:10.1007/s00220-017-2849-2

[32] Šnirel′man, A. I. Ergodic properties of eigenfunctions. (Russian) *Uspekhi Mat. Nauk* **29** (1974), no. 6, 181–182.

[33] Sodin, S. The spectral edge of some random band matrices. *Ann. of Math. (2)* **172** (2010), no. 3, 2223–2251. doi:10.4007/annals.2010.172.2223

[34] Spencer, T. Random banded and sparse matrices. *The Oxford handbook of random matrix theory*, 471–488. Oxford University Press, Oxford, 2011.

[35] Tao, T.; Vu, V. Random matrices: universality of local eigenvalue statistics. *Acta Math.* **206** (2011), no. 1, 127–204. doi:10.1007/s11511-011-0061-3

[36] Wilkinson, M; Feingold, M; Leitner, D. M. Localization and spectral statistics in a banded random matrix ensemble. *J. Phys. A: Math. Gen.* **24** (1991), no. 1, 175–182. doi:10.1088/0305-4470/24/1/025

[37] Yang, F.; Yin, J. Random band matrices in the delocalized phase, III: Averaging fluctuations, In preparation, 2018.

[38] Zelditch, S. Uniform distribution of eigenfunctions on compact hyperbolic surfaces. *Duke Math. J.* **55** (1987), no. 4, 919–941. doi:10.1215/S0012-7094-87-05546-3

PAUL BOURGADE
Courant Institute
251 Mercer St.
New York, NY 10012
USA
E-mail: bourgade@cims.nyu.edu

JUN YIN
Department of Mathematics
6172 Math Science Hall
University of California, Los Angeles
Box 951555, MS 6304
Los Angeles, CA 90095
USA
E-mail: jyin@math.ucla.edu

Received August 2018.

HORNG-TZER YAU
Department of Mathematics
Harvard University
One Oxford Street
Cambridge, MA 02138-2901
USA
E-mail: htyau@math.harvard.edu