Universality of Wigner random matrices: 
a survey of recent results

L. Erdős

Abstract. This is a study of the universality of spectral statistics for large random matrices. Considered are $N \times N$ symmetric, Hermitian, or quaternion self-dual random matrices with independent identically distributed entries (Wigner matrices), where the probability distribution of each matrix element is given by a measure $\nu$ with zero expectation and with subexponential decay. The main result is that the correlation functions of the local eigenvalue statistics in the bulk of the spectrum coincide with those of the Gaussian Orthogonal Ensemble (GOE), the Gaussian Unitary Ensemble (GUE), and the Gaussian Symplectic Ensemble (GSE), respectively, in the limit as $N \to \infty$. This approach is based on a study of the Dyson Brownian motion via a related new dynamics, the local relaxation flow. As a main input, it is established that the density of the eigenvalues converges to the Wigner semicircle law, and this holds even down to the smallest possible scale. Moreover, it is shown that the eigenvectors are completely delocalized. These results hold even without the condition that the matrix elements are identically distributed: only independence is used. In fact, for the matrix elements of the Green function strong estimates are given that imply that the local statistics of any two ensembles in the bulk are identical if the first four moments of the matrix elements match. Universality at the spectral edges requires matching only two moments. A Wigner-type estimate is also proved, and it is shown that the eigenvalues repel each other on arbitrarily small scales.

Bibliography: 108 titles.

Keywords: Wigner random matrices, Dyson Brownian motion, semicircle law, sine kernel.

Contents

1. Introduction 508
   1.1. Summary of the main results: An orientation for the reader 509
   1.2. Wigner matrix ensembles 514
   1.3. Motivations: From Schrödinger operators to $\zeta$-functions 517
   1.4. Eigenvalue density and delocalization 520
   1.5. Local statistics of eigenvalues: Previous results 525

This work was partially supported by grant no. SFB-TR 12 of the German Research Council. 
AMS 2010 Mathematics Subject Classification. Primary 15B52, 82B44.
1. Introduction

This survey is based on lecture notes that the author prepared for the participants at the Arizona School of Analysis and Applications in Tucson in 2010. The style of the presentation is closer to the informal style of a lecture than to a formal research article. For the details and sometimes even for the precise formulations we refer to the original papers.

In the first introductory section we give an overview about universality of random matrices, including previous results, history, and motivations. We introduce some basic concepts such as Wigner matrix, Wigner semicircle law, Stieltjes transform, moment method, sine kernel, gap distribution, level repulsion, bulk and edge universality, invariant ensembles, Green function comparison theorem, four-moment theorem, local relaxation flow, and reverse heat flow. Some of these concepts will not be used for our main results, but we included them to help the orientation of the reader. The selection of material presented in the first section admittedly reflects the personal bias of the author and is not meant to be comprehensive. It is focused on the background material for the later sections, where we present our recent results on universality of random matrices.

There are several very active research directions connected with random matrices that are not mentioned in this survey at all, for example, supersymmetric methods or connections with free probability. Some other very rich topics, for example, edge universality with the moment method or connections with orthogonal polynomials, are mentioned only superficially. We refer the reader to
more comprehensive surveys on random matrices, especially the classical book of Mehta [1], the survey of Deift [2], the recent book of Anderson, Guionnet, and Zeitouni [3], and the forthcoming book of Forrester [4]. An excellent short summary about the latest developments is by Guionnet [5].

Starting from §2, we present our recent results which give the shortest and up to now the most powerful approach to bulk universality for $N \times N$ Wigner matrices. One of the main results is formulated in Theorem 5.2. These results were obtained in collaboration with J. Ramírez, S. Péché, B. Schlein, H.-T. Yau, and J. Yin; see the bibliography for precise references. In this part we strive for mathematical rigour, but for several details we refer to the original papers. The argument has three distinct steps:

1) local semicircle law (§2);
2) universality for Gaussian convolutions via the local reflection flow (§3);
3) Green function comparison theorem (§4).

Finally, in §5 we put together the proof from these ingredients. The main result on universality of local statistics in the bulk for Wigner matrices is formulated in Theorem 5.1. Some technical lemmas that can be neglected at first reading are collected in the appendices.

**Conventions.** Throughout the paper the letters $C$ and $c$ denote positive constants whose values may change from formula to formula, and they are independent of the other parameters. Since we will always take the limit as $N \to \infty$ at the end, all estimates are understood for sufficiently large $N$. In informal explanations we will often neglect logarithmic factors, by introducing the notation $\lesssim$ and $\ll$ to indicate inequality up to some factor of order $\log N$. More precisely, $A \lesssim B$ means that $A \leq (\log N)^C B$ with some non-negative constant $C$, and $A \ll B$ means that $A \leq (\log N)^{-C} B$ for some positive constant $C$.

The author thanks H.-T. Yau for suggestions on improving the presentation of this survey.

1.1. **Summary of the main results: An orientation for the reader.** We will consider $N \times N$ matrices $H = (h_{ij})_{i,j=1}^N$ whose entries are real or complex random variables. In most cases we assume that $H$ is Hermitian or symmetric, but our method applies to other ensembles as well (our results for matrices with quaternion entries will not be discussed here; see [6]). We assume that the entries are independent (up to the symmetry constraint $h_{ij} = \bar{h}_{ji}$), they are centered, $E h_{ij} = 0$, and their tail probability has a uniform subexponential decay (see (2.32) below). We do not assume that the matrix elements are identically distributed, but we assume that the variances $\sigma_{ij}^2 := E|h_{ij}|^2$ satisfy the normalization conditions

$$\sum_{j=1}^N \sigma_{ij}^2 = 1, \quad i = 1, \ldots, N, \quad (1.1)$$

that is, the deterministic $N \times N$ matrix of variances $\Sigma = (\sigma_{ij}^2)$ is symmetric and doubly stochastic. These conditions guarantee that $-1 \leq \Sigma \leq 1$. We will always assume that 1 is a simple eigenvalue of $\Sigma$ and there is a positive number $\delta_- > 0$ such that $-1 + \delta_- \leq \Sigma$. This assumption is satisfied for practically any random
matrix ensembles. Sometimes we will need a uniform gap condition, that is, that there exists a positive $\delta_+ > 0$ such that

$$\text{Spec } \Sigma \subset [-1 + \delta_-, 1 - \delta_+] \cup \{1\}.$$ 

For example, for the standard Wigner matrix $\sigma_{ij}^2 = N^{-1}$ and $\delta_- = \delta_+ = 1$. For random band matrices (see (1.18) for the precise definition) with band width $W$ satisfying $1 \ll W \ll N$ the gap $\delta_+$ goes to zero as the size of the matrix increases.

The normalization (1.1) ensures that the bulk of the spectrum of $H$ lies in the interval $[-2, 2]$ and the density of the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$ is given by the Wigner semicircle law as $N \to \infty$. Apart from neighbourhoods of the edges $\pm 2$, the typical spacing of neighbouring eigenvalues is of order $1/N$. We are interested in the statistics of the eigenvalues in the limit as $N \to \infty$.

1.1.1. Summary of §2: Main results on the local semicircle law. In §2 we prove that the density of the eigenvalues follows the semicircle law down to the smallest possible scale, that is, to scales only slightly larger than $1/N$. We will call it the local semicircle law. The local semicircle law is identified via the Stieltjes transform of the empirical density of the eigenvalues,

$$m(z) := m_N(z) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{\lambda_j - z}, \quad z = E + i\eta, \quad E \in \mathbb{R}, \quad \eta > 0,$$

and we show that $m_N(z)$ converges to the Stieltjes transform of the semicircle density,

$$m_{\text{sc}}(z) := \int_{\mathbb{R}} \frac{\varrho_{\text{sc}}(x) \, dx}{x - z}, \quad \varrho_{\text{sc}}(x) := \frac{1}{2\pi} \sqrt{4 - x^2}_+, \quad x > 0,$$

in the limit as $N \to \infty$. The imaginary part $\eta = \text{Im } z$ may depend on $N$, which corresponds to the local scale on which the density is identified. The precision of our approximation is of order $(N\eta)^{-1}$. Our best result in this direction is Theorem 2.1 of [7], which we will call the strong local semicircle law:

$$|m(z) - m_{\text{sc}}(z)| \leq C(\log N)^L \frac{N^L}{N\eta}$$

for some sufficiently large $L$ and with a very high probability (see §2.4). This result holds even for a more general class of Wigner matrices whose variances are comparable (see (1.17) for the precise definition); the key input is that in this case we have $\delta_+ > 0$.

For even more general Wigner matrices (they will be called universal Wigner matrices; see Definition 1.1 below), the key quantity that measures the precision is the spread of the matrix, defined by

$$M := \frac{1}{\max_{ij} \sigma_{ij}^2}.$$  

For typical random band matrices (see (1.18) for the precise definition), $M$ is comparable to the band width $W$. If $M \ll N$, then the precision of our estimates is
determined by $M$ instead of $N$; for example, in [8] we showed that
\[ |m(z) - m_{sc}(z)| \lesssim \frac{CN^\varepsilon}{M \eta \kappa^2}, \quad \kappa := ||E| - 2|, \tag{1.4} \]
for any $\varepsilon > 0$, with a very high probability (see Theorem 2.5 below), and we proved that
\[ |m(z) - m_{sc}(z)| \lesssim \frac{C(\log N)^L}{\sqrt{M \eta \kappa}} \tag{1.5} \]
in Theorem 2.1 of [9]. We remark that these estimates deteriorate near the spectral edges.

It is well known that the identification of the Stieltjes transform of a measure for the complex parameters $z = E + i\eta$, $E \in \mathbb{R}$, is equivalent to knowing the density down to scales essentially $O(\eta)$; we thereby have control on the density down to scales essentially of order $\eta \sim 1/M$.

The Stieltjes transform $m(z)$ can also be viewed as the normalized trace of the resolvent,
\[ m(z) = \frac{1}{N} G(z) = \frac{1}{N} \sum_{i=1}^{N} G_{ii}(z), \quad G(z) := \frac{1}{H - z}. \]
In addition to (1.2), we are able to prove that not only the sum but also each diagonal element $G_{ii}(z)$ is given by the semicircle law, but the precision is weaker:
\[ \max_i |G_{ii}(z) - m_{sc}(z)| \lesssim \frac{C}{\sqrt{N \eta}}, \quad z = E + i\eta. \tag{1.6} \]
Finally, we can also show that the off-diagonal resolvent elements are small:
\[ \max_{ij} |G_{ij}(z)| \lesssim \frac{C}{\sqrt{N \eta}} \tag{1.7} \]
with logarithmic corrections [7] (see Theorem 2.19 in § 2.4). In our previous papers [9], [8] the constant $C$ in (1.6) and (1.7) depended on $\kappa$, that is, the estimates deteriorated near the spectral edge as an inverse power of $\kappa$, the exponent depending on whether a positive uniform lower bound $\delta_+ > 0$ is available or not. For more general Wigner matrices, for instance, for band matrices, we obtain the same estimates, but $M$ replaces $N$ on the right-hand sides of (1.6) and (1.7), and $C$ depends on $\kappa$. The precise statements are given in Theorem 2.5.

The asymptotics of the Stieltjes transform can be translated into the asymptotics of the counting function (for example, see Theorem 2.6) or into a result on the location of the eigenvalues (Theorem 2.7). Moreover, the local semicircle law easily implies that the eigenvectors are completely delocalized (see § 2.5).

1.1.2. Summary of § 3: Main results on bulk universality with Gaussian component. Bulk universality refers to the fact that the local eigenvalue statistics, that is, the correlation functions of the eigenvalues rescaled by a factor $N$, or the distribution of the gap between consecutive eigenvalues, exhibit a universal behaviour which is solely determined by the symmetry class of the ensemble.
Bulk universality was first proved for Gaussian Wigner ensembles when the matrix elements $h_{ij}$ are independent identically distributed (i.i.d.) Gaussian random variables by Dyson [10] and Mehta [11]. The Gaussian character facilitates the explicit calculations that are needed to identify the limiting correlation functions (for example, the celebrated sine kernel for the Hermitian case). The key fact is that the joint distribution function for the eigenvalues of such ensembles is explicit, and it contains a Vandermonde determinant structure from which local universality can be deduced (see (1.44)).

It is a natural idea to consider a broader class of matrices that still have some Gaussian character. Useful here is the concept of Gaussian divisible ensembles, that is, where the probability law of each matrix element contains a Gaussian component (Gaussian convolution).

One approach with Gaussian convolutions is to push the explicit calculations further by finding a similar Vandermonde structure. Based on an earlier paper of Brézin and Hikami [12], Johansson [13] found a representation formula for correlation functions and was able to prove bulk universality for Gaussian divisible matrices. For an algebraic reason, this method is available for the Hermitian case only.

The size of the Gaussian component in [13] was substantial; it was of the same order as the non-Gaussian part. Using our local semicircle law and a slightly modified version of an explicit representation formula of Johansson [13], we were able to prove bulk universality for Hermitian Wigner matrices with a tiny Gaussian component of variance $O(N^{-1+\varepsilon})$ with an improved formula in [14] (§1.6.1).

The second approach (sketched in §1.6.2 and elaborated in §3) is to embed the Gaussian divisible ensemble into a stochastic flow of matrices, and use the key observation of Dyson [15] that under this flow the eigenvalues perform a specific stochastic dynamics with a logarithmic interaction, the celebrated Dyson Brownian motion. Eventually the dynamics relaxes to an equilibrium, which is the well-known Gaussian model (GUE, GOE, or GSE). The main idea is that the local relaxation is much faster, that is, the local statistics of the eigenvalues already reach their equilibrium within a very short time $t = N^{-\varepsilon}$ (with an explicit $\varepsilon > 0$). In fact, Dyson [15] predicted that the time scale to local equilibrium is of order $N^{-1}$, which we eventually proved in [7]. Our main result states that the local correlation functions of Gaussian divisible matrices with a small Gaussian component coincide with the correlation functions of purely Gaussian ensembles.

This result can be formulated in a general setup and viewed as strong local ergodicity of the Dyson Brownian motion or, in fact, of any one-dimensional stochastic dynamics of particles with logarithmic interaction. This general formulation appeared first in [6] and will be given in Theorem 3.3, but most of the key ideas were invented in [16]. For the application of this general principle to random matrices we need certain a priori information about the location of the eigenvalues, which we obtain from the local semicircle law. In particular, with the help of this idea the bulk universality for symmetric Wigner matrices was first proved in [16]. The cases of quaternion self-dual and sample covariance matrices were treated in [6].

1.1.3. Summary of §4: Main results on the removal of the Gaussian component. To prove the universality of any Wigner ensemble, we need to compare it with a Gaussian divisible ensemble for which universality has already been proved.
Such a comparison principle is plausible if the Gaussian component is small, and a perturbation argument can indeed be applied. It is essentially a density argument, stating that Gaussian divisible ensembles are sufficiently ‘dense’ in the space of all Wigner ensembles.

The first result of this type used a reverse heat flow argument [14], where we showed that any smooth distribution can be approximated with a very high precision by a Gaussian divisible distribution. Combining this method with the bulk universality for Hermitian Wigner matrices with a Gaussian component of variance of order $O(N^{-1+\epsilon})$, we were able to prove bulk universality for any Wigner ensemble under the condition that the distribution of the matrix elements is smooth.

A more robust approach is the Green function comparison theorem from [9], which states that for two matrix ensembles the joint distributions of the Green functions coincide, provided that the first four moments of the probability law of the matrix elements are identical or very close. The spectral parameter $z$ can have a very small imaginary part $\text{Im} \, z \sim N^{-1-\epsilon}$, that is, these Green functions can detect individual eigenvalues. The precise statement is given in Theorem 4.1. The key input is the local semicircle law involving individual matrix elements of the resolvent, (1.6)–(1.7).

The combination of the results in §1.1.2 on the bulk universality of Gaussian divisible matrices and the Green function comparison theorem gives bulk universality for any Wigner ensemble by a simple matching argument [8]. This method applies even to matrices with comparable variances (1.17). The only condition in the approach is a subexponential decay for the tail of the probability law of the matrix elements (2.32). In fact, this condition can be relaxed to a sufficiently fast polynomial decay, but for simplicity we will not pursue this direction.

The four-moment condition was first mentioned by Tao and Vu [17] in the four-moment theorem for eigenvalues (Theorem 1.5). Their key technical input is also the local semicircle law and its corollary on delocalization of eigenvectors. They used this result to prove the universality for Hermitian Wigner matrices without a smoothness condition but under some moment and support conditions that especially excluded the Bernoulli distribution. Bulk universality for Hermitian Wigner matrices, including the Bernoulli case, was first proved in [18] after combining the results of [14] and [17].

Finally, in §5 we state the main result (Theorem 5.1) on bulk universality and we summarize how its proof follows from the previous sections. Currently the local relaxation flow method combined with the Green function comparison theorem gives the most general approach to bulk universality. This path not only proves bulk universality for general Wigner ensembles but also offers a conceptual understanding of how universality emerges from simple principles.

In §§1.2–1.7 we review several facts, results, and methods in connection with Wigner random matrices and some related ensembles. These subsections are meant to provide general background information. In §1.6 we also explain the key new ideas listed above in more detail and give a summary of various results on bulk universality. The reader wishing to focus only on the most recent developments can skip §§1.2–1.7 and jump to §2.
1.2. Wigner matrix ensembles. A central question in probability theory is the universality of cumulative statistics of a large set of independent data. Given an array of \(N\) independent random variables
\[
(X_1, \ldots, X_N), \tag{1.8}
\]
one forms linear statistics like the mean or the fluctuation
\[
\bar{X}^{(N)} := \frac{1}{N} \sum_{j=1}^{N} X_j, \quad S^{(N)} := \frac{1}{\sqrt{N}} \sum_{j=1}^{N} (X_j - EX_j). \tag{1.9}
\]
Under very general conditions, a universal pattern emerges as \(N \to \infty\): the mean converges to its expectation, in particular, it becomes deterministic,
\[
\bar{X}^{(N)} \to \lim_{N \to \infty} EX^{(N)},
\]
assuming that the latter limit exists (law of large numbers). Moreover, the fluctuation \(S^{(N)}\) converges to a centered normal Gaussian random variable \(\xi\)
\[
S^{(N)} \to \xi \quad \text{(in distribution)} \tag{1.10}
\]
(central limit theorem), that is, the density function of \(\xi\) is given by
\[
f(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right).
\]
The variance \(\sigma^2\) of \(\xi\) is given by the average of the variances of the \(X_j\),
\[
\sigma^2 := \lim_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} \sigma_j^2, \quad \sigma_j^2 := E[(X_j - EX_j)^2].
\]
In particular, for independent identically distributed (i.i.d.) random variables, \(\bar{X}^{(N)}\) converges to the common expectation value of the \(X_j\), and \(S^{(N)}\) converges to the centered normal distribution with the common variance of the \(X_j\).

The emergence of a single universal distribution, the Gaussian distribution, is a remarkable fact of nature. It shows that large systems with many independent components in a certain sense behave identically, irrespective of the details of the distributions of the components.

It is natural to generalize this question of universality from arrays (1.8) to double arrays, that is, to matrices:
\[
X^{(N,M)} = \begin{pmatrix}
X_{11} & X_{12} & \cdots & X_{1N} \\
X_{21} & X_{22} & \cdots & X_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
X_{M1} & X_{M2} & \cdots & X_{MN}
\end{pmatrix} \tag{1.11}
\]
with independent entries. The statistics in question should involve a quantity which reflects the matrix character and is influenced by all entries, for example, the (Euclidean) norm of \(X^{(N,M)}\). Although the norm of each random realization of \(X^{(N,M)}\) may vary, it is known, for example, that in the limit as \(N, M \to \infty\) with
$N/M \to d$, where $0 < d \leq 1$ is fixed, it becomes deterministic, for instance, we have [19], [20]
\[
\frac{1}{\sqrt{M}} \|X^{(N,M)}\| \to \sigma(1 + \sqrt{d}),
\]  
(1.12)
assuming that the matrix elements are centered, $EX_{ij} = 0$, and their average variance is $\sigma^2$. Note that the typical size of the norm of $X^{(N,M)}$ is only of order $\sqrt{M}$ despite the fact that the matrix has dimension $M \times N$ and is filled with elements of size $O(1)$. If the matrix elements were strongly correlated, then the norm could be of order $M$. For example, in the extreme case, if all the elements were the same, $X_{ij} = X$, then $\|X^{(N,M)}\| \sim M$. The independence of the matrix elements prevents such a ‘conspiracy’ and reduces the typical size of the matrix norm by a factor of $\sqrt{M}$, which is analogous to what the central limit theorem asserts (note the $\sqrt{N}$ normalization in (1.9)).

Matrices offer a much richer structure than studying only their norm reveals. Under the assumption that $M = N$ the most important characteristics of a square matrix are the eigenvalues and eigenvectors. As (1.12) suggests, it is convenient to assume a zero expectation for the matrix elements and to rescale the matrix by a factor $N^{-1/2}$ so as to have a norm of order 1. For most of this survey we will therefore consider large $N \times N$ square matrices of the form
\[
H = H^{(N)} = \begin{pmatrix}
h_{11} & h_{12} & \cdots & h_{1N} \\
h_{21} & h_{22} & \cdots & h_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
h_{N1} & h_{N2} & \cdots & h_{NN}
\end{pmatrix}
\]  
(1.13)
with centered entries,
\[
Eh_{ij} = 0, \quad i, j = 1, \ldots, N.
\]  
(1.14)
As for the normalization, we assume that the matrix of variances
\[
\Sigma := \begin{pmatrix}
\sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1N}^2 \\
\sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2N}^2 \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{N1}^2 & \sigma_{N2}^2 & \cdots & \sigma_{NN}^2
\end{pmatrix}, \quad \sigma_{ij}^2 := E|h_{ij}|^2,
\]  
(1.15)
is doubly stochastic, that is, for every $i = 1, \ldots, N$ we have
\[
\sum_j \sigma_{ij}^2 = \sum_j \sigma_{ji}^2 = 1.
\]
The most natural example is the **mean-field model**, when
\[
\sigma_{ij}^2 = \frac{1}{N}, \quad i, j = 1, 2, \ldots, N,
\]
that is, each matrix element is of size $h_{ij} \sim N^{-1/2}$. This corresponds to the standard Wigner matrix. For most of this exposition the reader can restrict attention to this case.
Random matrices are typically subject to some symmetry restrictions, for example, we will consider symmetric \((h_{ij} = h_{ji} \in \mathbb{R})\) or Hermitian \((h_{ij} = h_{ji} \in \mathbb{C})\) random matrices. We will mostly assume that the matrix elements are independent, up to the symmetry requirement (that is, in the case of symmetric or Hermitian matrices, the variables \(\{h_{ij} : i \leq j\}\) are independent).

**Definition 1.1.** An \(N \times N\) symmetric or Hermitian random matrix \((1.13)\) is called a universal Wigner matrix (ensemble) if the entries are centered \((1.14)\), their variances \(\sigma_{ij}^2 = \mathbb{E}|h_{ij}|^2\) satisfy
\[
\sum_j \sigma_{ij}^2 = 1, \quad i = 1, \ldots, N, \tag{1.16}
\]
and \(\{h_{ij} : i \leq j\}\) are independent. An important subclass of universal Wigner ensembles is the class of so-called generalized Wigner matrices (ensembles) for which it is required in addition that the variances are comparable, that is,
\[
0 < C_{\text{inf}} \leq N\sigma_{ij}^2 \leq C_{\text{sup}} < \infty, \quad i, j = 1, \ldots, N, \tag{1.17}
\]
for some fixed positive constants \(C_{\text{inf}}\) and \(C_{\text{sup}}\). In the special case \(\sigma_{ij}^2 = 1/N\) we recover the original definition of Wigner matrices or Wigner ensembles [21].

The most prominent Wigner ensembles are the Gaussian Orthogonal Ensemble (GOE) and the Gaussian Unitary Ensemble (GUE); that is, symmetric and Hermitian Wigner matrices with rescaled matrix elements \(\sqrt{N} h_{ij}\) that are standard Gaussian variables (in the Hermitian case, \(\sqrt{N} h_{ij}\) is a standard complex Gaussian variable, that is, \(\mathbb{E}|\sqrt{N} h_{ij}|^2 = 1\)).

For simplicity of the presentation, we assume in the case of Wigner ensembles that the \(h_{ij}, i < j\), are identically distributed (that is, not just that their variances are the same). In this case we fix a distribution \(\nu\) and assume that the rescaled matrix elements \(\sqrt{N} h_{ij}\) are distributed according to \(\nu\). Depending on the symmetry type, the diagonal elements may have a slightly different distribution, but we will omit this subtlety from the discussion. The distribution \(\nu\) will be called the single-entry distribution of \(H\).

We will sometimes mention a special class of universal Wigner matrices that have a band structure; they will be called random band matrices. The variances are given by
\[
\sigma_{ij}^2 = W^{-1} f \left( \frac{|i - j|_N}{W} \right), \tag{1.18}
\]
where \(W \gg 1, f: \mathbb{R} \to \mathbb{R}_+\) is a bounded non-negative symmetric function with \(\int f(x) \, dx = 1\), and \([i - j]_N \in \mathbb{Z}\) is defined by the property that \([i - j]_N \equiv i - j \mod N\) and \(-N/2 < [i - j]_N \leq N/2\). We note that the relation \((1.16)\) holds only asymptotically as \(W \to \infty\), but this can be remedied by an appropriate (and irrelevant) rescaling. One can even consider \(d\)-dimensional band matrices, where the rows and columns are labelled by a finite lattice \(\Lambda \subset \mathbb{Z}^d\), and \(\sigma_{ij}^2\) depends only on the difference \(i - j\) for any \(i, j \in \Lambda\).

Another class of random matrices, predating even Wigner, consists of the random covariance matrices. These are matrices of the form
\[
H = X^* X, \tag{1.19}
\]
where \( X \) is a rectangular \( M \times N \) matrix of the form (1.11) with centered i.i.d. entries with variance \( \mathbb{E}|X_{ij}|^2 = M^{-1} \). Note that the matrix elements of \( H \) are not independent, but are generated from the independent matrix elements of \( X \) in a straightforward way. These matrices appear in statistical samples and were first considered by Wishart [22]. In the case when the \( X_{ij} \) are centered Gaussian variables, random covariance matrices are called Wishart matrices or ensembles.

1.3. Motivations: From Schrödinger operators to \( \zeta \)-functions. We will primarily study the eigenvalue statistics of large random matrices, though some results about eigenvectors will also be mentioned. The main physical motivation is that a random matrix can model the Hamilton operator of a disordered quantum system. The symmetry properties of \( H \) stem from this consideration: symmetric matrices represent Hamiltonians of systems with time reversal invariance (for instance, no magnetic field), Hermitian matrices correspond to systems without time reversal symmetry. (There is a third class of matrices, the quaternion self-dual matrices, most prominently modelled by the Gaussian Symplectic Ensemble (GSE), that describe systems with odd spin and no rotational symmetry, but we will not discuss it in detail.)

Wigner originally introduced random matrices to mimic the eigenvalues of the unknown Hamiltonian of heavy nuclei. Lacking any information, he assumed that the matrix elements are i.i.d. random variables subject to the Hermitian condition. His very bold vision was that, although such a crude approximation cannot predict individual energy levels (eigenvalues) of the nucleus, their statistical properties may be characteristic of some global feature shared by any nucleus. By comparing measured data of energy levels of nuclei with numerical calculations of eigenvalues of certain random matrices, he found that the level statistics, that is, the distribution of the energy gaps between neighbouring energy levels (eigenvalues), show remarkable coincidence and robustness. In particular, he observed that energy levels tend to repel each other, a significant difference from the level statistics of completely uncorrelated random points (Poisson point process). A similar feature was found for random matrices: even Wigner matrices that are ‘as stochastic as possible’ delivered plots of strongly correlated (repelling) eigenvalues. This correlation is due to the underlying fundamental symmetry of the matrix ensemble. In particular, symmetric and Hermitian matrices were found to have a different strength of level repulsion, but within a fixed symmetry class a universal pattern emerged. For more details on the history of this remarkable discovery, see [1].

Universality of local eigenvalue statistics is believed to hold for a much broader class of matrix ensembles than we have introduced. There is no reason to think that the matrix elements of the Hamiltonian of the heavy nuclei are indeed i.i.d. random variables. Conceivably, the matrix elements need not be completely independent or identically distributed for universality. There is little known about matrices with correlated entries, apart from the unitary invariant ensembles (§1.5.3) that represent a very specific correlation. In the case of a certain class of Wigner matrices with weakly correlated entries, the semicircle law and its Gaussian fluctuation have been proved [23], [24].

Various classes of random matrices with independent but not identically distributed entries have been much more studied. The most prominent example is
the Anderson tight binding model \([25]\), that is, a Schrödinger operator, \(-\Delta + \lambda V\), on a regular square lattice \(\mathbb{Z}^d\) with a random one-site potential \(V\) and disorder strength \(\lambda\). This model describes electron propagation (conductance) in an ionic lattice with a disordered environment. Restricted to a finite box, it can be represented by a matrix whose diagonal elements are i.i.d. random variables; the deterministic off-diagonal elements are given by the Laplacian.

The general formulation of the \textit{universality conjecture for random Schrödinger operators} states that there are two distinctive regimes, depending on the energy and the disorder strength. In the strong disorder regime, the eigenfunctions are localized and the local spectral statistics are Poisson. In the weak disorder regime, the eigenfunctions are delocalized and the local statistics coincide with those of a Gaussian matrix ensemble. Separate conjectures, which will not be discussed here, relate these two regimes to chaotic versus integrable behaviour of the underlying classical dynamical system. According to the Berry–Tabor conjecture \([26]\), Poisson statistics of eigenvalues should emerge from quantization of integrable classical dynamics, while random matrix theory stems from quantization of chaotic classical dynamics (Bohigas, Giannoni, Schmit \([27]\)).

Returning to the more concrete Anderson model, we note that in space dimensions three or higher and for weak randomness the model is conjectured to exhibit a metal-insulator transition, that is, in \(d \geq 3\) dimensions the eigenfunctions of \(-\Delta + \lambda V\) are delocalized for small \(\lambda\), while they are localized for large \(\lambda\). It is a fundamental open mathematical question to establish this transition.

The localization regime at large disorder or near the spectral edges has been well understood by Fröhlich and Spencer with the multiscale technique \([28]\), \([29]\) and later by Aizenman and Molchanov using the fractional-moment method \([30]\), and many other works have since contributed to this field. In particular, it has been established that the local eigenvalue statistics are Poisson \([31]\) and that the eigenfunctions are exponentially localized, with an upper bound on the localization length that diverges as the energy parameter approaches the presumed phase transition point \([32]\), \([33]\).

Progress in studying the delocalization regime has been much slower. For the Bethe lattice, corresponding to the infinite-dimensional case, delocalization was established in \([34]\)–\([36]\) (the eigenvalue statistics here are Poisson, in apparent contradiction to general conjectures, but for a quite specific and well-understood reason \([37]\)). In finite dimensions only partial results are available. The existence of an absolutely continuous spectrum (that is, extended states) has been shown for a rapidly decaying potential, corresponding to a scattering regime \([38]\)–\([40]\). Diffusion has been established for a heavy quantum particle immersed in a phonon field in \(d \geq 4\) dimensions \([41]\). For the original Anderson Hamiltonian with a small coupling constant \(\lambda\) the eigenfunctions have a localization length of at least \(\lambda^{-2}\) \([42]\). The time and space scale \(\lambda^{-2}\) corresponds to the kinetic regime where the quantum evolution can be modelled by a linear Boltzmann equation \([43]\), \([44]\). Beyond this time scale the dynamics is diffusive. This has been established in the scaling limit \(\lambda \to 0\) up to time scales \(t \sim \lambda^{-2-\kappa}\) with an explicit \(\kappa > 0\) in \([45]\).

There are no rigorous results on the local spectral statistics of the Anderson model in the disordered regime, but it is conjectured (and supported by numerous arguments in the physics literature, especially by supersymmetric methods; see \([46]\))
that the local correlation functions of the eigenvalues in the finite-volume Anderson model follow the GOE statistics in the thermodynamic limit. GUE statistics are expected if an additional magnetic field breaks the time-reversal symmetry of the Anderson Hamiltonian. Based on this conjecture, the local eigenvalue statistics are used to compute the phase diagram numerically. It is very remarkable that the random Schrödinger operator represented by a very sparse random matrix exhibits the same universality class as the full Wigner matrix, at least in a certain energy range.

Due to their mean-field character, Wigner matrices are simpler to study than the Anderson model, and they are always in the delocalization regime. In this survey we mainly focus on Wigner matrices, but we keep in mind the original motivation from general disordered systems. In particular, we will study not only eigenvalue statistics but also eigenvectors which are shown to be completely delocalized [16]. The local spectral statistics in the bulk are universal, that is, they agree with the statistics of the corresponding Gaussian ensemble (GOE, GUE, GSE), depending on the symmetry type of the matrix. This topic will be the main goal of this presentation.

In closing this subsection we mention some other possible research directions which we will not further pursue here. The list is incomplete.

A natural intermediate class of ensembles between the completely stochastic Wigner matrices and the Anderson model with diagonal randomness is the family of random band matrices. These are Hermitian or symmetric random matrices $H$ with independent but not identically distributed entries. The variance of $h_{ij}$ depends only on $|i - j|$, and it becomes negligible if $|i - j|$ exceeds a given parameter $W$, the band-width; for example, $\sigma_{ij}^2 = \mathbb{E}|h_{ij}|^2 \sim \exp(-|i - j|/W)$. It is conjectured [47] that the system is completely delocalized if $W \gg \sqrt{N}$, and otherwise the localization length is $W^2$. Moreover, for narrow bands, $W \ll \sqrt{N}$, the local eigenvalue statistics are expected to be Poisson, while for broad bands, $W \gg \sqrt{N}$, they should be given by GUE or GOE statistics, depending on the symmetry class. Localization properties of $H$ for $W \gg N^{1/8}$ and an $O(W^8)$ upper bound on the localization length have been shown by Schenker [48], but this result does not include local statistics. From the delocalization side, with Knowles we recently proved [40], [50] diffusion up to time scale $t \ll W^{1/3}$, which implies that the localization length is at least $W^{1+1/6}$.

We mention that universality of local eigenvalue statistics is often investigated by supersymmetric techniques in the physics literature. These methods are extremely powerful for extracting results by saddle point computations, but the analysis justifying the saddle point approximation still lacks mathematical rigour. So far only the density of states has been investigated rigorously by using this technique [51]. Quantum diffusion can also be studied by supersymmetry, and certain intermediate models can be rigorously analyzed [52], [53].

Finally, we point out that while we have focused on the physical motivations coming from disordered quantum systems, random matrices appear in many other branches of physics and mathematics. They form a fundamental object of nature with an extremely rich structure. The most remarkable connection is with the ζ-function. It is conjectured that the roots of the Riemann ζ-function $\zeta(s) := \sum_{n=1}^{\infty} n^{-s}$, which lie on the vertical line $\text{Re } s = 1/2$, have the same local statistics.
as the GUE (after appropriate rescaling). A survey and further references to many instances of numerical evidence are found in the classical book of Mehta [1].

1.4. Eigenvalue density and delocalization. For a symmetric or Hermitian matrix $H$ let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$ denote the eigenvalues. They form a random point process on the real line with a distribution generated from the joint probability law of the matrix elements. Since the functional relation between the matrix elements and the eigenvalues is highly non-trivial, the product measure on the entries turns out to generate a complicated and highly correlated measure for the eigenvalues. Our main goal is to understand this induced measure.

Under the chosen normalization (1.16) the typical size of the eigenvalues is of order 1. We will prove a much more precise statement later, but for now it is instructive to get a rough feeling about the size by computing $\text{Tr} \, H^2$ in two ways:

$$\sum_{i=1}^{N} \lambda_i^2 = \text{Tr} \, H^2 = \sum_{i,j=1}^{N} |h_{ij}|^2.$$ 

Taking the expectation and using (1.16), we have

$$\frac{1}{N} \sum_i \mathbb{E}\lambda_i^2 = \frac{1}{N} \sum_{i,j} \sigma_{ij}^2 = 1,$$

that is, in an average sense $\mathbb{E}\lambda_i^2 = 1$.

1.4.1. Wigner semicircle law and other canonical densities. The empirical distribution of eigenvalues follows a universal pattern, the Wigner semicircle law. To formulate it more precisely, we note that the typical spacing between neighbouring eigenvalues is of order $1/N$, so in a fixed interval $[a, b] \subset \mathbb{R}$, one expects macroscopically many (of order $N$) eigenvalues. More precisely, it can be shown (the first proof was given by Wigner [21]) that for any fixed real numbers $a \leq b$,

$$\lim_{N \to \infty} \frac{1}{N} \#\{i : \lambda_i \in [a, b]\} = \int_a^b \rho_{\text{sc}}(x) \, dx, \quad \rho_{\text{sc}}(x) := \frac{1}{2\pi} \sqrt{(4-x^2)^+}, \quad (1.20)$$

where $(a)^+ := \max\{a, 0\}$ denotes the positive part of the number $a$. Note the emergence of the universal density, the semicircle law, which is independent of the details of the distribution of the matrix elements.

The semicircle law is characteristic for universal Wigner matrices (see Definition 1.1). For random square matrices with independent entries but without symmetry (that is, $h_{ij}$ are independent for all $i, j$), a similar universal pattern emerges, the circular law. For example, if the $h_{ij}$ are centered i.i.d. random variables with a common variance $\sigma_{ij}^2 = N^{-1}$, then the empirical density of the eigenvalues converges to the uniform measure on the unit disk in the complex plane [54]. But if independence is dropped, one can get many different density profiles.

For example, in the case of the random covariance matrices (1.19) the empirical density of the eigenvalues $\lambda_i$ of $H$ converges to the Marchenko–Pastur law [19] in
the limit as \( M, N \to \infty \) with \( d = N/M \) fixed, \( 0 \leq d \leq 1 \):
\[
\lim_{N \to \infty} \frac{1}{N} \# \{ i : \lambda_i \in [a, b] \} = \int_a^b \varrho_{\text{MP}}(x) \, dx,
\]
\[
\varrho_{\text{MP}}(x) := \frac{1}{2\pi d} \sqrt{\frac{[(\lambda_+ - x)(x - \lambda_-)]_+}{x^2}},
\]
(1.21)
where \( \lambda_\pm := (1 \pm \sqrt{d})^2 \) are the spectral edges. We note that in the case \( M \leq N \) the matrix \( H \) has macroscopically many zero eigenvalues, for otherwise the spectra of \( XX^* \) and \( X^*X \) coincide, so that the Marchenko–Pastur law can be applied to all non-zero eigenvalues, with the roles of \( M \) and \( N \) exchanged.

1.4.2. The moment method. The eigenvalue density is commonly approached via the fairly robust moment method (see [3] for an exposé), which was also the original approach of Wigner in proving the semicircle law [21]. For example, for Hermitian Wigner matrices it consists of computing traces of high powers of \( H \), that is,
\[
E \text{Tr } H^{2k},
\]
by expanding the product as
\[
E \sum_{i_1,i_2,\ldots,i_{2k}} h_{i_1i_2}h_{i_2i_3}\cdots h_{i_{2k}i_1}
\]
and noting that each factor \( h_{xy} \) must be paired with at least one other copy \( h_{yx} = \bar{h}_{xy} \), for otherwise the expectation value is zero. The possible index sequences that satisfy these pairing conditions can be classified according to their complexity, and it turns out that the main contribution comes from the so-called backtracking paths. These are index sequences \( i_1i_2i_3\cdots i_{2k}i_1 \) returning to the original index \( i_1 \) that can be successively generated by a substitution rule
\[
a \to aba, \quad b \in \{1, \ldots, N\}, \quad b \neq a,
\]
with an arbitrary index \( b \). These index sequences satisfy the pairing condition in an obvious manner, and it turns out that they involve the largest possible number \( (N^k) \) of independent indices. The number of backtracking paths is explicitly given by the Catalan numbers \( C_k = \frac{1}{k+1} \binom{2k}{k} \), so \( E \text{Tr } H^{2k} \) can be computed fairly precisely for each finite \( k \):
\[
\frac{1}{N} E \text{Tr } H^{2k} = \frac{1}{k+1} \binom{2k}{k} + O_k(N^{-2}).
\]
(1.22)
We note that the number \( N^k \) of independent labels exactly cancels the size of the \( k \)-fold product of the variances \( (E|h|^2)^k = N^{-k} \). If the distribution of the matrix elements is symmetric, then the traces of odd powers all vanish, since they can never satisfy the pairing condition. Without the symmetry condition, the traces of odd powers are non-zero but negligible.
We will compute the trace of the resolvent, or the Stieltjes transform of the empirical density

\[ \varrho_N(dx) := \frac{1}{N} \sum_{j=1}^{N} \delta(x - \lambda_j) \]

of the eigenvalues, that is, we define

\[ m(z) = m_N(z) := \frac{1}{N} \operatorname{Tr} \frac{1}{H - z} = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{\lambda_j - z} = \int_{\mathbb{R}} \frac{d\varrho_N(x)}{x - z} \]  

(1.23)

for any \( z = E + i\eta, \ E \in \mathbb{R}, \ \eta > 0 \). For large \( z \) one can expand \( m_N \) as follows:

\[ m_N(z) = \frac{1}{N} \operatorname{Tr} \frac{1}{H - z} = -\frac{1}{Nz} \sum_{m=0}^{\infty} \left( \frac{H}{z} \right)^m, \]

(1.24)

so after taking the expectation using (1.22), and neglecting the error terms, we get that

\[ \mathbb{E} m_N(z) \approx -\sum_{m=0}^{\infty} \frac{1}{k+1} \binom{2k}{k} z^{-(2k+1)}, \]

(1.25)

which, after some calculus, can be identified as the power series of \( (-z + \sqrt{z^2 - 4})/2 \). The approximation becomes exact in the limit as \( N \to \infty \). Although the expansion (1.24) is valid only for large \( z \), given that the limit is an analytic function of \( z \), one can extend the relation

\[ \lim_{N \to \infty} \mathbb{E} m_N(z) = \frac{1}{2} \left( -z + \sqrt{z^2 - 4} \right) \]

by analytic continuation to the whole of the upper half-plane \( z = E + i\eta, \ \eta > 0 \). It is an easy exercise to see that this is exactly the Stieltjes transform of the semicircle density, that is,

\[ m_{sc}(z) := \frac{1}{2} \left( -z + \sqrt{z^2 - 4} \right) = \int_{\mathbb{R}} \frac{\varrho_{sc}(x) dx}{x - z}. \]

(1.26)

The square root function is chosen with a branch cut in the segment \([-2, 2]\), so that \( \sqrt{z^2 - 4} \sim z \) at infinity. This guarantees that \( \operatorname{Im} m_{sc}(z) > 0 \) for \( \operatorname{Im} z > 0 \). Since the Stieltjes transform determines the measure uniquely, and pointwise convergence of Stieltjes transforms implies weak convergence of measures, we get that

\[ \mathbb{E} d\varrho_N(x) \to \varrho_{sc}(x) dx. \]

(1.27)

With slightly more effort one can show that

\[ \lim_{N \to \infty} m_N(z) = \frac{1}{2} \left( -z + \sqrt{z^2 - 4} \right) \]

(1.28)

holds with high probability, that is, the convergence holds also in probability and not just in expectation. For more details, see [3].
1.4.3. The local semicircle law. The moment method can typically identify the resolvent for any fixed $z$ and thus give the semicircle law as a weak limit, that is, (1.20) will hold for any fixed interval $I := [a, b]$ as $N \to \infty$. However, a fixed interval $I$ with length $|I|$ typically contains of order $N|I|$ eigenvalues. It is natural to ask whether the semicircle law holds locally as well, that is, for intervals whose length may shrink with $N$, but still with $N|I| \gg 1$. Indeed, the semicircle law is a type of law of large numbers that should require only that the number of random objects in consideration goes to infinity. Due to the formula

$$\text{Im} \ m_N(z) = \frac{1}{N} \sum_{i=1}^{N} \frac{\eta}{(\lambda_i - E)^2 + \eta^2} \sim \frac{\pi}{N} \sum_{i=1}^{N} \delta_\eta(\lambda_i - E), \quad z = E + i\eta,$$

where $\delta_\eta$ denotes an approximate delta function on the scale $\eta$, we see that knowing the Stieltjes transform for some $z \in \mathbb{C}$ with $\text{Im} \ z = \eta$ is essentially equivalent to knowing the local density on the scale $\eta$, that is, in an interval of length $|I| \sim \eta$.

In [55]–[57] we proved that the local semicircle law holds on the smallest possible scale of $\eta \gg 1/N$, that is, the limit (1.20) holds even if the length of $I = [a, b]$ is essentially of order $1/N$, and hence it typically contains only a large but finite number of eigenvalues. This will be the key technical input for further investigations on local spectral statistics. There are several versions of the local semicircle law. We will give three precise statements: Theorem 1.9 (from [57]), Theorem 2.5 (from [8]), and Theorem 2.19 (from [7]).

The method of proof is different from the moment method, but we still work with the resolvent, or the Stieltjes transform. The key observation (see also several previous papers, for example, [58], [19]) is that the Stieltjes transform $m_{sc}(z)$ of the semicircle density $\varrho_{sc}$ satisfies the simple quadratic equation

$$m_{sc}(z) + \frac{1}{z + m_{sc}(z)} = 0,$$  \hspace{1cm} (1.29)

and $m_{sc}(z)$ is chosen from the two possible solutions as described after (1.26). The strategy is that by expanding the empirical Stieltjes transform $m_N(z)$ in (1.23) with respect to minors of $H$, we prove that $m_N$ satisfies the self-consistent equation (1.29) approximately and with a high probability:

$$m_N(z) + \frac{1}{z + m_N(z)} \approx 0.$$

Then we conclude the proof that $m_N \approx m_{sc}$ by invoking the stability of the equation (1.29). Since the stability deteriorates near the edges, $E = \text{Re} \ z \approx \pm 2$, the estimate will be weaker there, indicating that the eigenvalue fluctuation is larger at the edge.

Our best results in this direction were obtained in [7] (which is partly a streamlined version of [9], [8]), where not only the trace of the Green function (1.23) but also individual diagonal elements were shown to be given by the semicircle law. The results were already listed informally in §1.1.1, where we pointed out that they hold also for universal Wigner matrices (see Theorems 2.5 and 2.19).

For universal Wigner ensembles Guionnet [59] and Anderson–Zeitouni [60] had already proved that the density of the eigenvalues converges to the Wigner semicircle law on a large scale; our result improves this to small scales. For example,
for band matrices (for the definition see (1.18)) with band width $W$ we get that the semicircle law holds down to energy scales $1/W$. The delocalization length is shown to be at least as large as the band width $W$. We note that a certain three-dimensional version of Gaussian band matrices was also considered by Disertori, Pinson, and Spencer [51] using the supersymmetric method. They proved that the expectation of the density of the eigenvalues is smooth and it coincides with the Wigner semicircle law up to a precision determined by the bandwidth.

1.4.4. Density of eigenvalues for invariant ensembles. There is another natural way to define probability distributions on symmetric or Hermitian matrices, apart from directly imposing a given probability law $\nu$ on their entries. This way involves defining a density function directly on the set of matrices:

$$P(H) \, dH := Z^{-1} \exp(-N \operatorname{Tr} V(H)) \, dH. \quad (1.30)$$

Here $dH = \prod_{i < j} dH_{ij}$ is the flat Lebesgue measure (in the case of Hermitian matrices and $i < j$, $dH_{ij}$ is Lebesgue measure on the complex plane $\mathbb{C}$). The function $V: \mathbb{R} \to \mathbb{R}$ is assumed to grow mildly at infinity (some logarithmic growth would suffice) to ensure that the measure defined in (1.30) is finite, and $Z$ is the normalization factor. Probability distributions of the form (1.30) are called invariant ensembles, since they are invariant under orthogonal or unitary conjugation (in the case of symmetric or Hermitian matrices, respectively). For example, in the Hermitian case, for any fixed unitary matrix $U$ the transformation $H \to U^* H U$ leaves the distribution (1.30) invariant thanks to the identity $\operatorname{Tr} V(U^* H U) = \operatorname{Tr} V(H)$.

Wigner matrices and invariant ensembles form two different universes with quite different mathematical tools available for their studies. In fact, these two classes are almost disjoint, Gaussian ensembles being the only invariant Wigner matrices. This is the content of the following lemma ([2] or Theorem 2.6.3 in [1]).

**Lemma 1.2.** Suppose that the symmetric or Hermitian matrix ensemble given in (1.30) has independent elements $h_{ij}$, $i \leq j$. Then $V(x)$ is a quadratic polynomial, $V(x) = ax^2 + bx + c$, with $a > 0$. This means that apart from a trivial shift and normalization, the ensemble is the GOE or GUE.

The density of the eigenvalues of the invariant ensemble (1.30) is determined by a variational problem [2]. It is given by the equilibrium density of a gas with logarithmic self-interaction and external potential $V$, that is, as the solution of the problem

$$\inf_\rho \left\{ \int_{\mathbb{R}} \int_{\mathbb{R}} \log |s - t|^{-1} \, \rho(ds) \, \rho(dt) + \int V(t) \, \rho(dt) \right\},$$

where the infimum is taken over all probability measures $\rho$. Under some mild conditions on $V$ the equilibrium measure is absolutely continuous, $\rho_{\text{eq}}(dt) = \rho_{\text{eq}}(t) \, dt$, and it has compact support. If $V$ is a polynomial, then the support consists of finitely many intervals. The empirical density of the eigenvalues converges to $\rho_{\text{eq}}$ in the sense of (1.20), where $\rho_{\text{sc}}$ is replaced by the function $\rho_{\text{eq}}$. It is an easy exercise to check that the solution of this variational problem for the Gaussian case $V(x) = x^2/2$ is indeed $\rho_{\text{sc}}$. 

1.4.5. Delocalization of eigenvectors. Apart from the statistics of the eigenvalues, one may also study the eigenvectors of a random matrix. In light of the universality conjecture about disordered systems explained in §1.3, it is a challenging question to test this hypothesis on the level of eigenvectors as well. Wigner matrices are mean-field models, and from physics intuition they are always in the delocalized regime. Of course, they are still finite matrices, so they cannot have an absolutely continuous spectrum, a standard signature for delocalization that people working with random Schrödinger operators are often looking for. But the delocalization of eigenvectors is a perfectly meaningful question for large but finite matrices as well. Surprisingly, this question was largely neglected both by the random matrix community and the random Schrödinger operator community within mathematics until Spencer raised it recently. He pointed out in a lecture that in the case of Gaussian ensembles a simple invariance argument proves that the eigenvectors \( v \in \mathbb{C}^N \) are completely delocalized in the sense that their \( \ell^4 \)-norm is \( \|v\|_4 \sim N^{-1/4} \) (assuming that \( \|v\|_2 = 1 \)). This is a signature of strong delocalization, since, on the one hand, by the Cauchy–Schwarz inequality we have

\[
N^{-1/2} \|v\|_2 = \left( \frac{1}{N} \sum_{i=1}^{N} |v_i|^2 \right)^{1/2} \leq \left( \frac{1}{N} \sum_{i=1}^{N} |v_i|^4 \right)^{1/4} = N^{-1/4} \|v\|_4,
\]

that is, \( \|v\|_4 \geq N^{-1/4} \) always holds, while on the other hand, this inequality is close to equality if all the coordinates of the eigenvector are of approximately the same size, \( |v_i| \sim N^{-1/2} \).

The simple invariance argument works only for the Gaussian case, where unitary invariance is present, but it is natural to ask whether eigenvectors of Wigner ensembles are also delocalized, and the answer is affirmative. We have proved (Corollary 3.2 in [57]) that if \( v \) is an \( \ell^2 \)-normalized eigenvector of a Wigner matrix \( H \) with eigenvalue \( \lambda \) away from the spectral edge, that is,

\[
Hv = \lambda v, \quad \lambda \in [-2 + \kappa, 2 - \kappa],
\]

for some \( \kappa > 0 \), then the \( \ell^p \)-norm of \( v \) for any \( 2 < p < \infty \) is bounded by

\[
\|v\|_p \leq QN^{-1/2 + 1/p},
\]

(1.31)

with a very high probability: the set of exceptional events is subexponentially small with respect to \( Q \) for large \( Q \). A similar bound with a logarithmic correction holds for \( p = \infty \) as well. The precise statement will be given in Theorems 2.21 and 2.22. It is essentially a straightforward corollary of the local semicircle law, Theorem 2.5, which was informally outlined in §1.4.3.

We note that if \( v \) is an \( \ell^2 \)-normalized eigenvector, then the size of the \( \ell^p \)-norm of \( v \) for \( p > 2 \) gives information about delocalization. Complete delocalization occurs when \( \|v\|_p \lesssim N^{-1/2 + 1/p} \), since this corresponds to the \( \ell^p \)-norm of the completely delocalized vector \( v = (N^{-1/2}, \ldots, N^{-1/2}) \). In contrast, a completely delocalized eigenvector \( v = (0, \ldots, 0, 1, 0, \ldots, 0) \) has \( \ell^p \)-norm 1.

1.5. Local statistics of eigenvalues: Previous results. A central question concerning random matrices is the universality conjecture, which states that local
statistics of eigenvalues of large $N \times N$ square matrices $H$ are determined by the symmetry type of the ensembles but are otherwise independent of the details of the distributions. It turns out that local statistics exhibit even stronger universality features than the eigenvalue density.

The terminology ‘local statistics’ refers to observables that can distinguish among individual eigenvalues. For all the ensembles we have considered so far, we used a normalization such that the typical eigenvalues remain in a compact set as $N \to \infty$; in other words, the limiting density function $\varrho$ was compactly supported. In this case the typical spacing between neighbouring eigenvalues is of order $N^{-1}$. This holds in the bulk of the spectrum, that is, at a positive distance away the spectral edges. The spectral edges are characterized by the points where $\varrho$ goes to zero. For example, for the Wigner semicircle distribution $\varrho_{sc}$ they are at $\pm 2$, for the Marchenko–Pastur distribution (1.21) they are at $\lambda_{\pm}$, and for certain invariant ensembles the support of the eigenvalue density might consist of several intervals, that is, it can have more than two spectral edges.

1.5.1. Bulk universality: the sine kernel and the gap distribution. To see individual eigenvalues and their joint distribution in the bulk of the spectrum, one needs to ‘zoom out’ the point process of the eigenvalues by magnifying it by a factor of $N$. We fix two real numbers $\alpha_1$ and $\alpha_2$, and an energy $E$ with $\varrho(E) > 0$, and we ask for the probability that there is an eigenvalue at $E + \frac{\alpha_1}{N\varrho(E)}$ and simultaneously an eigenvalue at $E + \frac{\alpha_2}{N\varrho(E)}$ (the normalization is chosen such that the typical number of eigenvalues between these two points is independent of $E$). It turns out that the answer is independent of the details of the ensemble and of the energy $E$, but depends only on the symmetry type. For example, for the Hermitian case, it is given by

$$P\left\{ \text{there are eigenvalues } \lambda \in E + \frac{\alpha_1 + d\alpha_2}{N\varrho(E)} \text{ and } \lambda' \in E + \frac{\alpha_2 + d\alpha_2}{N\varrho(E)} \right\} = \left[ 1 - \left( \frac{\sin \pi (\alpha_1 - \alpha_2)}{\pi (\alpha_1 - \alpha_2)} \right)^2 \right] d\alpha_1 d\alpha_2.$$  

(1.32)

The function on the right-hand side is obtained from the celebrated sine kernel and should be viewed as a $2 \times 2$ determinant of the form

$$\det (K(\alpha_i - \alpha_j))_{i,j=1}^2, \quad K(x) := \frac{\sin \pi x}{\pi x}.$$  

(1.33)

The explicit formula for the kernel $K$ in the symmetric case is more complicated (see [2]), but it is universal, and the correlation function has the same determinantal structure.

We note that (1.32) contains much more delicate information about the eigenvalues than the semicircle law (1.20). First, it is local information after magnification to a scale where individual eigenvalues matter. Second, it expresses a correlation between two eigenvalues. For example, since $\frac{\sin y}{y} \to 1$ as $y \to 0$, we see that the eigenvalues repel each other.

In general, the $k$th correlation functions (or $k$-point marginals) give information about the joint behaviour of a $k$-tuple of eigenvalues. Their definition is as follows
Definition 1.3. Let \( p_N(\lambda_1, \ldots, \lambda_N) \) be the joint symmetrized probability distribution of the eigenvalues. For any \( k \geq 1 \), the \( k \)-point correlation function is defined by

\[
p_N^{(k)}(\lambda_1, \ldots, \lambda_k) := \int_{\mathbb{R}^{N-k}} p_N(\lambda_1, \ldots, \lambda_k, \lambda_{k+1}, \ldots, \lambda_N) \, d\lambda_{k+1} \cdots d\lambda_N. \tag{1.34}
\]

Remark. We usually label the eigenvalues in increasing order. For the purpose of this definition, however, we do not impose this restriction and regard \( p_N(\lambda_1, \ldots, \lambda_N) \) as a symmetric function of the \( N \) variables \( \lambda = (\lambda_1, \ldots, \lambda_N) \) on \( \mathbb{R}^N \). Alternatively, one could consider the density \( \tilde{p}_N(\lambda) = \frac{N!}{N^N} p_N(\lambda) \cdot 1(\lambda \in \Xi(N)) \), where

\[
\Xi(N) := \{ \lambda_1 < \cdots < \lambda_N \} \subset \mathbb{R}^N.
\]

The significance of the \( k \)-point correlation functions is that they give the expectation value of observables (functions) \( O \) depending on \( k \)-tuples of eigenvalues via the formula

\[
\frac{(N-k)!}{N!} \mathbb{E} \sum_{i_1, \ldots, i_k=1}^N O(\lambda_{i_1}, \ldots, \lambda_{i_k}) = \int_{\mathbb{R}^k} O(x_1, \ldots, x_k) p_N^{(k)}(x_1, \ldots, x_k) \, dx_1 \cdots dx_k,
\]

where the summation is over all distinct indices \( i_1, \ldots, i_k \), and the prefactor is a normalization of the sum.

For example, the one-point function \( p_N^{(1)} \) expresses the density. In particular, choosing the observable to be the characteristic function of \([a, b]\), \( O(x) = 1(\lambda \in [a, b]) \), we have

\[
\frac{1}{N} \# \{ i : \lambda_i \in [a, b] \} = \frac{1}{N} \sum_{i=1}^N O(\lambda_i) = \int O(x) p_N^{(1)}(x) \, dx = \int_a^b p_N^{(1)}(x) \, dx.
\]

Therefore, the Wigner semicircle law (1.20) states that \( p_N^{(1)} \) converges weakly to \( g_{\text{sc}} \) as \( N \to \infty \).

In the Hermitian case the sine-kernel universality expresses the fact that the (weak) limit of the rescaled \( k \)-point correlation function as \( N \to \infty \) is given by the determinant of \( K(x) \) in (1.33), that is,

\[
\frac{1}{[\varrho(E)]^k} p_N^{(k)} \left( E + \frac{\alpha_1}{N \varrho(E)}, \ldots, E + \frac{\alpha_k}{N \varrho(E)} \right) \to \det \left( K(\alpha_i - \alpha_j) \right)_{i,j=1}^k \tag{1.35}
\]

for any fixed \( E \), in the sense of weak convergence of functions of the variables \( (\alpha_1, \ldots, \alpha_k) \).

Once the \( k \)-point correlation functions are identified, it is easy to derive limit theorems for other quantities related to individual eigenvalues. The most interesting one is the gap distribution, that is, the distribution of the difference \( \lambda_{j+1} - \lambda_j \) between neighbouring eigenvalues. We note that it apparently involves only two eigenvalues, but it is not expressible solely by a two-point correlation function, since the two eigenvalues must be consecutive. Nevertheless, the gap distribution can be expressed in terms of all the correlation functions as follows.
Fix an energy $E$ with $|E| < 2$. For $s > 0$ and for some $N$-dependent parameter $t$ with $1/N \ll t \ll 1$ let

$$\Lambda(s) = \Lambda_N(s) := \frac{1}{2Nt\varrho_{sc}(E)} \# \left\{ 1 \leq j \leq N-1 : \lambda_{j+1} - \lambda_j \leq \frac{s}{N\varrho_{sc}(E)} , \| \lambda_j - E \| \leq t \right\},$$

that is, the proportion of rescaled eigenvalue differences below a threshold $s$ in a large but still microscopic neighbourhood of an energy $E$. Let $K_\alpha$ be the operator acting on $L^2((0, \alpha))$ with integral kernel $K(x, y) := \frac{\sin \pi(x - y)}{\pi(x - y)}$. Then for any $E$ with $|E| < 2$ and for any $s > 0$ we have

$$\lim_{N \to \infty} E\Lambda_N(s) = \int_0^s p(\alpha) \, d\alpha, \quad p(\alpha) := \frac{d^2}{d\alpha^2} \det(1 - K_\alpha),$$

where $\det$ denotes the Fredholm determinant of the operator $1 - K_\alpha$ (note that $K_\alpha$ is a compact operator). The density function $p(s)$ of the nearest-neighbour eigenvalue spacing behaves, with a very good but not exact approximation (called the Wigner surmise), like $p(s) \approx \pi se^{-\pi s^2/4}/2$ for the symmetric case and like $p(s) \approx 32\pi^{-2}s^2e^{-4s^2/\pi}$ for the Hermitian case \cite{1}.

Note that this behaviour is in sharp contrast to the level-spacing statistics of Poisson point processes, where the corresponding density is $p(s) = e^{-s}$ (after rescaling the process so that the mean distance is 1). In particular, random matrices exhibit a level repulsion whose strength depends on the symmetry class (note the different behaviour of $p(s)$ for $s \approx 0$).

For the proof of (1.36) we can use the exclusion-inclusion formula to express

$$E\Lambda(s) = \frac{1}{2Nt} \sum_{m=2}^{\infty} (-1)^m \binom{N}{m} \int_{-t}^{t} \cdots \int_{-t}^{t} \frac{d^m v_1 \cdots d^m v_m}{\times p_N^{(m)}(E + v_1, \ldots, E + v_m)},$$

where $\varrho = \varrho_{sc}(E)$. After a change of variables,

$$E\Lambda(s) = \frac{1}{2Nt} \sum_{m=2}^{\infty} (-1)^m \int_{-Nt}^{Nt} \cdots \int_{-Nt}^{Nt} d\zeta_1 \cdots d\zeta_m$$

$$\times \left( \frac{N}{m} \right) \frac{1}{(\varrho)^m} p_N^{(m)} \left( u + \frac{z_1}{N\rho}, \ldots, u + \frac{z_m}{N\rho} \right) 1\left\{ \max |z_i - z_j| \leq \frac{s}{N\varrho} \right\}$$

$$\times \left( \frac{N}{m} \right) \frac{1}{(\varrho)^m} p_N^{(m)} \left( u + \frac{z_1}{N\rho}, u + \frac{z_1 + a_2}{N\rho}, \ldots, u + \frac{z_1 + a_m}{N\rho} \right),$$

(1.38)
where the factor $m$ comes from considering the integration sector $z_1 \leq z_j$, $j \geq 2$. Taking $N \to \infty$ and using (1.35), we get that

$$
\lim_{N \to \infty} E \Lambda(s) = \sum_{m=2}^{\infty} \frac{(-1)^m}{(m-1)!} \int_0^s da_2 \cdots \int_0^s da_m \det \left( \frac{\sin \pi (a_i - a_j)}{\pi (a_i - a_j)} \right)_{i,j=1}^m, \quad (1.39)
$$

where in the last determinant we set $a_1 = 0$. The interchange of the limit and the summation can be easily justified by an alternating series argument. We note that the left-hand side of (1.39) is $\int_0^s p(\alpha) \, d\alpha$, where $p(\alpha)$ is the second derivative of the Fredholm determinant $\det(1 - K_\alpha)$ given in (1.36) (see [61] or [3] for more details).

We thus have

$$
\lim_{N \to \infty} E \Lambda_N(s) = \int_0^s p(\alpha) \, d\alpha. \quad (1.40)
$$

1.5.2. Edge universality: the Airy kernel. Near the spectral edges and under a different scaling another type of universality emerges. It also has a determinantal form, but the kernel is the Airy kernel,

$$
A(x, y) := \frac{\text{Ai}(x) \text{Ai}'(y) - \text{Ai}'(x) \text{Ai}(y)}{x-y},
$$

where $\text{Ai}(x)$ is the Airy function, that is,

$$
\text{Ai}(x) = \frac{1}{\pi} \int_0^\infty \cos \left( \frac{1}{3} t^3 + xt \right) dt.
$$

which is the solution to the second-order differential equation, $y'' - xy = 0$, with vanishing boundary condition at $x = \infty$. The result, analogous to (1.35), is that at the upper spectral edge $E = 2$ of the Hermitian Wigner matrices the following weak limit holds as $N \to \infty$:

$$
p_N^{(k)} \left( 2 + \frac{\alpha_1}{N^{2/3}}, 2 + \frac{\alpha_2}{N^{2/3}}, \ldots, 2 + \frac{\alpha_k}{N^{2/3}} \right) \to \det(A(\alpha_i, \alpha_j))_{i,j=1}^k. \quad (1.41)
$$

A similar statement holds at the lower spectral edge $E = -2$. For Wigner matrices this was first proved by Soshnikov [62] following the paper [63] of Sinai and Soshnikov. Recently, different proofs were given by Tao and Vu [64] and by Erdős, Yau, and Yin [7] (see §1.6.6). Note the different magnification factor $N^{2/3}$, which expresses the fact that near the edge the typical eigenvalue spacing is $N^{-2/3}$. Intuitively, this spacing is consistent with the semicircle law, since

$$
\# \{ \lambda_j \geq 2 - \varepsilon \} \approx \frac{N}{2\pi} \int_{2-\varepsilon}^2 \sqrt{4-x^2} \, dx = \frac{2}{3\pi} \varepsilon^{3/2} N,
$$

so we expect finitely many eigenvalues at a distance $\varepsilon \sim N^{-2/3}$ away from the edge. Note however, that this argument is not rigorous, since the semicircle law (1.20) requires the test interval $[a, b]$ to be fixed and independent of $N$. Recently we proved a strong form of the local semicircle law in [7] (see Theorem 2.19 below), which rigorously justifies this argument.
The largest eigenvalue $\lambda_N$ may extend over 2, but no more than by $O(N^{-2/3})$. More precisely, the distribution function of the largest eigenvalue is given by another universal function, the Tracy–Widom distribution [65]:

$$\lim_{N \to \infty} P\left(\lambda_N \leq 2 + \frac{s}{N^{2/3}}\right) = F_{2,1}(s) := \exp\left(-\int_s^\infty (x-s)q^2(x)\,dx\right),$$

where $q(s)$ is the solution to the Painlevé II differential equation $q''(s) = sq(s) + 2q^3(s)$ with asymptotics $q(s) \sim \text{Ai}(s)$ at $s = +\infty$ as a boundary condition. One can prove that

$$F_{2,1}(s) \sim 1 - \frac{1}{16\pi s^{3/2}} e^{-4s^{3/2}/3}$$

as $s \to \infty$, that is, eigenvalues beyond the $O(N^{-2/3})$ scale are superexponentially damped. A similar formula holds for symmetric matrices as well [66]. We note that, in particular, this result precisely identifies the limiting distribution of the norm of a large Wigner matrix.

The edge universality is commonly approached via the moment method presented in §1.4. The error term in (1.22) deteriorates as $k$ increases, but with a more careful classification and evaluation of the possible pairing structure it is possible to determine the moments up to order $k = O(N^{2/3})$ (see [62]). We just mention the simpler result

$$\frac{1}{N} \mathbb{E} \text{Tr} H^{2k} = \frac{2^k}{\sqrt{\pi k^3}} \left(1 + o(1)\right) \quad (1.42)$$

for $k = o(N^{2/3})$. Such precision is sufficient to determine the upper spectral edge of $H$ with a precision almost $N^{-2/3}$, since

$$P(\lambda_N \geq 2 + \varepsilon) \leq \frac{\mathbb{E} \text{Tr} H^{2k}}{(2 + \varepsilon)^{2k}} \leq \frac{CN}{k^{3/2}(1 + \varepsilon/2)^{2k}} = o(1)$$

if $\varepsilon \geq (\log N)/k \gg N^{-2/3} \log N$. The computation (1.42) can be refined to include powers of order $k \sim N^{2/3}$ and to precisely identify the joint distribution of the largest eigenvalues [62]. We remark that the original work of Soshnikov assumed that the single-entry distribution is symmetric and all its moments are finite; this condition was subsequently relaxed [67]–[69].

The moment method does not seem to be applicable beyond Soshnikov’s scale, that is, for $k$ much larger than $N^{2/3}$. On the other hand, bulk universality would require computing moments up to $k \sim O(N)$, since $1/k$ is essentially the resolution scale for which knowing the moments of order $k$ precisely still gives some information. The proof of the bulk universality requires completely new methods.

We mention a useful rule of thumb. There is a strong relation among controlling $e^{-itH}$, $H^k$, and $(H - z)^{-1}$. Modulo some technicalities and logarithmic factors, the following three statements are roughly equivalent for any $0 < \varepsilon \ll 1$:

$$e^{-itH} \text{ can be controlled up to times } |t| \leq \varepsilon^{-1},$$

$$H^k \text{ can be controlled up to powers } k \leq \varepsilon^{-1},$$

$$(H - z)^{-1} \text{ can be controlled down to } \text{Im } z = \eta \geq \varepsilon.$$
These relations follow from the standard identities
\[ \frac{1}{H - z} = i \int_0^\infty e^{-it(H - z)} \, dt, \quad z = E + i\eta, \quad \eta > 0, \]
\[ e^{-itH} = \sum_{k=0}^\infty \frac{(-itH)^k}{k!} = \frac{1}{2\pi i} \int_\gamma \frac{e^{-itz}}{H - z} \, dz, \]
(where the contour \(\gamma\) encircles the spectrum of \(H\)).

1.5.3. Invariant ensembles. For ensembles that remain invariant under the transformation \(H \rightarrow U^*HU\) for any unitary matrix \(U\) (or, in the case of symmetric matrices \(H\), for any orthogonal matrix \(U\)), the joint probability density function of all the \(N\) eigenvalues can be explicitly computed. These ensembles are typically given by the probability density (1.30). The eigenvalues are strongly correlated and they are distributed according to a Gibbs measure with a long-range logarithmic interaction potential (this connection was first exploited in [70]). The joint probability density of the eigenvalues of \(H\) can be computed explicitly:

\[ p_N(\lambda_1, \ldots, \lambda_N) = \text{const} \prod_{i<j} (\lambda_i - \lambda_j)^\beta \prod_{j=1}^N \exp \left( -N \sum_{j=1}^N V(\lambda_j) \right), \quad (1.43) \]

where \(\beta = 1\) for symmetric and \(\beta = 2\) for Hermitian ensembles. In particular, for the Gaussian case \(V\) is quadratic, and thus the joint distribution of the GOE (\(\beta = 1\)) and GUE (\(\beta = 2\)) eigenvalues is given by

\[ p_N(\lambda_1, \ldots, \lambda_N) = \text{const} \prod_{i<j} (\lambda_i - \lambda_j)^\beta \prod_{j=1}^N \exp \left( -\frac{1}{4} \frac{\beta N}{N} \sum_{j=1}^N \lambda_j^2 \right). \quad (1.44) \]

It is often useful to think of this measure as a Gibbs measure of the form

\[ \mu_N(d\lambda) = p_N(\lambda) \, d\lambda = \frac{e^{-N\mathcal{H}(\lambda)}}{Z}, \quad \mathcal{H}(\lambda) := \sum_{i=1}^N V(\lambda_i) - \frac{\beta}{N} \sum_{i<j} \log |\lambda_j - \lambda_i| \quad (1.45) \]

with the confining potential \(V(\lambda) = \beta\lambda^2/4\). The proof of (1.43) is a direct (but involved) calculation, based on a suitable change of variable. We sketch it for the Hermitian (unitary invariant) case. The key observation is that invariance of the measure under conjugation implies that the eigenvalues, organized in the diagonal matrix \(D = \text{diag}(\lambda_1, \ldots, \lambda_N)\), are independent of the eigenvectors, organized in a unitary matrix \(U\). Writing \(H = UDU^*\), one gets that \(\mathcal{P}(H) \, dH\) factorizes as

\[ \mathcal{P}(H) \, dH = e^{-N\text{Tr}V(H)} \, dH = [p_N(\lambda_1, \ldots, \lambda_N) \, d\lambda_1 \cdots d\lambda_N] \, dU, \]

where \(dU\) denotes the uniform (Haar) measure on the unitary group \(U(N)\) and \(p_N\) is the induced density function on the diagonal matrices (or their entries). Thus, computing the function \(p_N\) amounts to computing the Jacobian of the change of variables from the matrix elements of \(H\) to the parametrization coordinates in terms of eigenvalues and eigenvectors. The result is

\[ dH = \text{const} \, |\Delta_N(\lambda)|^\beta \, d\lambda \, dU, \quad \Delta_N(\lambda) := \prod_{i<j} (\lambda_i - \lambda_j), \quad (1.46) \]
where $\beta = 1$ is the symmetric case and $\beta = 2$ is the Hermitian case (see [3] or §§ 3.1–3.3 of [1] for details).

Especially remarkable is the emerging Vandermonde determinant in (1.43), which comes directly from integrating out the Haar measure. Note that the symmetry type of the ensemble appears through the exponent $\beta$. Only the $\beta = 1, 2, 4$ cases correspond to matrix ensembles of the form (1.30), namely, to the symmetric, Hermitian, and quaternion self-dual matrices. We will not give the precise definition of the latter (see, for example, Chap. 7 of [1] or [6]), but only mention that this is the natural generalization of symmetric or Hermitian matrices to matrices with quaternion entries and that they have real eigenvalues.

Irrespective of any underlying matrix ensemble, one can nevertheless study the distribution (1.43) for any $\beta > 0$; these are called general $\beta$-ensembles. In particular, for the Gaussian case $V(\lambda) = \lambda^2/2$ there are corresponding tridiagonal matrix ensembles for any $\beta > 0$, obtained from successive Householder transformations, whose eigenvalue distribution is described by (1.43) (see [71] for an overview). Using the tridiagonal structure, we can apply methods from the theory of Jacobi matrices. For example, universality at the edge eigenvalues is understood in the sense that they are shown to converge to the lowest eigenvalues for a one-dimensional Schrödinger operator with a white noise drift. In particular, the $\beta$-analogue of the Tracy–Widom distribution was determined in [72] and [73], following the conjectures in [74]. A different method, the Brownian carousel representation [75], has been used to generalize the tail distribution of large eigenvalue gaps (‘Wigner surprise’) for Gaussian $\beta$-ensembles [76]. More precisely, it has been shown that under the distribution (1.43) with $V(\lambda) = \lambda^2/2$ the probability that there is no point falling in a fixed interval of length $s$ (after locally rescaling the spectrum so that the typical distance is $2\pi$) is given by

$$q_s = (\kappa_\beta + o(1)) s^{\gamma_\beta} \exp \left( -\frac{\beta}{64} s^2 + \left( \frac{\beta}{8} - \frac{1}{4} \right) s \right), \quad \gamma_\beta := \frac{1}{4} \left( \frac{\beta}{2} + \frac{2}{\beta} - 3 \right), \quad \kappa_\beta > 0,$$

as $s \to \infty$ (after the limit as $N \to \infty$).

Bulk universality, that is, the analogue of the sine-kernel behaviour (1.35) for general $\beta$-ensembles, is unproven, even for the Gaussian case. The main difficulty is that (1.43) represents an $N$-particle system with a long-range interaction. We can write the joint density as a Gibbs measure (1.45), and then we have $N$ particles in a confining potential $V$ that repel each other with a potential having locally logarithmic repulsion, but also a large (in fact increasing) long-range component. Standard methods from statistical physics to construct and analyze Gibbs measures do not seem to apply. Although here we do not attempt to construct an analogue of an infinite-volume Gibbs measure but only want to compute the correlation functions, even this is a daunting task by standard methods unless some extra structure is found.

1.5.4. Universality of classical invariant ensembles via orthogonal polynomials. Much more is known about the classical invariant ensembles, that is, the $\beta = 1, 2, 4$ cases, with a general potential $V$. For these specific values an extra mathematical structure emerges, namely the orthogonal polynomials with respect to the weight function $e^{-NV(\lambda)}$ on the real line. This approach was originally applied by Mehta and
Gaudin [1], [77] to compute the gap distribution for the Gaussian case using classical Hermite orthonormal polynomials. Dyson [10] computed the local correlation functions for a related ensemble (circular ensemble), and then this was extended to the standard Gaussian ensembles by Mehta [11]. A general method using orthogonal polynomials was later developed to tackle a very general class of unitary ensembles (see, for example, [1], [2], [78]–[82] and references therein).

For simplicity, to illustrate this connection, we will consider the Hermitian case $\beta = 2$ with a Gaussian potential $V(\lambda) = \lambda^2/2$ (which, by Lemma 1.2, is also a Wigner matrix ensemble, namely, the GUE). To simplify the presentation further, for the purpose of this argument only, we rescale the eigenvalues $\lambda \rightarrow \sqrt{N} \lambda$, which effectively removes the factor $N$ from the exponent in (1.43). (This pure scaling works only in the Gaussian case, but it is only a technical convenience to simplify formulae.)

Let $P_k(x)$ be the $k$th orthogonal polynomial with respect to the weight function $e^{-x^2/2}$, with leading coefficient 1. Let

$$\psi_k(x) := \frac{e^{-x^2/4}P_k(x)}{\|e^{-x^2/4}P_k\|}$$

be the corresponding orthonormal functions, that is,

$$\int \psi_k(x)\psi_\ell(x)\,dx = \delta_{k,\ell}. \quad (1.47)$$

In the particular case of the Gaussian weight function, $P_k$ is given by the Hermite polynomials

$$P_k(x) = H_k(x) := (-1)^k e^{x^2/2} \frac{d^k}{dx^k} e^{-x^2/2},$$

and

$$\psi_k(x) = \frac{P_k(x)}{(2\pi)^{1/4}(k!)^{1/2}} e^{-x^2/4},$$

but for the following discussion we will not need these explicit formulae.

The key observation is that by simple properties of the Vandermonde determinant we have

$$\Delta_N(x) = \prod_{1 \leq i < j \leq N} (x_j - x_i) = \det(x_i^{j-1})_{i,j=1}^{N} = \det(P_{j-1}(x_i))_{i,j=1}^{N}, \quad (1.48)$$

where we have used the fact that $P_j(x) = x^j + \cdots$ is a polynomial of degree $j$ with leading coefficient equal to 1. Let us define the kernel

$$K_N(x, y) := \sum_{k=0}^{N-1} \psi_k(x)\psi_k(y),$$

that is, the projection kernel onto the subspace spanned by the first $N$ orthonormal functions. Then (1.48) immediately implies that

$$p_N(x_1, \ldots, x_N) = C_N \left[\det(P_{j-1}(x_i))_{i,j=1}^{N}\right]^2 \prod_{i=1}^{N} e^{-x_i^2/2}$$

$$= C'_N \left[\det(\psi_{j-1}(x_i))_{i,j=1}^{N}\right]^2 = C'_N \det(K_N(x_i, x_j))_{i,j=1}^{N},$$
where in the last step we used the fact that the square of the matrix \((\psi_{j-1}(x_i))^{N}_{i,j=1}\) is exactly \((K_N(x_i, x_j))^{N}_{i,j=1}\), and for simplicity we did not write out the exact constants.

To compute the correlation functions, we expand the determinant:

\[
p^{(k)}_N(x_1, \ldots, x_k) = C_{k,N} \int_{\mathbb{R}^{N-k}} \det(K_N(x_i, x_j))^{N}_{i,j=1} \prod_{i=k+1}^{N} dx_i
\]

\[
= C_{k,N} \sum_{\sigma, \tau \in S_N} (-1)^{\tau + \sigma} \int_{\mathbb{R}^{N-k}} \prod_{j=1}^{N} \psi_{\sigma(j)-1}(x_j)\psi_{\tau(j)-1}(x_j) \prod_{i=k+1}^{N} dx_i
\]

\[
= C_{k,N} \sum_{\alpha_1 < \cdots < \alpha_k} \sum_{\sigma, \tau \in S_k} (-1)^{\tau + \sigma} \prod_{j=1}^{k} \psi_{\alpha_{\sigma(j)}-1}(x_j)\psi_{\alpha_{\tau(j)}-1}(x_j)
\]

\[
= C_{k,N} \sum_{\alpha_1 < \cdots < \alpha_k} \left[ \det(\psi_{\alpha_{\tau(j)}-1}(x_j))^{k}_{i,j=1} \right]^2,
\]

where \(S_N\) is the permutation group on \(N\) elements and \((-1)^{\tau}\) is the parity of the permutation. In the third line we used (1.47) to compute the integrals, setting \(\sigma(j) = \tau(j)\) for all \(j \geq k + 1\) and denoting by \(\{\alpha_1, \ldots, \alpha_k\}\) the ordering of the set \(\{\sigma(1), \ldots, \sigma(k)\} = \{\tau(1), \ldots, \tau(k)\}\). Finally, using the fact that the matrix \([K_N(x_i, x_j)]^{k}_{i,j=1}\) can be written as \(A^tA\) with \(A_{ij} = \psi_{i-1}(x_j)\) and using the Cauchy–Binet expansion formula for the determinant of a product matrix, we get that

\[
\det[K_N(x_i, x_j)]^{k}_{i,j=1} = \sum_{\alpha_1 < \cdots < \alpha_k} \left[ \det(\psi_{\alpha_{\tau(j)}-1}(x_j))^{k}_{i,j=1} \right]^2.
\]

Apart from the constant, which can be computed, we thus proved that

\[
p^{(k)}_N(x_1, \ldots, x_k) = \frac{(N-k)!}{N!} \det[K_N(x_i, x_j)]^{k}_{i,j=1},
\]

that is, the correlation functions have a determinantal structure.

In order to see the sine kernel (1.33) emerging, we need a basic algebraic property of the orthogonal polynomials, the Christoffel–Darboux formula:

\[
K_N(x, y) = \sum_{j=0}^{N-1} \psi_j(x)\psi_j(y) = \sqrt{N} \left[ \frac{\psi_N(x)\psi_{N-1}(y) - \psi_N(y)\psi_{N-1}(x)}{x - y} \right].
\]

Furthermore, orthogonal polynomials of high degree have the following asymptotic behaviour as \(N \to \infty\):

\[
\psi_{2m}(x) \approx \frac{(-1)^m}{N^{1/4} \sqrt{\pi}} \cos(\sqrt{N} x) + o(N^{-1/4}),
\]

\[
\psi_{2m+1}(x) \approx \frac{(-1)^m}{N^{1/4} \sqrt{\pi}} \sin(\sqrt{N} x) + o(N^{-1/4})
\]

for any \(m\) such that \(|2m - N| \leq C\). The approximation is uniform for \(|x| \leq CN^{-1/2}\). These formulae will be useful if we set \(E = 0\) in (1.35), since we rescaled the
eigenvalues by a factor of $\sqrt{N}$, so the relation between the notation of (1.35) (for $k = 2$) and $x, y$ is
\[ x = \sqrt{N} \left( E + \frac{\alpha_1}{N \varrho(E)} \right), \quad y = \sqrt{N} \left( E + \frac{\alpha_2}{N \varrho(E)} \right). \] (1.50)

For different values of $E$ one needs somewhat different asymptotic formulae for the orthogonal polynomials. We can thus compute that
\[ K_N(x, y) \approx \frac{1}{\pi} \frac{\sin(\sqrt{N} x) \cos(\sqrt{N} y) - \sin(\sqrt{N} y) \cos(\sqrt{N} x)}{x - y} = \frac{\sin \sqrt{N} (x - y)}{\pi (x - y)}. \] (1.51)

Using (1.50) and the fact that $\varrho(0) = \pi^{-1}$, we have
\[ \frac{1}{\varrho(0) \sqrt{N}} K_N(x, y) \approx \frac{\sin \pi (\alpha_1 - \alpha_2)}{\pi (\alpha_1 - \alpha_2)}, \]
which gives (1.35) for $E = 0$ after undoing the $\lambda \to \sqrt{N} \lambda$ magnification.

The main technical tool is the refined asymptotic formulae (1.49) for orthogonal polynomials. In the case of the classical orthogonal polynomials (appearing in the standard Gaussian Wigner and Wishart ensembles) they are usually obtained by Laplace asymptotics from their integral representation. For a general potential $V$ the corresponding analysis is quite involved and depends on the regularity properties of $V$. One successful approach was initiated by Fokas, Its, and Kitaev [81] and by Deift and collaborators via the Riemann–Hilbert method (see [2] and references therein). An alternative method was presented in [83], [84] using more direct methods with orthogonal polynomials.

There have been many refinements and improvements of results in this very active area of research related to invariant ensembles, revealing fruitful interconnections between random matrices, orthogonal polynomials, complex analysis, and even combinatorics (see [2]). One common input, however, is the explicit formula (1.43) for the joint probability density, enabling one to bring in orthogonal polynomials. We now depart from this topic and focus on ensembles when such an explicit formula is not available, the most prominent example being the Wigner matrix. Apart from the Gaussian case, no explicit formula is available for the joint eigenvalue distribution. Thus, the basic algebraic connection between eigenvalue ensembles and orthogonal polynomials is lacking, and completely new methods needed to be developed. In the next section we summarize recent results in this direction.

1.6. Local statistics of eigenvalues: New results.

1.6.1. Hermitian matrices with Gaussian convolution. The first rigorous partial result for bulk universality in the non-unitary case was given by Johansson [13] (see also Ben Arous and Péché [85] for extending [13] to the full bulk spectrum, and the recent improvement [86] on weakening moment conditions). The main result states that bulk universality holds for Gaussian divisible Hermitian ensembles, that is, Hermitian ensembles of the form
\[ H = \sqrt{1 - \varepsilon} \hat{H} + \sqrt{\varepsilon} V, \] (1.52)
where $\hat{H}$ is a Hermitian Wigner matrix, $V$ is an independent standard GUE matrix, and $\varepsilon$ is a positive constant of order 1, independent of $N$.

We will often use the parametrization
\begin{equation}
H = e^{-t/2} \hat{H} + (1 - e^{-t})^{1/2} V.
\end{equation}

If embedded in a flow, then the variable $t$ can be interpreted as the time in an *Ornstein–Uhlenbeck (OU)* process. This formalism incorporates the idea that matrices with Gaussian convolutions can be obtained as a matrix-valued stochastic process, namely, as the solution of the stochastic differential equation
\begin{equation}
dH_t = \frac{1}{\sqrt{N}} d\beta_t - \frac{1}{2} H_t \, dt, \quad H_0 = \hat{H},
\end{equation}
where $\beta_t$ is a Hermitian matrix-valued process whose diagonal matrix elements are standard real Brownian motions and whose off-diagonal matrix elements are standard complex Brownian motions. The probability distribution for the solution to (1.54) for any fixed $t$ coincides with the distribution for (1.53). Note that infinite time, $t = \infty$, corresponds to the GUE ensemble, so the matrices (1.53) interpolate between the Wigner matrix $\hat{H}$ and the GUE. This point of view will be extremely useful below, since it allows us to compare Wigner matrices with Gaussian ones if the effect of time evolution is under control.

Alternatively, one can consider the density function $u_t$ of the real and imaginary parts of the matrix elements as evolving according to the *generator* of the OU process:
\begin{equation}
\partial_t u_t = Au_t, \quad A := \frac{1}{4} \frac{\partial^2}{\partial x^2} - \frac{x}{2} \frac{\partial}{\partial x},
\end{equation}
where the initial condition $u_0$ is the density (with respect to the reversible Gaussian measure) of the distribution of the real and imaginary parts of the matrix elements of $\sqrt{N} \, \hat{H}$. For the diagonal elements an OU process with a slightly different normalization is used. The OU process (1.55) keeps the expectation zero and the variance $1/2$ if the initial condition $u_0$ has these properties.

The joint distribution of the eigenvalues of Gaussian divisible Hermitian random matrices of the form (1.53) still has a certain determinantal structure. The formula is somewhat simpler if we write
\begin{equation}
H = \sqrt{1 - \varepsilon} (\hat{H} + a V),
\end{equation}
with $a = \sqrt{\varepsilon/(1 - \varepsilon)}$, that is, we use the standard Gaussian convolution
\begin{equation}
\tilde{H} = \hat{H} + a V
\end{equation}
and then rescale at the end. We note that (1.56) can be generated by evolving the matrix elements by standard Brownian motions $\beta$ up to time $t = a^2$, that is, by solving the equation
\begin{equation}
d\tilde{H}_t = \frac{1}{\sqrt{N}} d\beta_t, \quad \tilde{H}_0 = \hat{H}.
\end{equation}
Moreover, to be in line with the normalization convention of [14], which follows [13], we assume that the matrix elements of the Wigner matrix $\hat{H}$ and the GUE
matrix $V$ have variance $1/(4N)$ instead of $1/N$ as in Definition 1.1. This means that the eigenvalues are scaled by a factor $1/2$, compared with the convention in the previous subsections, and the semicircle law (1.20) is modified to $2\pi^{-1}\sqrt{(1-x^2)}_+$. This convention applies only up to the end of this section.

Let $y = (y_1, \ldots, y_N)$ denote the eigenvalues of $\bar{H}$ and let $x = (x_1, \ldots, x_N)$ denote the eigenvalues of $H$. Then we have the following representation formulae (a slight variant of these formulae was given and used by Johansson in [13], motivated by similar formulae of Brézin and Hikami in [12]).

**Lemma 1.4** ([14], Proposition 3.2). Let $V$ be a GUE matrix. For any fixed Hermitian matrix $\bar{H}$ with eigenvalues $y$, the density function of the eigenvalues $x$ of $H = \bar{H} + aV$ is given by

$$q_S(x; y) := \frac{1}{(2\pi S)^{N/2}} \frac{\Delta_N(x)}{\Delta_N(y)} \det(e^{-(x_j-y_k)^2/(2S)})_{j,k=1}^N,$$

(1.58)

where $S = a^2/N$ and $\Delta_N$ denotes the Vandermonde determinant (1.48). The $m$-point correlation functions of the eigenvalues of $H = \bar{H} + aV$,

$$p_{N,y}^{(m)}(x_1, \ldots, x_m) := \int_{\mathbb{R}^{N-m}} q_S(x_1, \ldots, x_N; y) \, dx_{m+1} \cdots dx_N,$$

are given by the formula

$$p_{N,y}^{(m)}(x_1, \ldots, x_m) = \frac{(N-m)!}{N!} \det(\mathcal{X}_N^S(x_i, x_j; y))_{i,j=1}^m, \quad S = \frac{a^2}{N}. \quad (1.59)$$

Here

$$\mathcal{X}_N^S(u, v; y) := \frac{1}{(2\pi i)^2(v-u)S} \int_{\gamma} \left( w - r + z - u \right) e^{-(v-u)(w-r)/S} \prod_{j=1}^N \frac{w-y_j}{z-y_j} \times \frac{1}{w-r} \left( w - r + z - u - S \sum_j \frac{y_j-r}{(w-y_j)(z-y_j)} \right) e^{(w^2-2uw+z^2+2uz)/(2S)},$$

(1.60)

where $r \in \mathbb{R}$ is an arbitrary constant. The integration curves $\gamma$ and $\Gamma$ in the complex plane are as follows: $\gamma$ is the union $\gamma_+ \cup \gamma_-$ of two lines $\gamma_+ : \tau \to -\tau + i\omega$ and $\gamma_- : \tau \to \tau - i\omega$ $(\tau \in \mathbb{R})$ for any fixed $\omega > 0$, and $\Gamma$ is $\tau \to i\tau$ $(\tau \in \mathbb{R})$.

We note that $\Gamma$ can be shifted to any vertical line, since the integrand is an entire function of $w$ and has a Gaussian decay as $|\text{Im } w| \to \infty$. The constants $r \in \mathbb{R}$ and $\omega > 0$ (appearing in the definition of the contour $\gamma$) can be arbitrary and can be appropriately chosen in the contour integral estimates.

The key step behind the proof of (1.58) is the Harish-Chandra–Itzykson–Zuber integral [87]:

$$\int_{U(N)} e^{Tr(UAU^*B)} \, dU = \text{const} \frac{\det(e^{a_i b_j})_{i,j=1}^N}{\Delta_N(a)\Delta_N(b)},$$

where $A$ and $B$ are two Hermitian matrices with eigenvalues $a = (a_1, \ldots, a_N)$ and $b = (b_1, \ldots, b_N)$, and the integration is over the unitary group $U(N)$ with
respect to the Haar measure. We note that this is the step where the unitary invariance (or the Hermitian character of $H$) is crucially used; an analogous simple formula is not known for other symmetry groups (see [88]).

For any test function $f$, and for a fixed matrix $\hat{H}$ we have

$$\int f(x)q_S(x; y)\,dx = \text{const} \int f(x)e^{-N \text{Tr}(\hat{H}-\hat{H})^2} \,d\hat{H},$$

where we used the fact that $V = a^{-1}(\hat{H} - \hat{H})$ is a GUE matrix with distribution

$$\mathcal{P}(V)\,dV = e^{-N\text{Tr}V^2/2} \,dV.$$ 

Letting $\hat{H} = UXU^*$ be the diagonalization of $\hat{H}$ with $X = \text{diag}(x)$ and using (1.46), we have

$$\int f(x)q_S(x; y)\,dx = \text{const} \int_{\mathbb{R}^N} f(x) \exp\left(-\frac{1}{2a^2} N \text{Tr}(UXU^* - \hat{H})^2\right) \,dU \Delta_N^2(x) \,dx$$

$$= \text{const} \int_{\mathbb{R}^N} f(x) \exp\left(-\frac{N}{a^2} \text{Tr}UXU^* \hat{H}\right) \,dU$$

$$\times f(x) \exp\left(-\frac{N}{2S} \sum_i (x_i^2 + y_i^2)\right) \Delta_N^2(x) \,dx$$

$$= \text{const} \int_{\mathbb{R}^N} f(x) \frac{\det(e^{x_iy_j/S})_{i,j=1}^N}{\Delta_N(x)\Delta_N(y)} \exp\left(-\frac{N}{2S} \sum_i (x_i^2 + y_i^2)\right) \Delta_N^2(x) \,dx$$

$$= \text{const} \int_{\mathbb{R}^N} f(x) \frac{\Delta_N(x)}{\Delta_N(y)} \det\left(\exp\left(-\frac{1}{2S}(x_i - y_j)^2\right)\right)_{i,j=1}^N \,dx,$$

which proves (1.58) (apart from the constant). This shows how the Vandermonde determinant structure emerges for Gaussian convolutions. The proof of the contour integral representation (1.60) from (1.58) is a bit more involved; see Proposition 3.2 in [14] (or Proposition 2.3 in [13]) for details.

Once (1.60) is given, the key idea is to view it as a complex integral suited for Laplace asymptotics or saddle point calculations. More precisely, after some straightforward algebraic steps, it can be put in the following form (see §3.1 of [14]), where we already made a change of variables in the argument to detect the microscopic structure. For any fixed $|u| < 1$ and $t = a^2$ we find from (1.60) that

$$\frac{1}{N\varrho(u)} \mathcal{K}_N^{T/N}(u, u + \frac{\tau}{N\varrho(u)}; y) = N\int_\gamma \frac{dz}{2\pi i} \int_\Gamma \frac{dw}{2\pi i} h_N(w)g_N(z, w)e^{N(f_N(w) - f_N(z))},$$

(1.61)
with
\[ f_N(z) := \frac{1}{2t} (z^2 - 2uz) + \frac{1}{N} \sum_j \log(z - y_j), \]
\[ g_N(z, w) := \frac{1}{t(w - r)} [w - r + z - u] - \frac{1}{N(w - r)} \sum_j \frac{y_j - r}{(w - y_j)(z - y_j)}, \]
\[ h_N(w) := \frac{1}{\tau} e^{-\tau(w - r)/(t\varphi(u))} - 1. \]

Note the factor \( N \) in front of the exponent in (1.61), indicating that the main contribution to the integral comes from the saddle points, that is, from \( z \) and \( w \) values with \( f'_N(z) = f'_N(w) = 0 \). We remark that
\[ f'_N(z) = \frac{z - u}{t} + \frac{1}{N} \sum_j \frac{1}{z - y_j}, \]  
(1.62)
that is, the derivative is actually given by the empirical Stieltjes transform (1.23) of the eigenvalues of \( \hat{H} \). Suppose that the Wigner semicircle law holds, or equivalently, (1.28). Then the saddle point \( z_N \), where \( f'_N(z_N) = 0 \), can be well approximated by the solution to the equation
\[ \frac{z - u}{t} + 2(z - \sqrt{z^2 - 1}) = 0 \]  
(1.63)
(the formula for the Stieltjes transform differs slightly from (1.28) because of the different normalization of \( \hat{H} \)). It is easy to check that there are two solutions \( z^\pm \), with imaginary parts given by \( \pm 2ti\sqrt{1 - u^2} + O(t^2) \) for small \( t \).

Once the saddle points are identified, the integration contours \( \gamma \) and \( \Gamma \) in (1.61) can be shifted to pass through the saddle points from a direction where \( f_N \) is real and its second derivative has a ‘good sign’, so that the usual saddle point approximation holds. There are altogether four pairs of saddles \( (z_N^\pm, w_N^\pm) \), but only two give contributions to leading order. Their explicit evaluation gives the sine kernel (1.33) for \( \mathcal{X}_N \) in the limit as \( N \to \infty \).

The key technical requirement is to justify the semicircle approximation used in passing from (1.62) to (1.63) near the saddle point (some estimates are needed away from the saddle as well, but those are typically easier to obtain). The standard argument for the semicircle law presented in §1.4.2 holds for any fixed \( z \) with \( \text{Im} \ z > 0 \), independently of \( N \); especially important is that \( \text{Im} \ z > 0 \) uniformly with respect to \( N \). Therefore, this argument can be used to justify the passage to (1.63) only for a fixed \( t > 0 \). Recall that \( t = a^2 \) is the variance of the Gaussian convolution. This was the path essentially followed by Johansson, who proved [13] that sine-kernel universality (1.35) holds if the Wigner matrix has a Gaussian component comparable to its total size.

One implication of the local semicircle law explained in §1.1.1 is that the approximation in passing from (1.62) to (1.63) could be justified even for very short times \( t \). Essentially, any time of order \( \gg 1/N \) (with some logarithmic factor) is allowed, but for technical reasons we carried out the estimates only for \( t = N^{-1+\varepsilon} \) for any \( \varepsilon > 0 \), and we showed that the contour integral (1.61) is given by the sine kernel even for
such short times \cite{erdos2009random}. This proved the sine-kernel universality in the form of (1.35) for any fixed $E$ in the bulk of the spectrum for Hermitian matrices with a Gaussian component of variance $N^{-1+\varepsilon}$.

1.6.2. Gaussian convolutions for arbitrary matrices: the local relaxation flow. The method presented in §1.6.1 relies heavily on the Brézin–Hikami type formula (1.60), which is available only for Hermitian matrices. For ensembles with other symmetries, in particular for symmetric Wigner matrices, a new method was necessary. In a series of papers \cite{erdos2009random}, \cite{erdos2010local}, \cite{erdos2011local} we developed an approach based on hydrodynamical ideas from interacting-particle systems. We will present it in more detail in §3, but here we summarize the key points.

The starting point is a key observation of Dyson \cite{dyson1962brownian} from 1962. Let $\tilde{H}$ be an arbitrary fixed matrix and consider the solution $\tilde{H}_t$ to (1.57). Recall that for each fixed $t$, $\tilde{H}_t$ has the same distribution as $\tilde{H} + \sqrt{t} V$, where $V$ is a standard GUE matrix (independent of $\tilde{H}$). Dyson noted that the evolution of the eigenvalues of the flow $\tilde{H}_t$ is given by a coupled system of stochastic differential equations, commonly called the Dyson Brownian motion (DBM in short). For convenience, we will replace the Brownian motions by OU processes to keep the variance constant, that is, we will use (1.54) instead of (1.57) to generate the matrix flow. The Dyson Brownian motion we will use is given by the following system of stochastic differential equations for the eigenvalues $\lambda(t) = (\lambda_1(t), \ldots, \lambda_N(t))$ (see, for instance, §4.3.1 of \cite{erdos2010random}):

$$\frac{d\lambda_i}{t} = \frac{dB_i}{\sqrt{N}} + \left[ -\frac{\beta}{4} \lambda_i + \frac{\beta}{2N} \sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} \right] \, dt, \quad 1 \leq i \leq N,$$

(1.64)

where $\{B_i : 1 \leq i \leq N\}$ is a collection of independent Brownian motions. The initial condition $\lambda(0)$ is given by the eigenvalues of $\tilde{H}$. The choice of parameter $\beta = 2$ corresponds to the Hermitian case, but the process (1.64) is well defined for any $\beta \geq 1$ and the eigenvalues do not cross due to the strong repulsion between them. The threshold $\beta = 1$ is critical for the non-crossing property. As $t \to \infty$ the distribution of $\lambda(t)$ converges to the Gaussian $\beta$-ensemble distribution (1.44) as the global invariant measure; for example, for $\beta = 2$, it converges to the GUE.

Using the Dyson Brownian motion, we can translate the question of universality for Gaussian divisible ensembles to the question of the time needed for the DBM to reach equilibrium. The time scale to reach global equilibrium is of order 1, but as we eventually proved in \cite{erdos2009random}, the decay to local equilibrium is much faster: it occurs already in a time scale of order $t \sim N^{-1}$. Since the local statistics of the eigenvalues depend exclusively on the local equilibrium, this means that the local statistics of Gaussian divisible ensembles with a tiny Gaussian component of size $N^{-1+\varepsilon}$ are already universal.

We remark that by using the relation between Gaussian divisible ensembles and the relaxation time of the DBM, we can interpret the result of Johansson \cite{johansson2004local} as stating that the local statistics of the GUE are reached via the DBM in a time of order at most 1. Our result from \cite{erdos2009random}, explained in the previous subsection, indicates that the decay to local equilibrium occurs already in time $t \sim N^{-1+\varepsilon}$. This is, however, only a reinterpretation of the results, since neither \cite{johansson2004local} nor \cite{erdos2009random}
used hydrodynamical ideas. In particular, these proofs are valid only for Hermitian matrices, because they use some version of the Brézin–Hikami formula.

To establish universality in full generality, we developed a purely hydrodynamical approach based on the relaxation speed of the DBM. The key point in this approach is that there is no restriction on the symmetry type; the argument works equally well for symmetric, Hermitian, or quaternion self-dual ensembles, and moreover, with some obvious modifications it also works for random covariance matrices.

Our first paper using hydrodynamical ideas is [89]. In this paper we extended Johansson’s result [13] to Hermitian ensembles with a Gaussian component of size \( t \gg N^{-3/4} \), by capitalizing on the fact that the local statistics of eigenvalues depend exclusively on the approach to local equilibrium, which in general is faster than reaching global equilibrium. Unfortunately, the description of local equilibria in [89] still used explicit representations of correlation functions by orthogonal polynomials (following, for instance, [82]), and the extension to other ensembles in this way is not a simple task (see [90] for extension of [82] to symmetric matrices to prove edge universality).

To depart from using orthogonal polynomials, we introduced new hydrodynamical tools in [16] which entirely eliminated explicit formulae and gave a unified proof for the universality of symmetric and Hermitian Wigner matrices with a small Gaussian convolution. The size of the Gaussian component, or equivalently, the time needed to reach the local equilibrium in a sufficiently strong sense, increased from \( N^{-3/4} \) to \( N^{-\xi} \) with a small positive \( \xi \), but the method became more general. The result was further generalized in [6] to quaternion self-dual Wigner matrices and sample covariance matrices, and even to generalized Wigner matrices in [9], [8] (see Definition 1.1). Finally, in [7] we showed that the local equilibrium is already reached in a time \( t \gg N^{-1+\varepsilon} \), which is essentially optimal. More importantly, the hydrodynamical method not only applies to all these specific ensembles, but also gives a conceptual interpretation of the fact that the occurrence of universality is due to the relaxation to local equilibrium of the DBM.

Our hydrodynamical approach consists of two parts. First, we have a general theorem stating that under certain structural and convexity conditions on the Hamiltonian \( \mathcal{H} \) of the equilibrium measure of the DBM (see (1.45) for a special case) and for fairly strong control on the local density of the eigenvalues, the local equilibrium is reached within a short time \( t \sim N^{-\xi} \), \( \xi > 0 \), in the sense that the local correlation functions, rescaled in the form of (1.35), coincide with the same correlation functions in equilibrium. By the general Bakry–Emery [91] criterion, the speed of convergence to global equilibrium for the DBM depends on the lower bound on the Hessian of the Hamiltonian \( \mathcal{H} \), which in our case is of order 1. The key idea is to speed up this convergence by modifying the Hamiltonian. We add to \( \mathcal{H} \) an auxiliary potential of the form

\[
W(\lambda) = \frac{1}{2R^2} \sum_j (\lambda_j - \gamma_j)^2,
\]

where \( R \ll 1 \) is a parameter depending on \( N \) and the \( \gamma_j \) are the classical locations of the eigenvalues, given by

\[
N \int_{-\infty}^{\gamma_j} \varrho(x) \, dx = j.
\] (1.65)
Here $\rho(x)$ is the limiting density, for example, $\rho(x) = \rho_{sc}(x)$ for Wigner matrices. The Hamiltonian $\tilde{H} := H + W$ generates a new stochastic flow of the eigenvalues, called the local relaxation flow. The equilibrium Gibbs measure given by $\tilde{H}$ will be called the pseudo-equilibrium measure. The convergence to equilibrium for this flow is faster: it occurs in a time scale of order $R^2$. In fact, due to the strong repulsion between eigenvalues (reflected by a singular repulsion potential (1.64)), the convergence is even faster for observables that depend only on eigenvalue differences. Then we show that the modified dynamics is, in fact, not far from the original one, using the fact that the typical size of the auxiliary potential $W$ is small. More precisely, we will need to prove that the eigenvalues $\lambda_j$ lie near $\gamma_j$ with precision $N^{-1/2-\varepsilon}$, that is, that

$$
E \frac{1}{N} \sum_{i=1}^{N} (\lambda_i - \gamma_i)^2 = N^{-1-2\varepsilon}
$$

holds with some $\varepsilon > 0$. This is the key input condition to our general theorem, and it will be proved by a strong control on the local density. The exponent $\xi$ in the time scale $t \sim N^{-\xi}$ is essentially the number $2\varepsilon$ appearing in the estimate (1.66).

The second part of the hydrodynamical approach is to prove the necessary input conditions on the local density for the general theorem, especially (1.66). This is the step where specific properties of the matrix ensemble come into the game. To obtain relaxation to local equilibrium on the time scale $t \sim N^{-\xi}$, $\xi > 0$, we need to locate the eigenvalues with precision at least $N^{-1/2-\varepsilon}$, $\varepsilon = \xi/2$. To obtain the optimal relaxation time $t \gg N^{-1}$, the eigenvalues need to be located essentially with precision $N^{-1}$, as in [14]. Very crudely, the precision of location of the eigenvalues corresponds to the scale $\eta$ on which the local semicircle law holds, so this will be the key input for verifying (1.66). The technical difficulty is that we need fairly good control on the local density near the spectral edges as well, since (1.66) involves all the eigenvalues. Although we are interested only in bulk universality, that is, the local behaviour away from the edges, we still need the global speed of convergence for the modified dynamics, which is influenced also by the eigenvalues at the edges. We recall that the control on the density near the edges becomes weaker, since the eigenvalues near the edges tend to fluctuate more.

A good control on the local density was developed in our previous work on Wigner matrices [55]–[57], but the edge behaviour was not optimal. Nevertheless, in [6] we succeeded in proving (1.66) in a somewhat complicated way, relying on some improvements of estimates in [55]–[57]. In [9] we found a more direct way to control the local density and prove (1.66) more efficiently, and a more streamlined version is given in [8] and will be sketched in §2. The strongest result [7], to be explained in §2.4, gives (1.66) with essentially $2\varepsilon = 1$.

We mention that these proofs also apply to generalized Wigner matrices where the variances satisfy (1.17). In this case we prove the local semicircle law down to essentially the smallest possible energy scale $N^{-1}$ (modulo log $N$ factors). This is sufficient to prove (1.66), and thus we can apply our general theorem and prove the bulk universality of local statistics for these matrices. Much more difficult is the case of Wigner band matrices (1.18) where, roughly speaking, $\sigma^2_{ij} = 0$ if $|i - j| > W$ for some $W \ll N$. In this case we obtain ([9], [8]) the local semicircle law to the
energy scale $W^{-1}$, which is not strong enough to prove (1.66) if $W$ is much smaller than $N$ (the case $W \geq N^{1-\delta}$ with some small $\delta$ still works).

1.6.3. **Removing the Gaussian convolution.** I. **The reverse heat flow.** In the previous two subsections we discussed how to prove bulk universality for matrices with a small Gaussian convolution. The method of [14] (§1.6.1) required only a very small Gaussian component (with variance $N^{-1+\varepsilon}$), but it was restricted to the Hermitian case. The hydrodynamical method [7] (§1.6.2) works in general (the earlier versions [9], [8] assumed a larger Gaussian component with variance $\sim N^{-\xi}$). Both methods, however, need to be complemented by a perturbation step to remove this small Gaussian component.

Two independent approaches have been developed to remove the restriction on Gaussian divisibility. The first method is the reverse heat flow argument that appeared first in [16] and then was streamlined in §6 of [6]. The advantage of this method is that it can prove universality for a fixed energy $E$ as formulated in (1.35), and moreover, it is also very simple. The disadvantage is that it requires some smoothness of the distribution $\nu$ of the variables $\sqrt{N} h$, the rescaled entries of the Wigner matrix. We always assume that $\nu$ has subexponential decay, that is, that there are constants $C, \vartheta > 0$ such that for any $s$

$$\int 1(|x| \geq s) \, d\nu(x) \leq C \exp(-s^\vartheta). \tag{1.67}$$

The second method, which appeared slightly after the first, is the Green function comparison theorem via a perturbation argument with the four-moment condition. The advantage of this approach is that it holds for any distribution, the only condition being a subexponential decay (1.67). The disadvantage is that it proves universality (1.35) only after some averaging over $E$.

The four-moment condition was originally introduced by Tao and Vu [17] and used later in [64] and [69] in their study of eigenvalue perturbation which focused on the joint statistics of eigenvalues with fixed indices. In §4 we will present our approach [9] based on resolvent perturbation. Our result does not identify fixed eigenvalues but it is sufficiently strong to identify the local statistics, and its proof is much simpler than in [17] (see §1.6.4 for more explanation).

In this subsection we sketch the method of reverse heat flow, and in the next subsection we explain the four-moment comparison principles. For simplicity of the presentation we consider the Hermitian case, but we emphasize that this method applies to other symmetry classes as well, unlike the method outlined in §1.6.1. Consider the OU process defined in (1.55) with the variance $1/2$ kept fixed, and let $\gamma(dx) = \gamma(x) \, dx := \pi^{-1/2} e^{-x^2} \, dx$ denote the reversible measure of this process. Let $\nu_0(dx) = u(x) \gamma(dx)$ be the initial measure of the real and imaginary parts of the rescaled entries of the Wigner matrix $\hat{H}$ (note that in most of this paper $\nu$ denotes the distribution of $\sqrt{N} h_{ij}$, while for this discussion we have introduced the notation $\nu_0$ for the common distribution of $\sqrt{N} \text{Re} \, h_{ij}$ and $\sqrt{N} \text{Im} \, h_{ij}$, $i \neq j$). We let the OU process (1.55) act on the matrix elements, that is, we consider $H_t$, the solution to (1.54). For a fixed $t > 0$ the distribution of $H_t$ is given by

$$e^{-t/2} \hat{H} + (1 - e^{-t})^{1/2} V, \tag{1.68}$$
where $V$ is a GUE matrix, independent of $\tilde{H}$. The distribution of the real and imaginary parts of the matrix elements of $H_t$ is then given by $u_t(x) \gamma(dx)$, where $u_t$ is the solution to (1.55) with initial data $u_0 = u$ (strictly speaking, these formulae hold for the off-diagonal elements, while the diagonal elements have twice the variance and are subject to a slightly different OU flow).

The main observation is that the arguments in §1.6.1 or §1.6.2 guarantee the presence of the sine kernel for any Hermitian matrix that has a Gaussian component of variance $N^{-1+\varepsilon}$ or $N^{-\xi}$, respectively (the method of §1.6.1 applies only to the Hermitian case, while that of §1.6.2 works in general). Given a Wigner matrix $bH$, we do not necessarily have to compare $bH$ with its Gaussian convolution (1.68); it is sufficient to find another Wigner matrix $eH$ such that

$$bH \approx e^{-t/2}H + (1 - e^{-t})^{1/2}V$$

(1.69)

with a very high precision. In fact, $\tilde{H}$ can even be chosen to be $t$-dependent. The following lemma shows that any Wigner matrix $\tilde{H}$ with a sufficiently smooth distribution can be arbitrarily well approximated by Gaussian divisible matrices of the form (1.69) if $t \sim N^{-\delta}$ for some $\delta > 0$.

We assume that the initial density is positive, $u(x) > 0$, and that it can be written as

$$u(x) = e^{-V(x)}, \quad \text{with } \sum_{j=1}^{2K} |V^{(j)}(x)| \leq C_K (1 + x^2)^{C_K}$$

(1.70)

with any $K \in \mathbb{N}$ and with sufficiently large constants $C_K$. Moreover, we assume that the initial single-entry distribution $d\nu_0 = u d\gamma$ has subexponential decay (1.67).

The key technical lemma is the following approximation statement.

**Lemma 1.5** ([6], Proposition 6.1). Suppose that for some $K > 0$ the measure $d\nu_0 = u d\gamma$ satisfies (1.67) and (1.70). Then there is a small constant $\alpha_K$ depending on $K$ such that for any $t \leq \alpha_K$ there exists a probability density $g_t$ with mean zero and variance $1/2$ such that

$$\int |e^{tA} g_t - u| d\gamma \leq C t^K$$

(1.71)

for some constant $C > 0$ depending on $K$.

Furthermore, let $A = A \otimes n$ and $F = u \otimes n$ for some $n \leq CN^2$, and let $G_t = g_t \otimes n$. Then

$$\int |e^{tA} G_t - F| d\gamma \otimes n \leq C N^2 t^K$$

(1.72)

for some constant $C > 0$ depending on $K$.

**Sketch of the proof.** Given $u$, we want to solve the equation

$$e^{At} g_t = u,$$

that is, formally we have $g_t = e^{-At} u$. However, the operator $e^{-At}$ is like running a heat flow (with an OU drift) in reverse time, and is typically undefined unless
$u$ is analytic. But we can define an approximate solution to the backward heat equation, that is, we set

$$g_t := \left( I - At + \frac{t^2 A^2}{2!} - \cdots + \frac{(-tA)^{K-1}}{(K-1)!} \right) u.$$  

Since $A$ is a second-order differential operator and $u$ is sufficiently smooth, this expression is well defined, and moreover,

$$e^{At} g_t = O(t^K A^K u) = O(t^K).$$

This proves (1.71), and (1.72) directly follows from it.

Armed with Lemma 1.5, we can prove the sine-kernel universality in the form of (1.35) for any Hermitian Wigner matrix satisfying (1.67) and (1.70) for any fixed $|E| < 2$. We choose $n \sim N^2$ to be the number of independent OU processes needed to generate the flow of the matrix elements. By choosing $K$ large enough, we can compare the two measures $e^{tA} G_t$ and $F$ in the total variational norm; for any observable $J: \mathbb{R}^n \to \mathbb{R}$ of the matrix elements, we have

$$\left| \int J(e^{tA} G_t - F) \, d\gamma \otimes n \right| \leq \|J\|_\infty C N^2 t^K.$$  

In order to prove (1.35), appropriate observables $J$ need to be chosen that depend on the matrix elements via the eigenvalues and that express the local correlation functions. It is easy to see that for such a $J$ its norm $\|J\|_\infty$ can grow at most polynomially with respect to $N$. But we can always choose $K$ large enough to compensate for it with the choice $t = N^{-\delta}$. Since the sine kernel applies for the distribution $e^{tA} G_t$ with $t = N^{-1+\varepsilon}$ ($\S$ 1.6.1) or $t = N^{-\xi}$ ($\S$ 1.6.2), it will also apply for the Wigner measure $F$.

For symmetric matrices the reverse heat flow argument is exactly the same, but then only $\S$ 1.6.2 is available to obtain universality for short time; in particular, the $E$ in (1.35) needs to be averaged.

1.6.4. Removing the Gaussian convolution II. The Green function comparison theorem. Let $H$ and $H'$ be two Wigner ensembles such that the first four moments of the single-entry distributions $\nu$ and $\nu'$ coincide:

$$m_j = m'_j, \quad j = 1, 2, 3, 4, \quad (1.73)$$

where

$$m_j := \int x^j \, d\nu(x) \quad \text{and} \quad m'_j := \int x^j \, d\nu'(x).$$

For complex entries one has to take the collection of all $j$-moments, that is, $m_j$ represents the collection of all $\int_C x^a \bar{x}^b \, d\nu(x)$ with $a + b = j$. Recall that $\nu$ is the distribution of $\sqrt{N} h_{ij}$. By our normalization of Wigner matrices, the first moment is always 0 and the second moment is 1, so (1.73) is really a condition on the third and fourth moments.

Our main result is the following comparison theorem for the joint distribution of the Green functions. Here we only state the result in a somewhat simplified form. A more detailed presentation will be given in $\S$ 4.
Theorem 1.6 (Green function comparison theorem, [9], Theorem 2.3). Consider two Wigner matrices $H$ and $H'$ with single-entry distributions $\nu$ and $\nu'$. Assume that (1.67) and (1.73) hold for $\nu$ and $\nu'$. Let $G(z) = (H - z)^{-1}$ and $G'(z) = (H' - z)^{-1}$ denote the resolvents. Fix $k$ and suppose that the function $F: \mathbb{R}^k \to \mathbb{R}$ satisfies
\begin{equation}
\sup_{x \in \mathbb{R}^k} |\nabla^j F(x)| \leq N^{\varepsilon'}, \quad 0 \leq j \leq 5. \tag{1.74}
\end{equation}

Fix small parameters $\kappa$ and $\varepsilon$. Then for sufficiently small $\varepsilon'$ there is a constant $c_0 > 0$ such that for any integers $\ell_1, \ldots, \ell_k$ and for spectral parameters $z_j^m = E_j^m \pm i\eta$, $1 \leq j \leq \ell_m$, $m = 1, \ldots, k$, with $E_j^m \in [-2 + \kappa, 2 - \kappa]$ and $\eta \geq N^{-1-\varepsilon}$ the following inequality holds:
\begin{equation}
\left| E \left( \frac{1}{N} \text{Tr} \left[ \prod_{j=1}^{\ell_1} G(z_j^1) \right], \ldots, \frac{1}{N} \text{Tr} \left[ \prod_{j=1}^{\ell_k} G(z_j^k) \right] \right) - E' F(G \to G') \right| \leq N^{-c_0}. \tag{1.75}
\end{equation}

Here the shorthand notation $F(G \to G')$ means that we consider the same argument of $F$ as in the first term in (1.75), but all $G$ terms are replaced by $G'$.

In fact, the condition (1.73) can be weakened to require that the third and fourth moments be only close:
\begin{equation}
m_j = m'_j, \quad j = 1, 2, \quad \text{and} \quad |m_3 - m'_3| \leq N^{-1/2-\delta}, \quad |m_4 - m'_4| \leq N^{-\delta} \tag{1.76}
\end{equation}
with some $\delta > 0$. Then (1.75) still holds, but $\varepsilon$ and $\varepsilon'$ have to be sufficiently small, depending on $\delta$, and $c_0$ will also depend on $\delta$. The precise estimate will be stated in Theorem 4.1.

In other words, under the four-moment matching condition for two Wigner ensembles, the expectations of the traces of any combination of resolvent products coincide if the spectral parameters in the resolvents are not closer than $\eta = N^{-1-\varepsilon}$ to the real axis. Such a small distance corresponds to spectral resolution on the scale $N^{-1-\varepsilon}$, that is, it can identify local correlation functions of individual eigenvalues. It is an easy algebraic identity to express the correlation functions from the traces of the resolvents, for example, the one-point correlation function (density) on the scale $\eta$ is approximated by
\begin{equation}
p_N^{(1)}(E) \sim \frac{1}{\pi N} \text{Im} \text{Tr} G(E + i\eta) = \frac{1}{2\pi} \left[ \frac{1}{N} \text{Tr} G(E + i\eta) - \frac{1}{N} \text{Tr} G(E - i\eta) \right].
\end{equation}
and higher point correlation functions involve higher-order polynomials in the resolvents. Thus, Theorem 1.6 directly compares correlation functions (for the precise statement, see Theorem 4.2). We remark that taking traces is not essential in (1.75); a similar comparison principle works for matrix elements of the resolvents as well (see [9] for the precise formulation). In fact, the proof of Theorem 1.6 is a perturbation argument directly involving matrix elements of the resolvent. The key ingredient is a stronger form of the local semicircle law that lets us directly estimate $G_{ii}$ and $G_{ij}$, $i \neq j$, and not just the normalized trace $m(z) = \frac{1}{N} \sum_i G_{ii}$ (see (2.35)–(2.36) and (2.111) for the stronger result).
A related theorem for eigenvalues was proved earlier by Tao and Vu [17]. Let \( \lambda_1 < \cdots < \lambda_N \) and \( \lambda'_1 < \cdots < \lambda'_N \) denote the eigenvalues of \( H \) and \( H' \), respectively. The following theorem states that the joint distributions of any \( k \)-tuples of eigenvalues on the scale \( 1/N \) are very close to each other.

**Theorem 1.7** (four-moment theorem for eigenvalues; [17], Theorem 15). Let \( H \) and \( H' \) be two Wigner matrices and assume that (1.67) and (1.73) hold for their single-entry distributions \( \nu \) and \( \nu' \). For any sufficiently small positive \( \varepsilon \) and \( \varepsilon' \), for any function \( F: \mathbb{R}^k \to \mathbb{R} \) satisfying (1.74), and for any \( k \) indices \( i_1, \ldots, i_k \in [\varepsilon N, (1 - \varepsilon) N] \) chosen away from the spectral edges,

\[
|E F(N \lambda_{i_1}, \ldots, N \lambda_{i_k}) - E' F(N \lambda'_{i_1}, \ldots, N \lambda'_{i_k})| \leq N^{-c_0} \tag{1.77}
\]

for some \( c_0 > 0 \). The condition (1.73) can be relaxed to (1.76), but then \( c_0 \) will depend on \( \delta \).

We note that the arguments in (1.77) are magnified by a factor \( N \), and \( F \) is allowed to be concentrated on a scale \( N^{-\varepsilon'/5} \), so the result is sufficiently precise to detect eigenvalue correlations on the scale \( N^{-1-\varepsilon'/5} \), that is, even somewhat smaller than the eigenvalue spacing. Therefore, Theorem 1.6 or 1.7 can prove bulk universality for a Wigner matrix \( H \) if another matrix \( H' \) is found with matching four moments for which universality has already been proved. In the Hermitian case the GUE matrices, or more generally the Gaussian divisible matrices (1.53), provide a good reference ensemble \( H' \). Matching with a GUE matrix requires that the third and the fourth moments match: \( m_3 = 0 \) and \( m_4 = 3 \). Since the location of the eigenvalues for the GUE is known very precisely [92], [93], (1.77) can be interpreted as the limit of the correlation functions (see (1.35)) even at a fixed energy \( E \). If one aims only at the limiting gap distribution (1.36) instead of (1.35), then one can directly use the Gaussian divisible matrix (1.53) for matching. It is easy to check [94] that for any probability distribution \( \nu \) with \( m_1 = 0 \) and \( m_2 = 1 \) that is supported on at least three points, there is a distribution with an order-1 Gaussian component so that the first four moments match. Therefore, \( H \) can be matched with a Gaussian divisible matrix, for which Johansson proved universality [13]. Using the result of [14] on universality of Hermitian Wigner matrices with a tiny Gaussian convolution (conclusion of §1.6.1), and using the fact that the exact moment matching (1.73) can be relaxed to (1.76), one can compare any Wigner matrix \( H \) with its Ornstein–Uhlenbeck convolution (1.53) with very short time \( t \sim N^{-1+\varepsilon} \). This removes the condition \( m_3 = 0 \) and the condition that the support has at least three points, and proves the universality of the correlation functions for any Hermitian Wigner matrix in the sense of (1.35) after a little averaging with respect to \( E \) [18]. The only technical condition is the subexponential decay of \( \nu \) (1.67).

In the symmetric case, the analogue of Johansson’s result is not available (unless one uses [16]), and the only reference ensemble is a GOE. Theorem 1.7 thus implies [17] universality for symmetric Wigner matrices whose single-entry distribution has first four moments matching with a GOE in the sense of (1.76).

The careful reader may notice a subtle difference between the observable in (1.77) and the local correlation functions (1.35). While both detect the structure on the scale \( 1/N \), in (1.77) the indices of the eigenvalues are fixed, while in (1.35)
their locations are fixed. Roughly speaking, (1.77) can answer the question, say, as
to where the \((N/2)\)th and the \((N/2 - 1)\)th eigenvalues are. The local correlation
function asks about the probability of the simultaneous event that there is an
eigenvalue at \(E\) and another one at \(E' = E + \alpha/N\). These questions can be related
only if some a priori information is known about the location of the eigenvalues.

Prior to [17], apart from the GUE case [92], [93] there were no other ensembles
for which one could locate the eigenvalues with a precision of \(N^{-1 + c_0}\) for small \(c_0\),
and such precision is needed to translate (1.77) into (1.35) for a fixed \(E\). Using
a three-moment matching version of Theorem 1.7, one can locate the eigenvalues
of any Wigner ensembles with such a precision, provided that the third moment
vanishes (that is, matches with a GUE). Given this information, one can proceed
to match the fourth moment by choosing an appropriate Gaussian divisible matrix.
This is possible if the original distribution is supported on at least three points.
This is why (1.35) was eventually proved in [17] under the condition that the third
moment vanishes and the support contains at least three points. If one accepts that
(1.35) will be proved after some averaging with respect to \(E\), then the necessary
information on the location of the eigenvalues is much weaker, and it can typically
be obtained from the local semicircle law.

In fact, tracking individual eigenvalues can be a difficult task. We note that The-
orem 1.7 in itself does not directly imply convergence of the correlation functions;
one still needs some information about the location of the \(i\)th eigenvalue. On the
other hand, Theorem 1.7 contains information about eigenvalues with fixed indices
which was not contained in Theorem 1.6. We remark that the local semicircle law
is an essential input for both theorems.

The main reason why the proof of Theorem 1.6 is shorter is the fact that the
correlation functions can be identified from observables involving traces of resolvents
\((H - z)^{-1}\), where \(\text{Im} \ z \sim N^{-1 - \varepsilon}\), and these resolvents have an a priori bound of order
\(|\text{Im} \ z|^{-1} \leq N^{1 + \varepsilon}\), so perturbation formulae involving resolvents do not blow up. On
the other hand, the individual eigenvalues tracked by Theorem 1.7 may produce
resonances which could render some terms even potentially infinite (we will sketch
the proof of Theorem 1.7 in §4). While level repulsion is a general feature of
Wigner ensembles and it strongly suppresses resonances, the direct proof of the
level repulsion is not an easy task. In fact, the most complicated technical estimate
in [17] is the lower tail estimate of the gap distribution (Theorem 17 of [17]). It
states that for any \(c_0 > 0\) there is a \(c_1\) such that

\[
P(\lambda_{i+1} - \lambda_i \leq N^{-1 - c_0}) \leq N^{-c_1},
\]

if the index \(i\) lies in the bulk \(\varepsilon N \leq i \leq (1 - \varepsilon)N\).

1.6.5. Summary of new results on bulk universality. Even the expert reader may
find the recent developments slightly confusing, in view of the many papers on
bulk universality of Wigner matrices under various conditions and with different
methods. Their interrelation has not always been optimally presented in research
publications, since it was, and still is, a fast developing story. In this subsection we
try to give some orientation to the reader for the recent literature.
As mentioned in the Introduction, the main guiding principle behind these proofs of universality of local eigenvalue statistics is to compare the local statistics of a Wigner matrix with another matrix having a Gaussian component. More precisely, our approach consists of three main steps:

1) the local semicircle law;

2) universality for Gaussian divisible ensembles, that is, when the probability law of the matrix elements contains a small Gaussian component;

3) universality for general ensembles, approximation by a Gaussian divisible ensemble to remove the small Gaussian component.

It was clear to us from the very beginning that a good local semicircle law must be the first step in any proof of universality. In fact, all proofs of universality rely heavily on the details of the estimates one can obtain for the local semicircle law. We now summarize the existing results according to these three steps.

Although the proof of the local semicircle law down to the shortest scale $\eta \sim 1/N$ is rather simple now, it was only gradually achieved. In our first paper [55] we gave an upper bound on the local density essentially down to the optimal energy scale $\eta \sim (\log N)/N$, but the local semicircle law was proved only on the scale $\eta \gg N^{-2/3}$. In the second paper [56] we proved the local semicircle law down to the scale $\eta \geq (\log N)^8/N$, almost optimal but still off by a logarithmic factor. The tail probability of violation of the local semicircle law was also far from optimal. Both defects were remedied in [57] (see Theorem 1.11 below), where in addition an optimal delocalization result for eigenvectors was also proved (Theorem 2.22). In the first paper [55] we assumed a strong (Gaussian) decay condition and a certain convexity property of the single-entry distribution that implies concentration (via either the Brascamp–Lieb or logarithmic Sobolev inequalities). These technical conditions were subsequently removed and the Gaussian decay condition was replaced by a subexponential decay. Finally, in [9] and in its improved and streamlined version in [8], we obtained a much stronger error estimate for the local semicircle law (see (1.4)–(1.7)), but these estimates still deteriorated at the edges. The optimal result [7], which we call the strong local semicircle law (Theorem 2.19), holds uniformly with respect to the energy parameter.

As for Step 2), the main point is that the Gaussian component enables one to exhibit the universal behaviour. There are two ways to implement this idea:

i) the contour integral representation following Johansson [13] and Ben-Arous and Péché [85], though this option is available only for the Hermitian case (or for the complex sample covariance case, [85]);

ii) the hydrodynamical approach with a small Gaussian component (equivalently, a small-time evolution of an OU process) driving the system to local equilibrium, an approach which applies to all ensembles, including symmetric, Hermitian, symplectic, and sample covariance ensembles, and which also gives a conceptual understanding that the universality arises from the Dyson Brownian motion.

Both approaches require a good local semicircle law. Additionally, the earlier papers [16] and [6] on the hydrodynamical methods also assumed the logarithmic Sobolev inequality (LSI) for the single-entry distribution $\nu$, essentially in order to verify (1.66) from the local semicircle law. In [8] we removed this last condition by using a strengthening of the local semicircle law which gave a simpler and more
powerful proof of (1.66) (Theorem 2.7 below). Finally, the strong local semicircle law in [7] (Theorem 2.19 below) provided the optimal exponent $2\varepsilon = 1$ in (1.66).

Summarizing the first two steps, we thus obtained bulk universality for generalized Wigner matrices (1.17) with a small Gaussian convolution under the sole condition of subexponential decay. This condition can be relaxed to a high-order polynomial decay, but we have not yet worked out the details. Furthermore, although the extension of the strong local semicircle law to sample covariance matrices is straightforward, these details have also not yet been worked out (the earlier detailed proof [6] required the LSI).

The first two steps provide a large class of matrix ensembles with universal local statistics. In Step 3) it remains to approximate arbitrary matrix ensembles by these matrices so that the local statistics are preserved. This approximation step can be done in two ways:

a) via reverse heat flow;

b) via the Green function comparison theorem.

The reverse heat flow argument is very simple, but it requires smoothness of the single-entry distribution. This approach was used in [14], [16], and [6] and leads to universality for all the ensembles mentioned under the smoothness condition. This smoothness condition was then removed in [9], where the Green function comparison theorem was first proved. Unfortunately, we still needed the LSI and universality was established only for matrices whose distribution $\nu$ is supported on at least three points.

A stronger version of the local semicircle law was proved in [8] and all smoothness and support conditions on the distributions were removed. In summary, in [8] we obtained bulk universality of correlation functions (1.35) and gap distribution (1.36) for all classical Wigner ensembles (including the generalized Wigner matrices, (1.17)). The universality in (1.35) is understood after a small averaging with respect to the energy parameter $E$. The only condition on the single-entry distribution $\nu$ is the subexponential decay (1.67).

The approach of Tao and Vu [17] uses a similar strategy of the three Steps 1)–3) at the beginning of this subsection. For Step 2) the universality for Hermitian Wigner matrices and complex sample covariance matrices was previously proved by Johansson [13] and Ben-Arous and Péché [85]. Step 3) follows from the Tao–Vu four-moment theorem (Theorem 1.7), whose proof uses the local semicircle law, Step 1). This leads to universality for Hermitian Wigner matrices [17] and complex sample covariance matrices satisfying the condition that the support of the distribution contains at least three points (and if one aims at a fixed energy result in (1.35), then the third moment also has to vanish). For the symmetric case, the matching of the first four moments was required. The Tao–Vu approach can also be applied to prove edge universality [64]. In [69] the subexponential decay was replaced by a condition with a sufficiently strong polynomial decay. The support condition and the third-moment condition can be removed by combining [17] with a result following from our approach [14], and this has led to the universality of Hermitian matrices [18] for any distribution, including the Bernoulli measure. On the other hand, even for Hermitian matrices, the variances of the matrix elements are required to be constant in this approach.
Historically, the first Tao–Vu paper on universality [17] appeared shortly after the paper [14] on the universality of Hermitian matrices. A common ingredient for both [14] and [17] is the local semicircle law and the eigenfunction delocalization estimates that were essentially available from [56], [57], but due to certain technical conditions they were re-proved in [17]. The local semicircle law for sample covariance matrices was first proved in [6], and a slightly different version was given in [69] with some changes in the technical assumptions tailored to applications.

The four-moment condition first appeared in the four-moment theorem of Tao–Vu ([17], Theorem 1.7) and it was used in the Green function comparison theorem ([9], Theorem 1.6). The four-moment theorem concerns individual eigenvalues and thus it directly contains information about the eigenvalue gap distribution. In order to translate this information into correlation functions, the locations of individual eigenvalues of the comparison ensemble are required. The Green function comparison theorem, on the other hand, can be used to compare the correlation functions directly, but the information on the individual eigenvalues is weaker. Nevertheless, a standard exclusion-inclusion principle argument (like the one presented in (1.38)) also gives universality of the gap distribution. Since individual eigenvalues tend to fluctuate and Green functions are more stable, this explains why the proof of the four-moment theorem for eigenvalues is quite involved, while the Green function comparison theorem is very simple. Furthermore, the Green function comparison theorem yields not only spectral information, but also information on matrix elements.

1.6.6. New results on edge universality. Recall that \( \lambda_N \) is the largest eigenvalue of the random matrix. The probability distribution functions of \( \lambda_N \) for the classical Gaussian ensembles were found by Tracy and Widom [65], [66] to be

\[
\lim_{N \to \infty} \mathbb{P}(N^{2/3}(\lambda_N - 2) \leq s) = F_\beta(s),
\]

where the function \( F_\beta(s) \) can be computed in terms of Painlevé equations and \( \beta = 1, 2, 4 \) correspond to the standard classical ensembles. The distribution of \( \lambda_N \) is believed to be universal and independent of the Gaussian structure. The strong local semicircle law, Theorem 2.19, combined with a modification of the Green function comparison theorem, Theorem 1.6, and tailored to the spectral edges, implies the following version of universality of the extreme eigenvalues.

**Theorem 1.8** (universality of extreme eigenvalues [7], Theorem 2.4). Suppose that \( H^{(v)} \) and \( H^{(w)} \) are two \( N \times N \) generalized Wigner matrices with matrix elements \( h_{ij} \) given by the random variables \( N^{-1/2}v_{ij} \) and \( N^{-1/2}w_{ij} \), respectively, where \( v_{ij} \) and \( w_{ij} \) satisfy the subexponential decay condition (1.67) uniformly with respect to all \( i, j \). Let \( \mathbb{P}^v \) and \( \mathbb{P}^w \) denote the probabilities and \( \mathbb{E}^v \) and \( \mathbb{E}^w \) the expectations with respect to these collections of random variables. If the first two moments of \( v_{ij} \) and \( w_{ij} \) are the same, that is,

\[
\mathbb{E}^v v_{ij}^l v_{ij}^u = \mathbb{E}^w w_{ij}^l w_{ij}^u, \quad 0 \leq l + u \leq 2,
\]

(1.80)
then there exist an $\varepsilon > 0$ and a $\delta > 0$ depending on $\vartheta$ in (1.67) such that for any $s \in \mathbb{R}$

$$
P^v(N^{2/3}(\lambda_N - 2) \leq s - N^{-\varepsilon}) - N^{-\delta}
\leq P^w(N^{2/3}(\lambda_N - 2) \leq s) \leq P^v(N^{2/3}(\lambda_N - 2) \leq s + N^{-\varepsilon}) + N^{-\delta}, \quad (1.81)
$$

where $N$ is sufficiently large and independent of $s$. An analogous result holds for the smallest eigenvalue $\lambda_1$.

Theorem 1.8 can be extended to finite correlation functions of the extreme eigenvalues. For example, we have the following extension of the relations (1.81):

$$
P^v(N^{2/3}(\lambda_N - 2) \leq s_1 - N^{-\varepsilon}, \ldots, N^{2/3}(\lambda_{N-k} - 2) \leq s_{k+1} - N^{-\varepsilon}) - N^{-\delta}
\leq P^w(N^{2/3}(\lambda_N - 2) \leq s_1, \ldots, N^{2/3}(\lambda_{N-k} - 2) \leq s_{k+1})
\leq P^v(N^{2/3}(\lambda_N - 2) \leq s_1 + N^{-\varepsilon}, \ldots, N^{2/3}(\lambda_{N-k} - 2)s_{k+1} + N^{-\varepsilon}) + N^{-\delta}
\quad (1.82)
$$

for any fixed $k$ and sufficiently large $N$.

Edge universality for Wigner matrices was first proved via the moment method by Soshnikov [62] (see also the earlier paper [63]) for Hermitian and symmetric ensembles with symmetric single-entry distribution $\nu$ to ensure that all odd moments vanish. By combining the moment method and Chebyshev polynomials [95], Sodin proved edge universality for band matrices and for a certain special class of sparse matrices [96], [97].

The removal of the symmetry assumption was not straightforward. The approach of [96], [97] is restricted to ensembles with symmetric distributions. The symmetry assumption was partially removed in [98] and [68], and significant progress was made in [64], where it is assumed only that the first three moments of two Wigner ensembles are identical. In other words, the symmetry assumption was replaced by the vanishing third-moment condition for Wigner matrices. For a special class of ensembles, the Gaussian divisible Hermitian ensembles, edge universality was proved in [86] under the sole condition that the second moment is finite. By a combination of methods from [86] and [64], the same result can be proved for all Hermitian Wigner ensembles with finite second moment [86].

In comparison with these results, Theorem 1.8 does not imply edge universality for band matrices or sparse matrices [96], [97], but it does imply in particular that, for the purpose of identifying the distribution of the top eigenvalue for a generalized Wigner matrix, it suffices to consider generalized Wigner ensembles with Gaussian distribution. Since the distributions of the top eigenvalues of the Gaussian Wigner ensembles are given by $F_\beta$ in (1.79), Theorem 1.8 shows edge universality for standard Wigner matrices under the subexponential decay assumption alone. We remark that one can use Theorem 2.20 as an input in the approach of [86] to prove that the distributions of the top eigenvalues of generalized Hermitian Wigner ensembles with Gaussian distributions are given by $F_2$. But for ensembles in a different symmetry class, there is no corresponding result letting us identify the distribution of the top eigenvalue with $F_\beta$.

Finally, we comment that the subexponential decay assumption in our approach, though it can be weakened, is far from optimal for edge universality [99], [100], [68].
1.7. Level repulsion and the Wegner estimate on very short scales. One of our earlier local semicircle laws for Wigner matrices,

\[ |m(z) - m_{sc}(z)| \lesssim \frac{C}{\sqrt{N \eta \kappa}}, \quad \kappa = ||E| - 2|, \tag{1.83} \]

proved in Theorem 4.1 of [89], can be turned into a direct estimate of the empirical density in the form

\[ |c_{\eta}(E) - c_{sc}(E)| \lesssim \frac{C}{\sqrt{N \eta \kappa}}, \quad \kappa = ||E| - 2|. \]

Here \( c_{\eta} \) denotes the empirical density \( c(x) = \frac{1}{N} \sum_{i} \delta(\lambda_i - x) \) smoothed out on a scale \( \eta \). This result asserts that the empirical density on scales \( \eta \gg O(1/N) \) is close to the semicircle density. On even smaller scales \( \eta \lesssim O(1/N) \) the empirical density fluctuates, but its average \( E c_{\eta}(E) \) remains bounded uniformly with respect to \( \eta \). This is a type of Wegner estimate that plays a central role in the localization theory of random Schrödinger operators. In particular, it says that the probability of finding at least one eigenvalue in an interval \( I \) of size \( \eta = \varepsilon/N \) is bounded by \( C\varepsilon \) uniformly with respect to \( \varepsilon \gg 1 \) and for all sufficiently large \( N \). The exponents here are optimal, as one can easily see from the Vandermonde determinant in the joint probability density (1.43) for invariant ensembles. The sine-kernel behaviour (1.32) indicates level repulsion and even gives a lower bound on \( P(\mathcal{N}_I \geq n) \), but usually not on arbitrarily small scales, since the presence of the sine kernel is typically proved only as a weak limit (see (1.35)).

We also mention that (1.78) (Theorem 17 from [17]) is also a certain type of level repulsion bound, but the exponents there are not optimal and it does not hold on arbitrarily small scales. However, the virtue of (1.78) is that it assumes no smoothness of the distribution, in particular, it holds for discrete distributions as well. It is clear that, say, for the Bernoulli distribution not even the Wegner estimate (1.84) for \( n = 1 \) can hold on superexponentially small scales \( \varepsilon \sim 2^{-N^2} \).
Sketch of the proof. The first step of the proof is to provide an upper bound on $\mathcal{N}_I$. Let $H^{(k)}$ denote the $(N-1) \times (N-1)$ minor of $H$ after removing the $k$th row and $k$th column. Let $\lambda^{(k)}_\alpha$, $\alpha = 1, \ldots, N-1$, denote the eigenvalues of $H^{(k)}$ and let $u^{(k)}_\alpha$ denote its eigenvectors. Computing the $(k,k)$ diagonal element of $(H - z)^{-1}$, we easily obtain the following expression for $m(z) = m_N(z)$:

$$m(z) = \frac{1}{N} \sum_{k=1}^{N} \left( \frac{1}{H - z} \right)_{k,k} = \frac{1}{N} \sum_{k=1}^{N} \left[ h_{kk} - z - \frac{1}{N} \sum_{\alpha=1}^{N-1} \frac{\xi^{(k)}_\alpha}{\lambda^{(k)}_\alpha - z} \right]^{-1},$$

where

$$\xi^{(k)}_\alpha := N|a^{(k)}_\alpha \cdot u^{(k)}_\alpha|^2,$$

and $a^{(k)}$ is the $k$th column of $H$ without the diagonal element $h_{kk}$. Taking the imaginary part and using the inequality

$$\mathcal{N}_I \leq CN \eta \Im m(z), \quad z = E + i\eta,$$

we have

$$\mathcal{N}_I \leq CN \eta^2 \sum_{k=1}^{N} \left( \sum_{\alpha: \lambda^{(k)}_\alpha \in I} \xi^{(k)}_\alpha \right)^{-1}.$$  

(1.88)

It is an elementary fact that for each fixed $k$ the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$ of $H$ and the eigenvalues $\mu_1 \leq \cdots \leq \mu_{N-1}$ of $H^{(k)}$ are interlaced, meaning that

$$\lambda_1 \leq \mu_1 \leq \lambda_2 \leq \mu_2 \leq \cdots \leq \mu_{N-1} \leq \lambda_N.$$  

(1.89)

This can be seen by analyzing the equation for the eigenvalues $\lambda$ in terms of the eigenvalues $\mu$:

$$\lambda - h_{ii} = \sum_{\alpha=1}^{N-1} \frac{|a^i \cdot u^i_\alpha|^2}{\lambda - \mu_\alpha},$$

where $u^i_\alpha$ is the normalized eigenvector of the minor $H^{(i)}$ belonging to $\mu_\alpha$.

The interlacing property clearly implies that the number of eigenvalues $\lambda^{(k)}_\alpha$ in $I$ is at least $\mathcal{N}_I - 1$. For each fixed $k$ the random variables $\{\xi^{(k)}_\alpha: \alpha = 1, \ldots, N-1\}$ are almost independent and have expectation value 1, and thus the probability of the event

$$\Omega_k := \left\{ \sum_{\alpha: \lambda^{(k)}_\alpha \in I} \xi^{(k)}_\alpha \leq \delta(\mathcal{N}_I - 1) \right\}$$

is negligible for small $\delta$ ([57], Lemma 4.7). On the complement of all the events $\Omega_k$ we thus have from (1.88) that

$$\mathcal{N}_I \leq \frac{CN^2 \eta^2}{\delta(\mathcal{N}_I - 1)},$$

from which it follows that $\mathcal{N}_I \leq CN \eta$ with very high probability. One of the precise results of this type is the following statement.
Lemma 1.10 ([57], Theorem 4.6). If the single-entry distribution \( \nu \) has a Gaussian decay, then for any interval \( I \) with \(|I| \geq (\log N)/N\)

\[
P(\mathcal{N}_I \geq KN|I|) \leq Ce^{-c\sqrt{KN|I|}}.
\]

We remark that the Gaussian decay condition can be weakened, and a somewhat weaker result holds also for even shorter intervals \(|I| \geq 1/N\) (see Theorem 5.1 in [57]).

The proof of Theorem 1.9 also starts with (1.85) and (1.87). They imply that

\[
\mathcal{N}_I \leq C\eta \sum_{k=1}^{N} \frac{1}{(a_k^2 + b_k^2)^{1/2}},
\]

with

\[
a_k := \eta + \frac{1}{N} \sum_{\alpha=1}^{N-1} \frac{\eta \xi^{(k)}(\alpha)}{\left(\lambda^{(k)}_{\alpha} - E\right)^2 + \varepsilon^2}, \quad b_k := h_{kk} - E - \frac{1}{N} \sum_{\alpha=1}^{N-1} \frac{(\lambda^{(k)}_{\alpha} - E)\xi^{(k)}(\alpha)}{\left(\lambda^{(k)}_{\alpha} - E\right)^2 + \varepsilon^2},
\]

that is, \(a_k\) and \(b_k\) are the imaginary part and real part, respectively, of the reciprocal of the \(k\)th summand in (1.85). The proof of Lemma 1.10 used only the imaginary part, that is, \(b_k\) in (1.90) was neglected in the estimate (1.88). In the proof of Theorem 1.9, however, we make essential use of \(b_k\) as well. Since typically \(1/N \lesssim |\lambda^{(k)}_{\alpha} - E|\), we note that \(a_k^2\) is much smaller than \(b_k^2\) if \(\eta \ll 1/N\), and this is the relevant regime for the Wegner estimate and for the level repulsion.

For a certain smoothness of the single-entry distribution \(d\nu\) the distribution of \(\xi^{(k)}_{\alpha}\) will also be smooth, even if we fix an index \(k\) and impose the condition on the minor \(H^{(k)}\), that is, if we fix the eigenvalues \(\lambda^{(k)}_{\alpha}\) and the eigenvectors \(u^{(k)}_{\alpha}\). Although the random variables \(\xi^{(k)}_{\alpha} = N|a^{(k)}_{\alpha} \cdot u^{(k)}_{\alpha}|^2\) are not independent for different \(\alpha\), they are sufficiently uncorrelated so that the distribution of \(b_k\) inherits some smoothness from that of \(a^{(k)}_{\alpha}\). Sufficient smoothness of the distribution of \(b_k\) makes the expectation of \((a_k^2 + b_k^2)^{-p/2}\) finite for any \(p > 0\). This will give a bound on the \(p\)th moment on \(\mathcal{N}_I\), which will imply (1.84).

We present this idea for Hermitian matrices and for the simplest case \(n = 1\). From (1.90) we have

\[
P(\mathcal{N}_I \geq 1) \leq E\mathcal{N}_I^2 \leq C(N\eta)^2 E \frac{1}{a_1^2 + b_1^2}.
\]

Dropping the index \(k = 1\) and introducing the notation

\[
d_{\alpha} = \frac{N(\lambda_{\alpha} - E)}{N^2(\lambda_{\alpha} - E)^2 + \varepsilon^2}, \quad c_{\alpha} = \frac{\varepsilon}{N^2(\lambda_{\alpha} - E)^2 + \varepsilon^2},
\]

we have

\[
P(\mathcal{N}_I \geq 1) \leq C\varepsilon^2 E \left[ \left( \sum_{\alpha=1}^{N-1} c_{\alpha} \xi_{\alpha} \right)^2 + \left( h - E - \sum_{\alpha=1}^{N-1} d_{\alpha} \xi_{\alpha} \right)^2 \right]^{-1}. \tag{1.91}
\]
From one version of the local semicircle law (see Theorem 1.11 below), we know that with a very high probability there are several eigenvalues $\lambda_\alpha$ within a distance of $O(1/N)$ from $E$. Choosing four such eigenvalues, we can guarantee that for some index $\gamma$ we have

$$c_\gamma, c_{\gamma+1} \geq C\varepsilon, \quad d_{\gamma+2}, d_{\gamma+3} \geq C$$

(1.92)

for some positive constant $C$. If the $\xi_\alpha$ were indeed independent and distributed according to the square of a complex random variable $z_\alpha$ with a smooth and decaying density $d\mu(z)$ on the complex plane, then the expectation in (1.91) would be bounded by

$$\sup_E \int \frac{1}{(c_\gamma |z_\gamma|^2 + c_{\gamma+1}|z_{\gamma+1}|^2)^2 + (E - d_{\gamma+2}|z_{\gamma+2}|^2 - d_{\gamma+3}|z_{\gamma+3}|^2)^2} \prod_{j=0}^3 d\mu(z_{\gamma+j}).$$

(1.93)

A simple calculation shows that this integral is bounded by $C\varepsilon^{-1}$, assuming the lower bounds (1.92). Combining this bound with (1.91), we obtain (1.84) for $n = 1$. The proof for general $n$ goes by induction. The difference between the Hermitian and the symmetric cases manifests itself in the fact that the $\xi_\alpha$ are squares of complex or real variables, respectively. This gives different estimates for integrals of the type (1.93), resulting in different exponents in (1.84).

In this proof we used the following version of the local semicircle law.

**Theorem 1.11** ([57], Theorem 3.1). Let $H$ be an $N \times N$ Hermitian or symmetric Wigner matrix with a single-entry distribution having a Gaussian decay. Let $\kappa > 0$ and fix an energy $E \in [-2 + \kappa, 2 - \kappa]$. Then there exist positive constants $C$ and $c$ depending only on $\kappa$, and a universal constant $c_1 > 0$ such that the following hold:

1) For any $\delta \leq c_1\kappa$ and $N \geq 2$,

$$P\left( |m(E + i\eta) - m_{\text{sc}}(E + i\eta)| \geq \delta \right) \leq C e^{-c\delta \sqrt{N\eta}}$$

(1.94)

for $K/(\sqrt{N}/E) \leq \eta \leq 1$, where $K$ is a large universal constant.

2) Let $N_{\eta^*}(E) = N_{I^*}$ denote the number of eigenvalues in the interval $I^* := [E - \eta^*/2, E + \eta^*/2]$. Then for any $\delta \leq c_1\kappa$ there is a constant $K_\delta$ depending only on $\delta$ such that

$$P\left\{ \left| \frac{N_{\eta^*}(E)}{N\eta^*} - \rho_{\text{sc}}(E) \right| \geq \delta \right\} \leq C e^{-c\delta^2 \sqrt{N\eta^*}}$$

(1.95)

holds for all $\eta^*$ with $K_\delta/N \leq \eta^* \leq 1$ and for all $N \geq 2$.

**2. Local semicircle law and delocalization**

Each approach that proves bulk universality for generalized Wigner matrices requires first of all an analysis of the local density of the eigenvalues. The Wigner semicircle law [21] (and its analogue for Wishart matrices, the Marchenko–Pastur law [19]) has traditionally been among the first results established for random matrices. Typically, however, the empirical density is shown to converge weakly on macroscopic scales, that is, on intervals that contain $O(N)$ eigenvalues. Based on our results [55]–[57], [89], [7]–[9] we show here that the semicircle law holds on
much smaller scales as well. In §2.2 we follow the formalism of [57], while in §2.3 we use [9] and [8]. The former formalism aims directly at the Stieltjes transform, or the trace of the resolvent, while the latter formalism is designed to establish the semicircle law for individual diagonal elements of the resolvent, and it also gives an estimate of the off-diagonal elements. The strongest result [7] that holds uniformly with respect to the energy parameter is presented in §2.4. Finally, in §2.5 we indicate how to prove delocalization of eigenvectors from the local semicircle law.

2.1. Resolvent formulae. For definiteness we present the proof for the Hermi-
tian case, but all formulae below carry over to the other symmetry classes with obvious modifications. We first collect a few useful formulae about resolvents. Their proofs are elementary results from linear algebra.

**Lemma 2.1.** Let $A, B, C$ be $n \times n$, $m \times n$, and $m \times m$ matrices. Define the $(m + n) \times (m + n)$ matrix $D$ as

$$D := \begin{pmatrix} A & B^* \\ B & C \end{pmatrix}$$

and the $n \times n$ matrix $\hat{D}$ as

$$\hat{D} := A - B^*C^{-1}B.$$ 

Then for any $1 \leq i, j \leq n$

$$(D^{-1})_{ij} = (\hat{D}^{-1})_{ij}.$$ 

Recall that $G_{ij} = G_{ij}(z)$ denotes the matrix element of the resolvent,

$$G_{ij} = \left( \frac{1}{H - z} \right)_{ij}.$$ 

Let $G^{(i)}$ denote the resolvent of $H^{(i)}$, which is the $(N - 1) \times (N - 1)$ minor of $H$ obtained by removing the $i$th row and column. Let $a^i = (h_{1i}, h_{2i}, \ldots, h_{Ni})^t$ be the $i$th column of $H$, sometimes after removing one or more elements. We always keep the original labelling of the rows and columns, so there will be no confusion: if $a^i$ is multiplied by a matrix whose $j$th column and row are removed, then we remove the $j$th entry from $a^i$ as well. With similar conventions we can define the matrix $G^{(ij)}$, and so on. The superscript in parentheses for resolvents always means after removing the corresponding row and column. In particular, by independence of the matrix elements, this means that the matrix $G^{(ij)}$, say, is independent of the $i$th and $j$th rows and columns of $H$. This helps in decoupling the dependencies in formulae.

Using Lemma 2.1 for $n = 1$, $m = N - 1$, we have

$$G_{ii} = \frac{1}{h_{ii} - z - a^i} \cdot \frac{1}{H^{(i)} - z} a^i = \frac{1}{h_{ii} - z - a^i \cdot G^{(i)} a^i},$$

where $a^i$ is the $i$th column with the $i$th entry $h_{ii}$ removed.
For the off-diagonal elements, one has to do a two-row expansion. In this case let $\mathbf{a}^1$ and $\mathbf{a}^2$ denote the first and the second columns of $H$ after removing the first and second elements, that is, $h_{11}$ and $h_{21}$, from the first column and $h_{12}$ and $h_{22}$ from the second. With the notation $D = H - z$, $B = [\mathbf{a}^1, \mathbf{a}^2]$, and $C = H^{12} - z$ in Lemma 2.1 with $n = 2$ and $m = N - 2$, we can compute the matrix $\hat{D}$, which in this case we will call $K^{(12)}$:

$$
\hat{D} = \begin{pmatrix}
    h_{11} - z - \mathbf{a}^1 \cdot G^{(12)} \mathbf{a}^1 & h_{12} - \mathbf{a}^1 \cdot G^{(12)} \mathbf{a}^2 \\
    h_{21} - \mathbf{a}^2 \cdot G^{(12)} \mathbf{a}^1 & h_{22} - z - \mathbf{a}^2 \cdot G^{(12)} \mathbf{a}^2
\end{pmatrix} =:
\begin{pmatrix}
    K_{11}^{(12)} & K_{12}^{(12)} \\
    K_{21}^{(12)} & K_{22}^{(12)}
\end{pmatrix},
$$

(2.5)

where for convenience we have introduced the notation

$$
K_{ij}^{(12)} := h_{ij} - z\delta_{ij} - \mathbf{a}^i \cdot G^{(12)} \mathbf{a}^j, \quad i, j = 1, 2.
$$

(2.6)

Thus, from Lemma 2.1 we have, for example,

$$
G_{11} = \frac{K_{22}^{(12)}}{K_{22}^{(12)} K_{11}^{(12)} - K_{12}^{(12)} K_{21}^{(12)}},
$$

(2.7)

and

$$
G_{12} = -\frac{K_{12}^{(12)}}{K_{22}^{(12)} K_{11}^{(12)} - K_{12}^{(12)} K_{21}^{(12)}} = -G_{22}^2 \frac{K_{12}^{(12)}}{K_{11}^{(12)}} = -G_{22}^2 G_{11}^{(2)} K_{12}^{(12)}.
$$

(2.8)

In the last step we used the identity

$$
G_{11}^{(2)} = \frac{1}{K_{11}^{(12)}},
$$

(2.9)

which is exactly the one-row expansion (2.4), applied to the minor $H^{(2)}$ of $H$ obtained by removing the second row and second column.

There is another set of formulae which express how to compare the resolvents of $H$ and $H^{(1)}$; for example, for any $i \neq j$

$$
G_{ii} = G_{ii}^{(j)} + \frac{G_{ij} G_{ji}}{G_{jj}}.
$$

(2.10)

This can easily be checked on a $2 \times 2$ matrix and its inverse:

$$
M = \begin{pmatrix}
    a & b \\
    c & d
\end{pmatrix}, \quad M^{-1} = \frac{1}{\Delta} \begin{pmatrix}
    d & -c \\
    -b & a
\end{pmatrix} \quad \text{with } \Delta = ad - bc,
$$

so checking (2.10), for instance, for $i = 1$ and $j = 2$ boils down to the identity

$$
\frac{d}{\Delta} = \frac{1}{a} + \frac{c}{a} \frac{b}{\Delta}.
$$
For larger matrices one just uses (2.3). Note that in formulae (2.7), (2.8), and (2.9) we already expressed all the resolvents appearing in (2.10) in terms of matrix elements of the $2 \times 2$ matrix $K^{(12)}$ in (2.5), which can play the role of $M$ above. Similarly, one has for any three different indices $i, j, k$ that

$$G_{ij} = G_{ij}^{(k)} + \frac{G_{ik}G_{kj}}{G_{kk}}. \quad (2.11)$$

This identity can be checked on $3 \times 3$ matrices and then proved by induction in the general case.

### 2.2. Semicircle law via resolvents: Sketch of a crude method

In this subsection we sketch the proof of the following result.

**Theorem 2.2.** Let $z = E + i\eta$, $1/N \ll \eta \ll 1$, and $\kappa := |E| - 2$. Let $H$ be a Wigner matrix and $G(z) = (H - z)^{-1}$ its resolvent, and let $m(z) := \frac{1}{N} \text{Tr} G(z)$. Assume that the single-entry distribution $\nu$ has Gaussian decay ($\vartheta = 2$ in (1.67)). Then the approximation

$$|m(z) - m_{sc}(z)| \leq \min \left\{ \frac{(\log N)^C}{\sqrt{N\eta\kappa}}, \frac{(\log N)^C}{(N\eta)^{1/4}} \right\} \quad (2.12)$$

holds with a very high probability.

We proved the local semicircle law in this form in Proposition 8.1 of [6] for sample covariance matrices (replacing the semicircle distribution by the Marchenko–Pastur distribution), but the same (or even easier) proof works for Wigner matrices. The original proof was presented with a Gaussian decay condition, but it can easily be relaxed to subexponential decay; this affects only the estimate of the probability that the event (2.12) is violated. [For technical experts: in our previous papers, up to [6], we typically used the Hanson–Wright theorem [101] to estimate large deviation probabilities of quadratic forms. This gives a very good control for the tail, but requires Gaussian decay. In our more recent papers we use Lemma 2.12, based on martingale inequalities, which requires only subexponential decay, and in fact can be relaxed to polynomial decay, though the tail probability estimate is weaker.]

For the proof we start with the identity (2.4) and express $G_{ii}$ in the form

$$G_{ii} = \frac{1}{h_{ii} - z - E_i a^i \cdot G^{(i)} a^i - Z_i}, \quad (2.13)$$

where we split the quantities

$$a^i \cdot G^{(i)} a^i = E_i a^i \cdot G^{(i)} a^i + Z_i, \quad Z_i := a^i \cdot G^{(i)} a^i - E_i a^i \cdot G^{(i)} a^i$$

into their expectations and fluctuations ($E_i$ denotes the expectation with respect to the variables in the $i$th column and $i$th row). In particular, $G^{(i)}$ is independent of $a^i$, so we need to compute the expectations and fluctuations of quadratic functions.
The expectation is easy:

\[ E_i a_i \cdot G^{(i)} a_i = E_i \sum_{k,l \neq i} \overline{a_k} G_{kl}^{(i)} a_i = \sum_{k,l \neq i} E_i \overline{h_{ik}} G_{kl}^{(i)} h_{il} = \frac{1}{N} \sum_{k \neq i} G_{kk}^{(i)}, \]

where in the last step we used the fact that different matrix elements are independent, that is, \( E_i \overline{h_{ik}} h_{il} = \frac{1}{N} \delta_{kl} \). The summation always runs over all indices from 1 to \( N \), apart from those that are explicitly excluded.

Just as we defined

\[ m(z) = \frac{1}{N} \text{Tr} G(z) = \frac{1}{N} \sum_{k=1}^N G_{kk}(z), \]

we define

\[ m^{(i)}(z) := \frac{1}{N-1} \text{Tr} G^{(i)}(z) = \frac{1}{N-1} \sum_{k \neq i} G_{kk}^{(i)}(z), \]

and we have the following lemma to compare the traces of \( G \) and \( G^{(i)} \).

**Lemma 2.3.** For any \( 1 \leq i \leq N \),

\[ |m(z) - m^{(i)}(z)| \leq \frac{C}{N \eta}, \quad \eta = \text{Im} z > 0. \quad (2.14) \]

**Proof.** Let

\[ F(x) := \frac{1}{N} \# \{ \lambda_j \leq x \}, \quad F^{(i)}(x) := \frac{1}{N-1} \# \{ \mu_j \leq x \} \]

de note the normalized counting functions of the eigenvalues. The interlacing property of the eigenvalues of \( H \) and \( H^{(i)} \) (see (1.89)) in terms of these functions means that

\[ \sup_x |NF(x) - (N-1)F^{(i)}(x)| \leq 1. \]

Then after integrating by parts,

\[ \left| m(z) - \left( 1 - \frac{1}{N} \right) m^{(i)}(z) \right| = \left| \int \frac{dF(x)}{x-z} - \left( 1 - \frac{1}{N} \right) \int \frac{dF^{(i)}(x)}{x-z} \right| = \frac{1}{N} \left| \int \frac{NF(x) - (N-1)F^{(i)}(x)}{(x-z)^2} \, dx \right| \leq \frac{1}{N} \int \frac{dx}{|x-z|^2} \leq \frac{C}{N \eta}, \quad (2.15) \]

and this, together with the trivial bound \(|m^{(i)}| \leq \eta^{-1}\), proves (2.14).

Returning to (2.13), we thus have

\[ G_{ii} = \frac{1}{-z - m(z) + \Omega_i}, \quad (2.16) \]

where

\[ \Omega_i := h_{ii} - Z_i + O\left( \frac{1}{N \eta} \right). \quad (2.17) \]
Summing up (2.16), we get that

\[ m = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{-z - m(z) + \Omega_i}. \]  

(2.18)

Suppose that \( \Omega := \max_i |\Omega_i| \) is small and \( |z + m| \geq C > 0 \), so that we can write

\[ \frac{1}{-z - m(z) + \Omega_i} = \frac{1}{-z - m(z)} + O(\Omega). \]

Then we have the following self-consistent equation for \( m \):

\[ m + \frac{1}{z + m} = O(\Omega). \]  

(2.19)

We recall that the Stieltjes transform of the semicircle law

\[ m_{sc}(z) = \int_{\mathbb{R}} \frac{\rho_{sc}(x) \, dx}{x - z} \]

can be characterized as the only solution to the quadratic equation

\[ m_{sc}(z) + \frac{1}{z + m_{sc}(z)} = 0 \]  

(2.20)

with \( \text{Im} \, m_{sc}(z) > 0 \) for \( \text{Im} \, z > 0 \). We can thus use the stability of the equation (2.20) to determine the solution \( m \) to (2.19). The stability deteriorates near the spectral edges \( z \sim \pm 2 \), and we have the following precise result, which can be proved by elementary calculus.

**Lemma 2.4** ([6], Lemma 8.4). Fix \( z = E + i\eta, \ \eta > 0, \) and let \( \kappa := |E| - 2 \).

Suppose that \( m = m(z) \) has a positive imaginary part and

\[ \left| m + \frac{1}{z + m} \right| \leq \delta. \]

Then

\[ |m - m_{sc}(z)| \leq \frac{C \delta}{\sqrt{\kappa + \delta}}. \]

Applying this result, we get that

\[ |m(z) - m_{sc}(z)| \leq \frac{C \Omega}{\sqrt{\kappa + \Omega}}. \]  

(2.21)

We now give a rough bound on the size of \( \Omega \). Clearly, \( |h_{ii}| \lesssim N^{-1/2} \). If the single-entry distribution has subexponential decay, then we can guarantee that all diagonal elements simultaneously satisfy essentially this estimate with a very high probability. Recall that (1.67) implies that

\[ P(|h_{ii}| \geq M^\alpha N^{-1/2}) \leq Ce^{-M} \]
for each $i$. Choosing $M = (\log N)^{1+\varepsilon}$, we have
\[ P(\exists i : |h_{ii}| \geq (\log N)^{(1+\varepsilon)\alpha} N^{-1/2}) \leq CN e^{-\left(\log N\right)^{1+\varepsilon}} \leq C e^{-\left(\log N\right)^{1+\varepsilon}}, \] (2.22)
which is faster than any polynomial decay.

To estimate $Z_i$, we compute its second moment:
\[ E|Z_i|^2 = \sum_{k,l \neq i} \sum_{k',l' \neq i} E_i([\bar{h}_{ik} G_{kl}^{(i)} h_{il} - E_i \bar{h}_{ik} G_{kl}^{(i)} h_{il}] [\bar{h}_{ik'} G_{kl'}^{(i)} \bar{h}_{il'} - E_i \bar{h}_{ik'} G_{kl'}^{(i)} \bar{h}_{il'}]). \] (2.23)

Since $E\bar{h} = 0$, the non-zero contributions to this sum come from index combinations when all the $h$ and $\bar{h}$ are paired. Assume for simplicity that $E\bar{h}^2 = 0$, which can always be achieved, for example, if the distributions of the real and imaginary parts are the same. Then the $h$ factors in the above expression have to be paired in such a way that $h_{ik} = h_{ik'}$ and $h_{il} = h_{il'}$, that is, $k = k'$ and $l = l'$. We note that the pairing $h_{ik} = h_{il}$ will give zero, because the expectation is subtracted. The result is
\[ E|Z_i|^2 = \frac{1}{N^2} \sum_{k,l \neq i} |G_{kl}^{(i)}|^2 + \frac{m_4 - 1}{N^2} \sum_{k \neq i} |G_{kk}^{(i)}|^2, \] (2.24)
where $m_4 = E|\sqrt{N} h|^4$ is the fourth moment of the single-entry distribution. The first term can be computed as follows:
\[ \frac{1}{N^2} \sum_{k,l \neq i} |G_{kl}^{(i)}|^2 = \frac{1}{N^2} \sum_{k \neq i} (|G^{(i)}|^2)_{kk} = \frac{1}{N\eta} \sum_{k} \text{Im} G_{kk}^{(i)} = \frac{1}{N\eta} \text{Im} m^{(i)}, \] (2.25)
where $|G|^2 = GG^*$ and we used the identity
\[ |G|^2 = \frac{1}{|H - E|^2 + \eta^2} = \frac{1}{\eta} \text{Im} G. \]

To estimate this quantity, we need an upper bound on the local density on the scale $\eta \gg 1/N$. For any interval $I \subset \mathbb{R}$, let $\mathcal{N}_I := \#\{\lambda_j \in I\}$ denote the number of eigenvalues in $I$. Lemma 1.10 in §1.7 shows that $\mathcal{N}_I \lesssim N|I|$ with a very high probability. Using this lemma for the matrix $H^{(i)}$ with eigenvalues $\mu_1, \mu_2, \ldots, \mu_{N-1}$, we have
\[ \text{Im} m^{(i)} = \frac{1}{N - 1} \sum_{\alpha} \frac{\eta}{(E - \mu_\alpha)^2 + \eta^2} \lesssim \int \frac{\eta}{(x - E)^2 + \eta^2} q_{sc}(x) \, dx \lesssim O(1). \] (2.26)
Recall that the symbol $\lesssim$ here means up to factors of order $\log N$.

We can estimate the second term in (2.24) by using the trivial bound $|G_{kk}^{(i)}| \leq \eta^{-1}$, and thus
\[ \frac{1}{N^2} \sum_{k \neq i} |G_{kk}^{(i)}|^2 \leq \frac{1}{N^2 \eta} \sum_{k \neq i} |G_{kk}^{(i)}| \leq \frac{1}{N^2 \eta} \sum_{k \neq i} \sum_{\alpha=1}^{N-1} \frac{|u_\alpha(k)|^2}{|\lambda_\alpha - z|} \leq \frac{1}{N \eta} \frac{1}{N} \sum_{\alpha=1}^{N-1} \frac{1}{|\lambda_\alpha - z|} \lesssim \frac{1}{N \eta}, \] (2.27)
where $u_\alpha$ is the (normalized) eigenvector for $\mu_\alpha$, and in the last step we used an estimate similar to (2.26).

The estimates (2.25) and (2.27) confirm that the size of $Z_i$ is roughly

$$|Z_i| \lesssim \frac{1}{\sqrt{N\eta}},$$

at least in the second-moment sense. One can compute higher moments or use even stronger concentration results (if, for example, the logarithmic Sobolev inequality is available) to strengthen (2.28) so that it holds in the sense of probability.

All together, (2.17), (2.22), and (2.28) give

$$\Omega \lesssim C\sqrt{N} + C\sqrt{N\eta} + C N\eta.$$ 

Since we are interested in $1/N \ll \eta \leq 1$, we get that $\Omega \lesssim (N\eta)^{-1/2}$. Combining this with the stability bound, (2.21), we have

$$|m(z) - m_{sc}(z)| \lesssim \min \left\{ \frac{1}{\sqrt{N\eta\kappa}}, \frac{1}{(N\eta)^{1/4}} \right\},$$

which proves Lemma 2.3.

The remarkable feature is that this method works down to scales $\eta \sim 1/N$. The factor $\kappa$ expresses the fact that the estimate deteriorates near the edge. The exponents are not optimal, and the difference $m - m_{sc}$ can be compared with a precision of order $(N\eta)^{-1}$. This will be discussed in the next subsection. The gain will come from the fact that the main error term $Z_i$ in (2.17) is fluctuating, one can show [9] that its contributions to $m - m_{sc}$ cancel to leading order, and it is smaller than the size of $Z_i$ predicted by the variance calculation (this effect was first exploited in Theorem 4.1 of [89], and substantially improved in [9] and [8]).

We emphasize that our presentation was very sketchy, with many technical issues neglected.

### 2.3. Semicircle law via resolvents: Refined method.

#### 2.3.1. Statement of the theorem and consequences.

Here we present a more refined method that can estimate matrix elements of the resolvent and also applies to universal Wigner matrices (Definition 1.1). This method is a key ingredient for the improved precision of the semicircle law in terms of both $(N\eta)$-power and edge behaviour. The main ingredient is the analysis of a self-consistent equation for the vector $(G_{11}, \ldots, G_{NN})$ of diagonal elements of the resolvent and not just of their sum, which led to (2.18). Again, for definiteness we formulate the result only for generalized Hermitian Wigner matrices. The extension to symmetric matrices is straightforward.

A key quantity will be the matrix $\Sigma$ of variances introduced in (1.15) Recall that $\Sigma$ is symmetric, doubly stochastic by (1.16), and in particular satisfies $-1 \leq \Sigma \leq 1$. Let the spectrum of $\Sigma$ satisfy

$$\text{Spec}(\Sigma) \subset [-1 + \delta_-, 1 - \delta_+] \cup \{1\},$$

(2.29)
with some non-negative constants $\delta_{\pm}$. We will always have the following spectral assumption:

$$1 \text{ is a simple eigenvalue of } \Sigma \text{ and } \delta_- \text{ is a positive constant independent of } N. \quad (2.30)$$

For Wigner matrices, all entries of $\Sigma$ are identical and $\delta_{\pm} = 1$. It is easy to prove (see Lemma A.1 of [9]) that (2.30) holds for random band matrices (see (1.18) for the definition), with $\delta_- > 0$ depending only on $f$. For generalized Wigner matrices, that is, Wigner matrices with comparable variances (see (1.17) for the definition), it is easy to check that

$$\delta_{\pm} \geq C_{\inf} > 0.$$ 

The fact that $\delta_{\pm}$ is $> 0$ for generalized Wigner matrices allows better control in terms of the edge behaviour of the estimates. This is the main reason why the statement below is different for universal Wigner matrices (see (1.1)) and generalized Wigner matrices, (1.17).

The precision of the local semicircle law depends on three factors. The first factor is the resolution (the scale on which the semicircle law holds), which is given by the imaginary part $\eta = \text{Im } z$ of the spectral parameter $z$ in the Stieltjes transform. The second factor is the distance to the spectral edge, measured by $\kappa = |E| - 2$. The last factor is the size of the typical matrix elements, measured by the quantity

$$M := \frac{1}{\max_{ij} \sigma_{ij}^2}, \quad (2.31)$$

called the spread of the matrix. For example, $M = N$ for Wigner matrices, $M \sim N$ for generalized Wigner matrices with (1.1), and $M \sim W$ for random band matrices (1.18).

**Theorem 2.5** (local semicircle law for universal Wigner matrices [8], Theorem 2.1). Let $H$ be a Hermitian $N \times N$ random matrix with $Eh_{ij} = 0$, $1 \leq i, j \leq N$, and assume that the variances $\sigma_{ij}^2$ satisfy (1.16) and (2.30). Suppose that the distributions of the matrix elements have a uniformly subexponential decay in the sense that there exist constants $C, \vartheta > 0$ independent of $N$ such that for any $x > 0$ and for each $(i, j)$

$$\mathbb{P}(|h_{ij}| \geq x|\sigma_{ij}|) \leq C \exp(-x^{\vartheta}). \quad (2.32)$$

The universal Wigner matrices and the special subclass of them consisting of the generalized Wigner matrices will be considered in parallel, with the two classes distinguished by a parameter $A$: $A = 2$ for universal Wigner matrices, and $A = 1$ for generalized Wigner matrices, where the results will be stronger.

Define the following domain in $\mathbb{C}$:

$$D := \{z = E + i\eta \in \mathbb{C} : |E| \leq 5, \ 0 < \eta < 10, \ \sqrt{M\eta} \geq (\log N)^{C_1}(\kappa + \eta)^{1/4-A}\}, \quad (2.33)$$

where $\kappa := ||E| - 2|$. Then there exist constants $C_1, C_2, \text{ and } c > 0$ depending only on $\vartheta$ and the number $\delta_-$ in (2.30) such that for any $\varepsilon > 0$ and $K > 0$ there
is a constant $C(\varepsilon, K)$ for which the Stieltjes transform of the empirical eigenvalue distribution of $H$ satisfies

$$
P\left( \bigcup_{z \in D} \left\{ \left| m(z) - m_{sc}(z) \right| \geq \frac{N^\varepsilon}{M\eta(\kappa + \eta)^A} \right\} \right) \leq \frac{C(\varepsilon, K)}{N^K} \tag{2.34}$$

for sufficiently large $N$, the diagonal matrix elements $G_{ii}(z) = (H - z)^{-1}(i, i)$ of the Green function satisfy

$$
P\left( \bigcup_{z \in D} \left\{ \max_i \left| G_{ii}(z) - m_{sc}(z) \right| \geq \frac{(\log N)^{C_2}}{\sqrt{M\eta}} \left( \kappa + \eta \right)^{1/4 - A/2} \right\} \right) \leq CN^{-c(\log \log N)}, \tag{2.35}$$

and the off-diagonal elements satisfy

$$
P\left( \bigcup_{z \in D} \left\{ \max_{i \neq j} \left| G_{ij}(z) \right| \geq \frac{(\log N)^{C_2}}{\sqrt{M\eta}} \left( \kappa + \eta \right)^{1/4} \right\} \right) \leq CN^{-c(\log \log N)} \tag{2.36}$$

for any sufficiently large $N$.

**Remark 1.** These estimates are optimal with respect to the power of $M\eta$, but they are not optimal as far as the edge behaviour (power of $\kappa$) and the control on the probability are concerned. Under stronger decay assumptions on the single-entry distributions it is possible to get subexponential bounds on the probability, for example, Theorem 1.11 (see Theorem 3.1 in [57]). On the other hand, the subexponential decay condition (2.32) can easily be weakened if we are not aiming at error estimates faster than any power law of $N$.

**Remark 2.** Concerning the edge behaviour, we remark that in our first two papers [55], [56] we simply assumed that $\kappa \geq \kappa_0$ for some positive constant $\kappa_0$. The edge behaviour was effectively treated first in [57] and substantially improved in Theorem 4.1 of [89], but the bounds were not optimal. The best result for universal Wigner matrices is Theorem 2.5. For generalized Wigner matrices, Theorem 2.19 (proved in [7]) gives an optimal estimate uniform with respect to $\kappa$.

The local semicircle estimates imply that the empirical counting function of the eigenvalues is close to the semicircle counting function (Theorem 2.6) and that the locations of the eigenvalues are close to their classical locations in the sense of mean square deviation (Theorem 2.7). This latter result will be used to verify (1.66), or more precisely, Condition III (3.9) below, which will be the key input to our hydrodynamical approach to universality.

To formulate these statements precisely, let $\lambda_1 \leq \cdots \leq \lambda_N$ be the ordered eigenvalues of a universal Wigner matrix. We define the **normalized empirical counting function** by

$$
n(E) := \frac{1}{N} \# \{ \lambda_j \leq E \} \tag{2.37}$$

and the **averaged counting function** by

$$
n(E) = \frac{1}{N} \mathbb{E} \# \{ \lambda_j \leq E \}. \tag{2.38}$$
Finally, let
\[ n_{sc}(E) := \int_{-\infty}^{E} \varrho_{sc}(x) \, dx \]  
(2.39)

be the distribution function of the semicircle law, which is very close to the counting function of the \( \gamma_j, \frac{1}{N} \# [\gamma_j \leq E] \approx n_{sc}(E) \). Recall that the \( \gamma_j \) are the classical locations of the eigenvalues, determined by the semicircle law (see (1.65)).

With this notation we have the two theorems below.

**Theorem 2.6** ([8], Theorem 6.3). Let \( A = 2 \) for universal Wigner matrices satisfying (1.16), (2.31), and (2.32) with \( M \geq (\log N)^{24+6\alpha} \). For generalized Wigner matrices, satisfying (1.16), (2.31), (1.17), and (2.32), let \( A = 1 \) and recall that \( M = N \) in this case. Then for any \( \varepsilon > 0 \) and \( K \geq 1 \) there exists a constant \( C(\varepsilon, K) \) such that
\[
P\left\{ \sup_{|E| \leq 3} \left| n(E) - n_{sc}(E) \right| [\kappa_E]^A \leq \frac{CN^\varepsilon}{M} \right\} \geq 1 - \frac{C(\varepsilon, K)}{NK},
\]
where the functions \( n(E) \) and \( n_{sc}(E) \) were defined in (2.37) and (2.39), and \( \kappa_E = |E| - 2 \).

**Theorem 2.7** ([8], Theorem 7.1). Let \( H \) be a generalized Wigner matrix with subexponential decay, that is, assume that (1.16), (2.31), (1.17), and (2.32) hold. Let \( \lambda_j \) denote the eigenvalues of \( H \) and \( \gamma_j \) their semiclassical locations, defined by (1.65). Then for any \( \varepsilon < 1/7 \) and any \( K > 1 \) there exists a constant \( C_K \) such that
\[
P\left\{ \sum_{j=1}^{N} |\lambda_j - \gamma_j|^2 \leq N^{-\varepsilon} \right\} \geq 1 - \frac{C_K}{NK}
\]
and
\[
\sum_{j=1}^{N} E|\lambda_j - \gamma_j|^2 \leq CN^{-\varepsilon}.
\]

These theorems are consequences of the local semicircle law, Theorem 2.5. We will not give the detailed proofs, but we mention a useful formula that allows one to translate Stieltjes transforms to densities. In this context the formula first appeared in [89].

**Lemma 2.8** ([102], Helffer–Sjöstrand formula). Let \( f \) be a real-valued \( C^1 \) function on \( \mathbb{R} \). Let \( \chi(y) \) be a smooth cutoff function with support in \([-1,1]\), with \( \chi(y) = 1 \) for \( |y| \leq 1/2 \), and with bounded derivatives. Let
\[
\tilde{f}(x + iy) := (f(x) + iyf'(x))\chi(y).
\]

Then
\[
f(\lambda) = \frac{1}{2\pi} \int_{\mathbb{R}^2} \frac{\partial_{\bar{z}} \tilde{f}(x + iy)}{\lambda - x - iy} \, dx \, dy
\]
\[
= \frac{1}{2\pi} \int_{\mathbb{R}^2} \frac{iyf''(x)\chi(y) + i(f(x) + iyf'(x))\chi'(y)}{\lambda - x - iy} \, dx \, dy.
\]
(2.42)
We will apply this lemma in the following form. Let \( \varrho \in L^1(\mathbb{R}) \) be a real function and let \( m(z) \) be its Stieltjes transform

\[
m(z) := \int_{\mathbb{R}} \frac{\varrho(\lambda)}{\lambda - z} d\lambda.
\]

Since \( f \) is real, we have

\[
\left| \int_{\mathbb{R}} f(\lambda) \varrho(d\lambda) \right| = \left| \text{Re} \int_{\mathbb{R}} f(\lambda) \varrho(d\lambda) \right| = \frac{1}{2\pi} \text{Re} \int_{\mathbb{R}^2} \partial_x f(x + iy) m(x + iy) dx dy \leq \frac{1}{2\pi} \int_{\mathbb{R}^2} y f''(x) \chi(y) \text{Im} m(x + iy) dx dy + C \int_{\mathbb{R}^2} |f(x)| + |y||f'(x)||\chi'(y)||m(x + iy)| dx dy. \tag{2.43}
\]

In order to get the counting function, we will choose

\[
f(\lambda) = f_{E,\eta}(\lambda),
\]

where \( f_{E,\eta} \) is the characteristic function of the semi-axis \((-\infty, E)\), smoothed out on a scale \( \eta \) (that is, \( f_{E,\eta}(\lambda) \equiv 1 \) for \( \lambda \leq E - \eta \), \( f_{E,\eta}(\lambda) \equiv 0 \) for \( \lambda \geq E + \eta \), and \( |f'| \leq C\eta^{-1} \) and \( |f''| \leq C\eta^{-2} \) in the interval \( E - \eta, E + \eta \)).

The second term in (2.43) is typically harmless, since \( \chi' \) is supported where \( |y| \geq 1/2 \), that is, this term requires information about the Stieltjes transform far away from the real axis. In the first term, we have only the imaginary part of the Stieltjes transform, and it is easy to see that

\[
y \text{Im} m(x + iy) \leq \int |\varrho(x)| dx.
\]

One can then perform an integration by parts, bringing the second derivative of \( f \) down to the first derivative, which is under control after the integration even if \( \eta \) is very small.

Applying these ideas to the measure \( \varrho \) that is the difference between the empirical density and \( \varrho_{sc} \), one can control \( \int f(\lambda) \varrho(d\lambda) \) (that is, essentially the difference between the counting functions) in terms of the size of \( m - m_{sc} \). For details, see [89] and [8].

2.3.2. Sketch of the proof of the semicircle law for matrix elements. To simplify the exposition we will ignore the edge problem, that is, we will assume that the function \( E = \text{Re} z \) always satisfies \( \kappa = |E| - 2 \geq \kappa_0 \) for some fixed \( \kappa_0 > 0 \) and we do not follow the dependence of the constants on \( \kappa_0 \). We will thus prove the following partial version of Theorem 2.5.

**Theorem 2.9.** Assume the conditions of Theorem 2.5. For some \( \kappa_0 > 0 \), define the following domain in \( \mathbb{C} \):

\[
D := D_{\kappa_0} = \{ z = E + i\eta \in \mathbb{C} : |\eta| \leq Q, \ |E| - 2 \geq \kappa_0, \ \sqrt{M\eta} \geq (\log N)^{C_1}, \ \}
\]

(2.44)


where \( C_1 \) and \( Q \) are sufficiently large. Then there exist constants \( C_2, C, \) and \( c > 0 \) depending only on \( \vartheta, \kappa_0, \) and the \( \delta_- \) in (2.30) such that the diagonal matrix elements \( G_{ii}(z) \) of the Green function satisfy
\[
P\left( \bigcup_{z \in D} \left\{ \max_i |G_{ii}(z) - m_{sc}(z)| \geq \frac{(\log N)^{C_2}}{\sqrt{M\eta}} \right\} \right) \leq CN^{-c \log \log N},
\]
and the off-diagonal elements satisfy
\[
P\left( \bigcup_{z \in D} \left\{ \max_{i \neq j} |G_{ij}(z)| \geq \frac{(\log N)^{C_2}}{\sqrt{M\eta}} \right\} \right) \leq CN^{-c \log \log N}
\]
for sufficiently large \( N \).

We start with a system of self-consistent equations for the diagonal matrix elements of the resolvent. The following lemma is a simple combination of the resolvent identities in § 2.1.

**Lemma 2.10.** The diagonal matrix elements \( G_{ii} = (H - z)^{-1}(i, i) \) of the resolvent satisfy the following system of self-consistent equations:
\[
G_{ii} = \frac{1}{-z - \sum_j \sigma_{ij}^2 G_{jj} + \Upsilon_i},
\]
where
\[
\Upsilon_j := A_i + h_{ii} - Z_i,
\]
with
\[
A_i := \sigma_{ii}^2 G_{ii} + \sum_{j \neq i} \sigma_{ij}^2 \frac{G_{ij} G_{ji}}{G_{ii}},
\]
\[
Z_i := Z_{ii}^{(i)} - E_{i} Z_{ii}^{(i)}, \quad Z_{ii}^{(i)} := a^i \cdot G^{(i)} a^i = \sum_{k,l \neq i} a_k^i G_{ki}^{(i)} a_l^i.
\]

**Proof.** We introduce the notation
\[
Z_{ij}^{(ij)} := a^i \cdot G^{(ij)} a^j, \quad Z_{ii}^{(i)} := a^i \cdot G^{(i)} a^i.
\]
Then from (2.6)
\[
K_{ij}^{(ij)} = h_{ij} - z_{ij} - Z_{ij}^{(ij)}, \quad K_{ii}^{(i)} = h_{ii} - z - Z_{ii}^{(i)}.
\]
We can write \( G_{ii} \) as follows (see (2.9)):
\[
G_{ii} = (K_{ii}^{(i)})^{-1} = \frac{1}{E_{a^i} K_{ii}^{(i)} + K_{ii}^{(i)} - E_{a^i} K_{ii}^{(i)}},
\]
where \( E_{a^i} = E_i \) denotes the expectation with respect to the elements in the \( i \)th column of the matrix \( H \).
Using the fact that \( G(i) = (H(i) - z)^{-1} \) is independent of \( a_i \) and \( E_a a_k^a a_l^a = \delta_{kl}\sigma_{ik}^2 \), we get that

\[
E_a a_i K(i)_{ii} = -z - \sum_{j \neq i} \sigma_{ij}^2 G(jj)_{jj}
\]

and

\[
K(i)_{ii} - E_a a_i K(i)_{ii} = h_{ii} - Z_i. \tag{2.52}
\]

Using the equality

\[
G_{kl} = G(i)_{kl} + \frac{G_{ki} G_{il}}{G_{ii}}
\]

in (2.10) and the notation in (2.49), we get the expression

\[
E_a a_i K(i)_{ii} = -z - \sum_{j \neq i} \sigma_{ij}^2 G(jj)_{jj} = -z - \sum_{j \neq i} \sigma_{ij}^2 G_{jj} + \sum_{j} \sigma_{ij}^2 \frac{G_{ji} G_{ij}}{G_{ii}} = -z - \sum_{j} \sigma_{ij}^2 G_{jj} + A_i.
\]

Combining this with (2.52), we eventually obtain (2.47) from (2.51).

We introduce the notation

\[
v_i := G_{ii} - m_{sc}, \quad m := \frac{1}{N} \sum_i G_{ii}, \quad [v] := \frac{1}{N} \sum_i v_i = \frac{1}{N} \sum_i (G_{ii} - m_{sc}).
\]

We will estimate the following key quantities:

\[
\Lambda_d := \max_k |v_k| = \max_k |G_{kk} - m_{sc}|, \quad \Lambda_o := \max_{k \neq \ell} |G_{k\ell}|, \tag{2.53}
\]

where the subscripts refer to ‘diagonal’ and ‘off-diagonal’ matrix elements. All the quantities defined so far depend on the spectral parameter \( z = E + i\eta \), but we will mostly omit this from the notation. The real part \( E \) will always be kept fixed. For the imaginary part we will use a continuity argument at the end of the proof, and then the dependence of \( \Lambda_d \) and \( \Lambda_o \) on \( z \) will be indicated.

Both \( \Lambda_d \) and \( \Lambda_o \) will be typically small for \( z \in D \). Eventually we will prove that their size is less than \( (M\eta)^{-1/2} \), modulo logarithmic corrections. We thus define the exceptional event

\[
\Omega_\Lambda = \Omega_\Lambda(z) := \{ \Lambda_d(z) + \Lambda_o(z) \geq (\log N)^{-C} \}
\]

with some \( C \) (in this presentation, we will not care about matching all the exponents). We will always work in \( \Omega_\Lambda^c \), and in particular, we will have

\[
\Lambda_d(z) + \Lambda_o(z) \ll 1.
\]

It is easy to check from the explicit formula for \( m_{sc} \) that

\[
c \leq |m_{sc}(z)| \leq C, \quad z \in D, \tag{2.54}
\]

with some positive constants, so from the relation \( G_{ii} = m_{sc}(z) + O(\Lambda_d) \) we have

\[
c \leq |G_{ii}(z)| \leq C, \quad z \in D. \tag{2.55}
\]
The equality (2.11), namely,

$$G_{kl}^{(i)} = G_{kl} - \frac{G_{ki}G_{il}}{G_{ii}}, \quad i \neq l, k,$$

together with (2.55) implies that for any $i$ and $z \in D$ the following inequalities hold:

$$\max_{k \neq l} |G_{kl}^{(i)}| \leq \Lambda_0 + C \Lambda_0^2 \leq C \Lambda_0 \quad \text{in } \Omega^c_A,$$

$$c \leq |G_{kk}^{(i)}| \leq C \quad \text{for all } k \neq i \text{ and in } \Omega^c_A,$$

$$|G_{kk}^{(i)} - m_{sc}| \leq \Lambda_d + C \Lambda_0^2 \quad \text{for all } k \neq i \text{ and in } \Omega^c_A,$$

and (see (2.49))

$$|A_i| \leq \frac{C}{M} + C \Lambda_0^2 \quad \text{in } \Omega^c_A.$$

Similarly, with one more expansion step and still for $z \in D$, we get that

$$\max_{i,j} \max_{k \neq i} |G_{kl}^{(ij)}| \leq C \Lambda_0, \quad \max_{i,j} \max_{k} |G_{kk}^{(ij)}| \leq C \quad \text{in } \Omega^c_A$$

and

$$|G_{kk}^{(ij)} - m_{sc}| \leq \Lambda_d + C \Lambda_0^2 \quad \text{for all } k \neq i, j \text{ and in } \Omega^c_A.$$

In view of these estimates, the following lemma shows that $Z_i$ and $Z_{ij}^{(ij)}$ are small if $\Lambda_d + \Lambda_0$ is small and the $h_{ij}$ are not too large. These bounds hold uniformly with respect to $D$.

**Lemma 2.11.** Define the exceptional events

$$\Omega_1 := \left\{ \max_{1 \leq i, j \leq N} |h_{ij}| \geq (\log N)^C |\sigma_{ij}| \right\},$$

$$\Omega_d(z) := \left\{ \max_i |Z_i(z)| \geq \frac{(\log N)^C}{\sqrt{M_\eta}} \right\},$$

$$\Omega_o(z) := \left\{ \max_{i \neq j} |Z_{ij}^{(ij)}(z)| \geq \frac{(\log N)^C}{\sqrt{M_\eta}} \right\},$$

and consider the set

$$\Omega := \Omega_1 \cup \bigcup_{z \in D} \left[ (\Omega_d(z) \cup \Omega_o(z)) \cap \Omega^c_A(z) \right]$$

of all exceptional events. Then

$$P(\Omega) \leq CN^{-c \log \log N}. \quad (2.62)$$

**Proof.** Under the assumption of (2.32), in analogy to (2.22) we have

$$P(\Omega_1) \leq CN^{-c \log \log N},$$

$$P(\Omega_d) \leq CN^{-c \log \log N},$$

and

$$P(\Omega_o) \leq CN^{-c \log \log N}.$$
so we can work in the complement set $\Omega^c_\Lambda$ and assume that

$$\max_{i,j} |h_{ij}| \leq \frac{(\log N)^C}{\sqrt{M}}. \quad (2.64)$$

We now prove that for any fixed $z \in D$

$$P\left( \Omega^c_\Lambda(z) \cap \left\{ \max_i |Z_i(z)| \geq \frac{(\log N)^C}{\sqrt{M\eta}} \right\} \right) \leq CN^{-c\log\log N}, \quad (2.65)$$

$$P\left( \Omega^c_\Lambda(z) \cap \left\{ \max_{i\neq j} |Z_{ij}(z)| \geq \frac{(\log N)^C}{\sqrt{M\eta}} \right\} \right) \leq CN^{-c\log\log N}. \quad (2.66)$$

We recall that $Z_{ij}$ is a quadratic form in the components of the random vectors $a^i$ and $a^j$. For such functions we have the following large deviation result. The proof is based on the Burkholder martingale inequality and will be given in Appendix A.

**Lemma 2.12** ([9], Lemmas B.1, B.2). Let $a_i$ ($1 \leq i \leq N$) be $N$ independent random complex variables with mean zero, variance $\sigma^2$, and uniform subexponential decay

$$P(|a_i| \geq x^\alpha) \leq Ce^{-x} \quad (2.67)$$

for some positive $\alpha$ and all $x$. Let $A_i$, $B_{ij} \in \mathbb{C}$ ($1 \leq i, j \leq N$). Then

$$P\left\{ \left| \sum_{i=1}^N a_i A_i \right| \geq (\log N)^{3/2+\alpha} \sigma \left( \sum_i |A_i|^2 \right)^{1/2} \right\} \leq CN^{-\log\log N}, \quad (2.68)$$

$$P\left\{ \left| \sum_{i=1}^N \bar{a}_i B_{ii} a_i - \sum_{i=1}^N \sigma^2 B_{ii} \right| \geq (\log N)^{3/2+2\alpha} \sigma^2 \left( \sum_{i=1}^N |B_{ii}|^2 \right)^{1/2} \right\} \leq CN^{-\log\log N}, \quad (2.69)$$

$$P\left\{ \left| \sum_{i \neq j} \bar{a}_i B_{ij} a_j \right| \geq (\log N)^{3+2\alpha} \sigma^2 \left( \sum_{i \neq j} |B_{ij}|^2 \right)^{1/2} \right\} \leq CN^{-\log\log N}. \quad (2.70)$$

To see (2.65), we apply the estimate (2.69) and get that

$$|Z_i| \leq (\log N)^C \left( \sum_{k,l \neq i} |\sigma_{ik} G_{kl}^{(i)} \sigma_{li}|^2 \right)^{1/2} \quad (2.71)$$

holds with a probability larger than $1 - CN^{-c\log\log N}$ for sufficiently large $N$.

Denote by $u^{(i)}_\alpha$ and $\lambda^{(i)}_\alpha$ ($\alpha = 1, \ldots, N - 1$) the eigenvectors and eigenvalues of $H^{(i)}$. Let $u^{(i)}_\alpha(k)$ denote the $k$th coordinate of $u^{(i)}_\alpha$. Then using the inequalities $\sigma^2_{\alpha} \leq 1/M$, (2.57), and (2.56), we have

$$\sum_{k,l \neq i} |\sigma_{ik} G_{kl}^{(i)} \sigma_{li}|^2 \leq \frac{1}{M} \sum_{k \neq i} \sigma^2_{ik} \sum_{l \neq i} |G_{kl}^{(i)}|^2 \leq \frac{1}{M} \sum_{k \neq i} \sigma^2_{ik} (|G^{(i)}|^2)_{kk}$$

$$= \frac{1}{M} \sum_{k \neq i} \sigma^2_{ik} \sum_{\alpha} |u^{(i)}_\alpha(k)|^2 \leq \frac{1}{M} \sum_{k \neq i} \sigma^2_{ik} \frac{\Im G^{(i)}_{kk}(z)}{\eta} \leq \frac{C}{M\eta} \quad \text{in } \Omega^c_\Lambda. \quad (2.72)$$
Here we defined $|A|^2 := A^*A$ for any matrix $A$. Together with (2.71) we have proved (2.65) for a fixed $z$. The off-diagonal estimate (2.66) (for $i \neq j$) is proved similarly, using (2.70) instead of (2.69).

We have thus proved that for each fixed $z \in D$ the sets $\Omega_d(z)$ and $\Omega_o(z)$ have a very small probability. In order to prove (2.62), we must in principle take the union of uncountably many events. But it is easy to see that the quantities $Z_i(z)$ and $Z^{(ij)}_{ij}(z)$ defining these events are Lipschitz continuous functions of $z$ with a Lipschitz constant $\eta^{-1} \leq N$. Thus, controlling them on a sufficiently dense but finite net of points will control them for every $z \in D$. The number of necessary points is only a power of $N$, while the probability bounds are smaller than any inverse power of $N$. This proves Lemma 2.11.

The key step in the proof of Theorem 2.9 is the following lemma, which says that if $\Lambda$ admits an upper bound of order $(\log N)^{-C}$, then the estimate can be boosted to show that $\Lambda$ is much smaller, of order $(M\eta)^{-1/2}$.

**Lemma 2.13.** Recall that $\Lambda_d$, $\Lambda_o$, and $\Omega$ are defined in (2.53) and (2.61), and that the set $D$ is defined in (2.44). Then for any $z \in D$ and in $\Omega^c$, if

$$\Lambda_o(z) + \Lambda_d(z) \leq (\log N)^{-C},$$  \hspace{1cm} (2.73)

then

$$\Lambda_o(z) + \Lambda_d(z) \leq \frac{(\log N)^C}{\sqrt{M\eta}}.$$ \hspace{1cm} (2.74)

**Proof of Lemma 2.13.** Choosing $C_1$ in (2.44) sufficiently large, we can ensure from Lemma 2.11 that $Z^{(ij)}_{ij} \ll 1$, $Z_i \ll 1$ in $\Omega^c$.

We first estimate the off-diagonal term $G_{ij}$. Under the condition (2.73), we get from (2.8), (2.55), and (2.56) that

$$|G_{ij}| = |G_{ii}| |G^{(ij)}_{jj}| |K^{(ij)}_{ij}| \leq C(|h_{ij}| + |Z^{(ij)}_{ij}|), \quad i \neq j.$$

By Lemma 2.11, under the condition (2.73) we have

$$|G_{ij}| \leq \frac{(\log N)^C}{\sqrt{M}} + \frac{(\log N)^C}{\sqrt{M\eta}} \leq \frac{(\log N)^C}{\sqrt{M\eta}} \ll 1 \text{ in } \Omega^c.$$ \hspace{1cm} (2.75)

This proves the estimate (2.74) for the summand $\Lambda_o$.

Now we estimate the diagonal terms. For the quantity $\Upsilon_i = A_i + h_{ii} - Z_i$ in (2.48) we get from (2.58), (2.75), and Lemma 2.11 that

$$\Upsilon := \max_i |\Upsilon_i| \leq C\frac{(\log N)^C}{\sqrt{M}} + \frac{(\log N)^C}{\sqrt{M\eta}} \leq \frac{(\log N)^C}{\sqrt{M\eta}} \ll 1 \text{ in } \Omega^c$$ \hspace{1cm} (2.76)

(in the last step we used the facts that $z \in D$ and $C_1$ is large). From (2.47) we have the identity

$$v_i = G_{ii} - m_{sc} = \frac{1}{-z - m_{sc} - (\sum_j \sigma_{ij}^2 v_j - \Upsilon_i)} - m_{sc},$$ \hspace{1cm} (2.77)
Using the facts that \((m_{\text{sc}} + z) = -m_{\text{sc}}^{-1}, |m_{\text{sc}} + z| \geq 1,\) and \(\Lambda_d + \Upsilon \ll 1,\) we can expand (2.77) as

\[
v_i = m_{\text{sc}}^2 \left( \sum_j \sigma_{ij}^2 v_j - \Upsilon_i \right) + O\left( \sum_j \sigma_{ij}^2 v_j - \Upsilon_i \right)^2
= m_{\text{sc}}^2 \left( \sum_j \sigma_{ij}^2 v_j - \Upsilon_i \right) + O((\Lambda_d + \Upsilon)^2). \tag{2.78}
\]

Summing this formula over all \(i\) and recalling the definition \([v] := \frac{1}{N} \sum_i v_i = m - m_{\text{sc}},\) we find that

\[
[v] = m_{\text{sc}}^2 (z)[v] - \frac{m_{\text{sc}}^2(z)}{N} \sum_i \Upsilon_i + O((\Lambda_d + \Upsilon)^2).
\]

Introducing the notation \(\zeta := m_{\text{sc}}^2(z)\) and \([\Upsilon] := \frac{1}{N} \sum_i \Upsilon_i\) for simplicity (and using the inequalities \(\Lambda_d \leq 1\) and \(|\zeta| \leq 1),\) we have

\[
(1 - \zeta)[v] = -\zeta[\Upsilon] + O((\Lambda_d + \Upsilon)^2) = O(\Lambda_d^2 + \Upsilon). \tag{2.79}
\]

Recall that \(\Sigma\) denotes the matrix of covariances, \(\Sigma_{ij} = \sigma_{ij}^2,\) and that 1 is a simple eigenvalue of \(\Sigma\) with the constant vector \(e = N^{-1/2}(1, \ldots, 1)\) as eigenvector. Let \(Q := I - |e\rangle\langle e|\) be the projection onto the orthogonal complement of \(e,\) and note that \(\Sigma\) and \(Q\) commute. Let \(\| \cdot \|_{\infty \to \infty}\) denote the \(\ell^\infty \to \ell^\infty\) matrix norm. With this notation we can combine (2.79) and (2.78) to get that

\[
v_i - [v] = \zeta \sum_j \Sigma_{ij} (v_j - [v]) - \zeta (\Upsilon_i - [\Upsilon]) + O(\Lambda_d^2 + \Upsilon).
\]

When summing over all \(i,\) all three explicit terms sum up to zero, hence so do the error terms. Therefore, \(Q\) acts as the identity on the vector of error terms, and we have

\[
v_i - [v] = -\sum_j \left( \frac{\zeta}{1 - \zeta \Sigma} \right)_{ij} (\Upsilon_j - [\Upsilon]) + O\left( \left\| \frac{\zeta Q}{1 - \zeta \Sigma} \right\|_{\infty \to \infty} (\Lambda_d^2 + \Upsilon) \right)
= \left\| \frac{\zeta Q}{1 - \zeta \Sigma} \right\|_{\infty \to \infty} O(\Lambda_d^2 + \Upsilon). \tag{2.80}
\]

Combining (2.79) with (2.80), we have

\[
\Lambda_d = \max_i |v_i| \leq C\left( \left\| \frac{\zeta Q}{1 - \zeta \Sigma} \right\|_{\infty \to \infty} + \frac{1}{|1 - \zeta|} \right)(\Lambda_d^2 + \Upsilon). \tag{2.81}
\]

From this calculation it is clear how the estimates deteriorate at the edges. It is easy to check that

\[
1 - \zeta = 1 - m_{\text{sc}}^2(z) \sim \sqrt{\kappa + \eta}, \quad z = E + i\eta, \quad \kappa = |E| - 2.
\]

However, as we remarked at the beginning of the proof, we assume for simplicity that \(\kappa \geq \kappa_0 > 0,\) so we will not follow this deterioration.

To estimate the norm of the resolvent, we recall the following elementary lemma.
Lemma 2.14 ([9], Lemma 5.3). Let $\delta_- > 0$ be a given constant. Then there exist small real numbers $\tau > 0$ and $c_1 > 0$ depending only on $\delta_-$ such that for any positive number $\delta_+$

$$\max_{x \in [-1+\delta_-,1-\delta_+]} \{|\tau + x m_{sc}^2(z)|^2\} \leq (1 - c_1 q(z))(1 + \tau)^2,$$  

(2.82)

where

$$q(z) := \max\{\delta_+, |1 - \text{Re} m_{sc}^2(z)|\}.$$  

(2.83)

Lemma 2.15. Suppose that $\Sigma$ satisfies (2.29): $\text{Spec}(Q\Sigma) \subset [-1 + \delta_-, 1 - \delta_+]$. Then

$$Q 1 - m_{sc}^2(z)\Sigma \xrightarrow{\infty} \infty \leq C(\delta_-) \log N q(z),$$  

(2.84)

where $C(\delta_-)$ is a constant depending on $\delta_-$ and $q$ is defined in (2.83).

From this lemma one can see that the deterioration of the estimates at the edges due to the first term on the right-hand side of (2.81) can be offset by assuming that $\delta_+ > 0$.

Proof of Lemma 2.15. Let $\| \cdot \|$ denote the usual $\ell^2$ matrix norm and recall that $\zeta = m_{sc}^2(z)$. We write

$$\| Q 1 - \zeta \Sigma \| \leq \frac{1}{1 + \tau} \| Q 1 - (\zeta \Sigma + \tau)/(1 + \tau) \|$$

with $\tau$ given in (2.82). By (2.82),

$$\left\| \frac{\zeta \Sigma + \tau}{1 + \tau} Q \right\| \leq \sup_{x \in [-1+\delta_-,1-\delta_+]} \left| \frac{\zeta x + \tau}{1 + \tau} \right| \leq (1 - c_1 q(z))^{1/2}.$$

To estimate the $\ell^\infty \to \ell^\infty$ norm of this matrix, recall that $|\zeta| \leq 1$ and $\sum_j |\Sigma_{ij}| = \sum_j \sigma_{ij}^2 = 1$. Thus,

$$\left\| \frac{\zeta \Sigma + \tau}{1 + \tau} Q \right\|_{\infty \to \infty} = \max_i \sum_j \left| \left( \frac{\zeta \Sigma + \tau}{1 + \tau} \right)_{ij} \right|$$

$$\leq \frac{1}{1 + \tau} \max_i \sum_j |\zeta \Sigma_{ij} + \tau \delta_{ij}| \leq \frac{|\zeta| + \tau}{1 + \tau} \leq 1.$$  

To establish (2.84), we consider the following expansion up to an arbitrary threshold $n_0$:

$$\left\| \frac{1}{1 - (\zeta \Sigma + \tau)/(1 + \tau)} Q \right\|_{\infty \to \infty}$$

$$\leq \sum_{n < n_0} \left\| \frac{\zeta \Sigma + \tau}{1 + \tau} Q \right\|_{\infty \to \infty}^n + \sum_{n \geq n_0} \left\| \left( \frac{\zeta \Sigma + \tau}{1 + \tau} \right)^n Q \right\|_{\infty \to \infty}$$

$$\leq n_0 + \sqrt{N} \sum_{n \geq n_0} \left\| \left( \frac{\zeta \Sigma + \tau}{1 + \tau} \right)^n Q \right\| = n_0 + \sqrt{N} \sum_{n \geq n_0} (1 - c_1 q(z))^{n/2}$$

$$= n_0 + C \sqrt{N} \frac{(1 - c_1 q(z))^{n_0/2}}{q(z)} \leq C \log N q(z).$$
Choosing $n_0 = (C \log N)/q(z)$, we finish the proof of Lemma 2.15.

Let us now return to the proof of Lemma 2.13. Recall that we are in the set $\Omega^c$ and $\kappa \geq \kappa_0$, that is, the function $1 - \text{Re} m_{sc}^2(z)$ and thus also $q(z)$ are bounded away from zero. First, substituting (2.84) into (2.81), we get that

$$\Lambda_d \leq C(\Lambda_d^2 + \Upsilon) \log N.$$  

By the assumption (2.73), we have $C\Lambda_d \log N \leq 1/2$, so the quadratic term can be absorbed into the linear term on the right-hand side, and we have

$$\Lambda_d \leq C\Upsilon \log N.$$  

Using the estimate (2.76) of $\Upsilon$, we find that

$$\Lambda_d \leq \frac{(\log N)^C}{\sqrt{MN\eta}},$$

which, together with (2.75), completes the proof of (2.74) and thus the proof of Lemma 2.13.

**Proof of Theorem 2.9.** Introducing the functions

$$R(z) := (\log N)^{-C}, \quad S(z) := \frac{(\log N)^C}{\sqrt{MN\eta}},$$

we can reformulate Lemma 2.13 as saying that in the event $\Omega^c$ if $\Lambda_d(z) + \Lambda_o(z) \leq R(z)$ holds for some $z \in D$, then $\Lambda_d(z) + \Lambda_o(z) \leq S(z)$. By assumption (2.44) of Theorem 2.9, $S(z) < R(z)$ for any $z \in D$. Clearly, $\Lambda_d(z) + \Lambda_o(z) \leq 3/\eta \leq 3/10$ for $\text{Im} \ z = 10$. Using this information, one can mimic the proofs of Lemma 2.11 and Lemma 2.13 to get an a priori bound $\Lambda_d(z) + \Lambda_o(z) \leq R(z)$ for $\eta = 10$. Using the fact that $R(z)$, $S(z)$, $\Lambda_d(z)$, and $\Lambda_o(z)$ are continuous, and letting $z$ approach the real axis, we get by a continuity argument that $\Lambda_d(z) + \Lambda_o(z) \leq S(z)$ in $\Omega^c$, as long as the condition (2.44) is satisfied for $z$. This proves Theorem 2.9.

2.3.3. Sketch of the proof of the semicircle law for the Stieltjes transform. In this subsection we strengthen the estimate in Theorem 2.9 for the Stieltjes transform $m(z) = \frac{1}{N} \sum_i G_{ii}$. The key improvement is that $|m - m_{sc}|$ will be estimated with a precision $(M\eta)^{-1}$ while the quantity $|G_{ii} - m_{sc}|$ was controlled only with a precision $(M\eta)^{-1/2}$ (modulo logarithmic terms and terms expressing the deterioration of the estimate near the edges). In the following theorem we prove a partial version of the inequality (2.34) in Theorem 2.5.

**Theorem 2.16.** Assume the conditions of Theorem 2.5, let $\kappa_0 > 0$ be fixed, and recall the definition of the domain $D = D_{\kappa_0}$ in (2.44). Then for any $\varepsilon > 0$ and $K > 0$ there exists a constant $C = C(\varepsilon, K, \kappa_0)$ such that

$$P\left( \bigcup_{z \in D} \left\{ |m(z) - m_{sc}(z)| \geq \frac{N^\varepsilon}{M\eta} \right\} \right) \leq \frac{C(\varepsilon, K, \kappa_0)}{N^K}.$$  

(2.86)
Proof of Theorem 2.16. We mostly follow the proof of Theorem 2.9. Fix \( z \in D \). We can assume that we are working in the complement of the small probability events estimated in (2.45) and (2.46). In particular, the estimate (2.74) is available. As in (2.79) we have

\[
[v] = m - m_{sc} = -\frac{\zeta}{1 - \zeta} \frac{1}{N} \sum_i \Upsilon_i + O((\Lambda_d + \Upsilon)^2)
\]

holds with a very high probability, but now we keep the first term and estimate it better than the most trivial bound used in (2.79) to obtain some cancellation in the fluctuating sum. Then from (2.76) and (2.85) we have

\[
m - m_{sc} = O\left(\frac{1}{N} \sum_j \Upsilon_j\right) + O\left(\frac{N^\varepsilon}{M\eta}\right)
\]

with a very high probability for any small \( \varepsilon > 0 \). Recall that \( \Upsilon_i = A_i + h_{ii} - Z_i \). From (2.49), (2.55), and \( \sigma^2_{ij} \leq M^{-1} \) we find that

\[
A_j \leq \frac{C}{M} + C\Lambda_0^2 \leq \frac{CN^\varepsilon}{M\eta},
\]

where we used (2.74) to estimate \( \Lambda_0 \).

We thus get that

\[
m - m_{sc} = O\left(\frac{1}{N} \sum_i Z_i - \frac{1}{N} \sum_i h_{ii}\right) + O\left(\frac{N^\varepsilon}{M\eta}\right)
\]

with a very high probability. Since the \( h_{ii} \) are independent, by applying the first estimate in the large deviation Lemma 2.12 we get that

\[
P\left(\left|\frac{1}{N} \sum_i h_{ii}\right| \geq (\log N)^{C_\alpha} \frac{1}{\sqrt{MN}}\right) \leq CN^{-cloglogN}.
\]

On the complementary event we can include the bound \((\log N)^{C_\alpha}(MN)^{-1/2}\) in the last correction term in (2.87). It only remains to estimate

\[
[Z] := \frac{1}{N} \sum_i Z_i,
\]

whose moments are estimated in the next lemma, and we will comment on its proof below.

**Lemma 2.17** ([8], Lemma 5.2 and [7], Lemma 4.1). Recall the definition of \( Z_i \) in (2.50). Then for any fixed \( z \) in the domain \( D \) and any natural number \( s \)

\[
E\left|\frac{1}{N} \sum_{i=1}^N Z_i\right|^{2s} \leq C_s \left(\frac{(\log N)^C}{M\eta}\right)^{2s}.
\]
Using this lemma, we get that for any \( \varepsilon > 0 \) and \( K > 0 \)
\[
P\left( \frac{1}{N} \left| \sum_{i=1}^{N} Z_i \right| \geq \frac{N^\varepsilon}{M \eta} \right) \leq N^{-K}
\] (2.89)
for sufficiently large \( N \). Combining this with (2.88) and (2.87), we obtain (2.86) and complete the proof of Theorem 2.16.

**Sketch of the proof of Lemma 2.17.** We have two different proofs for this lemma, both quite involved. Here we present the ideas of the first proof, from [8]. The argument is a long and carefully organized high-moment calculation, similar to the estimate of the second moment in (2.23), but now we extract an additional factor from the sum. We note that by generalizing the second-moment calculation (2.23) to higher moments (and recalling that for universal Wigner matrices \( M \) replaces \( N \)) we can prove that
\[
|Z_i| \lesssim \frac{1}{\sqrt{M \eta}}
\] (2.90)
(we will indicate the proof in Lemma 2.18 below). If the \( Z_i \) were independent, then the central limit theorem would imply that
\[
\left| \frac{1}{N} \sum_i Z_i \right| \lesssim \frac{1}{\sqrt{N}} \frac{1}{\sqrt{M \eta}},
\]
which would be more than enough. They are not quite independent, but almost. The dependence among the different \( Z_i \) can be extracted by using the resolvent formulae in § 2.1. We will sketch the second-moment calculation, that is, the case \( s = 1 \). More details and higher moments are given in §§ 8–9 of [8].

Since \( E Z_i = 0 \), the variance of \( |Z| \) is given by
\[
\frac{1}{N^2} E \left| \sum_{i=1}^{N} Z_i \right|^2 = \frac{1}{N^2} E \sum_{\alpha \neq \beta} Z_{\alpha} Z_{\beta} + \frac{1}{N^2} E \sum_{\alpha} |Z_{\alpha}|^2.
\] (2.91)
We start by estimating the first term of (2.91) for \( \alpha = 1 \) and \( \beta = 2 \). The basic idea is to rewrite \( G_{kl}^{(1)} \) as
\[
G_{kl}^{(1)} = P_{kl}^{(12)} + P_{kl}^{(1)},
\] (2.92)
with \( P_{kl}^{(12)} \) independent of \( a^1 \) and \( a^2 \), and \( P_{kl}^{(1)} \) independent of \( a^1 \) (recall the notational convention: the superscript indicates independence from the corresponding column of \( H \)). To construct this decomposition for \( k, l \notin \{1, 2\} \), we use (2.10) or (2.11) and rewrite \( G_{kl}^{(1)} \) as
\[
G_{kl}^{(1)} = G_{kl}^{(12)} + \frac{G_{k2}^{(1)} G_{2l}^{(1)}}{G_{22}^{(1)}}, \quad k, l \notin \{1, 2\}.
\] (2.93)
The first term on the right-hand side is independent of \( a^2 \). Applying Theorem 2.9 to the minors, we get that \( G_{kl}^{(1)} \lesssim (M \eta)^{-1/2} \) for \( k \neq l \neq 1 \) and \( G_{kk}^{(2)} \geq c > 0 \), and thus
\[
\left| \frac{G_{k2}^{(1)} G_{2l}^{(1)}}{G_{22}^{(1)}} \right| \lesssim \frac{1}{M \eta}
\] (2.94)
holds with a very high probability. Note that this bound is the square of the bound in $G_{kl}^{(1)} \lesssim (M\eta)^{-1/2}$ in Theorem 2.9.

Now we define $P^{(1)}$ and $P^{(12)}$ (for $k, l \neq 1$) as follows:

1) if $k, l \neq 2$

$$P^{(12)}_{kl} := G_{kl}^{(12)}, \quad P^{(1)}_{kl} := \frac{G_{k2}^{(1)} G_{2l}^{(1)}}{G_{22}^{(1)}} = G_{kl}^{(1)} - G_{kl}^{(12)}; \quad (2.95)$$

2) if $k = 2$ or $l = 2$,

$$P^{(12)}_{kl} := 0, \quad P^{(1)}_{kl} := G_{kl}^{(1)}. \quad (2.96)$$

Hence (2.92) holds, and $P_{kl}^{(12)}$ is independent of $a^2$.

The size of the quadratic forms $a^1 \cdot P^{(1)}a^1$ and $a^1 \cdot P^{(12)}a^1$ is estimated in the following lemma, whose proof is postponed.

**Lemma 2.18.** For $N^{-1} \leq \eta \leq 1$ and fixed $p \in \mathbb{N}$,

$$\mathbb{E}|a^1 \cdot P^{(1)}a^1|^p \lesssim \frac{C_p}{(M\eta)^{p/2}}, \quad \mathbb{E}|a^1 \cdot P^{(12)}a^1|^p \lesssim \frac{C_p}{(M\eta)^{p/2}}. \quad (2.97)$$

Note that the first quadratic form is smaller, but the second one is independent of the second column of $H$. We introduce the operator $I_E_i := I - E_{a_i}$, where $I$ is the identity operator. With this notation we have the following expansion of $Z_1$:

$$Z_1 = I_{E_1}a^1 \cdot P^{(12)}a^1 + I_{E_1}a^1 \cdot P^{(1)}a^1. \quad (2.98)$$

Exchanging the indices 1 and 2, we can define $P^{(21)}$ and $P^{(2)}$ and expand $Z_2$ as

$$Z_2 = I_{E_2}a^2 \cdot P^{(21)}a^2 + I_{E_2}a^2 \cdot P^{(2)}a^2. \quad (2.99)$$

Here $P^{(21)}_{kl}$ is independent of $a^2$ and $a^1$, and $P^{(2)}_{kl}$ is independent of $a^2$. Combining (2.99) with (2.98), we get for the $\alpha = 1, \beta = 2$ cross term in (2.91) that

$$\mathbb{E}Z_1Z_2 = \mathbb{E}\left[(I_{E_1}\{a^1 \cdot P^{(12)}a^1 + a^1 \cdot P^{(1)}a^1\}) (I_{E_2}\{a^2 \cdot P^{(21)}a^2 + a^2 \cdot P^{(2)}a^2\})\right]. \quad (2.100)$$

Note that if $X^{(i)}$ is a random variable independent of $a^i$, then for any random variable $Y$

$$\mathbb{E}_i[YX^{(i)}] = X^{(i)}\mathbb{E}_i Y.$$

In particular, $\mathbb{E}_i X^{(i)} = 0$ (with $Y = 1$). Thus,

$$\mathbb{E}\left[(\mathbb{E}_1X^{(2)}) (\mathbb{E}_2X^{(1)})\right] = \mathbb{E}\left[\mathbb{E}_1\left[(\mathbb{E}_2X^{(1)})X^{(2)}\right]\right] = 0,$$

since $\mathbb{E}\mathbb{E}_i = 0$. Using this idea, one can easily see that the only non-vanishing term on the right-hand side of (2.100) is

$$\mathbb{E}(\mathbb{E}_1a^1P^{(1)}a^1)(\mathbb{E}_2a^2P^{(2)}a^2). \quad (2.101)$$

By the Cauchy–Schwarz inequality and Lemma 2.18 we get that

$$|\mathbb{E}Z_1Z_2| \lesssim \frac{1}{(M\eta)^2}. \quad (2.102)$$
By (2.92), Lemma 2.18 also implies that
\[ E|Z_i|^p \lesssim \frac{C_p}{(M \eta)^{p/2}}, \quad 1 \leq i \leq N, \quad (2.103) \]
that is, it also proves (2.90), and it estimates the second term in (2.91) by the number \( N^{-1}(M \eta)^{-1} \leq (M \eta)^{-2} \). Since the indices 1 and 2 in (2.102) can be replaced by any indices \( \alpha \neq \beta \), together with (2.91) we have thus proved Lemma 2.17 for \( s = 1 \).

**Proof of Lemma 2.18.** First we rewrite \( a_1 \cdot P^{(1)} a_1 \) as follows:
\[
a_1 \cdot P^{(1)} a_1 = \sum_{k,l \neq 2} \overline{a}_{k}^{1} \cdot \frac{G_{k2}^{(1)} G_{2l}^{(1)}}{G_{22}^{(1)}} a_{l}^{1} + \sum_{k \neq 2} \overline{a}_{k}^{1} \cdot G_{k2}^{(1)} a_{2}^{1} + \sum_{l \neq 2} \overline{a}_{l}^{1} \cdot G_{2l}^{(1)} a_{1}^{1} + \overline{a}_{2}^{1} G_{22}^{(1)} a_{2}^{1}. \quad (2.104)\]
By the large deviation estimate (2.70) and the inequality (2.94),
\[
P\left( \left| \sum_{k,l \neq 2} \overline{a}_{k}^{1} \left( \frac{G_{k2}^{(1)} G_{2l}^{(1)}}{G_{22}^{(1)}} \right) a_{l}^{1} \right| \geq \frac{(\log N)^{C}}{M \eta} \right) \leq N^{-c \log \log N}. \quad (2.105)\]
Similarly, from (2.68) and the fact that \( \|a_1\|_{\infty} \lesssim M^{-1/2} \) we get that the second and third terms in (2.104) are bounded by \( \frac{1}{\sqrt{M}} \frac{1}{\sqrt{M \eta}} \lesssim (M \eta)^{-1} \). The last term is even smaller, of order \( 1/M \), with a very high probability. We have thus proved that
\[
P\left( |a_1 \cdot P^{(1)} a_1| \leq \frac{(\log N)^{C}}{M \eta} \right) \geq 1 - N^{-c \log \log N}. \quad (2.106)\]
This inequality implies the first desired inequality in (2.97) except on the exceptional set with probability less than any power of \( 1/N \). Since all the Green functions are bounded by \( \eta^{-1} \leq N \), the contribution from the exceptional set is negligible, and this proves the first estimate in (2.97) The second estimate is proved similarly.

**2.4. Strong local semicircle law.** In this subsection we present our latest results from [7], which remove the \( \kappa \) dependence from Theorem 2.5, Theorem 2.6, and Theorem 2.7 for ensembles with spread \( M = N \), in particular, for generalized Wigner matrices. Recall the notation
\[
\Lambda_d := \max_k |G_{kk} - m_{sc}|, \quad \Lambda_o := \max_{i \neq j} |G_{ij}|, \quad \Lambda := |m - m_{sc}|, \]
and recall also that these quantities depend on the spectral parameter \( z \) and on \( N \).

**Theorem 2.19** (strong local semicircle law [7], Theorem 2.1). Let \( H = (h_{ij}) \) be a Hermitian or symmetric \( N \times N \) random matrix, \( N \geq 3 \), with \( \mathbb{E} h_{ij} = 0 \) for \( 1 \leq i, j \leq N \). Assume that the variances \( \sigma_{ij}^2 \) satisfy (1.16) and (2.29) with some positive constants \( \delta_{\pm} > 0 \) and have the estimate
\[
\sigma_{ij}^2 \leq \frac{C_0}{N}. \quad (2.107)\]
Suppose further that the distributions of the matrix elements have uniform subexponential decay in the sense that there exists a constant $\vartheta > 0$ independent of $N$ such that for any $x \geq 1$ and $1 \leq i, j \leq N$

$$P(|h_{ij}| \geq x\sigma_{ij}) \leq \vartheta^{-1} \exp(-x^\vartheta). \tag{2.108}$$

Then for any constant $\xi > 1$ there exist positive constants $L$, $C$, and $c$ depending only on $\xi$ and the $\vartheta$ in (2.108), on the $\delta_{\pm}$ in (2.29), and on the $C_0$ in (2.107) such that the Stieltjes transform of the empirical eigenvalue distribution of $H$ satisfies the inequality

$$P\left( \bigcup_{z \in \mathcal{S}_L} \left\{ \Lambda(z) \geq \frac{(\log N)^4L}{N\eta} \right\} \right) \leq C \exp[-c(\log N)^\xi], \tag{2.109}$$

where

$$\mathcal{S} := \mathcal{S}_L = \{ z = E + i\eta : |E| \leq 5, \ N^{-1}(\log N)^{10L} < \eta \leq 10 \}. \tag{2.110}$$

The individual matrix elements of the Green function satisfy the inequality

$$P\left( \bigcup_{z \in \mathcal{S}_L} \left\{ \Lambda_{d}(z) + \Lambda_{o}(z) \geq \frac{\log N}{N\eta} \sqrt{\frac{\Im m_{sc}(z)}{N\eta}} + \frac{(\log N)^4L}{N\eta} \right\} \right) \leq C \exp[-c(\log N)^\xi]. \tag{2.111}$$

Furthermore, interior to the set

$$\mathcal{O}_L := \{ z = E + i\eta : N\eta\sqrt{\kappa} \geq (\log N)^{4L}, \ \kappa \geq \eta, \ |E| > 2 \ \} \ \text{with} \ \kappa := ||E|-2| \tag{2.112}$$

and outside of the limiting spectrum there is the stronger estimate

$$P\left( \bigcup_{z \in \mathcal{O}_L} \left\{ \Lambda(z) \geq \frac{(\log N)^4L}{N\kappa} \right\} \right) \leq C \exp[-c(\log N)^\xi]. \tag{2.113}$$

The subexponential decay condition (2.108) can be weakened if we are not aiming at error estimates decaying faster than any power law of $N$. This can easily be carried out and we will not pursue it in this paper.

Prior to our results in [9] and [8], a central limit theorem for the semicircle law on a macroscopic scale for band matrices was established by Guionnet [59] and Anderson and Zeitouni [60], and a semicircle law for Gaussian band matrices was proved by Disertori, Pinson, and Spencer [51]. For a survey on band matrices, see the recent article [103] by Spencer.

As before, the local semicircle estimates imply that the empirical counting function of the eigenvalues is close to the semicircle counting function and that the locations of the eigenvalues are close to their classical locations. We have the following improved results (cf. Theorems 2.6–2.7).

**Theorem 2.20** ([7], Theorem 2.2). Assume the conditions of Theorem 2.19, that is, (1.16), (2.29) with some positive constants $\delta_{\pm} > 0$, (2.107), and (2.108). Then
for any constant $\xi > 1$ there exist constants $L_1$, $C$, and $c > 0$ depending only on $\xi$, $\vartheta$, $\delta_\pm$, and $C_0$ such that

\[
P\{ \exists j : |\lambda_j - \gamma_j| \geq (\log N)^{L_1} [\min\{j, N - j + 1\}]^{-1/3} N^{-2/3} \} \leq C \exp[-c(\log N)^\xi]. \tag{2.114}
\]

and

\[
P\left\{ \sup_{|E| \leq 5} |n(E) - n_{sc}(E)| \geq \frac{(\log N)^{L_1}}{N} \right\} \leq C \exp[-c(\log N)^\xi]. \tag{2.115}
\]

For Wigner matrices the estimate (2.115) with the factor $N^{-1}$ replaced by $N^{-2/5}$ was established in [104] (in a weaker sense with some modifications in the statement), and a stronger control with the factor $N^{-1/2}$ was proved for the difference $E_n(E) - n_{sc}(E)$. Theorem 1.3 of [105] contains the estimate (in our scaling)

\[
(E[|\lambda_j - \gamma_j|^2])^{1/2} \leq [\min\{j, N - j + 1\}]^{-1/3} N^{-1/6 - \varepsilon_0} \tag{2.116}
\]

with a small positive $\varepsilon_0$ for standard Wigner matrices under the assumption that the third moment of the matrix elements vanishes. In the same paper it was conjectured that the factor $N^{-1/6 - \varepsilon_0}$ on the right-hand side of (2.116) should be replaced by $N^{-2/3 + \varepsilon}$. Prior to the work [105], the estimate (2.114) was proved in [8] away from the edges with a slightly weaker probability estimate and with the $(\log N)^{L_1}$ factor replaced by $N^\delta$ for arbitrary $\delta > 0$ (see the equation before (7.8) in [8]).

We remark that all results are stated for both the Hermitian and symmetric cases, but the statements and the proofs hold for quaternion self-dual random matrices as well (see, for example, [6], §3.1).

There are several possible improvements of the arguments in §§2.3.2 and 2.3.3, and these led to the proof of the optimal Theorem 2.19 for the $M = N$ case. Here we mention only the most important ones.

First, note that the last part of the proof of (2.72) can be carried out more efficiently for $M = N$ and $\sigma_{ik}^2 \leq C/N$ using the fact that

\[
\frac{1}{N} \sum_k \text{Im} G^{(i)}_{kk} \leq \frac{1}{N} \sum_k \text{Im} G_{kk} + CA_o^2 = \text{Im} m + CA_o^2 \leq \Lambda + \Lambda_o^2 + \text{Im} m_{sc}.
\]

The gain here is that $\text{Im} m_{sc}(z) \sim \sqrt{\kappa + \eta}$, which is a better estimate near the edges than just $O(1)$. We therefore introduce

\[
\Psi = \Psi(z) := \sqrt{\frac{\Lambda(z) + \text{Im} m_{sc}(z)}{N \eta}}
\]

as our main control parameter, and note that this is random, but it depends only on $\Lambda$. In a way similar to that for (2.75) and (2.76), one can then prove that

\[
\Lambda_o + \max_i \Upsilon_i \leq \Psi \tag{2.117}
\]

with very high probability.
Second, a more careful algebraic analysis of the self-consistent equation (2.78) yields the following identity:

\[(1 - m_{sc}^2)[v] = m_{sc}^3[v]^2 + m_{sc}^2[Z] + O\left(\frac{\Lambda^2}{\log N}\right) + O((\log N)\Psi^2),\]  

(2.118)

where \([Z] := N^{-1} \sum_{i=1}^{N} Z_i\). The advantage of this formula is that it enables one not only to express \([v]\) from the left-hand side (after dividing through by \(1 - m_{sc}^2\)), but also to express \([v]\) from the right-hand side in the case when

\[|(1 - m_{sc}^2)[v]| \ll |m_{sc}^3[v]|^2\]

(which typically happens exactly near the edges, where \(1 - m_{sc}^2\) is small). This makes a dichotomy estimate possible by noting that if

\[(1 - m_{sc}^2)[v] = m_{sc}^3[v]^2 + \text{a small term},\]

or, in other words, if

\[\alpha(z) \Lambda = \Lambda^2 + \beta(z), \quad \alpha(z) := \left|\frac{1 - m_{sc}^2}{m_{sc}^3}\right| \sim \sqrt{\kappa + \eta}, \quad \beta \text{ a small term,}\]

then for some sufficiently large constant \(U\) and another constant \(C_1(U)\), we have

\[\Lambda(z) \lessgtr U \beta(z) \quad \text{or} \quad \Lambda(z) \gtrless \frac{\alpha(z)}{U} \quad \text{if} \quad \alpha \gtrless U^2 \beta, \quad (2.119)\]
\[\Lambda(z) \lessgtr C_1(U) \beta(z) \quad \text{if} \quad \alpha \lessgtr U^2 \beta. \quad (2.120)\]

The bad case \(\Lambda \gtrsim \alpha/U\) is excluded by a continuity argument: we can easily show that for \(\eta = 10\) this does not happen, and then we decrease \(\eta\) and see that either we are always in the first case (when \(\alpha \gtrsim U^2 \beta\)), or we automatically have \(\Lambda \lessgtr \beta\) (when \(\alpha \lessgtr U^2 \beta\)). The actual proof is more complicated, since the ‘small term’ itself depends on \(\Lambda\), but this dependence can be absorbed into the other terms. Moreover, all these estimates hold only with a very high probability, so exceptional events have to be tracked.

Finally, the estimate in Lemma 2.17 has to incorporate the improved control on the edges. Inspecting the proof, one sees that the gain \((M\eta)^{-2s}\) comes from the off-diagonal resolvent elements. In fact, Lemma 2.17 is better written in the following form:

\[E \left| \frac{1}{N} \sum_{i=1}^{N} Z_i \right|^{2s} \leq C_s E[\Lambda_o^{2s} + N^{-s}].\]

As before, this can be turned into a probability estimate by taking a large power, \(s \sim (\log N)^{\xi}\). Using (2.117) to estimate \(\Lambda_o\) in terms of \(\Psi\) and the fact that \(\Psi^2 \lessgtr o(\Lambda) + N^{-1}\) (since \(N\eta \gg 1\)), one can show that the term \(m_{sc}^2[Z]\) on the right-hand side of (2.118) is also a ‘small term’. The details are given in §§3 and 4 of [7].
2.5. Delocalization of eigenvectors. Let $H$ be a universal Wigner matrix with subexponential decay (2.32). If $v$ is an $\ell^2$-normalized eigenvector of $H$, then for $p > 2$ the size of the $\ell^p$-norm of $v$ gives information about delocalization of $v$. We say that complete delocalization occurs when $\|v\|_p \lesssim N^{-1/2+1/p}$ (note that $\|v\|_p \gtrsim CN^{-1/2+1/p}\|v\|_2$). The following result shows that for generalized Wigner matrices (1.17), the eigenvectors are completely delocalized with a very high probability. For universal Wigner matrices with spread $M$ in (2.31) the eigenvectors are delocalized on a scale at least $M$.

**Theorem 2.21.** Under the conditions of Theorem 2.5,

$$\mathbb{P}\left\{ \exists v : Hv = \lambda v, \; |\lambda - E| \leq \frac{1}{M}, \; \|v\|_2 = 1, \; \|v\|_\infty \geq \frac{(\log N)^C}{\sqrt{M}} \right\} \lesssim C N^{-c \log \log N}$$

for any $E$ with $|\kappa = |E - 2| \geq \kappa_0$.

We remark that $\|v\|_\infty \lesssim M^{-1/2}$ indicates that the eigenvector has to be supported in a region of size at least $M$, that is, the localization length of $v$ is at least $M$. We note also that the delocalization conjecture predicts that the localization length is in fact $M^2$, that is, the optimal estimate should be

$$\|v\|_\infty \lesssim \frac{1}{M}$$

with a high probability. As explained in §1.3, this is an open question. Only some partial results are available, for example, we proved in [49], [50] that for random band matrices (1.18) with band width $W \sim M$, the localization length is at least $M^{1+1/6}$.

**Proof.** We again neglect the dependence of the estimate on $\kappa_0$ (this can be traced back from the proof). The estimate (2.45) guarantees that

$$|G_{ii}(z)| \leq C$$

for any $z = E + i\eta$ with $M\eta \geq (\log N)^C$, with a very high probability. Choose $\eta = (\log N)^C/M$. Let $v_\alpha$ be the eigenvectors of $H$, and let $v$ be an eigenvector with eigenvalue $\lambda$, where $|\lambda - E| \leq 1/M$. Thus,

$$|v(i)|^2 \leq \frac{2\eta^2 |v(i)|^2}{M^{-2} + \eta^2} \leq 2 \sum_\alpha \frac{\eta^2 |v_\alpha(i)|^2}{(\lambda_\alpha - E)^2 + \eta^2} = 2\eta |\text{Im} G_{ii}| \leq C \eta,$$

that is,

$$\|v\|_\infty \leq \frac{(\log N)^C}{\sqrt{M}}.$$

We remark that this proof was very easy because pointwise bounds on the diagonal elements of the resolvent were available. It is possible to prove this theorem using only the local semicircle law, which is a conceptually simpler input; in fact, this was our traditional path in [55]–[57]. For example, in [57] we proved the following theorem.
Theorem 2.22 ([57], Corollary 3.2). Let $H$ be a Wigner matrix with single-entry distribution having Gaussian decay. Then for any $|E| < 2$, any fixed $K$, and any $2 < p < \infty$

$$
\mathbb{P}\left\{ \exists \mathbf{v} : H\mathbf{v} = \lambda\mathbf{v}, |\lambda - E| \leq \frac{K}{N}, \|\mathbf{v}\|_2 = 1, \|\mathbf{v}\|_p \geq QN^{-1/2+1/p} \right\} \leq C e^{-c\sqrt{Q}}
$$

for sufficiently large $Q$ and $N$.

Sketch of the proof. We will give the proof of a weaker result, where logarithmic factors are allowed. Suppose that $H\mathbf{v} = \lambda\mathbf{v}$ and $\lambda \in [-2 + \kappa_0, 2 - \kappa_0]$ with $\kappa_0 > 0$. Consider the decomposition

$$
H = \begin{pmatrix} h & \mathbf{a}^* \\ \mathbf{a} & H^{(1)} \end{pmatrix},
$$

introduced in §2.1, that is, here $\mathbf{a} = (h_{1,2}, \ldots, h_{1,N})^*$ and $H^{(1)}$ is the $(N - 1) \times (N - 1)$ matrix obtained by removing the first row and first column from $H$. Let $\mu_\alpha$ and $\mathbf{u}_\alpha$ (for $\alpha = 1, \ldots, N - 1$) denote the eigenvalues and the normalized eigenvectors of $H^{(1)}$. From the eigenvalue equation $H\mathbf{v} = \lambda\mathbf{v}$ we find that

$$(h - \lambda)v_1 + \mathbf{a} \cdot v' = 0,$$

$$v_1\mathbf{a} + (H^{(1)} - \lambda)v' = 0,$$

where we have decomposed the eigenvector as $\mathbf{v} = (v_1, v')$, $v_1 \in \mathbb{R}$, $v' \in \mathbb{R}^{N-1}$. Solving the second equation for $v'$, we get $v' = v_1(\lambda - H^{(1)})^{-1}\mathbf{a}$. From the normalization condition $\|\mathbf{v}\|^2 = v_1^2 + \|v'\|^2 = 1$ we thus obtain for the first component of $\mathbf{v}$ that

$$|v_1|^2 = \frac{1}{1 + \mathbf{a} \cdot (\lambda - H^{(1)})^{-2}\mathbf{a}} = \frac{1}{1 + N^{-1} \sum_\alpha \xi_\alpha/(\lambda - \mu_\alpha)^2} \leq \frac{4N \eta^2}{\sum_{\mu_\alpha \in I} \xi_\alpha},$$

where in the second equality we let $\xi_\alpha = |\sqrt{N} \mathbf{a} \cdot \mathbf{u}_\alpha|^2$ and used the spectral representation of $H^{(1)}$. We also chose an interval $I$ of length $\eta = |I| = Q/N$. It is easy to check that $\mathbb{E}\xi_\alpha = 1$ and that the different $\xi_\alpha$ are actually independent and satisfy the following large deviation estimate:

$$\mathbb{P}\left( \sum_{\alpha \in \mathcal{I}} \xi_\alpha \leq \frac{m}{2} \right) \leq e^{-c\sqrt{m}},$$

where $m = |\mathcal{I}|$ is the cardinality of the index set. There are several proofs of this fact, depending on the condition on the single-entry distribution. For example, under the Gaussian decay condition it was proved in Lemma 4.7 of [57] using the Hanson–Wright theorem [101].

Now let $\mathcal{N}_I$ denote the number of eigenvalues of $H$ in $I$. From the local semicircle law (for instance, Theorem 1.11) we know that $\mathcal{N}_I$ is of order $N|I|$ for any interval $I$ away from the edges such that $|I| \gg 1/N$. We recall that the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$ of $H$ and the eigenvalues of $H^{(1)}$ are interlaced. This means that
there exist at least $N_I - 1$ eigenvalues of $H^{(1)}$ in $I$. Therefore, using the fact that the components of any eigenvector are identically distributed, we have

$$P\left( \exists \mathbf{v} \text{ with } H \mathbf{v} = \lambda \mathbf{v}, \| \mathbf{v} \| = 1, \lambda \in I \text{ and } \| \mathbf{v} \|_{\infty} \geq \frac{Q}{N^{1/2}} \right)$$

$$\leq N^2 P\left( \exists \mathbf{v} \text{ with } H \mathbf{v} = \lambda \mathbf{v}, \| \mathbf{v} \| = 1, \lambda \in I \text{ and } |v_1|^2 \geq \frac{Q^2}{N} \right)$$

$$\leq CN^2 P\left( \sum_{\mu_\alpha \in I} \xi_\alpha \leq \frac{4N^2 \eta^2}{Q^2} \right)$$

$$\leq CN^2 P\left( \sum_{\mu_\alpha \in I} \xi_\alpha \leq \frac{4N^2 \eta^2}{Q^2} \text{ and } N_I \geq cN|I| \right) + CN^2 P(N_I \leq cN|I|)$$

$$\leq CN^2 e^{-c\sqrt{N|I|}} + CN^2 e^{-c\sqrt{N|I|}} \leq Ce^{-c\sqrt{Q}},$$

(2.124)

assuming that $4N^2 \eta^2/Q^2 \leq cN|I| = cN|I|$, that is, that $Q \geq \sqrt{N\eta}$.

Here we used the fact that the deviation from the semicircle law is subexponentially suppressed, that is,

$$P(N_I \leq cN|I|) \leq e^{-c'\sqrt{N|I|}}$$

for sufficiently small $c$ and $c'$ if $I$ is away from the spectral edges. Such a strong subexponential bound does not directly follow from the local semicircle law Theorem 2.5 whose proof was outlined in §2.3, but it can be proved for Wigner matrices with Gaussian decay (Theorem 1.11).

3. Universality for Gaussian convolutions

3.1. Strong local ergodicity of the Dyson Brownian motion. In this subsection we consider the following general question. Suppose that $\mu = e^{-N\mathcal{H}}/Z$ is a probability measure on the configuration space $\mathbb{R}^N$, characterized by some Hamiltonian $\mathcal{H}: \mathbb{R}^N \to \mathbb{R}$, where $Z = \int e^{-N\mathcal{H}(x)} \, dx < \infty$ is the normalization. We will always assume that $\mathcal{H}$ is symmetric under permutations of the variables $x = (x_1, \ldots, x_N) \in \mathbb{R}^N$. The typical example to keep in mind is the Hamiltonian of the general $\beta$-ensembles (1.45) or the specific GOE ($\beta = 1$) and GUE ($\beta = 2$) cases.

Let us consider the time-dependent permutation-symmetric probability density $f_t(x)$, $t \geq 0$, with respect to the measure $\mu(dx) = \mu(x) \, dx$, that is, $\int f_t(x) \mu(dx) = 1$. The dynamics is characterized by the forward equation

$$\partial_t f_t = L f_t, \quad t \geq 0,$$

(3.1)

with given permutation-symmetric initial data $f_0$. The generator $L$ is defined via the Dirichlet form as

$$D(f) := D_\mu(f) = -\int f L f \, d\mu = \sum_{j=1}^N \frac{1}{2N} \int (\partial_j f)^2 \, d\mu, \quad \partial_j = \partial_{x_j}.$$

(3.2)
Formally, we have $L = \frac{1}{2N} \Delta - \frac{1}{2} (\nabla \mathcal{H}) \nabla$. We will ignore questions concerning the domain, noting only that $D(f)$ is a semibounded quadratic form, so $L$ can be defined via the Friedrichs extension on $L^2(d\mu)$, and it can be extended to $L^1$ as well. The dynamics is well defined for any $f_0 \in L^1(d\mu)$. For more details, see Appendix A of [6].

Strictly speaking, we will consider a sequence of Hamiltonians $\mathcal{H}_N$ and corresponding dynamics $L_N$ and $f_{t,N}$ parametrized by $N$, but the $N$-dependence will be omitted. All results will concern the limit as $N \to \infty$.

Alternatively to (3.1), one could describe the dynamics by a coupled system of stochastic differential equations (1.64), as mentioned in §1.6.2, but we will not use this formalism here.

For any $k \geq 1$ we define the $k$-point correlation functions (marginals) of the probability measure $f_t d\mu$ by

$$p^{(k)}_{t,N}(x_1, \ldots, x_k) = \int_{\mathbb{R}^{N-k}} f_t(x) \mu(x) \, dx_{k+1} \cdots dx_N. \quad (3.3)$$

The correlation functions of the equilibrium measure are denoted by

$$p^{(k)}_{\mu,N}(x_1, \ldots, x_k) = \int_{\mathbb{R}^{N-k}} \mu(x) \, dx_{k+1} \cdots dx_N.$$ 

We now list our main assumptions on the initial distribution $f_0$ and on its evolution $f_t$. This formalism is adjusted to generalized Wigner matrices; random covariance matrices require some minor modifications (see [6] for details). We first define the subdomain

$$\Sigma_N := \{ x \in \mathbb{R}^N : x_1 < x_2 < \cdots < x_N \} \quad (3.4)$$

of ordered tuples $x$.

**Condition I.** The Hamiltonian $\mathcal{H}$ of the equilibrium measure has the form

$$\mathcal{H} = \mathcal{H}_N(x) = \beta \left[ \sum_{j=1}^{N} U(x_j) - \frac{1}{N} \sum_{i<j} \log |x_i - x_j| \right], \quad (3.5)$$

where $\beta \geq 1$. The function $U : \mathbb{R} \to \mathbb{R}$ is smooth, with $U'' \geq 0$, and

$$U(x) \geq C|x|^\delta \quad \text{for some } \delta > 0 \text{ and large } |x|.$$ 

This assumption is automatic for symmetric and Hermitian Wigner matrices, with the GOE or GUE invariant measure, respectively (see (1.44)–(1.45)).

Near the $x_{i+1} = x_i$ boundary component of $\Sigma_N$ the generator has the form

$$L = \frac{1}{4N} \left( \partial_u^2 + \frac{\beta}{u} \partial_u \right) + \text{regular operator}$$

in the relative coordinates $u = (x_{i+1} - x_i)/2$ when $u \ll 1$. It is known (Appendix A of [6]) that for $\beta \geq 1$ the repelling force of the singular diffusion is sufficiently strong to prevent the particle from falling into the origin, that is, in the original
coordinates the trajectories of the points do not cross. In particular, the ordering of the points will be preserved under the dynamics (for a stochastic proof, see Lemma 4.3.3 of [3]). In what follows we will thus assume that $f_t$ is a probability measure on $\Sigma_N$. We continue to use the notation $f$ and $\mu$ for the restricted measure.

Note that the correlation functions $p^{(k)}$ in (3.3) are still defined on $\mathbb{R}^k$, that is, their arguments remain unordered.

It follows from Condition I that the Hessian matrix of $\mathcal{H}$ satisfies the following estimate:

$$\langle v, \nabla^2 \mathcal{H}(x)v \rangle \geq \frac{\beta}{N} \sum_{i<j} \frac{(v_i - v_j)^2}{(x_i - x_j)^2}, \quad v = (v_1, \ldots, v_N) \in \mathbb{R}^N, \quad x \in \Sigma_N. \quad (3.6)$$

This convexity bound is the key assumption; our method works for a broad class of general Hamiltonians as long as (3.6) holds. In particular, an arbitrary many-body potential function $V(x)$ can be added to the Hamiltonian (3.5) as long as $V$ is convex on $\Sigma_N$.

**Condition II.** There exists on the real line a continuous compactly supported density function $\varrho(x) \geq 0$, $\int_\mathbb{R} \varrho = 1$, that is independent of $N$ and such that for any fixed $a, b \in \mathbb{R}$

$$\lim_{N \to \infty} \sup_{t \geq 0} \left| \frac{1}{N} \sum_{j=1}^{N} 1(x_j \in [a, b]) f_t(x) \, d\mu(x) - \int_a^b \varrho(x) \, dx \right| = 0. \quad (3.7)$$

In other words, we assume that a limiting density exists; for Wigner matrices this is the semicircle law. Let $\gamma_j = \gamma_{j,N}$ denote the location of the $j$th point under the limiting density, that is, $\gamma_j$ is defined by

$$N \int_{-\infty}^{\gamma_j} \varrho(x) \, dx = j, \quad 1 \leq j \leq N, \quad \gamma_j \in \text{supp} \varrho. \quad (3.8)$$

We will call $\gamma_j$ the classical location of the $j$th point. Note that $\gamma_j$ may not be uniquely defined if the support of $\varrho$ is not connected, but in this case Condition III below will not be satisfied anyway.

**Condition III.** There exists an $a > 0$ such that

$$\sup_{t \geq N^{-2a}} \int \frac{1}{N} \sum_{j=1}^{N} (x_j - \gamma_j)^2 f_t(dx) \, d\mu(dx) \leq CN^{-1-2a}, \quad (3.9)$$

with a constant $C$, uniformly with respect to $N$.

Under Condition II the typical spacing between neighbouring points is of order $1/N$ away from the spectral edges, that is, in a neighbourhood of any energy $E$ with $\varrho(E) > 0$. Condition III guarantees that the random points $x_j$ typically remain in the $N^{-1/2-a}$-neighbourhood of their classical locations.

The final assumption is an upper bound on the local density. For any interval $I \subset \mathbb{R}$ let $\mathcal{N} := \sum_{i=1}^{N} 1 \ (x_i \in I)$ denote the number of points in $I$. 

Condition IV. For any compact subinterval $I_0 \subset \{ E : \varrho(E) > 0 \}$ and for any $\delta > 0$ and $\sigma > 0$ there are constants $C_n$, $n \in \mathbb{N}$, depending on $I_0$, $\delta$, and $\sigma$ such that for any interval $I \subset I_0$ with $|I| \geq N^{-1+\sigma}$ and for any $K \geq 1$

$$\sup_{\tau \geq N^{-2a}} \int 1\{N_I \geq KN|I|\}f_\tau \, d\mu \leq C_n K^{-n}, \quad n = 1, 2, \ldots,$$  \hspace{1cm} (3.10)

where $a$ is the exponent in Condition III.

We note that for symmetric or Hermitian Wigner matrices, Condition I is automatic, Condition II is the (global) semicircle law (1.27), and Condition IV is the upper bound on the density (Lemma 1.10). The really serious condition to check is (3.9).

The following main general theorem asserts that the local statistics of the points $x_j$ in the bulk with respect to the time-evolving distribution $f_t$ coincide with the local statistics with respect to the equilibrium measure $\mu$ as long as $t \gg N^{-2a}$.

**Theorem 3.1** ([6], Theorem 2.1). Suppose that the Hamiltonian given in (3.5) satisfies Condition I and that Conditions II, III, and IV hold for the solution $f_t$ of the forward equation (3.1) with exponent $a$. Assume that at time $t_0 = N^{-2a}$ the density $f_{t_0}$ satisfies a bounded entropy condition, that is,

$$S_\mu(f_{t_0}) := \int f_{t_0} \log f_{t_0} \, d\mu \leq CN^m$$  \hspace{1cm} (3.11)

with some fixed exponent $m$. Let $E \in \mathbb{R}$ and $b > 0$ be such that $\min\{\varrho(x) : x \in [E-b, E+b]\} > 0$. Then for any $\delta > 0$, any integer $k \geq 1$, and any compactly supported continuous test function $O : \mathbb{R}^k \to \mathbb{R}$

$$\lim_{N \to \infty} \sup_{t \geq \tau} \int_{E-b}^{E+b} \frac{dE'}{2b} \int_{\mathbb{R}^k} d\alpha_1 \cdots d\alpha_k O(\alpha_1, \ldots, \alpha_k)$$

$$\times \frac{1}{\varrho(E)^k} \left( p_{t,N}^{(k)} - p_{\mu,N}^{(k)} \right) \left( E' + \frac{\alpha_1}{N\varrho(E)} , \ldots , E' + \frac{\alpha_k}{N\varrho(E)} \right) = 0,$$  \hspace{1cm} (3.12)

where $\tau := N^{-2a+\delta}$.

We remark that this limit can be effectively controlled. Indeed, in [6] we proved that, before the limit $N \to \infty$, the left-hand side of (3.12) is bounded by

$$CN^{2\epsilon'}[b^{-1}N^{-(1+2a)/3} + b^{-1/2}N^{-\delta/2}].$$

In many applications the local equilibrium statistics can be explicitly computed, and in the limit as $b \to 0$ they become independent of $E$. In particular, this is the case for the classical matrix ensembles. The simplest explicit formula is for the GUE case, when the correlation functions are given by the sine kernel (1.35).

### 3.2. The local relaxation flow

The main idea behind the proof of Theorem 3.1 is to analyze the relaxation to equilibrium of the dynamics (3.1). The equilibrium is given by an essentially convex Hamiltonian $\mathcal{H}$, so the Bakry–Emery method [91] applies. This method was first used in the context of the Dyson Brownian motion in §5.1 of [89]; the presentation here follows [6].
To explain the idea, assume for the time being that the potential $U$ in (3.5) is uniformly convex, that is,

$$U''(x) \geq U_0 > 0.$$  

This is certainly the case for the Gaussian ensembles, when $U(x) = x^2/4$. Then we have the following lower bound on the Hessian of $\mathcal{H}$:

$$\text{Hess} \mathcal{H} \geq \beta U_0$$  \hspace{1cm} (3.13)

on the set $\Sigma_N$ in (3.4), since the logarithmic potential is convex. It is essential to stress at this point that (3.13) holds only in the open set $\Sigma_N$, since the second derivatives of the logarithmic interactions have a delta-function singularity (with the ‘wrong’ sign) on the boundary. It requires a separate technical argument to show that for $\beta \geq 1$ the points sufficiently repel each other so that the Dyson Brownian motion never leaves the open set $\Sigma_N$ and thus the Bakry–Emery method applies. See the remark after Theorem 3.2 below.

In the following pedagogical subsection we formulate the Bakry–Emery criterion in a general setting on $\mathbb{R}^N$.

3.2.1. Bakry–Emery method. Let the probability measure $\mu$ on $\mathbb{R}^N$ be given by a strictly convex Hamiltonian $\mathcal{H}$:

$$d\mu(x) = \frac{e^{-\mathcal{H}(x)}}{Z} \, dx,$$

with some constant $K$, and let $L$ be the generator of the dynamics associated with the Dirichlet form

$$D(f) = D_\mu(f) = -\int f Lf \, d\mu := \frac{1}{2} \sum_j \int (\partial_j f)^2 \, d\mu, \quad \partial_j = \partial_{x_j}$$

(note that in this presentation we neglect the prefactor $1/N$ originally present in (3.2)). Formally, we have $L = \frac{1}{2} \Delta - \frac{1}{2} (\nabla \mathcal{H}) \nabla$. The operator $L$ is symmetric with respect to the measure $d\mu$, that is,

$$\int f Lg \, d\mu = \int (Lf)g \, d\mu = -\frac{1}{2} \int \nabla f \cdot \nabla g \, d\mu.$$  \hspace{1cm} (3.15)

We define the relative entropy of any probability density $f$ with $\int f \, d\mu = 1$ by

$$S_\mu(f) = S(f) = \int f(\log f) \, d\mu.$$  

Both the Dirichlet form and the entropy are non-negative, they are zero only for $f \equiv 1$, and they measure the distance of $f$ from the equilibrium $f = 1$. The entropy can be used to control the total variation norm directly via the entropy inequality

$$\int |f - 1| \, d\mu \leq \sqrt{2S_\mu(f)}.$$  \hspace{1cm} (3.16)

Let $f_t$ be the solution to the evolution equation

$$\partial_t f_t = Lf_t, \quad t > 0,$$  \hspace{1cm} (3.17)
with a given initial condition \( f_0 \), and consider the evolution of the entropy \( S(f_t) \) and the Dirichlet form \( D(\sqrt{f_t}) \). A simple calculation shows that

\[
\partial_t S(f_t) = \int (L f_t) \log f_t \, d\mu + \int f_t L f_t \, d\mu = -\frac{1}{2} \int \frac{(\nabla f_t)^2}{f_t} \, d\mu = -4D(\sqrt{f_t}),
\]

where we have used the fact that \( \int L f_t \, d\mu = 0 \) by (3.15). Similarly, we can compute the evolution of the Dirichlet form. Let \( h := \sqrt{f} \) for simplicity; then

\[
\partial_t h_t = \frac{1}{2} \partial_h \int (\nabla h)^2 \, d\mu = \frac{1}{2} \left[ \left( \partial_i h \right) \left( \partial_j h \right) - \frac{\left( \partial_i h \right)^2 \left( \partial_j h \right)}{h^2} \right] \, d\mu.
\]

In the last step we used the equality \( L h^2 = (\nabla h)^2 + 2h L h \), which can be seen either directly from \( L = \frac{1}{2} \Delta - \frac{1}{2} (\nabla \mathcal{H}) \nabla \), or from the following identity for any test function \( g \):

\[
\int g L h^2 \, d\mu = -\frac{1}{2} \int \nabla g \cdot \nabla (h^2) \, d\mu = -\int h(\nabla g)(\nabla h) \, d\mu = \int [-\nabla (hg) + g \nabla h] \nabla h \, d\mu = \int g[(\nabla h)^2 + 2h L h] \, d\mu.
\]

We compute (dropping the \( t \) subscript for brevity)

\[
\partial_t D(\sqrt{f_t}) = \frac{1}{2} \partial_t \int (\nabla h)^2 \, d\mu = \int (\nabla h)(\nabla L h) \, d\mu + \frac{1}{2} \int (\nabla h) \cdot \nabla \left( \frac{(\nabla h)^2}{h} \right) \, d\mu
\]

\[
= \int (\nabla h)[\nabla, L] h \, d\mu + \int (\nabla h) L(\nabla h) \, d\mu
\]

\[
+ \frac{1}{2} \int \sum_{i,j} \partial_i h \left[ \frac{2(\partial_j h) \partial_i \partial_j h}{h} - \frac{(\partial_j h)^2 \partial_i h}{h^2} \right] \, d\mu
\]

\[
= -\frac{1}{2} \int (\nabla h)(\nabla^2 \mathcal{H}) \nabla h \, d\mu - \frac{1}{2} \int \sum_{i,j} (\partial_i \partial_j h)^2 \, d\mu
\]

\[
+ \frac{1}{2} \int \sum_{i,j} \left[ \frac{2(\partial_j h) \partial_i h}{h} - \frac{(\partial_j h)^2 \partial_i h}{h^2} \right] \, d\mu
\]

\[
= -\frac{1}{2} \int (\nabla h)(\nabla^2 \mathcal{H}) \nabla h \, d\mu - \frac{1}{2} \int \sum_{i,j} \left( \partial_{ij} h - \frac{(\partial_i h)(\partial_j h)}{h} \right)^2 \, d\mu,
\]

where we used the commutator

\[
[\nabla, L] = -\frac{1}{2} (\nabla^2 \mathcal{H}) \nabla.
\]

Therefore, under the convexity condition (3.14), we have

\[
\partial_t D(\sqrt{f_t}) \leq -KD(\sqrt{f_t}).
\]

Combining (3.18) and (3.20),

\[
\partial_t D(\sqrt{f_t}) \leq \frac{K}{4} \partial_t S(f_t).
\]
At $t = \infty$ the equilibrium is reached, $f_\infty = 1$, and both the entropy and the Dirichlet form are zero. After integrating (3.21) backwards from $t = \infty$, we get the \textit{logarithmic Sobolev inequality}

$$S(f_t) \leq \frac{4}{K} D(\sqrt{f_t})$$

(3.22)

for any $t \geq 0$; in particular, this holds for any initial distribution $f = f_0$. Inserting this back into (3.18), we have

$$\partial_t S(f_t) \leq -KS(f_t).$$

Integrating from time zero, we obtain \textit{exponential relaxation of the entropy on the time scale $t \sim 1/K$}:

$$S(f_t) \leq e^{-tK} S(f_0).$$

(3.23)

Finally, we can integrate (3.18) from time $t/2$ to $t$ and get that

$$S(f_t) - S(f_{t/2}) = -4 \int_{t/2}^{t} D(h\sqrt{f_\tau}) \, d\tau.$$ 

From the positivity of the entropy $S(f_t) \geq 0$ on the left-hand side and the monotonicity of the Dirichlet form (from (3.20)) on the right-hand side, we get that

$$D(\sqrt{f_t}) \leq \frac{2}{t} S(f_{t/2}),$$

(3.24)

and thus by (3.23) we have \textit{exponential relaxation of the Dirichlet form on the time scale $t \sim 1/K$}:

$$D(\sqrt{f_t}) \leq \frac{2}{t} e^{-tK/2} S(f_0).$$

(3.25)

We summarize the results of these calculations.

\textbf{Theorem 3.2} [91]. \textit{Under the assumption of the convexity bound $\nabla^2 H \geq K$ on the Hamiltonian for some positive constant $K$, the measure $\mu = e^{-H}/Z$ satisfies the logarithmic Sobolev inequality}

$$S(f) \leq \frac{4}{K} D(\sqrt{f}) \quad \text{for any density } f \text{ with } \int f \, d\mu = 1,$$

(3.25)

and the dynamics (3.17) relaxes to equilibrium on the time scale $t \sim 1/K$ in the sense of both the entropy and the Dirichlet form:

$$S(f_t) \leq e^{-tK} S(f_0), \quad D(h \sqrt{f_t}) \leq \frac{2}{t} e^{-tK/2} S(f_0).$$

(3.26)

\textit{Technical remark}. In our applications the dynamics will be restricted to the subset $\Sigma_N = \{x : x_1 < \cdots < x_N\}$, and thus we need to check that the boundary term in the integration by parts

$$\int_{\partial \Sigma} \partial_i h \partial^2 h e^{-H} \, dx = 0$$

(3.27)
(from the third line to the fourth line in (3.19)) vanishes. The role of $\mathcal{H}$ will be played by $N\mathcal{H}_N$, where $\mathcal{H}_N$ is defined in (3.5). Although the density function of the measure $e^{-N\mathcal{H}_N}$ behaves like $(x_{i+1} - x_i)^\beta$ near the $x_{i+1} = x_i$ component of the boundary, and hence it vanishes at the boundary, the function $f_t$ is the solution to a parabolic equation with a singular drift, so in principle it may blow up at the boundary. A further complication is the fact that $h = \sqrt{f}$, and the derivative of the square root is singular. Nevertheless, by using parabolic regularity theory and cutoff functions properly subordinated to the geometry of the set $\Sigma$, we can prove that the integral in (3.27) vanishes. This is one reason why the restriction $\beta \geq 1$ is necessary. For details, see Appendix B of [6].

3.2.2. Universality of the gap distribution of the Dyson Brownian motion for short time. With the convexity bound (3.13), for the Dyson Brownian motion the Bakry–Emery method guarantees that $\mu$ satisfies the logarithmic Sobolev inequality, and the relaxation time to equilibrium is of order 1.

The following result is the main theorem in $\S$ 3.2. It shows that the relaxation time is in fact much shorter than order 1, at least locally and for observables that depend only on the eigenvalue differences.

**Theorem 3.3** (universality of the Dyson Brownian motion for short time [6], Theorem 4.1). Suppose that the Hamiltonian $\mathcal{H}$ given in (3.5) satisfies the convexity bound (3.6) with $\beta \geq 1$. Let $f_t$ be the solution of the forward equation (3.1), so that after time $t_0 = N^{-2a}$ it satisfies the inequality $S_\mu(f_{t_0}) := \int f_{t_0}(\log f_{t_0}) \, d\mu \leq CN^m$ for some fixed $m$. Let

$$Q := \sup_{t \geq t_0} \sum_j \int (x_j - \gamma_j)^2 f_t \, d\mu$$

and assume that $Q \leq CN^m$ for some exponent $m$. Fix $n \geq 1$ and an array $m = (m_1, \ldots, m_n) \in \mathbb{N}_1^n$ of increasing positive integers. Let $G: \mathbb{R}^n \to \mathbb{R}$ be a bounded smooth function with compact support and let

$$\mathcal{G}_{i,m}(x) := G(N(x_i - x_{i+m_1}), N(x_{i+m_1} - x_{i+m_2}), \ldots, N(x_{i+m_{n-1}} - x_{i+m_n})).$$

Then for any sufficiently small $\varepsilon' > 0$, there exist constants $C, c > 0$ depending only on $\varepsilon'$ and $G$ such that for any $J \subset \{1, 2, \ldots, N - m_n\}$ and any $\tau > 3t_0 = 3N^{-2a}$

$$\left| \int \frac{1}{N} \sum_{i \in J} \mathcal{G}_{i,m}(x) f_\tau \, d\mu - \int \frac{1}{N} \sum_{i \in J} \mathcal{G}_{i,m}(x) \, d\mu \right| \leq CN^{\varepsilon'} \sqrt{|J|Q(N\tau)^{-1}} + Ce^{-cN\varepsilon'}.$$  

(3.30)

In $\S$ 3.2.3 we explain the intuitive idea behind the proof. The precise proof will be given in $\S$ 3.2.4.

3.2.3. Intuitive proof of Theorem 3.3. The key idea is that we can ‘speed up’ the convergence to local equilibrium if we modify the dynamics by adding an auxiliary potential $W(x)$ to the Hamiltonian. It will have the form

$$W(x) := \sum_{j=1}^N W_j(x_j), \quad W_j(x) := \frac{1}{2R^2} (x_j - \gamma_j)^2,$$  

(3.31)
that is, it is a quadratic potential of order $R$ confining each eigenvalue close to its classical location, and we define

$$
\widetilde{\mathcal{H}} := \mathcal{H} + W.
$$

The new measure is denoted by

$$
d\omega := \omega(x) \, dx, \quad \omega := e^{-N\widetilde{\mathcal{H}}/Z},
$$

and will be called the pseudo-equilibrium measure (with a slight abuse of notation we will denote by $\omega$ both the measure and its density function with respect to Lebesgue measure). The corresponding generator is denoted by $\widetilde{L}$. We will typically choose $R \ll 1$, so that the additional term $W$ substantially increases the lower bound in (3.13) on the Hessian, hence speeding up the relaxation time of the dynamics to equilibrium from order $O(1)$ to order $O(R^2)$. This is the first step of the proof, and it will be formulated in Theorem 3.4, whose proof basically follows the Bakry–Emery argument in §3.2.1.

In the second step we consider an arbitrary probability measure of the form $q\omega$ with some function $q$, and we control the difference of the expectation values

$$
\int Gq \, d\omega - \int G \, d\omega
$$

of the observable

$$
G := \frac{1}{N} \sum_{i \in J} q_{i,m}
$$

in terms of the entropy and the Dirichlet form of $q$ with respect to $\omega$. Eventually this will enable us to compare the expectations,

$$
\int Gq \, d\omega - \int Gq' \, d\omega,
$$

for any two measures $q\omega$ and $q'\omega$, in particular, for the measures $f_\tau \mu$ and $\mu$, which can be written in this form by defining $q = f_\tau \mu / \omega$ and $q' = \mu / \omega$.

Here we are faced with an $N$-problem: both the entropy and the Dirichlet form are extensive quantities. A direct application of the entropy inequality (3.16) to (3.33) would estimate the observable $G$, an order-1 quantity, by a quantity of order $O(\sqrt{N})$. Instead, we can run the new dynamics up to some time $\tau$ and write

$$
\int Gq \, d\omega - \int G \, d\omega = \int G(q - q_\tau) \, d\omega + \int G(q_\tau - 1) \, d\omega.
$$

If $\tau$ is larger than the relaxation time of the new dynamics, then the second term is exponentially small by the entropy inequality, and this exponential smallness suppresses the $N$-problem.

To estimate the first term, we want to capitalize on the fact that $\tau$ is small. By integrating back the time derivative, $q_\tau - q = \int_0^\tau \partial_t q_t \, dt$, we could extract a factor proportional to $\tau$, but after using the equality $\partial_t q_t = \widetilde{L} q_t$ and integrating by parts we will have to differentiate the observable, which brings in an additional factor $N$ due
to its scaling. It seems that this method estimates a quantity of order 1 by a quantity of order $N$. However, we have proved a new estimate (see (3.45) below) that controls $\int \mathbf{G}(q-q_\tau) \, d\omega$ by $(\tau D_\omega(\sqrt{q})/N)^{1/2}$; note the additional factor $1/N$. The key reason for this improvement is that the dynamics relaxes to equilibrium much faster in certain directions, namely, for observables depending only on differences of the $x_i$. To extract this mechanism, we use the fact that the lower bound in (3.6) on the Hessian is of order $N$ in the difference variables $v_i - v_j$, and this estimate can be used to gain an additional $N$-factor; this is the content of Theorem 3.5.

The estimate will have a free parameter $\tau$ that can be optimized. This parameter stems from the method of the proof: we prove a time-independent inequality by a dynamical method, that is, we run the flow up to some time $\tau$ and we estimate the differences $q - q_\tau$ and $q_\tau - q_\infty$ differently.

Finally, in the third step we have to compare the original dynamics with the new one in the sense of entropy and Dirichlet form, since $D_\omega(\sqrt{f_\tau \mu/\omega})$ and $S_\omega(f_\tau \mu/\omega)$ must be computed for the estimates in the second step. These would be given by the standard decay to the equilibrium estimates (3.26) if $f_\tau \mu/\omega$ were evolving with respect to the modified dynamics but $f_\tau$ evolves by the original dynamics. We thus need to show that the error due to the modification of the dynamics by adding $W$ is negligible.

It turns out that the key quantity that determines how much error was made by the additional potential is the $H^1$ norm of $W$, that is,

$$\Lambda_t := \int (\nabla W)^2 f_t \, d\mu.$$ 

Due to the explicit form of $W$, we have

$$\Lambda_t = R^{-4} \sum_i (x_i - \gamma_i)^2 f_t \, d\mu \leq CN^{-2a} R^{-4}, \quad (3.34)$$

using Condition III (3.9). Given $a > 0$, we can therefore choose an $R \ll 1$ so that we still have $\Lambda \ll 1$. This will complete the proof.

Note that the speed of convergence is determined by the second derivative of the auxiliary potential, while the modification in the Dirichlet form and the entropy is determined by $(\nabla W)^2$. So one can speed up the dynamics and still compare the Dirichlet forms and entropies of the two equilibrium measures if a strong a priori bound (3.34) on $\Lambda$ is given. This is one of the reasons why the method works.

The other key observation is the effective use of the convexity bound in (3.6), which exploits a crucial property of the dynamics of the Dyson Brownian motion (1.64). The logarithmic interaction potential gives rise to a singular force

$$F(x_i) = -\frac{1}{4} x_i - \frac{1}{2N} \sum_{j \neq i} \frac{1}{x_i - x_j}, \quad (3.35)$$

acting on the $i$th particle. Formally, $F(x_i)$ is a mean field force, and if the $x_j$ were distributed according to the semicircle law, then the bulk of the sum would cancel the $-x_i/4$ term. However, the effect of the neighbouring particles, with $j = i \pm 1$, is huge: they exert a force of order 1 on $x_i$. Such a force may move the particle $x_i$
by a distance of order $1/N$ within a very short time of order $1/N$. Note that by the order-preserving property of the dynamics, an interval of size $O(1/N)$ is roughly the whole space available for $x_i$, at least in the bulk. Thus, $x_i$ is likely to relax to its equilibrium within a time scale of order $1/N$ due to the strong repulsive force from its neighbours. Of course this argument is not a proof, since the other particles move as well. However, our observables involve only eigenvalue differences and in the difference coordinates the strong drift is persistently present. This indicates that the eigenvalue differences may relax to their local equilibrium on a time scale almost of order $1/N$.

For other observables the relaxation time is necessary longer. In particular, it can happen that there is no precise cancellation from the bulk in (3.35), in which case the neighbouring particles all feel the same mean field drift and will move collectively. In fact, if the initial density profile differs substantially from the semicircle density, then the relaxation to the semicircle law may even take order-1 time (although such a scenario is excluded in our case by (3.9)).

3.2.4. Detailed proof of Theorem 3.3. Every constant in this proof depends on $\varepsilon'$ and $G$, and we will not indicate the precise dependence. Given $\tau > 0$, we define $R := \tau^{1/2} N^{-\varepsilon'/2}$.

We now introduce the pseudo-equilibrium measure $\omega_N = \omega = \psi\mu$ defined by

$$
\psi := Z \exp(-NW),
$$

where $Z$ is chosen such that $\omega$ is a probability measure; in particular, $\omega = e^{-N\mathcal{H}/Z}$ with

$$
\mathcal{H} = \mathcal{H} + W.
$$

The potential $W$ was defined in (3.31) and it confines the $j$th point $x_j$ to be near its classical location $\gamma_j$.

The local relaxation flow is defined to be the reversible dynamics with respect to $\omega$. This dynamics is described by the generator $\tilde{L}$ defined by

$$
\int f \tilde{L} g d\omega = -\frac{1}{2N} \sum_j \int (\partial_j f)(\partial_j g) d\omega.
$$

Explicitly, $\tilde{L}$ is given by

$$
\tilde{L} = L - \sum_j b_j \partial_j, \quad b_j = W'_j(x_j) = \frac{x_j - \gamma_j}{R^2}.
$$

Since the additional potential $W_j$ is uniformly convex with

$$
\inf_j \inf_{x \in \mathbb{R}} W''_j(x) \geq R^{-2},
$$

we get from (3.6) and the condition $\beta \geq 1$ that

$$
\langle \mathbf{v}, \nabla^2 \mathcal{H}(\mathbf{x}) \mathbf{v} \rangle \geq \frac{1}{R^2} \|\mathbf{v}\|^2 + \frac{1}{N} \sum_{i<j} \frac{(v_i - v_j)^2}{(x_i - x_j)^2}, \quad \mathbf{v} \in \mathbb{R}^N.
$$
The $R^{-2}$ in the first term comes from the additional convexity of the local interaction, and it enhances the local Dirichlet form dissipation. In particular, we have the uniform lower bound

$$\nabla^2 \tilde{\mathcal{E}} = \text{Hess}(-\log \omega) \geq R^{-2}.$$  

This guarantees that the relaxation time to the equilibrium $\omega$ for the $\tilde{L}$-dynamics is bounded above by $CR^2$.

The first ingredient in proving Theorem 3.3 is an analysis of the local relaxation flow, which satisfies the logarithmic Sobolev inequality and the following dissipation estimate.

**Theorem 3.4.** Suppose that (3.39) holds. Consider the forward equation

$$\partial_t q_t = \tilde{L} q_t, \quad t \geq 0,$$

with initial condition $q_0 = q$ and with reversible measure $\omega$. Assume that $q \in L^\infty(d\omega)$. Then the following estimates hold:

$$\partial_t D_\omega(\sqrt{q_t}) \leq -\frac{1}{R^2} D_\omega(\sqrt{q_t}) - \frac{1}{2N^2} \int \sum_{i,j=1}^N \frac{(\partial_i \sqrt{q_t} - \partial_j \sqrt{q_t})^2}{(x_i - x_j)^2} \, d\omega,$$

$$\frac{1}{2N^2} \int_0^\infty ds \int \sum_{i,j=1}^N \frac{(\partial_i \sqrt{q_s} - \partial_j \sqrt{q_s})^2}{(x_i - x_j)^2} \, d\omega \leq D_\omega(\sqrt{q}),$$

together with the logarithmic Sobolev inequality

$$S_\omega(q) \leq CR^2 D_\omega(\sqrt{q}),$$

where $C$ is a universal constant. Thus, the time to equilibrium is of order $R^2$:

$$S_\omega(q_t) \leq e^{-Ct/R^2} S_\omega(q).$$

**Proof.** This theorem can be proved following the standard argument presented in §3.2.1. The key additional input is the convexity bound in (3.6), which gives rise to the second term on the right-hand side of (3.41) from the last line of (3.19). In particular, this serves as an upper bound on $\partial_t D_\omega(\sqrt{q_t})$, and thus by integrating (3.41) one obtains (3.42).

The estimate (3.42) of the second term in (3.39) plays a key role in the next theorem.

**Theorem 3.5.** Suppose that Condition I holds and let $q \in L^\infty$ be a density, $\int q \, d\omega = 1$. Fix $n \geq 1$, let $m \in \mathcal{A}_+^n$, let $G: \mathbb{R}^n \to \mathbb{R}$ be a bounded smooth function with compact support, and recall the definition of $\mathcal{G}_{i,m}$ in (3.29). Then for any $J \subset \{1,2,\ldots,N-m_n\}$ and any $\tau > 0$

$$\left| \int \frac{1}{N} \sum_{i \in J} \mathcal{G}_{i,m}(x) q \, d\omega - \int \frac{1}{N} \sum_{i \in J} \mathcal{G}_{i,m}(x) \, d\omega \right| \leq C \left( \frac{|J|D_\omega(\sqrt{q}) \tau}{N^2} \right)^{1/2} + C \sqrt{S_\omega(q)} e^{-\epsilon \tau / R^2}.$$  

(3.45)
Proof. For simplicity we will consider the case when $m = (1, 2, \ldots, n)$; the general case follows easily by appropriately redefining the function $G$. Let $q_t$ satisfy

$$\partial_t q_t = \tilde{L} q_t, \quad t \geq 0,$$

with an initial condition $q_0 = q$. We write

$$\int \left[ \frac{1}{N} \sum_{i \in J} g_{i,m} \right] (q-1) \, d\omega = \int \left[ \frac{1}{N} \sum_{i \in J} g_{i,m} \right] (q-q_{\tau}) \, d\omega + \int \left[ \frac{1}{N} \sum_{i \in J} g_{i,m} \right] (q_{\tau}-1) \, d\omega. \tag{3.46}$$

The second term can be estimated by the entropy inequality (3.16), the decay of the entropy to equilibrium (3.44), and the boundedness of $G$. This gives the second term in (3.45).

To estimate the first term in (3.46), we get the following by differentiation, using the equation $\partial q_t = e^{Lq_t}$ and (3.36):

$$\int \frac{1}{N} \sum_i g_{i,m}(x) q_{\tau} \, d\omega - \int \frac{1}{N} \sum_i g_{i,m}(x) q_0 \, d\omega = \int_0^\tau ds \int \frac{1}{N} \sum_i \sum_{k=1}^n \partial_k G \left( N(x_i - x_{i+1}), \ldots, N(x_{i+n-1} - x_{i+n}) \right) \times [\partial_{i+k-1} q_s - \partial_{i+k} q_s] \, d\omega.$$

From the Cauchy–Schwarz inequality and the equality $\partial q = 2\sqrt{q} \partial\sqrt{q}$, the last term is bounded by

$$2 \sum_{k=1}^n \left[ \int_0^\tau ds \int \sum_i \left[ \partial_k G \left( N(x_i - x_{i+1}), \ldots, N(x_{i+n-1} - x_{i+n}) \right) \right]^2 \right]^{1/2} \times \left( x_{i+k-1} - x_{i+k} \right)^2 q_s \, d\omega \right]^{1/2} \times \left[ \int_0^\tau ds \int \frac{1}{N^2} \sum_i \frac{1}{(x_{i+k-1} - x_{i+k})^2} \left[ \partial_{i+k-1} \sqrt{q_s} - \partial_{i+k} \sqrt{q_s} \right]^2 \, d\omega \right]^{1/2} \leq C \left( \frac{|J|D \omega \sqrt{q_0} \tau}{N^2} \right)^{1/2}, \tag{3.47}$$

where we have used (3.42) and the fact that

$$\left[ \partial_k G \left( N(x_i - x_{i+1}), \ldots, N(x_{i+k-1} - x_{i+k}), \ldots, N(x_{i+n-1} - x_{i+n}) \right) \right]^2 \times (x_{i+k-1} - x_{i+k})^2 \leq CN^{-2},$$

since $G$ is smooth and compactly supported.

For comparison with Theorem 3.5 we state the following result, which can be proved in a similar way.
Lemma 3.6. Let $G: \mathbb{R} \to \mathbb{R}$ be a bounded smooth function with compact support and let a sequence $E_i$ be fixed. Then for any $\tau > 0$
\[
\left| \frac{1}{N} \sum_i \int G(N(x_i - E_i)) q \, d\omega - \frac{1}{N} \sum_i \int G(N(x_i - E_i)) \, d\omega \right| \leq C \sqrt{S_\omega(q)} \tau + C \sqrt{S_\omega(q)} e^{-c\tau/R^2}.
\] (3.48)

Note that by exploiting the local Dirichlet form dissipation coming from the second term on the right-hand side of (3.41), we have gained the crucial factor $N^{-1/2}$ in the estimate (3.45) compared with (3.48).

The final ingredient in proving Theorem 3.3 is provided by the following entropy and Dirichlet form estimates.

Theorem 3.7. Suppose that (3.6) holds and recall that $\tau = R^2 N^{t_0} \geq 3t_0$ with $t_0 = N^{-2a}$. Assume that $S_\mu(f_t) \leq CN^m$ with some fixed $m$. Let $g_t := f_t/\psi_t$. Then the entropy and the Dirichlet form satisfy the estimates
\[
S_\omega(g_{\tau/2}) \leq CNR^{-2}Q, \quad D_\omega(\sqrt{g}) \leq CNR^{-4}Q.
\] (3.49)

Proof. Recall that $\partial_t f_t = Lf_t$. The standard estimate of the entropy of $f_t$ with respect to the invariant measure is obtained by differentiating the entropy twice and using the logarithmic Sobolev inequality (see §3.2.1). The entropy and the Dirichlet form in (3.49) are, however, computed with respect to the measure $\omega$. This yields the additional second term in (3.18), and we use the following identity [106] which is valid for any probability density $\psi_t$:
\[
\partial_t S_\mu(f_t|\psi_t) = -\frac{2}{N} \sum_j \int (\partial_j \sqrt{g_t})^2 \psi_t \, d\mu + \int g_t(L - \partial_t)\psi_t \, d\mu,
\]
where $g_t := f_t/\psi_t$ and
\[
S_\mu(f_t|\psi_t) := \int f_t \log \frac{f_t}{\psi_t} \, d\mu
\]
is the relative entropy.

In our applications we take $\psi_t$ to be time independent, $\psi_t = \psi = \omega/\mu$, so that $S_\mu(f_t|\psi) = S_\omega(g_t)$. By (3.37), we have
\[
\partial_t S_\omega(g_t) = -\frac{2}{N} \sum_j \int (\partial_j \sqrt{g_t})^2 \, d\omega + \int \tilde{L} g_t \, d\omega + \sum_j \int b_j \partial_j g_t \, d\omega.
\]
Since $\omega$ is $\tilde{L}$-invariant, the middle term on the right-hand side vanishes, and from the Cauchy–Schwarz inequality we find that
\[
\partial_t S_\omega(g_t) \leq -D_\omega(\sqrt{g_t}) + CN \sum_j \int b_j^2 g_t \, d\omega \leq -D_\omega(\sqrt{g_t}) + CN \Lambda,
\] (3.50)

where we have defined
\[
\Lambda := QR^{-4} = \sup_{t \geq 0} R^{-4} \sum_j \int (x_j - \gamma_j)^2 f_t \, d\mu.
\] (3.51)
Together with (3.43), we have
\[ \partial_t S_\omega(g_t) \leq -CR^{-2}S_\omega(g_t) + CNA, \quad t \geq N^{-2a}. \] (3.52)

To obtain the first inequality in (3.49), we integrate (3.52) from \( t_0 = N^{-2a} \) to \( \tau/2 \), using the fact that \( \tau = R^2 N^{\varepsilon'} \) and \( S_\omega(g_t) \leq C N^m + N^2 Q \) with some finite \( m \) depending on \( a \). This a priori bound follows from the inequality
\[ S_\omega(g_t_0) = S_\mu(f_{t_0} | \psi) = S_\mu(f_{t_0}) - \log Z + \log \tilde{Z} + N \int f_{t_0} W \, d\mu \leq C N^m + N^2 Q, \] (3.53)
where we used the facts that \( |\log Z| \leq C N^m \) and \( |\log \tilde{Z}| \leq C N^m \), which can easily be checked. The second inequality in (3.49) can be obtained from the first one by integrating (3.50) from \( t = \tau/2 \) to \( t = \tau \) and using the monotonicity of the Dirichlet form with respect to time.

Finally, we complete the proof of Theorem 3.3. Recall that \( \tau = R^2 N^{\varepsilon'} \) and \( t_0 = N^{-2a} \). Choose \( q_\tau := g_\tau = f_\tau/\psi \) to be the density \( q \) in Theorem 3.5. The condition \( q_\tau \in L^\infty \) can be guaranteed by an approximation argument. Then Theorem 3.7 and Theorem 3.5 together with (3.53) and the equality \( \Lambda \tau = Q \tau^{-1} N^{2\varepsilon'} \) directly imply that
\[ \left| \int \frac{1}{N} \sum_{i \in J} \mathcal{G}_{i,m} f_\tau \, d\mu - \int \frac{1}{N} \sum_{i \in J} \mathcal{G}_{i,m} \, d\omega \right| \leq C N^{\varepsilon'} \sqrt{|J| Q (\tau N)^{-1}} + C e^{-c N^{\varepsilon'}}, \] (3.54)
that is, the local statistics for \( f_\tau \mu \) and \( \omega \) can be compared. Clearly, (3.54) also holds for the special choice \( f_0 = 1 \) (for which \( f_\tau = 1 \)), that is, the local statistics for \( \mu \) and \( \omega \) can also be compared. This completes the proof of Theorem 3.3.

3.3. From gap distribution to correlation functions: Sketch of the proof of Theorem 3.1. Our main result Theorem 3.1 will follow from Theorem 3.3 and from the fact that in the case when \( \tau \geq N^{-2a + \delta} \), the assumption (3.9) guarantees that
\[ N^{\varepsilon'} \sqrt{|J| Q (\tau N)^{-1}} \leq N^{\varepsilon' - \delta/2} = N^{-\delta/6} \to 0 \]
with the choice \( \epsilon' = \delta/3 \) and using the fact that \( |J| \leq N \). Therefore, the local statistics of observables involving eigenvalue differences coincide in the limit as \( N \to \infty \).

To complete the proof of Theorem 3.1, we will have to show that the convergence of the observables \( \mathcal{G}_{i,m} \) is sufficient to identify the correlation functions of the \( x_i \) in the sense described in Theorem 3.1. This is a fairly standard technical argument, and the details will be given in Appendix B. Here we just summarize the main points.

Theorem 3.3 detects the joint behaviour of the eigenvalue differences on the correct scale \( 1/N \), due to the factor \( N \) in the argument of \( \mathcal{G} \) in (3.29). The slight technical problem is that the observable (3.29), and its averaged version (3.30), involve fixed indices of eigenvalues, while the correlation functions involve cumulative statistics.
To understand this subtlety, consider the case $n = 1$ for simplicity and let $m_1 = 1$, say. The observable (3.30) answers the question as to what the empirical distribution is of the differences between consecutive eigenvalues; in other words, (3.30) directly identifies the gap distribution defined in §1.5.1. The correlation functions answer the question as to what the probability is that there are two eigenvalues at a fixed distance away from each other; in other words, they are not directly sensitive to possible other eigenvalues in between. Of course, these two questions are closely related and it is easy to deduce the answers from each other. This is exactly what the calculation (1.38) has achieved in one direction, and now we need to go in the other direction: identify the correlation functions from the (generalized) gap distributions.

In fact, this direction is easier, and the essence is given by the following formula:

$$\int_{E-b}^{E+b} \frac{dE'}{2b} \int_{\mathbb{R}^n} d\alpha_1 \cdots d\alpha_n O(\alpha_1, \ldots, \alpha_n) p_{\tau, N}^{(n)} \left( E' + \frac{\alpha_1}{N\rho(E)}, \ldots, E' + \frac{\alpha_n}{N\rho(E)} \right)$$

$$= C_{N,n} \int_{E-b}^{E+b} \frac{dE'}{2b} \int \sum_{i_1 \neq i_2 \neq \cdots \neq i_n} \tilde{O}(N(x_{i_1} - E'), N(x_{i_1} - x_{i_2}), \ldots, N(x_{i_{n-1}} - x_{i_n})) f_\tau d\mu$$

$$= C_{N,n} \int_{E-b}^{E+b} \frac{dE'}{2b} \int \sum_{m \in S_n} \sum_{i=1}^{N} Y_{i,m}(E', x) f_\tau d\mu,$$

(3.55)

with $C_{N,n} := N^n (N - n)! / N! = 1 + O_n(N^{-1})$, where we let $S_n$ denote the set of tuples of increasing positive integers, $m = (m_2, m_3, \ldots, m_n) \in \mathbb{N}_{m_2}^{n-1}$, $m_2 < m_3 < \cdots < m_n$, and we defined the quantities

$$Y_{i,m}(E', x) := \tilde{O}(N(x_i - E'), N(x_i - x_{i+m_2}), \ldots, N(x_i - x_{i+m_n})),$$

$$\tilde{O}(u_1, u_2, \ldots, u_n) := O(\varrho(E)u_1, \varrho(E)(u_1 - u_2), \ldots).$$

(3.56)

We also let $Y_{i,m} = 0$ for $i + m_n > N$. The first equality in (3.55) is just the definition of the correlation function after a trivial rescaling. In going from the second to the third line, note first that by permutational symmetry of $p_{\tau, N}^{(n)}$ we can assume that $O$ is symmetric, and thus we can restrict the summation to $i_1 < i_2 < \cdots < i_n$ with an overall factor of $n!$. Then we changed the indices $i = i_1$, $i_2 = i + m_2$, $i_3 = i + m_3$, \ldots and performed a resummation over all differences of indices in $m$. Apart from the first variable $N(x_{i_1} - E')$, the function $Y_{i,m}$ is of the form (3.29), so Theorem 3.3 will apply. The dependence on the first variable will be negligible after the $dE'$ integration on a macroscopic interval.

To control the error terms in this argument, and especially to show that even the error terms in the potentially infinite sum over $m \in S_n$ converge, one needs an a priori bound on the local density. This is where Condition IV (3.10) is used. For details, see Appendix B.

### 4. The Green function comparison theorems

A simplified version of the Green function comparison theorem was already stated in Theorem 1.6. Here we state the complete version, Theorem 4.1. It will lead
quickly to Theorem 4.2 stating that the correlation functions of the eigenvalues of two matrix ensembles are identical on the scale $1/N$ provided that the first four moments of all matrix elements of these two ensembles are almost identical. Here we do not assume that the real and imaginary parts are i.i.d. variables, hence the $k$th moment of $h_{ij}$ is understood as the collection of numbers $\int h^s h^{k-s} \nu_{ij}(dh)$, $s = 0,1,\ldots,k$. The related Theorem 1.7 (proved in [17]) compares the joint distributions of the individual eigenvalues—which is not covered by our Theorem 4.1—but it does not address directly the matrix elements of the Green functions. In § 4.3 we will sketch some ideas in the proof of Theorem 1.7 to point out the differences between the two results. The key input for both theorems is the local semicircle law on the almost optimal scale $N^{-1+\varepsilon}$. The eigenvalue perturbation used in Theorem 1.7 requires certain estimates on the level repulsion. At the same time, the proof of Theorem 4.1 is a straightforward perturbation theory argument for the resolvent.

**Theorem 4.1** (Green function comparison [9], Theorem 2.3). Let $H^{(v)}$ and $H^{(w)}$ be two generalized $N \times N$ Wigner matrices with matrix elements $h_{ij}$ given by the random variables $N^{-1/2}v_{ij}$ and $N^{-1/2}w_{ij}$, respectively, where $v_{ij}$ and $w_{ij}$ satisfy the uniform subexponential decay condition

$$P(|v_{ij}| \geq x) \leq C \exp(-x^\vartheta), \quad P(|w_{ij}| \geq x) \leq C \exp(-x^\vartheta)$$

for some $C$, $\vartheta > 0$. Fix a bijective ordering map

$$\phi: \{(i,j) : 1 \leq i \leq j \leq N\} \to \{1,\ldots,\gamma(N)\}, \quad \gamma(N) := \frac{N(N+1)}{2},$$

on the index set of the independent matrix elements, and denote by $H_\gamma$ the general-
ized Wigner matrix whose matrix elements $h_{ij}$ have the $v$-distribution if $\phi(i,j) \leq \gamma$ and the $w$-distribution otherwise. In particular, $H^{(v)} = H_0$ and $H^{(w)} = H_{\gamma(N)}$. Let $\kappa > 0$ be arbitrary and suppose that, for any small parameter $\tau > 0$ and for any $y \geq N^{-1+\tau}$, the following estimate holds for the diagonal elements of the resolvent:

$$P(\max_{0 \leq \gamma \leq \gamma(N)} \max_{1 \leq k \leq N} \max_{|E| \leq 2-\kappa} \left| \left( \frac{1}{H_\gamma - E - i\gamma} \right)_{kk} \right| \leq N^2\tau) \geq 1 - CN^{-c \log \log N} \quad (4.1)$$

with some constants $C$ and $c$ depending only on $\tau$ and $\kappa$. Moreover, assume that the first three moments of $v_{ij}$ and $w_{ij}$ are the same, that is,

$$E\tilde{v}_{ij}^s v_{ij}^u = E\tilde{w}_{ij}^s w_{ij}^u, \quad 0 \leq s + u \leq 3,$$

and the difference between the fourth moments of $v_{ij}$ and $w_{ij}$ is much less than 1, say

$$|E\tilde{v}_{ij}^s v_{ij}^{4-s} - E\tilde{w}_{ij}^s w_{ij}^{4-s}| \leq N^{-\delta}, \quad s = 0,1,2,3,4, \quad (4.2)$$

for some given $\delta > 0$. Let $\varepsilon > 0$ be arbitrary and choose an $\eta$ with $N^{-1-\varepsilon} \leq \eta \leq N^{-1}$. For any sequence of positive integers $k_1,\ldots,k_n$ consider complex parameters

$$z_j^m = E_j^m \pm i\eta, \quad j = 1,\ldots,k_m, \quad m = 1,\ldots,n, \quad |E_j^m| \leq 2 - 2\kappa$$

and with an arbitrary choice of the signs $\pm$. Let $G^{(v)}(z) = (H^{(v)} - z)^{-1}$ denote the resolvent
and let $F(x_1, \ldots, x_n)$ be a function such that for any multi-index $\alpha = (\alpha_1, \ldots, \alpha_n)$ with $1 \leq |\alpha| \leq 5$ and for any sufficiently small $\varepsilon' > 0$ the inequalities

$$\max\left\{ |\partial^\alpha F(x_1, \ldots, x_n)| : \max_j |x_j| \leq N^{\varepsilon'} \right\} \leq N^{C_0 \varepsilon'}$$

(4.3)

and

$$\max\left\{ |\partial^\alpha F(x_1, \ldots, x_n)| : \max_j |x_j| \leq N^{2} \right\} \leq N^{C_0}$$

(4.4)

hold for some constant $C_0$.

Then there is a constant $C_1$ depending on $\vartheta, \sum_m k_m$, and $C_0$ such that for any $\eta$ with $N^{-1-\varepsilon} \leq \eta \leq N^{-1}$ and for any choice of signs of the imaginary parts of the parameters $z_j^m$

$$\left| \mathbb{E} F\left( \frac{1}{N} \text{Tr} \left[ \prod_{j=1}^{k_1} G^{(v)}(z_j^1) \right], \ldots, \frac{1}{N} \text{Tr} \left[ \prod_{j=1}^{k_n} G^{(v)}(z_j^n) \right] \right) - \mathbb{E} F(G^{(v)} \rightarrow G^{(w)}) \right|$$

$$\leq C_1 N^{-1/2+\varepsilon} + C_1 N^{-\delta+\varepsilon},$$

(4.5)

where in the second term in the argument of $F$ the Green functions of $H^{(v)}$ are replaced by those of $H^{(w)}$ and all the other parameters remain unchanged.

**Remark 3.** We formulated Theorem 4.1 for functions of the traces of monomials of the Green functions because this is the form we need in the applications. However, the result (and the proof we are going to present) also holds directly for matrix elements of monomials of the Green functions (for the precise statement, see [9]). We also remark that Theorem 4.1 holds for generalized Wigner matrices if $C_{\text{sup}} = \sup_{i,j} N \sigma_{i,j}^2 < \infty$. The positive lower bound on the variances, that is, the condition $C_{\text{inf}} > 0$ in (1.17), is not necessary for this theorem.

**Remark 4.** Although we state Theorem 4.1 for Hermitian and symmetric ensembles, similar results hold for real and complex sample covariance ensembles, with obvious modifications in the proof.

The following result is the main corollary of Theorem 4.1 and will be proved later in this section. The important statement is Theorem 4.1, the proof of its corollary is a fairly straightforward technicality.

**Theorem 4.2** (correlation function comparison [9], Theorem 6.4). Suppose that the assumptions of Theorem 4.1 hold. Let $p_{v,N}^{(k)}$ and $p_{w,N}^{(k)}$ be the $k$-point functions of the eigenvalues with respect to the probability laws of the matrices $H^{(v)}$ and $H^{(w)}$, respectively. Then for any $|E| < 2$, any $k \geq 1$, and any compactly supported continuous test function $O: \mathbb{R}^k \rightarrow \mathbb{R}$

$$\int_{\mathbb{R}^k} d\alpha_1 \cdots d\alpha_k O(\alpha_1, \ldots, \alpha_k) \left( p_{v,N}^{(k)} - p_{w,N}^{(k)} \right) \left( E + \frac{\alpha_1}{N}, \ldots, E + \frac{\alpha_k}{N} \right) = 0.$$
4.1. Proof of the Green function comparison Theorem 4.1. The basic idea is that we estimate the effect of changing matrix elements of the resolvent one by one with the help of a resolvent expansion. Since each matrix element has a typical size of $N^{-1/2}$, and the resolvents are essentially bounded due to (4.1), a resolvent expansion up to the fourth order will identify the change of each element with a precision of $O(N^{-5/2})$ (modulo some tiny corrections of order $N^{O(\tau)}$). The expectation values of the terms up to fourth order involve only the first four moments of the single-entry distribution, which can be directly compared. The error terms are negligible even when we sum them up $N^2$ times, corresponding to the number of comparison steps needed to replace all the matrix elements.

To start the detailed proof, we first need an estimate of the type (4.1) for the resolvents of all intermediate matrices. From the trivial estimate comparison steps needed to replace all the matrix elements.

From (4.1) we have the following a priori estimate:

$$\max_{0 \leq \gamma \leq \gamma(N)} \max_{1 \leq k \leq N} \max_{|E| \leq 2^{-\kappa}} \sup_{\eta \geq N^{-1-\epsilon}} \left| \operatorname{Im} \left( \frac{1}{H - E - i\eta} \right) \right| \leq N^{3\tau + \epsilon}$$

Note that the supremum over $\eta$ can be included by establishing the estimate first for a fine grid of the $\eta$ with spacing $N^{-10}$ and then extending the estimate for all $\eta$ by using the fact that the Green functions are Lipschitz continuous with respect to $\eta$ with a Lipschitz constant $\eta^{-2}$.

Let $\lambda_m$ and $u_m$ denote the eigenvalues and eigenvectors of $H_\gamma$. Then by the definition of the Green function we have

$$\left| \left( \frac{1}{H_\gamma - z} \right)_{jk} \right| \leq \sum_{m=1}^{N} \frac{|u_m(j)| |u_m(k)|}{|\lambda_m - z|} \leq \left[ \sum_{m=1}^{N} \frac{|u_m(j)|^2}{|\lambda_m - z|} \right]^{1/2} \left[ \sum_{m=1}^{N} \frac{|u_m(k)|^2}{|\lambda_m - z|} \right]^{1/2}.$$  

We define a dyadic decomposition

$$U_n = \{ m : 2^n - 1 < |\lambda_m - E| < 2^n \}, \quad n = 1, 2, \ldots, n_0 := C \log N,$$

$$U_0 = \{ m : |\lambda_m - E| < \eta \}, \quad U_\infty := \{ m : 2^{n_0} \eta < |\lambda_m - E| \}$$

and break up the summation over $m$ using $\bigcup_n U_n$:

$$\sum_{m=1}^{N} \frac{|u_m(j)|^2}{|\lambda_m - z|} = \sum_{n} \sum_{m \in U_n} \frac{|u_m(j)|^2}{|\lambda_m - z|} \leq C \sum_{n} \sum_{m \in U_n} \operatorname{Im} \left( \frac{1}{H_\gamma - E - i \cdot 2^n \eta} \right)_{jj} \leq C \sum_{n} \operatorname{Im} \left( \frac{1}{H_\gamma - E - i \cdot 2^n \eta} \right)_{jj}.$$  

Using the estimate (4.1) for $n = 0, 1, \ldots, n_0$ and a trivial bound of $O(1)$ for $n = \infty$, we prove that

$$\max_{0 \leq \gamma \leq \gamma(N)} \max_{1 \leq k, \ell \leq N} \max_{|E| \leq 2^{-\kappa}} \sup_{\eta \geq N^{-1-\epsilon}} \left| \operatorname{Im} \left( \frac{1}{H_\gamma - E - i\eta} \right)_{k\ell} \right| \leq N^{4\tau + \epsilon}$$

$$\max_{0 \leq \gamma \leq \gamma(N)} \max_{1 \leq k, \ell \leq N} \max_{|E| \leq 2^{-\kappa}} \sup_{\eta \geq N^{-1-\epsilon}} \left| \operatorname{Im} \left( \frac{1}{H_\gamma - E - i\eta} \right)_{k\ell} \right| \leq N^{4\tau + \epsilon}$$

$$\max_{0 \leq \gamma \leq \gamma(N)} \max_{1 \leq k, \ell \leq N} \max_{|E| \leq 2^{-\kappa}} \sup_{\eta \geq N^{-1-\epsilon}} \left| \operatorname{Im} \left( \frac{1}{H_\gamma - E - i\eta} \right)_{k\ell} \right| \leq N^{4\tau + \epsilon}$$

$$\max_{0 \leq \gamma \leq \gamma(N)} \max_{1 \leq k, \ell \leq N} \max_{|E| \leq 2^{-\kappa}} \sup_{\eta \geq N^{-1-\epsilon}} \left| \operatorname{Im} \left( \frac{1}{H_\gamma - E - i\eta} \right)_{k\ell} \right| \leq N^{4\tau + \epsilon}$$

$$\max_{0 \leq \gamma \leq \gamma(N)} \max_{1 \leq k, \ell \leq N} \max_{|E| \leq 2^{-\kappa}} \sup_{\eta \geq N^{-1-\epsilon}} \left| \operatorname{Im} \left( \frac{1}{H_\gamma - E - i\eta} \right)_{k\ell} \right| \leq N^{4\tau + \epsilon}$$

$$\max_{0 \leq \gamma \leq \gamma(N)} \max_{1 \leq k, \ell \leq N} \max_{|E| \leq 2^{-\kappa}} \sup_{\eta \geq N^{-1-\epsilon}} \left| \operatorname{Im} \left( \frac{1}{H_\gamma - E - i\eta} \right)_{k\ell} \right| \leq N^{4\tau + \epsilon}$$

$$\max_{0 \leq \gamma \leq \gamma(N)} \max_{1 \leq k, \ell \leq N} \max_{|E| \leq 2^{-\kappa}} \sup_{\eta \geq N^{-1-\epsilon}} \left| \operatorname{Im} \left( \frac{1}{H_\gamma - E - i\eta} \right)_{k\ell} \right| \leq N^{4\tau + \epsilon}$$
Now we turn to the one-by-one replacement. For notational simplicity, we will consider the case when the test function $F$ has only $n = 1$ variable and $k_1 = 1$, that is, we consider the trace of a first-order monomial; the general case follows analogously. Consider the telescopic sum of the differences of the expectations:

$$EF\left(\frac{1}{N} \text{Tr} \frac{1}{H^{(v)} - z}\right) - EF\left(\frac{1}{N} \text{Tr} \frac{1}{H^{(w)} - z}\right) = \gamma^{(N)} \sum_{\gamma=1}^{\gamma^{(N)}} \left[ EF\left(\frac{1}{N} \text{Tr} \frac{1}{H_\gamma - z}\right) - EF\left(\frac{1}{N} \text{Tr} \frac{1}{H_{\gamma-1} - z}\right) \right].$$

(4.10)

Let $E^{(ij)}$ denote the matrix whose elements are zero everywhere except at the $(i, j)$ position, where it is 1, that is, $E^{(ij)}_{k\ell} = \delta_{ik}\delta_{j\ell}$. Fix a $\gamma > 1$, and let $(i, j)$ be determined by $\phi(i, j) = \gamma$. We will compare $H_{\gamma-1}$ with $H_{\gamma}$. Note that these two matrices differ only in the $(i, j)$ and $(j, i)$ matrix elements, and they can be written as

$$H_{\gamma-1} = Q + \frac{1}{\sqrt{N}} V, \quad V := v_{ij} E^{(ij)} + v_{ji} E^{(ji)},$$

$$H_{\gamma} = Q + \frac{1}{\sqrt{N}} W, \quad W := w_{ij} E^{(ij)} + w_{ji} E^{(ji)},$$

with a matrix $Q$ that has a zero matrix element at the $(i, j)$ and $(j, i)$ positions and where we set $v_{ji} := \bar{v}_{ij}$ for $i < j$, and similarly for $w$. The Green functions are defined by

$$R := \frac{1}{Q - z}, \quad S := \frac{1}{H_\gamma - z}.$$

We first assert that the estimate (4.9) holds for the Green function $R$ as well. To see this, we have from the resolvent expansion that

$$R = S + N^{-1/2} SV S + \cdots + N^{-9/5}(SV)^9 S + N^{-5}(SV)^{10} R.$$

Since $V$ has at most two non-zero elements, in the computation of the $(k, \ell)$ matrix element of this matrix identity each term is a finite sum involving matrix elements of $S$ or $R$ and $v_{ij}$, for example,

$$(SVS)_{k\ell} = S_{ki} v_{ij} S_{j\ell} + S_{kj} v_{ji} S_{i\ell}.$$

Using the estimate (4.9) for the $S$ matrix elements, the subexponential decay for $v_{ij}$, and the trivial estimate $|R_{ij}| \leq \eta^{-1}$, we obtain the estimate (4.9) for $R$.

We can now start proving the main result by comparing the resolvents of $H^{(\gamma-1)}$ and $H^{(\gamma)}$ with the resolvent $R$ of the reference matrix $Q$. By the resolvent expansion

$$S = R - N^{-1/2} RV R + N^{-1}(RV)^2 R - N^{-3/2}(RV)^3 R + N^{-2}(RV)^4 R - N^{-5/2}(RV)^5 S,$$

we can write

$$\frac{1}{N} \text{Tr} S = \hat{R} + \xi, \quad \xi := \sum_{m=1}^{4} N^{-m/2} \hat{R}^{(m)} + N^{-5/2} \Omega,$$
where
\[
\hat{R} := \frac{1}{N} \text{Tr} \, R, \quad \hat{R}^{(m)} := (-1)^m \frac{1}{N} \text{Tr}(RV)^m R, \quad \Omega := -\frac{1}{N} \text{Tr}(RV)^5 S.
\]

For each diagonal element in the computation of these traces, the contribution to \( \hat{R}, \hat{R}^{(m)}, \) and \( \Omega \) is a sum of several terms. For example,
\[
\hat{R}^{(2)} = \frac{1}{N} \sum_k [R_{ki} v_{ij} R_{jj} v_{ji} R_{ik} + R_{ki} v_{ij} R_{ji} v_{ij} R_{jk} + R_{kj} v_{ji} R_{ij} v_{ij} R_{ik}],
\]
and similar formulae hold for the other terms. Then
\[
EF\left( \frac{1}{N} \text{Tr} \frac{1}{H_γ - z} \right) = EF(\hat{R} + ξ)
= E[F(\hat{R}) + F'(\hat{R})ξ + F''(\hat{R})\xi^2 + \cdots + F^{(5)}(\hat{R} + ξ')\xi^5]
= \sum_{m=0}^{5} N^{-m/2} E A^{(m)},
\]
where \( ξ' \) is a number between 0 and \( ξ \) depending on \( \hat{R} \) and \( ξ \), and the \( A^{(m)} \) are defined as follows:
\[
A^{(0)} = F(\hat{R}), \quad A^{(1)} = F'(\hat{R}) A^{(1)}, \quad A^{(2)} = F''(\hat{R})(\hat{R}^{(1)})^2 + F'(\hat{R}) A^{(2)},
\]
and similarly for \( A^{(3)} \) and \( A^{(4)} \). Finally,
\[
A^{(5)} = F'(\hat{R}) Ω + F^{(5)}(\hat{R} + ξ')(\hat{R}^{(1)})^5 + \cdots.
\]
The expectation values of the terms \( A^{(m)} \), \( m \leq 4 \), with respect to \( v_{ij} \) are determined by the first four moments of \( v_{ij} \), for example,
\[
E A^{(2)} = F'(\hat{R}) \left[ \frac{1}{N} \sum_k R_{ki} R_{jj} R_{ik} + \cdots \right] E|v_{ij}|^2 + F''(\hat{R}) \left[ \frac{1}{N^2} \sum_{k,ℓ} R_{ki} R_{jℓ} R_{ℓj} R_{ik} + \cdots \right] E|v_{ij}|^2
+ F'(\hat{R}) \left[ \frac{1}{N} \sum_k R_{ki} R_{jℓ} + \cdots \right] E v_{ij}^2
+ F''(\hat{R}) \left[ \frac{1}{N^2} \sum_{k,ℓ} R_{ki} R_{jℓ} R_{ℓi} R_{jk} + \cdots \right] E v_{ij}^2.
\]
Note that the coefficients involve up to four derivatives of \( F \) and normalized sums of matrix elements of \( \hat{R} \). Using the estimate (4.9) for \( R \) and the derivative bounds (4.3) for the typical values of \( \hat{R} \), we see that all these coefficients are bounded by \( N^C(τ + ε) \) with a very high probability, where \( C \) is an explicit constant. We use the bound (4.4) for the extreme values of \( \hat{R} \), but this event has a very small probability.
in view of (4.9). Therefore, the coefficients of the moments \( E \bar{v}_{ij}^s v_{ij}^u \), \( u + s \leq 4 \), in the quantities \( A^{(0)}, \ldots, A^{(4)} \) are essentially bounded, modulo a factor \( N^{C(\tau + \varepsilon)} \). Note that the fourth moment of \( v_{ij} \) appears only in the \( m = 4 \) term, which already has a prefactor \( N^{-2} \) in (4.11). Therefore, to compute the \( m \leq 4 \) terms in (4.11) up to a precision of \( o(N^{-2}) \), it is sufficient to know the first three moments of \( v_{ij} \) exactly and the fourth moment only with a precision \( N^{-\delta} \). If \( \tau \) and \( \varepsilon \) are chosen such that \( C(\tau + \varepsilon) < \delta \), then the discrepancy in the fourth moment is irrelevant.

Finally, we have to estimate the error term \( A^{(5)} \). All terms not containing \( \Omega \) can be dealt with as before: after estimating the derivatives of \( F \) by \( N^{C(\tau + \varepsilon)} \), one can take the expectation with respect to \( v_{ij} \), which is independent of \( R^m \). For the terms involving \( \Omega \) one can argue similarly, by appealing to the fact that the matrix elements of \( S \) are also essentially bounded by \( N^{C(\tau + \varepsilon)} \) (see (4.9)) and that \( v_{ij} \) has subexponential decay. Alternatively, one can use the Hölder inequality to decouple \( S \) from the rest and then use (4.9) directly, for example:

\[
E|F'(\hat{R})\Omega| = \frac{1}{N} E|F'(\hat{R}) \operatorname{Tr}(RV)^5 S| \\
\leq \frac{1}{N} \left[ E(F'(\hat{R}))^2 \operatorname{Tr} S^2 \right]^{1/2} \left[ E \operatorname{Tr}(RV)^5 (VR^*)^5 \right]^{1/2} \leq CN^{-5/2+C(\tau + \varepsilon)}.
\]

We note that exactly the same perturbation expansion holds for the resolvent of \( H_{\gamma-1} \), just with \( v_{ij} \) replaced by \( w_{ij} \) everywhere. By the moment matching condition, the expectation values \( E A^{(m)} \) of terms with \( m \leq 3 \) in (4.11) are identical, and the \( m = 4 \) terms differ by \( N^{-\delta+C(\tau + \varepsilon)} \). Choosing \( \tau = \varepsilon \), we have

\[
EF\left( \frac{1}{N} \operatorname{Tr} \frac{1}{H_{\gamma} - z} \right) - EF\left( \frac{1}{N} \operatorname{Tr} \frac{1}{H_{\gamma-1} - z} \right) \leq CN^{-5/2+C\varepsilon} + CN^{-2-\delta+C\varepsilon}.
\]

After the summation in (4.10) we have thus proved that

\[
EF\left( \frac{1}{N} \operatorname{Tr} \frac{1}{H^{(v)} - z} \right) - EF\left( \frac{1}{N} \operatorname{Tr} \frac{1}{H^{(w)} - z} \right) \leq CN^{-1/2+C\varepsilon} + CN^{-\delta+C\varepsilon}.
\]

The proof can easily be generalized to functions of several variables. This concludes the proof of Theorem 4.1.

### 4.2. Proof of the correlation function comparison Theorem 4.2

We define an approximate delta function (times \( \pi \)) on the scale \( \eta \) by

\[
\theta_\eta(x) := \operatorname{Im} \frac{1}{x - i\eta}.
\]

We will choose \( \eta \sim N^{-1-\varepsilon} \), that is, slightly smaller than the typical eigenvalue spacing. This means that an observable of the form \( \theta_\eta \) has sufficient resolution to detect individual eigenvalues. Moreover, polynomials in such observables detect correlation functions. On the other hand,

\[
\frac{1}{N} \operatorname{Im} \operatorname{Tr} G(E + i\eta) = \frac{1}{N} \sum_i \theta_\eta(\lambda_i - E),
\]
and therefore the expectation values of such observables are covered by the inequality (4.5) in Theorem 4.1. The rest of the proof amounts to making this idea precise. There are two technicalities to resolve. First, the correlation functions involve distinct eigenvalues (see (4.12) below), while the polynomials of the resolvent include an overcounting of coinciding eigenvalues. Thus, an exclusion-inclusion formula will be needed. Second, although $\eta$ is much smaller than the relevant scale $1/N$, it still does not give pointwise information about the correlation functions. However, the correlation functions in (1.35) are identified only as weak limits, that is, tested against a continuous function $O$. The continuity of $O$ can be used to show that the difference between the exact correlation functions and the smeared-out ones on the scale $\eta \sim N^{-1-\varepsilon}$ is negligible. This last step requires an a priori upper bound on the density to ensure that not too many eigenvalues fall into a negligibly small interval. Such a bound is given in (4.1), and it will eventually be verified by the local semicircle law.

For notational simplicity the detailed proof will be given only for the case of three-point correlation functions; the proof is analogous for the general case. By the definition of the correlation function, for any fixed $E$, $\alpha_1$, $\alpha_2$, and $\alpha_3$ we have

$$
\mathbb{E}^w \frac{1}{N(N-1)(N-2)} \sum_{i \neq j \neq k} \theta_\eta(\lambda_i - E - \frac{\alpha_1}{N}) \theta_\eta(\lambda_j - E - \frac{\alpha_2}{N}) \theta_\eta(\lambda_k - E - \frac{\alpha_3}{N})
$$

$$
= \int dx_1 dx_2 dx_3 p_{w,N}(x_1, x_2, x_3) \theta_\eta(x_1 - E_1) \theta_\eta(x_2 - E_2) \theta_\eta(x_3 - E_3),
$$

(4.12)

where $E_j := E + \alpha_j/N$ and $\mathbb{E}^w$ indicates the expectation with respect to the variables $w$. By the exclusion-inclusion principle,

$$
\mathbb{E}^w \frac{1}{N(N-1)(N-2)} \sum_{i \neq j \neq k} \theta_\eta(x_1 - E_1) \theta_\eta(x_2 - E_2) \theta_\eta(x_3 - E_3) = \mathbb{E}^w A_1 + \mathbb{E}^w A_2 + \mathbb{E}^w A_3,
$$

(4.13)

where

$$
A_1 := \frac{1}{N(N-1)(N-2)} \prod_{j=1}^3 \left[ \frac{1}{N} \sum_i \theta_\eta(\lambda_i - E_j) \right],
$$

$$
A_3 := \frac{2}{N(N-1)(N-2)} \sum_i \theta_\eta(\lambda_i - E_1) \theta_\eta(\lambda_i - E_2) \theta_\eta(\lambda_i - E_3) + \cdots
$$

and

$$
A_2 := B_1 + B_2 + B_3,
$$

with

$$
B_3 = - \frac{1}{N(N-1)(N-2)} \sum_i \theta_\eta(\lambda_i - E_1) \theta_\eta(\lambda_i - E_2) \sum_k \theta_\eta(\lambda_k - E_3),
$$

and similarly, $B_1$ consists of terms with $j = k$, while $B_2$ consists of terms with $i = k$. 

We note that, modulo a trivial change in the prefactor, $E^w A_1$ can be approximated by

$$E^w F\left(\frac{1}{N} \operatorname{Im} \operatorname{Tr} \frac{1}{H^{(v)} - z_1}, \frac{1}{N} \operatorname{Im} \operatorname{Tr} \frac{1}{H^{(v)} - z_2}, \frac{1}{N} \operatorname{Im} \operatorname{Tr} \frac{1}{H^{(v)} - z_3}\right),$$

where the function $F$ is chosen to be $F(x_1, x_2, x_3) := x_1 x_2 x_3$ for $\max_j |x_j| \leq N^\varepsilon$ and is smoothly cut off to go to zero in the regime $\max_j |x_j| \geq N^{2\varepsilon}$. The difference between the expectations of $F$ and $A_1$ is negligible, since it comes from the regime where $N^\varepsilon \leq \max_j \frac{1}{N} |\operatorname{Im} \operatorname{Tr}(H^{(v)} - z_j)^{-1}| \leq N^2$, which has an exponentially small probability by (4.9) (the upper bound on the Green function always holds since $\eta \geq N^{-2}$). Here the arguments of $F$ are the imaginary parts of the traces of the Green functions, but this type of function is allowed when applying Theorem 4.1, since

$$\operatorname{Im} \operatorname{Tr} G(z) = \frac{1}{2} [\operatorname{Tr} G(z) - \operatorname{Tr} G(\bar{z})].$$

We remark that the main assumption (4.1) for Theorem 4.1 is satisfied, as follows from one of the theorems on the local semicircle law (for instance, Theorem 2.5 with the choice of $M \sim N$, or Theorem 2.19).

Similarly, we can approximate $E^w B_3$ by

$$E^w G\left(\frac{1}{N^2} \operatorname{Tr} \left\{ \frac{1}{N} \operatorname{Im} \operatorname{Tr} \frac{1}{H^{(v)} - z_1} \frac{1}{N} \operatorname{Im} \operatorname{Tr} \frac{1}{H^{(v)} - z_2} \right\}, \frac{1}{N} \operatorname{Im} \operatorname{Tr} \frac{1}{H^{(v)} - z_3}\right),$$

where $G(x_1, x_2) = x_1 x_2$ with an appropriate cutoff for large arguments. There are similar expressions for $B_1$ and $B_2$, and also for $A_3$, the latter involving the trace of the product of three resolvents. By Theorem 4.1, these expectations with respect to $w$ in the approximations of $E^w A_i$ can be replaced by the expectations with respect to $v$ with only negligible errors, provided that $\eta \geq N^{-1-\varepsilon}$. We have thus proved that

$$\lim_{N \to \infty} \int dx_1 dx_2 dx_3 [p^{(3)}_{w,N}(x_1, x_2, x_3) - p^{(3)}_{v,N}(x_1, x_2, x_3)]$$

$$\times \theta_{\eta}(x_1 - E_1) \theta_{\eta}(x_2 - E_2) \theta_{\eta}(x_3 - E_3) = 0. \quad (4.14)$$

Let $\eta = N^{-1-\varepsilon}$ for the rest of the proof. We now show that the validity of (4.14) for any choice of $E$, $\alpha_1$, $\alpha_2$, $\alpha_3$ (recall that $E_j = E + \alpha_j/N$) implies that the rescaled correlation functions $p^{(3)}_{w,N}(E + \beta_1/N, E + \beta_2/N, E + \beta_3/N)$ and $p^{(3)}_{v,N}(E + \beta_1/N, E + \beta_2/N, E + \beta_3/N)$, as functions of the variables $\beta_1$, $\beta_2$, $\beta_3$, have the same weak limit.

Let $O$ be a smooth, compactly supported test function, and let

$$O_{\eta}(\beta_1, \beta_2, \beta_3) := \frac{1}{(\pi N)^3} \int_{\mathbb{R}^3} d\alpha_1 d\alpha_2 d\alpha_3 O(\alpha_1, \alpha_2, \alpha_3)$$

$$\times \theta_{\eta}\left(\frac{\beta_1 - \alpha_1}{N}\right) \theta_{\eta}\left(\frac{\beta_2 - \alpha_2}{N}\right) \theta_{\eta}\left(\frac{\beta_3 - \alpha_3}{N}\right)$$
be a smoothing of it on the scale $N \eta$. Then we can write
\[
\int_{\mathbb{R}^3} d\beta_1 d\beta_2 d\beta_3 O(\beta_1, \beta_2, \beta_3) p_w^{(3)}(\frac{E + \beta_1}{N}, \frac{E + \beta_2}{N}, \frac{E + \beta_3}{N}) = \int_{\mathbb{R}^3} d\beta_1 d\beta_2 d\beta_3 O(\beta_1, \beta_2, \beta_3) p_w^{(3)}(\frac{E + \beta_1}{N}, \frac{E + \beta_2}{N}, \frac{E + \beta_3}{N}) + \int_{\mathbb{R}^3} d\beta_1 d\beta_2 d\beta_3 (O - O_\eta)(\beta_1, \beta_2, \beta_3) p_w^{(3)}(\frac{E + \beta_1}{N}, \frac{E + \beta_2}{N}, \frac{E + \beta_3}{N}).
\]

(4.15)

The first term on the right-hand side, after the change of variables $x_j = E + \beta_j/N$, takes the form
\[
\int_{\mathbb{R}^3} d\alpha_1 d\alpha_2 d\alpha_3 O(\alpha_1, \alpha_2, \alpha_3) \int_{\mathbb{R}^3} dx_1 dx_2 dx_3 p_w^{(3)}(x_1, x_2, x_3) \times \theta_\eta(x_1 - E_1) \theta_\eta(x_2 - E_2) \theta_\eta(x_3 - E_3),
\]

(4.16)

that is, it can be written as an integral of expressions of the form (4.14) for which limits with $p_w,N$ and $p_v,N$ coincide.

Finally, the second term on the right-hand side of (4.15) is negligible. To see this, note that for any test function $Q$
\[
\int_{\mathbb{R}^3} d\beta_1 d\beta_2 d\beta_3 Q(\beta_1, \beta_2, \beta_3) p_w^{(3)}(\frac{E + \beta_1}{N}, \frac{E + \beta_2}{N}, \frac{E + \beta_3}{N}) = N^3 \int_{\mathbb{R}^3} dx_1 dx_2 dx_3 Q(N(x_1 - E), (N x_2 - E), N(x_3 - E)) p_w^{(3)}(x_1, x_2, x_3) = \left(1 - \frac{1}{N}\right) \left(1 - \frac{2}{N}\right) \mathbb{E}^w \sum_{i \neq j \neq k} Q(N(\lambda_i - E), N(\lambda_j - E), N(\lambda_k - E)).
\]

(4.17)

If the test function $Q$ was supported on a ball of size $N \varepsilon'$ with $\varepsilon' > 0$, then this last term would be bounded by
\[
\|Q\|_\infty \mathbb{E}^w \mathcal{N}_E^{C N^{-1+\varepsilon'}}(E) \leq C\|Q\|_\infty N^{4\varepsilon'}.
\]

Here $\mathcal{N}_E(D)$ denotes the number of eigenvalues in the interval $[E - \tau, E + \tau]$, and in this estimate we used the local semicircle law on intervals of size $\tau \geq N^{-1+\varepsilon'}$.

Now let $Q := O - O_\eta$. From the definition of $O_\eta$ it is easy to see that the function
\[
Q_1(\beta_1, \beta_2, \beta_3) = O(\beta_1, \beta_2, \beta_3) - O_\eta(\beta_1, \beta_2, \beta_3) \prod_{j=1}^3 1(|\beta_j| \leq N^{\varepsilon'})
\]

satisfies the estimate $\|Q_1\|_\infty \leq \|Q\|_\infty = \|O - O_\eta\|_\infty \leq C N \eta = CN^{-\varepsilon}$. So choosing $\varepsilon' < \varepsilon/4$, we see that the contribution of $Q_1$ is negligible. Finally, $Q_2 = Q - Q_1$ is given by
\[
Q_2(\beta_1, \beta_2, \beta_3) = -O_\eta(\beta_1, \beta_2, \beta_3) \left[1 - \prod_{j=1}^3 1(|\beta_j| \leq N^{\varepsilon'})\right],
\]
and we have
\[
\left| Q_2 \right| \leq C \left[ \frac{1}{1 + \beta_1^2} \right] \left[ \frac{1}{1 + \beta_2^2} \right] \left[ \frac{1}{1 + \beta_3^2} \right] \{1(\left| \beta_1 \right| \geq N^{\varepsilon'}) + \cdots \}
\]
\[
\leq C \left\{ N^{-\varepsilon'} \right\} \left[ \frac{N^{\varepsilon'}}{N^{2\varepsilon'} + \beta_1^2} \right] \left[ \frac{1}{1 + \beta_2^2} \right] + \cdots \}. 
\tag{4.18}
\]
Hence the contribution of \( Q_2 \) in the last term of (4.17) is bounded by
\[
CN^{-3-\varepsilon'} \mathbb{E}w \sum_{i,j,k} \left\{ \frac{N^{-1+\varepsilon'}}{N^{-2+2\varepsilon'} + (\lambda_i - E)^2} \left[ \frac{N^{-1}}{N^{-2} + (\lambda_j - E)^2} \right] \times \left[ \frac{N^{-1}}{N^{-2} + (\lambda_k - E)^2} \right] + \cdots \right\}. 
\tag{4.19}
\]
From the local semicircle law, that is, Theorem 2.5 or Theorem 2.19, the last term is bounded by \( N^{-\varepsilon'} \) up to some logarithmic factor. To see this, note that the Riemann sums for the eigenvalues in (4.19) can be replaced by an integral, because the resolution scale of the functions involved is at least \( N^{-1} \). This completes the proof of Theorem 4.2.

### 4.3. Sketch of the proof of Theorem 1.7

We again interpolate between \( H \) and \( H' \) step by step, changing the distribution of the matrix elements of \( H \) from \( \nu \) to \( \nu' \) one by one according to a fixed ordering. Let \( H^{(\tau)}(h) \) be the matrix where the first \( \tau - 1 \) elements have the distribution \( \nu' \), the \( \tau \)th entry is \( h \), and the remaining elements have the distribution \( \nu \). Denote by \( \lambda_i(H^{(\tau)}(h)) \) the \( i \)th eigenvalue of \( H^{(\tau)}(h) \). We define
\[
\mathcal{F}_\tau(h) := F(N\lambda_1(H^{(\tau)}(h)), N\lambda_2(H^{(\tau)}(h)), \ldots)
\]
and prove that
\[
\left| \mathbb{E} \mathcal{F}_\tau(h) - \mathbb{E}' \mathcal{F}_\tau(h') \right| \leq CN^{-2-c_0}. 
\tag{4.20}
\]
Since the number of replacement steps is of order \( N^2 \), this will prove (1.77). Let \( \tau \) represent the \((p,q)\)th matrix element and then drop the index \( \tau \).

We will prove that
\[
\left| \frac{\partial^n \mathcal{F}}{\partial h^n} \right| \leq CN^{O(c_0) + o(1)} \tag{4.21}
\]
for any \( n \leq 5 \). Then by the Taylor expansion,
\[
\mathcal{F}(h) = \sum_{n=0}^{4} \frac{1}{n!} \frac{\partial^n \mathcal{F}}{\partial h^n}(0) h^n + N^{-5/2+O(c_0)+o(1)}, 
\tag{4.22}
\]
since \( |h| \leq N^{-5/2+o(1)} \) with very high probability. After taking the expectations for \( h \) with respect to \( \nu \) and \( \nu' \), we see because the first four moments match that the contributions of the fully expanded terms in (4.22) coincide, and this proves (4.20).
To see (4.21), we assume for simplicity that $F$ has only one variable and $i_1 = i$. Then
\[ \frac{\partial F}{\partial h}(h) = N F'(\lambda) \frac{\partial \lambda_i}{\partial h}(h). \]

By standard first-order perturbation theory we get that for $h = h_{pq}$
\[ \frac{\partial \lambda_i}{\partial h} = 2 \text{Re} u_i(p) \bar{u}_i(q), \tag{4.23} \]
where $u_i = (u_i(1), \ldots, u_i(N))$ is the eigenfunction belonging to $\lambda_i$. Since the eigenvectors are delocalized,
\[ \|u_i\|_\infty \approx N^{-1/2} \tag{4.24} \]
(modulo logarithmic corrections; see (1.31)), so we get that
\[ \left| \frac{\partial \lambda_i}{\partial h} \right| \lesssim O(N^{-1}), \tag{4.25} \]
and thus
\[ \left| \frac{\partial F}{\partial h} \right| \lesssim N N^{c_0} N^{-1} = N^{c_0}. \]

For higher-order derivatives, we have to differentiate the eigenfunctions as well. This gives rise to resonances, for example, with $h = h_{pq}$ we have
\[ \frac{\partial u_i(p)}{\partial h_{pq}} = \sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} u_j(p)[\bar{u}_j(p)u_i(q) + \bar{u}_j(q)u_i(p)]. \]
Thus,
\[ \left| \frac{\partial u_i(p)}{\partial h} \right| \leq N^{-1/2} \frac{1}{N} \sum_{j \neq i} \frac{1}{|\lambda_i - \lambda_j|} \lesssim N^{-1/2+c_0}, \]
assuming that the eigenvalues regularly follow the semicircle law and that no two neighbouring eigenvalues get closer than $N^{-1-c_0}$ (see (1.78)). Substituting this bound into the derivative of (4.23), we have
\[ \left| \frac{\partial^2 \lambda_i}{\partial h^2} \right| \lesssim C N^{-1+c_0}. \]

Combining this bound with (4.25) and reinstating the general case when $F$ has more than one variable, we have
\[ \left| \frac{\partial^2 F}{\partial h^2} \right| \lesssim C N |\partial_{ii} F| |\partial^2 \lambda_i| + C N^2 |\partial_{ij} F| |\partial \lambda_j| |\partial \lambda_i| \leq C N^{2c_0}. \]

The argument for the higher derivatives is similar. The key technical inputs are the delocalization bound on the eigenvectors (4.24), which can be obtained from the local semicircle law, and the lower tail estimate (1.78).
5. Universality for Wigner matrices: putting it together

In this section we put the previous information together to prove our main result, Theorem 5.1 below. We will state our most general result from [8]. The same result was proved under somewhat more restrictive conditions in our previous papers, for example, Theorem 2.3 in [16], Theorem 3.1 in [6], and Theorem 2.2 in [9].

Recall that \( p_N(\lambda_1, \ldots, \lambda_N) \) denotes the symmetric joint density of the eigenvalues of the \( N \times N \) Wigner matrix \( H \). For simplicity we will use the formalism as if the joint distribution of the eigenvalues were absolutely continuous with respect to Lebesgue measure, but this is not necessary for the proof. Recall the definition of the \( k \)-point correlation functions (marginals) \( p^{(k)}_N \) in (1.34). We will use the notation \( p^{(k)}_N, \text{GUE} \) and \( p^{(k)}_N, \text{GOE} \) for the correlation functions of the GUE and GOE ensembles.

We consider the rescaled correlation functions near a fixed energy \( E \) under a scaling for which the local density is 1. The sine-kernel universality for the GUE ensemble states that the rescaled correlation functions converge weakly to the determinant of the sine kernel, \( K(x) = \frac{\sin \pi x}{\pi x} \), that is,

\[
\frac{1}{[\varrho_{\text{sc}}(E)]^k} p^{(k)}_N(\text{GUE}) \left( E + \frac{\alpha_1}{N \varrho_{\text{sc}}(E)}, \ldots, E + \frac{\alpha_k}{N \varrho_{\text{sc}}(E)} \right) \rightarrow \det(K(\alpha_\ell - \alpha_j))_{\ell,j=1}^k \tag{5.1}
\]

as \( N \rightarrow \infty \) for any fixed energy \( |E| < 2 \) in the bulk of the spectrum [77], [2]. A similar result holds for the GOE case, with the sine kernel replaced by a similar but somewhat more complicated universal function (see [1]). Our main result is that the universality (5.1) holds for Hermitian or symmetric generalized Wigner matrices after averaging a bit with respect to the energy \( E \).

**Theorem 5.1** ([8], Theorem 2.2). Let \( H \) be an \( N \times N \) symmetric or Hermitian generalized Wigner matrix. In the Hermitian case the real and imaginary parts are assumed to be i.i.d. Suppose that the distribution \( \nu \) of the rescaled matrix elements \( \sqrt{N} h_{ij} \) has subexponential decay (2.32). Let \( k \geq 1 \) and \( O: \mathbb{R}^k \rightarrow \mathbb{R} \) be a continuous, compactly supported function. Then for any \( |E| < 2 \)

\[
\lim_{b \rightarrow 0} \lim_{N \rightarrow \infty} \frac{1}{2b} \int_{E-b}^{E+b} dv \int_{\mathbb{R}^k} d\alpha_1 \cdots d\alpha_k O(\alpha_1, \ldots, \alpha_k) \times \frac{1}{[\varrho_{\text{sc}}(v)]^k} \left( p^{(k)}_N - p^{(k)}_N, \# \right) \left( v + \frac{\alpha_1}{N \varrho_{\text{sc}}(v)}, \ldots, v + \frac{\alpha_k}{N \varrho_{\text{sc}}(v)} \right) = 0, \tag{5.2}
\]

where \( \# \) stands for GOE or GUE for the symmetric or Hermitian cases, respectively.

**Proof.** For definiteness, we consider the symmetric case, that is, the limit will be the Gaussian Orthogonal Ensemble (GOE), corresponding to the parameter \( \beta = 1 \) in the general formalism. The joint distribution of the eigenvalues \( x = (x_1, \ldots, x_N) \) is given by the measure

\[
\mu = \mu_N(dx) = \frac{e^{-N \mathcal{H}(x)}}{Z} dx, \quad \mathcal{H}(x) = \sum_{i=1}^N x_i^2 - \frac{1}{N} \sum_{i<j} \log |x_j - x_i|, \tag{5.3}
\]
and we assume that the eigenvalues are ordered, that is, \( \mu \) is restricted to the subset \( \Sigma_N = \{ x \in \mathbb{R}^N : x_1 < \cdots < x_N \} \).

Let \( \hat{H} \) be a symmetric Wigner matrix with single-entry distribution satisfying the subexponential decay (2.32). We let the matrix evolve according to the matrix-valued Ornstein–Uhlenbeck process (1.54):

\[
dH_t = \frac{1}{\sqrt{N}} d\beta_t - \frac{1}{2} H_t \, dt, \quad H_0 = \hat{H},
\]

and we recall that the distribution of \( H_t \) for each fixed \( t > 0 \) is the same as

\[
e^{-t/2} \hat{H} + (1 - e^{-t})^{1/2} V,
\]

where \( V \) is an independent GOE matrix. The distribution \( \nu_t(dx) = u_t(x) \, dx \) of the matrix elements evolves according to the Ornstein–Uhlenbeck process on \( \mathbb{R} \), that is,

\[
\partial_t u_t = Au_t, \quad A = \frac{1}{2} \partial^2_x - \frac{x}{2} \partial_x.
\]

Note that the initial distribution \( \nu = \nu_0 \) may be singular, but for any \( t > 0 \) the distribution \( \nu_t \) is absolutely continuous.

The Ornstein–Uhlenbeck process (5.5) induces the Dyson Brownian motion on the eigenvalues [15], with generator given by

\[
L = \sum_{i=1}^N \frac{1}{2N} \partial_i^2 + \sum_{i=1}^N \left( -\frac{1}{4} x_i + \frac{1}{2N} \sum_{j \neq i} \frac{1}{x_i - x_j} \right) \partial_i,
\]

acting on \( L^2(\mu) \). The measure \( \mu \) is invariant and reversible with respect to the dynamics generated by \( L \).

Denote the distribution of the eigenvalues at time \( t \) by \( f_t(x) \, \mu(dx) \). Then \( f_t \) satisfies the equation

\[
\partial_t f_t = L f_t
\]

with initial condition \( f_0 \) given by the eigenvalue density of the Wigner matrix \( \hat{H} \). With the previous notation, we have \( p_N = f_0 \mu_N \), where \( p_N \), and hence \( f_0 \), may be singular with respect to Lebesgue measure. Due to the condition \( \beta \geq 1 \), the eigenvalues do not cross, that is, the dynamics (5.7) is well defined on \( \Sigma_N \). By using a straightforward symmetrization, one can extend the equilibrium measure, the density functions, and the dynamics to the whole of \( \mathbb{R}^N \). We will use the formalism of ordered eigenvalues everywhere, except in the definition (1.34) of the correlation functions, where the symmetrized version is easier. With a slight abuse of notation we will disregard this difference.

Theorem 5.1 was originally proved in [16] for standard Wigner matrices and under more restrictive conditions on the single-entry distribution. Here we present a more streamlined proof, following [9], but for notational simplicity we consider only the case of standard Wigner matrices. The main part of the proof of Theorem 5.1 consists of three steps:

**Step 1.** First we show that there exists an \( \varepsilon_0 > 0 \) such that the correlation functions of any Wigner ensemble with a Gaussian convolution of variance \( t \sim N^{-\varepsilon_0} \) coincide
with the GOE. In other words, any ensemble of the form (5.4) with \( t \geq N^{-\varepsilon_0} \) (and with subexponential decay of the matrix elements of \( \hat{H} \)) has universal local statistics.

**Step 2.** Let \( t = N^{-\varepsilon_0} \). Here we show that for any given Wigner matrix \( H \) we can find another Wigner matrix \( \hat{H} \) such that the first three moments of \( H \) and \( H_t = e^{-t/2}\hat{H} + (1 - e^{-t})^{1/2}V \) coincide and the fourth moments are close of order \( O(N^{-\varepsilon_0}) \).

**Step 3.** Theorem 4.2, which was a corollary of the Green function comparison theorem, shows that the local correlation functions of \( H \) and \( H_t \) in Step 2 coincide. Together with Step 1, this will complete the proof of Theorem 5.1.

Now we formulate the statements in Step 1 and Step 2 more precisely, Step 3 has already been completed.

**5.1. Step 1: Universality for Gaussian convolutions.** This step is just an application of Theorem 3.1, and we formulate it for our special case.

**Theorem 5.2.** Suppose that the probability distribution of the initial symmetric Wigner matrix \( \hat{H} \) has subexponential decay (2.32) with some exponent \( \vartheta_1 \), and let \( H_t \) be given by the Gaussian convolution (5.4). Let \( p^{(k)}_{t,N} \) denote the \( k \)-point correlation function of the eigenvalues of \( H_t \). Then there exists an \( \varepsilon_0 > 0 \) depending on the parameters in (2.32) such that for any \( t \geq N^{-\varepsilon_0} \)

\[
\lim_{b \to 0} \lim_{N \to \infty} \frac{1}{2b} \int_{E-b}^{E+b} dv \int_{\mathbb{R}^k} d\alpha_1 \cdots d\alpha_k O(\alpha_1, \ldots, \alpha_k) \times \frac{1}{[\varrho_{sc}(v)]^k} \left( p_{t,N}^{(k)} - P_{N,\text{GOE}}^{(k)} \right) \left( v + \frac{\alpha_1}{N\varrho_{sc}(v)}, \ldots, v + \frac{\alpha_k}{N\varrho_{sc}(v)} \right) = 0 \quad (5.8)
\]

for any continuous compactly supported test function \( O \).

We remark that the threshold exponent \( \varepsilon_0 \) can be given explicitly. If we use the local semicircle law in Theorem 2.5 and its corollary Theorem 2.7, then \( \varepsilon_0 \) can be chosen as any number smaller than 1/7. By use of the strong local semicircle law Theorem 2.19, the exponent \( \varepsilon_0 \) can be chosen to be any number smaller than 1.

**Proof.** We just have to check that the assumptions of Theorem 3.1 are satisfied. First, the Hamiltonian of the equilibrium measure is given by (5.3), and it is clearly of the form (3.5), so Condition I is automatic. The entropy condition (3.11) about the initial state \( f_0 \) may not be satisfied, since \( f_0 \) can even be singular. However, by the semigroup property of the OU flow, one can consider the initial condition \( f_{t_0} \) for the flow \( f_t \), \( t \geq t_0 \), for some \( t_0 \leq N^{-\varepsilon_0} \), because the statement of Theorem 3.1 concerns only the time \( t \geq N^{-\varepsilon_0} \). Thus, it is sufficient to show that the entropy condition is satisfied for some very small \( t_0 \ll N^{-\varepsilon_0} \).

To see this, let \( \nu_t \) denote the single-entry distribution of \( H_t \) and \( \bar{\nu}_t \) the probability measure of the matrix \( H_t \). Let \( \nu_{\text{GOE}} \) denote the probability measure of the GOE ensemble and \( \nu_{\text{GOE}} \) the probability measure of its \((i, j)\)th element, which is a Gaussian measure with mean zero and variance \( 1/N \). Since the dynamics of the matrix elements are independent (subject to the symmetry condition), and the
entropy is additive, we have the identity
\[
\int \log \left( \frac{d\tilde{\nu}_t}{d\tilde{\nu}_{\text{GOE}}} \right) d\tilde{\nu}_t = \sum_{i \leq j} \int \log \left( \frac{d\nu_t}{d\nu_{\text{GOE}}} \right) d\nu_t \leq CN^2 \int \log \left( \frac{d\nu_t}{d\nu_{\text{GOE}}} \right) d\nu_t, \tag{5.9}
\]
because the summation runs over the indices of all the independent elements $1 \leq i \leq j \leq N$. Clearly, the process $t \to \nu_t$ is an Ornstein–Uhlenbeck process, and each entropy term on the right-hand side of (5.9) is bounded by $CN$, provided that $t > t_0 := 1/N$ and $\nu_0$ has subexponential decay. Since the entropy of the marginal distribution on the eigenvalues is bounded by the entropy of the total measure on the matrix, we have proved that
\[
\int f_1/N \log f_1/N d\mu \leq CN^3, \tag{5.10}
\]
and this verifies (3.11). Therefore, in order to apply Theorem 3.1, we only have to verify Conditions II, III, and IV. Clearly, Conditions II and IV follow from the local semicircle law Theorem 2.5 with $\varrho(E) = \varrho_{\text{sc}}(E)$ (note that $M \sim N$ in the bounded variance case), and Condition III was proved in Theorem 2.7. Now we can apply Theorem 3.1 and we get (5.8) with any $\varepsilon_0 < \varepsilon$, where $\varepsilon$ is obtained from Theorem 2.7, that is, $\varepsilon_0$ can be any number smaller than $1/7$. If we use the strong local semicircle law Theorem 2.19, then (2.114) implies that
\[
E \sum_j (\lambda_j - \gamma_j)^2 \lesssim N^{-1},
\]
that is, Condition III (3.9) holds with any $a < 1/2$, and thus $\varepsilon_0$ can be any number smaller than $1$.

5.2. Step 2: Matching Lemma. For any real random variable $\xi$, denote by $m_k(\xi) = \mathbb{E} \xi^k$ its $k$th moment. By the Cauchy–Schwarz inequality, the sequence of moments $m_1, m_2, \ldots$ are not arbitrary numbers, for example, $|m_1|^2 \leq m_2$, $m_2^2 \leq m_4$, and so on, but there are more subtle relations. For example, if $m_1 = 0$, then
\[
m_4m_2 - m_3^2 \geq m_2^3, \tag{5.11}
\]
which can be obtained using the inequality
\[
m_3^2 = [\mathbb{E} \xi^3]^2 = [\mathbb{E} \xi(\xi^2 - 1)]^2 \leq [\mathbb{E} \xi^2][\mathbb{E}(\xi^2 - 1)] = m_2(m_4 - 2m_2^2 + 1)
\]
and noting that (5.11) is scale invariant, so it is sufficient to prove it for $m_2 = 1$. In fact, it is easy to see that (5.11) becomes an equality if and only if the support of $\xi$ consists of exactly two points (apart from the trivial case when $\xi \equiv 0$).

This restriction shows that for a given sequence of four admissible moments $m_1 = 0$, $m_2 = 1$, $m_3$, $m_4$ there may not exist a Gaussian divisible random variable $\xi$ with these moments; for instance, the moment sequence $(m_1, m_2, m_3, m_4) = (0, 1, 0, 1)$ uniquely characterizes the standard Bernoulli variable ($\xi = \pm 1$ with probabilities $1/2$, $1/2$). However, if we allow a bit of variability in the fourth moment, then one can match any four admissible moments with a small Gaussian convolution. This is the content of the next lemma, which completes Step 2.
Lemma 5.3 ([9], Lemma 6.5). Let $m_3$ and $m_4$ be two real numbers such that

$$m_4 - m_3^2 - 1 \geq 0, \quad m_4 \leq C_2 \quad (5.12)$$

for some positive constant $C_2$. Let $\xi^G$ be a Gaussian random variable with mean 0 and variance 1. Then for any sufficient small $\gamma > 0$ (depending on $C_2$) there exists a real random variable $\xi_\gamma$ having subexponential decay and independent of $\xi^G$ such that the first four moments of

$$\xi' = (1 - \gamma)^{1/2}\xi + \gamma^{1/2}\xi^G \quad (5.13)$$

are $m_1(\xi') = 0$, $m_2(\xi') = 1$, $m_3(\xi') = m_3$, and $m_4(\xi')$, with

$$|m_4(\xi') - m_4| \leq C\gamma \quad (5.14)$$

for some $C$ depending on $C_2$.

Proof. It is easy to see the following by an explicit construction.

Claim 5.4. For any given numbers $m_3$ and $m_4$ with $m_4 - m_3^2 - 1 \geq 0$ there is a random variable $X$ with first four moments 0, 1, $m_3$, $m_4$ and with subexponential decay.

For any real random variable $\zeta$ independent of $\xi^G$ and with first 4 moments 0, 1, $m_3(\zeta)$, and $m_4(\zeta) < \infty$ the first 4 moments of

$$\zeta' = (1 - \gamma)^{1/2}\zeta + \gamma^{1/2}\xi^G \quad (5.15)$$

are 0, 1,

$$m_3(\zeta') = (1 - \gamma)^{3/2}m_3(\zeta), \quad (5.16)$$

and

$$m_4(\zeta') = (1 - \gamma)^2m_4(\zeta) + 6\gamma - 3\gamma^2. \quad (5.17)$$

Using Claim 5.4, we get that for any $\gamma > 0$ there exists a real random variable $\xi_\gamma$ with first four moments 0, 1,

$$m_3(\xi_\gamma) = (1 - \gamma)^{-3/2}m_3, \quad (5.18)$$

and

$$m_4(\xi_\gamma) = m_3(\xi_\gamma)^2 + (m_4 - m_3^2).$$

From $m_4 \leq C_2$ it follows that $m_3^2 \leq C_2^{3/2}$, and thus

$$|m_4(\xi_\gamma) - m_4| \leq C\gamma \quad (5.19)$$

for some $C$ depending on $C_2$. Hence, from (5.16) and (5.17) we get that $\xi' = (1 - \gamma)^{1/2}\xi_\gamma + \gamma^{1/2}\xi^G$ satisfies $m_3(\xi') = m_3$ and (5.14). This completes the proof of Lemma 5.3.
Appendix A. Large Deviation Estimates: proof of Lemma 2.12

The estimates in Lemma 2.12 are weaker than the corresponding results of Hanson and Wright [101] used in [57] and [6], but they require only independent not necessarily identically distributed random variables with subexponential decay, and moreover, the proofs are much simpler. Thus, the Gaussian decay requirement of Hanson and Wright is relaxed to subexponential decay in our case, but the tail probability estimate is weaker.

Proof of (2.68). Without loss of generality we may assume that \( \sigma = 1 \). The assumption (2.67) implies that the \( k \)th moment of \( a_i \) has the estimate

\[
E |a_i|^k \leq (Ck)^{\alpha k}
\]  

(A.1)

for some \( C > 0 \).

First, for \( p \in \mathbb{N} \) we estimate

\[
E \left| \sum_{i=1}^N a_i A_i \right|^p.
\]  

(A.2)

From the Marcinkiewicz–Zygmund inequality, we get that for \( p \geq 2 \)

\[
E \left| \sum_{i} a_i A_i \right|^p \leq (Cp)^{p/2}E \left( \sum_{i} |a_i A_i|^2 \right)^{p/2}
\]  

(A.3)

(for an estimate of the constant see, for example, [107], Exercise 2.2.30). Substituting (A.1) into (A.3), we have

\[
E \left| \sum_{i} a_i A_i \right|^p \leq (Cp^{\frac{1}{2}+\alpha})^p \left( \sum_{i} |A_i|^2 \right)^{p/2},
\]  

(A.4)

which implies (2.68) for the choice \( p = \log N \) using a high-moment Markov inequality.

Proof of (2.69). We note that the random variables \( |a_i|^2 - 1 \) \( (1 \leq i \leq N) \) are independent and have mean 0 and variances bounded by some constant \( C \). Furthermore, the \( k \)th moment of \( |a_i|^2 - 1 \) has the upper estimate

\[
E(|a_i|^2 - 1)^k \leq (Ck)^{2\alpha k}.
\]  

(A.5)

Then repeating the proof of (2.68) with \( |a_i|^2 - 1 \) replacing \( a_i \), we obtain (2.69).

Proof of (2.70). For any \( p \in \mathbb{N} \) with \( p \geq 2 \) we estimate

\[
E \left| \sum_{i>j} \bar{a}_i \xi_i \right|^p \equiv E \left| \sum_{i>j} \bar{a}_i B_{ij} a_j \right|^p,
\]  

(A.6)

where \( \xi_i := \sum_{j<i} B_{ij} a_j \). Note that \( a_i \) and \( \xi_i \) are independent for any fixed \( i \). By definition,

\[
X_n \equiv \sum_{i=1}^n \bar{a}_i \xi_i
\]  

(A.7)
is a martingale. From the Burkholder inequality we have
\[
E \left| \sum_i \tilde{a}_i \xi_i \right|^p \leq (Cp)^{3p/2} E \left( \sum_i |\tilde{a}_i \xi_i|^2 \right)^{p/2} \tag{A.8}
\]
(for the constant, see § VII.3 of [108]). In view of the generalized Minkowski inequality, the independence of \(a_i\) and \(\xi_i\), and (A.1),
\[
E \left| \sum_i \tilde{a}_i \xi_i \right|^2 \leq \sum_i E[|\tilde{a}_i \xi_i|^p]^{2/p} = \sum_i \left[ E(|\tilde{a}_i|^p) E(|\xi_i|^p) \right]^{2/p} \leq (Cp)^{2\alpha} \sum_i [E(|\xi_i|^p)]^{2/p}.
\]
By (A.4),
\[
E(|\xi_i|^p) \leq (Cp^{1/2+\alpha})^p \left( \sum_j |B_{ij}|^2 \right)^{p/2}.
\]
Combining this with (A.8), we get that
\[
E \left| \sum_i \tilde{a}_i \xi_i \right|^p \leq (Cp)^{2p(1+\alpha)} \left( \sum_i \sum_j |B_{ij}|^2 \right)^{p/2}. \tag{A.9}
\]

Then choosing \(p = \log N\) and using the Markov inequality, we obtain (2.70).

**Appendix B. Proof of Theorem 3.1**

Recalling the notation in (3.56) and nearby, we start with the identity (3.55):
\[
\int_{E-b}^{E+b} \frac{dE'}{2b} \int_{R^n} d\alpha_1 \cdots d\alpha_n O(\alpha_1, \ldots, \alpha_n) p^{(n)}_{\tau, N}(E' + \frac{\alpha_1}{N\varrho(E)}, \ldots, E' + \frac{\alpha_n}{N\varrho(E)})
= \int_{E-b}^{E+b} \frac{dE'}{2b} \sum_{m \in S_n} \sum_{i=1}^N Y_{i,m}(E', x). \tag{B.1}
\]

It must be shown that
\[
\lim_{N \to \infty} \left| \int_{E-b}^{E+b} \frac{dE'}{2b} \int_{m \in S_n} \sum_{i=1}^N Y_{i,m}(E', x)(f_\tau - 1) \, d\mu \right| = 0. \tag{B.2}
\]

Let \(M\) be an \(N\)-dependent parameter to be chosen at the end of the proof. We set
\[
S_n(M) := \{m \in S_n : m_n \leq M\}, \quad S_n^c(M) := S_n \setminus S_n(M)
\]
and note that \(|S_n(M)| \leq M^{n-1}\). To prove (B.2), it is sufficient to show that
\[
\lim_{N \to \infty} \left| \int_{E-b}^{E+b} \frac{dE'}{2b} \int_{m \in S_n} \sum_{i=1}^N Y_{i,m}(E', x)(f_\tau - 1) \, d\mu \right| = 0, \tag{B.3}
\]
and that
\[
\lim_{N \to \infty} \sum_{\mathbf{m} \in S_0(M)} \left| \int_{E-b}^{E+b} \frac{dE'}{2b} \int_{E-b}^E \sum_{i=1}^N Y_{i,\mathbf{m}}(E', x) f_\tau \, d\mu \right| = 0 \quad \text{(B.4)}
\]
for any \( \tau > N^{-2\varepsilon+\delta} \) (note that \( \tau = \infty \) corresponds to the equilibrium, that is, \( f_\infty = 1 \)).

**Step 1: small \( \mathbf{m} \) case; proof of (B.3).** After performing the \( dE' \) integration, we will eventually apply Theorem 3.3 to the function
\[
G(u_1, u_2, \ldots) := \int_{\mathbb{R}} \tilde{O}(y, u_1, u_2, \ldots) \, dy,
\]
that is, to the quantity
\[
\int_{\mathbb{R}} dE' Y_{i,\mathbf{m}}(E', x) = \frac{1}{N} G(N(x_i - x_{i+m_2}), \ldots) \quad \text{(B.5)}
\]
for each fixed \( i \) and \( \mathbf{m} \).

For any \( E \) and \( 0 < \xi < b \) we define sets of integers \( J = J_{E,b,\xi} \) and \( J^\pm = J_{E,b,\pm\xi}^\pm \) by
\[
J := \{ i : \gamma_i \in [E-b, E+b] \}, \quad J^\pm := \{ i : \gamma_i \in [E-(b \pm \xi), E+b \pm \xi], \}
\]
where \( \gamma_i \) was defined in (3.8). Clearly, \( J^- \subset J \subset J^+ \). With this notation we have
\[
\int_{E-b}^{E+b} \frac{dE'}{2b} \sum_{i=1}^N Y_{i,\mathbf{m}}(E', x) = \int_{E-b}^{E+b} \frac{dE'}{2b} \sum_{i \in J^+} Y_{i,\mathbf{m}}(E', x) + \Omega_{J,\mathbf{m}}^+(x). \quad \text{(B.6)}
\]

The error term \( \Omega_{J,\mathbf{m}}^+ \) defined implicitly by (B.6) comes from the indices \( i \notin J^+ \) for which \( x_i \in [E-b, E+b] + O(N^{-1}) \), since \( Y_{i,\mathbf{m}}(E', x) = 0 \) unless \( |x_i - E'| \leq C/N \), the constant depending on the support of \( O \). Thus,
\[
|\Omega_{J,\mathbf{m}}^+(x)| \leq CN^{-1}b^{-1} \# \{ i : |x_i - \gamma_i| \geq \xi/2 \}
\]
for any sufficiently large \( N \), assuming that \( \xi \gg 1/N \) and using the fact that \( O \) is a bounded function. The additional factor \( N^{-1} \) comes from the \( dE' \) integration. Taking the expectation with respect to the measure \( f_\tau \, d\mu \), we get that
\[
\int |\Omega_{J,\mathbf{m}}^+(x)| f_\tau \, d\mu \leq Cb^{-1}\xi^{-2}N^{-1} \int \sum_i (x_i - \gamma_i)^2 f_\tau \, d\mu = Cb^{-1}\xi^{-2}N^{-1-2a}, \quad \text{(B.7)}
\]
using Condition III (3.9). We also have the estimate
\[
\int_{E-b}^{E+b} \frac{dE'}{2b} \sum_{i \in J^+} Y_{i,\mathbf{m}}(E', x) \leq \int_{E-b}^{E+b} \frac{dE'}{2b} \sum_{i \in J^-} Y_{i,\mathbf{m}}(E', x) + CN^{-1}|J^+ \setminus J^-|
\]
\[
= \int_{\mathbb{R}} \frac{dE'}{2b} \sum_{i \in J^-} Y_{i,\mathbf{m}}(E', x) + C(Nb)^{-1}|J^+ \setminus J^-| + \Xi_{J,\mathbf{m}}^+(x)
\]
\[
\leq \int_{\mathbb{R}} \frac{dE'}{2b} \sum_{i \in J} Y_{i,\mathbf{m}}(E', x) + C(Nb)^{-1}|J^+ \setminus J^-| + C(Nb)^{-1}|J \setminus J^-| + \Xi_{J,\mathbf{m}}^+(x), \quad \text{(B.8)}
\]
where the error term \( \Xi_{J,m}^+ \) determined by (B.8) comes from the indices \( i \in J^- \) such that \( x_i \notin [E-b, E+b]+O(1/N) \). It satisfies the same estimate (B.7) as \( \Omega_{J,m}^+ \). By the continuity of \( \varrho \), the density of the \( \gamma_i \) is bounded by \( CN \), and thus \( |J^+ \setminus J^-| \leq CN\xi \) and \( |J \setminus J^-| \leq CN\xi \). Therefore, summing the formula (B.5) over \( i \in J \), we get from (B.6) and (B.8) that

\[
\int_{E-b}^{E+b} \frac{dE'}{2b} \int \sum_{i=1}^{N} Y_{i,m}(E',x)f_{\tau} \, d\mu \\
\leq \frac{1}{2b} \int \frac{1}{N} \sum_{i \in J} G(N(x_i - x_{i+m_2}, \ldots) f_{\tau} \, d\mu + Cb^{-1}\xi + Cb^{-1}\xi^{-2}N^{-1-2a}
\]

for each \( m \in S_n \). A similar lower bound can be obtained in the same way, and after choosing \( \xi = N^{-1/3} \) we get that

\[
\int_{E-b}^{E+b} \frac{dE'}{2b} \int \sum_{i=1}^{N} Y_{i,m}(E',x)f_{\tau} \, d\mu \\
- \int \frac{1}{N} \sum_{i \in J} G(N(x_i - x_{i+m_2}, \ldots) f_{\tau} \, d\mu \leq CN^{-1/3} \quad (B.9)
\]

for each \( m \in S_n \), where \( C \) depends on \( b \). It is possible to optimize the choice of \( \xi \), depending on \( b \) and \( a \), and this would yield the effective bound mentioned after Theorem 3.1, but in this exposition we will not pursue the effective bound (see [6] for more details).

Summing (B.9) over all \( m \in S_n(M) \), we get that

\[
\left| \int_{E-b}^{E+b} \frac{dE'}{2b} \int \sum_{m \in S_n(M)} \sum_{i=1}^{N} Y_{i,m}(E',x)f_{\tau} \, d\mu \\
- \int \sum_{m \in S_n(M)} \frac{1}{N} \sum_{i \in J} G(N(x_i - x_{i+m_2}, \ldots) f_{\tau} \, d\mu \right| \leq CM^{n-1}N^{-1/3}, \quad (B.10)
\]

and the same estimate holds for the equilibrium, that is, if we set \( \tau = \infty \) in (B.10). Subtracting these two formulae and applying the inequality (3.30) in Theorem 3.3 to each summand in the second term in (B.9), we conclude that

\[
\left| \int_{E-b}^{E+b} \frac{dE'}{2b} \int \sum_{m \in S_n(M)} \sum_{i=1}^{N} Y_{i,m}(E',x)(f_{\tau} - 1) \, d\mu \right| \leq CM^{n-1}(N^{-1/3} + N^{-\delta/6}). \quad (B.11)
\]

Choosing

\[
M := N^{\min\{1/3, \delta/6\}/n}, \quad (B.12)
\]

we get that (B.11) vanishes as \( N \to \infty \), and this proves (B.3).

**Step 2: large \( m \) case; proof of (B.4).** For a fixed \( y \in \mathbb{R} \) and \( \ell > 0 \), let

\[
\chi(y, \ell) := \sum_{i=1}^{N} \mathbb{1}\left\{ x_i \in \left[ y - \frac{\ell}{N}, y + \frac{\ell}{N} \right] \right\}
\]
denote the number of points in the interval \([y - \ell/N, y + \ell/N]\). Note that for a fixed \(m = (m_2, \ldots, m_n)\) we have

\[
\sum_{i=1}^{N} |Y_{i,m}(E', x)| \leq C \chi(E', \ell) \mathbf{1}(\chi(E', \ell) \geq m_n) \leq C \sum_{m=m_n}^{\infty} m \mathbf{1}(\chi(E', \ell) \geq m),
\]

where \(\ell\) denotes the maximum of \(|u_1| + \cdots + |u_n|\) in the support of \(\tilde{O}(u_1, \ldots, u_n)\).

Since the summation over all increasing sequences \(m = (m_2, \ldots, m_n) \in \mathbb{N}_+^{n-1}\) with a fixed \(m_n\) contains at most \(m_n^{n-2}\) terms, we have

\[
\sum_{m \in S_n^c(M)} \left| \int_{E-b}^{E+b} \frac{dE'}{2b} \sum_{i=1}^{N} |Y_{i,m}(E', x)| f_{\tau} \, d\mu \right| \leq C \int_{E-b}^{E+b} \frac{dE'}{2b} \sum_{m=M}^{\infty} m^{n-1} \int \mathbf{1}(\chi(E', \ell) \geq m) f_{\tau} \, d\mu. \tag{B.14}
\]

Now we use Condition IV for the interval \(I = [E' - N^{-1+\sigma}, E' + N^{-1+\sigma}]\) with \(\sigma := (2n)^{-1} \min\{1/3, \delta/6\}\). Clearly, \(\mathcal{N}_I \geq \chi(E', \ell)\) for sufficiently large \(N\), and thus we get from (3.10) that

\[
\sum_{m=M}^{\infty} m^{n-1} \int \mathbf{1}(\chi(E', \ell) \geq m) f_{\tau} \, d\mu \leq C_a \sum_{m=M}^{\infty} m^{n-1} \left( \frac{m}{N \sigma} \right)^{-a}
\]

holds for any \(a \in \mathbb{N}\). By the choice of \(\sigma\), we get that \(\sqrt{m} \geq N^{\sigma}\) for any \(m \geq M\) (see (B.12)), and thus, choosing \(a = 2n + 2\), we get that

\[
\sum_{m=M}^{\infty} m^{n-1} \int \mathbf{1}(\chi(E', \ell) \geq m) f_{\tau} \, d\mu \leq \frac{C_a}{M} \to 0
\]

as \(N \to \infty\). Substituting this in (B.14), we complete the proof of (B.4) and thereby the proof of Theorem 3.1.

Bibliography

[1] M. L. Mehta, *Random matrices*, Academic Press, Boston, MA 1991, xviii+562 pp.
[2] P. A. Deift, *Orthogonal polynomials and random matrices: A Riemann–Hilbert approach*, Courant Lect. Notes Math., vol. 3, New York Univ., Courant Institute of Mathematical Sciences, New York; Amer. Math. Soc., Providence, RI 1999, viii+273 pp.
[3] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*, Cambridge Stud. Adv. Math., vol. 118, Cambridge Univ. Press, Cambridge 2010, xiv+492 pp.
[4] P. J. Forrester, *Log-gases and random matrices*, London Math. Soc. Monogr. Ser., vol. 34, Princeton Univ. Press, Princeton, NJ 2010, 791 pp.
[5] A. Guionnet, *Grandes matrices aléatoires et théorèmes d’universalité*, Séminaire Bourbaki, 62ème année, no. 1019, 2009–2010, 33 pp.
[6] L. Erdős, B. Schlein, H.-T. Yau, and J. Yin, “The local relaxation flow approach to universality of the local statistics for random matrices”, Ann. Inst. Henri Poincaré (to appear); arXiv:0911.3687.
[7] L. Erdős, H.-T. Yau, and J. Yin, Rigidity of eigenvalues of generalized Wigner matrices, arXiv:1007.4652.
[8] L. Erdős, H.-T. Yau, and J. Yin, Universaliversality for generalized Wigner matrices with Bernoulli distribution, arXiv:1003.3813.
[9] L. Erdős, H.-T. Yau, and J. Yin, Bulk universality for generalized Wigner matrices, arXiv:1001.3453.
[10] F. J. Dyson, “Correlations between eigenvalues of a random matrix”, Comm. Math. Phys. 19:3 (1970), 235–250.
[11] M. L. Mehta, “A note on correlations between eigenvalues of a random matrix”, Comm. Math. Phys. 20:3 (1971), 245–250.
[12] E. Brézin and S. Hikami, “Correlations of nearby levels induced by a random potential”, Nuclear Phys. B 479:3 (1996), 697–706; “Spectral form factor in a random matrix theory”, Phys. Rev. E (3) 55:4 (1997), 4067–4083.
[13] K. J. Johansson, “Universality of the local spacing distribution in certain ensembles of Hermitian Wigner matrices”, Comm. Math. Phys. 215:3 (2001), 683–705.
[14] L. Erdős, S. Péché, J. A. Ramírez, B. Schlein, and H.-T. Yau, “Bulk universality for Wigner matrices”, Comm. Pure Appl. Math. 63:7 (2010), 895–925; arXiv:0905.4176.
[15] F. J. Dyson, “A Brownian-motion model for the eigenvalues of a random matrix”, J. Math. Phys. 3:6 (1962), 1191–1198.
[16] L. Erdős, B. Schlein, and H.-T. Yau, Universaliversality of random matrices and local relaxation flow, arXiv:0907.5605.
[17] T. Tao and V. Vu, “Random matrices: Universality of local eigenvalue statistics”, Acta Math. 206:1 (2011), 127–204; arXiv:0906.0510.
[18] L. Erdős, J. Ramírez, B. Schlein, T. Tao, V. Vu, and H.-T. Yau, “Bulk universality for Wigner Hermitian matrices with subexponential decay”, Math. Res. Lett. 17:4 (2010), 667–674, http://www.mathjournals.org/mrl/2010-017-004/2010-017-004-007.html.
[19] В.А. Марченко, Л.А. Пастур, “Распределение собственных значений в некоторых ансамблях случайных матриц”, Матем. сб. 72:4 (1967), 507–536; English transl., V. A. Marchenko and L. A. Pastur, “Distribution of eigenvalues for some sets of random matrices”, Math. USSR-Sb. 1:4 (1967), 457–483.
[20] K.W. Wachter, “Strong limits of random matrix spectra for sample covariance matrices of independent elements”, Ann. Probab. 6:1 (1978), 1–18.
[21] E.P. Wigner, “Characteristic vectors of bordered matrices with infinite dimensions”, Ann. of Math. (2) 62:3 (1955), 548–564.
[22] J. Wishart, “The generalized product moment distribution in samples from a normal multivariate population”, Biometrika 20A (1928), 32–52.
[23] J.H. Schenker and H. Schulz-Baldes, “Semicircle law and freeness for random matrices with symmetries or correlations”, Math. Res. Lett. 12:4 (2005), 531–542, http://www.mathjournals.org/mrl/2005-012-004/2005-012-004-007.html.
[24] J. Schenker and H. Schulz-Baldes, “Gaussian fluctuations for random matrices with correlated entries”, Int. Math. Res. Not. IMRN 15 (2007), rnm047, 36 pp.
[25] P.W. Anderson, “Absences of diffusion in certain random lattices”, Phys. Rev. 109:5 (1958), 1492–1505.
Universality of Wigner random matrices

[26] M. V. Berry and M. Tabor, "Level clustering in the regular spectrum", Proc. R. Soc. Lond. Ser. A 356:1686 (1977), 375–394.

[27] O. Bohigas, M. J. Giannoni, and C. Schmit, "Characterization of chaotic quantum spectra and universality of level fluctuation laws", Phys. Rev. Lett. 52:1 (1984), 1–4.

[28] J. Fröhlich and T. Spencer, "Absence of diffusion in the Anderson tight binding model for large disorder or low energy", Comm. Math. Phys. 88:2 (1983), 151–184.

[29] J. Fröhlich, F. Martinelli, E. Scoppola, and T. Spencer, "Constructive proof of localization in the Anderson tight binding model", Comm. Math. Phys. 101:1 (1985), 21–46.

[30] M. Aizenman and S. Molchanov, "Localization at large disorder and at extreme energies: an elementary derivation", Comm. Math. Phys. 157:2 (1993), 245–278.

[31] M. Aizenman, R. Sims, and S. Warzel, "Absolutely continuous spectra of quantum tree graphs with weak disorder", Comm. Math. Phys. 264:2 (2006), 371–389.

[32] T. Spencer, "Lifshitz tails and localization", Preprint, 1993.

[33] A. Elgart, "Lifshitz tails and localization in the three-dimensional Anderson model", Duke Math. J. 146:2 (2009), 331–360.

[34] A. Klein, "Absolutely continuous spectrum in the Anderson model on the Bethe lattice", Math. Res. Lett. 1:4 (1994), 399–407, http://www.mrlonline.org/mrl/1994-001-004/1994-001-004-001.html.

[35] M. Aizenman, R. Sims, and S. Warzel, "Absolutely continuous spectra of quantum tree graphs with weak disorder", Comm. Math. Phys. 264:2 (2006), 371–389.

[36] R. Froese, D. Hasler, and W. Spitzer, "Transfer matrices, hyperbolic geometry and absolutely continuous spectrum for some discrete Schrödinger operators on graphs", J. Funct. Anal. 230:1 (2006), 184–221.

[37] M. Aizenman and S. Warzel, "The canopy graph and level statistics for random operators on trees", Math. Phys. Anal. Geom. 9:4 (2007), 291–333.

[38] I. Rodnianski and W. Schlag, "Classical and quantum scattering for a class of long range potentials", Int. Math. Res. Not. 2003:5 (2003), 243–300.

[39] J. Bourgain, "Random lattice Schrödinger operators with decaying potential: some higher dimensional phenomena", Geometric aspects of functional analysis, Lecture Notes in Math., vol. 1807, Springer, Berlin 2003, pp. 70–98.

[40] S. Denisov, "Absolutely continuous spectrum of multidimensional Schrödinger operator", Int. Math. Res. Not. 2004:74, 3963–3982; arXiv: math/0408376.

[41] W. De Roeck and J. Fröhlich, "Diffusion of a massive quantum particle coupled to a quasi-free thermal medium", Comm. Math. Phys. 303:3 (2011), 613–707; arXiv: 0906.5178.

[42] T. Chen, "Localization lengths and Boltzmann limit for the Anderson model at small disorders in dimension 3", J. Stat. Phys. 120:1–2 (2005), 279–337.

[43] H. Spohn, "Derivation of the transport equation for electrons moving through random impurities", J. Statist. Phys. 17:6 (1977), 385–412.

[44] L. Erdös and H.-T. Yau, "Linear Boltzmann equation as the weak coupling limit of the random Schrödinger equation", Comm. Pure Appl. Math. 53:6 (2000), 667–735.

[45] L. Erdös, M. Salmhofer, and H.-T. Yau, "Quantum diffusion for the Anderson model in the scaling limit", Ann. Henri Poincaré 8:4 (2007), 621–685.

[46] K. B. Efetov, Supersymmetry in disorder and chaos, Cambridge Univ. Press, Cambridge 1997, xiv+441 pp.

[47] Y. V. Fyodorov and A. D. Mirlin, "Scaling properties of localization in random band matrices: A σ-model approach", Phys. Rev. Lett. 67:18 (1991), 2405–2409.
[48] J. Schenker, “Eigenvector localization for random band matrices with power law band width”, *Comm. Math. Phys.* **290**:3 (2009), 1065–1097.

[49] L. Erdős and A. Knowles, “Quantum diffusion and eigenfunction delocalization in a random band matrix model”, *Comm. Math. Phys.* **303**:2 (2011), 509–554; arXiv: 1002.1695.

[50] L. Erdős and A. Knowles, *Quantum diffusion and delocalization for band matrices with general distribution*, arXiv: 1005.1838.

[51] M. Disertori, H. Pinson, and T. Spencer, “Density of states for random band matrices”, *Comm. Math. Phys.* **232**:1 (2002), 83–124.

[52] M. Disertori and T. Spencer, “Anderson localization for a supersymmetric sigma model”, *Comm. Math. Phys.* **303**:2 (2011), 509–554; arXiv: 0910.3325.

[53] M. Disertori, T. Spencer, and M.R. Zirnbauer, “Quasi-diffusion in a 3D supersymmetric hyperbolic sigma model”, *Comm. Math. Phys.* **300**:2 (2010), 435–486; arXiv: 0901.1652.

[54] T. Tao, V. Vu, and M. Krishnapur, “Random matrices: universality of ESD’s and the circular law”, *Ann. Probab.* **38**:5 (2010), 2023–2065; arXiv: 0807.4898.

[55] L. Erdős, B. Schlein, and H.-T. Yau, “Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices”, *Ann. Probab.* **37**:3 (2009), 815–852.

[56] L. Erdős, B. Schlein, and H.-T. Yau, “Local semicircle law and complete delocalization for Wigner random matrices”, *Comm. Math. Phys.* **287**:2 (2009), 641–655.

[57] L. Erdős, B. Schlein, and H.-T. Yau, “Wegner estimate and level repulsion for Wigner random matrices”, *Int. Math. Res. Not. IMRN* **2010**:3 (2010), 436–479; arXiv: 0811.2591.

[58] Z. D. Bai and Y. Q. Yin, “Limit of the smallest eigenvalue of a large dimensional sample covariance matrix”, *Ann. Probab.* **21**:3 (1993), 1275–1294.

[59] A. Guionnet, “Large deviations upper bounds and central limit theorems for non-commutative functionals of Gaussian large random matrices”, *Ann. Inst. H. Poincaré Probab. Statist.* **38**:3 (2002), 341–384.

[60] G. W. Anderson and O. Zeitouni, “A CLT for a band matrix model”, *Probab. Theory Related Fields* **134**:2 (2006), 283–338.

[61] B. Simon, *Trace ideals and their applications*, 2nd ed., Math. Surveys Monogr., vol. 120, Amer. Math. Soc., Providence, RI 2005, viii+150 pp.

[62] A. Soshnikov, “Universality at the edge of the spectrum in Wigner random matrices”, *Comm. Math. Phys.* **207**:3 (1999), 697–733.

[63] Я. Г. Синай, А. Б. Сосников, “Уточнение полукругового закона Вигнера в окрестности края спектра для случайных симметричных матриц”, *Функц. анализ и его прир.* **32**:2 (1998), 56–79; English transl., Ya. G. Sinai and A. B. Soshnikov, “A refinement of Wigner’s semicircle law in a neighbourhood of the spectrum edge for random symmetric matrices”, *Funct. Anal. Appl.* **32**:2 (1998), 114–131.

[64] T. Tao and V. Vu, “Random matrices: universality of local eigenvalue statistics up to the edge”, *Comm. Math. Phys.* **298**:2 (2010), 549–572; arXiv: 0908.1892.

[65] C. A. Tracy and H. Widom, “Level-spacing distributions and the Airy kernel”, *Comm. Math. Phys.* **159**:1 (1994), 151–174.

[66] C. A. Tracy and H. Widom, “On orthogonal and symplectic matrix ensembles”, *Comm. Math. Phys.* **177**:3 (1996), 727–754.

[67] A. Ruzmaikina, “Universality of the edge distribution of eigenvalues of Wigner random matrices with polynomially decaying distributions of entries”, *Comm. Math. Phys.* **261**:2 (2006), 277–296.
[68] S. Péché and A. Soshnikov, “Wigner random matrices with non-symmetrically distributed entries”, *J. Stat. Phys.* 129:5–6 (2007), 857–884.

[69] T. Tao and V. Vu, “Random covariance matrices: Universality of local statistics of eigenvalues”, *Ann. Probab.* (to appear); arXiv: 0912.0966.

[70] F. J. Dyson, “Statistical theory of energy levels of complex systems. I”, *J. Math. Phys.* 3:1 (1962), 140–156; F. J. Dyson, “Statistical theory of energy levels of complex systems. II”, *J. Math. Phys.* 3:1 (1962), 157–165; F. J. Dyson, “Statistical theory of energy levels of complex systems. III”, *J. Math. Phys.* 3:1 (1962), 166–175.

[71] I. Dumitriu and A. Edelman, “Matrix models for beta ensembles”, *J. Math. Phys.* 43:11 (2002), 5830–5847.

[72] J. Ramírez, B. Rider, and B. Virag, *Beta ensembles, stochastic Airy spectrum, and a diffusion*, arXiv: math/0607331.

[73] J. A. Ramírez and B. Rider, “Diffusion at the random matrix hard edge”, *Comm. Math. Phys.* 288:3 (2009), 887–906.

[74] P. Deift, T. Kriecherbauer, K.T.-R. McLaughlin, S. Venakides, and X. Zhou, “Uniform asymptotics for polynomials orthogonal with respect to varying exponential weights and applications to universality questions in random matrix theory”, *Comm. Pure Appl. Math.* 52:11 (1999), 1335–1425.

[75] P. Deift, T. Kriecherbauer, K.T.-R. McLaughlin, S. Venakides, and X. Zhou, “Strong asymptotics of orthogonal polynomials with respect to exponential weights”, *Comm. Pure Appl. Math.* 52:12 (1999), 1491–1552.

[76] L. Pastur and M. Shcherbina, “Bulk universality and related properties of Hermitian matrix models”, *J. Stat. Phys.* 130:2 (2008), 205–250.

[77] D.S. Lubinsky, “A new approach to universality limits involving orthogonal polynomials”, *Ann. of Math.* (2) 170:2 (2009), 915–939.

[78] E. Levin and D.S. Lubinsky, “Universality limits in the bulk for varying measures”, *Adv. Math.* 219:3 (2008), 743–779.
[89] L. Erdős, J. Ramírez, B. Schlein, and H.-T. Yau, “Universality of sine-kernel for Wigner matrices with a small Gaussian perturbation”, Electron. J. Probab. 15 (2010), 526–603, Paper 18, http://www.math.washington.edu/~ejpecp/viewarticle.php?id=2084&layout=abstract; arXiv:0905.2089.

[90] M. Shcherbina, “Edge universality for orthogonal ensembles of random matrices”, J. Stat. Phys. 136:1 (2009), 35–50; arXiv:0812.3228.

[91] D. Bakry and M. Émery, “Diffusions hypercontractives”, Séminaire de probabilités XIX, Lecture Notes in Math., vol. 1123, Springer, Berlin 1985, pp. 177–206.

[92] J. Gustavsson, “Gaussian fluctuations of eigenvalues in the GUE”, Ann. Inst. H. Poincaré Probab. Statist. 41:2 (2005), 151–178.

[93] S. O’Rourke, “Gaussian fluctuations of eigenvalues in Wigner random matrices”, J. Stat. Phys. 138:6 (2010), 1045–1066.

[94] R. Curto and L. Fialkow, “Recursiveness, positivity and truncated moment problems”, Houston J. Math. 17:4 (1991), 603–635.

[95] O.N. Feldheim and S. Sodin, “A universality result for the smallest eigenvalues of certain sample covariance matrices”, Geom. Funct. Anal. 20:1 (2010), 88–123; arXiv:0812.1961.

[96] S. Sodin, “The spectral edge of some random band matrices”, Ann. of Math. (2) 172:3 (2010), 2223–2251; arXiv:0906.4047.

[97] S. Sodin, “The Tracy–Widom law for some sparse random matrices”, J. Stat. Phys. 136:5 (2010), 834–841; arXiv:0903.4295.

[98] S. Péché and A. Soshnikov, “On the lower bound of the spectral norm of symmetric random matrices with independent entries”, Electron. Commun. Probab. 13 (2008), 280–290, http://www.math.washington.edu/~ejpecp/ECP/viewarticle.php?id=2004&layout=abstract.

[99] A. Auffinger, G. Ben Arous, and S. Péché, “Poisson convergence for the largest eigenvalues of heavy tailed random matrices”, Ann. Inst. Henri Poincaré Probab. Statist. 45:3 (2009), 589–610.

[100] G. Biroli, J.-P. Bouchaud, and M. Potters, “On the top eigenvalue of heavy-tailed random matrices”, Europhys. Lett. EPL 78:1 (2007), Art. 10001, 5 pp.

[101] D.L. Hanson and F.T. Wright, “A bound on tail probabilities for quadratic forms in independent random variables”, Ann. Math. Statist. 42:3 (1971), 1079–1083.

[102] E. B. Davies, “The functional calculus”, J. London Math. Soc. (2) 52:1 (1995), 166–176.

[103] T. Spencer, “Random banded and sparse matrices (Chapter 23)”, Oxford handbook of random matrix theory (G. Akemann, J. Baik, P. Di Francesco, eds.), Oxford Univ. Press, Oxford 2011, 960 pp.

[104] Z. D. Bai, B. Miao, and J. Tsay, “Convergence rates of the spectral distributions of large Wigner matrices”, Int. Math. J. 1:1 (2002), 65–90.

[105] T. Tao and V. Vu, Random matrices: localization of the eigenvalues and the necessity of four moments, arXiv:1005.2901.

[106] H.-T. Yau, “Relative entropy and the hydrodynamics of Ginzburg–Landau models”, Lett. Math. Phys. 22:1 (1991), 63–80.

[107] D.W. Stroock, Probability theory: an analytic view, Cambridge Univ. Press, Cambridge 1993, xvi+512 pp.

[108] А.Н. Ширяев, Вероятность, 3-е изд., МЦНМО, М. 2007; 1-е изд., Наука, М. 1980; English transl. of 1st ed., A.N. Shiryaev, Probability, Grad. Texts in Math., vol. 95, Springer-Verlag, New York 1984, xi+577 pp.

L. Erdős
Ludwig-Maximilians-Universität München
E-mail: lerdos@mathematik.uni-muenchen.de

Received 07/APR/10