The Local Semicircle Law for a General Class of Random Matrices

László Erdős[1*] Antti Knowles[2†] Horng-Tzer Yau[3‡] Jun Yin[4§]

Institute of Science and Technology Austria, Am Campus 1, A-3400 Klosterneuburg, Austria
lerdos@ist.ac.at [1]

Courant Institute, New York University, 251 Mercer Street, New York, NY 10012, USA
knowles@cims.nyu.edu [2]

Department of Mathematics, Harvard University, Cambridge MA 02138, USA
htyau@math.harvard.edu [3]

Department of Mathematics, University of Wisconsin, Madison, WI 53706, USA
jyin@math.uwisc.edu [4]

May 24, 2013

We consider a general class of $N \times N$ random matrices whose entries h_{ij} are independent up to a symmetry constraint, but not necessarily identically distributed. Our main result is a local semicircle law which improves previous results [17] both in the bulk and at the edge. The error bounds are given in terms of the basic small parameter of the model, $\max_{i,j} \mathbb{E}|h_{ij}|^2$. As a consequence, we prove the universality of the local n-point correlation functions in the bulk spectrum for a class of matrices whose entries do not have comparable variances, including random band matrices with band width $W \gg N^{1-\varepsilon_n}$ with some $\varepsilon_n > 0$ and with a negligible mean-field component. In addition, we provide a coherent and pedagogical proof of the local semicircle law, streamlining and strengthening previous arguments from [6, 17, 19].

Keywords: Random band matrix, local semicircle law, universality, eigenvalue rigidity.

*Partially supported by SFB-TR 12 Grant of the German Research Council. On leave from Institute of Mathematics, University of Munich, Germany.

†Partially supported by NSF grant DMS-0757425.

‡Partially supported by NSF grants DMS-0804279 and Simons Investigator Award.

§Partially supported by NSF grants DMS-1001655 and DMS-1207961.

1. Introduction

Since the pioneering work [31] of Wigner in the fifties, random matrices have played a fundamental role in modelling complex systems. The basic example is the Wigner matrix ensemble, consisting of $N \times N$ symmetric or Hermitian matrices $H = (h_{ij})$ whose matrix entries are identically distributed random variables that are independent up to the symmetry constraint $H = H^*$. From a physical point of view, these matrices represent Hamilton operators of disordered mean-field quantum systems, where the quantum transition rate from state i to state j is given by the entry h_{ij}.

A central problem in the theory or random matrices is to establish the *local universality of the spectrum*. Wigner observed that the distribution of the distances between consecutive eigenvalues (the gap distribution) in complex physical systems follows a universal pattern. The Wigner-Dyson-Gaudin-Mehta conjecture, formalized in [25], states that this 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 [7, 14, 19]; an alternative approach was given in [29] for the special Wigner Hermitian case. The general approach of [7, 14, 19] to prove universality consists of three steps: (i) establish a local semicircle law for the density of eigenvalues; (ii) prove universality of Wigner matrices with a small Gaussian component by analysing the convergence of Dyson Brownian motion to local equilibrium; (iii) remove the small Gaussian component by comparing Green functions of Wigner ensembles with a few matching moments. For an overview of recent results and this three-step strategy, see [16].

Wigner's vision was not restricted to Wigner matrices. In fact, he predicted that universality should hold for any quantum system, described by a large Hamiltonian H, of sufficient complexity. In order to make such complexity mathematically tractable, one typically replaces the detailed structure of H with a statistical description. In this phenomenological model, H is drawn from a random ensemble whose distribution mimics the true complexity. One prominent example where random matrix statistics are expected to hold is the random Schrödinger operator in the delocalized regime. The random Schrödinger operator differs greatly from Wigner matrices in that most of its entries vanish. It describes a model with *spatial structure*, in contrast to the mean-field Wigner matrices where all matrix entries are of comparable size. In order to address the question of universality of general disordered quantum systems, and in particular to probe Wigner's vision, one therefore has to break the mean-field permutational symmetry of Wigner's original model, and hence to allow the distribution of h_{ij} to depend on i and j in a nontrivial fashion. For example, if the matrix entries are labelled by a discrete torus $\mathbb{T} \subset \mathbb{Z}^d$ on the d-dimensional lattice, then the distribution of h_{ij} may depend on the Euclidean distance $|i - j|$ between sites i and j, thus introducing a nontrivial spatial structure into the model. If $h_{ij} = 0$ for $|i - j| > 1$ we essentially obtain the random Schrödinger operator. A random Schrödinger operator models a physical system with a *short-range* interaction, in contrast to the infinite-range, mean-field interaction described by Wigner matrices. More generally, we may consider a *band matrix*, characterized by the property that h_{ij} becomes negligible if $|i - j|$ exceeds a certain parameter, W, called the *band width*, describing the range of the interaction. Hence, by varying the band width W, band matrices naturally interpolate between mean-field Wigner matrices and random Schrödinger operators; see [28] for an overview.

For definiteness, let us focus on the case of a one-dimensional band matrix H. A fundamental conjecture, supported by nonrigorous supersymmetric arguments as well as numerics [23], is that the local spectral statistics of H are governed by random matrix statistics for large W and by Poisson statistics for small W. This transition is in the spirit of the Anderson metal-insulator transition [23, 28], and is conjectured to be sharp around the critical value $W = \sqrt{N}$. In other words, if $W \gg \sqrt{N}$, we expect the universality results of [17–19] to hold. In addition to a transition in the local spectral statistics, an accompanying transition is

conjectured to occur in the behaviour *localization length* of the eigenvectors of H, whereby in the large-W regime they are expected to be completely delocalized and in the small-W regime exponentially localized. The localization length for band matrices was recently investigated in great detail in [8].

Although the Wigner-Dyson-Gaudin-Mehta conjecture was originally stated for Wigner matrices, the methods of [7,14,19] also apply to certain ensembles with independent but not identically distributed entries, which however retain the mean-field character of Wigner matrices. More precisely, they yield universality provided the variances

$$s_{ij} := \mathbb{E}|h_{ij}|^2$$

of the matrix entries are only required to be of comparable size (but not necessarily equal):

$$\frac{c}{N} \leqslant s_{ij} \leqslant \frac{C}{N} \tag{1.1}$$

for some positive constants c and C. (Such matrices were called *generalized Wigner matrices* in [19].) This condition admits a departure from spatial homogeneity, but still imposes a mean-field behaviour and hence excludes genuinely inhomogeneous models such as band matrices.

In the three-step approach to universality outlined above, the first step is to establish the semicircle law on very short scales. In the scaling of H where its spectrum is asymptotically given by the interval $[-2, 2]$, the typical distance between neighbouring eigenvalues is of order $1/N$. The number of eigenvalues in an interval of length η is typically of order $N\eta$. Thus, the smallest possible scale on which the empirical density may be close to a deterministic density (in our case the semicircle law) is $\eta \gg 1/N$. If we characterize the empirical spectral density around an energy E on scale η by its Stieltjes transform, $m_N(z) = N^{-1} \operatorname{Tr}(H - z)^{-1}$ for $z = E + i\eta$, then the local semicircle law around the energy E and in a spectral window of size η is essentially equivalent to

$$|m_N(z) - m(z)| = o(1) \tag{1.2}$$

as $N \to \infty$, where $m(z)$ is the Stieltjes transform of the semicircle law. For any $\eta \gg 1/N$ (up to logarithmic corrections) the asymptotics (1.2) in the bulk spectrum was first proved in [13] for Wigner matrices. The optimal error bound of the form $O((N\eta)^{-1})$ (with an N^{ε} correction) was first proved in [18] in the bulk. (Prior to this work, the best results were restricted to regime $\eta \geqslant N^{-1/2}$; see Bai et al. [1] as well as related concentration bounds in [20].) This result was then extended to the spectral edges in [19]. (Some improvements over the estimates from [13] at the edges, for a special class of ensembles, were obtained in [30].) In [19], the identical distribution of the entries of H was not required, but the upper bound in (1.1) on the variances was necessary. Band matrices in d dimensions with band width W satisfy the weaker bound $s_{ij} \leqslant C/W^d$. (Note that the band width W is typically much smaller than the linear size L of the configuration space $\mathbb{T}$, i.e. the bound W^{-d} is much larger than the inverse number of lattice sites, $L^{-d} = |\mathbb{T}|^{-1} = N^{-1}$.) This motivates us to consider even more general matrices, with the sole condition

$$s_{ij} \leqslant C/M \tag{1.3}$$

on the variances (instead of (1.1)). Here M is a new parameter that typically satisfies $M \ll N$. (From now on, the relation $A \ll B$ for two N-dependent quantities A and B means that $A \leqslant N^{-\varepsilon} B$ for some positive $\varepsilon > 0$.) The question of the validity of the local semicircle law under the assumption (1.3) was initiated in [17], where (1.2) was proved with an error term of order $(M\eta)^{-1/2}$ away from the spectral edges.

The purpose of this paper is twofold. First, we prove a local semicircle law (1.2), under the variance condition (1.3), with a stronger error bound of order $(M\eta)^{-1}$, including energies E near the spectral edge. Away from the spectral edge (and from the origin $E = 0$ if the matrix does not have a band structure), the

result holds for any $\eta \gg 1/M$. Near the edge there is a restriction on how small η can be. This restriction depends explicitly on a norm of the resolvent of the matrix of variances, $S = (s_{ij})$; we give explicit bounds on this norm for various special cases of interest.

As a corollary, we derive bounds on the eigenvalue counting function and rigidity estimates on the locations of the eigenvalues for a general class of matrices. Combined with an analysis of Dyson Brownian motion and the Green function comparison method, this yields bulk universality of the local eigenvalue statistics in a certain range of parameters, which depends on the matrix S. In particular, we extend bulk universality, proved for generalized Wigner matrices in [17], to a large class of matrix ensembles where the upper and lower bounds on the variances (1.1) are relaxed.

The main motivation for the generalizations in this paper is the Anderson transition for band matrices outlined above. While not optimal, our results nevertheless imply that band matrices with a sufficiently broad band plus a negligible mean-field component exhibit bulk universality: their local spectral statistics are governed by random matrix statistics. For example, the local two-point correlation functions coincide if $W \gg N^{33/34}$. Although eigenvector delocalization and random matrix statistics are conjectured to occur in tandem, delocalization was actually proved in [8] under more general conditions than those under which we establish random matrix statistics. In fact, the delocalization results of [8] hold for a mean-field component as small as $(N/W^2)^{2/3}$, and, provided that $W \gg N^{4/5}$, the mean-field component may even vanish (resulting in a genuine band matrix).

The second purpose of this paper is to provide a coherent, pedagogical, and self-contained proof of the local semicircle law. In recent years, a series of papers [6, 12, 13, 17–19] with gradually weaker assumptions, was published on this topic. These papers often cited and relied on the previous ones. This made it difficult for the interested reader to follow all the details of the argument. The basic strategy of our proof (that is, using resolvents and large deviation bounds) was already used in [6, 12, 13, 17–19]. In this paper we not only streamline the argument for generalized Wigner matrices (satisfying (1.1)), but we also obtain sharper bounds for random matrices satisfying the much weaker condition (1.3). This allows us to establish universality results for a class of ensembles beyond generalized Wigner matrices.

Our proof is self-contained and simpler than those of [6, 17–19]. In particular, we give a proof of the *Fluctuation Averaging Theorem*, Theorems 4.6 and 4.7 below, which is considerably simpler than that of its predecessors in [6, 18, 19]. In addition, we consistently use fluctuation averaging at several key steps of the main argument, which allows us to shorten the proof and relax previous assumptions on the variances s_{ij}. The reader who is mainly interested in the pedagogical presentation should focus on the simplest choice of S, $s_{ij} = 1/N$, which corresponds to the standard Wigner matrix (for which $M = N$), and focus on Sections 2, 4, 5, and 6, as well as Appendix B.

We conclude this section with an outline of the paper. In Section 2 we define the model, introduce basic definitions, and state the local semicircle law in full generality (Theorem 2.3). Section 3 is devoted to some examples of random matrix models that satisfy our assumptions; for each example we give explicit bounds on the spectral domain on which the local semicircle law holds. Sections 4, 5, and 6 are devoted to the proof of the local semicircle law. Section 4 collects the basic tools that will be used throughout the proof. The purpose of Section 5 is mainly pedagogical; in it, we state and prove a weaker form of the local semicircle law, Theorem 5.1. The error bounds in Theorem 5.1 are identical to those of Theorem 2.3, but the spectral domain on which they hold is smaller. Provided one stays away from the spectral edge, Theorems 5.1 and 2.3 are equivalent; near the edge, Theorem 2.3 is stronger. The proof of Theorem 5.1 is very short and contains several key ideas from the proof of Theorem 2.3. The expert reader may therefore want to skip Section 5, but for the reader looking for a pedagogical presentation we recommend first focusing on Sections 4 and 5 (along with Appendix B). The full proof of our main result, Theorem 2.3, is given in Section 6. In Sections

7 and 8 we draw consequences from Theorem 2.3. In Section 7 we derive estimates on the density of states and the rigidity of the eigenvalue locations. In Section 8 we state and prove the universality of the local spectral statistics in the bulk, and give applications to some concrete matrix models. In Appendix A we derive explicit bounds on relevant norms of the resolvent of S (denoted by the abstract control parameters $\widetilde{\Gamma}$ and Γ), which are used to define the domains of applicability of Theorems 2.3 and 5.1. Finally, Appendix B is devoted to the proof of the fluctuation averaging estimates, Theorems 4.6 and 4.7.

We use C to denote a generic large positive constant, which may depend on some fixed parameters and whose value may change from one expression to the next. Similarly, we use c to denote a generic small positive constant.

2. Definitions and the main result

Let $(h_{ij} : i \leqslant j)$ be a family of independent, complex-valued random variables $h_{ij} \equiv h_{ij}^{(N)}$ satisfying $\mathbb{E}h_{ij} = 0$ and $h_{ii} \in \mathbb{R}$ for all i. For $i > j$ we define $h_{ij} := \bar{h}_{ji}$, and denote by $H \equiv H_N = (h_{ij})_{i,j=1}^N$ the $N \times N$ matrix with entries h_{ij}. By definition, H is Hermitian: $H = H^*$. We stress that all our results hold not only for complex Hermitian matrices but also for real symmetric matrices. In fact, the symmetry class of H plays no role, and our results apply for instance in the case where some off-diagonal entries of H are real and some complex-valued. (In contrast to some other papers in the literature, in our terminology the concept of Hermitian simply refers to the fact that $H = H^*$.)

We define

$$s_{ij} \;:=\; \mathbb{E}|h_{ij}|^2\,, \qquad M \;\equiv\; M_N \;:=\; \frac{1}{\max_{i,j} s_{ij}}\,. \tag{2.1}$$

In particular, we have the bound

$$s_{ij} \;\leqslant\; M^{-1} \tag{2.2}$$

for all i and j. We regard N as the fundamental parameter of our model, and M as a function of N. We introduce the $N \times N$ symmetric matrix $S \equiv S_N = (s_{ij})_{i,j=1}^N$. We assume that S is (doubly) stochastic:

$$\sum_j s_{ij} \;=\; 1 \tag{2.3}$$

for all i. For simplicity, we assume that S is irreducible, so that 1 is a simple eigenvalue. (The case of non-irreducible S may be trivially dealt with by considering its irreducible components separately.) We shall always assume the bounds

$$N^\delta \;\leqslant\; M \;\leqslant\; N \tag{2.4}$$

for some fixed $\delta > 0$.

It is sometimes convenient to use the normalized entries

$$\zeta_{ij} \;:=\; (s_{ij})^{-1/2} h_{ij}\,, \tag{2.5}$$

which satisfy $\mathbb{E}\zeta_{ij} = 0$ and $\mathbb{E}|\zeta_{ij}|^2 = 1$. (If $s_{ij} = 0$ we set for convenience ζ_{ij} to be a normalized Gaussian, so that these relations continue hold. Of course in this case the law of ζ_{ij} is immaterial.) We assume that the random variables ζ_{ij} have finite moments, uniformly in N, i, and j, in the sense that for all $p \in \mathbb{N}$ there is a constant μ_p such that

$$\mathbb{E}|\zeta_{ij}|^p \;\leqslant\; \mu_p \tag{2.6}$$

for all N, i, and j. We make this assumption to streamline notation in the statements of results such as Theorem 2.3 and the proofs. In fact, our results (and our proof) also cover the case where (2.6) holds for some finite large p; see Remark 2.4.

Throughout the following we use a spectral parameter $z \in \mathbb{C}$ satisfying $\operatorname{Im} z > 0$. We use the notation

$$z \;=\; E + \mathrm{i}\eta$$

without further comment, and always assume that $\eta > 0$. Wigner semicircle law ϱ and its Stieltjes transform m are defined by

$$\varrho(x) \;:=\; \frac{1}{2\pi}\sqrt{(4-x^2)_+}\,, \qquad m(z) \;:=\; \frac{1}{2\pi}\int_{-2}^{2} \frac{\sqrt{4-x^2}}{x-z}\,\mathrm{d}x\,. \tag{2.7}$$

To avoid confusion, we remark that m was denoted by m_{sc} in the papers [6,7,12–15,17–19], in which m had a different meaning from (2.7). It is well known that the Stieltjes transform m is the unique solution of

$$m(z) + \frac{1}{m(z)} + z \;=\; 0 \tag{2.8}$$

satisfying $\operatorname{Im} m(z) > 0$ for $\operatorname{Im} z > 0$. Thus we have

$$m(z) \;=\; \frac{-z + \sqrt{z^2-4}}{2}. \tag{2.9}$$

Some basic estimates on m are collected in Lemma 4.3 below.

An important parameter of the model is[1]

$$\Gamma_N(z) \;\equiv\; \Gamma(z) \;:=\; \left\| \left(1 - m(z)^2 S\right)^{-1} \right\|_{\ell^\infty \to \ell^\infty}. \tag{2.10}$$

A related quantity is obtained by restricting the operator $\left(1 - m(z)^2 S\right)^{-1}$ to the subspace $\mathbf{e}^\perp$ orthogonal to the constant vector $\mathbf{e} := N^{-1/2}(1, 1, \ldots, 1)^*$. Since S is stochastic, we have the estimate $-1 \leqslant S \leqslant 1$ and 1 is a simple eigenvalue of S with eigenvector $\mathbf{e}$. Set

$$\widetilde{\Gamma}_N(z) \;\equiv\; \widetilde{\Gamma}(z) \;:=\; \left\| \left(1 - m(z)^2 S\right)^{-1} \Big|_{\mathbf{e}^\perp} \right\|_{\ell^\infty \to \ell^\infty}, \tag{2.11}$$

the norm of $(1 - m(z)^2 S)^{-1}$ restricted to the subspace orthogonal to the constants. Clearly, $\widetilde{\Gamma}(z) \leqslant \Gamma(z)$. Basic estimates on Γ and $\widetilde{\Gamma}$ are collected in Proposition A.2 below. Many estimates in this paper depend critically on Γ and $\widetilde{\Gamma}$. Indeed, these parameters quantify the stability of certain self-consistent equations that underlie our proof. However, Γ and $\widetilde{\Gamma}$ remain bounded (up to a factor $\log N$) provided $E = \operatorname{Re} z$ is separated from the set $\{-2, 0, 2\}$; for band matrices (see Example 3.2) it suffices that E be separated from the spectral edges $\{-2, 2\}$; see Appendix A. At a first reading, we recommend that the reader neglect Γ and $\widetilde{\Gamma}$ (i.e. replace them with a constant). For band matrices, this amounts to focusing on the local semicircle law in the bulk of the spectrum.

We define the *resolvent* or *Green function* of H through

$$G(z) \;:=\; (H - z)^{-1}\,,$$

[1] Here we use the notation $\|A\|_{\ell^\infty \to \ell^\infty} = \max_i \sum_j |A_{ij}|$ for the operator norm on $\ell^\infty(\mathbb{C}^N)$.

and denote its entries by $G_{ij}(z)$. The Stieltjes transform of the empirical spectral measure of H is

$$m_N(z) \;:=\; \frac{1}{N}\operatorname{Tr} G(z)\,. \tag{2.12}$$

The following definition introduces a notion of a high-probability bound that is suited for our purposes. It was introduced (in a slightly different form) in [9].

Definition 2.1 (Stochastic domination). *Let*

$$X = \big(X^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}\big)\,, \qquad Y = \big(Y^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}\big)$$

be two families of nonnegative random variables, where $U^{(N)}$ is a possibly N-dependent parameter set. We say that X is stochastically dominated by Y, uniformly in u, *if for all (small) $\varepsilon > 0$ and (large) $D > 0$ we have*

$$\sup_{u \in U^{(N)}} \mathbb{P}\Big[X^{(N)}(u) > N^{\varepsilon} Y^{(N)}(u)\Big] \;\leqslant\; N^{-D}$$

for large enough $N \geqslant N_0(\varepsilon, D)$. Unless stated otherwise, throughout this paper the stochastic domination will always be uniform in all parameters apart from the parameter δ in (2.4) *and the sequence of constants μ_p in* (2.6)*; thus, $N_0(\varepsilon, D)$ also depends on δ and μ_p. If X is stochastically dominated by Y, uniformly in u, we use the notation $X \prec Y$. Moreover, if for some complex family X we have $|X| \prec Y$ we also write $X = O_{\prec}(Y)$.*

For example, using Chebyshev's inequality and (2.6) one easily finds that

$$|h_{ij}| \;\prec\; (s_{ij})^{1/2} \;\prec\; M^{-1/2}\,, \tag{2.13}$$

so that we may also write $h_{ij} = O_{\prec}((s_{ij})^{1/2})$. Another simple, but useful, example is a family of events $\Xi \equiv \Xi^{(N)}$ with asymptotically very high probability: If $\mathbb{P}(\Xi^c) \leqslant N^{-D}$ for any $D > 0$ and $N \geqslant N_0(D)$, then the indicator function $\mathbf{1}(\Xi)$ of Ξ satisfies $1 - \mathbf{1}(\Xi) \prec 0$.

The relation $\prec$ is a partial ordering, i.e. it is transitive and it satisfies the familiar arithmetic rules of order relations. For instance if $X_1 \prec Y_1$ and $X_2 \prec Y_2$ then $X_1 + X_2 \prec Y_1 + Y_2$ and $X_1 X_2 \prec Y_1 Y_2$. More general statements in this spirit are given in Lemma 4.4 below.

Definition 2.2 (Spectral domain). *We call an N-dependent family*

$$\mathbf{D} \;\equiv\; \mathbf{D}^{(N)} \;\subset\; \big\{z : |E| \leqslant 10\,,\, M^{-1} \leqslant \eta \leqslant 10\big\}$$

a spectral domain. *(Recall that $M \equiv M_N$ depends on N.)*

In this paper we always consider families $X^{(N)}(u) = X_i^{(N)}(z)$ indexed by $u = (z, i)$, where z takes on values in some spectral domain $\mathbf{D}$, and i takes on values in some finite (possibly N-dependent or empty) index set. The stochastic domination $X \prec Y$ of such families will always be uniform in z and i, and we usually do not state this explicitly. Usually, which spectral domain $\mathbf{D}$ is meant will be clear from the context, in which case we shall not mention it explicitly.

In this paper we shall make use of two spectral domains, $\mathbf{S}$ defined in (5.2) and $\widetilde{\mathbf{S}}$ defined in (2.17). Our main result is formulated on the larger of these domains, $\widetilde{\mathbf{S}}$. In order to define it, we introduce an

E-dependent lower boundary $\widetilde{\eta}_E$ on the spectral domain. We choose a (small) positive constant γ, and define for each $E \in [-10, 10]$

$$\widetilde{\eta}_E \;:=\; \min\left\{ \eta \;:\; \frac{1}{M\eta} \leqslant \min\left\{ \frac{M^{-\gamma}}{\widetilde{\Gamma}(z)^3} \,,\, \frac{M^{-2\gamma}}{\widetilde{\Gamma}(z)^4 \operatorname{Im} m(z)} \right\} \text{ for all } z \in [E + \mathrm{i}\eta, E + 10\mathrm{i}] \right\}. \tag{2.14}$$

Note that $\widetilde{\eta}_E$ depends on γ, but we do not explicitly indicate this dependence since we regard γ as fixed. At a first reading we advise the reader to think of γ as being zero. Note also that the lower bound in (A.3) below implies that $\widetilde{\eta}_E \geqslant M^{-1}$. We also define the distance to the spectral edge,

$$\kappa \;\equiv\; \kappa_E \;:=\; \big||E| - 2\big| \,. \tag{2.15}$$

Finally, we introduce the fundamental control parameter

$$\Pi(z) \;:=\; \sqrt{\frac{\operatorname{Im} m(z)}{M\eta}} + \frac{1}{M\eta} \,, \tag{2.16}$$

which will be used throughout this paper as a sharp, deterministic upper bound on the entries of G. Note that the condition in the definition of $\widetilde{\eta}_E$ states that the first term of Π is bounded by $M^{-\gamma}\widetilde{\Gamma}^{-2}$ and the second term by $M^{-\gamma}\widetilde{\Gamma}^{-3}$. We may now state our main result.

THEOREM 2.3 (LOCAL SEMICIRCLE LAW). *Fix $\gamma \in (0, 1/2)$ and define the spectral domain*

$$\widetilde{\mathbf{S}} \;\equiv\; \widetilde{\mathbf{S}}^{(N)}(\gamma) \;:=\; \big\{ E + \mathrm{i}\eta : |E| \leqslant 10 \,,\, \widetilde{\eta}_E \leqslant \eta \leqslant 10 \big\} \,. \tag{2.17}$$

We have the bounds

$$\max_{i,j} \big| G_{ij}(z) - \delta_{ij} m(z) \big| \;\prec\; \Pi(z) \tag{2.18}$$

uniformly in $z \in \widetilde{\mathbf{S}}$, as well as

$$\big| m_N(z) - m(z) \big| \;\prec\; \frac{1}{M\eta} \tag{2.19}$$

uniformly in $z \in \widetilde{\mathbf{S}}$. Moreover, outside of the spectrum we have the stronger estimate

$$\big| m_N(z) - m(z) \big| \;\prec\; \frac{1}{M(\kappa + \eta)} + \frac{1}{(M\eta)^2 \sqrt{\kappa + \eta}} \tag{2.20}$$

uniformly in $z \in \widetilde{\mathbf{S}} \cap \{ z : |E| \geqslant 2 \,,\, M\eta\sqrt{\kappa + \eta} \geqslant M^{\gamma} \}$.

We remark that the main estimate for the Stieltjes transform m_N is (2.19). The other estimate (2.20) is mainly useful for controlling the norm of H, which we do in Section 7. We also recall that uniformity for the spectral parameter z means that the threshold $N_0(\varepsilon, D)$ in the definition of $\prec$ is independent of the choice of z within the indicated spectral domain. As stated in Definition 2.1, this uniformity holds for all statements containing $\prec$, and is not explicitly mentioned in the following; all of our arguments are trivially uniform in z and any matrix indices.

REMARK 2.4. Theorem 2.3 has the following variant for matrix entries where the condition (2.6) is only imposed for some large but fixed p. More precisely, for any $\varepsilon > 0$ and $D > 0$ there exists a constant $p(\varepsilon, D)$ such that if (2.6) holds for $p = p(\varepsilon, D)$ then

$$\mathbb{P}\big(|m_N(z) - m(z)| > N^{\varepsilon}(M\eta)^{-1}\big) \;\leqslant\; N^{-D}$$

for all $z \in \widetilde{\mathbf{S}}$ and $N \geqslant N_0(\varepsilon, D)$. An analogous estimate replaces (2.18) and (2.20). The proof of this variant is the same as that of Theorem 2.3.

REMARK 2.5. Most of the previous works [6,12,13,17–19] assumed a stronger, subexponential decay condition on ζ_{ij} instead of (2.6). Under the subexponential decay condition, certain probability estimates in the results were somewhat stronger and precise tolerance thresholds were sharper. Roughly, this corresponds to operating with a modified definition of $\prec$, where the factors N^{ε} are replaced by high powers of $\log N$ and the polynomial probability bound N^{-D} is replaced with a subexponential one. The proofs of the current paper can be easily adjusted to such a setup, but we shall not pursue this further.

A local semicircle law for Wigner matrices on the optimal scale $\eta \gtrsim 1/N$ was first obtained in [13]. The optimal error estimates in the bulk were proved in [18], and extended to the edges in [19]. These estimates underlie the derivation of rigidity estimates for individual eigenvalues, which in turn were used in [19] to prove Dyson's conjecture on the optimal local relaxation time for the Dyson Brownian motion.

Apart from the somewhat different assumption on the tails of the entries of H (see Remark 2.5), Theorem 2.3, when restricted to generalized Wigner matrices, subsumes all previous local semicircle laws obtained in [12, 13, 18, 19]. For band matrices, a local semicircle law was proved in [17]. (In fact, in [17] the band structure was not required; only the conditions (2.2), (2.3), and the subexponential decay condition for the matrix entries (instead of (2.6)) were used.) Theorem 2.3 improves this result in several ways. First, the error bounds in (2.18) and (2.19) are uniform in E, even for E near the spectral edge; the corresponding bounds in Theorem 2.1 of [17] diverged as κ^{-1}. Second, the bound (2.19) on the Stieltjes transform is better than (2.16) in [17] by a factor $(M\eta)^{-1/2}$. This improvement is due to exploiting the fluctuation averaging mechanism of Theorem 4.6. Third, the domain of η for which Theorem 2.3 applies is essentially $\eta \gg \kappa^{-7/2}M^{-1}$, which is somewhat larger than the domain $\eta \gg \kappa^{-4}M^{-1}$ of [17].

While Theorem 2.3 subsumes several previous local semicircle laws, two previous results are not covered. The local semicircle law for sparse matrices proved in [6] does not follow from Theorem 2.3. However, the argument of this paper may be modified so as to include sparse matrices as well; we do not pursue this issue further. The local semicircle law for one-dimensional band matrices given in Theorem 2.2 of [8] is, however, of a very different nature, and may not be recovered using the methods of the current paper. Under the conditions $W \gg N^{4/5}$ and $\eta \gg N^2/W^3$, Theorem 2.2 of [8] shows that (focusing for simplicity on the one-dimensional case)

$$\big|G_{ij}(z) - \delta_{ij}m(z)\big| \;\prec\; \frac{1}{(N\eta)^{1/2}} + \frac{1}{(W\sqrt{\eta})^{1/2}} \tag{2.21}$$

in the bulk spectrum, which is stronger than the bound of order $(W\eta)^{-1/2}$ in (2.18). The proof of (2.21) relies on a very general fluctuation averaging result from [9], which is considerably stronger than Theorems 4.6 and 4.7; see Remark 4.8 below. The key open problem for band matrices is to establish a local semicircle law on a scale η below W^{-1}. The estimate (2.21) suggests that the resolvent entries should remain bounded throughout the range $\eta \gtrsim \max\{N^{-1}, W^{-2}\}$.

The local semicircle law, Theorem 2.3, has numerous consequences, several of which are formulated in Sections 7 and 8. Here we only sketch them. Theorem 7.5 states that the empirical counting function

converges to the counting function of the semicircle law. The precision is of order M^{-1} provided that we have the lower bound $s_{ij} \geqslant c/N$ for some constant $c > 0$. As a consequence, Theorem 7.6 states that the bulk eigenvalues are rigid on scales of order M^{-1}. Under the same condition, in Theorem 8.2 we prove the universality of the local two-point correlation functions in the bulk provided that $M \gg N^{33/34}$; we obtain similar results for higher order correlation functions, assuming a stronger restriction on M. These results generalize the earlier theorems from [6,7,19], which were valid for generalized Wigner matrices satisfying the condition (1.1), under which M is comparable to N. We obtain similar results if the condition $s_{ij} \geqslant c/N$ in (1.1) is relaxed to $s_{ij} \geqslant N^{-1-\xi}$ with some small ξ. The exponent ξ can be chosen near 1 for band matrices with a broad band $W \asymp N$. In particular, we prove universality for such band matrices with a rapidly vanishing mean-field component. These applications of the general Theorem 8.2 are listed in Corollary 8.3.

3. Examples

In this section we give some important example of random matrix models H. In each of the examples, we give the deterministic matrix $S = (s_{ij})$ of the variances of the entries of H. The matrix H is then obtained from $h_{ij} = s_{ij}\zeta_{ij}$. Here (ζ_{ij}) is a Hermitian matrix whose upper-triangular entries are independent and whose diagonal entries are real; moreover, we have $\mathbb{E}\zeta_{ij} = 0$, $\mathbb{E}|\zeta_{ij}|^2 = 1$, and the condition (2.6) for all p, uniformly in N, i, and j.

Definition 3.1 (Full and flat Wigner matrices). *Let $a \equiv a_N$ and $b \equiv b_N$ be possibly N-dependent positive quantities. We call H an* a-full Wigner matrix *if S satisfies* (2.3) *and*

$$s_{ij} \;\geqslant\; \frac{a}{N}\,. \tag{3.1}$$

Similarly, we call H a b-flat Wigner matrix *if S satisfies* (2.3) *and*

$$s_{ij} \;\leqslant\; \frac{b}{N}\,.$$

(Note that in this case we have $M \geqslant N/b$.)

If a and b are independent of N we call an a-full Wigner matrix simply full *and a b-flat Wigner matrix simply* flat. *In particular, generalized Wigner matrices, satisfying* (1.1)*, are full and flat Wigner matrices.*

Definition 3.2 (Band matrix). *Fix $d \in \mathbb{N}$. Let f be a bounded and symmetric (i.e. $f(x) = f(-x)$) probability density on $\mathbb{R}^d$. Let L and W be integers satisfying*

$$L^{\delta'} \;\leqslant\; W \;\leqslant\; L$$

for some fixed $\delta' > 0$. Define the d-dimensional discrete torus

$$\mathbb{T}_L^d \;=\; [-L/2, L/2)^d \cap \mathbb{Z}^d\,.$$

Thus, $\mathbb{T}_L^d$ has $N = L^d$ lattice points; and we may identify $\mathbb{T}_L^d$ with $\{1, \ldots, N\}$. We define the canonical representative of $i \in \mathbb{Z}^d$ through

$$[i]_L \;:=\; (i + L\mathbb{Z}^d) \cap \mathbb{T}_L^d\,.$$

Then H is a d-dimensional band matrix *with band width W and profile function f if*

$$s_{ij} \;=\; \frac{1}{Z_L}\, f\biggl(\frac{[i-j]_L}{W}\biggr)\,,$$

where Z_L is a normalization chosen so that (2.3) *holds.*

Definition 3.3 (Band matrix with a mean-field component). *Let H_B a d-dimensional band matrix from Definition 3.2. Let H_W be an independent a-full Wigner matrix indexed by the set $\mathbb{T}_L^d$. The matrix $H := \sqrt{1-\nu}H_B + \sqrt{\nu}H_W$, with some $\nu \in [0,1]$, is called a* band matrix with a mean-field component.

The example of Definition 3.3 is a mixture of the previous two. We are especially interested in the case $\nu \ll 1$, when most of the variance comes from the band matrix, i.e. the profile of S is very close to a sharp band.

We conclude with some explicit bounds for these examples. The behaviour of Γ and $\widetilde{\Gamma}$ near the spectral edge is governed by the parameter

$$\theta \;\equiv\; \theta(z) \;:=\; \begin{cases} \kappa + \frac{\eta}{\sqrt{\kappa+\eta}} & \text{if } |E| \leqslant 2 \\ \sqrt{\kappa+\eta} & \text{if } |E| > 2\,, \end{cases} \tag{3.2}$$

where we set, as usual, $\kappa \equiv \kappa_E$ and $z = E + i\eta$. Note that the parameter θ may be bounded from below by $(\operatorname{Im} m)^2$. The following results follow immediately from Propositions A.2 and A.3 in Appendix A. They hold for an arbitrary spectral domain $\mathbf{D}$.

(i) For general H and any constant $c > 0$, there is a constant $C > 0$ such that

$$C^{-1} \;\leqslant\; \widetilde{\Gamma} \;\leqslant\; \Gamma \;\leqslant\; C \log N$$

provided $\operatorname{dist}(E, \{-2, 0, 2\}) \geqslant c$.

(ii) For a full Wigner matrix we have

$$c \;\leqslant\; \widetilde{\Gamma} \;\leqslant\; C \log N\,, \qquad \frac{c}{\sqrt{\kappa+\eta}} \;\leqslant\; \Gamma \;\leqslant\; \frac{C \log N}{\theta}\,,$$

where C depends on the constant a in Definition 3.1 but c does not.

(iii) For a band matrix with a mean-field component, as in Definition 3.3, we have

$$c \;\leqslant\; \widetilde{\Gamma} \;\leqslant\; \frac{C \log N}{(W/L)^2 + \nu a + \theta}\,.$$

The case $\nu = 0$ corresponds to a band matrix from Definition 3.2.

4. Tools

In this subsection we collect some basic facts that will be used throughout the paper. For two positive quantities A_N and B_N we use the notation $A_N \asymp B_N$ to mean $cA_N \leqslant B_N \leqslant CA_N$. Throughout the following we shall frequently drop the arguments z and N, bearing in mind that we are dealing with a function on some spectral domain $\mathbf{D}$.

Definition 4.1 (Minors). *For $\mathbb{T} \subset \{1, \ldots, N\}$ we define $H^{(\mathbb{T})}$ by*

$$(H^{(\mathbb{T})})_{ij} \;:=\; \mathbf{1}(i \notin \mathbb{T})\mathbf{1}(j \notin \mathbb{T})h_{ij}\,.$$

Moreover, we define the resolvent of $H^{(\mathbb{T})}$ *through*

$$G_{ij}^{(\mathbb{T})}(z) \;:=\; \big(H^{(\mathbb{T})} - z\big)^{-1}_{ij} .$$

We also set

$$\sum_{i}^{(\mathbb{T})} \;:=\; \sum_{i : i \notin \mathbb{T}} .$$

When $\mathbb{T} = \{a\}$*, we abbreviate* $(\{a\})$ *by* (a) *in the above definitions; similarly, we write* (ab) *instead of* $(\{a, b\})$*.*

Definition 4.2 (Partial expectation and independence). *Let* $X \equiv X(H)$ *be a random variable. For* $i \in \{1, \ldots, N\}$ *define the operations* P_i *and* Q_i *through*

$$P_i X \;:=\; \mathbb{E}(X|H^{(i)}) , \qquad Q_i X \;:=\; X - P_i X .$$

We call P_i partial expectation *in the index* i*. Moreover, we say that* X *is* independent of $\mathbb{T} \subset \{1, \ldots, N\}$ *if* $X = P_i X$ *for all* $i \in \mathbb{T}$*.*

We introduce the random z-dependent control parameters

$$\Lambda_o \;:=\; \max_{i \neq j} |G_{ij}| , \qquad \Lambda_d \;:=\; \max_{i} |G_{ii} - m| , \qquad \Lambda \;:=\; \max\{\Lambda_o, \Lambda_d\} , \qquad \Theta \;:=\; |m_N - m| . \tag{4.1}$$

We remark that the letter Λ had a different meaning in several earlier papers, such as [19]. The following lemma collects basic bounds on m.

Lemma 4.3. *There is a constant* $c > 0$ *such that for* $E \in [-10, 10]$ *and* $\eta \in (0, 10]$ *we have*

$$c \;\leqslant\; |m(z)| \;\leqslant\; 1 - c\eta , \tag{4.2}$$

$$|1 - m^2(z)| \;\asymp\; \sqrt{\kappa + \eta} , \tag{4.3}$$

as well as

$$\operatorname{Im} m(z) \;\asymp\; \begin{cases} \sqrt{\kappa + \eta} & \text{if } |E| \leqslant 2 \\ \frac{\eta}{\sqrt{\kappa + \eta}} & \text{if } |E| \geqslant 2 . \end{cases} \tag{4.4}$$

Proof. The proof is an elementary exercise using (2.9). □

In particular, recalling that $-1 \leqslant S \leqslant 1$ and using the upper bound $|m| \leqslant C$ from (4.2), we find that there is a constant $c > 0$ such that

$$c \;\leqslant\; \widetilde{\Gamma} \;\leqslant\; \Gamma . \tag{4.5}$$

The following lemma collects basic algebraic properties of stochastic domination $\prec$. Roughly, it states that $\prec$ satisfies the usual arithmetic properties of order relations. We shall use it tacitly throughout the following.

Lemma 4.4. *(i) Suppose that* $X(u, v) \prec Y(u, v)$ *uniformly in* $u \in U$ *and* $v \in V$*. If* $|V| \leqslant N^C$ *for some constant* C *then*

$$\sum_{v \in V} X(u, v) \;\prec\; \sum_{v \in V} Y(u, v)$$

uniformly in u*.*

(ii) Suppose that $X_1(u) \prec Y_1(u)$ uniformly in u and $X_2(u) \prec Y_2(u)$ uniformly in u. Then $X_1(u)X_2(u) \prec Y_1(u)Y_2(u)$ uniformly in u.

(iii) If $X \prec Y + N^{-\varepsilon}X$ for some $\varepsilon > 0$ then $X \prec Y$.

PROOF. The claims (i) and (ii) follow from a simple union bound. The claim (iii) is an immediate consequence of the definition of $\prec$. □

The following resolvent identities form the backbone of all of our calculations. The idea behind them is that a resolvent matrix element G_{ij} depends strongly on the i-th and j-th columns of H, but weakly on all other columns. The first identity determines how to make a resolvent matrix element G_{ij} independent of an additional index $k \neq i, j$. The second identity expresses the dependence of a resolvent matrix element G_{ij} on the matrix elements in the i-th or in the j-th column of H.

LEMMA 4.5 (RESOLVENT IDENTITIES). *For any Hermitian matrix H and $\mathbb{T} \subset \{1, \ldots, N\}$ the following identities hold. If $i, j, k \notin \mathbb{T}$ and $i, j \neq k$ then*

$$G_{ij}^{(\mathbb{T})} \;=\; G_{ij}^{(\mathbb{T}k)} + \frac{G_{ik}^{(\mathbb{T})} G_{kj}^{(\mathbb{T})}}{G_{kk}^{(\mathbb{T})}}\,, \qquad \frac{1}{G_{ii}^{(\mathbb{T})}} \;=\; \frac{1}{G_{ii}^{(\mathbb{T}k)}} - \frac{G_{ik}^{(\mathbb{T})} G_{ki}^{(\mathbb{T})}}{G_{ii}^{(\mathbb{T})} G_{ii}^{(\mathbb{T}k)} G_{kk}^{(\mathbb{T})}}\,. \tag{4.6}$$

If $i, j \notin \mathbb{T}$ satisfy $i \neq j$ then

$$G_{ij}^{(\mathbb{T})} \;=\; -G_{ii}^{(\mathbb{T})} \sum_k^{(\mathbb{T}i)} h_{ik} G_{kj}^{(\mathbb{T}i)} \;=\; -G_{jj}^{(\mathbb{T})} \sum_k^{(\mathbb{T}j)} G_{ik}^{(\mathbb{T}j)} h_{kj}\,. \tag{4.7}$$

PROOF. This is an exercise in linear algebra. The first identity (4.6) was proved in Lemma 4.2 of [17] and the second is an immediate consequence of the first. The identity (4.7) is proved in Lemma 6.10 of [7]. □

Our final tool consists of the following results on *fluctuation averaging*. They exploit cancellations in sums of fluctuating quantities involving resolvent matrix entries. A very general result was obtained in [9]; in this paper we state a special case sufficient for our purposes here, and give a relatively simple proof in Appendix B. We consider weighted averages of diagonal resolvent matrix entries G_{kk}. They are weakly dependent, but the correlation between G_{kk} and G_{mm} for $m \neq k$ is not sufficiently small to apply the general theory of sums of weakly dependent random variables; instead, we need to exploit the precise form of the dependence using the resolvent structure.

It turns out that the key quantity that controls the magnitude of the fluctuations is Λ. However, being a random variable, Λ itself is unsuitable as an upper bound. For technical reasons (our proof relies on a high-moment estimate combined with Chebyshev's inequality), it is essential that Λ be estimated by a *deterministic* control parameter, which we call Ψ. The error terms are then estimated in terms of powers of Ψ. We shall always assume that Ψ satisfies

$$M^{-1/2} \;\leqslant\; \Psi \;\leqslant\; M^{-c} \tag{4.8}$$

in the spectral domain $\mathbf{D}$, where $c > 0$ is some constant. We shall perform the averaging with respect to a family of complex weights $T = (t_{ik})$ satisfying

$$0 \;\leqslant\; |t_{ik}| \;\leqslant\; M^{-1}\,, \qquad \sum_k |t_{ik}| \;\leqslant\; 1\,. \tag{4.9}$$

Typical example weights are $t_{ik} = s_{ik}$ and $t_{ik} = N^{-1}$. Note that in both of these cases T commutes with S. We introduce the *average* of a vector $(a_i)_{i=1}^N$ through

$$[a] \;:=\; \frac{1}{N}\sum_i a_i\,. \tag{4.10}$$

THEOREM 4.6 (FLUCTUATION AVERAGING). *Fix a spectral domain* $\mathbf{D}$ *and a deterministic control parameter* Ψ *satisfying* (4.8). *Suppose that* $\Lambda \prec \Psi$ *and the weight* $T = (t_{ik})$ *satisfies* (4.9). *Then we have*

$$\sum_k t_{ik} Q_k \frac{1}{G_{kk}} \;=\; O_\prec(\Psi^2)\,, \qquad \sum_k t_{ik} Q_k G_{kk} \;=\; O_\prec(\Psi^2)\,. \tag{4.11}$$

If T *commutes with* S *then*

$$\sum_k t_{ik} v_k \;=\; O_\prec(\Gamma\Psi^2)\,. \tag{4.12}$$

Finally, if T *commutes with* S *and*

$$\sum_k t_{ik} \;=\; 1 \tag{4.13}$$

for all i *then*

$$\sum_k t_{ik}(v_k - [v]) \;=\; O_\prec(\widetilde{\Gamma}\Psi^2)\,, \tag{4.14}$$

where we defined $v_i := G_{ii} - m$. *The estimates* (4.11), (4.12), *and* (4.14) *are uniform in the index* i.

In fact, the first bound of (4.11) can be improved as follows.

THEOREM 4.7. *Fix a spectral domain* $\mathbf{D}$ *deterministic control parameters* Ψ *and* Ψ_o, *both satisfying* (4.8). *Suppose that* $\Lambda \prec \Psi$, $\Lambda_o \prec \Psi_o$, *and that the weight* $T = (t_{ik})$ *satisfies* (4.9). *Then*

$$\sum_k t_{ik} Q_k \frac{1}{G_{kk}} \;=\; O_\prec(\Psi_o^2)\,. \tag{4.15}$$

REMARK 4.8. The first instance of the fluctuation averaging mechanism appeared in [18] for the Wigner case, where $[Z] = N^{-1}\sum_k Z_k$ was proved to be bounded by Λ_o^2. Since $Q_k[G_{kk}]^{-1}$ is essentially Z_k (see (5.6) below), this corresponds to the first bound in (4.11). A different proof (with a better bound on the constants) was given in [19]. A conceptually streamlined version of the original proof was extended to sparse matrices [6] and to sample covariance matrices [26]. Finally, an extensive analysis in [9] treated the fluctuation averaging of general polynomials of resolvent entries and identified the order of cancellations depending on the algebraic structure of the polynomial. Moreover, in [9] an additional cancellation effect was found for the quantity $Q_i|G_{ij}|^2$. These improvements played a key role in obtaining the diffusion profile for the resolvent of band matrices and the estimate (2.21) in [8].

All proofs of the fluctuation averaging theorems rely on computing expectations of high moments of the averages, and carefully estimating the resulting terms. In [9], a diagrammatic representation was developed for bookkeeping such terms, but this is necessary only for the case of general polynomials. For the special cases given in Theorem 4.6, the proof is relatively simple and it is presented in Appendix B. Compared with [6, 18, 19], the algebra of the decoupling of the randomness is greatly simplified in the current paper. Moreover, unlike their counterparts from [6, 18, 19], the fluctuation averaging results of Theorems 4.6 and 4.7 do not require conditioning on the complement of some "bad" low-probability event, because such events are automatically accounted for by the definition of $\prec$; this leads to further simplifications in the proofs of Theorems 4.6 and 4.7.

5. A simpler proof using Γ instead of $\widetilde{\Gamma}$

In this section we prove the following weaker version of Theorem 2.3. In analogy to (2.14), we introduce the lower boundary

$$\eta_E \;:=\; \min\left\{ \eta \;:\; \frac{1}{M\eta} \leqslant \min\left\{ \frac{M^{-\gamma}}{\Gamma(z)^3}\,,\; \frac{M^{-2\gamma}}{\Gamma(z)^4\,\mathrm{Im}\,m(z)} \right\} \text{ for all } z \in [E+\mathrm{i}\eta, E+10\mathrm{i}] \right\}. \tag{5.1}$$

THEOREM 5.1. *Fix $\gamma \in (0,1/2)$ and define the spectral domain*

$$\mathbf{S} \;\equiv\; \mathbf{S}^{(N)}(\gamma) \;:=\; \big\{ E+\mathrm{i}\eta : |E| \leqslant 10\,,\; \eta_E \leqslant \eta \leqslant 10 \big\}. \tag{5.2}$$

We have the bounds

$$\big|G_{ij}(z) - \delta_{ij} m(z)\big| \;\prec\; \Pi(z) \tag{5.3}$$

uniformly in i,j and $z \in \mathbf{S}$, as well as

$$\big|m_N(z) - m(z)\big| \;\prec\; \frac{1}{M\eta} \tag{5.4}$$

uniformly in $z \in \mathbf{S}$.

Note that the only difference between Theorems 2.3 and 5.1 is that $\widetilde{\Gamma}$ was replaced with the larger quantity Γ in the definition of the threshold η_E and the spectral domain, so that

$$\frac{1}{M} \;\leqslant\; \widetilde{\eta}_E \;\leqslant\; \eta_E\,, \qquad \mathbf{S} \;\subset\; \widetilde{\mathbf{S}}\,. \tag{5.5}$$

Hence Theorem 5.1 is indeed weaker than Theorem 2.3, since it holds on a smaller spectral domain. As outlined after (2.11) and discussed in detail in Appendix A, Theorems 5.1 and 2.3 are equivalent provided E is separated from the set $\{-2,0,2\}$ (for band matrices they are equivalent provided E is separated from the spectral edges ± 2).

The rest of this section is devoted to the proof of Theorem 5.1. We give the full proof of Theorem 5.1 for pedagogical reasons, since it is simpler than that of Theorem 2.3 but already contains several of its key ideas. Theorem 2.3 will be proved in Section 6. One big difference between the two proofs is that in Theorem 5.1 the main control parameter is Λ, while in Theorem 2.3 we have to keep track of two control parameters, Λ and the smaller Θ.

5.1. The self-consistent equation. The key tool behind the proof is a self-consistent equation for the diagonal entries of G. The starting point is Schur's complement formula, which we write as

$$\frac{1}{G_{ii}} \;=\; h_{ii} - z - \sum_{k,l}^{(i)} h_{ik} G^{(i)}_{kl} h_{li}\,. \tag{5.6}$$

The partial expectation with respect to the index i (see Definition 4.2) of the last term on the right-hand side reads

$$P_i \sum_{k,l}^{(i)} h_{ik} G^{(i)}_{kl} h_{li} \;=\; \sum_{k}^{(i)} s_{ik} G^{(i)}_{kk} \;=\; \sum_{k}^{(i)} s_{ik} G_{kk} - \sum_{k}^{(i)} s_{ik} \frac{G_{ik}G_{ki}}{G_{ii}} \;=\; \sum_{k} s_{ik} G_{kk} - \sum_{k} s_{ik} \frac{G_{ik}G_{ki}}{G_{ii}}\,,$$

where in the first step we used (2.1) and in the second (4.6). Introducing the notation

$$v_i \;:=\; G_{ii} - m$$

and recalling (2.3), we therefore get from (5.6) that

$$\frac{1}{G_{ii}} \;=\; -z - m + \Upsilon_i - \sum_k s_{ik} v_k \,, \tag{5.7}$$

where we introduced the fluctuating error term

$$\Upsilon_i \;:=\; A_i + h_{ii} - Z_i \,, \qquad A_i \;:=\; \sum_k s_{ik} \frac{G_{ik} G_{ki}}{G_{ii}} \,, \qquad Z_i \;:=\; Q_i \sum_{k,l}^{(i)} h_{ik} G^{(i)}_{kl} h_{li} \,. \tag{5.8}$$

Using (2.8), we therefore get the *self-consistent equation*

$$-\sum_k s_{ik} v_k + \Upsilon_i \;=\; \frac{1}{m + v_i} - \frac{1}{m} \,. \tag{5.9}$$

Notice that this is an equation for the family $(v_i)_{i=1}^N$, with random error terms Υ_i.

Self-consistent equations play a crucial role in analysing resolvents of random matrices. The simplest one is the *scalar (or first level) self-consistent equation* for $m_N(z)$, the Stieltjes transform of the empirical density (2.12). By averaging the inverse of (5.7) and neglecting the error terms, one obtains that m_N approximately satisfies the equation $m = -(m+z)^{-1}$, which is the defining relation for the Stieltjes transform of the semicircle law (2.8).

The *vector (or second level) self-consistent equation*, as given in (5.9), allows one to control not only fluctuations of $m_N - m$ but also those of $G_{ii} - m$. The equation (5.9) first appeared in [17], where a systematic study of resolvent entries of random matrices was initiated.

For completeness, we mention that a *matrix (or third level) self-consistent equation* for local averages of $|G_{ij}|^2$, was introduced in [8]. This equation constitutes the backbone of the study of the diffusion profile of the resolvent entries of random band matrices.

5.2. Estimate of the error Υ_i in terms of Λ.

LEMMA 5.2. *The following statements hold for any spectral domain* $\mathbf{D}$. *Let* ϕ *be the indicator function of some (possibly* z*-dependent) event. If* $\phi\Lambda \prec M^{-c}$ *for some* $c > 0$ *then*

$$\phi\big(\Lambda_o + |Z_i| + |\Upsilon_i|\big) \;\prec\; \sqrt{\frac{\operatorname{Im} m + \Lambda}{M\eta}} \tag{5.10}$$

uniformly in $z \in \mathbf{D}$. *Moreover, for any fixed (*N*-independent)* $\eta > 0$ *we have*

$$\Lambda_o + |Z_i| + |\Upsilon_i| \;\prec\; M^{-1/2} \tag{5.11}$$

uniformly in $z \in \{w \in \mathbf{D} : \operatorname{Im} w = \eta\}$.

PROOF. We begin with the first statement. We shall often use the fact that, by the lower bound of (4.2) and the assumption $\phi\Lambda \prec M^{-c}$, we have

$$\phi/|G_{ii}| \;\prec\; 1\,. \tag{5.12}$$

First we estimate Z_i, which we split as

$$\phi|Z_i| \;\leqslant\; \phi\biggl|\sum_k^{(i)}\bigl(|h_{ik}|^2 - s_{ik}\bigr)G_{kk}^{(i)}\biggr| + \phi\biggl|\sum_{k\neq l}^{(i)} h_{ik}G_{kl}^{(i)}h_{li}\biggr|\,. \tag{5.13}$$

We estimate each term using the large deviation estimates from Theorem C.1, by conditioning on $G^{(i)}$ and using the fact that the family $(h_{ik})_{k=1}^N$ is independent of $G^{(i)}$. By (C.2), the first term of (5.13) is stochastically dominated by $\phi\bigl(\sum_k^{(i)} s_{ik}^2\bigl|G_{kk}^{(i)}\bigr|^2\bigr)^{1/2} \prec M^{-1/2}$, where we used the estimate (2.2) and $\phi\bigl|G_{kk}^{(i)}\bigr| \prec 1$, as follows from (4.6), (5.12), and the assumption $\phi\Lambda \prec M^{-c}$. For the second term of (5.13) we apply (C.4) with $a_{kl} = s_{ik}^{1/2}G_{kl}^{(i)}s_{li}^{1/2}$ and $X_k = \zeta_{ik}$ (see (2.5)). We find

$$\phi\sum_{k,l}^{(i)} s_{ik}\bigl|G_{kl}^{(i)}\bigr|^2 s_{li} \;\leqslant\; \phi\frac{1}{M}\sum_{k,l}^{(i)} s_{ik}\bigl|G_{kl}^{(i)}\bigr|^2 \;=\; \phi\frac{1}{M\eta}\sum_k^{(i)} s_{ik}\operatorname{Im}G_{kk}^{(i)} \;\prec\; \frac{\operatorname{Im}m+\Lambda}{M\eta}\,, \tag{5.14}$$

where the second step follows by spectral decomposition of $G^{(i)}$, and in the last step we used (4.6) and (5.12). Thus we get

$$\phi|Z_i| \;\prec\; \sqrt{\frac{\operatorname{Im}m+\Lambda}{M\eta}}\,, \tag{5.15}$$

where we absorbed the bound $M^{-1/2}$ on the first term of (5.13) into the right-hand side of (5.15), using $\operatorname{Im}m \geqslant \eta$ as follows from (4.4).

Next, we estimate Λ_o. We can iterate (4.7) once to get, for $i \neq j$,

$$G_{ij} \;=\; -G_{ii}\sum_k^{(i)} h_{ik}G_{kj}^{(i)} \;=\; -G_{ii}G_{jj}^{(i)}\biggl(h_{ij} - \sum_{k,l}^{(ij)} h_{ik}G_{kl}^{(ij)}h_{lj}\biggr)\,. \tag{5.16}$$

The term h_{ij} is trivially $O_\prec(M^{-1/2})$. In order to estimate the other term, we invoke (C.3) with $a_{kl} = s_{ik}^{1/2}G_{kl}^{(ij)}s_{lj}^{1/2}$, $X_k = \zeta_{ik}$, and $Y_l = \zeta_{lj}$. As in (5.14), we find

$$\phi\sum_{k,l}^{(ij)} s_{ik}\bigl|G_{kl}^{(ij)}\bigr|^2 s_{lj} \;\prec\; \frac{\operatorname{Im}m+\Lambda}{M\eta}\,.$$

Thus we find

$$\phi\Lambda_o \;\prec\; \sqrt{\frac{\operatorname{Im}m+\Lambda}{M\eta}}\,, \tag{5.17}$$

where we again absorbed the term $h_{ij} \prec M^{-1/2}$ into the right-hand side.

In order to estimate A_i and h_{ii} in the definition of Υ_i, we use (5.12) to estimate

$$\phi\bigl(|A_i| + |h_{ii}|\bigr) \;\prec\; \phi\Lambda_o^2 + M^{-1/2} \;\leqslant\; \phi\Lambda_o + C\sqrt{\frac{\operatorname{Im}m}{M\eta}} \;\prec\; \sqrt{\frac{\operatorname{Im}m+\Lambda}{M\eta}}\,,$$

where the second step follows from $\operatorname{Im} m \geqslash \eta$ (recall (4.4)). This completes the proof of (5.10).

The proof of (5.11) is almost identical to that of (5.10). The quantities $\big|G^{(i)}_{kk}\big|$ and $\big|G^{(ij)}_{kk}\big|$ are estimated by the trivial deterministic bound η^{-1}. We omit the details. □

5.3. A rough bound on Λ. The next step in the proof of Theorem 5.1 is to establish the following rough bound on Λ.

PROPOSITION 5.3. *We have $\Lambda \prec M^{-\gamma/3}\Gamma^{-1}$ uniformly in $\mathbf{S}$.*

The rest of this subsection is devoted to the proof of Proposition 5.3. The core of the proof is a *continuity argument.* Its basic idea is to establish a *gap* in the range of Λ of the form $\mathbf{1}(\Lambda \leqslant M^{-\gamma/4}\Gamma^{-1})\Lambda \prec M^{-\gamma/2}\Gamma^{-1}$ (Lemma 5.4 below). In other words, for all $z \in \mathbf{S}$, with high probability either $\Lambda \leqslant M^{-\gamma/2}\Gamma^{-1}$ or $\Lambda \geqslant M^{-\gamma/4}\Gamma^{-1}$. For z with a large imaginary part η, the estimate $\Lambda \leqslant M^{-\gamma/2}\Gamma^{-1}$ is easy to prove using a simple expansion (Lemma 5.5 below). Thus, for large η the parameter Λ is below the gap. Using the fact that Λ is continuous in z and hence cannot jump from one side of the gap to the other, we then conclude that with high probability Λ is below the gap for all $z \in \mathbf{S}$. See Figure 5.1 for an illustration of this argument.

LEMMA 5.4. *We have the bound*

$$\mathbf{1}\big(\Lambda \leqslant M^{-\gamma/4}\Gamma^{-1}\big)\Lambda \;\prec\; M^{-\gamma/2}\Gamma^{-1}$$

uniformly in $\mathbf{S}$.

PROOF. Set

$$\phi \;:=\; \mathbf{1}\big(\Lambda \leqslant M^{-\gamma/4}\Gamma^{-1}\big).$$

Then by definition we have $\phi\Lambda \leqslant M^{-\gamma/4}\Gamma^{-1} \leqslant CM^{-\gamma/4}$, where in the last step we used (4.5). Hence we may invoke (5.10) to estimate Λ_o and Υ_i. In order to estimate Λ_d, we expand the right-hand side of (5.9) in v_i to get

$$\phi\bigg(-\sum_k s_{ik}v_k + \Upsilon_i\bigg) \;=\; \phi\big(-m^{-2}v_i + O(\Lambda^2)\big)\,,$$

where we used (4.2) and that $|v_i| \leqslant CM^{-\gamma/4}$ on the event $\{\phi = 1\}$. Using (5.10) we therefore have

$$\phi\bigg(v_i - m^2\sum_k s_{ik}v_k\bigg) \;=\; O_\prec\bigg(\Lambda^2 + \sqrt{\frac{\operatorname{Im} m + \Lambda}{M\eta}}\bigg)\,.$$

We write the left-hand side as $\phi[(1 - m^2S)\mathbf{v}]_i$ with the vector $\mathbf{v} = (v_i)_{i=1}^N$. Inverting the operator $1 - m^2S$, we therefore conclude that

$$\phi\Lambda_d \;=\; \phi\max_i|v_i| \;\prec\; \Gamma\bigg(\Lambda^2 + \sqrt{\frac{\operatorname{Im} m + \Lambda}{M\eta}}\bigg)\,.$$

Recalling (4.5) and (5.10), we therefore get

$$\phi\Lambda \;\prec\; \phi\Gamma\bigg(\Lambda^2 + \sqrt{\frac{\operatorname{Im} m + \Lambda}{M\eta}}\bigg)\,. \tag{5.18}$$

Next, by definition of ϕ we may estimate

$$\phi\Gamma\Lambda^2 \;\leqslant\; M^{-\gamma/2}\Gamma^{-1}\,.$$

Moreover, by definitions of $\mathbf{S}$ and ϕ we have

$$\phi\Gamma\sqrt{\frac{\operatorname{Im} m+\Lambda}{M\eta}} \;\leqslant\; \Gamma\sqrt{\frac{\operatorname{Im} m}{M\eta}}+\Gamma\sqrt{\frac{\Gamma^{-1}}{M\eta}} \;\leqslant\; M^{-\gamma}\Gamma^{-1}+M^{-\gamma/2}\Gamma^{-1} \;\leqslant\; 2M^{-\gamma/2}\Gamma^{-1}\,.$$

Plugging this into (5.18) yields $\phi\Lambda \prec M^{-\gamma/2}\Gamma^{-1}$, which is the claim. □

In order to start the continuity argument underlying the proof of Proposition 5.3, we need the following bound on Λ for large η.

LEMMA 5.5. *We have $\Lambda \prec M^{-1/2}$ uniformly in $z \in [-10,10]+2\mathrm{i}$.*

PROOF. We shall make use of the trivial bounds

$$\big|G^{(\mathbb{T})}_{ij}\big| \;\leqslant\; \frac{1}{\eta} \;=\; \frac{1}{2}\,, \qquad |m| \;\leqslant\; \frac{1}{\eta} \;=\; \frac{1}{2}\,. \tag{5.19}$$

From (5.11) we get

$$\Lambda_o+|Z_i| \;\prec\; M^{-1/2}\,. \tag{5.20}$$

Moreover, we use (4.6) and (5.16) to estimate

$$|A_i| \;\leqslant\; \sum_j s_{ij}\left|\frac{G_{ij}G_{ji}}{G_{ii}}\right| \;\leqslant\; M^{-1}+\sum_j^{(i)} s_{ij}|G_{ji}G^{(i)}_{jj}|\left|h_{ij}-\sum_{k,l}^{(ij)} h_{ik}G^{(ij)}_{kl}h_{lj}\right| \;\prec\; M^{-1/2}\,,$$

where the last step follows using (C.3), exactly as the estimate of the right-hand side of (5.16) in the proof of Lemma 5.2. We conclude that $|\Upsilon_i| \prec M^{-1/2}$.

Next, we write (5.9) as

$$v_i \;=\; \frac{m\big(\sum_k s_{ik}v_k-\Upsilon_i\big)}{\big(m^{-1}-\sum_k s_{ik}v_k+\Upsilon_i\big)}\,.$$

Using $|m^{-1}| \geqslant 2$ and $|v_k| \leqslant 1$ as follows from (5.19), we find

$$\left|m^{-1}+\sum_k s_{ik}v_k-\Upsilon_i\right| \;\geqslant\; 1+O_\prec(M^{-1/2})\,.$$

Using $|m| \leqslant 1/2$ we therefore conclude that

$$\Lambda_d \;\leqslant\; \frac{\Lambda_d+O_\prec(M^{-1/2})}{2+O_\prec(M^{-1/2})} \;=\; \frac{\Lambda_d}{2}+O_\prec(M^{-1/2})\,,$$

from which the claim follows together with the estimate on Λ_o from (5.20). □

We may now conclude the proof of Proposition 5.3 by a continuity argument in $\eta = \operatorname{Im} z$. The gist of the continuity argument is depicted in Figure 5.1.

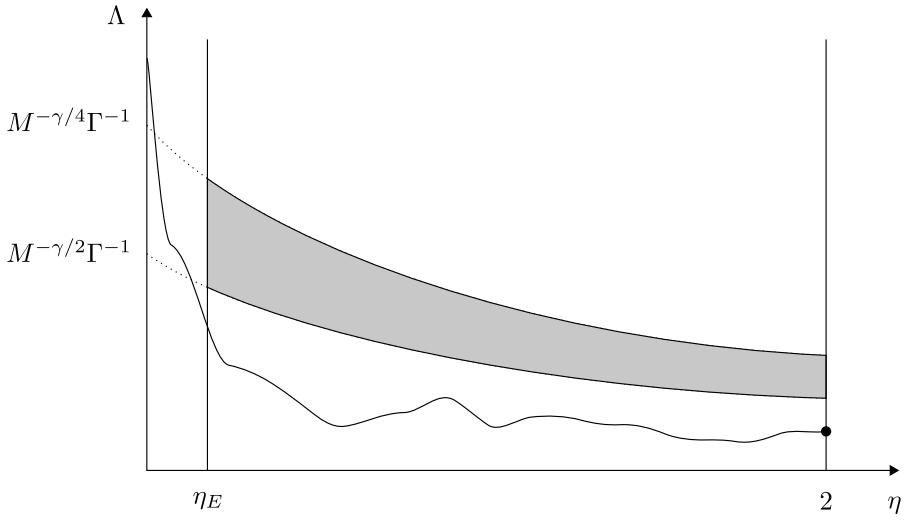

FIGURE 5.1. The (η, Λ)-plane for a fixed E. The shaded region is forbidden with high probability by Lemma 5.4. The initial estimate, given by Lemma 5.5, is marked with a black dot. The graph of $\Lambda = \Lambda(E + i\eta)$ is continuous and lies beneath the shaded region. Note that this method does not control $\Lambda(E + i\eta)$ in the regime $\eta \leqslant \eta_E$.

PROOF OF PROPOSITION 5.3. Fix $D > 10$. Lemma 5.4 implies that for each $z \in \mathbf{S}$ we have

$$\mathbb{P}\Big(M^{-\gamma/3}\Gamma(z)^{-1} \leqslant \Lambda(z) \leqslant M^{-\gamma/4}\Gamma(z)^{-1}\Big) \;\leqslant\; N^{-D} \tag{5.21}$$

for $N \geqslant N_0$, where $N_0 \equiv N_0(\gamma, D)$ does not depend on z.

Next, take a lattice $\Delta \subset \mathbf{S}$ such that $|\Delta| \leqslant N^{10}$ and for each $z \in \mathbf{S}$ there exists a $w \in \Delta$ such that $|z - w| \leqslant N^{-4}$. Then (5.21) combined with a union bounds gives

$$\mathbb{P}\Big(\exists w \in \Delta \,:\, M^{-\gamma/3}\Gamma(w)^{-1} \leqslant \Lambda(w) \leqslant M^{-\gamma/4}\Gamma(w)^{-1}\Big) \;\leqslant\; N^{-D+10} \tag{5.22}$$

for $N \geqslant N_0$. From the definitions of $\Lambda(z)$, $\Gamma(z)$, and $\mathbf{S}$ (recall (4.5)), we immediately find that Λ and Γ are Lipschitz continuous on $\mathbf{S}$, with Lipschitz constant at most M^2. Hence (5.22) implies

$$\mathbb{P}\Big(\exists z \in \mathbf{S} \,:\, 2M^{-\gamma/3}\Gamma(z)^{-1} \leqslant \Lambda(z) \leqslant 2^{-1}M^{-\gamma/4}\Gamma(z)^{-1}\Big) \;\leqslant\; N^{-D+10}$$

for $N \geqslant N_0$. We conclude that there is an event Ξ satisfying $\mathbb{P}(\Xi) \geqslant 1 - N^{-D+10}$ such that, for each $z \in \mathbf{S}$, either $\mathbf{1}(\Xi)\Lambda(z) \leqslant 2M^{-\gamma/3}\Gamma(z)^{-1}$ or $\mathbf{1}(\Xi)\Lambda(z) \geqslant 2^{-1}M^{-\gamma/4}\Gamma(z)^{-1}$. Since Λ is continuous and $\mathbf{S}$ is by definition connected, we conclude that either

$$\forall z \in \mathbf{S} \,:\, \mathbf{1}(\Xi)\Lambda(z) \;\leqslant\; 2M^{-\gamma/3}\Gamma(z)^{-1} \tag{5.23}$$

or

$$\forall z \in \mathbf{S} \,:\, \mathbf{1}(\Xi)\Lambda(z) \;\geqslant\; 2^{-1}M^{-\gamma/4}\Gamma(z)^{-1}\,. \tag{5.24}$$

(Here the bounds (5.23) and (5.24) each hold surely, i.e. for every realization of $\Lambda(z)$.)

It remains to show that (5.24) is impossible. In order to do so, it suffices to show that there exists a $z \in \mathbf{S}$ such that $\Lambda(z) < 2^{-1} M^{-\gamma/4} \Gamma(z)^{-1}$ with probability greater than $1/2$. But this holds for any z with $\operatorname{Im} z = 2$, as follows from Lemma 5.5 and the bound $\Gamma \leqslant C\eta^{-1}$, which itself follows easily by a simple expansion of $(1 - m^2 S)^{-1}$ combined with the bounds $\|S\|_{\ell^\infty \to \ell^\infty} \leqslant 1$ and (4.2). This concludes the proof. □

5.4. Iteration step and conclusion of the proof of Theorem 5.1. In the following a key role will be played by *deterministic* control parameters Ψ satisfying

$$cM^{-1/2} \;\leqslant\; \Psi \;\leqslant\; M^{-\gamma/3}\Gamma^{-1}\,. \tag{5.25}$$

(Using the definition of $\mathbf{S}$ and (4.4) it is not hard to check that the upper bound in (5.25) is always larger than the lower bound.) Suppose that $\Lambda \prec \Psi$ in $\mathbf{S}$ for some deterministic parameter Ψ satisfying (5.25). For example, by Proposition 5.3 we may choose $\Psi = M^{-\gamma/3}\Gamma^{-1}$.

We now improve the estimate $\Lambda \prec \Psi$ iteratively. The iteration step is the content of the following proposition.

Proposition 5.6. *Let Ψ be a control parameter satisfying* (5.25) *and fix $\varepsilon \in (0, \gamma/3)$. Then*

$$\Lambda \;\prec\; \Psi \qquad \Longrightarrow \qquad \Lambda \;\prec\; F(\Psi)\,, \tag{5.26}$$

where we defined

$$F(\Psi) \;:=\; M^{-\varepsilon}\Psi + \sqrt{\frac{\operatorname{Im} m}{M\eta}} + \frac{M^\varepsilon}{M\eta}\,.$$

For the proof of Proposition 5.6 we need the following averaging result, which is a simple corollary of Theorem 4.6.

Lemma 5.7. *Suppose that $\Lambda \prec \Psi$ for some deterministic control parameter Ψ satisfying* (4.8). *Then $[\Upsilon] = O_\prec(\Psi^2)$ (recall the definition of the average $[\cdot]$ from* (4.10)*).*

Proof. The claim easily follows from Schur's complement formula (5.6) written in the form

$$\Upsilon_i \;=\; A_i + Q_i \frac{1}{G_{ii}}\,.$$

We may therefore estimate $[\Upsilon]$ using the trivial bound $|A_i| \prec \Psi^2$ as well as the fluctuation averaging bound from the first estimate of (4.11) with $t_{ik} = 1/N$. □

Proof of Proposition 5.6. Suppose that $\Lambda \prec \Psi$ for some deterministic control parameter Ψ satisfying (5.25). We invoke Lemma 5.2 with $\phi = 1$ (recall the bound (4.5)) to get

$$\Lambda_o + |Z_i| + |\Upsilon_i| \;\prec\; \sqrt{\frac{\operatorname{Im} m + \Lambda}{M\eta}} \;\prec\; \sqrt{\frac{\operatorname{Im} m + \Psi}{M\eta}}\,. \tag{5.27}$$

Next, we estimate Λ_d. Define the z-dependent indicator function

$$\psi \;:=\; \mathbf{1}(\Lambda \leqslant M^{-\gamma/4})\,.$$

By (5.25), (4.5), and the assumption $\Lambda \prec \Psi$, we have $1-\psi \prec 0$. On the event $\{\psi = 1\}$, we expand the right-hand side of (5.9) to get the bound

$$\psi|v_i| \;\leqslant\; C\psi\Big|\sum_k s_{ik}v_k - \Upsilon_i\Big| + C\psi\Lambda^2\,.$$

Using the fluctuation averaging estimate (4.12) as well as (5.27), we find

$$\psi|v_i| \;\prec\; \Gamma\Psi^2 + \sqrt{\frac{\operatorname{Im} m + \Psi}{M\eta}}\,, \tag{5.28}$$

where we again used the lower bound from (4.5). Using $1-\psi \prec 0$ we conclude

$$\Lambda_d \;\prec\; \Gamma\Psi^2 + \sqrt{\frac{\operatorname{Im} m + \Psi}{M\eta}}\,, \tag{5.29}$$

which, combined with (5.27), yields

$$\Lambda \;\prec\; \Gamma\Psi^2 + \sqrt{\frac{\operatorname{Im} m + \Psi}{M\eta}}\,. \tag{5.30}$$

Using Young's inequality and the assumption $\Psi \leqslant M^{-\gamma/3}\Gamma^{-1}$ we conclude the proof. □

For the remainder of the proof of Theorem 5.1 we work on the spectral domain $\mathbf{S}$. We claim that if Ψ satisfies (5.25) then so does $F(\Psi)$. The lower bound $F(\Psi) \geqslant cM^{-1/2}$ is a consequence of the estimate $\operatorname{Im} m/\eta \geqslant c$, which follows from (4.4). The upper bound $M^{-\gamma/3-\varepsilon}\Gamma^{-1}$ on the first term of $F(\Psi)$ is trivial by assumption on Ψ. Moreover, the second term of $F(\Psi)$ satisfies $\sqrt{\operatorname{Im} m/(M\eta)} \leqslant M^{-\gamma}\Gamma^{-2} \leqslant CM^{-\gamma}\Gamma^{-1} \leqslant M^{-\gamma/3-\varepsilon}\Gamma^{-1}$ by definition of $\mathbf{S}$ and the lower bound (4.5). Similarly, the last term of $F(\Psi)$ satisfies $M^{\varepsilon}/(M\eta) \leqslant CM^{\varepsilon-\gamma}\Gamma^{-1} \leqslant M^{-\gamma/3-\varepsilon}\Gamma^{-1}$ by definition of $\mathbf{S}$.

We may therefore iterate (5.26). This yields a bound on Λ that is essentially the fixed point of the map $\Psi \mapsto F(\Psi)$, which is Π (up to the factor M^{ε}). More precisely, the iteration is started with $\Psi_0 := M^{-\gamma/3}\Gamma^{-1}$; the initial hypothesis $\Lambda \prec \Psi_0$ is provided by the rough bound from Proposition 5.3. For $k \geqslant 1$ we set $\Psi_{k+1} := F(\Psi_k)$. Hence from (5.26) we conclude that $\Lambda \prec \Psi_k$ for all k. Choosing $k := \lceil \varepsilon^{-1} \rceil$ yields

$$\Lambda \;\prec\; \sqrt{\frac{\operatorname{Im} m}{M\eta}} + \frac{M^{\varepsilon}}{M\eta}\,.$$

Since ε was arbitrary, we have proved that

$$\Lambda \;\prec\; \Pi\,, \tag{5.31}$$

which is (5.3).

What remains is to prove (5.4), i.e. to estimate Θ. We expand (5.9) on $\{\psi = 1\}$ to get

$$\psi m^2\Big(-\sum_k s_{ik}v_k + \Upsilon_i\Big) \;=\; -\psi v_i + O(\psi\Lambda^2)\,. \tag{5.32}$$

Averaging in (5.32) yields

$$\psi m^2\big(-[v] + [\Upsilon]\big) \;=\; -\psi[v] + O(\psi\Lambda^2)\,.$$

By (5.31) and (5.27) with $\Psi = \Pi$, we have $\Lambda + |\Upsilon_i| \prec \Pi$. Moreover, by Lemma 5.7 we have $|[\Upsilon]| \prec \Pi^2$. Thus we get

$$\psi[v] \;=\; m^2\psi[v] + O_\prec(\Pi^2)\,.$$

Since $1 - \psi \prec 0$, we conclude that $[v] = m^2[v] + O_\prec(\Pi^2)$. Therefore

$$|[v]| \;\prec\; \frac{\Pi^2}{|1-m^2|} \;\leqslant\; \left(\frac{\operatorname{Im} m}{|1-m^2|} + \frac{1}{|1-m^2|M\eta}\right)\frac{2}{M\eta} \;\leqslant\; \left(C + \frac{\Gamma}{M\eta}\right)\frac{2}{M\eta} \;\leqslant\; \frac{C}{M\eta}\,.$$

Here in the third step we used (4.3), (4.4), and the bound $\Gamma \geqslant |1-m^2|^{-1}$ which follows from the definition of Γ by applying the matrix $(1-m^2S)^{-1}$ to the vector $\mathbf{e} = N^{-1/2}(1,1,\dots,1)^*$. The last step follows from the definition of $\mathbf{S}$. Since $\Theta = |[v]|$, this concludes the proof of (5.4), and hence of Theorem 5.1.

6. Proof of Theorem 2.3

The key novelty in this proof is that we solve the self-consistent equation (5.9) separately on the subspace of constants (the span of the vector $\mathbf{e}$) and on its orthogonal complement $\mathbf{e}^\perp$. On the space of constant vectors, it becomes a scalar equation for the average $[v]$, which can be expanded up to second order. Near the spectral edges ± 2, the resulting quadratic self-consistent scalar equation (given in (6.2) below) is more effective than its linearized version. On the space orthogonal to the constants, we still solve a self-consistent vector equation, but the stability will now be quantified using $\widetilde{\Gamma}$ instead of the larger quantity Γ.

Accordingly, the main control parameter in this proof is $\Theta = |[v]|$, and the key iterative scheme (Lemma 6.7 below) is formulated in terms of Θ. However, many intermediate estimates still involve Λ. In particular, the self-consistent equation (5.9) is effective only in the regime where v_i is already small. Hence we need two preparatory steps. In Section 6.1 we will prove an apriori bound on Λ, essentially showing that $\Lambda \ll 1$. This proof itself is a continuity argument (see Figure 6.1 for a graphical illustration) similar to the proof of Proposition 5.3; now, however, we have to follow Λ and Θ in tandem. The main reason why Θ is already involved in this part is that we work in larger spectral domain $\widetilde{\mathbf{S}}$ defined using $\widetilde{\Gamma}$. Thus, already in this preparatory step, the self-consistent equation has to be solved separately on the subspace of constants and its orthogonal complement.

In Section 6.2, we control Λ in terms of Θ, which allows us to obtain a self-consistent equation involving only Θ. In this step we use the Fluctuation Averaging Theorem to obtain a quadratic estimate which, very roughly, states that $\Lambda \lesssim \Theta + \Lambda^2$ (see (6.20) below for the precise statement). This implies $\Lambda \lesssim \Theta$ in the regime $\Lambda \ll 1$.

Finally, in Section 6.3, we solve the quadratic iteration for Θ. Since the corresponding quadratic equation has a dichotomy and for large $\eta = \operatorname{Im} z$ we know that Θ is small by direct expansion, a continuity argument similar to the proof of Proposition 5.3 will complete the proof.

6.1. A rough bound on Λ. In this section we prove the following apriori bounds on both control parameters, Λ and Θ.

PROPOSITION 6.1. *In $\widetilde{\mathbf{S}}$ we have the bounds*

$$\Lambda \;\prec\; M^{-\gamma/4}\widetilde{\Gamma}^{-1}\,, \qquad \Theta \;\prec\; (M\eta)^{-1/3}\,.$$

Before embarking on the proof of Proposition 6.1, we state some preparatory lemmas. First, we derive the key equation for $[v] = N^{-1}\sum_i v_i$, the average of v_i.

LEMMA 6.2. *Define the z-dependent indicator function*

$$\phi \;:=\; \mathbf{1}(\Lambda \leqslant M^{-\gamma/4}\widetilde{\Gamma}^{-1}) \tag{6.1}$$

and the random control parameter

$$q(\Theta) \;:=\; \sqrt{\frac{\operatorname{Im} m + \Theta}{M\eta}} + \frac{\widetilde{\Gamma}}{M\eta}\,.$$

Then we have

$$\phi\Big((1-m^2)[v] - m^{-1}[v]^2\Big) \;=\; \phi\, O_{\prec}\big(q(\Theta) + M^{-\gamma/4}\Theta^2\big) \tag{6.2}$$

and

$$\phi\Lambda \;\prec\; \Theta + \widetilde{\Gamma}\, q(\Theta)\,. \tag{6.3}$$

PROOF. For the whole proof we work on the event $\{\phi = 1\}$, i.e. every quantity is multiplied by ϕ. We consistently drop these factors ϕ from our notation in order to avoid cluttered expressions. In particular, $\Lambda \leqslant CM^{-\gamma/4}$ throughout the proof.

We begin by estimating Λ_o and Λ_d in terms of Θ. Recalling (4.5), we find that ϕ satisfies the hypotheses of Lemma 5.2, from which we get

$$\Lambda_o + |\Upsilon_i| \;\prec\; r(\Lambda)\,, \qquad r(\Lambda) \;:=\; \sqrt{\frac{\operatorname{Im} m + \Lambda}{M\eta}}\,. \tag{6.4}$$

In order to estimate Λ_d, we expand the self-consistent equation (5.9) (on the event $\{\phi = 1\}$) to get

$$v_i - m^2 \sum_k s_{ik} v_k \;=\; O_{\prec}\big(\Lambda^2 + r(\Lambda)\big)\,; \tag{6.5}$$

here we used the bound (6.4) on $|\Upsilon_i|$. Next, we subtract the average $N^{-1}\sum_i$ from each side to get

$$(v_i - [v]) - m^2 \sum_k s_{ik}(v_k - [v]) \;=\; O_{\prec}\big(\Lambda^2 + r(\Lambda)\big)\,.$$

Note that the average of the left-hand side vanishes, so that the average of the right-hand side also vanishes. Hence the right-hand side is perpendicular to $\mathbf{e}$. Inverting the operator $1 - m^2 S$ on the subspace $\mathbf{e}^{\perp}$ therefore yields

$$\big|v_i - [v]\big| \;\prec\; \widetilde{\Gamma}\big(\Lambda^2 + r(\Lambda)\big)\,. \tag{6.6}$$

Combining with the bound $\Lambda_o \prec r(\Lambda)$ from (6.4), we therefore get

$$\Lambda \;\prec\; \Theta + \widetilde{\Gamma}\Lambda^2 + \widetilde{\Gamma} r(\Lambda)\,. \tag{6.7}$$

By definition of ϕ we have $\widetilde{\Gamma}\Lambda^2 \leqslant M^{-\gamma/4}\Lambda$, so that by Lemma 4.4 (iii) the second term on the right-hand side of (6.7) may be absorbed into the left-hand side:

$$\Lambda \;\prec\; \Theta + \widetilde{\Gamma} r(\Lambda)\,. \tag{6.8}$$

Now we claim that

$$r(\Lambda) \;\prec\; q(\Theta)\,. \tag{6.9}$$

If (6.9) is proved, clearly (6.3) follows from (6.8). In order to prove (6.9), we use (6.8) and the Cauchy-Schwarz inequality to get

$$r(\Lambda) \;\leqslant\; \sqrt{\frac{\operatorname{Im} m}{M\eta}} + \sqrt{\frac{\Lambda}{M\eta}} \;\prec\; \sqrt{\frac{\operatorname{Im} m}{M\eta}} + \sqrt{\frac{\Theta}{M\eta}} + \sqrt{\frac{\widetilde{\Gamma}\, r(\Lambda)}{M\eta}} \;\leqslant\; \sqrt{\frac{\operatorname{Im} m}{M\eta}} + \sqrt{\frac{\Theta}{M\eta}} + M^{-\varepsilon} r(\Lambda) + M^{\varepsilon} \frac{\widetilde{\Gamma}}{M\eta}$$

for any $\varepsilon > 0$. We conclude that

$$r(\Lambda) \;\prec\; \sqrt{\frac{\operatorname{Im} m}{M\eta}} + \sqrt{\frac{\Theta}{M\eta}} + M^{\varepsilon} \frac{\widetilde{\Gamma}}{M\eta}\,.$$

Since $\varepsilon > 0$ was arbitrary, (6.9) follows.

Next, we estimate Θ. We expand (5.9) to second order:

$$-\sum_k s_{ik} v_k + \Upsilon_i \;=\; -\frac{1}{m^2} v_i + \frac{1}{m^3} v_i^2 + O(\Lambda^3)\,. \tag{6.10}$$

In order to take the average and get a closed equation for $[v]$, we write, using (6.6),

$$v_i^2 \;=\; \big([v] + v_i - [v]\big)^2 \;=\; [v]^2 + 2[v](v_i - [v]) + O_{\prec}\Big(\widetilde{\Gamma}^2 \big(\Lambda^2 + r(\Lambda)\big)^2\Big)\,.$$

Plugging this back into (6.10) and taking the average over i gives

$$-m^2[v] + m^2[\Upsilon] \;=\; -[v] + m^{-1}[v]^2 + O_{\prec}\Big(\Lambda^3 + \widetilde{\Gamma}^2\Lambda^4 + \widetilde{\Gamma}^2 r(\Lambda)^2\Big)\,.$$

Estimating $[\Upsilon]$ by $\max|\Upsilon_i| \prec r(\Lambda)$ (recall (6.4)) yields

$$(1 - m^2)[v] - m^{-1}[v]^2 \;=\; O_{\prec}\Big(r(\Lambda) + \Lambda^3 + \widetilde{\Gamma}^2\Lambda^4 + \widetilde{\Gamma}^2 r(\Lambda)^2\Big)\,.$$

By definitions of $\widetilde{\mathbf{S}}$ and ϕ, we have $\widetilde{\Gamma}^2 r(\Lambda) \leqslant 1$. Therefore we may absorb the last error term into the first. For the second and third error terms we use (6.8) to get

$$(1 - m^2)[v] - m^{-1}[v]^2 \;=\; O_{\prec}\Big(r(\Lambda) + \Theta^3 + \widetilde{\Gamma}^3 r(\Lambda)^3 + \widetilde{\Gamma}^2\Theta^4 + \widetilde{\Gamma}^6 r(\Lambda)^4\Big)\,.$$

In order to conclude the proof of (6.2), we observe that, by the estimates $\Theta \leqslant \Lambda \leqslant CM^{-\gamma/4}$, $\widetilde{\Gamma}^2 r(\Lambda) \leqslant 1$, and $\Lambda \leqslant M^{-\gamma/4}\widetilde{\Gamma}^{-1}$, we have

$$\Theta^3 \;\leqslant\; CM^{-\gamma/4}\Theta^2\,, \qquad \widetilde{\Gamma}^3 r(\Lambda)^3 \;\leqslant\; r(\Lambda)\,, \qquad \widetilde{\Gamma}^2\Theta^4 \;\leqslant\; \widetilde{\Gamma}^2\Lambda^2\Theta^2 \;\leqslant\; M^{-\gamma/2}\Theta^2\,, \qquad \widetilde{\Gamma}^6 r(\Lambda)^4 \;\leqslant\; r(\Lambda)\,.$$

Putting everything together, we have

$$(1 - m^2)[v] - m^{-1}[v]^2 \;=\; O_{\prec}\big(r(\Lambda) + M^{-\gamma/4}\Theta^2\big)\,.$$

Hence (6.2) follows from (6.9). □

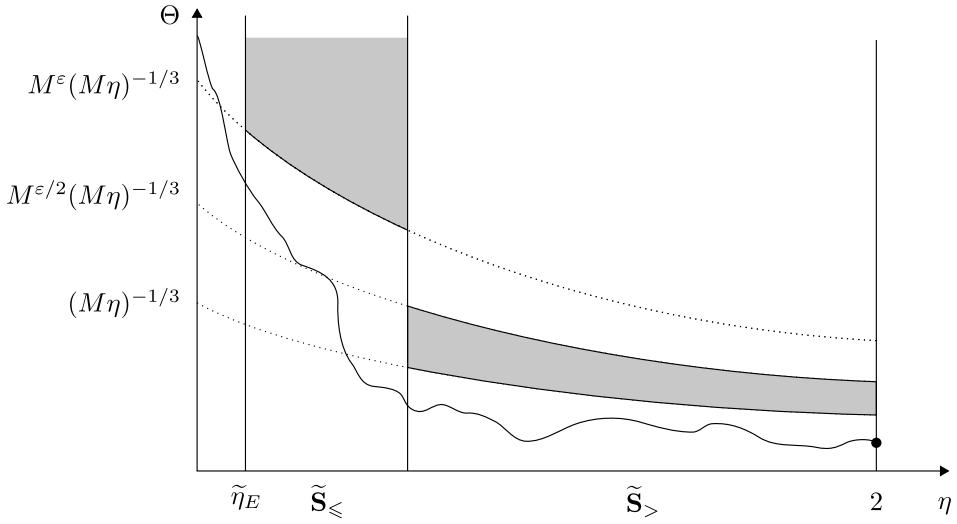

FIGURE 6.1. The (η, Θ)-plane for a fixed E near the edge (i.e. with small κ). The shaded regions are forbidden with high probability by Lemmas 6.3 and 6.4. The initial estimate, given by Lemma 5.5, is marked with a black dot. The graph of $\Theta = \Theta(E + \mathrm{i}\eta)$ is continuous, and hence lies beneath the shaded regions.

Next, we establish a bound analogous to Lemma 5.4, establishing gaps in the ranges of Λ and Θ. To that end, we need to partition $\widetilde{\mathbf{S}}$ in two. For the following we fix $\varepsilon \in (0, \gamma/12)$ and partition $\widetilde{\mathbf{S}} = \widetilde{\mathbf{S}}_{>} \cup \widetilde{\mathbf{S}}_{\leqslant}$, where

$$\widetilde{\mathbf{S}}_{>} \;:=\; \big\{z \in \widetilde{\mathbf{S}} : \sqrt{\kappa+\eta} > M^{\varepsilon}(M\eta)^{-1/3}\big\}\,, \qquad \widetilde{\mathbf{S}}_{\leqslant} \;:=\; \big\{z \in \widetilde{\mathbf{S}} : \sqrt{\kappa+\eta} \leqslant M^{\varepsilon}(M\eta)^{-1/3}\big\}.$$

The bound relies on (6.2), whereby one of the two terms on the left-hand side of (6.2) is estimated in terms of all the other terms, which are regarded as an error. In $\widetilde{\mathbf{S}}_{>}$ we shall estimate the first term on the left-hand side of (6.2), and in $\widetilde{\mathbf{S}}_{\leqslant}$ the second. Figure 6.1 summarizes the estimates on Θ of Lemma 6.3 and 6.4.

We begin with the domain $\widetilde{\mathbf{S}}_{>}$. In this domain, the following lemma roughly says that if $\Theta \leqslant M^{\varepsilon/2}(M\eta)^{-1/3}$ and $\Lambda \leqslant M^{-\gamma/4}\widetilde{\Gamma}^{-1}$ then we get the improved bounds $\Theta \prec (M\eta)^{-1/3}$, $\Lambda \prec M^{-\gamma/3}\widetilde{\Gamma}^{-1}$, i.e. we gain a small power of M. These improvements will be fed into the continuity argument as before.

LEMMA 6.3. *Let $\varepsilon \in (0, \gamma/12)$. Define the z-dependent indicator function*

$$\chi \;:=\; \mathbf{1}\Big(\Theta \leqslant M^{\varepsilon/2}(M\eta)^{-1/3}\Big)$$

and recall the indicator function ϕ from (6.1). *In $\widetilde{\mathbf{S}}_{>}$ we have the bounds*

$$\phi\chi\,\Theta \;\prec\; (M\eta)^{-1/3}\,, \qquad \phi\chi\,\Lambda \;\prec\; M^{-\gamma/3}\widetilde{\Gamma}^{-1}\,. \tag{6.11}$$

PROOF. From the definition of $\widetilde{\mathbf{S}}_{>}$ and (4.3) we get

$$\phi\chi\,|[v]| \;=\; \phi\chi\,\Theta \;\leqslant\; M^{\varepsilon/2}(M\eta)^{-1/3} \;\leqslant\; M^{-\varepsilon/2}\sqrt{\kappa+\eta} \;\leqslant\; CM^{-\varepsilon/2}|1-m^2|\,.$$

Therefore, on the event $\{\phi\chi = 1\}$, in (6.2) we may absorb the second term on the left-hand side and the second term on the right-hand side into the first term on the left-hand side:

$$\phi\chi\,(1-m^2)[v] \;=\; \phi O_{\prec}\big(q(\Theta)\big)\,.$$

Recalling $|1-m^2| \asymp \sqrt{\kappa+\eta}$ (see (4.3)), $\operatorname{Im} m \leqslant C\sqrt{\kappa+\eta}$ (see (4.4)), (6.9), $|[v]| = \Theta$, and the definition of $\widetilde{\mathbf{S}}_{>}$, we get

$$\begin{aligned}
\phi\chi\,\Theta \;&\prec\; \phi\chi\,(\kappa+\eta)^{-1/2}\left(\sqrt{\frac{\operatorname{Im} m}{M\eta}}+\sqrt{\frac{\Theta}{M\eta}}+\frac{\widetilde{\Gamma}}{M\eta}\right)\\
&\leqslant\; (\kappa+\eta)^{-1/4}(M\eta)^{-1/2}+(\kappa+\eta)^{-1/2}M^{\varepsilon/2}(M\eta)^{-2/3}+(\kappa+\eta)^{-1/2}\widetilde{\Gamma}(M\eta)^{-1}\\
&\leqslant\; (M\eta)^{-1/3}\,.
\end{aligned}$$

What remains is to estimate Λ. From (6.3), the bound $\widetilde{\Gamma}^2\sqrt{\operatorname{Im} m(M\eta)^{-1}} \leqslant M^{-\gamma}$ from the definition of $\widetilde{\mathbf{S}}$, and the estimate $\phi\widetilde{\Gamma}\Theta \leqslant \phi\widetilde{\Gamma}\Lambda \leqslant 1$ we get

$$\begin{aligned}
\phi\chi\,\Lambda \;&\prec\; \phi\chi\,\Theta + M^{-\gamma}\widetilde{\Gamma}^{-1}+\widetilde{\Gamma}\sqrt{\widetilde{\Gamma}^{-1}(M\eta)^{-1}}+\widetilde{\Gamma}^2(M\eta)^{-1}\\
&\prec\; (M\eta)^{-1/3}+M^{-\gamma/2}\widetilde{\Gamma}^{-1}+M^{-\gamma}\widetilde{\Gamma}^{-1}\\
&\leqslant\; 2M^{-\gamma/3}\widetilde{\Gamma}^{-1}\,.
\end{aligned}$$

This concludes the proof. □

Next, we establish a gap in the range of Λ, in the domain $\widetilde{\mathbf{S}}_{\leqslant}$. To that end, we improve the estimate on Λ from $\Lambda \leqslant M^{-\gamma/4}\widetilde{\Gamma}^{-1}$ to $\Lambda \prec M^{\varepsilon-\gamma/3}\widetilde{\Gamma}^{-1}$ as before. In this regime there is no need for a gap in Θ, i.e. the continuity argument will be performed on the value of Λ only.

LEMMA 6.4. *In $\widetilde{\mathbf{S}}_{\leqslant}$ we have the bounds*

$$\phi\Theta \;\prec\; M^{\varepsilon}(M\eta)^{-1/3}\,, \qquad \phi\Lambda \;\prec\; M^{\varepsilon-\gamma/3}\widetilde{\Gamma}^{-1}\,. \tag{6.12}$$

PROOF. We write (6.2) as

$$\phi[v](1-m^2-m^{-1}[v]) \;=\; \phi\, O_{\prec}\big(q(\Theta)+M^{-\gamma/4}\Theta^2\big).$$

Solving this quadratic relation for $[v]$, we get

$$\phi\Theta \;\prec\; |1-m^2|+\phi\sqrt{q(\Theta)+M^{-\gamma/4}\Theta^2}\,. \tag{6.13}$$

Using (4.4), the bound $\widetilde{\Gamma} \leqslant M^{-\gamma/3}(M\eta)^{1/3} \leqslant (M\eta)^{1/3}$ from the definition of $\widetilde{\mathbf{S}}$, and Young's inequality, we estimate

$$\begin{aligned}
\sqrt{q(\Theta)+M^{-\gamma/4}\Theta^2} \;&\leqslant\; (\operatorname{Im} m)^{1/4}(M\eta)^{-1/4}+\Theta^{1/4}(M\eta)^{-1/4}+\widetilde{\Gamma}^{\,1/2}(M\eta)^{-1/2}+M^{-\gamma/8}\Theta\\
&\leqslant\; C\sqrt{\kappa+\eta}+CM^{\varepsilon}(M\eta)^{-1/3}+CM^{-\varepsilon}\Theta\,.
\end{aligned}$$

Plugging this bound into (6.13), together with (4.3) and the definition of $\widetilde{\mathbf{S}}_{\leqslant}$, we find

$$\phi\Theta \;\prec\; \sqrt{\kappa+\eta} + M^{\varepsilon}(M\eta)^{-1/3} \;\leqslant\; 2M^{\varepsilon}(M\eta)^{-1/3}\,.$$

This proves the first bound of (6.12).

What remains is the estimate of Λ. From (6.3) and the bounds $\widetilde{\Gamma} \leqslant M^{-\gamma/3}(M\eta)^{1/3}$ and $\widetilde{\Gamma}^2\sqrt{\operatorname{Im} m(M\eta)^{-1}} \leqslant M^{-\gamma}$ from the definition of $\widetilde{\mathbf{S}}$, we get

$$\begin{aligned}\phi\Lambda \;&\prec\; \phi\Theta + M^{-\gamma}\widetilde{\Gamma}^{-1} + \widetilde{\Gamma}\sqrt{\widetilde{\Gamma}^{-1}(M\eta)^{-1}} + \widetilde{\Gamma}^2(M\eta)^{-1} \\ &\prec\; M^{\varepsilon}(M\eta)^{-1/3} + M^{-\gamma/2}\widetilde{\Gamma}^{-1} + M^{-\gamma}\widetilde{\Gamma}^{-1} \\ &\leqslant\; 2M^{\varepsilon-\gamma/3}\widetilde{\Gamma}^{-1}\,.\end{aligned}$$

This concludes the proof. □

We now have all of the ingredients to complete the proof of Proposition 6.1.

Proof of Proposition 6.1. The proof is a continuity argument similar to the proof of Proposition 5.3. In a first step, we prove that

$$\Lambda \;\prec\; M^{-\gamma/3}\widetilde{\Gamma}^{-1}\,, \qquad \Theta \;\prec\; (M\eta)^{-1/3}\,. \tag{6.14}$$

in $\widetilde{\mathbf{S}}_{>}$. The continuity argument is almost identical to that following (5.21); the only difference is that we keep track of the two parameters Λ and Θ. The required gaps in the ranges of Λ and Θ are provided by (6.11), and the argument is closed using the large-η estimate from Lemma 5.5, which yields $\Theta \leqslant \Lambda \prec M^{-1/2}$ for $\eta = 2$.

In a second step, we prove that

$$\Lambda \;\prec\; M^{\varepsilon-\gamma/4}\widetilde{\Gamma}^{-1}\,, \qquad \Theta \;\prec\; M^{\varepsilon}(M\eta)^{-1/3}$$

in $\widetilde{\mathbf{S}}_{\leqslant}$. This is again a continuity argument almost identical to that following (5.21). Now we establish a gap only in the range of Λ. The gap is provided by (6.12) (recall that by definition of ε we have $\varepsilon - \gamma/3 < -\gamma/4$), and the argument is closed using the bound (6.14) at the boundary of the domains $\widetilde{\mathbf{S}}_{>}$ and $\widetilde{\mathbf{S}}_{\leqslant}$.

The claim now follows since we may choose $\varepsilon \in (0, \gamma/12)$ to be arbitrarily small. This concludes the proof of Proposition 6.1. □

6.2. An improved bound on Λ in terms of Θ. In (6.3) we already estimated Λ in terms of Θ; the goal of this section is to improve this bound by removing the factor $\widetilde{\Gamma}$ from that estimate. We do this using the Fluctuation Averaging Theorem, but we stress that the removal of a factor $\widetilde{\Gamma}$ is not the main rationale for using the fluctuation averaging mechanism. Its fundamental use will take place in Lemma 6.6 below. A technical consequence of invoking fluctuation averaging is that we have to use deterministic control parameters instead of random ones. Thus, we introduce a deterministic control parameter Φ that captures the size of the random control parameter Θ through the relation $\Theta \prec \Phi$. Throughout the following we shall make use of the control parameter

$$p(\Phi) \;:=\; \sqrt{\frac{\operatorname{Im} m + \Phi}{M\eta}} + \frac{1}{M\eta}\,,$$

which differs from $q(\Phi)$ only by a factor $\widetilde{\Gamma}$ in the second term.

LEMMA 6.5. *Suppose that* $\Lambda \prec \Psi$ *and* $\Theta \prec \Phi$ *in* $\widetilde{\mathbf{S}}$ *for some deterministic control parameters* Ψ *and* Φ *satisfying*

$$cM^{-1/2} \leqslant \Psi \leqslant CM^{-\gamma/4}\widetilde{\Gamma}^{-1}, \qquad \Phi \leqslant CM^{-\gamma/4}\widetilde{\Gamma}^{-1}. \tag{6.15}$$

Then

$$\Lambda_o + |Z_i| \prec p(\Phi), \qquad \Lambda \prec p(\Phi) + \Phi. \tag{6.16}$$

We remark that, by Proposition 6.1, the choice $\Psi = M^{-\gamma/4}\widetilde{\Gamma}^{-1}$ and $\Phi = (M\eta)^{-1/3} \leqslant M^{-\gamma/4}\widetilde{\Gamma}^{-1}$ satisfies the assumptions of Lemma 6.5.

PROOF OF LEMMA 6.5. Choosing $\phi = 1$ in Lemma 5.2 and recalling (4.5), we get

$$\Lambda_o + |\Upsilon_i| \prec r(\Psi), \qquad r(\Psi) := \sqrt{\frac{\operatorname{Im} m + \Psi}{M\eta}}. \tag{6.17}$$

In order to estimate Λ_d, as in (5.32), we expand (5.9) to get

$$-\sum_k s_{ik}v_k + \Upsilon_i = -m^{-2}v_i + O_\prec(\Psi^2). \tag{6.18}$$

As in the proof of (5.32) and (6.5), the expansion of (5.9) is only possible on the event $\{\Lambda \leqslant M^{-\delta}\}$ for some $\delta > 0$. By $\Lambda \prec \Psi$ and (6.15), the indicator function of this event is $1 + O_\prec(0)$; the contribution $O_\prec(0)$ of the complementary event can be absorbed in the error term $O_\prec(\Psi^2)$.

Subtracting the average $N^{-1}\sum_i$ from both sides of (6.18) and estimating m^2 by a constant (see (4.2)) yields

$$\big|v_i - [v]\big| \leqslant C\bigg|\sum_k s_{ik}\big(v_k - [v]\big) - \big(\Upsilon_i - [\Upsilon]\big)\bigg| + O_\prec(\Psi^2) \prec \widetilde{\Gamma}\Psi^2 + r(\Psi), \tag{6.19}$$

where in the last step we used the fluctuation averaging estimate (4.14) and $|\Upsilon_i| \prec r(\Psi)$ from (6.17). Together with $|[v]| = \Theta \prec \Phi$, this gives the estimate $\Lambda_d \prec \Gamma\Psi^2 + \Phi + r(\Psi)$. Combining it with the bound (6.17), we conclude that

$$\Lambda \prec \widetilde{\Gamma}\Psi^2 + \Phi + r(\Psi). \tag{6.20}$$

Now fix $\varepsilon \in (0, \gamma/4)$. Using the assumption $\widetilde{\Gamma}\Psi \leqslant CM^{-\gamma/4} \leqslant M^{-\varepsilon}$, we conclude: if Ψ and Φ satisfy (6.15) then

$$\Lambda \prec \Psi \qquad \Longrightarrow \qquad \Lambda \prec F(\Psi, \Phi), \tag{6.21}$$

where we defined

$$F(\Psi, \Phi) := M^{-\varepsilon}\Psi + \Phi + \sqrt{\frac{\operatorname{Im} m}{M\eta}} + \frac{M^\varepsilon}{M\eta},$$

which plays a role similar to $F(\Psi)$ in Proposition 5.6. (Here we estimated $\sqrt{\Psi(M\eta)^{-1}}$ in $r(\Psi)$ by $M^{-\varepsilon}\Psi + M^\varepsilon(M\eta)^{-1}$.) From (4.4) and the definition of $\widetilde{\mathbf{S}}$ it easily follows that if (Ψ, Φ) satisfy (6.15) then so do $(F(\Psi, \Phi), \Phi)$. Therefore iterating (6.21) $\lceil \varepsilon^{-1} \rceil$ times and using the fact that $\varepsilon \in (0, \gamma/4)$ was arbitrary yields

$$\Lambda \prec \sqrt{\frac{\operatorname{Im} m}{M\eta}} + \frac{1}{M\eta} + \Phi. \tag{6.22}$$

This implies the claimed bound (6.16) on Λ. Calling the right-hand side of (6.22) Ψ, we find

$$r(\Psi) \leqslant Cp(\Phi). \tag{6.23}$$

Hence the claimed bound (6.16) on Λ_o and Z_i follows from (6.17). □

6.3. Iteration for Θ and conclusion of the proof of Theorem 2.3. Next, we prove the following version of (5.9), which is the key tool for estimating Θ.

LEMMA 6.6. *Let Φ be some deterministic control parameter satisfying $\Theta \prec \Phi$ in $\widetilde{\mathbf{S}}$. Then*

$$(1-m^2)[v] - m^{-1}[v]^2 \;=\; O_\prec\big(p(\Phi)^2 + M^{-\gamma/4}\Phi^2\big)\,. \tag{6.24}$$

Notice that this bound is stronger than the previous formula (6.2), as the power of $p(\Phi)$ is two instead of one. The improvement is due to using fluctuation averaging in $[\Upsilon]$. Otherwise the proof is very similar to that of (6.2).

PROOF. By Proposition 6.1, we may assume that

$$\Phi \;\leqslant\; M^{-\gamma/4}\widetilde{\Gamma}^{-1} \tag{6.25}$$

since $\Theta \leqslant \Lambda \prec M^{-\gamma/4}\widetilde{\Gamma}^{-1}$. From Lemma 6.5 we get $\Lambda_o + |Z_i| \prec p(\Phi)$ and $\Lambda \prec \Psi$, where

$$\Psi \;:=\; p(\Phi) + \Phi\,. \tag{6.26}$$

By definition of $\widetilde{\mathbf{S}}$ and (6.25), we find that $\Psi \leqslant 2M^{-\gamma/4}\widetilde{\Gamma}^{-1}$.

Now we expand the right-hand side of (5.9) exactly as in (6.10) to get

$$-m^2\sum_k s_{ik}v_k + m^2\Upsilon_i \;=\; -v_i + m^{-1}v_i^2 + O_\prec(\Psi^3)\,. \tag{6.27}$$

Using Theorem 4.7 and the bound $\Lambda_o \prec p(\Phi)$ from Lemma 6.5, we may prove, exactly as in Lemma 5.7, that $|[\Upsilon]| \prec p(\Phi)^2$. Taking the average over i in (6.27) therefore yields

$$(1-m^2)[v] - m^{-1}\frac{1}{N}\sum_i v_i^2 \;=\; -m^2[\Upsilon] + O_\prec(\Psi^3) \;=\; O_\prec\big(p(\Phi)^2 + \Psi^3\big)\,. \tag{6.28}$$

Using the estimates (6.19) and (6.23), we write the quadratic term on the left-hand side as

$$\frac{1}{N}\sum_i v_i^2 \;=\; [v]^2 + \frac{1}{N}\sum_i \big(v_i - [v]\big)^2 \;=\; [v]^2 + O_\prec\Big(\big(\widetilde{\Gamma}\Psi^2 + p(\Phi)\big)^2\Big) \;=\; [v]^2 + O_\prec\big(M^{-\gamma/2}\Psi^2 + p(\Phi)^2\big)$$

where we also used $\widetilde{\Gamma}\Psi \leqslant 2M^{-\gamma/4}$, as observed after (6.26). From (6.28) we therefore get

$$(1-m^2)[v] - m^{-1}[v]^2 \;=\; O_\prec\big(p(\Phi)^2 + M^{-\gamma/4}\Psi^2\big)\,.$$

The claim follows from (6.26). □

The bound on Θ will follow by iterating the following estimate.

LEMMA 6.7. *Fix $\varepsilon \in (0, \gamma/12)$ and suppose that $\Theta \prec \Phi$ in $\widetilde{\mathbf{S}}$ for some deterministic control parameter Φ.*

(i) If $\Phi \geqslant M^{3\varepsilon}(M\eta)^{-1}$ then

$$\Theta \;\prec\; M^{-\varepsilon}\Phi\,. \tag{6.29}$$

(ii) If $|E| \geqslant 2$, $\frac{M^{3\varepsilon}}{M(\kappa+\eta)} \leqslant \Phi \leqslant M^{\varepsilon}\sqrt{\kappa+\eta}$, *and* $M\eta\sqrt{\kappa+\eta} \geqslant M^{2\varepsilon}$, *then*

$$\Theta \;\prec\; \frac{1}{(M\eta)^2\sqrt{\kappa+\eta}} + M^{-\varepsilon}\Phi\,. \tag{6.30}$$

Proof. We begin by partitioning $\widetilde{\mathbf{S}} = \widetilde{\mathbf{S}}^{>} \cup \widetilde{\mathbf{S}}^{\leqslant}$. This partition is analogous to the partition $\widetilde{\mathbf{S}} = \widetilde{\mathbf{S}}_{>} \cup \widetilde{\mathbf{S}}_{\leqslant}$ from Section 6.1, and will determine which of the two terms in the left-hand side of (6.24) is estimated in terms of the others. Here

$$\widetilde{\mathbf{S}}^{>} \;:=\; \big\{z \in \widetilde{\mathbf{S}} : \sqrt{\kappa+\eta} > M^{-\varepsilon}\Phi\big\}\,, \qquad \widetilde{\mathbf{S}}^{\leqslant} \;:=\; \big\{z \in \widetilde{\mathbf{S}} : \sqrt{\kappa+\eta} \leqslant M^{-\varepsilon}\Phi\big\}.$$

We begin with the domain $\widetilde{\mathbf{S}}^{>}$. Let $K > 0$ be a constant large enough that

$$\sqrt{\kappa+\eta} \;\leqslant\; \frac{K}{2}\big|1-m^2\big||m|\,;$$

such constant exists by (4.2) and (4.3). Define the indicator function

$$\psi \;:=\; \mathbf{1}\big(\Theta \leqslant \sqrt{\kappa+\eta}/K\big)\,. \tag{6.31}$$

Hence on the event $\{\psi = 1\}$ we may absorb the quadratic term on the left-hand side of (6.24) into the linear term, to get the bound

$$\psi\Theta \;\prec\; (\kappa+\eta)^{-1/2}\bigg(\frac{\operatorname{Im} m + \Phi}{M\eta} + \frac{1}{(M\eta)^2} + M^{-\gamma/4}\Phi^2\bigg) \;\leqslant\; C\frac{M^{\varepsilon}}{M\eta} + M^{\varepsilon-\gamma/4}\Phi \;\leqslant\; CM^{-2\varepsilon}\Phi\,, \tag{6.32}$$

where in the second step we used (4.4), the assumption $(M\eta)^{-1} \leqslant M^{-3\varepsilon}\Phi \leqslant \Phi$, and the definition of $\widetilde{\mathbf{S}}^{>}$. We conclude that in $\widetilde{\mathbf{S}}^{>}$ we have

$$\psi\Theta \;\prec\; M^{-2\varepsilon}\Phi \;\leqslant\; M^{-\varepsilon}\sqrt{\kappa+\eta}\,, \tag{6.33}$$

where in the last step we used the definition of $\widetilde{\mathbf{S}}^{>}$. This means that there is a gap of order $\sqrt{\kappa+\eta}$ between the bound in the definition of ψ in (6.31) and the right-hand side of (6.33). Moreover, by Proposition 6.1 we have $\Theta \prec M^{-\varepsilon}\sqrt{\kappa+\eta}$ for $\eta = 2$. Hence a continuity argument on Θ, similar to the proof of Proposition 5.3, yields (6.29) in $\widetilde{\mathbf{S}}^{>}$.

Let us now consider the domain $\widetilde{\mathbf{S}}^{\leqslant}$. We write the left-hand side of (6.24) as $(1-m^2-m^{-1}[v])[v]$. Solving the resulting equation for $[v]$, as in the proof of (6.13), yields the bound

$$\Theta \;\prec\; |1-m^2| + p(\Phi) + M^{-\gamma/8}\Phi \;\leqslant\; CM^{-\varepsilon}\Phi + \sqrt{\frac{\operatorname{Im} m + \Phi}{M\eta}} + \frac{1}{M\eta} \;\leqslant\; CM^{-\varepsilon}\Phi + \frac{M^{\varepsilon}}{M\eta} \;\leqslant\; CM^{-\varepsilon}\Phi\,, \tag{6.34}$$

where we used the definition of $\widetilde{\mathbf{S}}^{\leqslant}$ and the bounds (4.3) and (4.4). This proves (6.29) in $\widetilde{\mathbf{S}}^{\leqslant}$, and hence completes the proof of part (i) of Lemma 6.7.

The proof of part (ii) is analogous. In this case we are in the domain $\widetilde{\mathbf{S}}^{>}$, and use the estimate $\operatorname{Im} m \leqslant C\eta(\kappa+\eta)^{-1/2}$ from (4.4) instead of $\operatorname{Im} m \leqslant C\sqrt{\kappa+\eta}$ in (6.32). Using the other assumptions in part (ii), we have

$$\psi\Theta \;\prec\; \frac{1}{(M\eta)^2\sqrt{\kappa+\eta}} + CM^{-2\varepsilon}\Phi \;\leqslant\; M^{-\varepsilon}\sqrt{\kappa+\eta}\,, \tag{6.35}$$

which replaces (6.32) and (6.33). The rest of the argument is unchanged. □

Armed with Lemma 6.7, we may now complete the proof of Theorem 2.3. Fix $\varepsilon \in (0, \gamma/12)$. From Proposition 6.1 we get that $\Theta \prec \Phi_0$ for $\Phi_0 := (M\eta)^{-1/3} + M^{3\varepsilon}(M\eta)^{-1}$. Iteration of Lemma 6.7 therefore implies that, for all $k \in \mathbb{N}$, we have $\Theta \prec \Phi_k$ where

$$\Phi_{k+1} \;:=\; \frac{M^{3\varepsilon}}{M\eta} + M^{-\varepsilon}\Phi_k \;\leqslant\; C_k\biggl(\frac{M^{3\varepsilon}}{M\eta} + M^{-\varepsilon k}\Phi_0\biggr)\,.$$

Choosing $k = \lceil \varepsilon^{-1} \rceil$ yields $\Theta \prec M^{3\varepsilon}(M\eta)^{-1}$. Since ε can be made as small as desired, we therefore obtain $\Theta \prec (M\eta)^{-1}$. This is (2.19).

In the regime $|E| \geqslant 2$, the same argument with the better iteration bound (6.30) yields (2.20). The iteration can be started with $\Phi_0 = M^{3\varepsilon}(M\eta)^{-1}$ from (2.19).

Finally, the bound $\Lambda \prec \Pi$ in (2.18) follows from (2.19) and Lemma 6.5. This concludes the proof of Theorem 2.3.

7. Density of states and eigenvalue locations

In this section we apply the local semicircle law to obtain information on the density of states and on the location of eigenvalues. The techniques used here have been developed in a series of papers [6, 13, 15, 19].

The first result is to translate the local semicircle law, Theorem 2.3, into a statement on the counting function of the eigenvalues. Let $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_N$ denote the ordered eigenvalues of H, and recall the semicircle density ϱ defined in (2.7). We define the distribution functions

$$n(E) \;:=\; \int_{-\infty}^{E} \varrho(x)\,\mathrm{d}x\,, \qquad \mathfrak{n}_N(E) \;:=\; \frac{1}{N}\bigl|\{\alpha : \lambda_\alpha \leqslant E\}\bigr| \tag{7.1}$$

for the semicircle law and the empirical eigenvalue density of H. Recall also the definition (2.15) of κ_x for $x \in \mathbb{R}$ and the definition (2.14) of $\widetilde{\eta}_x$ for $|x| \leqslant 10$. The following result is proved in Section 7.1 below.

Lemma 7.1. *Suppose that* (2.19) *holds uniformly in* $z \in \widetilde{\mathbf{S}}$, *i.e. for* $|E| \leqslant 10$ *and* $\widetilde{\eta}_E \leqslant \eta \leqslant 10$ *we have*

$$|m_N(z) - m(z)| \;\prec\; \frac{1}{M\eta}\,. \tag{7.2}$$

For given $E_1 < E_2$ *in* $[-10, 10]$ *we abbreviate*

$$\widetilde{\eta} \;:=\; \max\bigl\{\widetilde{\eta}_E : E \in [E_1, E_2]\bigr\}. \tag{7.3}$$

Then, for $-10 \leqslant E_1 < E_2 \leqslant 10$, *we have*

$$\Bigl|\bigl(\mathfrak{n}_N(E_2) - \mathfrak{n}_N(E_1)\bigr) - \bigl(n(E_2) - n(E_1)\bigr)\Bigr| \;\prec\; \widetilde{\eta}\,. \tag{7.4}$$

The accuracy of the estimate (7.4) depends on $\widetilde{\Gamma}$ (see (A.3) for explicit bounds on $\widetilde{\Gamma}$), since $\widetilde{\Gamma}$ determines $\widetilde{\eta}_E$, the smallest scale on which the local semicircle law (Theorem 2.3) holds around the energy E. In the regime away from the spectral edges $E = \pm 2$ and away from $E = 0$, the parameter $\widetilde{\Gamma}$ is essentially bounded (see the example (i) from Section 3); in this case $\widetilde{\eta}_E \asymp M^{-1}$ (up to an irrelevant logarithmic factor). For E near 0, the parameter $\widetilde{\Gamma}$ blows up as E^{-2}, so that $\widetilde{\eta}_E \sim E^{-12}M^{-1}$; however, if S has a positive gap δ_- at

the bottom of its spectrum, $\widetilde{\Gamma}$ remains bounded in the vicinity of $E = 0$ (see (A.3)). See Definition A.1 in Appendix A for the definition of the spectral gaps $\delta_\pm$.

A typical example of S without a positive gap δ_- is a 2×2 block matrix with zero diagonal blocks, i.e. $s_{ij} = 0$ if $i, j \leqslant L$ or $L + 1 \leqslant i, j \leqslant N$. In this case, the vector $\mathbf{v} = (1, 1, \dots 1, -1, -1, \dots -1)$ consisting of L ones and $N - L$ minus ones satisfies $S\mathbf{v} = -\mathbf{v}$, so that -1 is in fact an eigenvalue of S. Since at energy $E = 0$ we have $m^2(z) = m^2(i\eta) = -1 + O(\eta)$, the inverse matrix $(1 - m^2 S)^{-1}$, even after restricting it to $\mathbf{e}^\perp$, becomes singular as $\eta \to 0$. Thus, $\widetilde{\Gamma}(i\eta) \sim \eta^{-1}$, and the estimates leading to Theorem 2.3 become unstable. The corresponding random matrix has the form

$$H \;=\; \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix}$$

where A is an $L \times (N - L)$ rectangular matrix with independent centred entries. The eigenvalues of H are the square roots (with both signs) of the eigenvalues of the random covariance matrices AA^* and A^*A, whose spectral density is asymptotically given by the *Marchenko-Pastur law* [24]. The instability near $E = 0$ arises from the fact that H has a macroscopically large kernel unless $L/N \to 1/2$. In the latter case the support of the Marchenko-Pastur law extends to zero and in fact the density diverges as $E^{-1/2}$. We remark that a local version of the Marchenko-Pastur law was given in [15] for the case when the limit of L/N differs from 0, $1/2$ and ∞; the "hard edge" case, $L/N \to 1/2$, in which the density near the lower spectral edge is singular, was treated in [2].

This example shows that the vanishing of δ_- may lead to a very different behaviour of the spectral statistics. Although our technique is also applicable to random covariance matrices, for simplicity in this section we assume that $\delta_- \geqslant c$ for some positive constant c. By Proposition A.3, this holds for random band matrices, for full Wigner matrices (see Definition 3.1), and for their combinations; these examples are our main interest in this paper.

Under the condition $\delta_- \geqslant c$, the upper bound of (A.3) yields

$$\widetilde{\Gamma}(E + i\eta) \;\leqslant\; \frac{C \log N}{\delta_+ + \theta}\,, \tag{7.5}$$

where θ was defined in (3.2) and δ_+ is the upper gap of the spectrum of S given in Definition A.1. Notice that θ vanishes near the spectral edge $E = \pm 2$ as $\eta \to 0$. For the purpose of estimating $\widetilde{\Gamma}$, this deterioration is mitigated if the upper gap δ_+ is non-vanishing. While full Wigner matrices satisfy $\delta_+ \geqslant c$, the lower bound on δ_+ for band matrices is weaker; see Proposition A.3 for a precise statement.

We first give an estimate on $\widetilde{\eta}_x$ using the explicit bound (7.5). While not fully optimal, this estimate is sufficient for our purposes and in particular reproduces the correct behaviour when $\delta_+ \geqslant c$.

LEMMA 7.2. *Suppose that $\delta_- \geqslant c$ (so that (7.5) holds). Then we have for any $|x| \leqslant 2$*

$$\widetilde{\eta}_x \;\leqslant\; \frac{CM^{3\gamma}}{M(\kappa_x + \delta_+ + M^{-1/5})^{7/2}}\,. \tag{7.6}$$

In the regime $2 \leqslant |x| \leqslant 10$ we have the improved bound

$$\widetilde{\eta}_x \;\leqslant\; \frac{CM^{3\gamma}}{M(\sqrt{\kappa_x} + \delta_+ + M^{-1/5})^{3}}\,. \tag{7.7}$$

PROOF. For any $|x| \leqslant 2$ define η_x' as the solution of the equation

$$\sqrt{\frac{\sqrt{\kappa_x+\eta}}{M\eta}}\frac{1}{(\kappa_x+\eta^{2/3}+\delta_+)^2}+\frac{1}{M\eta}\frac{1}{(\kappa_x+\eta^{2/3}+\delta_+)^3} \;=\; M^{-\frac{3\gamma}{2}}. \tag{7.8}$$

This solution is unique since the left-hand side is decreasing in η. An elementary but tedious analysis of (7.8) yields

$$\eta_x' \;\leqslant\; \frac{CM^{3\gamma}}{M(\kappa_x+\delta_++M^{-1/5})^{7/2}}\,. \tag{7.9}$$

(The calculation is based on the observation that if $\eta(a+\eta^\alpha) \leqslant b$ for some $a,b>0$ and $\alpha \geqslant 0$, then $\eta \leqslant 2b(b^{\frac{\alpha}{1+\alpha}}+a)^{-1}$.) From (7.5), $\operatorname{Im} m(x+i\eta) \leqslant C\sqrt{\kappa_x+\eta}$ (see (4.4)) and the simple bound $\theta(x+i\eta) \geqslant c(\kappa_x+\eta^{2/3})$, we get for $\eta \geqslant \eta_x'$

$$\sqrt{\frac{\operatorname{Im} m(x+i\eta)}{M\eta}}\widetilde{\Gamma}^2(x+i\eta)+\frac{1}{M\eta}\widetilde{\Gamma}^3(x+i\eta) \;\leqslant\; C(\log N)^3M^{-\frac{3\gamma}{2}}.$$

From the definition (2.17) of $\widetilde{\mathbf{S}}$, we therefore get $\widetilde{\eta}_x \leqslant \eta_x'$, which proves (7.6).

The proof of (7.7) is similar, but we use $\theta=\sqrt{\kappa+\eta}$ and the stronger bound $\operatorname{Im} m \leqslant \eta/\sqrt{\kappa+\eta}$ available in the regime $|x| \geqslant 2$. For $2 \leqslant |x| \leqslant 10$, define η_x' to be the solution of the equation

$$\sqrt{\frac{1}{M\sqrt{\kappa_x+\eta}}}\frac{1}{(\sqrt{\kappa_x+\eta}+\delta_+)^2}+\frac{1}{M\eta}\frac{1}{(\sqrt{\kappa_x+\eta}+\delta_+)^3} \;=\; M^{-\frac{3\gamma}{2}}. \tag{7.10}$$

As for (7.9), a tedious calculation yields

$$\eta_x' \;\leqslant\; \frac{CM^{3\gamma}}{M(\sqrt{\kappa_x}+\delta_++M^{-1/5})^3}\,.$$

This concludes the proof. □

Next, we obtain an estimate on the extreme eigenvalues.

THEOREM 7.3 (EXTREMAL EIGENVALUES). *Suppose that $\delta_- \geqslant c$ (so that (7.5) holds) and that $N^{3/4} \leqslant M \leqslant N$. Then we have*

$$\|H\| \;\leqslant\; 2+O_\prec(X)\,, \tag{7.11}$$

where we introduced the control parameter

$$X \;:=\; \frac{N^2}{M^{8/3}}+\left(\frac{N}{M^2}\right)^2\left[\delta_++\left(\frac{N}{M^2}\right)^{1/7}\right]^{-12}. \tag{7.12}$$

In particular, if $\delta_+ \geqslant c$ then

$$\|H\| \;\leqslant\; 2+O_\prec\left(\frac{N^2}{M^{8/3}}\right). \tag{7.13}$$

Note that (7.13) yields the optimal error bound $O_\prec(N^{-2/3})$ in the case of a full and flat Wigner matrix (see Definition 3.1). Under stronger assumptions on the law of the entries of H, Theorem 7.3 can be improved as follows.

THEOREM 7.4. *Suppose that the matrix elements h_{ij} have a uniform subexponential decay, i.e. that there exist positive constants C and ϑ such that*

$$\mathbb{P}\big(|h_{ij}| \geqslant x^{\vartheta}\sqrt{s_{ij}}\big) \;\leqslant\; C\,\mathrm{e}^{-x}\,. \tag{7.14}$$

Then (7.11) *holds with*

$$X \;:=\; M^{-1/4}\,. \tag{7.15}$$

If in addition the law of each matrix entry is symmetric (i.e. h_{ij} and $-h_{ij}$ have the same law), then (7.11) *holds with*

$$X \;:=\; M^{-2/3}\,. \tag{7.16}$$

We remark that (7.15) can obtained via a relatively standard moment method argument combined with refined combinatorics. Obtaining the bound (7.16) is fairly involved; it makes use of the Chebyshev polynomial representation first used by Feldheim and Sodin [22,27] in this context for a special distribution of h_{ij}, and extended in [5] to general symmetric entries.

PROOF OF THEOREM 7.3. We shall prove a lower bound on the smallest eigenvalue λ_1 of H; the largest eigenvalue λ_N may be estimated similarly from above. Fix a small $\gamma > 0$ and set

$$\ell \;:=\; M^{6\gamma}\frac{N^2}{M^{8/3}}\,.$$

We distinguish two regimes depending on the location of λ_1, i.e. we decompose

$$\mathbf{1}(\lambda_1 \leqslant -2-\ell) \;=\; \phi_1 + \phi_2\,,$$

where

$$\phi_1 \;:=\; \mathbf{1}(-3 \leqslant \lambda_1 \leqslant -2-\ell)\,, \qquad \phi_2 \;:=\; \mathbf{1}(\lambda_1 \leqslant -3)\,.$$

In the first regime we further decompose the probability space by estimating

$$\phi_1 \;\leqslant\; \sum_{k=0}^{k_0}\phi_{1,k}\,, \qquad \phi_{1,k} \;:=\; \mathbf{1}\bigg(-2-\ell-\frac{k+1}{N} \leqslant \lambda_1 \leqslant -2-\ell-\frac{k}{N}\bigg)\,.$$

The upper bound k_0 is the smallest integer such that $2+\ell+\frac{k_0+1}{N} \geqslant 3$; clearly $k_0 \leqslant N$. For any $k \leqslant k_0$ we set

$$z_k \;:=\; E_k + \mathrm{i}\eta_k\,, \qquad E_k \;:=\; -2-\kappa_k\,, \qquad \kappa_k \;:=\; \ell + \frac{k}{N}\,, \qquad \eta_k \;:=\; M^{4\gamma}\frac{N}{M^2\sqrt{\kappa_k}}\,.$$

Clearly, $\eta_k \leqslant \kappa_k$ since $M \leqslant N$. On the support of $\phi_{1,k}$ we have $|\lambda_1 - E_k| \leqslant C/N \leqslant \eta_k$, so that we get the lower bound

$$\phi_{1,k}\,\mathrm{Im}\,m_N(z_k) \;=\; \phi_{1,k}\frac{1}{N}\sum_{\alpha=1}^{N}\frac{\eta_k}{(\lambda_\alpha - E_k)^2+\eta_k^2} \;\geqslant\; \phi_{1,k}\frac{1}{N}\frac{\eta_k}{(\lambda_1-E_k)^2+\eta_k^2} \;\geqslant\; \frac{c}{N\eta_k} \tag{7.17}$$

for some positive constant c. On the other hand, by (4.4), we have

$$\mathrm{Im}\,m(z_k) \;\leqslant\; \frac{C\eta_k}{\sqrt{\kappa_k}}\,.$$

Therefore we get

$$\phi_{1,k}\big|\operatorname{Im} m_N(z_k) - \operatorname{Im} m(z_k)\big| \;\geqslant\; \frac{c}{N\eta_k} - \frac{C\eta_k}{\sqrt{\kappa_k}} \;\geqslant\; \frac{c'}{N\eta_k} \tag{7.18}$$

for some positive constant c'. Here in the second step we used that $\eta_k/\sqrt{\kappa_k} \leqslant M^{-\gamma}(N\eta_k)^{-1}$.

Suppose for now that $\delta_+ \geqslant c$. Then by (7.6) we have the upper bound $\widetilde{\eta}_x \leqslant CM^{3\gamma-1}$, uniformly for $|x| \leqslant 10$. Since $\eta_k \geqslant CM^{4\gamma-1}$ we find that $z_k \in \widetilde{\mathbf{S}}$ with $|\operatorname{Re} z_k| \geqslant 2$. Hence (2.20) applies for $z = z_k$ and we get

$$\big|\operatorname{Im} m_N(z_k) - \operatorname{Im} m(z_k)\big| \;\prec\; \frac{1}{M\kappa_k} + \frac{1}{(M\eta_k)^2\sqrt{\kappa_k}} \;\leqslant\; CM^{-\gamma}\frac{1}{N\eta_k}\,. \tag{7.19}$$

Comparing this bound with (7.18) we conclude that $\phi_{1,k} \prec 0$ (i.e. the event $\{\phi_{1,k} = 1\}$ has very small probability). Summing over k yields $\phi_1 \prec 0$. Note that in this proof the stronger bound (2.20) outside of the spectrum was essential; the general bound of order $(M\eta_k)^{-1}$ from (2.19) is not smaller than the right-hand side of (7.18).

The preceding proof of $\phi_1 \prec 0$ assumed the existence of a spectral gap $\delta_+ \geqslant c$. The above argument easily carries over to the case without a gap of constant size, in which case we choose

$$\ell \;:=\; M^{6\gamma}\left(\frac{N^2}{M^{8/3}} + \left(\frac{N}{M^2}\right)^2\left[\delta_+ + \left(\frac{N}{M^2}\right)^{1/7}\right]^{-12}\right),$$

$$E_k \;:=\; -2 - \kappa_k\,, \qquad \kappa_k \;:=\; \ell + \frac{k}{N}\,, \qquad \eta_k \;:=\; M^{4\gamma}\left(\frac{N}{M^2\sqrt{\kappa_k}} + \frac{1}{M(\sqrt{\kappa_k}+\delta_+)^3}\right).$$

The last term in η_k guarantees that $z_k \in \widetilde{\mathbf{S}}$, by (7.7). Then we may repeat the above proof to get $\phi_1 \prec 0$ for the new function ϕ_1.

All that remains to complete the proof of (7.11) and (7.13) is the estimate $\phi_2 \prec 0$. Clearly

$$\mathbb{P}(\lambda_1 \leqslant -3) \;\leqslant\; \mathbb{E}\big|\{j : \lambda_j \leqslant -3\}\big|\,.$$

In part (2) of Lemma 7.2 in [17] it was shown, using the moment method, that the right-hand side is bounded by $CN^{-c\log\log N}$ provided the matrix entries h_{ij} have subexponential decay, i.e.

$$\mathbb{P}(|\zeta_{ij}| \geqslant x^\alpha) \;\leqslant\; \beta \mathrm{e}^{-x} \qquad (x > 0)\,,$$

for some constants α, β (recall the notation (2.5)). In this paper we only assume polynomial decay, (2.6). However, the subexponential decay assumption of [17] was only used in the first truncation step, Equations (7.28)–(7.29) in [17], where a new set of independent random variables $\widehat{h}_{ij}$ was constructed with the properties that

$$\mathbb{P}\big(\zeta_{ij} = \widehat{\zeta}_{ij}\big) \;\geqslant\; 1 - \mathrm{e}^{-n}\,, \qquad \big|\widehat{\zeta}_{ij}\big| \;\leqslant\; n\,, \qquad \mathbb{E}\zeta_{ij} = 0\,, \qquad \mathbb{E}\big|\widehat{\zeta}_{ij}\big|^2 \;\leqslant\; \mathbb{E}|\zeta_{ij}|^2 + \mathrm{e}^{-n} \tag{7.20}$$

for $n = (\log N)(\log\log N)$. Under the condition (2.6) the same truncation can be performed, but the estimates in (7.20) will be somewhat weaker; instead of the exponent $n = (\log N)(\log\log N)$ we get $n = D\log N$ for any fixed $D > 0$. The conclusion of the same proof is that, assuming only (2.6), we have

$$\mathbb{E}\big|\{j : \lambda_j \leqslant -3\}\big| \;\leqslant\; N^{-D} \tag{7.21}$$

for any positive number D and for any $N \geqslant N_0(D)$. This guarantees that $\phi_2\|H\| \prec 0$. Together with the estimate $\phi_1\|H\| \leqslant 3\phi_1 \prec 0$ established above, this completes the proof of Theorem 7.3. □

PROOF OF THEOREM 7.4. The estimate of $\|H\|$ with $X = M^{-1/6}$ follows from the proof of part (2) of Lemma 7.2 in [17], by choosing $k = M^{-1/6-\varepsilon}$ with any small $\varepsilon > 0$ in (7.32) of [17]. This argument can be improved to $X = M^{-1/4}$ by the remark after (7.18) in [17]. Finally, the bound with $X = M^{-2/3}$ under the symmetry condition on the entries of H is proved in Theorem 3.4 of [5]. □

Next, we establish an estimate on the normalized counting function $\mathfrak{n}_N$ defined in (7.1). As above, the exponents are not expected to be optimal, but the estimate is in general sharp if $\delta_+ \geqslant c$.

THEOREM 7.5 (EIGENVALUE COUNTING FUNCTION). *Suppose that $\delta_- \geqslant c$ (so that (7.5) holds). Then*

$$\sup_{E\in\mathbb{R}} |\mathfrak{n}_N(E) - n(E)| = O_\prec(Y), \tag{7.22}$$

where we introduced the control parameter

$$Y := \frac{1}{M}\left(\frac{1}{\delta_+ + M^{-1/5}}\right)^{7/2}. \tag{7.23}$$

PROOF. First we prove the bound (7.22) for any fixed $E \in [-10, 10]$. Define the dyadic energies $E_k := -2 - 2^k(\delta_+ + M^{-1/5})$. By (7.6) we have for all $k \geqslant 0$

$$\max\{\widetilde{\eta}_E : E \in [E_{k+1}, E_k]\} \leqslant \frac{CM^{-1+4\gamma}}{\left[2^k(\delta_+ + M^{-1/5})\right]^{7/2}}.$$

A similar bound holds for $E'_k := -2 + 2^k(\delta_+ + M^{-1/5})$. For any $E \in [-10, 0]$, we express $\mathfrak{n}_N(E) - n(E)$ as a telescopic sum and use (7.4) to get

$$\begin{aligned}
|\mathfrak{n}_N(E) - n(E)| \leqslant\ & |\mathfrak{n}_N(-10) - n(-10)| + \sum_{k\geqslant 0}\Big|\big(\mathfrak{n}_N(E_{k+1}) - \mathfrak{n}_N(E_k)\big) - \big(n(E_{k+1}) - n(E_k)\big)\Big| \\
&+ \sum_{k\geqslant 0}\Big|\big(\mathfrak{n}_N(E'_{k+1}) - \mathfrak{n}_N(E'_k)\big) - \big(n(E'_{k+1}) - n(E'_k)\big)\Big| \\
\prec\ & M^{-1+4\gamma}(\delta_+ + M^{-1/5})^{-7/2}.
\end{aligned} \tag{7.24}$$

Here we used that $n(-10) = 0$ and $\mathfrak{n}_N(-10) \leqslant \mathfrak{n}_N(-3) \prec 0$ by (7.21). In fact, (7.24) easily extends to any $E < -10$. By an analogous dyadic analysis near the upper spectral edge, we also get (7.21) for any $E \geqslant 0$. Since this holds for any $\gamma > 0$, we thus proved

$$|\mathfrak{n}_N(E) - n(E)| \prec Y \tag{7.25}$$

for any fixed $E \in [-10, 10]$.

To prove the statement uniformly in E, we define the *classical location of the α-th eigenvalue* γ_α through

$$\int_{-\infty}^{\gamma_\alpha} \varrho(x)\,\mathrm{d}x = \frac{\alpha}{N}. \tag{7.26}$$

Applying (7.25) for the N energies $E = \gamma_1, \ldots, \gamma_N$, we get

$$\left|\mathfrak{n}_N(\gamma_\alpha) - \frac{\alpha}{N}\right| \prec Y \tag{7.27}$$

uniformly in $\alpha = 1, \dots, N$. Since $\mathfrak{n}_N(E)$ and $n(E)$ are nondecreasing and $Y \geqslash 1/N$, we find

$$\sup\{\mathfrak{n}_N(E) - n(E) : \gamma_{\alpha-1} \leqslant E \leqslant \gamma_\alpha\} \;\leqslant\; \mathfrak{n}_N(\gamma_\alpha) - n(\gamma_{\alpha-1}) \;=\; \mathfrak{n}_N(\gamma_\alpha) - n(\gamma_\alpha) + \frac{1}{N} \;=\; O_\prec(Y)$$

uniformly in $\alpha = 2, 3, \dots$. Below γ_1 we use (7.27) to get

$$\sup_{E \leqslant \gamma_1} \big(\mathfrak{n}_N(E) - n(E)\big) \;\leqslant\; \mathfrak{n}_N(\gamma_1) \;=\; O_\prec(Y)\,.$$

Finally, for any $E \geqslant \gamma_N$, we have $\mathfrak{n}_N(E) - n(E) = \mathfrak{n}_N(E) - 1 \leqslant 0$ deterministically. Thus we have proved

$$\sup_{E \in \mathbb{R}} \big(\mathfrak{n}_N(E) - n(E)\big) \;=\; O_\prec(Y)\,.$$

A similar argument yields $\inf_{E\in\mathbb{R}}\big(\mathfrak{n}_N(E) - n(E)\big) = O_\prec(Y)$. This concludes the proof of Theorem 7.5. □

Next, we derive rigidity bounds on the locations of the eigenvalues. Recall the definition of γ_α from (7.26).

THEOREM 7.6 (EIGENVALUE LOCATIONS). *Suppose that $\delta_- \geqslant c$ (so that (7.5) holds) and that (7.11) and (7.22) hold with some positive control parameters $X, Y \leqslant C$. Define $\widehat{\alpha} := \min\{\alpha, N+1-\alpha\}$ and let $\varepsilon > 0$ be arbitrary. Then*

$$|\lambda_\alpha - \gamma_\alpha| \;\prec\; Y\left(\frac{N}{\widehat{\alpha}}\right)^{1/3} \qquad \text{for} \quad \widehat{\alpha} \geqslant M^\varepsilon NY\,, \tag{7.28}$$

and

$$|\lambda_\alpha - \gamma_\alpha| \;\prec\; X + (M^\varepsilon Y)^{2/3} \qquad \text{for} \quad \widehat{\alpha} \leqslant M^\varepsilon NY\,. \tag{7.29}$$

PROOF. To simplify notation, we assume that $\alpha \leqslant N/2$ so that $\widehat{\alpha} = \alpha$; the other eigenvalues are handled analogously. Without loss of generality we assume that $\lambda_{N/2} \leqslant 1$. Indeed, the condition $\lambda_{N/2} \leqslant 1$ is equivalent to $\mathfrak{n}(1) \geqslant 1/2$, which holds with very high probability by Theorem 7.5 and the fact that $n_{sc}(1) > 1/2$.

The key relation is

$$\frac{\alpha}{N} \;=\; n(\gamma_\alpha) \;=\; \mathfrak{n}_N(\lambda_\alpha) \;=\; n(\lambda_\alpha) + O_\prec(Y), \tag{7.30}$$

where in the last step we used Theorem 7.5. By definition of $n(x)$ we have for $-2 \leqslant x \leqslant 1$ that

$$n(x) \;\asymp\; (2+x)^{3/2} \;\asymp\; \kappa_x^{3/2}\,, \qquad n'(x) \;\asymp\; n(x)^{1/3}\,. \tag{7.31}$$

Hence for $\alpha \leqslant N/2$ we have

$$\gamma_\alpha + 2 \;\asymp\; \left(\frac{\alpha}{N}\right)^{2/3}, \qquad n(\gamma_\alpha) = \frac{\alpha}{N}\,, \qquad n'(\gamma_\alpha) \;\asymp\; \left(\frac{\alpha}{N}\right)^{1/3}. \tag{7.32}$$

Suppose first that $\alpha \geqslant \alpha_0 := M^\varepsilon NY$. Then $n(\gamma_\alpha) \geqslant M^\varepsilon Y$, so that the relation (7.30) implies

$$\big|n(\gamma_\alpha) - n(\lambda_\alpha)\big| \;\prec\; Y \leqslant M^{-\varepsilon} n(\gamma_\alpha)\,,$$

which yields $n(\gamma_\alpha) \asymp n(\lambda_\alpha)$. By (7.31), we we therefore get that $n'(\gamma_\alpha) \asymp n'(\lambda_\alpha)$ as well. Since n' is nondecreasing, we get $n'(x) \asymp n'(\gamma_\alpha) \asymp n'(\lambda_\alpha)$ for any x between γ_α and λ_α. Therefore, by the mean value theorem, we have

$$|\gamma_\alpha - \lambda_\alpha| \;\leqslant\; \frac{C|n(\gamma_\alpha) - n(\lambda_\alpha)|}{n'(\gamma_\alpha)} \;\prec\; Y\biggl(\frac{N}{\alpha}\biggr)^{1/3},$$

where in the last step we used (7.30) and (7.32). This proves (7.28) for $\alpha \geqslant M^\varepsilon NY$.

For the remaining indices, $\alpha < \alpha_0$, we get from (7.30) the upper bound

$$2 + \lambda_\alpha \;\leqslant\; 2 + \lambda_{\alpha_0} \;=\; 2 + \gamma_{\alpha_0} + O_\prec(Y^{2/3}) \;\prec\; (M^\varepsilon Y)^{2/3},$$

where in the second step we used (7.28) and in the last step (7.32). In order to obtain a lower bound, we use Theorem 7.3 to get

$$-(2 + \lambda_\alpha) \;\leqslant\; -(2 + \lambda_1) \;\prec\; X.$$

Similar bounds hold for γ_α as well:

$$0 \;\leqslant\; 2 + \gamma_\alpha \;\leqslant\; 2 + \gamma_{\alpha_0} \;\leqslant\; (M^\varepsilon Y)^{2/3}.$$

Combining these bounds, we obtain

$$|\lambda_\alpha - \gamma_\alpha| \;\prec\; X + (M^\varepsilon Y)^{2/3}.$$

This concludes the proof. □

Finally, we state a trivial corollary of Theorem 7.6.

COROLLARY 7.7. *Suppose that $\delta_- \geqslant c$ and that* (7.11) *and* (7.22) *hold with some positive control parameters $X, Y \leqslant C$. Then*

$$\sum_{\alpha=1}^{N} |\lambda_\alpha - \gamma_\alpha|^2 \;\prec\; NY(Y + X^2).$$

7.1. Local density of states: proof of Lemma 7.1. In this section we prove Lemma 7.1. Define the empirical eigenvalue distribution

$$\varrho_N(x) \;=\; \frac{1}{N} \sum_{\alpha=1}^{N} \delta(x - \lambda_\alpha),$$

so that we may write

$$\mathfrak{n}_N(E) \;=\; \frac{1}{N} \bigl|\{\alpha : \lambda_\alpha \leqslant E\}\bigr| \;=\; \int_{-\infty}^{E} \varrho_N(x)\,\mathrm{d}x, \qquad m_N(z) \;=\; \frac{1}{N} \operatorname{Tr} G(z) \;=\; \int \frac{\varrho_N(x)\,\mathrm{d}x}{x - z}.$$

We introduce the differences

$$\varrho^\Delta \;:=\; \varrho_N - \varrho, \qquad m^\Delta \;:=\; m_N - m.$$

Following [11], we use the Helffer-Sjöstrand functional calculus [4,21]. Introduce $\mathcal{E} := \max\bigl\{E_2 - E_1, \widetilde{\eta}\bigr\}$. Let χ be a smooth cutoff function equal to 1 on $[-\mathcal{E}, \mathcal{E}]$ and vanishing on $[-2\mathcal{E}, 2\mathcal{E}]^c$, such that $|\chi'(y)| \leqslant C\mathcal{E}^{-1}$. Let f be a characteristic function of the interval $[E_1, E_2]$ smoothed on the scale $\widetilde{\eta}$: $f(x) = 1$ on $[E_1 + \widetilde{\eta}, E_2 - \widetilde{\eta}]$, $f(x) = 0$ on $[E_1, E_2]^c$, $|f'(x)| \leqslant C\widetilde{\eta}^{-1}$, and $|f''(x)| \leqslant C\widetilde{\eta}^{-2}$. Note that the supports of f' and f'' have measure $O(\widetilde{\eta})$.

Then we have the estimate (see Equation (B.13) in [11])

$$\left|\int f(\lambda)\,\varrho^{\Delta}(\lambda)\,\mathrm{d}\lambda\right| \;\leqslant\; C\left|\int \mathrm{d}x\int_0^{\infty}\mathrm{d}y\,(f(x)+yf'(x))\,\chi'(y)\,m^{\Delta}(x+\mathrm{i}y)\right| + C\left|\int \mathrm{d}x\int_0^{\widetilde{\eta}}\mathrm{d}y\,f''(x)\chi(y)\,y\,\mathrm{Im}\,m^{\Delta}(x+\mathrm{i}y)\right| + C\left|\int \mathrm{d}x\int_{\widetilde{\eta}}^{\infty}\mathrm{d}y\,f''(x)\chi(y)\,y\,\mathrm{Im}\,m^{\Delta}(x+\mathrm{i}y)\right|. \quad (7.33)$$

Since χ' vanishes away from $[\mathcal{E},2\mathcal{E}]$ and f vanishes away from $[E_1,E_2]$, we may apply (7.2) to get

$$|m_N(x+iy)-m(x+iy)| \;\prec\; \frac{1}{My} \quad (7.34)$$

uniformly for $x\in[E_1,E_2]$ and $y\geqslant\widetilde{\eta}$. Thus the first term on the right-hand side of (7.33) is bounded by

$$\frac{C}{M\mathcal{E}}\int \mathrm{d}x\int_{\mathcal{E}}^{2\mathcal{E}}\mathrm{d}y\,|f(x)+yf'(x)| \;\prec\; \frac{1}{M}\,. \quad (7.35)$$

In order to estimate the two remaining terms of (7.33), we estimate $\mathrm{Im}\,m^{\Delta}(x+\mathrm{i}y)$. If $y\geqslant\widetilde{\eta}$ we may use (7.34). Consider therefore the case $0<y\leqslant\widetilde{\eta}$. From Lemma 4.3 we find

$$|\mathrm{Im}\,m(x+\mathrm{i}y)| \;\leqslant\; C\sqrt{\kappa_x+y}\,. \quad (7.36)$$

By spectral decomposition of H, it is easy to see that the function $y\mapsto y\,\mathrm{Im}\,m_N(x+\mathrm{i}y)$ is monotone increasing. Thus we get, using (7.36), $x+\mathrm{i}\widetilde{\eta}\in\widetilde{\mathbf{S}}$, and (7.2), that

$$y\,\mathrm{Im}\,m_N(x+\mathrm{i}y) \;\leqslant\; \widetilde{\eta}\,\mathrm{Im}\,m_N(x+\mathrm{i}\widetilde{\eta}) \;\prec\; \widetilde{\eta}\left(\sqrt{\kappa_x+\widetilde{\eta}}+\frac{1}{M\widetilde{\eta}}\right) \;\prec\; \widetilde{\eta}\sqrt{\kappa_x+\widetilde{\eta}}+\frac{1}{M}\,, \quad (7.37)$$

for $y\leqslant\widetilde{\eta}$ and $x\in[E_1,E_2]$. Using $m^{\Delta}=m_N-m$ and recalling (7.36), we therefore get

$$|y\,\mathrm{Im}\,m^{\Delta}(x+\mathrm{i}y)| \;\prec\; \widetilde{\eta}\sqrt{\kappa_x+\widetilde{\eta}}+\frac{1}{M}\,, \quad (7.38)$$

for $y\leqslant\widetilde{\eta}$ and $x\in[E_1,E_2]$. The second term of (7.33) is therefore bounded by

$$\left(\widetilde{\eta}\sqrt{\kappa_x+\widetilde{\eta}}+\frac{1}{M}\right)\int \mathrm{d}x\,|f''(x)|\int_0^{\widetilde{\eta}}\mathrm{d}y\,\chi(y) \;\leqslant\; \widetilde{\eta}\sqrt{\kappa_x+\widetilde{\eta}}+\frac{1}{M}\,.$$

In order to estimate the third term on the right-hand side of (7.33), we integrate by parts, first in x and then in y, to obtain the bound

$$C\left|\int \mathrm{d}x\,f'(x)\,\widetilde{\eta}\,\mathrm{Re}\,m^{\Delta}(x+\mathrm{i}\widetilde{\eta})\right| + C\left|\int \mathrm{d}x\int_{\widetilde{\eta}}^{\infty}\mathrm{d}y\,f'(x)\chi'(y)y\,\mathrm{Re}\,m^{\Delta}(x+\mathrm{i}y)\right| + C\left|\int \mathrm{d}x\int_{\widetilde{\eta}}^{\infty}\mathrm{d}y\,f'(x)\chi(y)\,\mathrm{Re}\,m^{\Delta}(x+\mathrm{i}y)\right|. \quad (7.39)$$

The second term of (7.39) is similar to the first term on the right-hand side of (7.33), and is easily seen to be bounded by $1/M$ as in (7.35).

In order to bound the first and third terms of (7.39), we estimate, for any $y \leqslant \widetilde{\eta}$,

$$\big|m^{\Delta}(x+\mathrm{i}y)\big| \;\leqslant\; \big|m^{\Delta}(x+\mathrm{i}\widetilde{\eta})\big| + \int_y^{\widetilde{\eta}} \mathrm{d}u \left(\big|\partial_u m_N(x+\mathrm{i}u)\big| + \big|\partial_u m(x+\mathrm{i}u)\big| \right). \tag{7.40}$$

Moreover, using the monotonicity of $y \mapsto y \operatorname{Im} m_N(x+\mathrm{i}y)$ and the identity $\sum_j |G_{ij}|^2 = \eta^{-1} \operatorname{Im} G_{ii}$, we find for any $u \leqslant \widetilde{\eta}$ that

$$\big|\partial_u m_N(x+\mathrm{i}u)\big| \;=\; \left|\frac{1}{N} \operatorname{Tr} G^2(x+\mathrm{i}u)\right| \;\leqslant\; \frac{1}{N}\sum_{i,j} \big|G_{ij}(x+\mathrm{i}u)\big|^2 \;=\; \frac{1}{u} \operatorname{Im} m_N(x+\mathrm{i}u) \;\leqslant\; \frac{1}{u^2}\, \widetilde{\eta} \operatorname{Im} m_N(x+\mathrm{i}\widetilde{\eta})\,.$$

Similarly, we find from (2.7) that

$$\big|\partial_u m(x+\mathrm{i}u)\big| \;\leqslant\; \frac{1}{u^2}\, \widetilde{\eta} \operatorname{Im} m(x+\mathrm{i}\widetilde{\eta}) \leqslant \frac{C\widetilde{\eta}}{u^2} \qquad (u \leqslant \widetilde{\eta}).$$

Thus (7.40) and (7.34) yield

$$\big|m^{\Delta}(x+\mathrm{i}y)\big| \;\prec\; \frac{1}{M\widetilde{\eta}} + \int_y^{\widetilde{\eta}} \mathrm{d}u\, \frac{\widetilde{\eta}}{u^2}\Big(1 + \frac{1}{M\widetilde{\eta}}\Big) \;\prec\; \frac{\widetilde{\eta}}{y} \qquad (y \leqslant \widetilde{\eta})\,, \tag{7.41}$$

where we also used that $\widetilde{\eta} \geqslant M^{-1}$. Using (7.41) for $y = \widetilde{\eta}$, we may now estimate the first term of (7.39) by $\widetilde{\eta}$.

What remains is the third term of (7.39), which can be estimated, using (7.34), by

$$\int \mathrm{d}x \int_{\widetilde{\eta}}^{2\mathcal{E}} \mathrm{d}y\, |f'(x)| \frac{1}{My} \;\leqslant\; CM^{-1}(1 + |\log \widetilde{\eta}|) \;\leqslant\; CM^{-1} \log M\,.$$

Summarizing, we have proved that

$$\left|\int f(\lambda)\, \varrho^{\Delta}(\lambda)\, \mathrm{d}\lambda\right| \;\prec\; \frac{1}{M} + \widetilde{\eta}\sqrt{\kappa_x + \widetilde{\eta}} + \widetilde{\eta} + \frac{\log M}{M} \;\prec\; \widetilde{\eta} + \frac{1}{M}. \tag{7.42}$$

Since $\operatorname{Im} m_N(x+\mathrm{i}\widetilde{\eta})$ controls the local density on scale $\widetilde{\eta}$, we may estimate $|\mathfrak{n}_N(E) - n(E)|$ using (7.37) according to

$$|\mathfrak{n}_N(x+\widetilde{\eta}) - \mathfrak{n}_N(x-\widetilde{\eta})| \;\leqslant\; C\widetilde{\eta} \operatorname{Im} m_N(x+\mathrm{i}\widetilde{\eta}) \prec \widetilde{\eta}\sqrt{\kappa_x + \widetilde{\eta}} + \frac{1}{M}.$$

Thus we get

$$\left|\mathfrak{n}_N(E_1) - \mathfrak{n}_N(E_2) - \int f(\lambda)\, \varrho_N(\lambda)\, \mathrm{d}\lambda\right| \;\leqslant\; C \sum_{i=1,2} \big(\mathfrak{n}(E_i + \widetilde{\eta}) - \mathfrak{n}(E_i - \widetilde{\eta})\big) \;\prec\; \widetilde{\eta}\sqrt{\kappa_x + \widetilde{\eta}} + \frac{1}{M}\,.$$

Similarly, since ϱ has a bounded density, we find

$$\left|n(E_1) - n(E_2) - \int f(\lambda)\, \varrho(\lambda)\, \mathrm{d}\lambda\right| \;\leqslant\; C\widetilde{\eta}\,.$$

Together with (7.42) and recalling $\widetilde{\eta} \geqslant M^{-1}$, we therefore get (7.4). This concludes the proof of Lemma 7.1.

8. Bulk universality

Local eigenvalue statistics are described by correlation functions on the scale $1/N$. Fix an integer $n \geqslant 2$ and an energy $E \in (-2, 2)$. Abbreviating $\mathbf{x} = (x_1, x_2, \dots x_n)$, we define the *local correlation function*

$$f_N^{(n)}(E, \mathbf{x}) \;:=\; \frac{1}{\varrho(E)^n} p_N^{(n)}\left(E + \frac{x_1}{N\varrho(E)}, E + \frac{x_2}{N\varrho(E)}, \dots, E + \frac{x_n}{N\varrho(E)}\right), \tag{8.1}$$

where $p_N^{(n)}$ is the n-point correlation function of the N eigenvalues and $\varrho(E)$ is the density of the semicircle law defined in (2.7). Universality of the local eigenvalue statistics means that, for any fixed n, the limit as $N \to \infty$ of the local correlation function $f_N^{(n)}$ only depends on the symmetry class of the matrix entries, and is otherwise independent of their distribution. In particular, the limit of $f_N^{(n)}$ coincides with that of a GOE or GUE matrix, which is explicitly known. In this paper, we consider local correlation functions averaged over a small energy interval of size $\ell = N^{-\varepsilon}$,

$$\widetilde{f}_N^{(n)}(E, \mathbf{x}) \;:=\; \frac{1}{2\ell} \int_{E-\ell}^{E+\ell} f_N^{(n)}(E', \mathbf{x})\, \mathrm{d}E' \,. \tag{8.2}$$

Universality is understood in the sense of the weak limit, as $N \to \infty$ for fixed $|E| < 2$, of $\widetilde{f}_N^{(n)}(E, \mathbf{x})$ in the variables $\mathbf{x}$.

The general approach developed in [14, 15, 17] to prove the universality of the local eigenvalue statistics in the bulk spectrum of a general Wigner-type matrix consists of three steps.

(i) A rigidity estimate on the locations of the eigenvalues, in the sense of a quadratic mean.

(ii) The spectral universality for matrices with a small Gaussian component, via local ergodicity of the Dyson Brownian motion (DBM).

(iii) A perturbation argument that removes the small Gaussian component by comparing Green functions.

In this paper we do not give the details of steps (ii) and (iii), since they have been concisely presented elsewhere, e.g. in [16]. Here we only summarize the results and the key arguments of steps (ii) and (iii) for the general class of matrices we consider. In this section we assume that H is either real symmetric or complex Hermitian. The former case means that the entries of H are real. The latter means, loosely, that its off-diagonal entries have a nontrivial imaginary part. More precisely, in the complex Hermitian case we shall replace the lower bound on the variances s_{ij} from Definition 3.1 with the following, stronger, condition.

Definition 8.1. *We call the Hermitian matrix H a* complex a-full Wigner matrix *if for each i, j the 2×2 covariance matrix*

$$\sigma_{ij} \;=\; \begin{pmatrix} \mathbb{E}(\operatorname{Re} h_{ij})^2 & \mathbb{E}(\operatorname{Re} h_{ij})(\operatorname{Im} h_{ij}) \\ \mathbb{E}(\operatorname{Re} h_{ij})(\operatorname{Im} h_{ij}) & \mathbb{E}(\operatorname{Im} h_{ij})^2 \end{pmatrix}$$

satisfies

$$\sigma \;\geqslant\; \frac{a}{N}$$

as a symmetric matrix. Note that this condition implies that H is a-full, but the converse is not true.

We consider a stochastic flow of Wigner-type matrices generated by the Ornstein-Uhlenbeck equation

$$\mathrm{d}H_t \;=\; \frac{1}{\sqrt{N}}\,\mathrm{d}B_t - \frac{1}{2}H_t\mathrm{d}t$$

with some given initial matrix H_0. Here B is an $N \times N$ matrix-valued standard Brownian motion with the same symmetry type as H. The resulting dynamics on the level of the eigenvalues is Dyson Brownian motion (DBM). It is well known that H_t has the same distribution as the matrix

$$\mathrm{e}^{-t/2}H_0 + (1-\mathrm{e}^{-t})^{1/2}U\,, \tag{8.3}$$

where U is an independent standard Gaussian Wigner matrix of the same symmetry class as H. In particular, H_t converges to U as $t \to \infty$. The eigenvalue distribution converges to the Gaussian equilibrium measure, whose density is explicitly given by

$$\mu(\boldsymbol{\lambda}) \;=\; \frac{1}{Z}\,\mathrm{e}^{-\beta N\mathcal{H}(\boldsymbol{\lambda})}\mathrm{d}\boldsymbol{\lambda}\,, \qquad \mathcal{H}(\boldsymbol{\lambda}) \;:=\; \sum_{i=1}^{N}\frac{\lambda_i^2}{4} - \frac{1}{N}\sum_{i<j}\log|\lambda_i-\lambda_j|\,;$$

here $\beta = 1$ for the real symmetric case (GOE) and $\beta = 2$ for the complex Hermitian case (GUE).

The matrix $S^{(t)}$ of variances of H_t is given by

$$S^{(t)} \;=\; \mathrm{e}^{-t}S^{(0)} + (1-\mathrm{e}^{-t})\mathbf{e}\mathbf{e}^*,$$

where $S^{(0)}$ is the matrix of variances of H_0. It is easy to see that the gaps $\delta_\pm(t)$ of $S^{(t)}$ satisfy $\delta_\pm(t) \geqslant \delta_\pm(0)$; therefore the corresponding parameters (2.11) satisfy $\widetilde{\Gamma}_t(z) \leqslant \widetilde{\Gamma}_0(z)$. Since all estimates behind our main theorems in Sections 2 and 7 improve if $\delta_\pm$ increase, it is immediate that all results in these sections hold for H_t provided they hold for H_0.

The key quantity to be controlled when establishing bulk universality is the mean quadratic distance of the eigenvalues from their classical locations,

$$Q \;:=\; \max_{t\geqslant 0}\,\mathbb{E}^{(t)}\frac{1}{N}\sum_i(\lambda_i-\gamma_i)^2\,, \tag{8.4}$$

where $\mathbb{E}^{(t)}$ denotes the expectation with respect to the ensemble H_t. By Corollary 7.7 we have

$$Q \;\leqslant\; N^{\varepsilon}Y(Y+X^2)$$

for any $\varepsilon > 0$ and $N \geqslant N_0(\varepsilon)$. Here we used that the estimate from Corollary 7.7 is uniform in t, by the remark in the previous paragraph.

We modify the original DBM by adding a local relaxation term of the form $\frac{1}{2\tau}\sum_i(\lambda_i-\gamma_i)^2$ to the original Hamiltonian $\mathcal{H}$, which has the effect of artificially speeding up the relaxation of the dynamics. Here $\tau \ll 1$ is a small parameter, the relaxation time of the modified dynamics. We choose $\tau := N^{1+4\varepsilon}Q$ for some $\varepsilon > 0$. As Theorem 4.1 of [15] (see also Theorem 2.2 of [16]) shows, the local statistics of the eigenvalue gaps of H_t and GUE/GOE coincide if $t \geqslant N^{\varepsilon}\tau = N^{1+4\varepsilon}Q$, i.e. if

$$t \;\geqslant\; N^{1+5\varepsilon}Y(Y+X^2)\,. \tag{8.5}$$

The local statistics are averaged over $N^{1-\varepsilon}$ consecutive eigenvalues or, alternatively, in the energy parameter E over an interval of length $N^{-\varepsilon}$.

To complete the programme (i)–(iii), we need to compare the local statistics of the original ensemble H and H_t, i.e. perform step (iii). We first recall the Green function comparison theorem from [17] for the case $M \asymp N$ (generalized Wigner). The result states, roughly, that expectations of Green functions with spectral parameter z satisfying $\operatorname{Im} z \geqslant N^{-1-\varepsilon}$ are determined by the first four moments of the single-entry distributions. Therefore the local eigenvalue statistics on a very small scale, $\eta = N^{-1-\varepsilon}$, of two Wigner ensembles are indistinguishable if the first four moments of their matrix entries match. More precisely, for the local n-point correlation functions (8.1) to match, one needs to compare expectations of n-th order monomials of the form

$$\prod_{k=1}^{n} m_N(E_k + i\eta)\,, \tag{8.6}$$

where the energies E_k are chosen in the bulk spectrum with $E_k - E_{k'} = O(1/N)$. (Recall that $m_N(z) = \frac{1}{N}\operatorname{Tr} G(z)$.)

The proof uses a Lindeberg-type replacement strategy to change the distribution of each matrix entry h_{ij} one by one in a telescopic sum. The idea of applying Lindeberg's method in random matrices was recently used by Chatterjee [3] for comparing the traces of the Green functions; the idea was also used by Tao and Vu [29] in the context of comparing individual eigenvalue distributions. The error resulting from each replacement is estimated using a fourth order resolvent expansion, where all resolvents $G(z) = (H-z)^{-1}$ with $z = E_k + i\eta$ appearing in (8.6) are expanded with respect to the single matrix entry h_{ij} (and its conjugate $h_{ji} = \bar{h}_{ij}$). If the first four moments of the two distributions match, then the terms of at most fourth order in this expansion remain unchanged by each replacement. The error term is of order $\mathbb{E}|h_{ij}|^5 \asymp N^{-5/2}$, which is negligible even after summing up all N^2 pairs of indices (i,j). This estimate assumes that the resolvent entries in the expansion (and hence all factors $m_N(z)$ in (8.6)) are essentially bounded.

The Green function comparison method therefore has two main ingredients. First, a high probability apriori estimate is needed on the resolvent entries at any spectral parameter z with imaginary part η slightly below $1/N$:

$$\max_{i,j} |G_{ij}(E+i\eta)| \;\prec\; N^{2\varepsilon} \qquad (\eta \geqslant N^{-1-\varepsilon}) \tag{8.7}$$

for any small $\varepsilon > 0$. Clearly, the same estimate also holds for $m_N(E+i\eta)$. The bound (8.7) is typically obtained from the local semicircle law for the resolvent entries, (2.18). Although the local semicircle law is effective only for $\operatorname{Im} z \gg 1/N$, it still gives an almost optimal bound for a somewhat smaller η by using the trivial estimate

$$\max_{i,j} |G_{ij}(E+i\eta)| \;\leqslant\; \log N \left(\frac{\eta'}{\eta}\right) \sup_{\eta'' \geqslant \eta'} \max_i \operatorname{Im} G_{ii}(E+i\eta'') \qquad (\eta \leqslant \eta') \tag{8.8}$$

with the choice of $\eta' = N^{-1+\varepsilon}$. The proof of (8.8) follows from a simple dyadic decomposition; see the proof of Theorem 2.3 in Section 8 of [17] for details.

The second ingredient is the construction of an initial ensemble H_0 whose time evolution H_t for some $t \leqslant 1$ satisfying (8.5) is close to H; here closeness is measured by the *matching of moments* of the matrix entries between the ensembles H and H_t. We shall choose H_0, with variance matrix $S^{(0)}$, so that the second moments of H and H_t match,

$$S \;=\; \mathrm{e}^{-t} S^{(0)} + (1 - \mathrm{e}^{-t}) \mathbf{e}\mathbf{e}^*\,, \tag{8.9}$$

and the third and fourth moments are close. We remark that the matching of higher moments was introduced in the work of [29], while the idea of approximating a general matrix ensemble by an appropriate Gussian one appeared earlier in [10]. They have to be so close that even after multiplication with at most five resolvent

entries and summing up for all i, j indices, their difference is still small. (Five resolvent entries appear in the fourth order of the resolvent expansion of G.) Thus, given (8.7), we require that

$$\max_{i,j} \left| \mathbb{E} h_{ij}^s - \mathbb{E}^{(t)} h_{ij}^s \right| \;\leqslant\; N^{-2-(2n+9)\varepsilon} \qquad (s = 3, 4) \tag{8.10}$$

to ensure that the expectations of the n-fold product in (8.6) are close. This formulation holds for the real symmetric case; in the complex Hermitian case all moments of order $s = 3, 4$ involving the real and imaginary parts of h_{ij} have to be approximated. To simplify notation, we work with the real symmetric case in the sequel.

The matching can be done in two steps. In the first we construct a matrix of variances $S^{(0)}$ such that (8.9) holds. This first step is possible if, given S associated with H, (8.9) can be satisfied for a doubly stochastic $S^{(0)}$, i.e. if H is an a-full Wigner matrix and

$$a \;\geqslant\; Ct \tag{8.11}$$

with some large constant C. For the complex Hermitian case, the condition (8.11) is the same but H has to be complex a-full Wigner matrix (see Definition 8.1).

In the second step of moment matching, we use Lemma 3.4 of [18] to construct an ensemble H_0 with variances $S^{(0)}$, such that the entries of H and H_t satisfy

$$\mathbb{E} h_{ij} \;=\; \mathbb{E}^{(t)} h_{ij} \;=\; 0\,, \qquad \mathbb{E} h_{ij}^2 \;=\; \mathbb{E}^{(t)} h_{ij}^2 \;=\; s_{ij}\,, \qquad \mathbb{E} h_{ij}^3 \;=\; \mathbb{E}^{(t)} h_{ij}^3\,, \qquad \left| \mathbb{E} h_{ij}^4 - \mathbb{E}^{(t)} h_{ij}^4 \right| \;\leqslant\; Cts_{ij}^2\,.$$

This means that (8.10) holds if

$$Cts_{ij}^2 \;\leqslant\; N^{-2-(2n+9)\varepsilon}\,.$$

Suppose that H is b-flat, i.e. that $s_{ij} \leqslant b/N$. Then this condition holds provided

$$Ctb^2 \;\leqslant\; N^{-(2n+9)\varepsilon}\,. \tag{8.12}$$

The argument so far assumed that $M \asymp N$ (H is a generalized Wigner matrix), in which case $G_{ij}(E + i\eta')$ remains essentially bounded down to the scale $\eta' \approx 1/N$. If $M \ll N$, then (2.18) provides control only down to scale $\eta' \gg 1/M$ and (8.8) gives only the weaker bound

$$|G_{ij}(E + \mathrm{i}\eta)| \;\prec\; \frac{1}{M\eta}\,, \tag{8.13}$$

for any $\eta \leqslant 1/M$, which replaces (8.7). Using this weaker bound, the condition (8.12) is replaced with

$$Ctb^2 \;\prec\; (M\eta)^{n+4}\,, \tag{8.14}$$

which is needed for n-fold products of the form (8.6) to be close. (For convenience, here we use the notation $A_N \prec B_N$ even for deterministic quantities to indicate that $A_N \leqslant N^{\varepsilon} B_N$ for any $\varepsilon > 0$ and $N \geqslant N_0(\varepsilon)$.) The bound (8.14) thus guarantees that, for any fixed n, the expectations of the n-fold products of the form (8.6) with respect to the ensembles H and H_t are close. Following the argument in the proof of Theorem 6.4 of [17], this means that for any smooth, compactly supported function $O : \mathbb{R}^n \to \mathbb{R}$, the expectations of observables

$$\sum_{i_1 \neq i_2 \neq \ldots \neq i_n} O_\eta\Big(N(\lambda_{i_1} - E), N(\lambda_{i_2} - E), \ldots, N(\lambda_{i_n} - E) \Big) \tag{8.15}$$

are close, where the smeared out observable O_η on scale η is defined through

$$O_\eta(\beta_1,\dots,\beta_n) \;:=\; \frac{1}{(\pi N)^n}\int_{\mathbb{R}^n} \mathrm{d}\alpha_1\cdots\mathrm{d}\alpha_n\, O(\alpha_1,\dots,\alpha_n)\prod_{j=1}^n \theta_\eta\Big(\frac{\beta_j-\alpha_j}{N}\Big)\,, \qquad \theta_\eta(x) \;:=\; \operatorname{Im}\frac{1}{x-i\eta}\,.$$

To conclude the result for observables with O instead of O_η in (8.15), we need to estimate, for both ensembles, the difference

$$\mathbb{E}\sum_{i_1\neq i_2\neq\ldots\neq i_n}(O-O_\eta)\Big(N(\lambda_{i_1}-E),N(\lambda_{i_2}-E),\dots,N(\lambda_{i_n}-E)\Big)\,. \tag{8.16}$$

Due to the smoothness of O, we can decompose $O-O_\eta = Q_1+Q_2$, where

$$|Q_1(\beta_1,\dots,\beta_n)| \;\leqslant\; CN\eta\prod_{j=1}^n \mathbf{1}(|\beta_j|\leqslant K)$$

and

$$|Q_2(\beta_1,\dots,\beta_n)| \;\leqslant\; C\sum_{j=1}^n \mathbf{1}(|\beta_j|\geqslant K)\prod_{j=1}^n \frac{1}{1+\beta_j^2}\,,$$

with an arbitratry parameter $K\gg N/M$. Here the constants depend on O. The contribution from Q_1 to (8.16) can thus be estimated by

$$\mathbb{E}\sum_{i_1\neq i_2\neq\ldots\neq i_n} Q_1\Big(\dots\Big) \;\prec\; CN\eta K^n\,,$$

where we used that the expected number of eigenvalues in the interval $[E-K/N,E+K/N]$ is $O_\prec(K)$, since (8.13) guarantees that the density is bounded on scales larger than $1/M$. The contribution from Q_2 to (8.16) is estimated by

$$\mathbb{E}\sum_{i_1\neq i_2\neq\ldots\neq i_n} Q_2\Big(\dots\Big) \;\prec\; CK^{-1}\Big(\frac{N}{M}\Big)^n\,. \tag{8.17}$$

In the last step we used (8.13) to estimate

$$\sum_{k=1}^N \frac{1}{1+N^2(\lambda_k-E)^2} \;=\; \frac{1}{N}\operatorname{Im}\operatorname{Tr}G\Big(E+\frac{\mathrm{i}}{N}\Big) \;\prec\; \frac{N}{M}\,. \tag{8.18}$$

Optimizing the choice of K and η, (8.14) becomes

$$Ctb^2 \;\prec\; \Big(\frac{M}{N}\Big)^{(n^2+1)(n+4)}\,. \tag{8.19}$$

Summarizing the conditions (8.5), (8.11), and (8.19), we require that

$$N^{1+5\varepsilon}Y(Y+X^2) \;\prec\; \min\Big\{a, b^{-2}\Big(\frac{M}{N}\Big)^{(n^2+1)(n+4)}\Big\}$$

in order to have bulk universality. We have therefore proved the following result.

THEOREM 8.2. *Suppose that H is N/M-flat and a-full (in the real symmetric case) or complex a-full (in the complex Hermitian case). Suppose moreover that* (7.11) *and* (7.22) *hold with some positive control parameters $X, Y \leqslant C$. Fix an arbitrary positive parameter $\varepsilon > 0$. Then the local n-point correlation functions of H, averaged over the energy parameter in an interval of size $N^{-\varepsilon}$ around $|E| < 2$ (see* (8.2)*), coincide with those of GOE or GUE provided that*

$$N^{1+6\varepsilon} Y(Y + X^2) \;\leqslant\; \min\left\{ a, \left(\frac{M}{N}\right)^{(n^2+1)(n+4)+2} \right\}. \tag{8.20}$$

In particular, if $N^{3/4} \leqslant M \leqslant N$ then (7.11) *and* (7.22) *hold with X and Y defined in* (7.12) *and* (7.23)*.*

We conclude with a few examples illustrating Theorem 8.2.

COROLLARY 8.3. *Fix an integer $n \geqslant 2$. There exists a positive number $p(n) \geqslant cn^{-3}$ with the following property. Suppose that H satisfies* ***any*** *of the following conditions for some sufficiently small $\xi > 0$.*

(i) $cN^{-1-\xi} \leqslant s_{ij} \leqslant CN^{-1+p(n)-\xi}$.

(ii) $cN^{-\frac{9}{8}+\xi} \leqslant s_{ij} \leqslant CN^{-1}$.

(iii) H is a one-dimensional band matrix with band width W with a mean-field component of size ν (see Definition 3.3) such that $W \geqslant N^{1-p(n)+\xi}$ and $\nu \geqslant N^{15+\xi} W^{-16}$.

Then there exists an $\varepsilon > 0$ (depending on ξ and n) such that the local n-point correlation functions of H, averaged over the energy parameter in an interval of size $N^{-\varepsilon}$ around $|E| < 2$, coincide with those of GOE or GUE (depending on the symmetry class of H).

We remark that the conditions for the upper bound on s_{ij} in parts (i) and (iii) are similar. But the band structure in (iii) allows one to choose a much smaller mean-field component than in (i).

PROOF. In Case (i), we have $a = cN^{-\xi}$ and $b = N/M$ in Definition 3.1; hence $\delta_\pm \geqslant cN^{-\xi}$ by Proposition A.3. Therefore $Y = M^{-1}N^{-7\xi/2}$ and $X = N^2 M^{-8/3}$ from (7.12) and (7.23), so that (8.20) reads

$$\frac{N}{M}\Big(\frac{1}{M} + \frac{N^4}{M^{16/3}}\Big) \;\leqslant\; N^{-(1+6\varepsilon)} N^{7\xi} \min\left\{ N^{-\xi}, \Big(\frac{M}{N}\Big)^{(n^2+1)(n+4)+2} \right\}.$$

By Theorem 8.2 bulk universality therefore holds provided that $M \geqslant N^{1-p(n)+\xi}$ with any sufficiently small positive $\xi > 0$ (and ε chosen appropriately, depending on ξ and n). The function $p(n)$ can be easily computed.

We remark that if we additionally assume that h_{ij} has a symmetric law with subexponential decay (7.14), then by Theorem 7.4 we can use the improved control parameter $X = M^{-2/3}$. This yields a better threshold $p(n)$. For example, for $n = 2$ we obtain $p(n) = \frac{1}{34}$.

In Case (ii) we take $M = N$, i.e. $b = c$ and $\delta_+ \geqslant a = N^{-1/8+\xi}$. Then with the choice (7.12) and (7.23) we have $Y \leqslant CN^{-1}\delta_+^{-7/2}$, $X \leqslant CN^{-2/3} + CN^{-2}(\delta_+ + N^{-1/7})^{-12}$, so that (8.20) reads

$$\delta_+^{-7/2}\Big(N^{-1}\delta_+^{-7/2} + N^{-4/3} + N^{-4}(\delta_+ + N^{-1/7})^{-24}\Big) \;\ll\; a\,,$$

which holds since $\delta_+ \geqslant a \geqslant N^{-1/8}$.

Finally, in Case (iii) we have $W \asymp M$, $b = N/M$, $a = \nu$, $\delta_+ \geqslant c\nu + c(M/N)^2$, and $\delta_- \geqslant c$. Since $M \geqslant N^{22/23}$ we have $\delta_+ \geqslant cM^{-1/5}$, Thus, with the choice (7.12) and (7.23), we have

$$Y \asymp \frac{1}{M\delta_+^{7/2}} \leqslant C\frac{N^7}{M^8}, \qquad X \leqslant C\frac{N^2}{M^{8/3}} + C\frac{N^{26}}{M^{28}} \asymp \frac{N^{26}}{M^{28}},$$

and (8.20) reads

$$\frac{N^8}{M^8}\Big(\frac{N^7}{M^8} + \frac{N^{52}}{M^{56}}\Big) \ll \min\Big\{\nu, \Big(\frac{M}{N}\Big)^{(n^2+1)(n+4)+2}\Big\}.$$

This leads to the conditions

$$\nu \gg \frac{N^{15}}{M^{16}}, \qquad M \gg N^{1-p(n)}, \tag{8.21}$$

with some positive $p(n)$, which concludes the proof. □

A. Behaviour of Γ and $\widetilde{\Gamma}$

In this section we give basic bounds on the parameters Γ and $\widetilde{\Gamma}$. As it turns out, their behaviour is intimately linked with the spectrum of S, more precisely with its spectral gaps. Recall that the spectrum of S lies in $[-1, 1]$, with 1 being a simple eigenvalue.

Definition A.1. *Let δ_- be the distance from -1 to the spectrum of S, and δ_+ the distance from 1 to the spectrum of S restricted to $\mathbf{e}^\perp$. In other words, $\delta_\pm$ are the largest numbers satisfying*

$$S \geqslant -1 + \delta_-, \qquad S\big|_{\mathbf{e}^\perp} \leqslant 1 - \delta_+.$$

The following proposition gives explicit bounds on Γ and $\widetilde{\Gamma}$ depending on the spectral gaps $\delta_\pm$. We recall the notations $z = E + \mathrm{i}\eta$, $\kappa := \big||E| - 2\big|$ and the definition of θ from (3.2).

Proposition A.2. *There is a universal constant C such that the following holds uniformly in the domain $\big\{z = E + i\eta : |E| \leqslant 10,\ M^{-1} \leqslant \eta \leqslant 10\big\}$, and in particular in any spectral domain $\mathbf{D}$.*

(i) We have the estimate

$$\frac{1}{C\sqrt{\kappa+\eta}} \leqslant \Gamma(z) \leqslant \frac{C\log N}{1 - \max_\pm\big|\frac{1\pm m^2}{2}\big|} \leqslant \frac{C\log N}{\min\{\eta + E^2, \theta\}}. \tag{A.1}$$

(ii) In the presence of a gap δ_- we may improve the upper bound to

$$\Gamma(z) \leqslant \frac{C\log N}{\min\{\delta_- + \eta + E^2, \theta\}}. \tag{A.2}$$

(iii) For $\widetilde{\Gamma}$ we have the bounds

$$C^{-1} \leqslant \widetilde{\Gamma}(z) \leqslant \frac{C\log N}{\min\{\delta_- + \eta + E^2, \delta_+ + \theta\}}. \tag{A.3}$$

PROOF. The first bound of (A.1) follows from $(1-m^2S)^{-1}\mathbf{e} = (1-m^2)^{-1}\mathbf{e}$ combined with (4.3). In order to prove the second bound of (A.1), we write

$$\frac{1}{1-m^2S} \;=\; \frac{1}{2}\frac{1}{1-\frac{1+m^2S}{2}}$$

and observe that

$$\left\|\frac{1+m^2S}{2}\right\|_{\ell^2\to\ell^2} \;\leqslant\; \max_{\pm}\left|\frac{1\pm m^2}{2}\right| \;=:\; q\,. \tag{A.4}$$

Therefore

$$\begin{aligned}\left\|\frac{1}{1-m^2S}\right\|_{\ell^\infty\to\ell^\infty} &\leqslant \sum_{n=0}^{n_0-1}\left\|\frac{1+m^2S}{2}\right\|_{\ell^\infty\to\ell^\infty}^n + \sqrt{N}\sum_{n=n_0}^{\infty}\left\|\frac{1+m^2S}{2}\right\|_{\ell^2\to\ell^2}^n \\ &\leqslant\; n_0 + \sqrt{N}\frac{q^{n_0}}{1-q} \\ &\leqslant\; \frac{C\log N}{1-q}\,,\end{aligned}$$

where in the last step we chose $n_0 = \frac{C_0\log N}{1-q}$ for large enough C_0. Here we used that $\|S\|_{\ell^\infty\to\ell^\infty}\leqslant 1$ and (4.2) to estimate the summands in the first sum. This concludes the proof of the second bound of (A.1). The third bound of (A.1) follows from the elementary estimates

$$\left|\frac{1-m^2}{2}\right| \;\leqslant\; 1-c(\eta+E^2)\,, \qquad \left|\frac{1+m^2}{2}\right| \;\leqslant\; 1-c\bigg((\operatorname{Im} m)^2+\frac{\eta}{\operatorname{Im} m+\eta}\bigg) \;\leqslant\; 1-c\theta \tag{A.5}$$

for some universal constant $c>0$, where in the last step we used Lemma 4.3.

The estimate (A.2) follows similarly. Due to the gap δ_- in the spectrum of S, we may replace the estimate (A.4) with

$$\left\|\frac{1+m^2S}{2}\right\|_{\ell^2\to\ell^2} \;\leqslant\; \max\left\{1-\delta_- -\eta-E^2\,,\,\left|\frac{1+m^2}{2}\right|\right\}. \tag{A.6}$$

Hence (A.2) follows using (A.5).

The lower bound of (A.3) was proved in (4.5). The upper bound is proved similarly to (A.2), except that (A.6) is replaced with

$$\left\|\frac{1+m^2S}{2}\Big|_{\mathbf{e}^\perp}\right\|_{\ell^2\to\ell^2} \;\leqslant\; \max\left\{1-\delta_- -\eta-E^2\,,\,\min\left\{1-\delta_+\,,\,\left|\frac{1+m^2}{2}\right|\right\}\right\}.$$

This concludes the proof of (A.3). □

The following proposition gives the behaviour of the spectral gaps $\delta_\pm$ for the example matrices from Section 3.

PROPOSITION A.3 (SPECTRUM OF S FOR EXAMPLE MATRICES). *(i) If H is an a-full Wigner matrix then $\delta_-\geqslant a$ and $\delta_+\geqslant a$.*

(ii) If H is a band matrix there is a positive constant c, depending on the dimension d and the profile function f, such that $\delta_-\geqslant c$ and $\delta_+\geqslant c(W/L)^2$.

(iii) If $H = \sqrt{1-\nu}H_B + \sqrt{\nu}H_W$, where H_B is a band matrix, H_W is an a-full Wigner matrix independent of H_B, and $\nu \in [0,1]$ (see Definition 3.3), then there is a constant c depending only on the dimension d and the profile function f of H_B, such that $\delta_- \geqslant c$ and $\delta_+ \geqslant c(W/L)^2 + \nu a$.

PROOF. For the case where H is an a-full Wigner matrix, the claim easily follows by splitting

$$S \;=\; (S - a\mathbf{e}\mathbf{e}^*) + a\mathbf{e}\mathbf{e}^* \,.$$

By assumption, the first term is $(1-a)$ times a doubly stochastic matrix. Hence its spectrum lies in $[-1+a, 1-a]$. The claims on $\delta_\pm$ now follow easily.

The claims about band matrices were proved in Lemma A.1 of [17] and Equation (5.16) of [8], respectively. Finally, (iii) easily follows from (i) and (ii). □

B. Proof of Theorems 4.6 and 4.7

Theorems 4.6 and 4.7 are essentially simple special cases of the much more involved, and general, fluctuation averaging estimate from [9]. Nevertheless, here we give the details of the proofs because (a) they do not strictly follow from the formulation of the result in [9], and (b) their proof is much easier than that of [9], so that the reader only interested in the applications of fluctuation averaging to the local semicircle law need not read the lengthy proof of [9]. We start with a simple lemma which summarizes the key properties of $\prec$ when combined with expectation.

LEMMA B.1. *Suppose that the deterministic control parameter Ψ satisfies $\Psi \geqslant N^{-C}$, and that for all p there is a constant C_p such that the nonnegative random variable X satisfies $\mathbb{E}X^p \leqslant N^{C_p}$. Suppose moreover that that $X \prec \Psi$. Then for any fixed $n \in \mathbb{N}$ we have*

$$\mathbb{E}X^n \;\prec\; \Psi^n \,. \tag{B.1}$$

(Note that this estimate involves deterministic quantities only, i.e. it means that $\mathbb{E}X^n \leqslant N^\varepsilon \Psi^n$ for any $\varepsilon > 0$ if $N \geqslant N_0(n,\varepsilon)$.) Moreover, we have

$$P_i X^n \;\prec\; \Psi^n \,, \qquad Q_i X^n \;\prec\; \Psi^n \tag{B.2}$$

uniformly in i. If $X = X(u)$ and $\Psi = \Psi(u)$ depend on some parameter u and the above assumptions are uniform in u, then so are the conclusions.

PROOF OF LEMMA B.1. It is enough to consider the case $n = 1$; the case of larger n follows immediately from the case $n = 1$, using the basic properties of $\prec$ from Lemma 4.4.

For the first claim, pick $\varepsilon > 0$. Then

$$\mathbb{E}X \;=\; \mathbb{E}X\mathbf{1}(X \leqslant N^\varepsilon\Psi) + \mathbb{E}X\mathbf{1}(X > N^\varepsilon\Psi) \;\leqslant\; N^\varepsilon\Psi + \sqrt{\mathbb{E}X^2}\sqrt{\mathbb{P}(X > N^\varepsilon\Psi)} \;\leqslant\; N^\varepsilon\Psi + N^{C_2/2 - D/2} \,,$$

for arbitrary $D > 0$. The first claim therefore follows by choosing D large enough.

The second claim follows from Chebyshev's inequality, using a high-moment estimate combined with Jensen's inequality for partial expectation. We omit the details, which are similar to those of the first claim. □

We shall apply Lemma B.1 to resolvent entries of G. In order to verify its assumptions, we record the following bounds.

LEMMA B.2. *Suppose that $\Lambda \prec \Psi$ and $\Lambda_o \prec \Psi_o$ for some deterministic control parameters Ψ and Ψ_o both satisfying (4.8). Fix $p \in \mathbb{N}$. Then for any $i \neq j$ and $\mathbb{T} \subset \{1, \dots, N\}$ satisfying $|\mathbb{T}| \leqslant p$ and $i, j \notin \mathbb{T}$ we have*

$$G_{ij}^{(\mathbb{T})} \;=\; O_{\prec}(\Psi_o)\,, \qquad \frac{1}{G_{ii}^{(\mathbb{T})}} \;=\; O_{\prec}(1)\,. \tag{B.3}$$

Moreover, we have the rough bounds $\big|G_{ij}^{(\mathbb{T})}\big| \leqslant M$ and

$$\mathbb{E}\left|\frac{1}{G_{ii}^{(\mathbb{T})}}\right|^n \;\leqslant\; N^{\varepsilon} \tag{B.4}$$

for any $\varepsilon > 0$ and $N \geqslant N_0(n, \varepsilon)$.

PROOF. The bounds (B.3) follow easily by a repeated application of (4.6), the assumption $\Lambda \prec M^{-c}$, and the lower bound in (4.2). The deterministic bound $\big|G_{ij}^{(\mathbb{T})}\big| \leqslant M$ follows immediately from $\eta \geqslant M^{-1}$ by definition of a spectral domain.

In order to prove (B.4), we use Schur's complement formula (5.6) applied to $1/G_{ii}^{(\mathbb{T})}$, where the expectation is estimated using (2.6) and $\big|G_{ij}^{(\mathbb{T})}\big| \leqslant M$. (Recall (2.4).) This gives

$$\mathbb{E}\left|\frac{1}{G_{ii}^{(\mathbb{T})}}\right|^p \;\prec\; N^{C_p}$$

for all $p \in \mathbb{N}$. Since $1/G_{ii}^{(\mathbb{T})} \prec 1$, (B.4) therefore follows from (B.1). □

PROOF OF THEOREM 4.7. First we claim that, for any fixed $p \in \mathbb{N}$, we have

$$\left|Q_k \frac{1}{G_{kk}^{(\mathbb{T})}}\right| \;\prec\; \Psi_o \tag{B.5}$$

uniformly for $\mathbb{T} \subset \{1, \dots, \mathbb{N}\}$, $|\mathbb{T}| \leqslant p$, and $k \notin \mathbb{T}$. To simplify notation, for the proof we set $\mathbb{T} = \emptyset$; the proof for nonempty $\mathbb{T}$ is the same. From Schur's complement formula (5.6) we get $|Q_k(G_{kk})^{-1}| \leqslant |h_{kk}| + |Z_k|$. The first term is estimated by $|h_{kk}| \prec M^{-1/2} \leqslant \Psi_o$. The second term is estimated exactly as in (5.13) and (5.14):

$$|Z_k| \;\prec\; \left(\sum_{x \neq y}^{(k)} s_{kx} \big|G_{xy}^{(k)}\big|^2 s_{yk}\right)^{1/2} \;\prec\; \Psi_o\,,$$

where in the last step we used that $\big|G_{xy}^{(k)}\big| \prec \Psi_o$ as follows from (B.3), and the bound $1/|G_{kk}| \prec 1$ (recall that $\Lambda \prec \Psi \leqslant M^{-c}$). This concludes the proof of (B.5).

Abbreviate $X_k := Q_k(G_{kk})^{-1}$. We shall estimate $\sum_k t_{ik} X_k$ in probability by estimating its p-th moment by Ψ_o^{2p}, from which the claim will easily follow using Chebyshev's inequality. Before embarking on the estimate for arbitrary p, we illustrate its idea by estimating the variance

$$\mathbb{E}\left|\sum_k t_{ik} X_k\right|^2 \;=\; \sum_{k,l} t_{ik} \bar{t}_{il}\, \mathbb{E} X_k \overline{X}_l \;=\; \sum_k |t_{ik}|^2\, \mathbb{E} X_k \overline{X}_k + \sum_{k \neq l} t_{ik} \bar{t}_{il}\, \mathbb{E} X_k \overline{X}_l\,. \tag{B.6}$$

Using Lemma B.1 and the bounds (4.9) on t_{ik}, we find that the first term on the right-hand side of (B.6) is $O_\prec(M^{-1}\Psi_o^2) = O_\prec(\Psi_o^4)$, where we used the estimate (4.8). Let us therefore focus on the second term of (B.6). Using the fact that $k \neq l$, we apply (4.6) to X_k and X_l to get

$$\mathbb{E}X_k\overline{X}_l \;=\; \mathbb{E}Q_k\bigg(\frac{1}{G_{kk}}\bigg)Q_l\overline{\bigg(\frac{1}{G_{ll}}\bigg)} \;=\; \mathbb{E}Q_k\bigg(\frac{1}{G_{kk}^{(l)}} - \frac{G_{kl}G_{lk}}{G_{kk}G_{kk}^{(l)}G_{ll}}\bigg)Q_l\overline{\bigg(\frac{1}{G_{ll}^{(k)}} - \frac{G_{lk}G_{kl}}{G_{ll}G_{ll}^{(k)}G_{kk}}\bigg)}\,. \tag{B.7}$$

We multiply out the parentheses on the right-hand side. The crucial observation is that if the random variable Y is independent of i (see Definition 4.2) then $\mathbb{E}Q_i(X)Y = \mathbb{E}Q_i(XY) = 0$. Hence out of the four terms obtained from the right-hand side of (B.7), the only nonvanishing one is

$$\mathbb{E}Q_k\bigg(\frac{G_{kl}G_{lk}}{G_{kk}G_{kk}^{(l)}G_{ll}}\bigg)Q_l\overline{\bigg(\frac{G_{lk}G_{kl}}{G_{ll}G_{ll}^{(k)}G_{kk}}\bigg)} \;\prec\; \Psi_o^4\,.$$

Together with (4.9), this concludes the proof of $\mathbb{E}\big|\sum_k t_{ik}X_k\big|^2 \prec \Psi_o^4$.

After this pedagogical interlude we move on to the full proof. Fix some even integer p and write

$$\mathbb{E}\bigg|\sum_k t_{ik}X_k\bigg|^p \;=\; \sum_{k_1,\dots,k_p} t_{ik_1}\cdots t_{ik_{p/2}}\bar{t}_{ik_{p/2+1}}\cdots\bar{t}_{ik_p}\,\mathbb{E}X_{k_1}\cdots X_{k_{p/2}}\overline{X}_{k_{p/2+1}}\cdots\overline{X}_{k_p}\,.$$

Next, we regroup the terms in the sum over $\mathbf{k} := (k_1,\dots,k_p)$ according to the partition of $\{1,\dots,p\}$ generated by the indices $\mathbf{k}$. To that end, let $\mathfrak{P}_p$ denote the set of partitions of $\{1,\dots,p\}$, and $\mathcal{P}(\mathbf{k})$ the element of $\mathfrak{P}_p$ defined by the equivalence relation $r \sim s$ if and only if $k_r = k_s$. In short, we reorganize the summation according to coincidences among the indices $\mathbf{k}$. Then we write

$$\mathbb{E}\bigg|\sum_k t_{ik}X_k\bigg|^p \;=\; \sum_{P\in\mathfrak{P}_p}\sum_{\mathbf{k}} t_{ik_1}\cdots t_{ik_{p/2}}\bar{t}_{ik_{p/2+1}}\cdots\bar{t}_{ik_p}\,\mathbf{1}(\mathcal{P}(\mathbf{k}) = P)V(\mathbf{k})\,, \tag{B.8}$$

where we defined

$$V(\mathbf{k}) \;:=\; \mathbb{E}X_{k_1}\cdots X_{k_{p/2}}\overline{X}_{k_{p/2+1}}\cdots\overline{X}_{k_p}\,.$$

Fix $\mathbf{k}$ and set $P := \mathcal{P}(\mathbf{k})$ to be partition induced by the coincidences in $\mathbf{k}$. For any $r \in \{1,\dots,p\}$, we denote by $[r]$ the block of r in P. Let $L \equiv L(P) := \{r : [r] = \{r\}\} \subset \{1,\dots,p\}$ be the set of "lone" labels. We denote by $\mathbf{k}_L := (k_r)_{r\in L}$ the summation indices associated with lone labels.

The resolvent entry G_{kk} depends strongly on the randomness in the k-column of H, but only weakly on the randomness in the other columns. We conclude that if r is a lone label then all factors X_{k_s} with $s \neq r$ in $V(\mathbf{k})$ depend weakly on the randomness in the k_r-th column of H. Thus, the idea is to make all resolvent entries inside the expectation of $V(\mathbf{k})$ as independent of the indices $\mathbf{k}_L$ as possible (see Definition 4.2), using the identity (4.6). To that end, we say that a resolvent entry $G_{xy}^{(\mathbb{T})}$ with $x, y \notin \mathbb{T}$ is *maximally expanded* if $\mathbf{k}_L \subset \mathbb{T} \cup \{x, y\}$. The motivation behind this definition is that using (4.6) we cannot add upper indices from the set $\mathbf{k}_L$ to a maximally expanded resolvent entry. We shall apply (4.6) to all resolvent entries in $V(\mathbf{k})$. In this manner we generate a sum of monomials consisting of off-diagonal resolvent entries and inverses of diagonal resolvent entries. We can now repeatedly apply (4.6) to each factor until either they are all maximally expanded or a sufficiently large number of off-diagonal resolvent entries has been generated. The cap on the number of off-diagonal entries is introduced to ensure that this procedure terminates after a finite number of steps.

In order to define the precise algorithm, let $\mathcal{A}$ denote the set of monomials in the off-diagonal entries $G^{(\mathbb{T})}_{xy}$, with $\mathbb{T} \subset \mathbf{k}_L$, $x \neq y$, and $x, y \in \mathbf{k} \setminus \mathbb{T}$, as well as the inverse diagonal entries $1/G^{(\mathbb{T})}_{xx}$, with $\mathbb{T} \subset \mathbf{k}_L$ and $x \in \mathbf{k} \setminus \mathbb{T}$. Starting from $V(\mathbf{k})$, the algorithm will recursively generate sums of monomials in $\mathcal{A}$. Let $d(A)$ denote the number of off-diagonal entries in $A \in \mathcal{A}$. For $A \in \mathcal{A}$ we shall define $w_0(A), w_1(A) \in \mathcal{A}$ satisfying

$$A \;=\; w_0(A) + w_1(A)\,, \qquad d(w_0(A)) \;=\; d(A)\,, \qquad d(w_1(A)) \;\geqslant\; \max\bigl\{2, d(A)+1\bigr\}\,. \tag{B.9}$$

The idea behind this splitting is to use (4.6) on one entry of A; the first term on the right-hand side of (4.6) gives rise to $w_0(A)$ and the second to $w_1(A)$. The precise definition of the algorithm applied to $A \in \mathcal{A}$ is as follows.

(1) If all factors of A are maximally expanded or $d(A) \geqslant p+1$ then stop the expansion of A. In other words, the algorithm cannot be applied to A in the future.

(2) Otherwise choose some (arbitrary) factor of A that is not maximally expanded. If this entry is off-diagonal, $G^{(\mathbb{T})}_{xy}$, write

$$G^{(\mathbb{T})}_{xy} \;=\; G^{(\mathbb{T}u)}_{xy} + \frac{G^{(\mathbb{T})}_{xu} G^{(\mathbb{T})}_{uy}}{G^{(\mathbb{T})}_{uu}} \tag{B.10}$$

for the smallest $u \in \mathbf{k}_L \setminus (\mathbb{T} \cup \{x, y\})$. If the chosen entry is diagonal, $1/G^{(\mathbb{T})}_{xx}$, write

$$\frac{1}{G^{(\mathbb{T})}_{xx}} \;=\; \frac{1}{G^{(\mathbb{T}u)}_{xx}} - \frac{G^{(\mathbb{T})}_{xu} G^{(\mathbb{T})}_{ux}}{G^{(\mathbb{T})}_{xx} G^{(\mathbb{T}u)}_{xx} G^{(\mathbb{T})}_{uu}} \tag{B.11}$$

for the smallest $u \in \mathbf{k}_L \setminus (\mathbb{T} \cup \{x\})$. Then the splitting $A = w_0(A) + w_1(A)$ is defined by the splitting induced by (B.10) or (B.11), in the sense that we replace the factor $G^{(\mathbb{T})}_{xy}$ or $1/G^{(\mathbb{T})}_{xx}$ in the monomial A by the right-hand sides of (B.10) or (B.11).

(This algorithm contains some arbitrariness in the choice of the factor of A to be expanded. It may be removed for instance by first fixing some ordering of all resolvent entries $G^{(\mathbb{T})}_{ij}$. Then in (2) we choose the first factor of A that is not maximally expanded.) Note that (B.10) and (B.11) follow from (4.6). It is clear that (B.9) holds with the algorithm just defined.

We now apply this algorithm recursively to each entry $A^r := 1/G_{k_r k_r}$ in the definition of $V(\mathbf{k})$. More precisely, we start with A^r and define $A^r_0 := w_0(A^r)$ and $A^r_1 := w_1(A^r)$. In the second step of the algorithm we define four monomials

$$A^r_{00} \;:=\; w_0(A^r_0)\,, \qquad A^r_{01} \;:=\; w_0(A^r_1)\,, \qquad A^r_{10} \;:=\; w_1(A^r_0)\,, \qquad A^r_{11} \;:=\; w_1(A^r_1)\,,$$

and so on, at each iteration performing the steps (1) and (2) on each new monomial independently of the others. Note that the lower indices are binary sequences that describe the recursive application of the operations w_0 and w_1. In this manner we generate a binary tree whose vertices are given by finite binary strings σ. The associated monomials satisfy $A^r_{\sigma i} := w_i(A^r_\sigma)$ for $i = 0, 1$, where σi denotes the binary string obtained by appending i to the right end of σ. See Figure B.1 for an illustration of the tree.

We stop the recursion of a tree vertex whenever the associated monomial satisfies the stopping rule of step (1). In other words, the set of leaves of the tree is the set of binary strings σ such that either all factors of A^r_σ are maximally expanded or $d(A^r_\sigma) \geqslant p+1$. We claim that the resulting binary tree is finite, i.e. that the algorithm always reaches step (1) after a finite number of iterations. Indeed, by the stopping rule in (1),

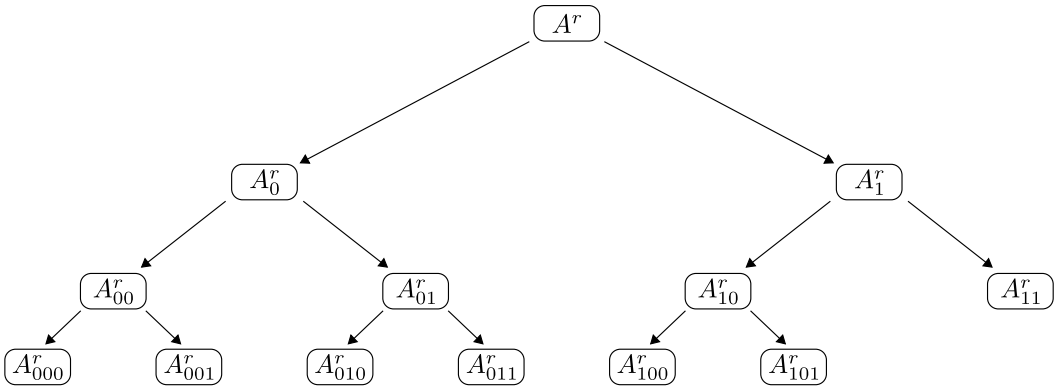

FIGURE B.1. The binary tree generated by applying the algorithm (1)–(2) to a monomial A^r. Each vertex of the tree is indexed by a binary string σ, and encodes a monomial A^r_σ. An arrow towards the left represents the action of w_0 and an arrow towards the right the action of w_1. The monomial A^r_{11} satisfies the assumptions of step (1), and hence its expansion is stopped, so that the tree vertex 11 has no children.

we have $d(A^r_\sigma) \leqslant p+1$ for any vertex σ of the tree. Since each application of w_1 increases $d(\cdot)$ by at least one, and in the first step (i.e. when applied to A^r) by two, we conclude that the number of ones in any σ is at most p. Since each application of w_1 increases the number of resolvent entries by at most four, and the application of w_0 does not change this number, we find that the number of resolvent entries in A^r_σ is bounded by $4p+1$. Hence the maximal number of upper indices in A^r_σ for any tree vertex σ is $(4p+1)p$. Since each application of w_0 increases the total number of upper indices by one, we find that σ contains at most $(4p+1)p$ zeros. We conclude that the maximal length of the string σ (i.e. the depth of the tree) is at most $(4p+1)p+p = 4p^2+2p$. A string σ encoding a tree vertex contains at most p ones. Denoting by k the number of ones in a string encoding a leaf of the tree, we find that the number of leaves is bounded by $\sum_{k=0}^{p} \binom{4p^2+2p}{k} \leqslant (Cp^2)^p$. Therefore, denoting by $\mathcal{L}_r$ the set of leaves of the binary tree generated from A^r, we have $|\mathcal{L}_r| \leqslant (Cp^2)^p$.

By definition of the tree and w_0 and w_1, we have the decomposition

$$X_{k_r} \;=\; Q_{k_r} \sum_{\sigma \in \mathcal{L}_r} A^r_\sigma \,. \tag{B.12}$$

Moreover, each monomial A^r_σ for $\sigma \in \mathcal{L}_r$ either consists entirely of maximally expanded resolvent entries or satisfies $d(A^r_\sigma) = p+1$. (This is an immediate consequence of the stopping rule in (1)).

Next, we observe that for any string σ we have

$$A^k_\sigma \;=\; O_\prec\big(\Psi_o^{b(\sigma)+1}\big)\,, \tag{B.13}$$

where $b(\sigma)$ is the number ones in the string σ. Indeed, if $b(\sigma) = 0$ then this follows from (B.5); if $b(\sigma) \geqslant 1$ this follows from the last statement in (B.9) and (B.3).

Using (B.8) and (B.12) we have the representation

$$V(\mathbf{k}) \;=\; \sum_{\sigma_1 \in \mathcal{L}_1} \cdots \sum_{\sigma_p \in \mathcal{L}_p} \mathbb{E}\big(Q_{k_1} A^1_{\sigma_1}\big) \cdots \big(Q_{k_p} \overline{A^p_{\sigma_p}}\big)\,. \tag{B.14}$$

We now claim that any nonzero term on the right-hand side of (B.14) satisfies

$$\big(Q_{k_1} A^1_{\sigma_1}\big) \cdots \big(Q_{k_p} \overline{A^p_{\sigma_p}}\big) \;=\; O_\prec\big(\Psi_o^{p+|L|}\big)\,. \tag{B.15}$$

Proof of (B.15). Before embarking on the proof, we explain its idea. By (B.13), the naive size of the left-hand side of (B.15) is Ψ_o^p. The key observation is that each lone label $s \in L$ yields one extra factor Ψ_o to the estimate. This is because the expectation in (B.14) would vanish if all other factors $\big(Q_{k_r} A^r_{\sigma_r}\big)$, $r \neq s$, were independent of k_s. The expansion of the binary tree makes this dependence explicit by exhibiting k_s as a lower index. But this requires performing an operation w_1 with the choice $u = k_s$ in (B.10) or (B.11). However, w_1 increases the number of off-diagonal element by at least one. In other words, every index associated with a lone label must have a "partner" index in a different resolvent entry which arose by application of w_1. Such a partner index may only be obtained through the creation of at least one off-diagonal resolvent entry. The actual proof below shows that this effect applies *cumulatively* for all lone labels.

In order to prove (B.15), we consider two cases. Consider first the case where for some $r = 1, \dots, p$ the monomial $A^r_{\sigma_r}$ on the left-hand side of (B.15) is not maximally expanded. Then $d(A^r_{\sigma_r}) = p + 1$, so that (B.3) yields $A^r_{\sigma_r} \prec \Psi_o^{p+1}$. Therefore the observation that $A^s_{\sigma_s} \prec \Psi_o$ for all $s \neq r$, together with (B.2) implies that the left-hand side of (B.15) is $O_\prec\big(\Psi_o^{2p}\big)$. Since $|L| \leqslant p$, (B.15) follows.

Consider now the case where $A^r_{\sigma_r}$ on the left-hand side of (B.15) is maximally expanded for all $r = 1, \dots, p$. The key observation is the following claim about the left-hand side of (B.15) with a nonzero expectation.

($*$) For each $s \in L$ there exists $r = \tau(s) \in \{1, \dots, p\} \setminus \{s\}$ such that the monomial $A^r_{\sigma_r}$ contains a resolvent entry with lower index k_s.

In other words, after expansion, the lone label s has a "partner" label $r = \tau(s)$, such that the index k_s appears also in the expansion of A^r (note that there may be several such partner labels r). To prove ($*$), suppose by contradiction that there exists an $s \in L$ such that for all $r \in \{1, \dots, p\} \setminus \{s\}$ the lower index k_s does not appear in the monomial $A^r_{\sigma_r}$. To simplify notation, we assume that $s = 1$. Then, for all $r = 2, \dots, p$, since $A^r_{\sigma_r}$ is maximally expanded, we find that $A^r_{\sigma_r}$ is independent of k_1 (see Definition 4.2). Therefore we have

$$\mathbb{E}\big(Q_{k_1} A^1_{\sigma_1}\big)\big(Q_{k_2} A^2_{\sigma_2}\big) \cdots \big(Q_{k_p} \overline{A^p_{\sigma_p}}\big) \;=\; \mathbb{E} Q_{k_1}\Big(A^1_{\sigma_1}\big(Q_{k_2} A^2_{\sigma_2}\big) \cdots \big(Q_{k_p} \overline{A^p_{\sigma_p}}\big)\Big) \;=\; 0\,,$$

where in the last step we used that $\mathbb{E} Q_i(X)Y = \mathbb{E} Q_i(XY) = 0$ if Y is independent of i. This concludes the proof of ($*$).

For $r \in \{1, \dots, p\}$ we define $\ell(r) := \sum_{s \in L} \mathbf{1}(\tau(s) = r)$, the number of times that the label r was chosen as a partner to some lone label s. We now claim that

$$A^r_{\sigma_r} \;=\; O_\prec\big(\Psi_o^{1+\ell(r)}\big)\,. \tag{B.16}$$

To prove (B.16), fix $r \in \{1, \dots, p\}$. By definition, for each $s \in \tau^{-1}(\{r\})$ the index k_s appears as a lower index in the monomial $A^r_{\sigma_r}$. Since $s \in L$ is by definition a lone label and $s \neq r$, we know that k_s does not appear as an index in A^r. By definition of the monomials associated with the tree vertex σ_r, it follows that $b(\sigma_r)$, the number of ones in σ_r, is at least $\big|\tau^{-1}(\{r\})\big| = \ell(r)$ since each application of w_1 adds precisely one new (lower) index. Note that in this step it is crucial that $s \in \tau^{-1}(\{r\})$ was a lone label. Recalling (B.13), we therefore get (B.16).

Using (B.16) and Lemma B.1 we find

$$\Big|\big(Q_{k_1} A^1_{\sigma_1}\big) \cdots \big(Q_{k_p} \overline{A^p_{\sigma_p}}\big)\Big| \;\prec\; \prod_{r=1}^p \Psi_o^{1+\ell(r)} \;=\; \Psi_o^{p+|L|}\,.$$

This concludes the proof of (B.15). □

Summing over the binary trees in (B.14) and using Lemma B.1, we get from (B.15)

$$V(\mathbf{k}) \;=\; O_{\prec}\big(\Psi_o^{p+|L|}\big)\,. \tag{B.17}$$

We now return to the sum (B.8). We perform the summation by first fixing $P \in \mathfrak{P}_p$, with associated lone labels $L = L(P)$. We find

$$\left|\sum_{\mathbf{k}} \mathbf{1}(\mathcal{P}(\mathbf{k}) = P)\, t_{ik_1} \cdots t_{ik_{p/2}} \bar{t}_{ik_{p/2+1}} \cdots \bar{t}_{ik_p}\right| \;\leqslant\; (M^{-1})^{p-|P|} \;\leqslant\; (M^{-1/2})^{p-|L|}\,;$$

in the first step we used (4.9) and the fact that the summation is performed over $|P|$ free indices, the remaining $p - |P|$ being estimated by M^{-1}; in the second step we used that each block of P that is not contained in L consists of at least two labels, so that $p - |P| \geqslant (p - |L|)/2$. From (B.8) and (B.17) we get

$$\mathbb{E}\left|\sum_k t_{ik} X_k\right|^p \;\prec\; \sum_{P \in \mathfrak{P}_p} (M^{-1/2})^{p-|L(P)|}\, \Psi_o^{p+|L(P)|} \;\leqslant\; C_p \Psi_o^{2p}\,,$$

where in the last step we used the lower bound from (4.8) and estimated the summation over $\mathfrak{P}_p$ with a constant C_p (which is bounded by $(Cp^2)^p$). Summarizing, we have proved that

$$\mathbb{E}\left|\sum_k t_{ik} X_k\right|^p \;\prec\; \Psi_o^{2p} \tag{B.18}$$

for any $p \in 2\mathbb{N}$.

We conclude the proof of Theorem 4.7 with a simple application of Chebyshev's inequality. Fix $\varepsilon > 0$ and $D > 0$. Using (B.18) and Chebyshev's inequality we find

$$\mathbb{P}\left(\left|\sum_k t_{ik} X_k\right| > N^{\varepsilon} \Psi_o^2\right) \;\leqslant\; N\, N^{-\varepsilon p}$$

for large enough $N \geqslant N_0(\varepsilon, p)$. Choosing $p \geqslant \varepsilon^{-1}(1 + D)$ concludes the proof of Theorem 4.7. □

Remark B.3. The identity (4.6) is the only identity about the entries of G that is needed in the proof of Theorem 4.7. In particular, (4.7) is never used, and the actual entries of H never appear in the argument.

Proof of Theorem 4.6. The first estimate of (4.11) follows from Theorem 4.7 and the simple bound $\Lambda_o \leqslant \Lambda \prec \Psi$. The second estimate of (4.11) may be proved by following the proof of Theorem 4.7 verbatim; the only modification is the bound

$$\big|Q_k G_{kk}^{(\mathbb{T})}\big| \;=\; \big|Q_k\big(G_{kk}^{(\mathbb{T})} - m\big)\big| \;\prec\; \Psi\,,$$

which replaces (B.5). Here we again use the same upper bound $\Psi_o = \Psi$ for Λ and Λ_o.

In order to prove (4.12), we write Schur's complement formula (5.6) using (2.8) as

$$\frac{1}{G_{ii}} \;=\; \frac{1}{m} + h_{ii} - \left(\sum_{k,l}^{(i)} h_{ik} G_{kl}^{(i)} h_{li} - m\right). \tag{B.19}$$

Since $|h_{ii}| \prec M^{-1/2} \leqslant \Psi$ and $|1/G_{ii} - 1/m| \prec \Psi$, we find that the term in parentheses is stochastically dominated by Ψ. Therefore we get, inverting (B.19) and expanding the right-hand side, that

$$v_i \;=\; G_{ii} - m \;=\; m^2 \biggl(-h_{ii} + \sum_{k,l}^{(i)} h_{ik} G^{(i)}_{kl} h_{li} - m \biggr) + O_\prec(\Psi^2) \,.$$

Taking the partial expectation P_i yields

$$P_i v_i \;=\; m^2 \biggl(\sum_k^{(i)} s_{ik} G^{(i)}_{kk} - m \biggr) + O_\prec(\Psi^2) \;=\; m^2 \sum_k s_{ik} v_k + O_\prec(\Psi^2) \,,$$

where in the second step we used (4.6), (2.2), and (B.3). Therefore we get, using (4.11) and $Q_i G_{ii} = Q_i (G_{ii} - m) = Q_i v_i$,

$$w_a \;:=\; \sum_i t_{ai} v_i \;=\; \sum_i t_{ai} P_i v_i + \sum_i t_{ai} Q_i v_i \;=\; m^2 \sum_{i,k} t_{ai} s_{ik} v_k + O_\prec(\Psi^2) \;=\; m^2 \sum_{i,k} s_{ai} t_{ik} v_k + O_\prec(\Psi^2) \,,$$

where in the last step we used that the matrices T and S commute by assumption. Introducing the vector $\mathbf{w} = (w_a)_{a=1}^N$ we therefore have the equation

$$\mathbf{w} = m^2 S \mathbf{w} + O_\prec(\Psi^2) \,, \tag{B.20}$$

where the error term is in the sense of the ℓ^∞-norm (uniform in the components of the vector $\mathbf{w}$). Inverting the matrix $1 - m^2 S$ and recalling the definition (2.10) yields (4.12).

The proof of (4.14) is similar, except that we have to treat the subspace $\mathbf{e}^\perp$ separately. Using (4.13) we write

$$\sum_i t_{ai} (v_i - [v]) \;=\; \sum_i t_{ai} v_i - \sum_i \frac{1}{N} v_i \,,$$

and apply the above argument to each term separately. This yields

$$\sum_i t_{ai} (v_i - [v]) \;=\; m^2 \sum_i t_{ai} \sum_k s_{ik} v_k - m^2 \sum_i \frac{1}{N} \sum_k t_{ik} v_k + O_\prec(\Psi^2) \;=\; m^2 \sum_{i,k} s_{ai} t_{ik} (v_k - [v]) + O_\prec(\Psi^2) \,,$$

where we used (2.3) in the second step. Note that the error term on the right-hand side is perpendicular to $\mathbf{e}$ when regarded as a vector indexed by a, since all other terms in the equation are. Hence we may invert the matrix $(1 - m^2 S)$ on the subspace $\mathbf{e}^\perp$, as above, to get (4.14). □

We conclude this section with an alternative proof of Theorem 4.7. While the underlying argument remains similar, the following proof makes use of an additional decomposition of the space of random variables, which avoids the use of the stopping rule from Step (1) in the above proof of Theorem 4.7. This decomposition may be regarded as an abstract reformulation of the stopping rule.

Alternative proof of Theorem 4.7. As before, we set $X_k := Q_k (G_{kk})^{-1}$. For simplicity of presentation, we set $t_{ik} = N^{-1}$. The decomposition is defined using the operations P_i and Q_i, introduced in Definition 4.2. It is immediate that P_i and Q_i are projections, that $P_i + Q_i = 1$, and that all of these projections commute with each other. For a set $A \subset \{1, \dots, N\}$ we use the notations $P_A := \prod_{i \in A} P_i$ and $Q_A := \prod_{i \in A} Q_i$.

Let p be even and introduce the shorthand $\widetilde{X}_{k_s} := X_{k_s}$ for $s \leqslant p/2$ and $\widetilde{X}_{k_s} := \overline{X}_{k_s}$ for $s > p/2$. Then we get

$$\mathbb{E}\Big|\frac{1}{N}\sum_k X_k\Big|^p \;=\; \frac{1}{N^p}\sum_{k_1,\dots,k_p}\mathbb{E}\prod_{s=1}^p \widetilde{X}_{k_s} \;=\; \frac{1}{N^p}\sum_{k_1,\dots,k_p}\mathbb{E}\prod_{s=1}^p\Bigg(\prod_{r=1}^p (P_{k_r}+Q_{k_r})\widetilde{X}_{k_s}\Bigg)\,.$$

Introducing the notations $\mathbf{k} = (k_1,\dots,k_p)$ and $[\mathbf{k}] = \{k_1,\dots,k_p\}$, we therefore get by multiplying out the parentheses

$$\mathbb{E}\Big|\frac{1}{N}\sum_k X_k\Big|^p \;=\; \frac{1}{N^p}\sum_{\mathbf{k}}\sum_{A_1,\dots,A_p\subset[\mathbf{k}]}\mathbb{E}\prod_{s=1}^p\big(P_{A_s^c}Q_{A_s}\widetilde{X}_{k_s}\big)\,. \tag{B.21}$$

Next, by definition of $\widetilde{X}_{k_s}$, we have that $\widetilde{X}_{k_s} = Q_{k_s}\widetilde{X}_{k_s}$, which implies that $P_{A_s^c}\widetilde{X}_{k_s} = 0$ if $k_s \notin A_s$. Hence may restrict the summation to A_s satisfying

$$k_s \;\in\; A_s \tag{B.22}$$

for all s. Moreover, we claim that the right-hand side of (B.21) vanishes unless

$$k_s \;\in\; \bigcup_{q\neq s} A_q \tag{B.23}$$

for all s. Indeed, suppose that $k_s \in \bigcap_{q\neq s} A_q^c$ for some s, say $s=1$. In this case, for each $s = 2,\dots,p$, the factor $P_{A_s^c}Q_{A_s}\widetilde{X}_{k_s}$ is independent of k_1 (see Definition 4.2). Thus we get

$$\begin{aligned}\mathbb{E}\prod_{s=1}^p\big(P_{A_s^c}Q_{A_s}\widetilde{X}_{k_s}\big) \;&=\; \mathbb{E}\big(P_{A_1^c}Q_{A_1}Q_{k_1}\widetilde{X}_{k_1}\big)\prod_{s=2}^p\big(P_{A_s^c}Q_{A_s}\widetilde{X}_{k_s}\big)\\ &=\; \mathbb{E}Q_{k_1}\Bigg(\big(P_{A_1^c}Q_{A_1}\widetilde{X}_{k_1}\big)\prod_{s=2}^p\big(P_{A_s^c}Q_{A_s}\widetilde{X}_{k_s}\big)\Bigg) \;=\; 0\,,\end{aligned}$$

where in the last step we used that $\mathbb{E}Q_i(X) = 0$ for any i and random variable X.

We conclude that the summation on the right-hand side of (B.21) is restricted to indices satisfying (B.22) and (B.23). Under these two conditions we have

$$\sum_{s=1}^p |A_s| \;\geqslant\; 2\,|[\mathbf{k}]|\,, \tag{B.24}$$

since each index k_s must belong to at least two different sets A_q: to A_s (by (B.22)) as well as to some A_q with $q \neq s$ (by (B.23)).

Next, we claim that for $k \in A$ we have

$$|Q_A X_k| \;\prec\; \Psi_o^{|A|}\,. \tag{B.25}$$

(Note that if we were doing the case $X_k = Q_k G_{kk}$ instead of $X_k = Q_k(G_{kk})^{-1}$, then (B.25) would have to be weakened to $|Q_A X_k| \prec \Psi^{|A|}$, in accordance with (4.11). Indeed, in that case and for $A = \{k\}$, we only have the bound $|Q_k G_{kk}| \prec \Psi$ and not $|Q_k G_{kk}| \prec \Psi_o$.)

Before proving (B.25), we show it may be used to complete the proof. Using (B.21), (B.25), and Lemma B.1, we find

$$\mathbb{E}\left|\frac{1}{N}\sum_k X_k\right|^p \;\prec\; C_p\frac{1}{N^p}\sum_{\mathbf{k}}\Psi_o^{2|[k]|} \;=\; C_p\sum_{u=1}^{p}\Psi_o^{2u}\frac{1}{N^p}\sum_{\mathbf{k}}\mathbf{1}(|[\mathbf{k}]|=u)$$
$$\leqslant\; C_p\sum_{u=1}^{p}\Psi_o^{2u}N^{u-p} \;\leqslant\; C_p(\Psi_o+N^{-1/2})^{2p} \;\leqslant\; C_p\Psi_o^{2p}\,,$$

where in the first step we estimated the summation over the sets $A_1,\dots,A_p$ by a combinatorial factor C_p depending on p, in the forth step we used the elementary inequality $a^nb^m\leqslant(a+b)^{n+m}$ for positive a,b, and in the last step we used (4.8) and the bound $M\leqslant N$. Thus we have proved (B.18), from which the claim follows exactly as in the first proof of Theorem 4.7.

What remains is the proof of (B.25). The case $|A|=1$ (corresponding to $A=\{k\}$) follows from (B.5), exactly as in the first proof of Theorem 4.7. To simplify notation, for the case $|A|\geqslant 2$ we assume that $k=1$ and $A=\{1,2,\dots,t\}$ with $t\geqslant 2$. It suffices to prove that

$$\left|Q_t\cdots Q_2\frac{1}{G_{11}}\right| \;\prec\; \Psi_o^t\,. \tag{B.26}$$

We start by writing, using (4.6),

$$Q_2\frac{1}{G_{11}} \;=\; Q_2\frac{1}{G_{11}^{(2)}}+Q_2\frac{G_{12}G_{21}}{G_{11}G_{11}^{(2)}G_{22}} \;=\; Q_2\frac{G_{12}G_{21}}{G_{11}G_{11}^{(2)}G_{22}}\,,$$

where the first term vanishes since $G_{11}^{(2)}$ is independent of 2 (see Definition 4.2). We now consider

$$Q_3Q_2\frac{1}{G_{11}} \;=\; Q_2Q_3\frac{G_{12}G_{21}}{G_{11}G_{11}^{(2)}G_{22}}\,,$$

and apply (4.6) with $k=3$ to each resolvent entry on the right-hand side, and multiply everything out. The result is a sum of fractions of entries of G, whereby all entries in the numerator are diagonal and all entries in the denominator are diagonal. The leading order term vanishes,

$$Q_2Q_3\frac{G_{12}^{(3)}G_{21}^{(3)}}{G_{11}^{(3)}G_{11}^{(23)}G_{22}^{(3)}} \;=\; 0\,,$$

so that the surviving terms have at least three (off-diagonal) resolvent entries in the numerator. We may now continue in this manner; at each step the number of (off-diagonal) resolvent entries in the numerator increases by at least one.

More formally, we obtain a sequence $A_2,A_3,\dots,A_t$, where $A_2:=Q_2\frac{G_{12}G_{21}}{G_{11}G_{11}^{(2)}G_{22}}$ and A_i is obtained by applying (4.6) with $k=i$ to each entry of Q_iA_{i-1}, and keeping only the nonvanishing terms. The following properties are easy to check by induction.

(i) $A_i=Q_iA_{i-1}$.

(ii) A_i consists of the projection $Q_2\cdots Q_i$ applied to a sum of fractions such that all entries in the numerator are diagonal and all entries in the denominator are diagonal.

(iii) The number of (off-diagonal) entries in the numerator of each term of A_i is at least i.

By Lemma B.1 combined with (ii) and (iii) we conclude that $|A_i| \prec \Psi_o^i$. From (i) we therefore get

$$Q_t \cdots Q_2 \frac{1}{G_{11}} \;=\; A_t \;=\; O_\prec(\Psi_o^t)\,.$$

This is (B.26). Hence the proof is complete. □

C. Large deviation bounds

We consider random variables X satisfying

$$\mathbb{E}X \;=\; 0\,, \qquad \mathbb{E}|X|^2 \;=\; 1\,, \qquad (\mathbb{E}|X|^p)^{1/p} \;\leqslant\; \mu_p \tag{C.1}$$

for all $p \in \mathbb{N}$ and some constants μ_p.

THEOREM C.1 (LARGE DEVIATION BOUNDS). *Let $\big(X_i^{(N)}\big)$ and $\big(Y_i^{(N)}\big)$ be independent families of random variables and $\big(a_{ij}^{(N)}\big)$ and $\big(b_i^{(N)}\big)$ be deterministic; here $N \in \mathbb{N}$ and $i, j = 1, \ldots, N$. Suppose that all entries $X_i^{(N)}$ and $Y_i^{(N)}$ are independent and satisfy* (C.1)*. Then we have the bounds*

$$\sum_i b_i X_i \;\prec\; \bigg(\sum_i |b_i|^2\bigg)^{1/2}, \tag{C.2}$$

$$\sum_{i,j} a_{ij} X_i Y_j \;\prec\; \bigg(\sum_{i,j} |a_{ij}|^2\bigg)^{1/2}, \tag{C.3}$$

$$\sum_{i \neq j} a_{ij} X_i X_j \;\prec\; \bigg(\sum_{i \neq j} |a_{ij}|^2\bigg)^{1/2}. \tag{C.4}$$

If the coefficients $a_{ij}^{(N)}$ and $b_i^{(N)}$ depend on an additional parameter u, then all of these estimates are uniform in u (see Definition 2.1), i.e. the threshold $N_0 = N_0(\varepsilon, D)$ in the definition of $\prec$ depends only on the family μ_p from (C.1) *and δ from* (2.4)*; in particular, N_0 does not depend on u.*

PROOF. The estimates (C.2), (C.3), and (C.4) follow from Lemmas B.2, B.3, and B.4 of [8], combined with Chebyshev's inequality. □

References

[1] Bai, Z. D., Miao, B., Tsay, J.: Convergence rates of the spectral distributions of large Wigner matrices. *Int. Math. J.* **1** (2002), no. 1, 65–90.

[2] Cacciapuoti, C., Maltsev, A., Schlein, B.: Local Marchenko-Pastur Law at the Hard Edge of Sample Covariance Matrices. Preprint. arxiv:1206.1730

[3] Chatterjee, S.: A generalization of the Lindeberg principle. *Ann. Probab.* **34** (2006), no. 6, 2061–2076.

[4] Davies, E.B.: The Functional Calculus. *J. London Math. Soc.* **52**, 166–176 (1995).

[5] Erdős, L., A. Knowles, A.: Quantum Diffusion and Delocalization for Band Matrices with General Distribution. *Annales Inst. H. Poincaré*, **12** (7), 1227-1319 (2011)

[6] Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Spectral Statistics of Erdős-Rényi Graphs I: Local Semicircle Law. To appear in *Annals Prob.* Preprint. Arxiv:1103.1919

[7] Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Spectral Statistics of Erdős-Rényi Graphs II: Eigenvalue Spacing and the Extreme Eigenvalues. *Comm. Math. Phys.* **314** no. 3. 587–640 (2012)

[8] Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Delocalization and Diffusion Profile for Random Band Matrices. Preprint. Arxiv:1205.5669

[9] Erdős, L., Knowles, A., Yau, H.-T.: Averaging Fluctuations in Resolvents of Random Band Matrices. Preprint. Arxiv:1205.5664

[10] Erdős, L., Péché, G., Ramírez, J., Schlein, B., and Yau, H.-T., Bulk universality for Wigner matrices. *Commun. Pure Appl. Math.* **63**, No. 7, 895–925 (2010)

[11] Erdős, L., Ramirez, J., Schlein, B., Yau, H.-T.: Universality of sine-kernel for Wigner matrices with a small Gaussian perturbation. *Electr. J. Prob.* **15**, Paper 18, 526–604 (2010)

[12] Erdős, L., Schlein, B., Yau, H.-T.: Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices. *Ann. Probab.* **37**, No. 3, 815–852 (2009)

[13] Erdős, L., Schlein, B., Yau, H.-T.: Local semicircle law and complete delocalization for Wigner random matrices. *Commun. Math. Phys.* **287**, 641–655 (2009)

[14] Erdős, L., Schlein, B., Yau, H.-T.: Universality of random matrices and local relaxation flow. *Invent. Math.* **185** (2011), no.1, 75–119.

[15] Erdős, L., Schlein, B., Yau, H.-T., Yin, J.: The local relaxation flow approach to universality of the local statistics for random matrices. *Annales Inst. H. Poincaré (B), Probability and Statistics* **48**, no. 1, 1–46 (2012)

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

[17] Erdős, L., Yau, H.-T., Yin, J.: Bulk universality for generalized Wigner matrices. To appear in *Prob. Theor. Rel. Fields.* Preprint arXiv:1001.3453.

[18] Erdős, L., Yau, H.-T., Yin, J.: Universality for generalized Wigner matrices with Bernoulli distribution. *J. of Combinatorics,* **1** (2011), no. 2, 15–85

[19] Erdős, L., Yau, H.-T., Yin, J.: Rigidity of Eigenvalues of Generalized Wigner Matrices. *Adv. Math.* **229**, no. 3, 1435–1515 (2012)

[20] Guionnet, A., Zeitouni, O.: Concentration of the spectral measure for large matrices. *Electronic Comm. in Probability* **5** (2000) Paper 14.

[21] Helffer, B. and Sjöstrand, J.: Équation de Schrödinger avec champ magnétique et équation de Harper. *Schrödinger operators*, Lecture Notes in Physics 345 (eds. H. Holden and A. Jensen; Springer, Berlin, 1989) 118–197.

[22] Feldheim, O. and Sodin, S.: A universality result for the smallest eigenvalues of certain sample covariance matrices. *Geom. Funct. Anal.* **20** (2010), no.1, 88–123.

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

[24] V. A. Marčenko and L. A. Pastur, Distribution of eigenvalues for some sets of random matrices, *Sbornik: Mathematics* **1** (1967), 457–483.

[25] Mehta, M.L.: *Random Matrices.* Third Edition, Academic Press, New York, 1991.

[26] Pillai, N.S. and Yin, J.: Universality of covariance matrices. Preprint arXiv:1110.2501

[27] Sodin, S.: The spectral edge of some random band matrices. *Ann. Math.* (2) **172** (2010), no. 3, 2223-2251.

[28] Spencer, T.: Random banded and sparse matrices (Chapter 23), to appear in "Oxford Handbook of Random Matrix Theory" edited by G. Akemann, J. Baik, and P. Di Francesco

[29] Tao, T. and Vu, V.: Random matrices: Universality of the local eigenvalue statistics. *Acta Math.*, **206** (2011), no. 1, 127–204.

[30] Tao, T. and Vu, V.: Random matrices: Sharp concentration of eigenvalues. Preprint arXiv:1201.4789

[31] Wigner, E.: Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math.* **62** (1955), 548-564.