Universality for a class of random band matrices

P. Bourgade
New York University, Courant Institute
bourgade@cims.nyu.edu

L. Erdős
Institute of Science and Technology Austria
lerdos@ist.ac.at

H.-T. Yau
Harvard University
htyau@math.harvard.edu

J. Yin
University of Wisconsin, Madison
jyin@math.wisc.edu

We prove the universality for the eigenvalue gap statistics in the bulk of the spectrum for band matrices, in the regime where the band width is comparable with the dimension of the matrix, $W \sim N$. All previous results concerning universality of non-Gaussian random matrices are for mean-field models. By relying on a new mean-field reduction technique, we deduce universality from quantum unique ergodicity for band matrices.

Keywords: Universality, Band matrices, Dyson Brownian motion, Quantum unique ergodicity.

1 Introduction

1.1 Previous studies of Wigner and band matrices. There has been tremendous progress on the universality of non-invariant random matrices over the past decade. The basic model for such matrices, the Wigner ensemble, consists of $N \times N$ real symmetric or complex Hermitian matrices $H = (H_{ij})_{1 \leqslant i,j \leqslant N}$ whose matrix entries are identically distributed centered random variables that are independent up to the symmetry constraint $H = H^*$. The fundamental conjecture regarding the universality of the Wigner ensemble, the Wigner-Dyson-Mehta conjecture, states that the eigenvalue gap distribution is universal in the sense that it depends only on the symmetry class of the matrix, but is otherwise independent of the details of the distribution of the matrix entries. This conjecture has recently been established for all symmetry classes in a series of works [16, 19, 5] (see [14, 25, 38] for the Hermitian class of Wigner matrices). The approach initiated in [14, 16] to prove universality consists of three steps: (i) establish a local semicircle law for the density of eigenvalues (or more generally estimates on the Green functions); (ii) prove universality of Gaussian divisible ensembles, i.e., Wigner matrices with a small Gaussian component, by analyzing the convergence of Dyson Brownian motion to local equilibrium; (iii) remove the small Gaussian component by comparing Green functions of Wigner ensembles with those of Gaussian divisible ones. For an overview of universality results for Wigner matrices and this three-step strategy, see [18].

Wigner in fact predicted that universality should hold for any large quantum system, described by a Hamiltonian H, of sufficient complexity. One prominent example where random matrix statistics are ex-

The work of P. B. is partially supported by NSF grants DMS-1208859 and DMS-1513587. The work of L. E. is partially supported by ERC Advanced Grant, RANMAT 338804. The work of H.-T. Y. is partially supported by the NSF grant DMS-1307444 and the Simons investigator fellowship. The work of J. Y. is partially supported by NSF Grant DMS-1207961. The major part of this research was conducted when all authors were visiting IAS and were also supported by the NSF Grant DMS-1128255.

pected to hold is the random Schrödinger operator in the delocalized regime. The random Schrödinger operator describes a system with spatial structure, whereas Wigner matrices are mean-field models. Unfortunately, there has been virtually no progress in establishing the universality for the random Schrödinger operator in the delocalized regime. One prominent model interpolating between the Wigner matrices and the random Schrödinger operator is the *random band matrix*. In this model the physical state space, which labels the matrix elements, is equipped with a distance. Band matrices are characterized by the property that H_{ij} becomes negligible if $\mathrm{dist}(i,j)$ exceeds a certain parameter, W, called the *band width*. A fundamental conjecture [22] states that the local spectral statistics of a band matrix H are governed by random matrix statistics for large W and by Poisson statistics for small W. The transition is conjectured to be sharp [22, 37] for the band matrices in one spatial dimension around the critical value $W = \sqrt{N}$. In other words, if $W \gg \sqrt{N}$, we expect the universality results of [14, 16, 19, 5] to hold. Furthermore, the eigenvectors of H are expected to be completely delocalized in this range. For $W \ll \sqrt{N}$, one expects that the eigenvectors are exponentially localized. This is the analogue of the celebrated Anderson metal-insulator transition for random band matrices. The only rigorous work indicating the $\sqrt{N}$ threshold concerns the second mixed moments of the characteristic polynomial for a special class of Gaussian band matrices [33, 34]

The localization length for band matrices in one spatial dimension was recently investigated in numerous works. For general distribution of the matrix entries, eigenstates were proved to be localized [31] for $W \ll N^{1/8}$, and delocalization of most eigenvectors in a certain averaged sense holds for $W \gg N^{6/7}$ [13], improved to $W \gg N^{4/5}$ [12]. The Green's function $(H - z)^{-1}$ was controlled down to the scale $\mathrm{Im}\, z \gg W^{-1}$ in [20], implying a lower bound of order W for the localization length of all eigenvectors. When the entries are Gaussian with some specific covariance profiles, supersymmetry techniques are applicable to obtain stronger results. This approach has first been developed by physicists (see [11] for an overview); the rigorous analysis was initiated by Spencer (see [37] for an overview), with an accurate estimate on the expected density of states on arbitrarily short scales for a three-dimensional band matrix ensemble in [10]. More recent works include universality for $W = \Omega(N)$ [32], and the control of the Green's function down to the optimal scale $\mathrm{Im}\, z \gg N^{-1}$, hence delocalization in a strong sense for all eigenvectors, when $W \gg N^{6/7}$ [4] with first four moments matching the Gaussian ones (both results require a block structure and hold in part of the bulk spectrum). Our work is about statistics in the bulk of the spectrum, but we note that for universality at the spectral edge, much more is known [36]: extreme eigenvalues follow the Tracy-Widom law for $W \gg N^{5/6}$, an essentially optimal condition.

1.2 Difficulties and new ideas for general non mean-field models. In trying to use the above three-steps strategy for band matrices, let us first mention difficulties related to step (i), the local law. The Wigner-Dyson-Gaudin-Mehta conjecture was originally stated for Wigner matrices, but the methods of [14, 16] also apply to certain ensembles with independent but not identically distributed entries, which however retain the mean-field character of Wigner matrices. For generalized Wigner matrices with entries having varying variances, but still following the semicircle law, see [21], and more generally [1], where even the global density differs from the semicircle law. In particular, the local law up to the smallest scale N^{-1} can be obtained under the assumption that the entries of H satisfy

$$s_{ij} := \mathbb{E}(|H_{ij}|^2) \leqslant \frac{C}{N} \tag{1.1}$$

for some positive constant C. In this paper, we assume that $\sum_i s_{ij} = 1$; this normalization guarantees that the spectrum is supported on $[-2, 2]$. However, if the matrix entries vanish outside the band $|i-j| \lesssim W \ll N$, (1.1) cannot hold and the best known local semicircle law in this context [12] gives estimates only up to scale W^{-1}, while the optimal scale would be N^{-1}, comparable with the eigenvalue spacing. Hence for $W = N^{1-\delta}$, $\delta > 0$, the optimal local law is not known up to the smallest scale, which is a key source of difficulty for proving the delocalization of the band matrices. In this article, as $W = cN$ for some fixed small constant c, the local law holds up to the optimal scale.

While step (i) for the three-step strategy holds in this paper, steps (ii) and (iii) present a key hurdle to prove the universality. To explain this difficulty, consider Gaussian divisible matrices of the form $H_0 + \mathrm{GOE}(t)$, where H_0 is an arbitrary Wigner matrix and $\mathrm{GOE}(t)$ is a $N \times N$ Gaussian orthogonal ensemble

with matrix entries given by independent Brownian motions (up to the symmetry requirement) starting from 0. For any fixed time t, GOE(t) is a GOE matrix ensemble with variances of the matrix entries proportional to t. The basic idea for step (ii) is to prove the universality for matrices of the form $H_0 + \mathrm{GOE}(t)$ for t small, say, $t = N^{-1+\varepsilon}$ for some $\varepsilon > 0$. Finally, in step (iii), one shows that the eigenvalue statistics of the original matrix H can be approximated by $H_0 + \mathrm{GOE}(t)$ for a good choice of H_0. For $0 \leqslant \varepsilon < 1/2$ and H satisfying (1.1) with a matching lower bound $s_{ij} \geqslant c/N$, $c > 0$, up to a trivial rescaling we can choose $H_0 = H$ [7]. If $1/2 \leqslant \varepsilon < 1$, more complicated arguments requiring matching higher moments of the matrix entries are needed to choose an appropriate H_0 [20]. Unfortunately, both methods for this third step depend on the fact that the second moments of the entries of the original matrix match those of $H_0 + \mathrm{GOE}(t)$, up to rescaling. For band matrices, the variances outside the band vanish; therefore, the second moments of $H_0 + \mathrm{GOE}(t)$ and the band matrix H will never match outside the band. For the past years, this obstacle in step (iii) has been a major roadblock to extend the three-step strategy to the band matrices and to other non mean-field models. In this paper, we introduce a new method that overcomes this difficulty. In order to outline the main idea, we first need to describe the quantum unique ergodicity as proved in [7].

From the local law for band matrices [12] with $W = cN$, we have the complete delocalization of eigenvectors: with very high probability

$$\max |\psi_k(i)| \leqslant \frac{(\log N)^{C \log\log N}}{\sqrt{N}},$$

where C is a fixed constant and the maximum ranges over all coordinates i of all the ℓ^2-normalized eigenvectors, $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_N$. Although this bound prevents concentration of eigenvectors onto a set of size less than $N(\log N)^{-C \log\log N}$, it does not imply the "complete flatness" of eigenvectors in the sense that $|\psi_k(i)| \approx N^{-1/2}$. Recall the quantum ergodicity theorem (Shnirel'man [35], Colin de Verdière [8] and Zelditch [39]) asserts that "most" eigenfunctions for the Laplacian on a compact Riemannian manifold with ergodic geodesic flow are completely flat. For d-regular graphs under certain assumptions on the injectivity radius and spectral gap of the adjacency matrices, similar results were proved for eigenvectors of the adjacency matrices [3]. A stronger notion of quantum ergodicity, the quantum unique ergodicity (QUE) proposed by Rudnick-Sarnak [30] demands that *all* high energy eigenfunctions become completely flat, and it supposedly holds for negatively curved compact Riemannian manifolds. One case for which QUE was rigorously proved concerns arithmetic surfaces, thanks to tools from number theory and ergodic theory on homogeneous spaces [29, 23, 24].

For Wigner matrices, a probabilistic version of QUE was settled in [7]. In particular, it is known that there exists $\varepsilon > 0$ such that for any deterministic $1 \leqslant j \leqslant N$ and $I \subset [\![1, N]\!]$, for any $\delta > 0$ we have

$$\mathbb{P}\left(\Big| \sum_{i \in I} |\psi_j(i)|^2 - \frac{|I|}{N} \Big| \geqslant \delta \right) \leqslant N^{-\varepsilon}/\delta^2. \tag{1.2}$$

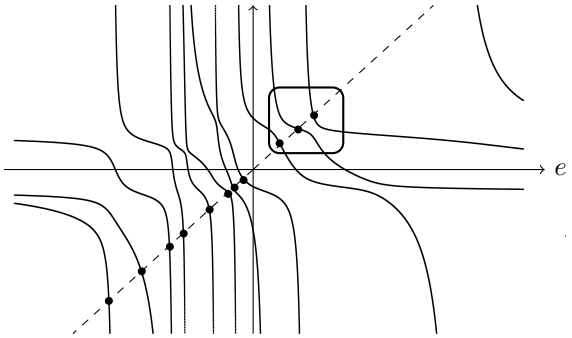
Our key idea for proving universality of band matrices is a *mean-field reduction*. In this method, the above probabilistic QUE will be a central tool. To explain the mean-field reduction and its link with QUE, we block-decompose the band matrix H and its eigenvectors as

$$H = \begin{pmatrix} A & B^* \\ B & D \end{pmatrix}, \quad \boldsymbol{\psi}_j := \begin{pmatrix} \mathbf{w}_j \\ \mathbf{p}_j \end{pmatrix}, \tag{1.3}$$

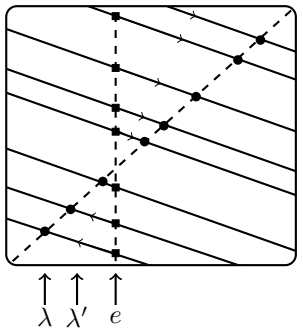
where A is a $W \times W$ matrix. From the eigenvector equation $H\boldsymbol{\psi}_j = \lambda_j \boldsymbol{\psi}_j$ we have

$$\Big(A - B^* \frac{1}{D - \lambda_j} B \Big) \mathbf{w}_j = \lambda_j \mathbf{w}_j, \tag{1.4}$$

i.e. $\mathbf{w}_j$ is an eigenvector of $A - B^*(D - \lambda_j)^{-1}B$, with corresponding eigenvalue λ_j. In agreement with the band structure, we may assume that the matrix elements of A do not vanish and thus the eigenvalue problem in (1.4) features a mean field random matrix (of smaller size).



(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 abscissa of the intersections with the diagonal.



(b) Zoom into the framed region of Figure (a), for large N, W: the curves ξ_j are almost parallel, with slope about $1-N/W$. The eigenvalues of $A-B^*(D-e)^{-1}B$ and those of H are related by a projection to the diagonal followed by a projection to the horizontal axis.

Figure 1: The idea of mean-field reduction: universality of gaps between eigenvalues for fixed e implies universality on the diagonal through parallel projection.

For a real parameter e, consider the following matrix

$$Q_e = A - B^* \frac{1}{D-e} B, \tag{1.5}$$

and let $\xi_k(e)$, $\mathbf{u}_k(e)$ be its sequence of eigenvalues and eigenvectors: $Q_e \mathbf{u}_k(e) = \xi_k(e)\mathbf{u}_k(e)$. Consider the curves $e \to \xi_k(e)$ (see Figure 1). By definition, the intersection points of these curves with the diagonal $e = \xi$ are eigenvalues for H, i.e., given j, we have $\xi_k(\lambda_j) = \lambda_j$ for some k. From this relation, we can find the eigenvalue λ_j near an energy e from the values of $\xi_k(e)$ provided that we know the slope of the curves $e \to \xi_k(e)$. It is a simple computation that this slope is given by $1 - (\sum_{i=1}^W \left|\psi_j'(i)\right|^2)^{-1}$, where $\boldsymbol{\psi}_j'$ is the eigenvector of H_e where H_e is the same as H except D is replaced by $D - e$ (see Subsection 2.2 for details). If the QUE in the sense of (1.2) holds for $\boldsymbol{\psi}_j'$, then $\sum_{i=1}^W \left|\psi_j'(i)\right|^2 \sim W/N$ and the leading order of the slope is a constant, independent of k. Therefore, the statistics of λ_j will be given by those of ξ_k up to a trivial scaling factor. Since ξ_k's are eigenvalues of a mean field random matrix, thanks to A, the universal statistics of ξ_k will follow from previous methods.

To summarize, our idea is to use the mean-field reduction to convert the problem of universality of the band matrices (H) to a matrix ensemble (Q_e) of the form $A + R$ with A a Wigner ensemble of the size of the band, independent of R. The key input for this mean-field reduction is the QUE for the big band matrix. This echoes the folklore belief that delocalization (or QUE) and random matrix statistics occur simultaneously. In fact, this is the first time that universality of random matrices is proved via QUE. We wish to emphasize that, as a tool for proving universality, we will need QUE while quantum ergodicity is not strong enough.

In order to carry out this idea, we need (i) to prove the QUE (2.7) for the band matrices; (ii) to show that the eigenvalue statistics of Q_e are universal. The last problem was recently studied in [17, 27] which can be applied to the current setting once some basic estimate for Q_e is obtained. The QUE for the band matrices, however, is a difficult problem. The method in [7] for proving QUE depends on analysis of the flow of the eigenvectors $H_0 + \mathrm{GOE}(t)$ and on the comparison between the eigenvectors of this matrix ensemble and those of the original matrices. Once again, due to vanishing matrix elements in H, we will not be able to use the comparison idea and the method in [7] cannot be applied directly. Our idea to resolve this difficulty is to use again the mean field reduction, this time for eigenvectors, and consider the eigenvector of the matrix Q_e. Recall the decomposition (1.3) of the band matrix. From (1.4), $\mathbf{w}_j$ is an eigenvector to Q_{λ_j}. Temporarily neglecting the fact that λ_j is random, we will prove that QUE holds for Q_e for any e fixed and thus $\mathbf{w}_j$ is

completely flat. This implies that the first W indices of $\boldsymbol{\psi}_j$ are completely flat. We now apply this procedure inductively to the decompositions of the band matrix where the role of $A = A_m$ will be played by the $W \times W$ minor on the diagonal of H between indices $mW/2+1$ and $(m+1)W/2$, where $m = 0, \dots, (2N-W)/W$ is an integer. Notice that the successively considered blocks $A_1, A_2, \dots, A_m$ overlap to guarantee consistency. Assuming QUE holds in each decomposition, we have concluded that $\boldsymbol{\psi}_j$ is completely flat by this patching procedure. This supplies the QUE we need for the band matrices, provided that we can resolve the technical problem that we need these results for $e = \lambda_j$, which is random. The resolution of this question relies on a new tool in analyzing non mean-field random matrices: an *uncertainty principle* asserting that whenever a vector is nearly an eigenvector, it is delocalized on macroscopic scales. This extends the delocalization estimate for eigenvectors to approximate eigenvectors and is of independent interest. This will be presented in Section 3.

Convention. We denote c (resp. C) a small (resp. large) constant which may vary from line to line but does not depend on other parameters. By $W = \Omega(N)$ we mean $W \geqslant cN$ and $[\![a, b]\!] := [a, b] \cap \mathbb{Z}$ refers to all integers between a and b.

2 Main results and sketch of the proof

2.1 The model and the results. Our method mentioned in the introduction applies to all symmetry classes, but for definiteness we will discuss the real symmetric case (in particular all eigenvectors are assumed to be real). Consider an $N \times N$ band matrix H with real centered entries that are independent up to the symmetry condition, and band width $4W-1$ (for notational convenience later in the paper) such that $N = 2Wp$ with some fixed $p \in \mathbb{N}$, i.e. in this paper we consider the case $W = \Omega(N)$. More precisely, we assume that

$$H_{ij} = 0, \text{ if } |i-j| > 2W, \tag{2.1}$$

where the distance $|\cdot|$ on $\{1, 2, \dots, N\}$ is defined by periodic boundary condition mod N. We set the variance

$$s_{ij} = \mathbb{E}\left(H_{ij}^2\right) = \frac{1}{4W-1}, \quad \text{if } |i-j| \leqslant 2W. \tag{2.2}$$

For simplicity, we assume identical variances within the band, but an upper and lower bound of order W^{-1} for each s_{ij} would suffice. We also assume that for some $\delta > 0$ we have

$$\sup_{N,i,j} \mathbb{E}\left(e^{\delta W H_{ij}^2}\right) < \infty. \tag{2.3}$$

This condition can be easily weakened to some finite moment assumption, we assume (2.3) mainly for the convenience of presentation. The eigenvalues of H are ordered, $\lambda_1 \leqslant \ldots \leqslant \lambda_N$, and we know that the empirical spectral measure $\frac{1}{N}\sum_{k=1}^N \delta_{\lambda_k}$ converges almost surely to the Wigner semicircle distribution with density

$$\rho_{\mathrm{sc}}(x) = \frac{1}{2\pi}\sqrt{(4-x^2)_+}. \tag{2.4}$$

Our main result is the universality of the gaps between eigenvalues: finitely many consecutive spacings between eigenvalues of H have the same limiting distribution as for the Gaussian Orthogonal Ensemble, GOE_N, which is known as the multi-dimensional Gaudin distribution.

Theorem 2.1. *Consider a band matrix H satisfying (2.1)–(2.3) with parameters $N = 2pW$. For any fixed $\kappa > 0$ and $n \in \mathbb{N}$ there exists an $\varepsilon = \varepsilon(p, \kappa, n) > 0$ such that for any smooth and compactly supported function O in $\mathbb{R}^n$, and $k \in [\![\kappa N, N - \kappa N]\!]$ we have*

$$\left| \left(\mathbb{E}^H - \mathbb{E}^{\mathrm{GOE}_N}\right) O\left(N\rho_{\mathrm{sc}}(\lambda_k)(\lambda_{k+1} - \lambda_k), \dots, N\rho_{\mathrm{sc}}(\lambda_i)(\lambda_{k+n} - \lambda_{k+n-1})\right) \right| \leqslant C_O N^{-\varepsilon}, \tag{2.5}$$

where the constant C_O depends only on κ and the test function O.

As mentioned in the introduction, a key ingredient for Theorem 2.1 is the quantum unique ergodicity of the eigenvectors of our band matrix model. In fact, we will need QUE for small perturbations of H on the diagonal: for any vector $\mathbf{g} = (g_1, \dots, g_N) \in \mathbb{R}^N$ we define

$$H^{\mathbf{g}} = H - \sum_{j=1}^{N} g_j \mathbf{e}_j \mathbf{e}_j^*, \tag{2.6}$$

where $\mathbf{e}_j$ is the j-th coordinate vector. Let $\lambda_1^{\mathbf{g}} \leqslant \ldots \leqslant \lambda_N^{\mathbf{g}}$ be the eigenvalues of $H^{\mathbf{g}}$ and $\boldsymbol{\psi}_k^{\mathbf{g}}$ be the corresponding eigenvectors, i.e. $H^{\mathbf{g}} \boldsymbol{\psi}_k^{\mathbf{g}} = \lambda_k^{\mathbf{g}} \boldsymbol{\psi}_k^{\mathbf{g}}$.

Theorem 2.2. *Consider a band matrix H satisfying (2.1)–(2.3) with parameters $N = 2pW$. Then for any small $\mathbf{g}$, $H^{\mathbf{g}}$ satisfies the QUE in the bulk. More precisely, there exists $\varepsilon, \zeta > 0$ such that for any fixed $\kappa > 0$, there exists $C_{\kappa,p} > 0$ such that for any $k \in [\![\kappa N, (1-\kappa)N]\!]$, $\delta > 0$, and $\mathbf{a} \in [-1,1]^N$, we have*

$$\sup_{\|\mathbf{g}\|_\infty \leqslant N^{-1+\zeta}} \mathbb{P}\left(\left| \sum_{i=1}^{N} \mathbf{a}(i) \left(|\psi_k^{\mathbf{g}}(i)|^2 - \frac{1}{N} \right) \right| \geqslant \delta \right) \leqslant C_{\kappa,p} N^{-\varepsilon} / \delta^2. \tag{2.7}$$

For the simplicity of exposition, we have stated the above result for QUE only at macroscopic scales (i.e., by choosing a bounded test vector $\mathbf{a}$), while it holds at any scale (like in [7]). The macroscopic scale will be enough for our proof of Theorem 2.1.

2.2 Sketch of the proof. In this outline of the proof, amongst other things we explain why QUE for small diagonal perturbation $H^{\mathbf{g}}$ of H is necessary to our mean-field reduction strategy. The role of other tools such as the uncertainty principle and the local law is also enlightened below.

We will first need some notation: we decompose $H^{\mathbf{g}}$ and its eigenvectors as

$$H^{\mathbf{g}} := \begin{pmatrix} A^{\mathbf{g}} & B^* \\ B & D^{\mathbf{g}} \end{pmatrix}, \quad \boldsymbol{\psi}_k^{\mathbf{g}} = \begin{pmatrix} \mathbf{w}_k^{\mathbf{g}} \\ \mathbf{p}_k^{\mathbf{g}} \end{pmatrix}, \quad k = 1, 2, \dots, W, \tag{2.8}$$

where $A^{\mathbf{g}}$ is a $W \times W$ matrix. The equation $H^{\mathbf{g}} \boldsymbol{\psi}_k^{\mathbf{g}} = \lambda_k^{\mathbf{g}} \boldsymbol{\psi}_k^{\mathbf{g}}$ then gives

$$\left(A^{\mathbf{g}} - B^* \frac{1}{D^{\mathbf{g}} - \lambda_k^{\mathbf{g}}} B \right) \mathbf{w}_k^{\mathbf{g}} = \lambda_k^{\mathbf{g}} \mathbf{w}_k^{\mathbf{g}}, \tag{2.9}$$

i.e. $\mathbf{w}_k^{\mathbf{g}}, \lambda_k^{\mathbf{g}}$ are the eigenvectors and eigenvalues of $Q^{\mathbf{g}}_{\lambda_k^{\mathbf{g}}}$ where we define

$$Q_e^{\mathbf{g}} := A^{\mathbf{g}} - B^* \frac{1}{D^{\mathbf{g}} - e} B \tag{2.10}$$

for any real parameter e. Notice that $A^{\mathbf{g}}$ depends only on $g_1, \cdots, g_W$ and $D^{\mathbf{g}}$ depends only on $g_{W+1}, \dots, g_N$. Let $\xi_1^{\mathbf{g}}(e) \leqslant \ldots \leqslant \xi_W^{\mathbf{g}}$ be the ordered sequence of eigenvalues of $Q_e^{\mathbf{g}}$ and $\mathbf{u}_k^{\mathbf{g}}(e)$ the corresponding eigenvectors:

$$Q_e^{\mathbf{g}} \mathbf{u}_k^{\mathbf{g}}(e) = \xi_k^{\mathbf{g}}(e) \mathbf{u}_k^{\mathbf{g}}(e). \tag{2.11}$$

We will be interested in a special class $g_i = g \mathbb{1}_{i>W}$ for some $g \in \mathbb{R}$, and we denote the matrix

$$H^g := \begin{pmatrix} A & B^* \\ B & D - g \end{pmatrix}, \tag{2.12}$$

and let ψ_j^g, λ_j^g be its eigenvectors and eigenvalues.

First step: From QUE of H^g to universality of H by mean-field reduction. Following Figure 1, we obtain eigenvalue statistics of H by parallel projection. Denote $\mathcal{C}_1, \dots, \mathcal{C}_N$ the continuous curves depicted in Figure 1b, labelled in increasing order of their intersection with the diagonal (see also Figure 3 and Section 4 for a formal definition of these curves).

Assume we are interested in the universality of the gap $\lambda_{k+1} - \lambda_k$ for some fixed $k \in [\![\kappa N, (1-\kappa)N]\!]$, and let $\xi > 0$ be a small constant. By some a priori local law, we know $|\lambda_k - e_0| \leqslant N^{-1+\xi}$ for some deterministic e_0, with overwhelming probability. Universality of the eigenvalue gaps around λ_k then follows from two facts: (i) universality of gaps between eigenvalues of Q_{e_0} in the local window $I = [e_0 - N^{-1+\xi}, e_0 + N^{-1+\xi}]$, (ii) the lines $(e \mapsto \mathcal{C}_j(e))_{j=k,k+1}$ have almost constant identical negative slope in the window $e \in I$.

For (i), note that the $Q_{e_0} = A + R$ where A is a mean-field, Wigner, random matrix and R is independent of A. For such matrices, bulk universality is known [28, 17, 27]. The key tools are some a priori rigidity estimates for the eigenvalues (see the fourth step), a coupling between Dyson Brownian motions [5] and Hölder estimates for a resulting parabolic equation [18].

For the key step (ii), the slopes are expressed through QUE properties of matrices of type H^g. More precisely, first note that any $e \in I$ can be written uniquely as

$$e = \lambda_k^g + g$$

for some $|g| \leqslant CN^{-1+\xi}$. Indeed, this is true for $e = \lambda_k$ with $g = 0$, and the function $g \to \lambda_k^g + g$ has a regular inverse, since by perturbative calculus $\partial(\lambda_k^g + g)/\partial g = \sum_{i=1}^W |\psi_k^g(i)|^2$, which is larger than some deterministic $c > 0$, with overwhelming probability, by the uncertainty principle detailed in the third step. Once such a writing of e is allowed, differentiating in g the identity $\mathcal{C}_k(\lambda_k^g + g) = \lambda_k^g$ (a direct consequence of (2.9)) gives

$$\frac{\partial}{\partial e}\mathcal{C}_k(e) = 1 - \left(\sum_{i=1}^W |\psi_k^g(i)|^2\right)^{-1}. \tag{2.13}$$

As a consequence, using QUE in the sense of Theorem 2.2, we know that $(\partial/\partial e)\mathcal{C}_k$ and $(\partial/\partial e)\mathcal{C}_{k+1}$ are almost constant, approximately $1 - (N/W)$. By parallel projection we obtain universality for H from universality of Q_{e_0}. In terms of scales, the average gap between eigenvalues of Q_{e_0} around e_0 is $(W\rho_{\mathrm{sc}}(e_0))^{-1}$, hence the average gap $\lambda_{k+1} - \lambda_k$ is $(N\rho_{\mathrm{sc}}(e_0))^{-1}$ as expected. This mean-field reduction strategy is detailed in Section 4.

Second step. Quantum unique ergodicity. The proof of Theorem 2.2 proceeds in four steps, with successive proofs of QUE for the following eigenvectors (k' is the unique index such that $\xi_{k'}$ lies on the curve $\mathcal{C}_k$):

(i) $\mathbf{u}_{k'}^g(e)$ ($\mathbf{a} \in [-1,1]^W$);

(ii) $\mathbf{u}_{k'}^g(\lambda_k^g)$ ($\mathbf{a} \in [-1,1]^W$);

(iii) $\mathbf{w}_k^g$ ($\mathbf{a} \in [-1,1]^W$);

(iv) $\boldsymbol{\psi}_k^g$ ($\mathbf{a} \in [-1,1]^N$).

In the parentheses we indicated the type of test vectors used in the QUE statement.

First, (i) is QUE for a matrix of type $Q_e = A + R$ where A is a mean-field, Wigner, random matrix and R is independent of A. For such matrices, QUE is known from the work [6], which made use of the local eigenvector moment flow method from [7]. For this step, some a priori information on location of eigenvalues of Q_e is necessary and given by the local law (see the the fourth step).

From (i) to (ii), some stability the eigenvectors of Q_e is required as e varies. Accurate estimates on $(\partial/\partial e)\mathbf{u}_{k'}^g(e)$ are given by the uncertainty principle (see the third step) and rigidity estimates of the eigenvalues (see the fourth step).

From (ii) to (iii), note that $\mathbf{w}_k^g$ and $\mathbf{u}_{k'}^g(\lambda_k^g)$ are collinear, so QUE for $\mathbf{w}_k^g$ will be proved provided it is properly normalized:

$$\|\mathbf{w}_k^g\|_{\ell^2}^2 \approx W/N. \tag{2.14}$$

This is proved by patching: in (ii), choosing $\mathbf{a}(i) = 1$ for $i \in [\![1, W/2]\!]$, -1 for $i \in [\![W/2+1, W]\!]$, and using translation invariance in our problem, we have $\sum_{i \in [\![1,W/2]\!] + \ell W/2} |\psi_k^g(i)|^2 \approx \sum_{i \in [\![1,W/2]\!] + (\ell+1)W/2} |\psi_k^g(i)|^2$ for any ℓ, so that (2.14) holds.

The final step from (iii) to (iv) is a simple consequence of translation invariance, as (iii) holds for any W successive coordinates of $\boldsymbol{\psi}_k^g$. These steps are detailed in Section 5.

Third step. Uncertainty principle. This important ingredient of the proof can be summarized as follows: any vector approximately satisfying the eigenvector equation of $H^{\mathbf{g}}$ or $D^{\mathbf{g}}$ is delocalized in the sense that macroscopic subsets of its coordinates carry a non-negligible portion of its ℓ^2 norm (see Proposition 3.1 for a precise statement). This information allows us to bound the slopes of the curves $e \mapsto \mathcal{C}_k(e)$ through (2.13). It is also important in the proof of the local law for matrices of type Q_e (see Lemma 6.5).

The proof of the uncertainty principle relies on an induction on q, where $N = qW$, classical large deviation estimates and discretization of the space arguments. Details are given in Section 3.

Fourth step. Local law. The local law for matrices of type Q_e is necessary for multiple purposes in the first two steps, most notably to establish universality of eigenvalues in a neighborhood of e and QUE for corresponding eigenvectors.

Note that the limiting empirical spectral distribution of Q_e is hard to be made explicit, and in this work we do not aim at describing it. Instead, we only prove bounds on the Green's function of Q_e *locally*, i.e.

$$(Q_e - z)_{ij}^{-1} \approx m(z)\delta_{ij}, \quad N^{-1+\omega} \leqslant \mathrm{Im}(z) \leqslant N^{-\omega},$$

in the range when $|\mathrm{Re}(z) - e|$ is small enough. Here $m(z)$ is the Stieltjes transform of the limiting spectral density whose precise form is irrelevant for our work. This estimate is obtained from the local law for the band matrix H [12] through Schur's complement formula. This *local* a priori information on eigenvalues (resp. eigenvectors) is enough to prove universality by Dyson Brownian motion coupling (resp. QUE through the eigenvector moment flow) strategy. The proof of the local law is given in Section 6.

In the above steps, we assumed that the entries of H have distribution which is a convolution with a small normal component (a *Gaussian-divisible ensemble*), so that the mean-field matrices Q_e are the result of a matrix Dyson Brownian motion evolution. This assumption is classically removed by density arguments such as the Green functions comparison theorem [20] or microscopic continuity of the Dyson Brownian motion [7], as will be appearent later along the proof.

3 Uncertainty principle

This section proves an uncertainty principle for our band matrices satisfying (2.1)–(2.3): if a vector approximately satisfies the eigenvalue equation, then it is delocalized on macroscopic scales.

Proposition 3.1. *Recall the notations (2.8). There exists $\mu > 0$ such that for any (small) $c > 0$ and (large) $D > 0$, we have, for large enough N,*

$$\mathbb{P}\left(\exists e \in \mathbb{R}, \exists \mathbf{u} \in \mathbb{R}^{N-W}, \exists \mathbf{g} \in \mathbb{R}^N \ : \ \|\mathbf{g}\|_\infty \leqslant N^{-c},\ \|\mathbf{u}\| = 1,\ \|(D^{\mathbf{g}} - e)\mathbf{u}\| \leqslant \mu,\ \sum_{1 \leqslant i \leqslant W} |u_i|^2 \leqslant \mu^2\right) \leqslant N^{-D}, \tag{3.1}$$

$$\mathbb{P}\left(\exists e \in \mathbb{R}, \exists \mathbf{g} \in \mathbb{R}^N \ : \ \|\mathbf{g}\|_\infty \leqslant N^{-c},\ B^* \frac{\mu^2}{(D^{\mathbf{g}} - e)^2} B \geqslant \left(B^* \frac{1}{D^{\mathbf{g}} - e} B\right)^2 + 1\right) \leqslant N^{-D} \tag{3.2}$$

This proposition gives useful information for two purposes.

(i) An a priori bound on the slopes of lines $e \mapsto \mathcal{C}_k^{\mathbf{g}}(e)$ (see Figure 3 in Section 4) will be provided by inequality (3.2).

(ii) The proof of the local law for the matrix $Q_e^{\mathbf{g}}$ will require the uncertainty principle (3.1).

For the proof, we first consider general random matrices in Subsection 3.1 before making an induction on the size of some blocks in Subsection 3.2.

3.1 Preliminary estimates. In this subsection, we consider a random matrix B of dimension $L \times M$ and a Hermitian matrix D of dimension $L \times L$ matrix where L and M are comparable. We have the decomposition (1.3) in mind and in the next subsection we will apply the results of this subsection, Lemma 3.2 and Proposition 3.3, for $M = W$ and $L = kW$ with some $k \in [\![1, 2p-1]\!]$. We assume that B has real independent, mean zero entries and, similarly to (2.3),

$$\sup_{M,i,j} \mathbb{E}\left(e^{\delta M B_{ij}^2}\right) < C_\delta < \infty \tag{3.3}$$

for some $\delta, C_\delta > 0$. In particular, we have the following bound:

$$\sup_{M,i,j} s_{ij} < \frac{C_\delta}{\delta M}, \qquad \text{where } s_{ij} := \mathbb{E}\left(|B_{ij}|^2\right). \tag{3.4}$$

The main technical tool, on which the whole section relies, is the following lemma.

Lemma 3.2. *Let B be an $L \times M$ random matrix satisfying the above assumptions and set $\beta := M/L$. Let S be a subspace of $\mathbb{R}^L$ with $\dim S =: \alpha L$. Then for any given γ and β, for small enough positive α, we have*

$$\mathbb{P}\Big(\exists \mathbf{u} \in S \,:\, \|\mathbf{u}\| = 1,\ \|B^*\mathbf{u}\| \leqslant \sqrt{\gamma}/4,\ \textit{and}\ \min_{1\leqslant j\leqslant M} \sum_{i=1}^{L} s_{ij}|u_i|^2 \geqslant \gamma M^{-1}\Big) \leqslant e^{-cL} \tag{3.5}$$

for large enough L. Here $0 < \alpha < \alpha_0(\beta, \gamma, \delta, C_\delta)$ and $c = c(\alpha, \beta, \gamma, \delta, C_\delta) > 0$.

Proof. With the replacement $B \to \sqrt{\gamma}B$, we only need to prove the case $\gamma = 1$ by adjusting δ to δ/γ. Hence in the following proof we set $\gamma = 1$.

First, we have an upper bound on the norm of BB^*. For any $T \geqslant T_0(\beta, \delta, C_\delta)$ (with δ, C_δ in (3.3)),

$$\mathbb{P}(\|BB^*\| \geqslant T) \leqslant e^{-c_1 TL} \tag{3.6}$$

for some small $c_1 = c_1(\beta) > 0$. This is a standard large deviation result, e.g. it was proved in [15, Lemma 7.3, part (i)] (this was stated when the B_{ij}'s are i.i.d, but only independence was used in the proof, the identical law was not).

Let $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_M \in \mathbb{R}^L$ be the columns of B, then $\|B^*\mathbf{u}\|^2 = \sum_{j=1}^M |\mathbf{b}_j \cdot \mathbf{u}|^2$. Since the $\mathbf{b}_j \cdot \mathbf{u}$ scalar products are independent, we have for any $g > 0$

$$\mathbb{P}\left(\|B^*\mathbf{u}\|^2 \leqslant 1/2\right) \leqslant e^{gM/2}\mathbb{E}\left(e^{-gM\|B^*\mathbf{u}\|^2}\right) = \prod_{j=1}^{M}\left(e^{g/2}\mathbb{E}\left(e^{-gM|\mathbf{b}_j\cdot\mathbf{u}|^2}\right)\right).$$

Since $e^{-gr} \leqslant 1 - gr + \frac{1}{2}g^2r^2$ for all $r > 0$, for $\|\mathbf{u}\| = 1$ we have

$$\mathbb{E}\left(e^{-gM|\mathbf{b}_j\cdot\mathbf{u}|^2}\right) \leqslant 1 - gM\mathbb{E}\left(|\mathbf{b}_j \cdot \mathbf{u}|\right)^2 + \frac{g^2M^2}{2}\mathbb{E}\left(|\mathbf{b}_j \cdot \mathbf{u}|^4\right) = 1 - Mg\sum_i \mathbb{E}\left(|B_{ij}|^2\right)|u_i|^2 + \mathrm{O}(g^2). \tag{3.7}$$

If $\mathbf{u}$ satisfies the last condition in the left hand side of (3.5), i.e. (with $\gamma = 1$) $\sum_i s_{ij}|u_i|^2 \geqslant M^{-1}$ for all $1 \leqslant j \leqslant M$ then (3.7) is bounded by $1 - g + \mathrm{O}(g^2) \leqslant \exp\big(-g + \mathrm{O}(g^2)\big)$. Choosing g sufficiently small, we have

$$\mathbb{P}(\|B^*\mathbf{u}\|^2 \leqslant 1/2) \leqslant \left(e^{-g/2+\mathrm{O}(g^2)}\right)^M \leqslant e^{-c_2 M} \tag{3.8}$$

where c_2 depends only on the constants δ, C_δ from (3.3).

Now we take an ε grid in the unit ball of S, i.e. vectors $\{\mathbf{u}_j \,:\, j \in I\} \subset S$ such that for any $\mathbf{u} \in S$, with $\|\mathbf{u}\| \leqslant 1$ we have $\|\mathbf{u} - \mathbf{u}_j\| \leqslant \varepsilon$ for some $j \in I$. It is well-known that $|I| \leqslant (c_3\varepsilon)^{-\dim S}$ for some constant c_3 of order one. We now choose $\varepsilon = (4\sqrt{T})^{-1}$ (where T is chosen large enough to satisfy (3.6)). If there

exists a $\mathbf{u}$ in the unit ball of S with $\|B^*\mathbf{u}\| \leqslant 1/4$ then by choosing j such that $\|\mathbf{u}-\mathbf{u}_j\| \leqslant \varepsilon$ we can bound $\|B^*\mathbf{u}_j\| \leqslant \|B^*\mathbf{u}\| + \sqrt{T}\|\mathbf{u}-\mathbf{u}_j\| \leqslant 1/2$, provided that $\|BB^*\| < T$. Hence together with (3.8), we have

$$\begin{aligned}\mathbb{P}\Big(\exists u \in S \,:\, \|B^*u\| \leqslant \frac{1}{4},\ \|u\|=1\Big) \leqslant\ & \mathbb{P}(\|BB^*\| \geqslant T) + \textstyle\sum_{j\in I}\mathbb{P}\Big(\|B^*u_j\| \leqslant \frac{1}{2}\Big) \\ \leqslant\ & e^{-c_1 TL} + (c_3\varepsilon)^{-\dim S} e^{-c_2 M} \leqslant e^{-cL},\end{aligned}$$

where the last estimate holds if

$$c \leqslant \alpha \log(c_3\varepsilon) + c_2\beta. \tag{3.9}$$

After the fixed choice of a sufficiently large constant T we have $\log(c_3\varepsilon) < 0$, and for small enough α there exists $c > 0$ such that (3.9) holds, and consequently (3.5) as well. □

Proposition 3.3. *Let D be an $L \times L$ deterministic matrix and B be a random matrix as in Lemma 3.2. Assume that D satisfies the following two conditions:*

$$\|D\| \leqslant C_D \tag{3.10}$$

for some large constant C_D (independent of L) and

$$\max_{a,b:|a-b|\leqslant (C_D \log L)^{-1}} \#\left\{\mathrm{Spec}(D) \cap [a,b]\right\} \leqslant \frac{L}{\log L}. \tag{3.11}$$

For any fixed $\gamma > 0$, there exists $\mu_0(\beta,\gamma,\delta,C_\delta,C_D) > 0$ such that if $\mu \leqslant \mu_0$, then for large enough L we have

$$\mathbb{P}\left(\exists e \in \mathbb{R},\ \exists \mathbf{u} \in \mathbb{R}^L \ :\ \|\mathbf{u}\|=1,\ \|B^*\mathbf{u}\| \leqslant \sqrt{\gamma}\mu,\ \min_{1\leqslant j\leqslant M}\sum_{i=1}^{L} s_{ij}|u_i|^2 \geqslant \gamma M^{-1}, \|(D-e)\mathbf{u}\| \leqslant \mu\right) \leqslant e^{-cL}. \tag{3.12}$$

Proof. We will first prove the following weaker statement: for any fixed $e \in \mathbb{R}$ and $\gamma > 0$, if $\mu \leqslant \mu_0(\beta,\gamma,C_\delta,C_D)$ is sufficiently small, then for large enough L we have

$$\mathbb{P}\left(\exists \mathbf{u} \ :\ \|\mathbf{u}\|=1,\ \|B^*\mathbf{u}\| \leqslant \sqrt{\gamma}\mu,\ \min_{j}\sum_{i} s_{ij}|u_i|^2 \geqslant \gamma M^{-1},\ \|(D-e)\mathbf{u}\| \leqslant \mu\right) \leqslant e^{-cL}. \tag{3.13}$$

As in the proof of Lemma 3.2, with the replacement $B \to \sqrt{\gamma}B$, we only need to prove the case $\gamma = 1$. Fix a small number ν and consider P to be the spectral projection

$$P := P_\nu := \mathbf{1}(|D-e| \leqslant \nu).$$

Assume there exists some $\mathbf{u}$ satisfying the conditions in the left hand side of (3.13). Then we have

$$\mu^2 \geqslant \|(D-e)\mathbf{u}\|^2 \geqslant \|(D-e)(1-P)\mathbf{u}\|^2 \geqslant \nu^2\|(1-P)\mathbf{u}\|^2.$$

Consequently, denoting $\mathbf{v} = P\mathbf{u}$ and $\mathbf{w} = (1-P)\mathbf{u}$, we have

$$\|\mathbf{w}\| \leqslant \frac{\mu}{\nu}, \qquad \|\mathbf{v}\|^2 \geqslant 1 - \frac{\mu^2}{\nu^2} \geqslant \frac{1}{2},$$

provided that $\mu^2 \leqslant \nu^2/2$. Using the bound $\|B^*\mathbf{u}\| \leqslant \mu$ in (3.13) and $\|\mathbf{v}\|^2 \geqslant 1/2$, assuming $\|B^*\| \leqslant C_1$ (this holds with probability e^{-cL} for large enough C_1, by (3.6)), we have

$$\|B^*\mathbf{v}\| \leqslant \|B^*\mathbf{u}\| + \|B^*\mathbf{w}\| \leqslant \mu + C_1\|\mathbf{w}\| \leqslant 2\mu\|\mathbf{v}\| + C_2\frac{\mu}{\nu}\|\mathbf{v}\| \tag{3.14}$$

with probability $1 - \mathrm{O}(e^{-cL})$. Moreover, by (3.4) and the assumption $\sum_i s_{ij}|u_i|^2 \geqslant M^{-1}$ in (3.13), we have

$$2\sum_i s_{ij}|v_i|^2 \geqslant \sum_i s_{ij}|u_i|^2 - 2\sum_i s_{ij}|w_i|^2 \geqslant M^{-1} - 2\widetilde{C}\|\mathbf{w}\|^2 L^{-1} \geqslant (2M)^{-1} \tag{3.15}$$

with $\widetilde{C} = C_\delta/(\delta\beta)$ (see (3.4)) and provided that $\nu^2 \geqslant 4\beta\widetilde{C}\mu^2$. Define $\widetilde{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|$, which is a unit vector in $\mathrm{Im}(P)$, the range of P. So far we proved that

$$\mathbb{P}\left(\exists \mathbf{u} \ : \ \|\mathbf{u}\| = 1,\ \|B^*\mathbf{u}\| \leqslant \mu,\ \min_j \sum_i s_{ij}|u_i|^2 \geqslant M^{-1},\ \|(D-e)\mathbf{u}\| \leqslant \mu\right)$$
$$\leqslant \mathbb{P}\left(\exists \widetilde{\mathbf{v}} \in \mathrm{Im}(P) \ : \ \|\widetilde{\mathbf{v}}\| = 1,\ \|B^*\widetilde{\mathbf{v}}\| \leqslant 2\mu + C_2\frac{\mu}{\nu},\quad \sum_i s_{ij}|\widetilde{v}_i|^2 \geqslant (4M)^{-1}\right) + e^{-cL}.$$

We now set μ and ν such that $2\mu + C_2\mu/\nu \leqslant 1/8$, $\mu^2 \leqslant \nu^2/2$ and $4\beta\widetilde{C}\mu^2 \leqslant \nu^2$. By Lemma 3.2, with $S := \mathrm{Im}(P)$ and $\gamma = 1/4$, the probability of the above event is exponentially small as long as

$$\mathrm{rank}(P)/L \quad i.e. \quad \#\left\{\mathrm{Spec}(D) \cap [e-\nu, e+\nu]\right\}/L$$

is sufficiently small (determined by β, δ, C_δ, see the threshold α_0 in Lemma 3.2). Together with (3.11), by writing the interval $[e-\nu, e+\nu]$ as a union of intervals of length $(C_D \log L)^{-1}$, by choosing small enough ν, then even smaller μ and finally a large L, we proved (3.13).

The proof of (3.12) follows by a simple grid argument. For fixed $\mu > 0$, consider a discrete set of energies $(e_i)_{i=1}^r$ such that (i) $r \leqslant 2(C_D+1)/\mu$, (ii) $|e_j| \leqslant C_D + 1$ for any $1 \leqslant j \leqslant r$ and (iii) for any $|e| \leqslant C_D + 1$, there is a $1 \leqslant j \leqslant r$ with $|e_j - e| \leqslant \mu$. If $|e| \leqslant C_D + 1$, we therefore have, for some $1 \leqslant j \leqslant r$,

$$\|(D-e_j)\mathbf{u}\| \leqslant \mu + |e - e_j| \leqslant 2\mu.$$

If $|e| > C_D + 1$, then $\|(D-e)\mathbf{u}\| \geqslant |e| - C_D > 1$. We therefore proved that, for any $\mu < 1$,

$$\mathbb{P}\left(\exists e \in \mathbb{R},\ \exists \mathbf{u} \in \mathbb{R}^L \ : \ \|\mathbf{u}\| = 1,\ \|B^*\mathbf{u}\| \leqslant \mu,\ \min_j \sum_i s_{ij}|u_i|^2 \geqslant M^{-1},\ \|(D-e)\mathbf{u}\| \leqslant \mu\right)$$
$$\leqslant \sum_{j=1}^r \mathbb{P}\left(\exists \mathbf{u} \in \mathbb{R}^L \ : \ \|\mathbf{u}\| = 1,\ \|B^*\mathbf{u}\| \leqslant \mu,\ \min_j \sum_i s_{ij}|u_i|^2 \geqslant M^{-1},\ \|(D-e_j)\mathbf{u}\| \leqslant 2\mu\right).$$

For large enough L, the right hand side is exponentially small by (3.13). □

3.2 Strong uncertainty principle. In this subsection, we study the matrix with the following block structure. Let $H = H_0$ be a $N \times N$ random matrix such that $\{H_{ij}\}_{i \leqslant j}$'s, are independent of each others. Consider the inductive decomposition

$$H_{m-1} = \begin{pmatrix} A_m & B_m^* \\ B_m & H_m \end{pmatrix}, \tag{3.16}$$

where A_m is a $W \times W$ matrix and H_m has dimensions $(N - mW) \times (N - mW)$. Remember that in our setting $N = 2pW$, so that the decomposition (3.16) is defined for $1 \leqslant m \leqslant 2p$ with $H_{2p-1} = A_{2p}$.

Lemma 3.4. *In addition to the previous assumptions, assume that the entries of B_m's, $1 \leqslant m \leqslant 2p$, satisfy* (3.3) *and*

$$\mathbb{E}|(B_m)_{ij}|^2 \geqslant \frac{\widehat{c}}{W} \quad \text{for all } 1 \leqslant i, j \leqslant W, \tag{3.17}$$

for some constant $\widehat{c} > 0$. For any $K > 0$, let $\Omega := \Omega_K(H)$ be the set of events such that

$$\|A_m\| + \|B_m\| + \|H_m\| \leqslant K, \tag{3.18}$$

and

$$\max_{a,b:|a-b| \leqslant K^{-1}(\log N)^{-1}} \#\left\{\mathrm{Spec}(H_m) \cap [a,b]\right\} \leqslant N/(\log N), \tag{3.19}$$

for all $0 \leqslant m \leqslant 2p$. Then there exist (small) μ_0 and c_0 depending on $(\widehat{c}, K, \delta, C_\delta, p)$, such that for any $0 < \mu < \mu_0$ and $0 \leqslant m \leqslant 2p-1$ we have

$$\mathbb{P}\Big(\exists e \in \mathbb{R},\ \exists \mathbf{u} \in \mathbb{R}^{N-mW} : \|\mathbf{u}\| = 1,\ \|(H_m - e)\mathbf{u}\| \leqslant \mu,\quad \sum_{1 \leqslant i \leqslant W} |u_i|^2 \leqslant \mu^2\Big) \leqslant e^{-c_0 N} + \mathbb{P}(\Omega^{\mathrm{c}}). \tag{3.20}$$

Proof. We will use an induction from $m = 2p-1$ to $m = 0$ to prove that for each $1 \leqslant m \leqslant 2p-1$ there exist two sequences of parameters $\mu_m \in (0,1)$ and $c_m > 0$, depending on $(\widehat{c}, K, \delta, C_\delta)$, such that

$$\mathbb{P}\left(\exists e \in \mathbb{R}, \mathbf{u} \in \mathbb{R}^{N-mW} :\ \|\mathbf{u}\| = 1,\ \|(H_m - e)\mathbf{u}\| \leqslant \mu_m,\ \sum_{1 \leqslant i \leqslant W} |u_i|^2 \leqslant \mu_m^2\right) \leqslant e^{-c_m N} + \mathbb{P}(\Omega^{\mathrm{c}}). \tag{3.21}$$

This would clearly imply (3.20). First the case $m = 2p-1$ is trivial, since we can choose $\mu_{2p-1} = 1/2$ and use $\sum_{1 \leqslant i \leqslant W} |u_i|^2 = \|\mathbf{u}\|^2 = 1$ in this case.

Now we assume that (3.21) has been proved for some $m+1$, and we need to prove it for m. Assume we are in Ω and there exists e and $\mathbf{u} \in \mathbb{R}^{N-mW}$ such that the event in the left hand side of (3.21) holds. We write $\mathbf{u} = \begin{pmatrix} \mathbf{v}' \\ \mathbf{v} \end{pmatrix}$ with $\mathbf{v}' \in \mathbb{R}^W$, $\|\mathbf{v}'\|^2 = \sum_{1 \leqslant i \leqslant W} |u_i|^2$. From $\|(H_m - e)\mathbf{u}\| \leqslant \mu_m$, we have

$$\|(A_{m+1} - e)\mathbf{v}' + B_{m+1}^* \mathbf{v}\| + \|B_{m+1}\mathbf{v}' + (H_{m+1} - e)\mathbf{v}\| \leqslant \sqrt{2}\mu_m.$$

Combining (3.18) with $\|(H_m - e)\mathbf{u}\| \leqslant \mu_m$, we have $|e| \leqslant K + \mu_m$. Inserting it in the above inequality together with $\|\mathbf{v}'\| \leqslant \mu_m$, and using (3.18) again, we obtain

$$\|B_{m+1}^* \mathbf{v}\| + \|(H_{m+1} - e)\mathbf{v}\| \leqslant \sqrt{2}\mu_m + (4K + 2\mu_m)\mu_m.$$

Since $\|\mathbf{v}\| \geqslant \sqrt{1 - \mu_m^2}$, denoting $\widetilde{\mathbf{v}} := \mathbf{v}/\|\mathbf{v}\|$ we have

$$\|B_{m+1}^* \widetilde{\mathbf{v}}\| + \|(H_{m+1} - e)\widetilde{\mathbf{v}}\| \leqslant \left(\sqrt{2}\mu_m + (4K + 2\mu_m)\mu_m\right)(1 - \mu_m^2)^{-1/2} =: \widetilde{\mu}_m.$$

We therefore proved

$$\begin{aligned}
&\mathbb{P}\left(\exists e \in \mathbb{R}, \exists \mathbf{u} \in \mathbb{R}^{N-mW} :\ \|\mathbf{u}\| = 1,\ \|(H_m - e)\mathbf{u}\| \leqslant \mu_m,\ \sum_{1 \leqslant i \leqslant W} |u_i|^2 \leqslant \mu_m^2\right) \\
\leqslant &\mathbb{P}\Big(\exists e \in \mathbb{R}, \exists \widetilde{\mathbf{v}} \in \mathbb{R}^{N-(m+1)W} :\ \|\widetilde{\mathbf{v}}\| = 1,\ \|B_{m+1}^* \widetilde{\mathbf{v}}\| + \|(H_{m+1} - e)\widetilde{\mathbf{v}}\| \leqslant \widetilde{\mu}_m\Big) + \mathbb{P}(\Omega^{\mathrm{c}}) \\
\leqslant &\mathbb{P}\left(\exists e \in \mathbb{R}, \exists \widetilde{\mathbf{v}} \in \mathbb{R}^{N-(m+1)W} :\ \|\widetilde{\mathbf{v}}\| = 1,\ \|B_{m+1}^* \widetilde{\mathbf{v}}\| + \|(H_{m+1} - e)\widetilde{\mathbf{v}}\| \leqslant \widetilde{\mu}_m,\ \sum_{1 \leqslant i \leqslant W} |\widetilde{v}_i|^2 \geqslant \mu_{m+1}^2\right) \\
&\quad + e^{-c_{m+1} N} + \mathbb{P}(\Omega^{\mathrm{c}}),
\end{aligned}$$

where in the last inequality we used the induction hypothesis (at rank $m+1$) and we assumed that $\widetilde{\mu}_m \leqslant \mu_{m+1}$, which holds by choosing μ_m small enough.

With (3.17) the last probability is bounded by

$$\mathbb{P}\Big(\exists e \in \mathbb{R},\ \exists \widetilde{\mathbf{v}} \in \mathbb{R}^{N-(m+1)W} :\ \|\widetilde{\mathbf{v}}\| = 1,\ \|B_{m+1}^* \widetilde{\mathbf{v}}\| + \|(H_{m+1} - e)\widetilde{\mathbf{v}}\| \leqslant \widetilde{\mu}_m, \\ \min_{1 \leqslant j \leqslant W} \sum_i \mathbb{E}\,|(B_{m+1})_{ij}|^2\, |\widetilde{v}_i|^2 \geqslant \mu_{m+1}^2 \frac{\widehat{c}}{W}\Big)$$

Applying (3.12) with $\mu = \widetilde{\mu}_m$ and $\gamma = \widehat{c}\mu_{m+1}^2$, together with assumption (3.19), we know that for small enough μ_m (and therefore small enough $\widetilde{\mu}_m$), the above probability is bounded by $e^{-\widetilde{c}N}$ for some $\widetilde{c} > 0$. Therefore (3.21) holds at rank m if we define c_m recursively backwards such that $c_m < \min\{c_{m+1}, \widetilde{c}\}$. The sequence μ_m may also be defined recursively backwards with an initial $\mu_{2p-1} = 1/2$ so that each $\widetilde{\mu}_m$ remains smaller than μ_{m+1} and the small threshold $\mu_0(\beta, \gamma = \widehat{c}\mu_{m+1}^2, \delta, C_\delta, C_D)$ from Proposition 3.3. □

Corollary 3.5. *Under the assumptions of Lemma 3.4, there exist (small) $\widetilde{\mu}_0$ and $\widetilde{c}$ depending on $(\widehat{c}, K, \delta, C_\delta, p)$, such that for any $0 < \mu < \mu_0$ we have*

$$\mathbb{P}\left(\exists e \in \mathbb{R} \ : \ B_1^* \frac{\mu^2}{(H_1 - e)^2} B_1 \geqslant \left(B_1^* \frac{1}{H_1 - e} B_1\right)^2 + 1\right) \leqslant \mathbb{P}(\Omega^{\mathrm{c}}) + e^{-\widetilde{c}N}. \tag{3.22}$$

Proof. By definition, the left hand side of (3.22) is

$$\mathbb{P}\left(\exists e \in \mathbb{R},\ \exists \mathbf{u} \in \mathbb{R}^W \ : \ \|\mathbf{u}\| = 1,\ \mu \Big\| \frac{1}{H_1 - e} B_1 \mathbf{u} \Big\| \geqslant \left\{ \Big\| B_1^* \frac{1}{H_1 - e} B_1 \mathbf{u} \Big\|^2 + 1 \right\}^{1/2}\right)$$

Define $\mathbf{v} := (H_1 - e)^{-1} B_1 \mathbf{u}$, and $\widetilde{\mathbf{v}} := \mathbf{v}/\|\mathbf{v}\|$. As $\|B_1\| \leqslant K$ in Ω, the above probability is bounded by

$$\begin{aligned}
&\mathbb{P}\left(\exists e \in \mathbb{R},\ \exists \mathbf{v} \in \mathbb{R}^{N-W} \ : \ \mu \|\mathbf{v}\| \geqslant \left(\|B_1^* \mathbf{v}\|^2 + 1\right)^{1/2},\ \ \|(H_1 - e)\mathbf{v}\| \leqslant K\right) + \mathbb{P}(\Omega^{\mathrm{c}}) \\
&\qquad \leqslant \mathbb{P}\left(\exists e \in \mathbb{R},\ \exists \widetilde{\mathbf{v}} \in \mathbb{R}^{N-W} \ : \ \|\widetilde{\mathbf{v}}\| = 1,\ \|B_1^* \widetilde{\mathbf{v}}\| \leqslant \mu,\ \ \|(H_1 - e)\widetilde{\mathbf{v}}\| \leqslant K\mu\right) + \mathbb{P}(\Omega^{\mathrm{c}}).
\end{aligned}$$

With (3.20) (choosing $m = 1$), for any $\mu \leqslant \mu_0$, where μ_0 was obtained in Lemma 3.4, the above expression is bounded by

$$\begin{aligned}
&\mathbb{P}\left(\exists e \in \mathbb{R},\ \exists \widetilde{\mathbf{v}} \in \mathbb{R}^{N-W} \ : \ \|\widetilde{\mathbf{v}}\| = 1,\ \|B_1^* \widetilde{\mathbf{v}}\| \leqslant \mu,\ \ \|(H_1 - e)\widetilde{\mathbf{v}}\| \leqslant K\mu,\ \ \sum_{1 \leqslant i \leqslant W} |\widetilde{\mathbf{v}}_i|^2 \geqslant \mu_0^2\right) + e^{-c_0 N} + \mathbb{P}(\Omega^{\mathrm{c}}) \\
&\qquad \leqslant \mathbb{P}\Bigg(\exists e \in \mathbb{R},\ \exists \widetilde{\mathbf{v}} \in \mathbb{R}^{N-W} \ : \ \|\widetilde{\mathbf{v}}\| = 1,\ \|B_1^* \widetilde{\mathbf{v}}\| \leqslant \mu,\ \ \|(H_1 - e)\widetilde{\mathbf{v}}\| \leqslant K\mu, \\
&\qquad\qquad \min_{1 \leqslant j \leqslant W} \sum_{1 \leqslant i \leqslant W} \mathbb{E}|(B_1)_{ij}|^2 |\widetilde{\mathbf{v}}_i|^2 \geqslant \mu_0^2 \frac{\widehat{c}}{W}\Bigg) + e^{-c_0 N} + \mathbb{P}(\Omega^{\mathrm{c}}).
\end{aligned}$$

For the last inequality we used (3.17). From (3.12) with $\gamma = \widehat{c}\mu_0^2$ and for small enough $\mu \leqslant \widetilde{\mu}_0 := \mu_0(\beta, \gamma = \widehat{c}\mu_0^2, \delta, C_\delta, C_D)$, the above term is bounded by $\mathbb{P}(\Omega) + e^{-\widetilde{c}N}$ for some $\widetilde{c} > 0$, which completes the proof of Corollary 3.5. □

Proof of Proposition 3.1. We write $H^{\mathbf{g}}$ in the from of (3.16). Then $H^{\mathbf{g}}$ satisfies the assumptions (3.3) and (3.17). Define $\Omega := \Omega_K(H)$ as in (3.18) and (3.19). Lemma 3.4 and Corollary 3.5 would thus immediately prove (3.1)–(3.2) if $\mathbf{g}$ were fixed. To guarantee the bound simultaneously for any $\mathbf{g}$, we only need to prove that there exists a fixed (large) $K > 0$ such that for any $D > 0$ we have

$$\mathbb{P}\Big(\bigcup_{\|\mathbf{g}\| \leqslant N^{-c}} \Omega_K(H^{\mathbf{g}})\Big) \leqslant N^{-D}$$

if N is large enough. This is just a crude bound on the norm of band matrices which can be proved by many different methods. For example, by perturbation theory, we can remove $\mathbf{g}$ and thus we only need to prove $\mathbb{P}\big(\Omega_K(H^{\mathbf{0}})\big) \leqslant N^{-D}$. This follows easily from the rigidity of the eigenvalues of the matrix H (see [1, Corollary 1.10]). □

4 Universality

In this section, we prove the universality of band matrix H (Theorem 2.1) assuming the QUE for the band matrices of type $H^{\mathbf{g}}$ (Theorem 2.2). In the first subsection, we remind some a priori information on the location of the eigenvalues of the band matrix. The following subsections give details for the mean-field reduction technique previously presented.

4.1 Local semicircle law for band matrices. We first recall several known results concerning eigenvalues and Green function estimates for band matrices. For $e \in \mathbb{R}$ and $\omega > 0$, we define

$$\mathbf{S}(e, N; \omega) = \big\{ z = E + \mathrm{i}\eta \in \mathbb{C} : |E - e| \leqslant N^{-\omega}\,,\, N^{-1+\omega} \leqslant \eta \leqslant N^{-\omega} \big\}, \tag{4.1}$$

$$\widehat{\mathbf{S}}(e, N; \omega) = \big\{ z = E + \mathrm{i}\eta \in \mathbb{C} : |E - e| \leqslant N^{-\omega}\,,\, N^{-1+\omega} \leqslant \eta \leqslant 1 \big\}. \tag{4.2}$$

In this section, we are interested only in $\widehat{\mathbf{S}}$; the other set $\mathbf{S}$ will be needed later on. We will drop the dependence in N whenever it is obvious. We view ω as an arbitrarily small number playing few active roles and we will put all these type of parameters after semicolon. In the statement below, we will also need $m(z)$, the Stieltjes transform of the semicircular distribution, i.e.

$$m(z) = \int \frac{\varrho_{\mathrm{sc}}(s)}{s - z} \mathrm{d}s = \frac{-z + \sqrt{z^2 - 4}}{2}, \tag{4.3}$$

where ϱ_{sc} is the semicircle distribution defined in (2.4) and the square root is chosen so that m is holomorphic in the upper half plane and $m(z) \to 0$ as $z \to \infty$. The following results on the Green function $G(z) = (H - z)^{-1}$ of the random band matrix H and its normalized trace $m(z) = m_N(z) = \frac{1}{N} \operatorname{Tr} G(z)$ have been proved in [12].

In the following theorem and more generally in this paper, the notation $A_N \prec B_N$ means that for any (small) $\varepsilon > 0$ and (large) $D > 0$ we have $\mathbb{P}(|A_N| > N^{\varepsilon} |B_N|) \leqslant N^{-D}$ for large enough $N \geqslant N_0(\varepsilon, D)$. When A_N and B_N depend on a parameter (typically on z in some set $\mathbf{S}$ or some label) then by $A_N(z) \prec B_N(z)$ uniformly in $z \in \mathbf{S}$ we mean that the threshold $N_0(\varepsilon, D)$ may be chosen independently of z.

Theorem 4.1 (Local semicircle law, Theorem 2.3 in [12])**.** *For the band matrix ensemble defined by (2.1) and (2.2), satisfying the tail condition* (2.3)*, uniformly in $z \in \widehat{\mathbf{S}}(e, W; \omega)$ we have*

$$\max_{i,j} \big| G_{ij}(z) - \delta_{ij} m(z) \big| \prec \sqrt{\frac{\operatorname{Im} m(z)}{W\eta}} + \frac{1}{W\eta}, \tag{4.4}$$

$$\big| m_N(z) - m(z) \big| \prec \frac{1}{W\eta}. \tag{4.5}$$

We now recall the following rigidity estimate of the eigenvalues for band matrices [12]. This estimate was first proved for generalized Wigner matrices in [21] (for our finite band case this latter result would be sufficient). We define the classical location of the j-th eigenvalue by the equation

$$\frac{j}{N} = \int_{-\infty}^{\gamma_j} \varrho_{\mathrm{sc}}(x) \mathrm{d}x. \tag{4.6}$$

Corollary 4.2 (Rigidity of eigenvalues, Theorem 2.2 of [21] or Theorem 7.6 in [12])**.** *Consider the band matrix ensemble defined by (2.1) and (2.2), satisfying the tail condition* (2.3)*, and $N = 2pW$ with p finite. Then, uniformly in $j \in [\![1, N]\!]$, we have*

$$|\lambda_j - \gamma_j| \prec \big(\min\big(j, N - j + 1 \big) \big)^{-1/3} N^{-2/3}. \tag{4.7}$$

4.2 Mean field reduction for Gaussian divisible band matrices. Recall the definition of H^g from (2.12),

$$H^g = \begin{pmatrix} A & B^* \\ B & D - g \end{pmatrix}, \tag{4.8}$$

i.e. A has dimensions $W \times W$, B has dimensions $(N - W) \times W$ and D has dimensions $(N - W) \times (N - W)$. Its eigenvalues and eigenvectors are denoted by λ_j^g and $\boldsymbol{\psi}_j^g$, $1 \leqslant j \leqslant N$.

Almost surely, there is no multiple eigenvalue for any g, i.e. the curves $g \to \lambda_j^g$ do not cross, as shown both by absolute continuity argument (we consider Gaussian-divisible ensembles) and by classical codimension

counting argument (see [9, Theorem 5.3]). In particular, the indexing is consistent, i.e. if we label them in increasing order for g near $-\infty$, $\lambda_1^{-\infty} < \lambda_2^{-\infty} < \ldots < \lambda_N^{-\infty}$, then the same order will be kept for any g:

$$\lambda_1^g < \lambda_2^g < \ldots < \lambda_N^g. \tag{4.9}$$

Moreover, the eigenfunctions are well defined (modulo a phase and normalization) and by standard perturbation theory, the functions $g \to \lambda_j^g$ and $g \to \boldsymbol{\psi}_j^g$ are analytic functions (very strictly speaking in the second case these are analytic functions into homogeneous space of the unit ball of $\mathbb{C}^N$ modulo $U(1)$). Moreover, by variational principle $g \to \lambda_j^g$ are decreasing. In fact, they are strictly decreasing (almost surely) and they satisfy

$$-1 < \frac{\partial \lambda_k^g}{\partial g} < 0 \tag{4.10}$$

since by perturbation theory we have

$$\frac{\partial \lambda_k^g}{\partial g} = -1 + \sum_{i=1}^{W} |\psi_k^g(i)|^2 \tag{4.11}$$

and $0 < \sum_{i=1}^{W} |\psi_k^g(i)|^2 < \|\boldsymbol{\psi}_k^g\|_2 = 1$ almost surely. We may also assume (generically), that A and D have simple spectrum, and denote their spectra

$$\sigma(A) = \{\alpha_1 < \alpha_2 < \ldots < \alpha_W\}, \qquad \sigma(D) = \{\delta_1 < \delta_2 < \ldots < \delta_{N-W}\}.$$

We claim the following behavior of λ_j^g for $g \to \pm\infty$ (see Figure 2a):

$$\lambda_j^g = \begin{cases} \alpha_j + \mathrm{O}\left(|g|^{-1}\right), & \text{for } j \leqslant W, \\ -g + \delta_{j-W} + \mathrm{O}\left(|g|^{-1}\right), & \text{for } W < j \leqslant N, \end{cases} \quad \text{as } g \to -\infty, \tag{4.12}$$

$$\lambda_j^g = \begin{cases} -g + \delta_j + \mathrm{O}\left(|g|^{-1}\right), & \text{for } j \leqslant N-W, \\ \alpha_{j-(N-W)} + \mathrm{O}\left(|g|^{-1}\right), & \text{for } N-W < j \leqslant N, \end{cases} \quad \text{as } g \to \infty. \tag{4.13}$$

Notice that the order of labels is consistent with (4.9). The above formulas are easy to derive by simple analytic perturbation theory. For example, for $g \to -\infty$ and $j \leqslant W$ we use

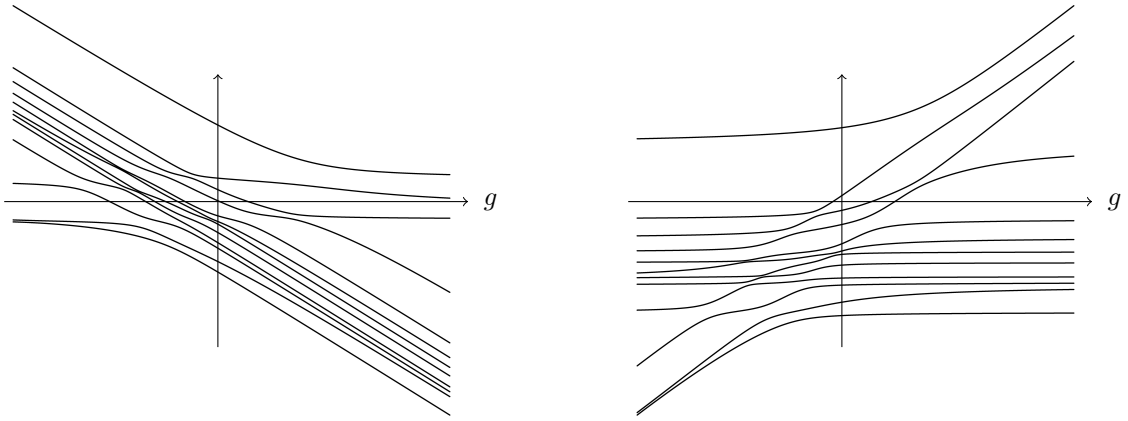
$$\Big(A - B^* \frac{1}{D - g - \lambda^g} B\Big)\mathbf{w}_j^g = \lambda_j^g \mathbf{w}_j^g.$$

Let $\mathbf{q}_j$ be the eigenvector of A corresponding to α_j, $A\mathbf{q}_j = \alpha_j \mathbf{q}_j$, then we can express $\mathbf{w}_j^g = \mathbf{q}_j + \Delta\mathbf{q}_j$ and $\lambda_j^g = \alpha_j + \Delta\alpha_j$, plug it into the formula above and get that $\Delta\mathbf{q}_j, \Delta\alpha_j = \mathrm{O}(|g|^{-1})$.

The formulas (4.12) and(4.13) together with the information that the eigenvalue lines do not cross and that the functions $g \to \lambda_j^g$ are strictly monotone decreasing, give the following picture. The lowest W lines, $g \to \lambda_j^g$, $j \leqslant W$ start at $g \to -\infty$ almost horizontally at the levels $\alpha_1, \alpha_2, \ldots, \alpha_W$ and go down linearly, shifted with $\delta_1, \ldots, \delta_W$ at $g \to \infty$. The lines $g \to \lambda_j^g$, $W < j \leqslant N - W$ start decreasing linearly at $g \to -\infty$, shifted with $\delta_1, \delta_2, \ldots, \delta_{N-W}$ (in this order) and continue to decrease linearly at $g \to \infty$ but shifted with $\delta_{W+1}, \delta_{W+2}, \ldots \delta_N$. Finally, the top lines, $g \to \lambda_j^g$, $N - W < j \leqslant N$, start decreasing linearly at $g \to -\infty$, shifted with $\delta_{N-2W+1}, \ldots, \delta_{N-W}$ and become almost horizontal at levels $\alpha_1, \alpha_2, \ldots, \alpha_W$ for $g \to \infty$.

Similarly one can draw the curves $g \to x_j(g) := \lambda_j^g + g$ (see Figure 2b) Since $x_j(g)$ is an increasing function w.r.t. $g \in \mathbb{R}$ by (4.10), with (4.12)-(4.13), it is easy to check that

$$\operatorname{Ran} x_j = \begin{cases} (-\infty, \delta_j), & j \leqslant W, \\ (\delta_{j-W}, \delta_j), & W < j \leqslant N - W, \\ (\delta_{j-W}, \infty), & N - W < j \leqslant N. \end{cases} \tag{4.14}$$



(a) The maps $g \mapsto \lambda_j^g$, $1 \leqslant j \leqslant N$.

(b) The maps $g \mapsto \lambda_j^g + g = x_j(g)$, $1 \leqslant j \leqslant N$.

Figure 2: The eigenvalues of H^g (left) and $H^g + g\,\mathrm{Id}$ (right) for $N = 12$ and $W = 3$.

From this description it is clear that for any $e \notin \sigma(D)$, the equation $x_j(g) = e$ has exactly W solutions, namely

$$\big\{j : \exists g, \text{ s.t. } x_j(g) = e\big\} = \begin{cases} [\![1, W]\!], & e < \delta_1, \\ [\![m+1, m+W]\!], & \delta_m < e < \delta_{m+1}, \\ [\![N-W+1, N]\!], & e > \delta_{N-W}. \end{cases} \tag{4.15}$$

For any such j, the corresponding g is unique by strict monotonicity of $x_j(g)$, thus this function can be locally inverted. Finally, for any j, we define the following curves:

$$\mathcal{C}_j(e) = \lambda_j^g, \quad s.t. \quad e = x_j(g) = \lambda_j^g + g.$$

Their domains are defined as follows:

$$\mathrm{Dom}\,\mathcal{C}_j = \begin{cases} (-\infty, \delta_j), & j \leqslant W, \\ (\delta_{j-W}, \delta_j), & W < j \leqslant N - W, \\ (\delta_{j-W}, \infty), & N - W < j \leqslant N. \end{cases} \tag{4.16}$$

From the definition of $\mathcal{C}$ it is clear that these are smooth functions, since they are compositions of two smooth functions: $g \to \lambda_j^g$ and the inverse of $x_j(g)$.

Finally, by just comparing the definition of $\xi_j(e)$ in (2.11) for $\mathbf{g} = 0$, we know that if $\mathcal{C}_k(e)$ exists then it is one of the eigenvalues of Q_e: $\mathcal{C}_k(e) = \xi_{k'}(e)$ for some k'. Moreover, we know that almost surely there is no e such that Q_e has multiple eigenvalues (see [9, Theorem 5.3]), so we can assume the curves $(\mathcal{C}_k)_{1 \leqslant k \leqslant N}$ do not intersect. This proves

$$\mathcal{C}_k(e) = \xi_{k'}(e), \text{with } k' = k'(e) = k - \mathcal{N}_D(e) \tag{4.17}$$

where we defined $\mathcal{N}_D(e) = |\sigma(D) \cap (-\infty, e)|$ the number of eigenvalues of D smaller than e.

The above discussion is summarized as follows, and extended to more general matrices, $A^{\mathbf{g}}$ and $D^{\mathbf{g}}$ instead of A and D. We stress that the parameter g in the above discussion was an auxiliary variable and its role independent of the fixed $\mathbf{g}$ in the definition below.

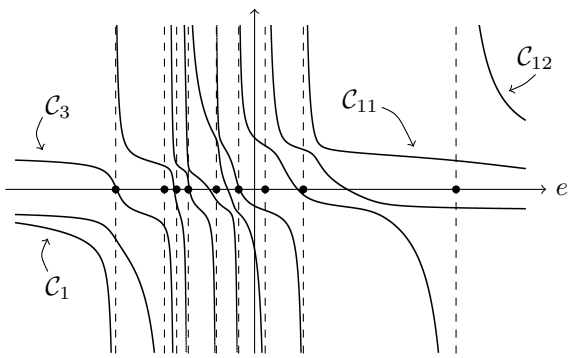


Figure 3: A sample of curves $(\mathcal{C}_k^{\mathbf{g}})_{1 \leqslant k \leqslant N}$ for $N = 12, W = 3$. Bullet points are eigenvalues of $D^{\mathbf{g}}$.

Definition 4.3 (Curves $\mathcal{C}_k^{\mathbf{g}}(e)$)**.** *Fix any $\mathbf{g} \in \mathbb{R}^N$ parameter vector. The curves $\mathcal{C}_k^{\mathbf{g}}(e)$ are the continuous extensions of $\xi_{k-\mathcal{N}_{D^{\mathbf{g}}}(e)}(e)$. More precisely, we have*

$$\mathcal{C}_k^{\mathbf{g}}(e) = \xi_{k'}^{\mathbf{g}}(e), \quad k' = k - \mathcal{N}_{D^{\mathbf{g}}}(e) \tag{4.18}$$

for any $e \notin \sigma(D^{\mathbf{g}})$.

The result below shows that the slopes of these curves are uniformly bounded, for ordinates on compact sets.

Lemma 4.4. *Consider any fixed (large) $K > 0$. There exists a constant C_K such that for any (small) $\zeta > 0$ and any (large) $D > 0$ we have, for large enough N,*

$$\mathbb{P}\left(\exists \mathbf{g} : \|\mathbf{g}\|_\infty \leqslant N^{-\zeta}, \sup_{e \notin \sigma(D), 1 \leqslant k \leqslant N} \mathbb{1}_{|\mathcal{C}_k^{\mathbf{g}}(e)| \leqslant K} \Big| \frac{\mathrm{d}\mathcal{C}_k^{\mathbf{g}}}{\mathrm{d}e}(e) \Big| \leqslant C_K \right) \leqslant N^{-D}. \tag{4.19}$$

Proof. We first note that for $e \notin \sigma(D)$,

$$\Big| \frac{\mathrm{d}\mathcal{C}_k^{\mathbf{g}}}{\mathrm{d}e} \Big| = \Big\| \frac{1}{D^{\mathbf{g}} - e} B \mathbf{u}_{k'}^{\mathbf{g}}(e) \Big\|^2, \quad k' = k - \mathcal{N}_{D^{\mathbf{g}}}(e) \tag{4.20}$$

by differentiating (2.11) w.r.t. e and multiplying it by $\mathbf{u}_{k'}^{\mathbf{g}}(e)$. Here we used that $k' = k'(e)$ is constant as e varies between two consecutive eigenvalues of D. By (2.10) and (2.11), with $\|\mathbf{u}_{k'}^{\mathbf{g}}(e)\| = 1$, we have that

$$\Big\| B^* \frac{1}{D^{\mathbf{g}} - e} B \mathbf{u}_{k'}^{\mathbf{g}}(e) \Big\| \leqslant \|A^{\mathbf{g}}\| + |\mathcal{C}_k^{\mathbf{g}}(e)|. \tag{4.21}$$

Using Proposition 3.1 and that $\|A^{\mathbf{g}}\| \leqslant C$ holds with high probability, we have for any $e \notin \sigma(D)$ that

$$\Big\| \frac{1}{D^{\mathbf{g}} - e} B \mathbf{u}_{k'}^{\mathbf{g}}(e) \Big\|^2 \leqslant \frac{2}{\mu^2} \Big\| B^* \frac{1}{D^{\mathbf{g}} - e} B \mathbf{u}_{k'}^{\mathbf{g}}(e) \Big\|^2 + \frac{1}{\mu^2} \leqslant C_\mu (1 + |\mathcal{C}_k^{\mathbf{g}}(e)|^2) \tag{4.22}$$

for all $\|\mathbf{g}\|_\infty \leqslant N^{-\zeta}$, with high probability, where in the last step we used (4.21). Together with (4.20), we have proved (4.19). □

The following theorem summarizes the key idea of the mean-field reduction.

Theorem 4.5. *Let $\theta \in (0,1)$ be fixed. Let H be a Gaussian divisible band matrix of type*

$$H = \sqrt{q} H_1 + \sqrt{1-q} H_2, \quad q = W^{-1+\theta}, \tag{4.23}$$

where H_1, H_2 are independent band matrices of width $4W-1$, satisfying (2.1)–(2.2), and let H_1 have Gaussian entries. Recall that $\boldsymbol{\psi}_j^g$ is the eigenvector of H^g defined in (2.12). Fix an energy $e_0 \in (-2,2)$ and let k satisfy $|\gamma_k - e_0| \leqslant N^{-1}(\log N)$. Suppose that all $\boldsymbol{\psi}_j^g$ are flat for $|j-k| \leqslant \log N$ and $|g| \leqslant N^{-1+\zeta}$ for some $\zeta > 0$, in the sense that

$$\sup_{\substack{j\,:\,|j-k| \leqslant \log N \\ |g| \leqslant N^{-1+\zeta}}} \mathbb{E} \left| \sum_{i=1}^W \big|\psi_j^g(i)\big|^2 - W/N \right| \leqslant N^{-\zeta}. \tag{4.24}$$

Then for any fixed constant C we have (here $\lambda_j = \lambda_j^{g=0}$), for large enough N,

$$\sup_{j:|j-k| \leqslant C} \mathbb{P}\left(\left| \mathcal{C}_j(e_0) - e_0 - \frac{N}{W}(\lambda_j - e_0) \right| \geqslant N^{-1-\zeta/5} \right) \leqslant N^{-\zeta/5}. \tag{4.25}$$

Proof. We will prove (4.25) only for $j = k$, the general case clearly follows a similar argument. Denote by $\mathcal{R}$ the set of matrices H such that

$$|\lambda_k - e_0| \leqslant N^{-1+\zeta/2}, \quad \text{and} \quad \sup_{e \notin \sigma(D)} \mathbb{1}_{|\mathcal{C}_k(e)| \leqslant 3} \Big| \frac{\mathrm{d}\mathcal{C}_k(e)}{\mathrm{d}e} \Big| \leqslant C.$$

By the assumption $|\gamma_k - e_0| \leqslant N^{-1}\log N$ and the rigidity of λ_k (see Corollary 4.2), for any $\zeta > 0$ the first condition above holds with high probability. As guaranteed by Lemma 4.4, the second condition in the definition of $\mathcal{R}$ holds with high probability for a large enough C. Hence, for such ζ and C, for any $D > 0$ and large enough N we have $\mathbb{P}(\mathcal{R}) \geqslant 1 - N^{-D}$.

In this proof, we will assume that $\lambda_k > e_0$ for simplicity of notations. In $\mathcal{R}$, we have

$$\Big\{ (e, \mathcal{C}_k(e)) \ : \ e \subset [e_0, \lambda_k] \Big\} \in \Big[e_0 - N^{-1+\zeta/2}, e_0 + N^{-1+\zeta/2} \Big]^2, \qquad \sup_{e \in [e_0,\lambda_k]\setminus\sigma(D)} \Big| \frac{\mathrm{d}\mathcal{C}_k(e)}{\mathrm{d}e} \Big| \leqslant C. \tag{4.26}$$

Recall that the function $\mathcal{C}_k$ satisfies the relation

$$\mathcal{C}_k\left(g + \lambda_k^g\right) = \lambda_k^g. \tag{4.27}$$

Differentiating (4.27) at the point $e = g + \lambda_k^g$, and using (4.11), we have

$$\frac{\mathrm{d}\mathcal{C}_k(e)}{\mathrm{d}e} = \frac{\partial_g \lambda_k^g}{1 + \partial_g \lambda_k^g} = \frac{\sum_{i=1}^W |\psi_k^g(i)|^2 - 1}{\sum_{i=1}^W |\psi_k^g(i)|^2}. \tag{4.28}$$

Hence there is a constant c such that in $\mathcal{R}$ we have

$$\inf_{e \in [e_0,\lambda_k]\setminus\sigma(D)} \sum_{i=1}^W |\psi_k^g(i)|^2 \geqslant c. \tag{4.29}$$

Since $\mathcal{C}_k(\lambda_k) = \lambda_k$ at $g = 0$, we have

$$\mathcal{C}_k(e_0) - e_0 = \int_{\lambda_k}^{e_0} \left(\frac{\mathrm{d}\mathcal{C}_k(e)}{\mathrm{d}e} - 1 \right) \mathrm{d}e = \int_{e_0}^{\lambda_k} \left(\sum_{i=1}^W |\psi_k^g(i)|^2 \right)^{-1} \mathrm{d}e,$$

where e and g are related by $g + \lambda_k^g = e$. Using the above equation, a simple calculation gives (remember $N = 2pW$ and c is defined in (4.29))

$$\mathbb{E}\mathbb{1}_{\mathcal{R}} \left| \mathcal{C}_k(e_0) - e_0 - \frac{N}{W}\left(\lambda_k - e_0\right) \right| \leqslant \frac{2p}{c} \mathbb{E}\mathbb{1}_{\mathcal{R}} \int_{e_0}^{\lambda_k} \left| \sum_{i=1}^W |\psi_k^g(i)|^2 - W/N \right| \mathrm{d}e. \tag{4.30}$$

The integration domain is over $e = g + \lambda_k^g \in [e_0, \lambda_k]$ with $g = 0$ when $e = \lambda_k$, and $g = g_0$ when $e = e_0$ with the g_0 that satisfies $e_0 = g_0 + \lambda_k^{g_0}$. Notice that in the set $\mathcal{R}$ we have

$$\frac{\mathrm{d}g}{\mathrm{d}e} = \left(\frac{\mathrm{d}\lambda_k^g}{\mathrm{d}g} + 1 \right)^{-1} = \left(\sum_{i=1}^W |\psi_k^g(i)|^2 \right)^{-1} \in [1, c^{-1}],$$

which implies $|g_0| \leqslant c^{-1}|\lambda_k - e_0| \leqslant c^{-1} N^{-1+\zeta/2}$, i.e., g is in the domain required for using (4.24). Therefore, we can insert the estimate (4.24) into (4.30) and conclude that

$$\mathbb{E}\mathbb{1}_{\mathcal{R}} \left| \mathcal{C}_k(e_0) - e_0 - \frac{N}{W}\left(\lambda_k - e_0\right) \right| \leqslant \frac{2p}{c} N^{-1-\zeta/2}.$$

This implies (4.25) and completes the proof of the theorem. □

4.3 Proof of Theorem 2.1. We will first prove Theorem 2.1 for the class of Gaussian divisible band matrix ensemble, which was defined in (4.23). We will prove general case at the end of this section. Recall $Q_e = A - B^*(D-e)^{-1}B$ where, for the Gaussian divisible band matrix ensemble, we can decompose A as

$$A = \sqrt{q}A_1 + \sqrt{1-q}A_2, \tag{4.31}$$

where A_1 and A_2 are independent and A_1 is a standard $W \times W$ GOE matrix. For a smooth test function O of n variables with compact support, define the following observable of the rescaled eigenvalue gaps of H:

$$O_{k,n}(\boldsymbol{\lambda}, N) := O\left(N\rho_{\mathrm{sc}}(\lambda_k)(\lambda_{k+1}-\lambda_k), \ldots, N\rho_{\mathrm{sc}}(\lambda_k)(\lambda_{k+n}-\lambda_{k+n-1})\right). \tag{4.32}$$

Our goal is to prove that for some $c > 0$, for any $k \in [\![\kappa N, (1-\kappa)N]\!]$ we have

$$(\mathbb{E}^H - \mathbb{E}^{\mathrm{GOE}_N})O_{k,n}(\boldsymbol{\lambda}, N) \leqslant N^{-c}.$$

Given k, let the (nonrandom) energy $e_0 \in (-2,2)$ be such that $|e_0 - \gamma_k| \leqslant (\log N)N^{-1}$. We claim that we can choose e_0 with $|e_0 - \gamma_k| \leqslant (\log N)N^{-1}$ such that

$$\mathbb{P}(\|(D-e_0)^{-1}\| \geqslant N^4) \leqslant N^{-1}. \tag{4.33}$$

To prove this, we note that $\|(D-e_0)^{-1}\| \geqslant N^4$ is equivalent to $\min_\ell |\delta_\ell - e_0| \leqslant N^{-4}$, where, remember, that the spectrum of D is denoted $\delta_1 < \delta_2 < \ldots < \delta_{N-W}$. For any $\sigma(D)$ fixed, we have the trivial bound

$$\int_{\gamma_k-(\log N)N^{-1}}^{\gamma_k+(\log N)N^{-1}} \mathbb{1}_{\min_\ell |\delta_\ell - e_0| \leqslant N^{-4}} \mathrm{d}e_0 \leqslant N^{-5/2}.$$

Taking expectation of the last inequality w.r.t the probability law of D and using the Markov inequality, we have proved (4.33). We remark that, by smoothness of O and by rigidity (Corollary 4.2), $\rho_{\mathrm{sc}}(\lambda_k)$ can be replaced by $\rho_{\mathrm{sc}}(e_0)$ in (4.32).

Denote $\mathbb{E}^{Q_e}$ the expectation w.r.t the law of Q_e induced from the distribution of the original band matrix H and let $\boldsymbol{\xi}(e) = (\xi_1(e), \xi_2(e), \ldots, \xi_W(e))$ be the ordered spectrum of Q_e. From the approximate affine transformation between the λ and ξ eigenvalues, guaranteed by Theorem 4.5, we have

$$\mathbb{E}^H O_{k,n}(\boldsymbol{\lambda}, N) = \mathbb{E}^{Q_{e_0}} O_{k-\alpha,n}(\boldsymbol{\xi}(e_0), W) + \mathrm{O}(N^{-c}), \quad \alpha := \mathcal{N}_D(e_0),$$

where we used Definition 4.3, and the definition

$$O_{k,n}(\boldsymbol{\xi}(e_0), W) := O\left(W\rho_\xi(\xi_k)(\xi_{k+1}-\xi_k), \ldots, W\rho_\xi(\xi_k)(\xi_{k+n}-\xi_{k+n-1})\right), \qquad \xi_i = \xi_i(e_0).$$

Here ρ_ξ denotes the limiting density of the eigenvalues Q_e. We also used $\rho_\xi(e_0) = \rho_{\mathrm{sc}}(e_0)$, and that ρ_ξ is smooth so $\rho_\xi(\xi_k)$ is very close to $\rho_\xi(e_0)$ by rigidity, both are easy consequences of the local law for Q_{e_0}, Theorem 6.1. We therefore now need to prove

$$\mathbb{E}^{Q_{e_0}} O_{k-\alpha,n}(\boldsymbol{\xi}(e_0), W) - \mathbb{E}^{\mathrm{GOE}_N} O_{k,n}(\boldsymbol{\lambda}, N) = \mathrm{O}(N^{-c}). \tag{4.34}$$

We now compute the left side of (4.34) by first conditioning on the law of A_2, B, D. Theorem 2.1 for Gaussian divisible matrices thus follows from (4.33) and the following lemma (proved in the next subsection), which asserts the local spectral statistics of the matrix Q_{e_0} are universal.

Lemma 4.6. *Under the assumptions of Theorem 2.1 and* (4.31)*, there exists $c > 0$ such that*

$$\mathbb{P}\left(\mathbb{1}_{\|(D-e_0)^{-1}\| \leqslant N^4} \left|\mathbb{E}^{A_1}\left(O_{k-\alpha,n}(\boldsymbol{\xi}(e_0), W)\middle| A_2, B, D\right) - \mathbb{E}^{\mathrm{GOE}_N} O_{k,n}(\boldsymbol{\lambda}, N)\right| \geqslant N^{-c}\right) \leqslant N^{-c}.$$

Theorem 2.1 for our band matrices with general entries follows from Lemma 4.6 and the following comparison result. Let $H_t = (H_{ij}(t))$ be a time dependent flow of symmetric $N \times N$ matrices with $H_0 = H$ our original band matrix. The dynamics of the matrix entries are given by the stochastic differential equations

$$\mathrm{d}H_{ij}(t) = \frac{\mathrm{d}\mathcal{B}_{ij}(t)}{\sqrt{N}} - \frac{1}{2Ns_{ij}} h_{ij}(t)\mathrm{d}t, \quad |i-j| \leqslant 2W, \tag{4.35}$$

where $\mathcal{B}$ is a symmetric matrix with $(\mathcal{B}_{ij})_{i \leqslant j}$ a family of independent Brownian motions. By definition, $H_{ij}(t) = 0$ for $|i-j| > 2W$. The parameter $s_{ij} > 0$ can take any positive values, but we choose s_{ij} to be the variance of $H_{ij}(0)$, i.e., $s_{ij} = 1/(4W-1)$. Clearly, for any $t \geqslant 0$ we have $\mathbb{E}(H_{ij}(t)^2) = s_{ij}$ for all i, j and thus the variance of the matrix element is preserved in this flow. This flow is similar to the Dyson Brownian motion but adapted to the band structure. For this flow, the following continuity estimate holds.

Lemma 4.7. *Let $\kappa > 0$ be arbitrarily small, $\delta \in (0, 1/2)$ and $t = N^{-1+\delta}$. Suppose that $W = cN$ for some constant c independent of N. Denote by H_t the solution of (4.35) with initial condition a symmetric band matrix H_0 as defined in (2.1), (2.2). Let m be any positive integer and $\Theta : \mathbb{R}^{m+m^2} \to \mathbb{R}$ be a smooth function with derivatives satisfying*

$$\sup_{k \in [\![0,5]\!], x \in \mathbb{R}^{m+m^2}} |\Theta^{(k)}(x)|(1+|x|)^{-C} < \infty \tag{4.36}$$

for some $C > 0$. Denote by $(\mathbf{u}_1(t), \ldots, \mathbf{u}_N(t))$ the eigenvectors of H_t associated with the eigenvalues $\lambda_1(t) \leqslant \ldots \leqslant \lambda_N(t)$, and $(u_k(t,\alpha))_{1 \leqslant \alpha \leqslant N}$ the coordinates of $\mathbf{u}_k(t)$. Then there exists $\varepsilon > 0$ (depending only on Θ, δ and κ) such that, for large enough N,

$$\sup_{I \subset [\![\kappa N, (1-\kappa)N]\!], |I| = m = |J|} \left| (\mathbb{E}^{H_t} - \mathbb{E}^{H_0}) \Theta\left((N(\lambda_k - \gamma_k), N u_k(\cdot, \alpha)^2)_{k \in I, \alpha \in J} \right) \right| \leqslant N^{-\varepsilon}.$$

The proof of this lemma is identical to that of the Corollary A.2 in [7] and we thus omit it. Instead of Lemma 4.7, the Green function comparison theorem from [20, 26] could be used as well to finish the proof.

We now complete the proof of Theorem 2.1. Recall that we have proved this theorem for Gaussian divisible ensembles of the form (4.23). At any time t, the entry $h_{ij}(t)$ of H_t for the flow (4.35) is distributed as

$$e^{-\frac{t}{2Ns_{ij}}} H_{ij}(0) + \left(s_{ij} \left(1 - e^{-\frac{t}{Ns_{ij}}} \right) \right)^{1/2} \mathscr{N}^{(ij)}, \quad |i-j| \leqslant 2W, \tag{4.37}$$

where $(\mathscr{N}^{(ij)})_{i \leqslant j}$ are independent standard Gaussian random variables. Hence H_t is Gaussian divisible and bulk universality holds for $t = N^{-1+\delta}$ with δ a small positive number. By Lemma 4.7, the bulk statistics of H_t and H_0 are the same up to negligible errors. We have thus proved Theorem 2.1.

4.4 Universality for mean-field perturbations. We now prove Lemma 4.6. We first recall a general theorem [27] concerning gap universality (see [17] for a related result). We start from the following definition. In the rest of the paper, we fix a small number $\mathfrak{a} > 0$, and define the control parameter

$$\varphi = W^{\mathfrak{a}}. \tag{4.38}$$

We will be interested in the deformed GOE defined by

$$\widetilde{H}_t = V + \sqrt{t} Z, \tag{4.39}$$

where V is a deterministic matrix and Z is a $W \times W$ GOE matrix. We now list the assumptions on the initial matrix V at some energy level E_0; in order to formulate them we will need two W-dependent mesoscopic scales $\eta_* \geqslant \varphi/W$ and $r \geqslant \varphi \eta_*$.

Assumption 1. *Let η_* and r be two W-dependent parameters, such that $\varphi/W \leqslant \eta_* \leqslant r/\varphi \leqslant 1$. We assume that there exist large positive constants C_1, C_2 such that*

(i) The norm of V is bounded, $\|V\| \leqslant W^{C_1}$.

(ii) The imaginary part of the Stieltjes transform of V is bounded from above and below, i.e.,

$$C_2^{-1} \leqslant \Im(m_V(z)) \leqslant C_2, \qquad m_V(z) := \frac{1}{W}\operatorname{Tr}(V-z)^{-1}, \tag{4.40}$$

uniformly for any $z \in \{E+\mathrm{i}\eta : E \in [E_0-r, E_0+r], \eta_* \leqslant \eta \leqslant 2\}$.

A deterministic matrix V satisfying these conditions will be called (η_*, r)-regular at E_0.

The following theorem was the main result of [27] (note that the size of the matrix W was replaced by N there).

Theorem 4.8 (Universality for mean-field perturbations [27])**.** *Suppose that V is (η_*, r)-regular at E_0 and set T such that $\eta_*\varphi \leqslant T \leqslant r^2/\varphi$ with $\varphi = W^{\mathfrak{a}}$. Let j be an index so that the j-th eigenvalue of V, $V_j \in [E_0 - r/3, E_0 + r/3]$. Denote the eigenvalues of $\widetilde{H}_T$ (defined in (4.39)) by $\boldsymbol{\lambda}_T = \{\lambda_{T,i}\}_{i=1}^W$ and let*

$$m_{\widetilde{H}_T}(z) = \frac{1}{W}\operatorname{Tr}(\widetilde{H}_T - z)^{-1}. \tag{4.41}$$

Recall the definition of the gap observable $O_{j,n}$ from (4.32) for some fixed n. For $\mathfrak{a}$ small enough, there is a constant $c > 0$ (depending on $C_1, C_2, \mathfrak{a}$) such that

$$\mathbb{E}^{\widetilde{H}_T} O_{j,n}\left(\boldsymbol{\lambda}_T, W\frac{\rho_T(\lambda_{T,j})}{\rho_{\mathrm{sc}}(\lambda_{T,j})}\right) - \mathbb{E}^{\mathrm{GOE}_W} O_{j,n}(\boldsymbol{\lambda}, W) = \mathrm{O}(W^{-c}), \tag{4.42}$$

where

$$\rho_T(\lambda_{T,j}) = \operatorname{Im} m_{\widetilde{H}_T}(\lambda_{T,j} + \mathrm{i}\eta), \quad \eta = T/\varphi.$$

Furthermore, for any $\delta > 0$ the following level repulsion estimate holds:

$$\mathbb{P}\left(|\lambda_{T,i} - \lambda_{T,i+1}| \leqslant x/W\right) \leqslant C_\delta W^\delta x^{2-\delta} \tag{4.43}$$

for any $x > 0$ (which can depend on W) and for all i such that $\lambda_{T,i} \in [E_0 - r/3, E_0 + r/3]$.

The compensating factor $\frac{\rho_T(\lambda_{T,j})}{\rho_{\mathrm{sc}}(\lambda_{T,j})}$ is due to our definition of the observable (4.32) with a scaling ρ_{sc}.

Proof of Lemma 4.6. We apply Theorem 4.8 to the matrix

$$\widetilde{H}_T = Q_{e_0} = \sqrt{q}A_1 + V \text{ where } V = \sqrt{1-q}A_2 - B^*(D-e_0)^{-1}B, \tag{4.44}$$

with the following choices:

$$T = q = N^{-1+\theta},\ \eta_* = N^{-1+\theta/2},\ r = N^{-1/2+\theta},\ E_0 = e_0,\ j = k - \alpha \quad (\alpha = \mathcal{N}_D(e_0)),\ \lambda_{T,k} = \xi_k(e_0),\ C_1 = 5, \tag{4.45}$$

and C_2 some large constant (in the regularity assumptions on V). Remember that $\xi_k(e_0)$ is the eigenvalue of Q_{e_0} and $\mathcal{N}_D(e_0)$ was defined below (4.17).

In order to verify the regularity assumption of Theorem 4.8, we need a local law for Q_{e_0}, which is stated and proved in Theorem 6.1: from (6.3), there exists some $c > 0$ such that for any $D > 0$ we have, for large enough N,

$$\mathbb{P}\left(\forall z = E + \mathrm{i}\eta :\ |E - e_0| \leqslant r;\ \eta_* \leqslant \eta \leqslant c, \quad \frac{1}{W}\Im\operatorname{Tr}(V-z)^{-1} \in [c, c^{-1}]\right) \geqslant 1 - N^{-D}.$$

This verifies that part *(ii)* of the assumption of Theorem 4.8.

Moreover, since the statement of Lemma 4.6 concerns only the set $\|(D-e_0)^{-1}\| \leqslant N^4$, together with the fact that A_2 and B are bounded with high probability, we have in this set

$$\|\sqrt{1-q}A_2 - B^*(D-e_0)^{-1}B\| \leqslant N^5$$

with high probability. This verifies that part (i) of the assumption of Theorem 4.8 with $C_1 = 5$.

Recall the mean field reduction from Section 4.2. By (4.17) and $\mathcal{C}_k(\lambda_k) = \lambda_k$, we have

$$|\xi_j(e_0) - \gamma_k| = |\xi_{k-\alpha}(e_0) - \gamma_k| = |\mathcal{C}_k(e_0) - \gamma_k| \leqslant |\mathcal{C}_k(e_0) - \mathcal{C}_k(\lambda_k)| + |\lambda_k - \gamma_k| \leqslant N^{-1+\omega} \tag{4.46}$$

with probability larger than $1 - N^{-D}$ for any small $\omega > 0$ and large $D > 0$. Here we have used the rigidity of λ_k, the assumption $|e_0 - \gamma_k| \leqslant (\log N)N^{-1}$ and the estimate (4.19) on $(\mathrm{d}/\mathrm{d}e)\mathcal{C}_k(e)$.

Since $\xi_j = \xi_j(e_0)$ is the j-th eigenvalue of $\sqrt{q}A_1 + V$ and let V_j be j-th eigenvalue of V, we have $\xi_j(e_0) - V_j = \mathrm{O}(\sqrt{q}N^{\omega})$ with probability larger than $1 - N^{-D}$. Therefore with high probability $V_j \in [e_0 - r/3, e_0 + r/3]$. Hence we can apply Theorem 4.8 to get

$$\mathbb{P}\left(\left|\mathbb{E}^{A_1}\left(O_{k-\alpha,n}\left(\boldsymbol{\xi}(e_0), W\frac{\rho_T(\xi_{k-\alpha})}{\rho_{\mathrm{sc}}(\xi_{k-\alpha})}\right)\Big|A_2, B, D\right) - \mathbb{E}^{\mathrm{GOE}_W} O_{k-\alpha,n}(\boldsymbol{\lambda}, W)\right| \geqslant N^{-c}\right) \leqslant N^{-c} \tag{4.47}$$

for some $c > 0$. By (4.46) and smoothness of ρ_{sc}, we can replace $\rho_{\mathrm{sc}}(\xi_{k-\alpha})$ with $\rho_{\mathrm{sc}}(\gamma_k)$ up to negligible error. Furthermore, by the local law (6.2) we have for some $c > 0$ that

$$\mathbb{P}\left(\forall z = E + \mathrm{i}\eta :\ |E - e_0| \leqslant N^{-1/2}, \quad \eta = T/\varphi, \quad \left|\frac{1}{W}\Im\,\mathrm{Tr}(Q_{e_0} - z)^{-1} - \rho_{\mathrm{sc}}(e_0)\right| \leqslant N^{-c}\right) \geqslant 1 - N^{-D}.$$

Therefore, we can replace $\rho_T(\xi_{k-\alpha})$ by $\rho_{\mathrm{sc}}(e_0)$, again up to negligible error. With this replacement, (4.47) is exactly the statement of Lemma 4.6, after noticing that $\mathbb{E}^{\mathrm{GOE}_W} O_{k-\alpha,n}(\boldsymbol{\lambda}, W)$ converges, as $W \to \infty$, to a limit independent of the bulk index $k - \alpha$. □

5 Quantum unique ergodicity

In this section, we prove Theorem 2.2, in particular we check that the assumption of Theorem 4.5 concerning the flatness of eigenvector holds. The following lemma implies the assumption (4.24) by choosing $\mathbf{a}(i) = 1$ for all $1 \leqslant i \leqslant W$, $\mathbf{g} = (g_1, \ldots, g_N)$ with $g_i = g\mathbb{1}_{i>W}$ and noticing $0 \leqslant \sum_{i=1}^{W} \left|\psi_j^g(i)\right|^2 \leqslant 1$. We will prove this lemma after completing the proof of Theorem 2.2.

Lemma 5.1 (Quantum unique ergodicity for Gaussian divisible band matrices)**.** *Recall that* $\boldsymbol{\psi}_k^{\mathbf{g}} = \begin{pmatrix} \mathbf{w}_k^{\mathbf{g}} \\ \mathbf{p}_k^{\mathbf{g}} \end{pmatrix}$ *is the k-th eigenvector of $H^{\mathbf{g}}$ with eigenvalue $\lambda_k^{\mathbf{g}}$. Suppose that* (4.23) *holds. Let $\kappa > 0$ be fixed. There exists $\varepsilon, \zeta > 0$ such that for any $k \in [\![\kappa N, (1-\kappa)N]\!]$, $\mathbf{a} \in [-1,1]^W$ and $\delta > 0$ we have*

$$\sup_{\|\mathbf{g}\|_\infty \leqslant N^{-1+\zeta}} \mathbb{P}\left(\left|\sum_{i=1}^{W} \mathbf{a}(i)\left(|w_k^{\mathbf{g}}(i)|^2 - \frac{1}{N}\right)\right| \geqslant \delta\right) \leqslant C_\kappa N^{-\varepsilon}/\delta^2. \tag{5.1}$$

Proof of Theorem 2.2. We will first prove Theorem 2.2 for the class of Gaussian divisible band matrix ensemble, which was defined in (4.23). With (5.1), we know that there exists $\zeta, \varepsilon > 0$ such that for any $k \in [\![\kappa N, (1-\kappa)N]\!]$, $\mathbf{a} \in [-1,1]^N$, $m \in [\![0, N/W - 1]\!]$, $\|\mathbf{g}\|_\infty < N^{-1+\zeta}$, and $\delta > 0$, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{W} \mathbf{a}(i + mW)\left(|\psi_k^{\mathbf{g}}(i + mW)|^2 - \frac{1}{N}\right)\right| \geqslant \delta\right) \leqslant C_\kappa N^{-\varepsilon}/\delta^2.$$

Then summing up $m \in [\![0, N/W - 1]\!] = 0, 1, \ldots 2p - 1$, we have proved Theorem 2.2 in the case of Gaussian divisible band matrix. For the general case, we consider $\mathbf{g} = 0$ for simplicity, without loss of generality. Recall the definition of H_t in (4.35). With (2.7) for any Gaussian divisible band matrix, we know that for some $\varepsilon > 0$,

$$\mathbb{E}^{H_t}\left|\sum_{i=1}^{N} \mathbf{a}(i)\left(|\psi_k(i)|^2 - \frac{1}{N}\right)\right|^2 \leqslant C_{\kappa,p} N^{-\varepsilon}. \tag{5.2}$$

Then comparing $H = H_0$ with H_t using Lemma 4.7, we have

$$\left|\left(\mathbb{E}^{H_t} - \mathbb{E}^{H_0}\right)|\psi_k(i)|^2\right| \leqslant C_\kappa N^{-1-\widetilde{\varepsilon}},\ \left|\left(\mathbb{E}^{H_t} - \mathbb{E}^{H_0}\right)|\psi_k(i)|^2|\psi_k(j)|^2\right| \leqslant C_\kappa N^{-2-\widetilde{\varepsilon}},$$

for some $\widetilde{\varepsilon} > 0$ and for any i, j. Together with (5.2), we therefore proved

$$\mathbb{E}^{H_0}\left|\sum_{i=1}^{N}\mathbf{a}(i)\left(|\psi_k^{\mathbf{g}}(i)|^2 - \frac{1}{N}\right)\right|^2 \leqslant C_{\kappa,p}(N^{-\varepsilon} + N^{-\widetilde{\varepsilon}}),$$

which implies the desired result (2.7) by Markov's inequality. □

We now prove Lemma 5.1. Recall the notations in (2.8)-(2.11), i.e. that $\mathbf{u}_j^{\mathbf{g}}(e)$, ($e \in \mathbb{R}$ and $j \in [\![1, W]\!]$) is a (real) eigenvector of the matrix

$$Q_e^{\mathbf{g}} = A^{\mathbf{g}} - B^*(D^{\mathbf{g}} - e)^{-1}B = \sqrt{q}A_1 + V^{\mathbf{g}}, \quad V^{\mathbf{g}} = \sqrt{1-q}A_2 + \sum_{1\leqslant i\leqslant W} g_i \mathbf{e}_i\mathbf{e}_i^* - B^*(D^{\mathbf{g}} - e)^{-1}B. \quad (5.3)$$

Note that not only A has a Gaussian divisible decomposition (4.31) but also B and D, however this latter fact is irrelevant and we will not follow it in the notation. With the labeling of eigenvalue convention in (4.17), we have the following relation between $\mathbf{u}^{\mathbf{g}}$ and $w^{\mathbf{g}}$.

$$\mathbf{u}_{\hat{k}}^{\mathbf{g}}(\lambda_k^{\mathbf{g}}) = \frac{w_k^{\mathbf{g}}}{\|w_k^{\mathbf{g}}\|}, \quad \hat{k} := k'(\lambda_k^{\mathbf{g}}) = k - \mathcal{N}_D(\lambda_k^{\mathbf{g}}). \quad (5.4)$$

To prove Lemma 5.1, we first claim that the following QUE for $\mathbf{u}_{\hat{k}}^{\mathbf{g}}(\lambda_k^{\mathbf{g}})$ holds. The challenge is that we consider the matrix $Q_e^{\mathbf{g}}$ with a random shift e, namely $e = \lambda_k^{\mathbf{g}}$, and the index $\hat{k}$ is also random.

Lemma 5.2 (Quantum unique ergodicity for mean-field matrices with random shift e)**.** *Let $\kappa > 0$ be fixed. Under the assumption of Lemma 5.1 and (4.31), there exists $\varepsilon, \zeta > 0$ such that for any $k \in [\![\kappa N, (1-\kappa)N]\!]$, $\mathbf{a} \in [-1, 1]^W$ and $\delta > 0$ we have ($[\mathbf{x}]_i$ denotes the i-th component of a vector $\mathbf{x}$)*

$$\sup_{\|\mathbf{g}\|_\infty \leqslant N^{-1+\zeta}} \mathbb{P}\left(\left|\sum_{i=1}^{W}\mathbf{a}(i)\left(\left[\mathbf{u}_{\hat{k}}^{\mathbf{g}}(\lambda_k^{\mathbf{g}})\right]_i^2 - \frac{1}{W}\right)\right| \geqslant \delta\right) \leqslant C_\kappa N^{-\varepsilon}/\delta^2, \qquad \hat{k} := k'(\lambda_k^{\mathbf{g}}) = k - \mathcal{N}_D(\lambda_k^{\mathbf{g}}). \quad (5.5)$$

Proof of Lemma 5.1. Clearly, to deduce (5.1) from (5.5), one only needs to show that there exists $\widetilde{\varepsilon} > 0$ such that

$$\sup_{\|\mathbf{g}\|_\infty \leqslant N^{-1+\zeta}} \mathbb{P}\left(\left|\sum_{1\leqslant i\leqslant W}\psi_k^{\mathbf{g}}(i)^2 - \frac{W}{N}\right| \geqslant N^{-\widetilde{\varepsilon}}\right) \leqslant C_\kappa N^{-\widetilde{\varepsilon}}. \quad (5.6)$$

To see this, we first note that by choosing $\mathbf{a}(i) = \mathbb{1}_{i\leqslant W/2} - \mathbb{1}_{i>W/2}$, and $\delta = N^{-\varepsilon/10}$ in (5.5), we have

$$\mathbb{P}\left(\left|\sum_{1\leqslant i\leqslant W/2}\psi_k^{\mathbf{g}}(i)^2 - \sum_{W/2<i\leqslant W}\psi_k^{\mathbf{g}}(i)^2\right| \geqslant N^{-\varepsilon/10}\right) \leqslant C_\kappa N^{-\varepsilon/10}.$$

In the above equation, the index set $[\![1, W]\!]$ which determines the decomposition (2.8) can be replaced by $[\![1+nW/2, W+nW/2]\!]$ with $n \in [\![0, 2(N/W-1)]\!]$. By a simple union bound, we can assume all these bounds hold simultaneously. In particular, the local ℓ^2-norms of $\psi_k^{\mathbf{g}}$ on each consecutive $W/2$ batches of indices coincide approximately. As $\boldsymbol{\psi}_k^{\mathbf{g}}$ is normalized, all these local norms are close to $W/(2N)$, which implies (5.6) and completes the proof. □

In Lemma 5.2, the energy $\lambda_k^{\mathbf{g}}$ is random and the index includes a random shift $\mathcal{N}_{D^{\mathbf{g}}}(\lambda_k^{\mathbf{g}})$. To prove Lemma 5.2, we need the following lemma (proved at the end of this section) which replaces the random parameter $\lambda_k^{\mathbf{g}}$ in (5.5) by a deterministic one.

Lemma 5.3 (Quantum unique ergodicity for mean-field matrices with fixed shift e)**.** *Let $\kappa > 0$ be fixed. Under the assumption of Lemma 5.1 and (4.31), there exists $\varepsilon, \zeta > 0$ such that for any $k \in [\![\kappa N, (1-\kappa)N]\!]$, $|e - \gamma_k| \leqslant N^{-1+2\zeta}$, $\|\mathbf{g}\|_\infty \leqslant N^{-1+\zeta}$, $\mathbf{a} \in [-1,1]^W$, and $\delta > 0$, we have*

$$\mathbb{P}\left(\exists j : |j - k'| \leqslant N^\zeta,\ \left|\sum_{i=1}^W \mathbf{a}(i)\left(\left[\mathbf{u}_j^{\mathbf{g}}(e)\right]_i^2 - \frac{1}{W}\right)\right| \geqslant \delta\right) \leqslant C_\kappa N^{-\varepsilon}/\delta^2$$

where $k' = k'(e) = k - \mathcal{N}_{D^{\mathbf{g}}}(e)$.

Proof of Lemma 5.2. Since $|\lambda_k^{\mathbf{g}} - \lambda_k| \leqslant \|\mathbf{g}\|_\infty$ and the rigidity estimate holds for λ_k (see (4.7)), with high probability we have

$$\lambda_k^{\mathbf{g}} - \gamma_k = \mathrm{O}(N^{-1+\zeta}) \tag{5.7}$$

for any $\|\mathbf{g}\|_\infty \leqslant N^{-1+\zeta}$.

We discretize the set of the parameter e. Denote $e_m = \gamma_k + mN^{-1-\zeta}$. For small enough $\varepsilon, \zeta > 0$, for any fixed $\mathbf{a}$ and $\|\mathbf{g}\|_\infty$ as in the assumptions of Lemma 5.3, we thus have, from this lemma (and a union bound),

$$\mathbb{P}\left(\exists m \in \mathbb{Z}, \exists j :\ |m| \leqslant N^{3\zeta},\ |j - k'(e_m)| \leqslant N^\zeta,\ \left|\sum_{i=1}^W \mathbf{a}(i)\left(\left[\mathbf{u}_j^{\mathbf{g}}(e_m)\right]_i^2 - \frac{1}{W}\right)\right| \geqslant \delta\right) \leqslant C_\kappa N^{-\varepsilon}/\delta^2. \tag{5.8}$$

Using (5.7), we have with high probability that there exists a random integer $|\widetilde{m}| \leqslant N^{3\zeta}$ such that

$$|e_{\widetilde{m}} - \lambda_k^{\mathbf{g}}| \leqslant N^{-1-\zeta}. \tag{5.9}$$

Defining the $W \times W$ matrix J by $J_{ij} := \mathbf{a}(i)\delta_{ij}$ and setting $e^* := \lambda_k^{\mathbf{g}}$, we have

$$\sum_i \mathbf{a}(i)\left[\mathbf{u}_{\hat{k}}^{\mathbf{g}}(\lambda_k^{\mathbf{g}})\right]_i^2 = \left(\mathbf{u}_{\hat{k}}^{\mathbf{g}}(e^*), J\mathbf{u}_{\hat{k}}^{\mathbf{g}}(e^*)\right) = \left(\mathbf{u}_{k'(e_{\widetilde{m}})}^{\mathbf{g}}(e_{\widetilde{m}}), J\mathbf{u}_{k'(e_{\widetilde{m}})}^{\mathbf{g}}(e_{\widetilde{m}})\right) + \int_{e_{\widetilde{m}}}^{e^*} \frac{\mathrm{d}}{\mathrm{d}e}\left(\mathbf{u}_{k'(e)}^{\mathbf{g}}(e), J\mathbf{u}_{k'(e)}^{\mathbf{g}}(e)\right)\mathrm{d}e. \tag{5.10}$$

From (5.8) and (5.9),

$$\mathbb{P}\left(\left|\sum_{i=1}^W \mathbf{a}(i)\left(\left[\mathbf{u}_{k'(e_{\widetilde{m}})}^{\mathbf{g}}(e_{\widetilde{m}})\right]_i^2 - \frac{1}{W}\right)\right| \geqslant \delta\right) \leqslant \widetilde{C}_\kappa N^{-\varepsilon}/\delta^2.$$

We therefore just need to bound the second term on the right hand side of (5.10). A simple calculation yields (we now abbreviate $k' = k'(e)$ and similarly $\ell' = \ell'(e) = \ell - \mathcal{N}_D(e)$)

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

Together with $\|\mathbf{a}\|_\infty \leqslant 1$, this gives

$$\left|\frac{\mathrm{d}}{\mathrm{d}e}\left(\mathbf{u}_{k'}^{\mathbf{g}}(e), J\,\mathbf{u}_{k'}^{\mathbf{g}}(e)\right)\right| \leqslant \sum_{\ell \neq k} \frac{C}{|\mathcal{C}_k^{\mathbf{g}}(e) - \mathcal{C}_\ell^{\mathbf{g}}(e)|}\left\|\frac{1}{D^{\mathbf{g}} - e} B\mathbf{u}_{\ell'}^{\mathbf{g}}(e)\right\|\left\|\frac{1}{D^{\mathbf{g}} - e} B\mathbf{u}_{k'}^{\mathbf{g}}(e)\right\|.$$

By (4.22), for all $e \notin \sigma(D^{\mathbf{g}})$, we can bound $\left\|\frac{1}{D^{\mathbf{g}} - e} B\mathbf{u}_{\ell'}^{\mathbf{g}}(e)\right\|$ by $C(1 + |\mathcal{C}_{\ell'}^{\mathbf{g}}(e)|)$ with high probability. Since for $e \in [e_1, e_{N^{3\zeta}}]$, $\mathcal{C}_k^{\mathbf{g}}(e) = \mathrm{O}(1)$ with high probability, we have

$$\left|\frac{\mathrm{d}}{\mathrm{d}e}\left(\mathbf{u}_{k'}^{\mathbf{g}}(e), J\,\mathbf{u}_{k'}^{\mathbf{g}}(e)\right)\right| \leqslant C\sum_{\ell \neq k} \frac{C(1 + |\mathcal{C}_\ell^{\mathbf{g}}(e)|)}{|\mathcal{C}_k^{\mathbf{g}}(e) - \mathcal{C}_\ell^{\mathbf{g}}(e)|}, \qquad e \in [e_1, e_{N^{3\zeta}}] \setminus \sigma(D^{\mathbf{g}}),$$

with high probability. Using (4.19) and (6.4) in Theorem 6.1 with $t = q$ (note that, with the notations of Theorem 6.1, we have $\mathcal{C}_k^{\mathbf{g}}(e) = \xi_{k'}^{\mathbf{g}}(e, q, q)$, $k' = k - \mathcal{N}_{D^{\mathbf{g}}}(e)$ and $\xi_k^{\mathbf{g}}(e, q, t)$ is the k-th eigenvalue of $Q_e^{\mathbf{g}}(t, q)$, which is defined in (6.1)), we have

$$\sum_{\ell:|\ell-k|\geqslant 2N^{2\omega}} \frac{C(1+|\mathcal{C}_\ell^{\mathbf{g}}(e)|)}{|\mathcal{C}_k^{\mathbf{g}}(e) - \mathcal{C}_\ell^{\mathbf{g}}(e)|} \leqslant CN^{1+3\omega}, \qquad e \in [e_1, e_{N^{3\zeta}}] \setminus \sigma(D^{\mathbf{g}})$$

with high probability for any small ω. We have thus proved that with high probability

$$\left|\frac{\mathrm{d}}{\mathrm{d}e}\Big(\mathbf{u}_{k'}^{\mathbf{g}}(e), J\,\mathbf{u}_{k'}^{\mathbf{g}}(e)\Big)\right| \leqslant \sum_{\ell:|\ell-k|\leqslant 2N^{2\omega}} \frac{1}{|\mathcal{C}_k^{\mathbf{g}}(e) - \mathcal{C}_\ell^{\mathbf{g}}(e)|} + CN^{1+3\omega}$$

for any $e \in [e_1, e_{N^{3\zeta}}] \setminus \sigma(D^{\mathbf{g}})$. Inserting the last equation into (5.10), using (5.9), the ordering of the curves $\mathcal{C}_k^{\mathbf{g}}$ in k, and choosing $\omega \leqslant \zeta/10$, we obtain that with high probability,

$$|(\mathbf{u}_{k'}^{\mathbf{g}}(\lambda_m^{\mathbf{g}}), J\mathbf{u}_{k'}^{\mathbf{g}}(\lambda_m^{\mathbf{g}})) - (\mathbf{u}_{k'}^{\mathbf{g}}(e_{\widetilde{m}}), J\mathbf{u}_{k'}^{\mathbf{g}}(e_{\widetilde{m}}))| \leqslant CN^{2\omega} \int_{e_{\widetilde{m}}}^{e^*} \sum_{\ell=k\pm 1} \frac{1}{|\mathcal{C}_k^{\mathbf{g}}(e) - \mathcal{C}_\ell^{\mathbf{g}}(e)|}\mathrm{d}e + N^{-\zeta/2}. \tag{5.11}$$

By Hölder's inequality, we have

$$\begin{aligned}\mathbb{E}\left|\int_{e_{\widetilde{m}}}^{e^*} \frac{1}{|\mathcal{C}_k^{\mathbf{g}}(e) - \mathcal{C}_\ell^{\mathbf{g}}(e)|}\mathrm{d}e\right| &\leqslant \left(\mathbb{E}\left|\int_{e_{\widetilde{m}}}^{e^*} \mathrm{d}e\right|\right)^{1/3} \left(\mathbb{E}\int_{e_{\widetilde{m}}}^{e^*} \frac{1}{|\mathcal{C}_k^{\mathbf{g}}(e) - \mathcal{C}_\ell^{\mathbf{g}}(e)|^{3/2}}\,\mathrm{d}e\right)^{2/3} \\ &\leqslant CN^{-\zeta-1/3}\left(N^{-1+\zeta} \max_{e:|e-\gamma_k|\leqslant N^{-1+2\zeta}} \mathbb{E}\,|\mathcal{C}_k^{\mathbf{g}}(e) - \mathcal{C}_\ell^{\mathbf{g}}(e)|^{-3/2}\right)^{2/3}.\end{aligned} \tag{5.12}$$

As in the proof of Lemma 4.6, we apply Theorem 4.8 to the operator $Q_e^{\mathbf{g}}$ in (5.3). We can similarly verify that Assumption 1 holds with high probability and thus the level repulsion estimate (4.43) holds. Since $|\mathbf{g}| \ll N^{-1/2} \ll r$ (r is chosen as in (4.45)), we have $V_k^{\mathbf{g}} \in [e - r/3, e + r/3]$ for any k such that $|e - \gamma_k| \leqslant N^{-1+2\zeta}$. Thus for any small ω we have

$$\max_{e:|e-\gamma_k|\leqslant N^{-1+2\zeta}} \mathbb{E}\,|\mathcal{C}_k^{\mathbf{g}}(e) - \mathcal{C}_\ell^{\mathbf{g}}(e)|^{-3/2} \leqslant C_\omega N^{3/2+\omega}, \qquad \ell = k \pm 1.$$

Together with the Markov inequality, (5.12) and (5.11), this concludes the proof of Lemma 5.2. □

Proof of Lemma 5.3. We will need the local QUE from [6]. Remember the notations from Subsection 4.4 and the control parameter $\varphi = W^{\mathfrak{a}} = cN^{\mathfrak{a}}$. Let $\mathbf{u}_1(t), \ldots, \mathbf{u}_W(t)$ be the real eigenvectors for the matrix $\widetilde{H}_t$ defined in (4.39) and let $u_j(i,t)$ be the i-th component of $\mathbf{u}_j(t)$. The following result is the content of Corollary 1.3 in [6].

Theorem 5.4 (Quantum unique ergodicity for mean-field perturbations [6])**.** *We assume the initial matrix $\widetilde{H}_0 = V$ satisfies Assumption 1 in Subsection 4.4. We further assume that there exists a small constant $\mathfrak{b}$ such that*

$$\left|(\widetilde{H}_0 - z)_{ij}^{-1} - m_0(z)\delta_{ij}\right| \leqslant \frac{1}{W^{\mathfrak{b}}}, \text{ with } m_0(z) = \frac{1}{N}\operatorname{Tr}(\widetilde{H}_0 - z)^{-1}, \tag{5.13}$$

uniformly in $\{z = E + \mathrm{i}\eta : E \in [E_0 - r, E_0 + r], \eta_* \leqslant \eta \leqslant r\}$ with E_0, η_* and r as in Assumption 1. Then the following quantum unique ergodicity holds: for any $\mu > 0$ there exists $\varepsilon, C_\mu > 0$ (depending also on $\mathfrak{a}, \mathfrak{b}$ and C_2 from Assumption 1) such that for any T with $\varphi\eta_* \leqslant T \leqslant r/\varphi$, $\mathbf{a} \in [-1, 1]^W$, and $\delta > 0$, we have

$$\sup_{j:|\lambda_{T,j}-E_0|<(1-\mu)r} \mathbb{P}\left(\left|\frac{1}{\|\mathbf{a}\|_1}\sum_{i=1}^{W} \mathbf{a}(i)(Wu_j^2(i,T) - 1)\right| > \delta\right) \leqslant C_\mu\ \left(W^{-\varepsilon} + \|\mathbf{a}\|_1^{-1}\right)/\delta^2. \tag{5.14}$$

We now return to the proof of Lemma 5.3. Theorem 5.4 implies in particular that

$$\sup_{j:|\lambda_{T,j}-E_0|<(1-\mu)r} \mathbb{P}\left(\left|\sum_{i=1}^{N} \mathbf{a}(i)\left(u_j^2(i,T)-N^{-1}\right)\right|>\delta\right) \leqslant C\ N^{-\varepsilon}/\delta^2. \tag{5.15}$$

Similarly to (4.44), we apply Theorem 5.4 to the matrix

$$\widetilde{H}_T = Q_e^{\mathbf{g}} = \sqrt{q}A_1 + V \text{ where } V = \sqrt{1-q}A_2 + \sum_{1\leqslant i\leqslant W} g_i \mathbf{e}_i\mathbf{e}_i^* - B^*(D^{\mathbf{g}}-e)^{-1}B,$$

with the following choices:

$$T = q = N^{-1+\theta},\ \eta_* = N^{-1+\theta/2},\ r = N^{-1/2+\theta},\ E_0 = e,$$

In particular, the supremum in (5.14) will cover all indices j such that $|j-(k-\mathcal{N}_D^{\mathbf{g}}(e))| \leqslant N^{\zeta}$ and recall that $u_j(i,T) = u_j^{\mathbf{g}}(i)$ for such j. Using the results from the next section, both requirements of Assumption 1 hold for our V, in particular (4.40) is satisfied by (6.3) for $q = t = 0$. Moreover (5.13) holds by (6.2). Hence the assumption for Theorem 5.4 are verified. Therefore with (5.15) we obtain that there exists $\varepsilon > 0$ such that for any δ,

$$\sup_{\ell:|\mathcal{C}_\ell^{\mathbf{g}}(e)-e|<(1-\mu)N^{-1/2+\theta}} \mathbb{P}\left(\left|\sum_{i=1}^{W} \mathbf{a}(i)\left(\left[\mathbf{u}_{\ell'}^{\mathbf{g}}(e)\right]_i^2 - N^{-1}\right)\right|>\delta\right) \leqslant CN^{\varepsilon}/\delta^2, \quad \ell' = \ell - \mathcal{N}_{D^{\mathbf{g}}}(e). \tag{5.16}$$

Moreover for any index k satisfying $|e-\gamma_k| \leqslant N^{-1+2\zeta}$ we have $|\mathcal{C}_k^{\mathbf{g}}(e)-e| < (1-\mu)N^{-1/2+\theta}$. Indeed, with the rigidity property (4.7) and the trivial perturbation estimate $|\lambda_k^{\mathbf{g}}-\lambda_k| \leqslant \|\mathbf{g}\|_\infty$, we know that

$$|\lambda_k^{\mathbf{g}} - e| \leqslant |\lambda_k^{\mathbf{g}}-\lambda_k| + |\lambda_k-\gamma_k| + |\gamma_k - e| \leqslant CN^{-1+\zeta}.$$

By definition, $\mathcal{C}_k^{\mathbf{g}}(\lambda_k^{\mathbf{g}}) = \lambda_k^{\mathbf{g}}$. Hence together with (4.19), we have $|\mathcal{C}_k^{\mathbf{g}}(e)-e| \leqslant CN^{-1+\zeta}$ with high probability.

Finally, after choosing such a k satisfying $|e-\gamma_k| \leqslant N^{-1+2\zeta}$, for any j such that $|j-k'(e)| \leqslant N^{\zeta}$ we have $j = \ell'(e)$ for some ℓ. Moreover, $|\mathcal{C}_\ell^{\mathbf{g}}(e)-e| \leqslant |\mathcal{C}_\ell^{\mathbf{g}}(e) - C_k^{\mathbf{g}}(e)| + CN^{-1+\zeta} \leqslant CN^{-1+\zeta}$, so that we can apply (5.16). This concludes the proof of Lemma 5.3 by a simple union bound over all j's such that $|j-k'(e)| \leqslant N^{\zeta}$. □

6 Local law

The main purpose of this section is to prove the local law of the Green's function of $H^{\mathbf{g}}$, $Q_e^{\mathbf{g}}$ and some variations of them (recall the notations from Section 2.2). As we have seen in the previous sections, these local laws are the basic inputs for proving universality and QUE of these matrices.

Theorem 6.1 (Local law for Q)**.** *Recall $\mathbf{S}(e,N;\omega)$, $\widehat{\mathbf{S}}(e,N;\omega)$ and $m(z)$ defined in (4.1)-(4.3). We fix a vector $\mathbf{g}\in\mathbb{R}^N$ with $\|\mathbf{g}\| \leqslant N^{-1/2}$, numbers $0 \leqslant t \leqslant q \leqslant N^{-1/2}$, a positive N-independent threshold $\kappa > 0$ and any energy e with $|e| \leqslant 2-\kappa$. Set*

$$Q_e^{\mathbf{g}}(t,q) := \sqrt{t}A_1 + \sqrt{1-q}A_2 - \sum_{1\leqslant i\leqslant W} g_i\mathbf{e}_i\mathbf{e}_i^* - B^*\frac{1}{D^{\mathbf{g}}-e}B. \tag{6.1}$$

For any (small) $\omega > 0$ and (small) $\zeta > 0$ and (large) D, we have

$$\mathbb{P}\left(\exists z\in\mathbf{S}(e,N;\omega)\ s.t.\ \max_{ij}\left|[Q_e^{\mathbf{g}}(t,q)-z]_{ij}^{-1} - m(z)\delta_{ij}\right| \geqslant N^{\zeta}\left((N\eta)^{-1/2}+|z-e|\right)\right) \leqslant N^{-D}, \tag{6.2}$$

and there exists $c > 0$ such that

$$\mathbb{P}\left(\exists z \in \widehat{\mathbf{S}}(e, N; \omega) \text{ s.t. } \frac{1}{W} \operatorname{Im} \sum_i \left[Q_e^{\mathbf{g}}(t,q) - z\right]_{ii}^{-1} \notin [c, c^{-1}]\right) \leqslant N^{-D}. \tag{6.3}$$

Notice that (6.3) *holds in* $\widehat{\mathbf{S}}(e, N; \omega)$*, which is larger than the set* $\mathbf{S}(e, N; \omega)$ *used in* (6.2)*. But instead of a precise error estimate as in* (6.2)*, here* (6.3) *only provides a rough bound.*

Let $\xi_k^{\mathbf{g}}(e,t,q)$ *be the* k*-th eigenvalue of* $Q_e^{\mathbf{g}}(t,q)$*. Then for any (small)* $\omega > 0$ *and (large)* D

$$\mathbb{P}\left(\exists k, \ell : \xi_k^{\mathbf{g}}(e,t,q), \xi_\ell^{\mathbf{g}}(e,t,q) \in [e - N^{-\omega}, e + N^{-\omega}],\ |\xi_k^{\mathbf{g}}(e,t,q) - \xi_\ell^{\mathbf{g}}(e,t,q)| \leqslant \frac{|\ell - k|}{N^{1+\omega}} - N^{-1+\omega}\right) \leqslant N^{-D} \tag{6.4}$$

Notice the minus sign in front of $-N^{-1+\omega}$ *so that the right hand side of the last inequality is positive only when* $|k - \ell| \geqslant N^{2\omega}$*.*

6.1 Local law for generalized Green's function. To prove Theorem 6.1, we start with a more general setting. Let $\widetilde{H}$ be an $N \times N$ real symmetric random matrix with centered and independent entries, up to symmetry. (Here we use a different notation since it is different from the H of main part. Moreover, this $\widetilde{H}$ is also different from the matrix defined in (4.39).) Define

$$\widetilde{s}_{ij} := \mathbb{E}\widetilde{H}_{ij}^2, \quad 1 \leqslant i, j \leqslant N.$$

Assume that $\widetilde{s}_{ij} = \mathrm{O}(N^{-1})$ and there exist s_{ij} such that for some $c > 0$,

$$\widetilde{s}_{ij} = (1 + \mathrm{O}(N^{-1/2-c}))s_{ij}, \tag{6.5}$$

and

$$s_{ij} = s_{ji}, \quad \sum_i s_{ij} = 1.$$

Note that the row sums of the matrix of variances of $\widetilde{H}$ is not exactly 1 any more, so this class of matrices $\widetilde{H}$ goes slightly beyond the concept of generalized Wigner matrices introduced in [20] but still remain in their perturbative regime. A detailed analysis of the general case was given in [1].

As in (2.8), we define

$$\widetilde{H}^{\mathbf{g}} = \widetilde{H} - \sum_i g_i \mathbf{e}_i \mathbf{e}_i^*, \quad \widetilde{H}^{\mathbf{g}} = \begin{pmatrix} \widetilde{A}^{\mathbf{g}} & \widetilde{B}^* \\ \widetilde{B} & \widetilde{D}^{\mathbf{g}} \end{pmatrix}, \quad \mathbf{g} = (g_1, \ldots, g_N) \in \mathbb{R}^N, \tag{6.6}$$

where $\widetilde{A}^{\mathbf{g}}$ is a $W \times W$ matrix. We define

$$\widetilde{Q}_e^{\mathbf{g}} = \widetilde{A}^{\mathbf{g}} - \widetilde{B}^* \left(\widetilde{D}^{\mathbf{g}} - e\right)^{-1} \widetilde{B}. \tag{6.7}$$

Clearly $Q_e^{\mathbf{g}}(t,q)$ defined in (6.1) equals to $\widetilde{Q}_e^{\mathbf{g}}(t,q)$ if we choose

$$\widetilde{H} = \widetilde{H}(t,q) = \begin{pmatrix} \sqrt{t}A_1 + \sqrt{1-q}A_2 & B^* \\ B & D \end{pmatrix}. \tag{6.8}$$

We now prove the local law of $\widetilde{Q}_e^{\mathbf{g}} = \widetilde{Q}_e^{\mathbf{g}}(t,q)$ by going to the large matrix. In the following everything depends on the parameters t, q but we will often omit this from the notation.

For any $\widetilde{H}$ and complex parameters $z, z' \in \mathbb{C}$ we define

$$\widetilde{G}^{\mathbf{g}}(z,z') := \begin{pmatrix} \widetilde{A}^{\mathbf{g}} - z & \widetilde{B}^* \\ \widetilde{B} & \widetilde{D}^{\mathbf{g}} - z' \end{pmatrix}^{-1} := \left(\widetilde{H}^{\mathbf{g}}(z,z')\right)^{-1} := \left(\widetilde{H}^{\mathbf{g}} - zJ - z'J'\right)^{-1} \tag{6.9}$$

with $J_{ij} = \delta_{ij}\mathbb{1}(i \leqslant W)$ and $J'_{ij} = \delta_{ij}\mathbb{1}(i > W)$. Clearly

$$(\widetilde{Q}_e^{\mathbf{g}} - z)^{-1}_{ij} = \widetilde{G}^{\mathbf{g}}(z, e)_{ij}, \quad 1 \leqslant i, j \leqslant W$$

Note that $\widetilde{G}^{\mathbf{g}}(z, z')$ is not a Green's function unless $z = z'$; we will call it *generalized Green function.* In Lemma 6.3 below we show that an analogue of the local law holds for $\widetilde{G}^{\mathbf{g}}(z, z')$ in a sense that its diagonal entries are well approximated by deterministic functions $M_i^{\mathbf{g}}(z, z')$ and the off diagonal entries are small. The functions $M_i^{\mathbf{g}}$ are defined via a self-consistent equation in the following lemma.

Lemma 6.2. *Recall $m(z)$ defined in* (4.3). *For $z \in \mathbb{C}$, such that $\operatorname{Im} z > 0$, $|z| \leqslant C$, and $|z^2 - 4| \geqslant \kappa$, for some fixed $C, \kappa > 0$, we define*

$$A(z, \zeta) := \left\{ z' \in \mathbb{C} : \operatorname{Im} z' > 0, \quad |z - z'| \leqslant N^{-\zeta} \right\} \subset \mathbb{C}.$$

For any $z' \in A(z, \zeta)$, $\|\mathbf{g}\|_\infty \leqslant N^{-\zeta}$, there is a unique solution $M_i^{\mathbf{g}}(z, z')$ to the equation

$$\frac{1}{M_i^{\mathbf{g}}(z, z')} = -(z' - z)\mathbb{1}_{i>W} - g_i - z - \sum_{j=1}^{N} s_{ij} M_j^{\mathbf{g}}(z, z'), \quad 1 \leqslant i \leqslant N \tag{6.10}$$

with the constraint

$$\max_i |M_i^{\mathbf{g}}(z, z') - m(z)| = \mathrm{O}(\log N)^{-1}. \tag{6.11}$$

Furthermore, $M_i^{\mathbf{g}}(z, z')$ is continuous w.r.t. to z' and $\mathbf{g}$, and it satisfies the following bound

$$\max_i |M_i^{\mathbf{g}}(z, z') - m(z)| = \mathrm{O}(\log N) \left(|z - z'| + \|\mathbf{g}\|_\infty \right), \tag{6.12}$$

in particular $M_i^{\mathbf{g}=0}(z, z) = m(z)$.

This theorem in a very general setup (without the restriction (6.5)) was proved in Lemma 4.4 of [2]. In particular, it showed the existence, uniqueness and stability for any small additive perturbation of the equation

$$\frac{1}{M_i} = -z - \sum_j s_{ij} M_j. \tag{6.13}$$

which has a unique solution $M_i = m(z)$ in the upper half plane. In other words, the solution $\mathbf{M}(\mathbf{d})$ of the perturbed equation

$$\frac{1}{M_i(\mathbf{d})} = -z - d_i - \sum_j s_{ij} M_j(\mathbf{d}) \tag{6.14}$$

depends analytically on the vector $\mathbf{d}$ for $\|\mathbf{d}\| \leqslant c/\log N$. (Thanks to (6.5), here we need only the special case when the perturbation is around the semicircle, $M_i = m$, this result was essentially contained in [20] although not stated explicitly.) The necessary input is a bound on the norm

$$\left\| \frac{1}{1 - m^2(z) S} \right\|_{\ell^\infty \to \ell^\infty} \leqslant C_\varepsilon \log N, \qquad \text{for } |z^2 - 4| \geqslant \varepsilon, \tag{6.15}$$

that was first proven in [20], see also part (ii) of Proposition A.2 in [12]. The bound (6.15) requires a spectral gap above -1 in the spectrum of S which is guaranteed by Lemma A.1 from [20] under the condition (2.2). In fact, for our band matrices the $\log N$ factor in (6.15) can be removed, see Lemma 2.11 in [2].

Lemma 6.3. *Recall $\widetilde{G}^{\mathbf{g}}(z, z')$, the generalized Green's function of $\widetilde{H}$ from* (6.9). *Let Ω be the subset of the probability space such that for any two complex numbers $y, y' \in \mathbb{C}$ satisfying $0 \leqslant \operatorname{Im} y' \leqslant \operatorname{Im} y$ and $|y|$, $|y'| \leqslant 3$, we have*

$$\|\widetilde{G}^{\mathbf{g}}(y, y')\| \leqslant C(\operatorname{Im} y)^{-1}. \tag{6.16}$$

Suppose that $\mathbb{P}(\Omega) \geqslant 1 - N^{-D}$ for any fixed $D > 0$. Assume that $\mathbf{g}$, z and z' satisfy

$$\|\mathbf{g}\|_\infty \leqslant N^{-1/2}, \qquad |z^2 - 4| \geqslant \kappa, \quad N^{-1+\zeta} \leqslant \operatorname{Im} z \leqslant \zeta^{-1}, \qquad \zeta, \kappa > 0$$

and

$$|z - z'| \leqslant N^{-\zeta}, \qquad 0 \leqslant \operatorname{Im} z' \leqslant \operatorname{Im} z.$$

Then for any small $\varepsilon > 0$, we have

$$\max_{ij} \left| \widetilde{G}^{\mathbf{g}}_{ij}(z, z') - M^{\mathbf{g}}_i(z, z')\delta_{ij} \right| \leqslant (N\eta)^{-1/2} N^{\varepsilon}, \qquad \eta = \operatorname{Im} z, \tag{6.17}$$

holds with probability greater than $1 - N^{-D}$ for any fixed $D > 0$.

Note that both the condition (6.16) and the estimate (6.17) are uniform in $\operatorname{Im} y'$ and $\operatorname{Im} z'$, respectively, in particular (6.17) holds even if z' is on the real axis. This is formulated more explicitly in the following:

Corollary 6.4. *In the setting of Lemma 6.3, we assume $\|\mathbf{g}\|_\infty \leqslant N^{-1/2}$ and pick an $e \in \mathbb{R}$ with $|e| \leqslant 2 - \kappa$ for some $\kappa > 0$. Then we have*

$$\max_{1 \leqslant i,j \leqslant W} \sup_{N^{-1+\zeta} \leqslant \operatorname{Im} z \leqslant N^{-\zeta}} \sup_{E : |E-e| \leqslant N^{-\zeta}} \left| \widetilde{G}^{\mathbf{g}}_{ij}(z, e) - M^{\mathbf{g}}_i(z, e)\delta_{ij} \right| \leqslant (N\eta)^{-1/2} N^{\varepsilon}, \quad z = E + i\eta \tag{6.18}$$

holds with probability $1 - N^{-D}$ for any fixed $D > 0$ and $\varepsilon, \zeta > 0$.

Proof. From Lemma 6.3, we know (6.18) holds for fixed z and e. Hence we only need to prove that that they hold at same time for all $z = E + i\eta : |E - e| \leqslant N^{-\zeta}$ and $N^{-1+\zeta} \leqslant \eta \leqslant N^{-\zeta}$. We choose an N^{-10}-grid in both parameter spaces so the validity of (6.18) can be simultaneously guaranteed for each element of this net. Since in Ω we have $\left| \partial_z \widetilde{G}^{\mathbf{g}}_{ij} \right| \leqslant \|\widetilde{G}^{\mathbf{g}}\|^2 \leqslant \eta^{-2} \leqslant N^2$, and the same bound holds for $\partial_e \widetilde{G}^{\mathbf{g}}_{ij}$, we can approximate $\widetilde{G}^{\mathbf{g}}_{ij}(z, e)$ at a nearby grid point with very high accuracy. The same argument holds for $M^{\mathbf{g}}_i(z, e)$ by the stability of its defining equation. This proves Corollary 6.4. □

Proof of Lemma 6.3. For the proof we proceed in three steps.

Step 1: We first consider the case $z = z'$. By (6.12), we only need to prove that for any small $\varepsilon > 0$

$$\max_{ij} \left| \widetilde{G}^{\mathbf{g}}_{ij}(z, z) - m(z)\delta_{ij} \right| \leqslant (N\eta)^{-1/2} N^{\varepsilon} \tag{6.19}$$

holds with probability greater than $1 - N^{-D}$. To prove this estimate, we claim that there exists a set Ξ so that $\mathbb{P}(\Xi) \geqslant 1 - N^{-D}$ for any $D > 0$, and in Ξ

$$\max_{ij} |\widetilde{G}^{\mathbf{g}}_{ij} - m(z)\delta_{ij}| \leqslant (\log N)^{-1}.$$

Furthermore an approximate self-consistent equation for $\widetilde{G}^{\mathbf{g}}_{ii}$ holds in Ξ; more precisely, we have

$$\mathbb{1}_\Xi \left| \left(\widetilde{G}^{\mathbf{g}}_{ii} \right)^{-1} - \widetilde{H}_{ii} - g_i + z + \sum_j \widetilde{s}_{ij} \widetilde{G}^{\mathbf{g}}_{jj} \right| \leqslant (N\eta)^{-1/2} N^{\varepsilon}.$$

These facts were shown in [21] with $\mathbf{g} = 0$ and the same argument holds to the letter including a small perturbation $\mathbf{g}$. Since by assumption $\|\mathbf{g}\|_\infty \leqslant N^{-1/2}$, and $\widetilde{s}_{ij} = (1 + \mathrm{O}(N^{-1/2-c}))s_{ij}$ we obtain

$$\mathbb{1}_\Xi \left| \left(\widetilde{G}^{\mathbf{g}}_{ii} \right)^{-1} - \widetilde{H}_{ii} + z + \sum_j s_{ij} \widetilde{G}^{\mathbf{g}}_{jj} \right| \leqslant (N\eta)^{-1/2} N^{\varepsilon}.$$

Since $|\widetilde{H}_{ii}| \leqslant N^{-1/2+\varepsilon}$ with very high probability, using the stability of the unperturbed self-consistent equation as in [21], we obtain (6.19).

Step 2: Proof of (6.17) *for* $z' \neq z$. Clearly the $\widetilde{G}^{\mathbf{g}}(z,z') - \widetilde{G}^{\mathbf{g}}(z,z)$ is a continuous function w.r.t. z' and it equals to zero at $z = z'$. We define the following interpolation between $y(0) = z$ to $y(1) = \operatorname{Re} z' + \mathrm{i}\operatorname{Im} z$ and then to $y(2) = z'$:

$$y(s) = \begin{cases} (1-s)\operatorname{Re} z + s\operatorname{Re} z' + \mathrm{i}\operatorname{Im} z & 0 \leqslant s \leqslant 1 \\ \operatorname{Re} z' + (2-s)\mathrm{i}\operatorname{Im} z + (s-1)\mathrm{i}\operatorname{Im} z' & 1 \leqslant s \leqslant 2. \end{cases}$$

Denote by $s_k = kN^{-4}$ and $y_k = y(s_k)$, and our goal is to prove that (6.17) holds for $z' = y_k$ for $k = 2N^4$. We have proved in Step 1 that (6.17) holds for $z' = y_{k=0}$ and we now apply induction. For any fixed $\alpha < 1/2$, we define the event $\Xi_k^{(\alpha)} \subset \Omega$ as

$$\Xi_k^{(\alpha)} := \Omega \cap \Big\{ \max_{ij} \Big| \widetilde{G}^{\mathbf{g}}_{ij}(z, y_k) - M^{\mathbf{g}}_i(z, y_k)\delta_{ij} \Big| \leqslant (N\eta)^{-\alpha} \Big\}.$$

Now we claim that, for any $1 \leqslant k \leqslant 2N^4$, any small $\varepsilon > 0$ and any large D, we have

$$\mathbb{P}\Big\{ \Big(\bigcap_{\ell \leqslant k} \Xi_\ell^{(1/4)} \Big) \setminus \Xi_k^{(1/2-\varepsilon)} \Big\} \leqslant N^{-D}. \tag{6.20}$$

Assuming this estimate is proved, we continue to prove (6.17). Recall the bound

$$\mathbb{P}(\Xi_0^{(1/2-\varepsilon)}) \geqslant 1 - N^{-D} \tag{6.21}$$

from Step 1. Simple calculus and (6.16) yield that

$$|\partial_{z'} \widetilde{G}^{\mathbf{g}}_{ij}| \leqslant \|\widetilde{G}^{\mathbf{g}}\|^2 \leqslant N^2$$

holds in the set Ω. Hence we can estimate the difference between $\widetilde{G}^{\mathbf{g}}_{ij}(z, y_{k+1})$ and $\widetilde{G}^{\mathbf{g}}_{ij}(z, y_k)$ by N^{-2}. Similar estimate holds between $M^{\mathbf{g}}_i(z, y_k)$ and $M^{\mathbf{g}}_i(z, y_{k+1})$ by the stability of the self-consistent equation (6.10) at the parameter (z, y_k), provided by Lemma 4.4 of [2].

These bounds easily imply that

$$\mathbb{P}\Big(\Xi_k^{(1/2-\varepsilon)} \setminus \Xi_{k+1}^{(1/4)} \Big) \leqslant N^{-D}. \tag{6.22}$$

It is clear that the initial bound (6.21) and the two estimates (6.20) and (6.22) allow us to use induction to conclude $\mathbb{P}(\Xi_k^{(1/2-\varepsilon)}) \geqslant 1 - N^{-D}$ for any $1 \leqslant k \leqslant 2N^4$. We have thus proved (6.17) assuming (6.20).

Step 3. Proof of (6.20). Recall $\widetilde{G}^{\mathbf{g}}$ is defined with $H^{\mathbf{g}}$ in (6.9). We define $H^{\mathbf{g},(i)}(z,z')$ as the matrix obtained by removing the i-th row and column of $H^{\mathbf{g}}(z,z')$ and set

$$\widetilde{G}^{\mathbf{g},(i)}(z,z') := \Big(H^{\mathbf{g},(i)}(z,z') \Big)^{-1}.$$

As in [21], the standard large deviation argument implies that for any $\varepsilon > 0$, in $\Xi_k^{1/4}$,

$$\frac{1}{\widetilde{G}^{\mathbf{g}}_{ii}(z, y_k)} = -(y_k - z)\mathbb{1}(i > W) - g_i - z - \sum_{ij} \widetilde{s}_{ij} \Big(\widetilde{G}^{\mathbf{g},(i)}(z, y_k) \Big)_{jj} + O\Big(N^{-1+\varepsilon} \|\widetilde{G}^{\mathbf{g},(i)}(z, y_k)\|_{HS} \Big)$$

holds with probability $1 - \mathrm{O}(N^{-D})$, where $\|\cdot\|_{HS}$ is the Hilbert-Schmidt norm. The matrix entries of $\widetilde{G}^{\mathbf{g},(i)}$ can be replaced by $\widetilde{G}^{\mathbf{g}}$ by using the identity (see [20, Lemma 4.2])

$$(\widetilde{G}^{\mathbf{g},(\ell)})_{ij} = \widetilde{G}^{\mathbf{g}}_{ij} - \frac{\widetilde{G}^{\mathbf{g}}_{i\ell}\widetilde{G}^{\mathbf{g}}_{\ell i}}{\widetilde{G}^{\mathbf{g}}_{\ell\ell}}, \qquad \ell \neq i, j, \tag{6.23}$$

and using that both off-diagonal matrix elements are bounded by $(N\eta)^{-1/4}$ on $\Xi_k^{(1/4)}$. Together with (6.5), we have obtained the self-consistent equation

$$\frac{1}{\widetilde{G}^{\mathbf{g}}_{ii}(z,y_k)} = -(y_k-z)\mathbb{1}(i>W) - g_i - z - \sum_j s_{ij}\widetilde{G}^{\mathbf{g}}_{jj}(z,y_k) + O\left(N^{-1+\varepsilon}\|\widetilde{G}^{\mathbf{g}}(z,y_k)\|_{HS} + (N\eta)^{-1/2}\right), \quad (6.24)$$

which holds with probability larger than $1-\mathrm{O}(N^{-D})$. The standard argument then uses the so-called Ward identity that the Green function $G=(H-z)^{-1}$ of any self-adjoint matrix H satisfies that

$$\|G(z)\|_{HS}^2 = \eta^{-1}\operatorname{Im}\operatorname{Tr} G(z), \qquad \eta = \operatorname{Im} z. \quad (6.25)$$

In our case, $\widetilde{G}^{\mathbf{g}}$ is not a Green function and this presents the major difficulty. The main idea is to write

$$\widetilde{G}^{\mathbf{g}}(z,y_k) = \widetilde{G}^{\mathbf{g}}(z,\widetilde{y}_k) + \widetilde{G}^{\mathbf{g}}(z,y_k)\mathrm{i}(\eta-\operatorname{Im} y_k)J\widetilde{G}^{\mathbf{g}}(z,\widetilde{y}_k), \quad \widetilde{y}_k = y_k + \mathrm{i}(\eta-\operatorname{Im} y_k),$$

where J is the matrix defined by $J_{ij}=1_{1\leqslant i\leqslant W}\delta_{ij}$ and the imaginary part of $\widetilde{y}_k$ equals $\eta=\operatorname{Im} z$. In particular, $\widetilde{G}^{\mathbf{g}}(z,\widetilde{y}_k)$ is a Green function of a self-adjoint matrix, hence the Ward identity is applicable. By definition, $\widetilde{y}_k\in\{y_\ell\,:\,\ell\leqslant k\}$. Hence in the set $\bigcap_{\ell\leqslant k}\Xi_\ell^{(1/4)}\subset\Omega$, we have

$$\|\widetilde{G}^{\mathbf{g}}(z,y_k)\|_{HS} \leqslant \|\widetilde{G}^{\mathbf{g}}(z,\widetilde{y}_k)\|_{HS} + \|\widetilde{G}^{\mathbf{g}}(z,y_k)\mathrm{i}(\eta-\operatorname{Im} y_k)J\widetilde{G}^{\mathbf{g}}(z,\widetilde{y}_k)\|_{HS} \quad (6.26)$$

$$\leqslant \|\widetilde{G}^{\mathbf{g}}(z,\widetilde{y}_k)\|_{HS} + \eta\|\widetilde{G}^{\mathbf{g}}(z,y_k)\|\|\widetilde{G}^{\mathbf{g}}(z,\widetilde{y}_k)\|_{HS} \leqslant C\left[\eta^{-1}\operatorname{Im}\operatorname{Tr}\widetilde{G}^{\mathbf{g}}(z,\widetilde{y}_k)\right]^{1/2}, \quad (6.27)$$

where we have used the Ward identity (6.25) for $\widetilde{G}^{\mathbf{g}}(z,\widetilde{y}_k)$ and (6.16) for $\widetilde{G}^{\mathbf{g}}(z,y_k)$. Inserting this bound into (6.24), we have that in $\bigcap_{\ell\leqslant k}\Xi_\ell^{(1/4)}$ with probability $1-\mathrm{O}(N^{-D})$ that for any $\varepsilon>0$,

$$\frac{1}{\widetilde{G}^{\mathbf{g}}_{ii}(z,y_k)} = -(y_k-z)\mathbb{1}(i>W) - g_i - z - \sum_{ij} s_{ij}\widetilde{G}^{\mathbf{g}}_{jj}(z,y_k) + O\left((N\eta)^{-1/2}N^{\varepsilon}\right).$$

Now we compare this equation with (6.10) and notice that both are perturbations of the equation (6.13) that is stable in an $O((\log N)^{-1})$ neighborhood of the vector $\mathbf{m}$. We obtain that in $\bigcap_{\ell\leqslant k}\Xi_\ell^{(1/4)}$ with probability $1-\mathrm{O}(N^{-D})$

$$\max_i\left|\widetilde{G}^{\mathbf{g}}_{ii}(z,y_k) - M_i^{\mathbf{g}}(z,y_k)\right| = O\left((N\eta)^{-1/2}N^{\varepsilon}\right).$$

For the off-diagonal terms i.e., $\widetilde{G}^{\mathbf{g}}_{ij}(z,y_k)$, similarly as in [20], we know that in $\Xi_k^{(1/4)}$ with probability $1-\mathrm{O}(N^{-D})$

$$\left|\widetilde{G}^{\mathbf{g}}_{ij}(z,y_k)\right| \leqslant \left|\widetilde{G}^{\mathbf{g}}_{ij}(z,y_k)\right|\left|\widetilde{G}^{\mathbf{g},(i)}_{jj}(z,y_k)\right|\left(N^{-1/2+\varepsilon} + N^{-1+\varepsilon}\|(\widetilde{G}^{\mathbf{g},(ij)}(z,y_k))\|_{HS}\right)$$

Then with (6.26) and (6.23), we obtain that in $\bigcap_{\ell\leqslant k}\Xi_\ell^{(1/4)}$ with probability $1-\mathrm{O}(N^{-D})$

$$\left|\widetilde{G}^{\mathbf{g}}_{ij}(z,y_k)\right| = O\left((N\eta)^{-1/2}N^{\varepsilon}\right).$$

This completes the proof of (6.20) and Lemma 6.3. □

6.2 Operator bound of $G(z,z')$. As explained in the beginning of this section, we are going to prove Theorem 6.1 with Corollary 6.4. For this purpose, we need to prove that band matrix satisfies the assumption (6.16). In this subsection, we prove a sufficient condition for (6.16). We formulate the result in a non-random setup and later we will check that the conditions hold with very high probability in case of our random band matrix.

Lemma 6.5. *Let H be a (non random) symmetric $N \times N$ matrix and consider its block decomposition as*

$$H = \begin{pmatrix} A & B^* \\ B & D \end{pmatrix}.$$

Suppose that for (small) $\mu > 0$ and C_0, the following holds:

(i) there does not exist $e \in \mathbb{R}$, $\mathbf{u} \in \mathbb{R}^N$ such that

$$\|\mathbf{u}\| = 1,\ \|B^*\mathbf{u}\| \leqslant \mu,\ \|(D-e)\mathbf{u}\| \leqslant \mu. \tag{6.28}$$

(ii) The submatrices are bounded:

$$\|A\| + \|B\| + \|D\| \leqslant C_0. \tag{6.29}$$

Define

$$G(z,z') := \begin{pmatrix} A-z & B^* \\ B & D-z' \end{pmatrix}^{-1}.$$

Then for any large $C'' > 0$, there exists $C' > 0$, depending only on C'', μ and C_0, such that if

$$z, z' \in \mathbb{C}, \quad 0 \leqslant \operatorname{Im} z' \leqslant \operatorname{Im} z, \quad |z| + |z'| \leqslant C'',$$

then we have

$$\|G(z,z')\| \leqslant \frac{C'}{\operatorname{Im} z}. \tag{6.30}$$

Proof of Lemma 6.5. Define the symmetric matrix

$$P := \begin{pmatrix} A - \operatorname{Re} z & B^* \\ B & D - \operatorname{Re} z' \end{pmatrix},$$

then by resolvent identity we have

$$G = G(z,z') = \frac{1}{P - \mathrm{i}\operatorname{Im} z} + \frac{1}{P - \mathrm{i}\operatorname{Im} z}(\operatorname{Im} z' - \operatorname{Im} z)\mathrm{i}JG,$$

where the matrix J was already defined by $J_{ij} = 1_{1 \leqslant i \leqslant W}\delta_{ij}$. Here W is the size of the block A. Then

$$\|G\| \leqslant (\operatorname{Im} z)^{-1} + \frac{\operatorname{Im} z - \operatorname{Im} z'}{\operatorname{Im} z}\|G\|,$$

which implies $\|G\| \leqslant (\operatorname{Im} z')^{-1}$. Furthermore, with [I removed a prime from J, I think J' was something obsolete]

$$\partial_{z'} G = GJG,$$

it is easy to see by integrating $\partial_{z'} G$ from $z' = e'$ to $z' = e' + i\eta'$ that we only need to prove (6.30) for the case $\operatorname{Im} z' = 0$. Hence from now on, we assume that

$$z = e + \mathrm{i}\eta, \quad z' = e'.$$

Applying Schur formula, we obtain

$$G = G(z,z') = \begin{pmatrix} A-z & B^* \\ B & D-z' \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{A-z-B^*(D-z')^{-1}B} & -\frac{1}{A-z}B^*\frac{1}{D-z'-B(A-z)^{-1}B^*} \\ -\frac{1}{D-z'-B(A-z)^{-1}B^*}B\frac{1}{A-z} & \frac{1}{D-z'-B(A-z)^{-1}B^*} \end{pmatrix}.$$

First with $\operatorname{Im} z' = 0$, we have the trivial bounds (which follows by $\|(P+i\eta)^{-1}\| \leqslant \eta^{-1}$ for any symmetric matrix P):

$$\left\| \left(A - z - B^* \frac{1}{D - z'} B \right)^{-1} \right\| \leqslant \frac{1}{|\operatorname{Im} z|}, \qquad \left\| (A-z)^{-1} \right\| \leqslant \frac{1}{|\operatorname{Im} z|}, \tag{6.31}$$

which controls the upper left corner of G. Second, we claim that

$$\| \frac{1}{A-z} B^* \frac{1}{D - z' - B \frac{1}{A-z} B^*} \|^2 \leqslant \frac{1}{|\operatorname{Im} z|} \left\| \frac{1}{D - e' - B \frac{1}{A-z} B^*} \right\|. \tag{6.32}$$

For the proof, picking any nonzero vector $\mathbf{v}$ and setting $\mathbf{u} = (D - e' - B\frac{1}{A-z}B^*)^{-1}\mathbf{v}$, we have

$$\begin{aligned}
\|\mathbf{v}\| &= \left\| \Big(D - e' - B \frac{1}{A-z} B^* \Big) \mathbf{u} \right\| \geqslant \frac{1}{\|\mathbf{u}\|} \left| \Big\langle \mathbf{u}, \Big(D - e' - B \frac{1}{A-z} B^* \Big) \mathbf{u} \Big\rangle \right| \\
&= \frac{1}{\|\mathbf{u}\|} \left| \Big\langle \mathbf{u}, \Big(D - e' - B \frac{A-e}{(A-e)^2 + \eta^2} B^* \Big) \mathbf{u} \Big\rangle - i \Big\langle \mathbf{u}, \Big(B \frac{\eta}{(A-e)^2 + \eta^2} B^* \Big) \mathbf{u} \Big\rangle \right| \\
&\geqslant \frac{\eta}{\|\mathbf{u}\|} \Big\langle u, B \frac{1}{(A-e)^2 + \eta^2} B^* \mathbf{u} \Big\rangle = \frac{\eta}{\|\mathbf{u}\|} \Big\langle u, B \frac{1}{|A-z|^2} B^* \mathbf{u} \Big\rangle \\
&= \frac{\eta}{\|\mathbf{u}\|} \left\| \frac{1}{A-z} B^* \mathbf{u} \right\|^2.
\end{aligned}$$

Changing the vector $\mathbf{u}$ back to $\mathbf{v}$, we have

$$\left\| \frac{1}{A-z} B^* \frac{1}{D - e' - B \frac{1}{A-z} B^*} \mathbf{v} \right\|^2 \leqslant \frac{1}{|\operatorname{Im} z|} \left\| \frac{1}{D - e' - B \frac{1}{A-z} B^*} \mathbf{v} \right\| \|\mathbf{v}\|,$$

which implies (6.32). Now it only remains to bound $\left\| \left(D - e' - B \frac{1}{A-z} B^* \right)^{-1} \right\|$ by C/η, which would then control all other three blocks of G. Suppose for some normalized vector $\mathbf{u}$ and small $\widetilde{\mu} > 0$, we have

$$\left\| (D - e')\mathbf{u} - B \frac{1}{A-z} B^* \mathbf{u} \right\| \leqslant \widetilde{\mu} \eta. \tag{6.33}$$

Then

$$\widetilde{\mu}\eta \geqslant \left| \operatorname{Im} \Big\langle \mathbf{u}, (D-e')\mathbf{u} + B \frac{1}{A-z} B^* \mathbf{u} \Big\rangle \right| = \Big\langle \mathbf{u}, \Big(B \frac{\eta}{(A-e)^2 + \eta^2} B^* \Big) \mathbf{u} \Big\rangle.$$

Then for some $C_1 > 0$, we have

$$\widetilde{\mu} \geqslant \Big\langle \mathbf{u}, B \frac{1}{|A-z|^2} B^* \mathbf{u} \Big\rangle \geqslant \frac{1}{C_1} \|B^* \mathbf{u}\|^2 \tag{6.34}$$

where we used that the fact $|A-z|^2$ is bounded. This shows that

$$\|B^* \mathbf{u}\| \leqslant \sqrt{C_1 \widetilde{\mu}}, \qquad \|B B^* \mathbf{u}\| \leqslant \sqrt{C_0 C_1 \widetilde{\mu}} \tag{6.35}$$

by (6.29). From (6.34), we also have

$$\left\| B \frac{1}{A-z} B^* \mathbf{u} \right\|^2 = \Big\langle \mathbf{u}, B \frac{1}{A-\bar{z}} B^* B \frac{1}{A-z} B^* \mathbf{u} \Big\rangle \leqslant C_0 \Big\langle \mathbf{u}, B \frac{1}{|A-z|^2} B^* \mathbf{u} \Big\rangle \leqslant C_0 \widetilde{\mu}.$$

Then with (6.33), for small enough $\widetilde{\mu}$, we have

$$\|(D-e')\mathbf{u}\| \leqslant \left\| B \frac{1}{A-z} B^* \mathbf{u} \right\| + \widetilde{\mu}\eta \leqslant \sqrt{C_0 \widetilde{\mu}} + \widetilde{\mu}\eta \leqslant C \sqrt{\widetilde{\mu}}. \tag{6.36}$$

Combining (6.35), (6.36) and (6.28), we obtain (6.33) does not hold for small enough $\widetilde{\mu}$. Together with (6.31) and (6.32), we completed the proof of Lemma 6.5. □

6.3 Proof of Theorem 6.1. Now we return to prove Theorem 6.1, i.e., the local law of the Green's function of some particular matrices which are derived from band matrix.

Proof of (6.2). As explained in (6.8), we know that $Q_e^{\mathbf{g}}(t,q)$ is a matrix of the form in (6.7). We will apply Lemma 3.2 with $M = W$ and $L = N - W$. Since H is a band matrix with band width $4W-1$, see (2.2), the upper $W \times W$ block of B has variance $(4W-1)^{-1}$, i.e.

$$s_{ij} = \mathbb{E}B_{ij}^2 = \frac{1}{4W-1}, \quad 1 \leqslant i,j \leqslant W.$$

Using this information in (3.5) to estimate $\sum_{1\leqslant i\leqslant W}|u_i|^2$ from below and inserting this into the last condition in (3.1), we learn that for some small $\mu > 0$ we have

$$\mathbb{P}\left(\exists e \in \mathbb{R},\ \exists \mathbf{u} \in \mathbb{R}^{N-W}\ :\ \|\mathbf{u}\| = 1,\ \|B^*\mathbf{u}\| \leqslant \mu,\ \|(D^{\mathbf{g}} - e)\mathbf{u}\| \leqslant \mu\right) \leqslant N^{-D}.$$

We also know that $\|H\|$, hence $\|A\|$, $\|B\|$ and $\|D\|$ are all bounded by a large constant with very high probability. Then using Lemma 6.5, we obtain that for some large $C > 0$, we have

$$\mathbb{P}\left(\exists z, z' \in \mathbb{C} : |z|, |z'| \leqslant 3,\ \operatorname{Im} z \geqslant \operatorname{Im} z' \geqslant 0, \|\widetilde{G}^{\mathbf{g}}(z,z')\| \geqslant C(\operatorname{Im} z)^{-1}\right) \leqslant N^{-D}.$$

With this bound, we can use Corollary 6.4. Together with (6.12), we complete the proof of (6.2). □

Proof of (6.3). Because of (6.2), it only remains to prove (6.3) for

$$|E - e| \leqslant N^{-\omega}, \quad N^{-\omega} \leqslant \eta \leqslant 1. \tag{6.37}$$

Recall $\xi_k^{\mathbf{g}}(e,t,q)$, $1 \leqslant k \leqslant W$ is the k-th eigenvalue of $Q_e^{\mathbf{g}}(t,q)$. Then

$$\operatorname{Im}\sum_j \left[Q_e^{\mathbf{g}}(t,q) - z\right]_{jj}^{-1} = \operatorname{Im}\operatorname{Tr}\frac{1}{Q_e^{\mathbf{g}}(t,q) - E - i\eta} = \sum_k \frac{\eta}{|\xi_k^{\mathbf{g}}(e,t,q) - E|^2 + \eta^2}, \quad z = E + i\eta.$$

In our case (6.37), we know

$$|\xi_k^{\mathbf{g}}(e,t,q) - E|^2 + \eta^2 \sim |\xi_k^{\mathbf{g}}(e,t,q) - e|^2 + \eta^2.$$

Therefore, we only need to prove that there exists $c > 0$ such that

$$\mathbb{P}\left(\exists \eta,\ N^{-\omega} \leqslant \eta \leqslant 1\ s.t.\ \frac{1}{W}\operatorname{Im}\operatorname{Tr}\frac{1}{Q_e^{\mathbf{g}}(t,q) - e - i\eta} \notin [c, c^{-1}]\right) \leqslant N^{-D}.$$

After adjusting the constant c, it will be implied by the following high probability bound on the eigenvalue density:

$$\mathbb{P}\left(\exists \eta,\ N^{-\omega} \leqslant \eta \leqslant 1\ s.t.\ (N\eta)^{-1}\#\left\{k : \xi_k^{\mathbf{g}}(e,t,q) \in [e-\eta, e+\eta]\right\} \notin [c, c^{-1}]\right) \leqslant N^{-D}. \tag{6.38}$$

From Section 4.2 recall the definition of curves $\mathcal{C}_k^{\mathbf{g}}(e)$ constructed from the matrix (4.8). Similarly, starting with the matrix $\widetilde{H}^{\mathbf{g}}$, see (6.6) and (6.8), we can define the curves $e \to \mathcal{C}_k^{\mathbf{g}}(e,t,q)$ for any fixed parameters t, q. As in Lemma 4.4, we have that for any K, there exists C_K such that

$$\mathbb{P}\left(\sup_{e\notin\sigma(D^{\mathbf{g}})}\sup_k \mathbb{1}(|\mathcal{C}_k^{\mathbf{g}}(e,t,q)| \leqslant K)\left|\frac{\mathrm{d}\mathcal{C}_k^{\mathbf{g}}}{\mathrm{d}e}(e,t,q)\right| \leqslant C_K\right) \leqslant N^{-D}. \tag{6.39}$$

It means the slopes of these curves are bounded in $[-K,K]^2$. The crossing points of these curves with $x = y$ line are exactly the points

$$(\lambda_k^{\mathbf{g}}(t,q), \lambda_k^{\mathbf{g}}(t,q)), \quad 1 \leqslant k \leqslant N,$$

where $\lambda_k^{\mathbf{g}}(t,q)$ is the k-th eigenvalue of $\widetilde{H}^{\mathbf{g}}$. By simple perturbation theory and using $|t-q| \leqslant N^{-1/2}$, $\|\boldsymbol{\sigma}\| \leqslant N^{-1/2}$, it is easy to see that with high probability, we have

$$|\lambda_k - \lambda_k^{\mathbf{g}}(t,q)| \ll N^{-\omega}, \quad \lambda_k := \lambda_k(q,q).$$

Note $\lambda_k(q,q)$ is the eigenvalue of a regular generalized Wigner matrix, i.e. $\widetilde{H}^{\mathbf{g}=0}$ at $t=q$ has variances summing up exactly to one in each row. Then together with the rigidity of λ_k, we know

$$\mathbb{P}\left(\exists N^{-\omega} \leqslant \eta \leqslant 1 \ s.t. \ (N\eta)^{-1}\#\left\{k : \mathcal{C}_k^{\mathbf{g}}(e,t,q) \in [e-\eta, e+\eta]\right\} \notin [c, c^{-1}]\right) \leqslant N^{-D}.$$

With (6.39), (note $\frac{\mathrm{d}\mathcal{C}_k^{\mathbf{g}}}{\mathrm{d}e} \leqslant 0$ as in (4.28)) we obtain (6.38) and complete the proof of (6.3). □

Proof of (6.4). With (6.3), we know that

$$\mathbb{P}\Big(\exists x, y \in [e-N^{-\omega}, e+N^{-\omega}], \ |x-y| \geqslant N^{-1+\omega}, \ \ N^{-1}\#\left\{k : \xi_k^{\mathbf{g}}(e,t,q) \in [x,y]\right\} \geqslant |x-y|\log N\Big) \leqslant N^{-D}.$$

It is easy to see that it implies (6.4), which completes the proof of Theorem 6.1. □

References

[1] O. Ajanki, L. Erdős, and T. Kruger, *Universality for general Wigner-type matrices*, prepublication, arXiv:1506.05098 (2015).

[2] ———, *Quadratic vector equations on complex upper half plane*, prepublication, arXiv:1506.05095 (2015).

[3] N. Anantharaman and E. Le Masson, *Quantum ergodicity on large regular graphs*, Duke Math. J. **164** (2015), no. 4, 723–765.

[4] Z. Bao and L. Erdős, *Delocalization for a class of random block band matrices*, to appear in Probab. Theory Related Fields (2016).

[5] P. Bourgade, L. Erdős, H.-T. Yau, and J. Yin, *Fixed energy universality for generalizd Wigner matrices*, to appear in Communications on Pure and Applied Mathematics (2016).

[6] P. Bourgade, J. Huang, and H.-T. Yau, *Eigenvector statistics of sparse random matrices* (2016).

[7] P. Bourgade and H.-T. Yau, *The Eigenvector Moment Flow and local Quantum Unique Ergodicity*, to appear in Commun. Math. Phys. (2016).

[8] Y. Colin de Verdière, *Ergodicité et fonctions propres du laplacien*, Comm. Math. Phys. **102** (1985), no. 3, 497–502 (French, with English summary).

[9] ———, *Spectres de graphes*, Cours Spécialisés [Specialized Courses], vol. 4, Société Mathématique de France, Paris, 1998 (French, with English and French summaries).

[10] M. Disertori, L. Pinson, and T. Spencer, *Density of states for random band matrices*, Commun. Math. Phys. **232** (2002), 83–124.

[11] K. Efetov, *Supersymmetry in disorder and chaos*, Cambridge University Press (1997).

[12] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *The local semicircle law for a general class of random matrices*, Elect. J. Prob. **18** (2013), no. 59, 1–58.

[13] L. Erdős and A. Knowles, *Quantum Diffusion and Delocalization for Band Matrices with General Distribution*, Ann. Inst. H. Poincaré **12** (2011), no. 7, 1227-1319.

[14] L. Erdős, S. Péché, J. A. Ramírez, B. Schlein, and H.-T. Yau, *Bulk universality for Wigner matrices*, Comm. Pure Appl. Math. **63** (2010), no. 7, 895–925.

[15] L. Erdős, B. Schlein, and H.-T. Yau, *Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices*, Annals of Probability **37** (2009), 815–852.

[16] ———, *Universality of random matrices and local relaxation flow*, Invent. Math. **185** (2011), no. 1, 75–119.

[17] L. Erdős and K. Schnelli, *Universality for Random Matrix Flows with Time-dependent Density*, preprint, arXiv:1504.00650 (2015).

[18] L. Erdős and H.-T. Yau, *Universality of local spectral statistics of random matrices*, Bull. Amer. Math. Soc. (N.S.) **49** (2012), no. 3, 377–414.

[19] ———, *Gap universality of generalized Wigner and beta ensembles*, J. Eur. Math. Soc. **17** (2015), 1927–2036.

[20] L. Erdős, H.-T. Yau, and J. Yin, *Bulk universality for generalized Wigner matrices*, Probab. Theory Related Fields **154** (2012), no. 1-2, 341–407.

[21] ———, *Rigidity of eigenvalues of generalized Wigner matrices*, Adv. Math. **229** (2012), no. 3, 1435–1515.

[22] Y.V. Fyodorov and A.D. Mirlin, *Scaling properties of localization in random band matrices: A σ-model approach.*, Phys. Rev. Lett. **67** (1991), 2405–2409.

[23] R. Holowinsky, *Sieving for mass equidistribution*, Ann. of Math. (2) **172** (2010), no. 2, 1499–1516.

[24] R. Holowinsky and K. Soundararajan, *Mass equidistribution for Hecke eigenforms*, Ann. of Math. (2) **172** (2010), no. 2, 1517–1528.

[25] K. Johansson, *Universality of the local spacing distribution in certain ensembles of Hermitian Wigner matrices*, Comm. Math. Phys. **215** (2001), no. 3, 683–705.

[26] A. Knowles and J. Yin, *Eigenvector distribution of Wigner matrices*, Probability Theory and Related Fields **155** (2013), no. 3, 543–582.

[27] B. Landon and H.-T. Yau, *Convergence of local statistics of Dyson Brownian motion*, preprint, arXiv:1504.03605 (2015).

[28] J.-O. Lee, K. Schnelli, B. Stetler, and H.-T Yau, *Bulk universality for deformed Wigner matrices*, to appear in Annals of Probability (2015).

[29] E. Lindenstrauss, *Invariant measures and arithmetic quantum unique ergodicity*, Ann. of Math. (2) **163** (2006), no. 1, 165–219.

[30] Z. Rudnick and P. Sarnak, *The behaviour of eigenstates of arithmetic hyperbolic manifolds*, Comm. Math. Phys. **161** (1994), no. 1, 195–213.

[31] J. Schenker, *Eigenvector localization for random band matrices with power law band width*, Comm. Math. Phys. **290** (2009), 1065–1097.

[32] T. Shcherbina, *Universality of the local regime for the block band matrices with a finite number of blocks*, J. Stat. Phys. **155** (2014), 466–499.

[33] ———, *On the Second Mixed Moment of the Characteristic Polynomials of 1D Band Matrices*, Communications in Mathematical Physics **328** (2014), 45–82.

[34] ———, *Universality of the second mixed moment of the characteristic polynomials of the 1D band matrices: Real symmetric case*, J. Math. Phys. **56** (2015).

[35] A. I. Shnirel'man, Uspekhi Mat. Nauk **29** (1974), no. 6, 181–182.

[36] S. Sodin, *The spectral edge of some random band matrices*, Ann. of Math. **173** (2010), no. 3, 2223-2251.

[37] T. Spencer, *Random banded and sparse matrices (Chapter 23)*, Oxford Handbook of Random Matrix Theory, edited by G. Akemann, J. Baik, and P. Di Francesco.

[38] T. Tao and V. Vu, *Random matrices: universality of local eigenvalue statistics*, Acta Math. **206** (2011), no. 1.

[39] S. Zelditch, *Uniform distribution of eigenfunctions on compact hyperbolic surfaces*, Duke Math. J. **55** (1987), no. 4, 919–941.