

## Universality for random matrices and log-gases

László Erdős

**ABSTRACT.** Eugene Wigner's revolutionary vision predicted that the energy levels of large complex quantum systems exhibit a universal behavior: the statistics of energy gaps depend only on the basic symmetry type of the model. These universal statistics show strong correlations in the form of level repulsion and they seem to represent a new paradigm of point processes that are characteristically different from the Poisson statistics of independent points.

Simplified models of Wigner's thesis have recently become mathematically accessible. For mean field models represented by large random matrices with independent entries, the celebrated Wigner-Dyson-Gaudin-Mehta (WDGM) conjecture asserts that the local eigenvalue statistics are universal. For invariant matrix models, the eigenvalue distributions are given by a log-gas with potential  $V$  and inverse temperature  $\beta = 1, 2, 4$ , corresponding to the orthogonal, unitary and symplectic ensembles. For  $\beta \notin \{1, 2, 4\}$ , there is no natural random matrix ensemble behind this model, but the analogue of the WDGM conjecture asserts that the local statistics are independent of  $V$ .

In these lecture notes we review the recent solution to these conjectures for both invariant and non-invariant ensembles. We will discuss two different notions of universality in the sense of (i) local correlation functions and (ii) gap distributions. We will demonstrate that the local ergodicity of the Dyson Brownian motion is the intrinsic mechanism behind the universality. In particular, we review the solution of Dyson's conjecture on the local relaxation time of the Dyson Brownian motion. Additionally, the gap distribution requires a De Giorgi-Nash-Moser type Hölder regularity analysis for a discrete parabolic equation with random coefficients. Related questions such as the local version of Wigner's semicircle law and delocalization of eigenvectors will also be discussed. We will also explain how these results can be extended beyond the mean field models, especially to random band matrices.

### CONTENTS

1. Introduction	60
2. Local semicircle law for general Wigner-type matrices	75
3. Universality of the correlation functions for Wigner matrices	97

---

2010 *Mathematics Subject Classification.* 15B52, 82B44.

*Key words and phrases.*  $\beta$ -ensemble, local semicircle law, Dyson Brownian motion. De Giorgi-Nash-Moser theory.

4. Universality of the correlation functions for $\beta$ -ensembles	109
5. Single gap universality	115
References	129

## 1. Introduction

### 1.1. The pioneering vision of Wigner.

Perhaps I am now too courageous when I try to guess the distribution of the distances between successive levels (of energies of heavy nuclei). Theoretically, the situation is quite simple if one attacks the problem in a simpleminded fashion. The question is simply what are the distances of the characteristic values of a symmetric matrix with random coefficients.

Eugene Wigner on the Wigner surmise, 1956

Large complex systems often exhibit remarkably simple universal patterns as the number of degrees of freedom increases. The simplest example is the central limit theorem: the fluctuation of the sums of independent random scalars, irrespective of their distributions, follows the Gaussian distribution. The other cornerstone of probability theory identifies the Poisson point process as the universal limit of many independent point-like events in space or time. These mathematical descriptions assume that the original system has independent (or at least weakly dependent) constituents. What if independence is not a realistic approximation and strong correlations need to be modelled? Is there a universality for strongly correlated models?

At first sight this seems an impossible task. While independence is a unique concept, correlations come in many different forms; a-priori there is no reason to believe that they all behave similarly. Nevertheless they do, according to the pioneering vision of Wigner [87] at least if they originate from certain physical systems and if the “right” question is asked. The actual correlated system he studied was the energy levels of heavy nuclei. Looking at spectral measurement data, it is obvious that the eigenvalue density (or density of states, as it is called in physics) heavily depends on the system. But Wigner asked a different question: what about the distribution of the rescaled energy gaps? He discovered that the difference of consecutive energy levels, after rescaling with the local density, shows a surprisingly universal behavior. He even predicted a universal law, given by the simple formula (called the *Wigner surmise*),

$$(1.1) \quad \mathbb{P}(\tilde{E}_j - \tilde{E}_{j-1} = s + ds) \approx \frac{\pi s}{2} \exp\left(-\frac{\pi}{4}s^2\right)ds,$$

where  $\tilde{E}_j = \varrho E_j$  denote the rescaling of the actual energy levels  $E_j$  by the density of states  $\varrho$  near the energy  $E_j$ . This law is characteristically different from the gap distribution of the Poisson process which is the exponential distribution,  $e^{-s}ds$ . The prefactor  $s$  in (1.1) indicates a *level repulsion* for the point process  $\tilde{E}_j$ , i.e. the eigenvalues are strongly correlated.

Comparing measurement data from various experiments, Wigner’s pioneering vision was that the energy gap distribution (1.1) of complicated quantum systems

is essentially universal; it depends only on the basic symmetries of model (such as time-reversal invariance). This thesis has never been rigorously proved for any realistic physical system but experimental data and extensive numerics leave no doubt on its correctness (see [65] for an overview).

Wigner not only predicted universality in complicated systems, but he also discovered a remarkably simple mathematical model for this new phenomenon: the eigenvalues of large random matrices. For practical purposes, Hamilton operators of quantum models are often approximated by large matrices that are obtained from some type of discretization of the original continuous model. These matrices have specific forms dictated by physical rules. Wigner's bold step was to neglect all details and consider the simplest random matrix whose entries are independent and identically distributed. The only physical property he retained was the basic symmetry class of the system; time reversal physical models were modelled by real symmetric matrices, while systems without time reversal symmetry (e.g. with magnetic fields) were modelled by complex Hermitian matrices. As far as the gap statistics are concerned, this simple-minded model reproduced the behavior of the complex quantum systems! The universal behavior extends to the joint statistics of several consecutive gaps which are essentially equivalent to the local correlation functions of the point process  $\tilde{E}_j$ . From mathematical point of view, a universal strongly correlated point process was found. The natural representatives of these universality classes are the random matrices with independent identically distributed Gaussian entries. These are called the *Gaussian orthogonal ensemble (GOE)* and the *Gaussian unitary ensemble (GUE)* in case of real symmetric and complex Hermitian matrices, respectively.

Since Wigner's discovery random matrix statistics are found everywhere in physics and beyond, wherever nontrivial correlations prevail. Among many other applications, random matrix theory (RMT) is present in chaotic quantum systems in physics, in principal component analysis in statistics, in communication theory and even in number theory. In particular, the zeros of the Riemann zeta function on the critical line are expected to follow RMT statistics due to a spectacular result of Montgomery [69].

In retrospect, Wigner's idea should have received even more attention. For centuries, the primary territory of probability theory was to model uncorrelated or weakly correlated systems. The surprising ubiquity of random matrix statistics is a strong evidence that it plays a similar fundamental role for correlated systems as Gaussian distribution and Poisson point process play for uncorrelated systems. RMT seems to provide essentially the only universal and generally computable pattern for complicated correlated systems.

In fact, a few years after Wigner's seminal paper [87], Gaudin [53] has discovered another remarkable property of this new point process: the correlation functions have a determinantal structure, at least if the distributions of the matrix elements are Gaussian. The algebraic identities within the determinantal form opened up the route to calculations and to obtain explicit formulas for local correlation functions. In particular, the gap distribution for the complex Hermitian case is given by a Fredholm determinant involving Hermite polynomials. In fact, Hermite polynomials were first introduced in the context of random matrices by Mehta and Gaudin [67] earlier. Dyson and Mehta [24, 26, 66] have later extended this exact calculation to correlation functions and to other symmetry classes. When compared with the

exact formula, the Wigner surmise (1.1), based upon a simple  $2 \times 2$  matrix model, turned out to be quite accurate. While the determinantal structure is present only in Gaussian Wigner matrices, the paradigm of local universality predicts that the formulas for the local eigenvalue statistics obtained in the Gaussian case hold for general distributions as well.

**1.2. Physical models.** The ultimate mathematical goal is to prove Wigner's vision for a very large class of realistic quantum mechanical models. This is extremely hard, since the local statistics involve tracking individual eigenvalues in the bulk spectrum. Wigner's original model, the energy levels of heavy nuclei, is a strongly interacting many-body quantum system. The rigorous analysis of such model with the required precision is beyond the reach of current mathematics.

A much simpler question is to neglect all interactions and to study the natural one-body quantum model, the Schrödinger operator  $-\Delta + V$  with a potential  $V$  on  $\mathbb{R}^d$ . The complexity comes from assuming that  $V$  is generic in some sense, in particular to exclude models with additional symmetries that may lead to non-universal eigenvalue correlations. Two well-studied examples are (i) the random Schrödinger operators where  $V = V(x)$  is a random field with a short range correlation, and (ii) quantum chaos models, where  $V$  is generic but fixed and the statistical ensemble is generated by sampling the spectrum in small spectral windows at high energies (an alternative formulation uses the semiclassical limit).

Unfortunately, there are essentially no rigorous results on local spectral universality even in these one-body models. Random Schrödinger operators are conjectured to exhibit a metal-insulator transition that was discovered by Anderson [4]. The high disorder regime is relatively well understood since the seminal work of Fröhlich and Spencer [51] (an alternative proof is given by Aizenman and Molchanov [1]). However, in this regime the eigenfunctions are localized and thus eigenfunctions belonging to neighboring eigenvalues are typically spatially separated, hence uncorrelated. Therefore, due to localization, the system does not have sufficient correlation to fall into the RMT universality class; in fact the local eigenvalue statistics follow the Poisson process [68]. In contrast, in the low disorder regime, starting from three spatial dimension and away from the spectral edges, the eigenfunctions are conjectured to be delocalized (*extended states conjecture*). Spatially overlapping eigenfunctions introduce correlations among eigenvalues and it is expected that the local statistics are given by RMT. In the theoretical physics literature, the existence of the delocalized regime and its RMT statistics are considered as facts, supported both by non-rigorous arguments and numerics. One of the most intriguing approach is via supersymmetric (SUSY) functional integrals that remarkably reproduce all formulas obtained by the determinantal calculations in much more general setup but in a non-rigorous way due to neglecting highly oscillatory terms. The rigorous mathematics seriously lags behind these developments; even the existence of the delocalized regime is not proven, let alone detailed spectral statistics.

Judged from the horizons of theoretical physics, rigorous mathematics does not fare much better in the quantum chaos models either. The grand vision is that the quantization of an integrable classical Hamiltonian system exhibits Poisson eigenvalue statistics and a chaotic classical system gives rise to RMT statistics [8, 10]. While Poisson statistics have been shown to emerge some specific integrable models [64, 77, 80], there is no rigorous result on the RMT statistics. Recently there

has been a remarkable mathematical progress in *quantum unique ergodicity (QUE)* that predicts that *all* eigenfunctions of chaotic systems are uniformly distributed all over the space, at least in some macroscopic sense. For arithmetic domains QUE has been proved in [62]. For general manifolds much less is known, but a lower bound on the topological entropy of the support of the limiting densities of eigenfunctions excludes that eigenfunctions are supported only on a periodic orbit [2]. Very roughly, QUE can be considered as the analogue of the extended states for random Schrödinger operators. Theoretically, the overlap of eigenfunctions should again lead to correlations between neighboring eigenvalues, but their direct quantitative analysis would require a much more precise understanding of the eigenfunctions.

**1.3. Random matrix ensembles.** In these lectures we consider even simpler models to test Wigner's universality hypothesis, namely the random matrix ensemble itself. The main goal is to show that their eigenvalues follow the local statistics of the Gaussian Wigner matrices which have earlier been computed explicitly by Dyson, Gaudin and Mehta. The statement that the local eigenvalue statistics is independent of the law of the matrix elements is generally referred to as the *universality conjecture of random matrices* and we will call it the *Wigner-Dyson-Gaudin-Mehta conjecture*. It was first formulated in Mehta's treatise on random matrices [65] in 1967 and has remained a key question in the subject ever since. The goal of these lecture notes is to review the recent progress that has led to the proof of this conjecture and we sketch some important ideas. We will, however, not be able to present all aspects of random matrices and we refer the reader to recent comprehensive books [3, 18, 20].

**1.3.1. Wigner ensembles.** To make the problem simpler, we restrict ourselves to either real symmetric or complex Hermitian matrices so that the eigenvalues are real. The standard model consists of  $N \times N$  square matrices  $H = (h_{ij})$  with matrix elements having mean zero and variance  $1/N$ , i.e.,

$$(1.2) \quad \mathbb{E} h_{ij} = 0, \quad \mathbb{E} |h_{ij}|^2 = \frac{1}{N} \quad i, j = 1, 2, \dots, N.$$

The matrix elements  $h_{ij}$ ,  $i, j = 1, \dots, N$ , are real or complex independent random variables subject to the symmetry constraint  $h_{ij} = \bar{h}_{ji}$ . These ensembles of random matrices are called (*standard*) *Wigner matrices*. We will always consider the limit as the matrix size goes to infinity, i.e.,  $N \rightarrow \infty$ . Every quantity related to  $H$  depends on  $N$ , so we should have used the notation  $H^{(N)}$  and  $h_{ij}^{(N)}$ , etc., but for simplicity we will omit  $N$  in the notation.

In Section 2 we will also consider generalizations of these ensembles, where we allow the matrix elements  $h_{ij}$  to have different distributions (but retaining independence). The main motivation is to depart from the mean-field character of the standard Wigner matrices, where the quantum transition amplitudes  $h_{ij}$  between any two sites  $i, j$  have the same statistics. The most prominent example is the random band matrix ensemble (see Example 2.1) that naturally interpolates between standard Wigner matrices and random Schrödinger operators with a short range hopping mechanism (see [81] for an overview).

The first rigorous result about the spectrum of a random matrix of this type is the famous *Wigner semicircle law* [87] which states that the empirical density of

the eigenvalues,  $\lambda_1, \lambda_2, \dots, \lambda_N$ , under the normalization (1.2), is given by

$$(1.3) \quad \varrho_N(x) := \frac{1}{N} \sum_{j=1}^N \delta(x - \lambda_j) \rightarrow \varrho_{sc}(x) := \frac{1}{2\pi} \sqrt{(4 - x^2)_+}$$

in the weak limit as  $N \rightarrow \infty$ . The limit density is independent of the details of the distribution of  $h_{ij}$ .

The Wigner surmise (1.1) is a much finer problem since it concerns individual eigenvalues and not only their behavior on macroscopic scale. To understand it, we introduce correlation functions. If  $p_N(\lambda_1, \lambda_2, \dots, \lambda_N)$  denotes the joint probability density of the (unordered) eigenvalues, then the  $n$ -point correlation functions (marginals) are defined by

$$(1.4) \quad p_N^{(n)}(\lambda_1, \lambda_2, \dots, \lambda_n) := \int_{\mathbb{R}^{N-n}} p_N(\lambda_1, \dots, \lambda_n, \lambda_{n+1}, \dots, \lambda_N) d\lambda_{n+1} \dots d\lambda_N.$$

To keep this introduction simple, we state the corresponding results in terms of the eigenvalue correlation functions for Hermitian  $N \times N$  matrices. In the Gaussian case (GUE) the joint probability density of the eigenvalues can be expressed explicitly as

$$(1.5) \quad p_N(\lambda_1, \lambda_2, \dots, \lambda_N) = \text{const.} \prod_{i < j} (\lambda_i - \lambda_j)^2 \prod_{j=1}^N e^{-\frac{1}{2} N \lambda_j^2},$$

where the normalization constant can be computed explicitly. The Vandermonde determinant structure allows one to compute the  $k$ -point correlation functions in the large  $N$  limit via Hermite polynomials that are the orthogonal polynomials with respect to the Gaussian weight function.

The result of Dyson, Gaudin and Mehta asserts that for any fixed energy  $E$  in the bulk of the spectrum, i.e.,  $|E| < 2$ , the small scale behavior of  $p_N^{(n)}$  is given explicitly by

$$(1.6) \quad \frac{1}{[\varrho_{sc}(E)]^n} p_N^{(n)}\left(E + \frac{\alpha_1}{N \varrho_{sc}(E)}, E + \frac{\alpha_2}{N \varrho_{sc}(E)}, \dots, E + \frac{\alpha_n}{N \varrho_{sc}(E)}\right) \rightarrow \det(K(\alpha_i - \alpha_j))_{i,j=1}^n$$

where  $K$  is the celebrated sine kernel

$$(1.7) \quad K(x, y) = \frac{\sin \pi(x - y)}{\pi(x - y)}.$$

Note that the limit in (1.6) is independent of the energy  $E$  as long as it lies in the bulk of the spectrum. The rescaling by a factor  $N^{-1}$  of the correlation functions in (1.6) corresponds to the typical distance between consecutive eigenvalues and we will refer to the law under such scaling as *local statistics*. Note that the correlation functions do not factorize, i.e. the eigenvalues are strongly correlated despite that the matrix elements are independent. Similar but more complicated formulas were obtained for symmetric matrices and also for the self-dual quaternion random matrices which is the third symmetry class of random matrix ensembles.

The convergence in (1.6) holds for each fixed  $|E| < 2$  and uniformly in  $(\alpha_1, \dots, \alpha_n)$  in any compact subset of  $\mathbb{R}^n$ . Fix now  $k$  compact subsets  $A_1, \dots, A_k$  in  $\mathbb{R}$ . From (1.6) one can compute the distribution of the number  $n_j$  of the rescaled eigenvalues  $\tilde{\lambda}_\alpha := N(\lambda_\alpha - E)\varrho_{sc}(E)$  in  $A_j$  around a fixed energy  $|E| < 2$ . The limit of the joint

probabilities

$$(1.8) \quad \mathbb{P}\left(\#\{\tilde{\lambda}_\alpha \in A_j\} = n_j, j = 1, 2, \dots, k\right)$$

is given as derivatives of a Fredholm determinant involving the sine kernel. Clearly (1.8) gives a complete local description of the rescaled eigenvalues as a point process around a fixed energy  $E$ . In particular it describes the distribution of the eigenvalue gap that contains a *fixed energy*  $E$ . However, (1.8) does not determine the distribution of the gap with a *fixed label*, e.g. the gap  $\lambda_{N/2+1} - \lambda_{N/2}$ . Only the cumulative statistics of many consecutive gaps can be deduced, see [18] for a precise formulation. The slight discrepancy between the statements at fixed energy and with fixed label leads to involved technical complications.

**1.3.2. Invariant ensembles.** The explicit formula (1.5) is special for Gaussian Wigner matrices; if  $h_{ij}$  are independent but non-Gaussian, then no analogous explicit formula is known for the joint probability density. Gaussian Wigner matrices have this special property because their distribution is *invariant* under base transformation. The derivation of (1.5) relies on the fact that in the diagonalization  $H = U\Lambda U^*$  of  $H$ , where  $\Lambda$  is diagonal and  $U$  is unitary, the distributions of  $U$  and  $\Lambda$  decouple. The Gaussian measure of  $h_{ij}$  with the normalization (1.2) can also be expressed as

$$(1.9) \quad \exp\left(-\frac{1}{2}N \operatorname{Tr} H^2\right) dH = \exp\left(-\frac{1}{2}N \operatorname{Tr} \Lambda^2\right) d(U\Lambda U^*),$$

where  $dH$  is the Lebesgue measure on hermitian matrices. The Vandermonde determinant in (1.5) originates from the integrating the Jacobian  $d(U\Lambda U^*)/d\Lambda$  over the unitary group. Similar argument holds for real symmetric matrices with orthogonal conjugations, the only difference is the exponent 2 of the Vandermonde determinant becomes 1. The exponent is 4 for the third symmetry class of Wigner matrices, the self-dual quaternion matrices with symmetry group being the symplectic matrices (*Gaussian symplectic ensemble, GSE*).

Starting from (1.5), there are two natural generalizations of Gaussian Wigner matrices. One direction is the Wigner matrices with non-Gaussian but independent entries that we have already introduced in Section 1.3.1. Another direction is to consider a more general real function  $V(H)$  of  $H$  instead of the quadratic  $H^2$  in (1.9). Since invariance still holds,  $\operatorname{Tr} V(H) = \operatorname{Tr} V(U\Lambda U^*) = \operatorname{Tr} V(\Lambda)$ , the same argument gives (1.5), with  $V(\lambda_i)$  instead of  $\lambda_i^2/2$ , for the correlation functions of  $\exp(-N \operatorname{Tr} V(H))$ . These are called *invariant ensembles* with potential  $V$ . Their matrix elements are in general correlated except in the Gaussian case.

Invariant ensembles in all three symmetry classes can be given simultaneously by the probability measure

$$Z^{-1} e^{-\frac{1}{2}N\beta \operatorname{Tr} V(H)} dH,$$

where  $N$  is the size of the matrix  $H$ ,  $V$  is a real valued potential and  $Z = Z_N$  is the normalization constant. The positive parameter  $\beta$  is determined by the symmetry class, its value is 1, 2 or 4, for real symmetric, complex hermitian and self-dual quaternion matrices, respectively. The Lebesgue measure  $dH$  is understood over the matrices in the same class. The probability distribution of the eigenvalues  $\lambda =$

$(\lambda_1, \dots, \lambda_N)$  is given by the explicit formula (c.f. (1.5))

$$(1.10) \quad \mu_{\beta, V}^{(N)}(\boldsymbol{\lambda}) d\boldsymbol{\lambda} \sim e^{-\beta N \mathcal{H}(\boldsymbol{\lambda})} d\boldsymbol{\lambda}$$

$$\text{with Hamiltonian } \mathcal{H}(\boldsymbol{\lambda}) := \sum_{k=1}^N \frac{1}{2} V(\lambda_k) - \frac{1}{N} \sum_{1 \leq i < j \leq N} \log(\lambda_j - \lambda_i).$$

The key structural ingredient of this formula, the logarithmic interaction that gives rise to the Vandermonde determinant, is the same as in the Gaussian case, (1.5). Thus all previous computations, developed for the Gaussian case, can be carried out for  $\beta = 1, 2, 4$ , provided that the Gaussian weight function for the orthogonal polynomials is replaced with the function  $e^{-\beta V(x)/2}$ . The analysis of the correlation functions depends critically on the asymptotic properties of the corresponding orthogonal polynomials.

While the asymptotics of the Hermite polynomial for the Gaussian case are well-known, the extension of the necessary analysis to a general potential is a demanding task; important progress was made since the late 1990's by Fokas-Its-Kitaev [50], Bleher-Its [9], Deift *et. al.* [18, 21, 22], Pastur-Shcherbina [72, 73] and more recently by Lubinsky [63]. These results concern the simpler  $\beta = 2$  case. For  $\beta = 1, 4$ , the universality was established only quite recently for analytic  $V$  with additional assumptions [19, 20, 60, 79] using earlier ideas of Widom [86]. The final outcome of these sophisticated analyses is that universality holds for the measure (1.10) in the sense that the short scale behavior of the correlation functions is independent of the potential  $V$  (with appropriate assumptions) provided that  $\beta$  is one of the classical values, i.e.,  $\beta \in \{1, 2, 4\}$ , that corresponds to an underlying matrix ensemble.

Notwithstanding matrix ensembles or orthogonal polynomials, the measure (1.10) on  $N$  points  $\lambda_1, \dots, \lambda_N$  is perfectly well defined for any  $\beta > 0$ . It can be interpreted as the Gibbs measure for a system of particles with external potential  $\frac{1}{2}V$  and with a logarithmic interaction (log-gas) at inverse temperature  $\beta$ . From this point of view  $\beta$  is a continuous parameter and the classical values  $\beta = 1, 2, 4$  play apparently no distinguished role. It is therefore natural to extend the universality problem to all non-classical  $\beta$  but the orthogonal polynomial methods are difficult to apply for this case. For any  $\beta > 0$  the local statistics for the Gaussian case  $V(x) = x^2/2$  is given by a point process, denoted by  $\text{Sine}_\beta$ . It can be obtained from a rescaling of the  $\text{Airy}_\beta$  process as  $\lim_{a \rightarrow \infty} \sqrt{a}(\text{Airy}_\beta + a) = \text{Sine}_\beta$ . The Airy process itself is the low lying eigenvalues of the one dimensional Schrödinger operator  $-\frac{d^2}{dx^2} + x + \frac{2}{\sqrt{\beta}}b'_x$  on the positive half line, where  $b'_x$  is the white noise. The relation between Gaussian random matrices and random Schrödinger operators is derived from a tridiagonal matrix representation [23]. Another convenient representation of the  $\text{Sine}_\beta$  process is given by the “Brownian carousel” [76, 85].

Beyond random matrices, the log-gas can also be viewed as the only interacting particle model with a scale-invariant interaction and with a single relevant parameter, the inverse temperature  $\beta$ . It is believed to be the canonical model for strongly correlated systems and thus to play a similarly fundamental role in probability theory as the Poisson process or the Brownian motion. Nevertheless, we still have very little information about its properties. Unlike the universality problem that is inherently analytical, many properties of the log-gas are destined, at the first sight, to be revealed by smart algebraic identities. Despite many trials by physicists and mathematicians, the log-gas with a general  $\beta$  seems to defy all algebraic attempts.



We do not really understand why the algebraic approach is suitable for  $\beta = 2$ , and to a lesser extent for  $\beta = 1, 4$ , but it fails for any other  $\beta$ , while from an analytical point of view there is no difference between various values of  $\beta$ . To understand this fascinating ensemble, a main goal is to develop general analytical methods that work for any  $\beta$ .

**1.4. Universality of the local statistics: the main results.** All universality results reviewed in the previous sections rely on some version of the explicit formula (1.10) that is not available for Wigner matrices with non-Gaussian matrix elements. The only result prior 2009 towards universality for Wigner matrices was the proof of Johansson [57] (extended by Ben Arous-Péché [7]) for complex Hermitian Wigner matrices with a substantial Gaussian component. The hermiticity is necessary, since the proof still relies on an algebraic formula, a modification of the Harish-Chandra/Itzykson/Zuber integral observed first by Brézin and Hikami in this context [14].

To indicate the restrictions imposed by the usage of explicit formulas, we note that previous methods were not suitable to deal even with very small perturbations of the Gaussian Wigner case. For example, universality was already not known if only a few matrix elements of  $H$  had a distribution different from Gaussian.

Given this background, the main challenge a few years ago was to develop a new approach to universality that does not rely on any algebraic identity. We believe that the genuine reason behind Wigner's universality is of analytic nature. Algebraic computations may be used to obtain explicit formulas for the most convenient representative of a universality class (typically the Gaussian case), but only analytical methods have the power to deal with the general case. In light of the two main classes of random matrix ensembles, we set the following two main problems.

*Problem 1:* Prove the Wigner-Dyson-Gaudin-Mehta conjecture, i.e. the universality for Wigner matrices with a general distribution for the matrix elements.

*Problem 2:* Prove the universality of the local statistics for the log-gas (1.10) for all  $\beta > 0$ .

We were able to solve Problem 1 for a very general class of distributions. As for Problem 2, we solved it for the case of real analytic potentials  $V$  assuming that the equilibrium measure is supported on a single interval, which, in particular, holds for any convex potential. We will give a historical overview of related results in Section 1.5.3.

The original universality conjectures, as formulated in Mehta's book [65], do not specify the type of convergence in (1.6). We focus on two types of results for both problems. First we show that universality holds in the sense that local correlation functions around an energy  $E$  converge weakly if  $E$  is averaged on a small interval of size  $N^{-1+\varepsilon}$ . Second, we prove the universality of the joint distribution of consecutive gaps with fixed labels.

We note that universality of the *cumulative* statistics of  $N^\varepsilon$  gaps directly follows from the weak convergence of the correlation functions but our result on a single gap requires a quite different approach. From the point of view of Wigner's original vision on the ubiquity of the random matrix statistics in seemingly disparate ensembles and physical systems, the issue of cumulative gap statistics versus single gap statistics is minuscule. Our main reason of pursuing the single gap universality is less for the result itself; more importantly, we develop new methods to analyze

the structure of the log-gases, which seem to represent the universal statistics of strongly correlated systems. In the next two sections we state the results precisely.

1.4.1. *Generalized Wigner matrices.* Our main results hold for a larger class of ensembles than the standard Wigner matrices, which we will call *generalized Wigner matrices*.

DEFINITION 1.1. ([46]) The real symmetric or complex Hermitian matrix ensemble  $H$  with centred and independent matrix elements  $h_{ij} = \bar{h}_{ji}$ ,  $i \leq j$ , is called generalized Wigner matrix if the following assumptions hold on the variances of the matrix elements  $s_{ij} = \mathbb{E}|h_{ij}|^2$ :

(A) For any  $j$  fixed

$$(1.11) \quad \sum_{i=1}^N s_{ij} = 1.$$

(B) There exist two positive constants,  $C_1$  and  $C_2$ , independent of  $N$  such that

$$(1.12) \quad \frac{C_1}{N} \leq s_{ij} \leq \frac{C_2}{N}.$$

The result on the correlation functions is the following theorem:

THEOREM 1.2 (Wigner-Dyson-Gaudin-Mehta conjecture for averaged correlation functions). [33, Theorem 7.2] Suppose that  $H = (h_{ij})$  is a complex Hermitian (respectively, real symmetric) generalized Wigner matrix. Suppose that for some constants  $\varepsilon > 0$ ,  $C > 0$ ,

$$(1.13) \quad \mathbb{E} \left| \sqrt{N} h_{ij} \right|^{4+\varepsilon} \leq C.$$

Let  $n \in \mathbb{N}$  and  $O : \mathbb{R}^n \rightarrow \mathbb{R}$  be compactly supported and continuous. Let  $E$  satisfy  $|E| < 2$  and let  $\xi > 0$ . Then for any sequence  $b_N$  satisfying  $N^{-1+\xi} \leq b_N \leq ||E| - 2|/2$  we have

$$(1.14) \quad \lim_{N \rightarrow \infty} \int_{E-b_N}^{E+b_N} \frac{dx}{2b_N} \int_{\mathbb{R}^n} d\alpha_1 \cdots d\alpha_n O(\alpha_1, \dots, \alpha_n) \\ \times \frac{1}{\varrho_{sc}(E)^n} \left( p_N^{(n)} - p_{G,N}^{(n)} \right) \left( x + \frac{\alpha_1}{N\varrho_{sc}(E)}, \dots, x + \frac{\alpha_n}{N\varrho_{sc}(E)} \right) = 0.$$

Here  $\varrho_{sc}$  is the semicircle law defined in (1.3),  $p_N^{(n)}$  is the  $n$ -point correlation function of the eigenvalue distribution of  $H$  (1.4), and  $p_{G,N}^{(n)}$  is the  $n$ -point correlation function of an  $N \times N$  GUE (respectively, GOE) matrix.

The condition (1.12) can be relaxed, see Corollary 8.3 [31]. For example, the lower bound can be changed to  $N^{-9/8+\varepsilon}$ . Alternatively, the upper bound  $C_2 N^{-1}$  can be replaced with  $N^{-1+\varepsilon_n}$  for some  $\varepsilon_n > 0$ . For band matrices, the upper and lower bounds can be simultaneously relaxed.

We remark that for the complex Hermitian case the convergence of the correlation functions can be strengthened to a convergence at each fixed energy, i.e. for

any fixed  $|E| < 2$  we have that

$$(1.15) \quad \int_{\mathbb{R}^n} d\alpha_1 \cdots d\alpha_n O(\alpha_1, \dots, \alpha_n) \\ \times \frac{1}{\varrho_{sc}(E)^n} \left( p_N^{(n)} - p_{G,N}^{(n)} \right) \left( E + \frac{\alpha_1}{N\varrho_{sc}(E)}, \dots, E + \frac{\alpha_n}{N\varrho_{sc}(E)} \right) = 0.$$

The main ideas leading to the results (1.14) and (1.15) have been developed in a series of papers. We will give a short overview of the key methods in Section 1.5.1 and of the related results in Section 1.5.3.

The second result on generalized Wigner matrices asserts that the local gap statistics in the bulk of the spectrum are universal for any general Wigner matrix, in particular they coincide with those of the Gaussian case. To formulate the statement, we need to introduce the notation  $\gamma_j$  for the  $j$ -th quantile of the semicircle density, i.e.  $\gamma_j = \gamma_j^{(N)}$  is defined by

$$(1.16) \quad \frac{j}{N} = \int_{-2}^{\gamma_j} \varrho_{sc}(x) dx.$$

We also introduce the notation  $\llbracket A, B \rrbracket := \{A, A+1, \dots, B\}$  for any integers  $A < B$ .

**THEOREM 1.3** (Gap universality for Wigner matrices). *[45, Theorem 2.2] Let  $H$  be a generalized real symmetric or complex Hermitian Wigner matrix with subexponentially decaying matrix elements, i.e. we assume that*

$$(1.17) \quad \mathbb{P}(\sqrt{N}|h_{ij}| \geq x) \leq C_0 \exp(-x^\vartheta)$$

*holds for any  $x > 0$  with some  $C_0, \vartheta$  positive constants. Fix a positive number  $\alpha > 0$ , an integer  $n \in \mathbb{N}$  and a smooth, compactly supported function  $O : \mathbb{R}^n \rightarrow \mathbb{R}$ . There exists an  $\varepsilon > 0$  and  $C > 0$ , depending only on  $C_0, \vartheta, \alpha$  and  $O$  such that*

$$(1.18) \quad \left| \left[ \mathbb{E} - \mathbb{E}^\mu \right] O(N(x_j - x_{j+1}), N(x_j - x_{j+2}), \dots, N(x_j - x_{j+n})) \right| \leq CN^{-\varepsilon},$$

*for any  $j \in \llbracket \alpha N, (1-\alpha)N \rrbracket$  and for any sufficiently large  $N \geq N_0$ , where  $N_0$  depends on all parameters of the model, as well as on  $n$  and  $\alpha$ . Here  $\mathbb{E}$  and  $\mathbb{E}^\mu$  denotes the expectation with respect to the Wigner ensemble  $H$  and the Gaussian equilibrium measure (see (1.5) for the Hermitian case), respectively.*

*More generally, for any  $k, m \in \llbracket \alpha N, (1-\alpha)N \rrbracket$  we have*

$$(1.19) \quad \left| \mathbb{E} O((N\varrho_k)(x_k - x_{k+1}), (N\varrho_k)(x_k - x_{k+2}), \dots, (N\varrho_k)(x_k - x_{k+n})) \right. \\ \left. - \mathbb{E}^\mu O((N\varrho_m)(x_m - x_{m+1}), (N\varrho_m)(x_m - x_{m+2}), \dots, (N\varrho_m)(x_m - x_{m+n})) \right| \\ \leq CN^{-\varepsilon},$$

*where the local density  $\varrho_k$  is defined by  $\varrho_k := \varrho_{sc}(\gamma_k)$ .*

As it was already mentioned, the gap universality with a certain local averaging, i.e. for the cumulative statistics of  $N^\varepsilon$  consecutive gaps, follows directly from the universality of the correlation functions, Theorem 1.2. The gap distribution for Gaussian random matrices, with a local averaging, can then be explicitly expressed via a Fredholm determinant, see [18–20]. The first result for a single gap, i.e. without local averaging, was only achieved recently in the special case of the Gaussian unitary ensemble (GUE) in [84], which statement then easily implies the same

results for complex Hermitian Wigner matrices satisfying the four moment matching condition.

1.4.2. *Log-gases.* In the case of invariant ensembles, it is well-known that for  $V$  satisfying certain mild conditions the sequence of one-point correlation functions, or densities, associated with  $\mu = \mu^{(N)}$  from (1.10) has a limit as  $N \rightarrow \infty$  and the limiting equilibrium density  $\varrho_V(s)$  can be obtained as the unique minimizer of the functional

$$I(\nu) = \int_{\mathbb{R}} V(t) \nu(t) dt - \int_{\mathbb{R}} \int_{\mathbb{R}} \log |t - s| \nu(s) \nu(t) dt ds.$$

We assume that  $\varrho = \varrho_V$  is supported on a single compact interval,  $[A, B]$  and  $\varrho \in C^2(A, B)$ . Moreover, we assume that  $V$  is *regular* in the sense that  $\varrho$  is strictly positive on  $(A, B)$  and vanishes as a square root at the endpoints, see (1.4) of [12]. It is known that these conditions are satisfied if, for example,  $V$  is strictly convex. In this case  $\varrho_V$  satisfies the equation

$$(1.20) \quad \frac{1}{2} V'(t) = \int_{\mathbb{R}} \frac{\varrho_V(s) ds}{t - s}$$

for any  $t \in (A, B)$ . For the Gaussian case,  $V(x) = x^2/2$ , the equilibrium density is given by the semicircle law,  $\varrho_V = \varrho_{sc}$ , see (1.3).

The following result was proven in Corollary 2.2 of [11] for convex potential  $V$  and it was generalized in Theorem 1.2 of [12] for the non-convex case.

**THEOREM 1.4** (Bulk universality of  $\beta$ -ensemble). *Assume  $V$  is real analytic with  $\inf_{x \in \mathbb{R}} V''(x) > -\infty$ . Let  $\beta > 0$ . Consider the  $\beta$ -ensemble  $\mu = \mu_{\beta, V}^{(N)}$  given in (1.10) and let  $p_N^{(n)}$  denote the  $n$ -point correlation functions of  $\mu$ , defined analogously to (1.4). For the Gaussian case,  $V(x) = x^2/2$ , the correlation functions are denoted by  $p_{G, N}^{(n)}$ . Let  $E \in (A, B)$  lie in the interior of the support of  $\varrho$  and similarly let  $E' \in (-2, 2)$  be inside the support of  $\varrho_{sc}$ . Let  $O : \mathbb{R}^n \rightarrow \mathbb{R}$  be a smooth, compactly supported function. Then for  $b_N = N^{-1+\xi}$  with any  $0 < \xi \leq 1/2$  we have*

$$(1.21) \quad \lim_{N \rightarrow \infty} \int d\alpha_1 \cdots d\alpha_n O(\alpha_1, \dots, \alpha_n) \left[ \int_{E-b_N}^{E+b_N} \frac{dx}{2b_N} \frac{1}{\varrho(E)^n} p_N^{(n)} \left( x + \frac{\alpha_1}{N\varrho(E)}, \dots, x + \frac{\alpha_n}{N\varrho(E)} \right) - \int_{E'-b_N}^{E'+b_N} \frac{dx}{2b_N} \frac{1}{\varrho_{sc}(E')^n} p_{G, N}^{(n)} \left( x + \frac{\alpha_1}{N\varrho_{sc}(E')}, \dots, x + \frac{\alpha_n}{N\varrho_{sc}(E')} \right) \right] = 0,$$

*i.e. the appropriately normalized correlation functions of the measure  $\mu_{\beta, V}^{(N)}$  at the level  $E$  in the bulk of the limiting density asymptotically coincide with those of the Gaussian case. In particular, they are independent of the value of  $E$ .*

For the corresponding theorem on the single gap we need to define the classical location of the  $j$ -th particle  $\gamma_{j, V}$  by

$$(1.22) \quad \frac{j}{N} = \int_A^{\gamma_{j, V}} \varrho_V(x) dx,$$

similarly to the quantiles  $\gamma_j$  of the semicircle law, see (1.16). We set

$$(1.23) \quad \varrho_j^V := \varrho_V(\gamma_{j,V}), \quad \text{and} \quad \varrho_j := \varrho_{sc}(\gamma_j)$$

to be the limiting densities at the classical location of the  $j$ -th particle. Our main theorem on the  $\beta$ -ensembles is the following.

**THEOREM 1.5** (Gap universality for  $\beta$ -ensembles). *[45, Theorem 2.3] Let  $\beta \geq 1$  and  $V$  be a real analytic potential with  $\inf V'' > -\infty$ , such that  $\varrho_V$  is supported on a single compact interval,  $[A, B]$ ,  $\varrho_V \in C^2(A, B)$ , and that  $V$  is regular. Fix a positive number  $\alpha > 0$ , an integer  $n \in \mathbb{N}$  and a smooth, compactly supported function  $O : \mathbb{R}^n \rightarrow \mathbb{R}$ . Let  $\mu = \mu_V = \mu_{\beta,V}^{(N)}$  be given by (1.10) and let  $\mu_G$  denote the same measure for the Gaussian case,  $V(x) = \frac{1}{2}x^2$ . Then there exist an  $\varepsilon > 0$ , depending only on  $\alpha, \beta$  and the potential  $V$ , and a constant  $C$  depending on  $O$  such that*

$$(1.24) \quad \left| \mathbb{E}^{\mu_V} O \left( (N\varrho_k^V)(x_k - x_{k+1}), (N\varrho_k^V)(x_k - x_{k+2}), \dots, (N\varrho_k^V)(x_k - x_{k+n}) \right) \right. \\ \left. - \mathbb{E}^{\mu_G} O \left( (N\varrho_m)(x_m - x_{m+1}), (N\varrho_m)(x_m - x_{m+2}), \dots, (N\varrho_m)(x_m - x_{m+n}) \right) \right| \\ \leq CN^{-\varepsilon}$$

for any  $k, m \in [\alpha N, (1 - \alpha)N]$  and for any sufficiently large  $N \geq N_0$ , where  $N_0$  depends on  $V, \beta$ , as well as on  $n$  and  $\alpha$ . In particular, the distribution of the rescaled gaps w.r.t.  $\mu_V$  does not depend on the index  $k$  in the bulk.

We point out that Theorem 1.4 holds for any  $\beta > 0$ , but Theorem 1.5 requires  $\beta \geq 1$ . Most likely this is only a technical restriction related to a certain condition in the De Giorgi-Nash-Moser regularity theory that is the backbone of our proof.

### 1.5. Some remarks on the general strategy and on related results.

**1.5.1. Strategy for the universality of correlation functions.** The proof of Theorem 1.2 consists of the following three steps, discussed in Sections 2, 3.1 and 3.2, respectively. This three-step strategy was first introduced in [35].

**Step 1. Local semicircle law and delocalization of eigenvectors:** It states that the density of eigenvalues is given by the semicircle law not only as a weak limit on macroscopic scales (1.3), but also in a strong sense and down to short scales containing only  $N^\varepsilon$  eigenvalues for all  $\varepsilon > 0$ . This will imply the *rigidity of eigenvalues*, i.e., that the eigenvalues are near their classical location in the sense to be made clear in Section 3.1. We also obtain precise estimates on the matrix elements of the Green function which in particular imply complete delocalization of eigenvectors.

**Step 2. Universality for Gaussian divisible ensembles:** The Gaussian divisible ensembles are complex or real Hermitian matrices of the form

$$H_t = e^{-t/2} H_0 + \sqrt{1 - e^{-t}} U,$$

where  $H_0$  is a Wigner matrix and  $U$  is an independent GUE/GOE matrix. The parametrization of  $H_t$  reflects that  $H_t$  is most conveniently obtained by an Ornstein-Uhlenbeck process. There are two methods and both methods imply the bulk universality of  $H_t$  for  $t = N^{-\tau}$  for the entire range of  $0 < \tau < 1$  with different estimates.

- 2a. *Proposition 3.1 of [35] which uses an extension of Johansson's formula [57].*
- 2b. *Local ergodicity of the Dyson Brownian motion (DBM).*

The approach in 2a yields a slightly stronger estimate (no local averaging in the energy) than the approach in 2b, but it works only in the complex Hermitian case. In these notes, we will focus on 2b. As time evolves, the eigenvalues of  $H_t$  evolve according to a system of stochastic differential equations, the Dyson Brownian motion. The distribution of the eigenvalues of  $H_t$  will be written as  $f_t\mu$ , where  $\mu$  is the equilibrium measure (1.5). We will study the evolution equation  $\partial_t f_t = \mathcal{L}f_t$ , where  $\mathcal{L}$  is the generator to the Dirichlet form  $\int |\nabla f|^2 d\mu$ . As time goes to infinity,  $f_t$  converges to constant, i.e. to equilibrium. The key technical question is the speed to local equilibrium.

**Step 3. Approximation by Gaussian divisible ensembles:** It is a simple density argument in the space of matrix ensembles which shows that for any probability distribution of the matrix elements there exists a Gaussian divisible distribution with a small Gaussian component, as in Step 2, such that the two associated Wigner ensembles have asymptotically identical local eigenvalue statistics. The first implementation of this approximation scheme was via a reverse heat flow argument [35]; it was later replaced by the *Green function comparison theorem* [46] that was motivated by the four moment matching condition of [82].

The proof of Theorem 1.4 consists of the following two steps that will be presented in Sections 4.1 and 4.2.

**Step 1. Rigidity of eigenvalues.** This establishes that the location of the eigenvalues are not too far from their classical locations  $\gamma_{j,V}$  determined by the equilibrium density  $\rho_V$ , see (1.22). At this stage the analyticity of  $V$  is necessary since we make use of the loop equation from Johansson [58] and Shcherbina [79].

**Step 2. Uniqueness of local Gibbs measures with logarithmic interactions.** With the precision of eigenvalue location estimates from the Step 1 as an input, the eigenvalue spacing distributions are shown to be given by the corresponding Gaussian ones. (We will take the uniqueness of the spacing distributions as our definition of the uniqueness of Gibbs state.)

There are several similarities and differences between the proofs of Theorem 1.2 and 1.4. Both start with rigidity estimates on eigenvalues and then establish that the local spacing distributions are the same as in the Gaussian cases. The Gaussian divisible ensembles, which play a key role in our theory for noninvariant ensembles, are completely absent for invariant ensembles. The key connection between the two methods, however, is the usage of DBM (or its analogue) in the Steps 2. In Section 3.1, we will first present this idea.

1.5.2. *Strategy for gap universality.* The proofs of Theorems 1.3 and 1.5 require several new ideas. The focus is to analyze the local conditional measures  $\mu_{\mathbf{y}}$  and  $f_{t,\mathbf{y}}\mu_{\mathbf{y}}$  instead of the equilibrium measure  $\mu$  and the DBM evolved measure  $f_t\mu$ . They are obtained by fixing all but  $K$  consecutive points, denoted by  $\mathbf{y}$ . The local measures are Gibbs measures on  $K$  points, denoted by  $\mathbf{x}$ , that are confined to an interval  $J = J_{\mathbf{y}}$  determined by the boundary points of  $\mathbf{y}$ . The external potential,  $V_{\mathbf{y}}$ , of the local measure contains not only the external potential  $V$  from  $\mu$ , but also the interactions between  $\mathbf{x}$  and  $\mathbf{y}$ .

The first step is again to establish rigidity, but this time with respect to the conditional measures  $\mu_{\mathbf{y}}$  and  $(f_t\mu)_{\mathbf{y}} =: f_{t,\mathbf{y}}\mu_{\mathbf{y}}$ , at least for most boundary conditions  $\mathbf{y}$ . Due to the logarithmic interactions,  $V_{\mathbf{y}}$  is not analytic any more and the loop equation is not available, but the rigidity information can still be extracted from the rigidity with respect to the global measure with some additional arguments.

In the second step, which is the key part of the argument, we establish the universality of the gap distribution w.r.t  $\mu_{\mathbf{y}}$  by interpolating between  $\mu_{\mathbf{y}}$  and  $\mu_{\tilde{\mathbf{y}}}$  with two different boundary conditions  $\mathbf{y}$  and  $\tilde{\mathbf{y}}$ . This amounts to estimating the correlation between a gap observable, say  $O(x_i - x_{i+1})$ , and  $V_{\mathbf{y}} - V_{\tilde{\mathbf{y}}}$ . The correlation between particles in log-gases decay only logarithmically, i.e. extremely slowly:

$$(1.25) \quad \frac{\langle x_i; x_j \rangle}{\sqrt{\langle x_i; x_i \rangle \langle x_j; x_j \rangle}} \sim \frac{1}{\log |i - j|}$$

at least if  $i, j$  are far from the boundaries. Here  $\langle \cdot; \cdot \rangle$  denotes the covariance with respect to  $\mu_{\mathbf{y}}$ . The key observation is that correlation between a *gap*  $x_i - x_{i+1}$  and a particle  $x_j$  decays much faster

$$(1.26) \quad \frac{\langle x_i - x_{i+1}; x_j \rangle}{\sqrt{\langle x_i - x_{i+1}; x_i - x_{i+1} \rangle \langle x_j; x_j \rangle}} \sim \frac{1}{|i - j|},$$

because it is essentially the derivative in  $i$  of (1.25). The decay of the gap-gap correlation is even faster.

While the formulas (1.25)–(1.26) are plausible, their rigorous proof is extremely difficult due to the very strong correlations in  $\mu_{\mathbf{y}}$ . We are able to prove a much weaker version of (1.26), practically a decay of order  $|i - j|^{-\varepsilon}$  for some small  $\varepsilon > 0$ , which is sufficient for our purposes. Even the proof of this weaker decay requires quite heavy tools.

We start with a classical observation by Helffer and Sjöstrand [56] that the covariance of any two observables  $f, g$  with respect to a Gibbs measure  $\mu = \exp(-\mathcal{H}(\mathbf{x}))d\mathbf{x}$  can be expressed as

$$(1.27) \quad \langle f(\mathbf{x}); g(\mathbf{x}) \rangle_{\mu} = \int_0^{\infty} \langle \mathbf{h}_t(\mathbf{x}), \nabla g(\mathbf{x}) \rangle_{\mu} dt, \quad \partial_t \mathbf{h}_t = -(\mathcal{L} + \mathcal{H}'')\mathbf{h}_t, \quad \mathbf{h}_0 = \nabla f,$$

where  $\mathcal{L} \geq 0$  is the generator to the Dirichlet form  $\int |\nabla f|^2 d\mu$  and  $\mathcal{H}''$  is the Hessian of the Hamiltonian. The generator  $\mathcal{L}$  in the heat equation in (1.27) creates a time dependent random environment  $\mathbf{x}(t)$  that makes the matrix entries  $(x_i - x_j)^{-2}$  of  $\mathcal{H}''$  time dependent. The solution  $\mathbf{h}_t$  to the equation in (1.27) can be thus represented as a random walk in a time dependent random environment, where the jump rate from site  $i$  to  $j$  is given by  $(x_i(t) - x_j(t))^{-2}$  at time  $t$ . On large scales and for typical realizations of  $\mathbf{x}(t)$ , this jump rate is close to a discretization of the  $\sqrt{-\Delta}$  operator. A discrete version of Di Giorgi-Nash-Moser partial regularity theory [15] then guarantees that the neighboring components of  $\mathbf{h}_t$  are close, which renders the covariance  $\langle \mathbf{h}_t(\mathbf{x}), \nabla g(\mathbf{x}) \rangle_{\mu}$  small, assuming that  $g$  is a function of  $x_i - x_{i+1}$ . In more general terms, the correlation decay (1.26) with  $|i - j|^{-\varepsilon}$  is equivalent to the Hölder regularity a discrete parabolic PDE with random coefficients. This approach has a considerable potential to study log-gases since it connects the problem with one of the deepest phenomena in PDE.

Finally, in the third step, we pass the information on the universality of the gap w.r.t. local measures to the global ones. For the invariant ensemble this step is fairly

straightforward, while for the Wigner ensemble we need to use an approximation step similar to Step 3 in Section 1.5.1.

1.5.3. *Historical remarks.* The method of the proof of Theorem 1.2 is extremely general and the result holds for a much larger class of matrix ensembles with independent entries. Adjacency matrices of the Erdős-Rényi graphs are also covered as long as the matrix is not too sparse, namely more than  $N^{2/3}$  entries of each row are non-zero on average [32, 33]. Although Theorem 1.2 in its current form was proved in [33], the key ideas have been developed through several important steps in [35, 41, 46–48]. In particular, the Wigner-Dyson-Gaudin-Mehta (WDGM) conjecture for complex Hermitian matrices in the form of (1.15) was first proved in Theorem 1.1 of [35]. This result holds whenever the distributions of the matrix elements are smooth. The smoothness requirement for (1.15) was partially removed in [82] and completely removed in [36] but only in the averaged convergence sense (1.14). For a general distribution (1.15) was proved in Theorem 5 in [83]. Although the proof in [83] took a slightly different path, this generalization is an immediate corollary of previous results [44]. These arguments are restricted to the complex Hermitian case since they still use some explicit formula.

The WDGM conjecture for real symmetric matrices in the averaged form of (1.14) was resolved in [41] (a special case, under a restrictive third moment matching condition, was treated in [82]). In [41], a novel idea based on Dyson Brownian motion was introduced. The most difficult case, the real symmetric Bernoulli matrices, was solved in [47], where a “Fluctuation Averaging Lemma” (Theorem 2.16 of the current paper) exploiting cancellation of matrix elements of the Green function was first introduced. A more detailed historical review on Theorem 1.2 was given in Section 11 of [43].

For  $\beta = 2$ , Theorem 1.4 was proved for very general potentials, the best results for  $\beta = 1, 4$  [19, 60, 79] are still restricted to analytic  $V$  with additional conditions. Prior to Theorem 1.4 there was no result for general  $\beta$ , except for the Gaussian case [85].

Given the historical importance of the Wigner surmise, it is somewhat surprising that single gap universality did not receive much attention until very recently. This is probably because our understanding of the Wigner-Dyson-Gaudin-Mehta universality became sufficiently sophisticated only in the last few years to realize the subtle difference between *fixed energy* and *fixed label* universality. In fact, even the GUE case was not known until the very recent paper by Tao [84]. In this work, the complex Hermitian Wigner case was also covered under the condition that the distribution matches that of the GUE to fourth order. Theorem 1.3 is considerable more general, as it applies to any symmetry classes and does not require moment matching. Finally, the single gap universality of the invariant ensembles has not been considered before Theorem 1.5.

1.5.4. *What will not be discussed.* In these lecture notes we focus on the four universality results, Theorem 1.2–1.5, and the necessary background material. There are many related questions on random matrix universality and several of them can be studied with the methods we present here. Here we just list them and give a few relevant references.

- Edge universality for Wigner matrices. See Section 9 of [43] for a summary and also the recent paper [61] that gives the optimal moment condition. The analogues of Theorems 1.3 and 1.5 are presented in [13].



- Universality of eigenvectors. See [59].
- Universality for sample covariance matrices. See [42, 74, 75].
- Sparse matrices and adjacency matrices of Erdős-Rényi graphs. See Section 10 of [43].

1.5.5. *Structure of the lecture notes.* A large part of presentation in these lecture notes is borrowed from other papers and reviews written on the subject [27, 31, 43] and sometimes whole paragraphs of the original articles are verbatim taken over. The overlap is especially large with the review paper [43]; Sections 3.1–4.2 on the Dyson Brownian motion, the Green function comparison theorem and on the analysis of the  $\beta$ -ensemble are repeated without much changes. The local semicircle law (Section 2) is presented here more generally than in [43], following the recent paper [31]. For pedagogical reasons, we will give the proof in a simplified form in Section 2.4 and we only comment on the general proof in Section 2.5. Sections 2.3.3–2.3.4 cover new results on random band matrices based upon the recent work [34]. Section 5 presents an extensive outline of the proofs of Theorems 1.3 and 1.5 on the single gap universality following the very recent paper [45].

We will use the convention that  $C$  and  $c$  denote generic positive constants whose actual values are irrelevant and may change from line to line. For two  $N$ -dependent quantities  $A_N$  and  $B_N$  we use the notation  $A_N \asymp B_N$  to express that  $c \leq A_N/B_N \leq C$ .

**Acknowledgments.** The results in these lecture notes were obtained in collaboration with Horng-Tzer Yau, Benjamin Schlein, Jun Yin, Antti Knowles and Paul Bourgade and in some work, also with Jose Ramirez and Sandrine Péche. This article reports the joint progress with these authors.

This work was partially supported by SFB-TR 12 Grant of the German Research Council.

## 2. Local semicircle law for general Wigner-type matrices

**2.1. Setup and the main results.** Let  $(h_{ij} : i \leq j)$  be a family of independent, complex-valued random variables satisfying  $\mathbb{E}h_{ij} = 0$  and  $h_{ii} \in \mathbb{R}$  for all  $i$ . For  $i > j$  we define  $h_{ij} := \bar{h}_{ji}$ , and denote by  $H = (h_{ij})_{i,j=1}^N$  the  $N \times N$  matrix with entries  $h_{ij}$ . By definition,  $H$  is Hermitian:  $H = H^*$ . (Note that this setup also includes the case of a real symmetric matrix  $H$ .) Such ensembles will be called *general Wigner-type matrices*. Note that we allow for the matrix elements having different distributions. This class of matrices is a natural generalization of the *standard real symmetric Wigner matrices* for which  $h_{ij} \in \mathbb{R}$  are identically distributed, and the *standard complex Hermitian Wigner matrices* for which the off-diagonal elements  $h_{ij} \in \mathbb{C}$  are identically distributed and the diagonal elements  $h_{ii} \in \mathbb{R}$  have their own, but still identical distribution.

The fundamental data of the model is the  $N \times N$  matrix of variances  $S = (s_{ij})$ , where

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

We introduce the parameter  $M := [\max_{i,j} s_{ij}]^{-1}$  that expresses the maximal size of  $s_{ij}$ :

$$(2.1) \quad s_{ij} \leq M^{-1}$$

for all  $i$  and  $j$ . We regard  $N$  as the fundamental parameter and  $M = M_N$  as a function of  $N$ .

$$(2.2) \quad N^\delta \leq M \leq N$$

for some fixed  $\delta > 0$ . We assume that  $S$  is (doubly) stochastic:

$$(2.3) \quad \sum_j s_{ij} = 1$$

for all  $i$ . For standard Wigner matrices,  $h_{ij}$  are identically distributed, hence  $s_{ij} = \frac{1}{N}$  and  $M = N$ . In this presentation, we allow for the matrix elements having different distributions but independence (up to the Hermitian symmetry) is always assumed.

EXAMPLE 2.1. Random band matrices are characterized by translation invariant variances of the form

$$(2.4) \quad s_{ij} = \frac{1}{W} f\left(\frac{|i-j|_N}{W}\right)$$

where  $f$  is a smooth, symmetric probability density on  $\mathbb{R}$ ,  $W$  is a large parameter, called the *band width*, and  $|i-j|_N$  denotes the periodic distance on the discrete torus  $\mathbb{T}$  of length  $N$ . The generalization in higher spatial dimensions is straightforward, in this case the rows and columns of  $H$  are labelled by a discrete  $d$  dimensional torus  $\mathbb{T}_L^d$  of length  $L$  with  $N = L^d$ .

For convenience we assume that the normalized entries

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

have a polynomial decay of arbitrary high degree, i.e. for all  $p \in \mathbb{N}$  there is a constant  $\mu_p$  such that

$$(2.6) \quad \mathbb{E}|\zeta_{ij}|^p \leq \mu_p$$

for all  $N$ ,  $i$ , and  $j$ . We make this assumption to streamline notation, but in fact, our results hold, with the same proof, provided (2.6) is valid for some large but fixed  $p$ . If we strengthen it to uniform subexponential decay, (1.17), then certain estimates will become stronger. In this paper we work with (2.6) for simplicity, but we remark that most of our previous work used (1.17).

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

$$z = E + i\eta$$

without further comment. The eigenvalues of  $H$  in the  $N \rightarrow \infty$  limit are distributed by the celebrated Wigner semicircle law,

$$(2.7) \quad \varrho(x) = \varrho_{sc}(x) := \frac{1}{2\pi} \sqrt{(4-x^2)_+}.$$

and its Stieltjes transform at spectral parameter  $z$  is defined by

$$(2.8) \quad m(z) := \int_{\mathbb{R}} \frac{\varrho(x)}{x-z} dx.$$

To avoid confusion, we remark that  $m$  was denoted by  $m_{sc}$  and  $\varrho$  by  $\varrho_{sc}$  in most of our previous papers. In this section we drop the subscript referring to “semicircle”.

It is well known that the Stieltjes transform  $m$  is the unique solution of

$$(2.9) \quad m(z) + \frac{1}{m(z)} + z = 0$$

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

$$(2.10) \quad m(z) = \frac{-z + \sqrt{z^2 - 4}}{2}.$$

We define the *resolvent* of  $H$  through

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

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

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

for the eigenvalues  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$  of  $H$  is

$$(2.11) \quad m_N(z) := \int_{\mathbb{R}} \frac{\varrho_N(dx)}{x - z} = \frac{1}{N} \text{Tr } G(z).$$

An important parameter of the model is

$$(2.12) \quad \Gamma(z) := \left\| \frac{1}{1 - m^2(z)S} \right\|_{\infty \rightarrow \infty}.$$

Note that  $S$ , being a stochastic matrix, satisfies  $-1 \leq S \leq 1$ , and 1 is an eigenvalue with eigenvector  $\mathbf{e} = N^{-1/2}(1, 1, \dots, 1)$ ,  $S\mathbf{e} = \mathbf{e}$ . We assume that 1 is simple for convenience. Another important parameter is

$$(2.13) \quad \tilde{\Gamma}(z) := \left\| \frac{1}{1 - m^2(z)S} \Big|_{\mathbf{e}^{\perp}} \right\|_{\infty \rightarrow \infty},$$

i.e. the norm of  $1 - m^2S$  restricted to the subspace orthogonal to the constants. Clearly  $\tilde{\Gamma} \leq \Gamma$ .

For standard Wigner matrices we easily obtain that

$$(2.14) \quad \Gamma(z) = \frac{1}{|1 - m^2(z)|} \asymp \frac{1}{\sqrt{\kappa_E + \eta}}, \quad \tilde{\Gamma}(z) = 1,$$

where  $\kappa_E := ||E| - 2|$  denotes the distance of  $E$  to the spectral edges. For generalized Wigner matrices (Definition 1.1) essentially the same relations hold:

$$(2.15) \quad \Gamma(z) = \frac{1}{|1 - m^2(z)|} \asymp \frac{1}{\sqrt{\kappa_E + \eta}}, \quad \tilde{\Gamma}(z) \asymp 1.$$

The following definition introduces a notion of a high-probability bound that is suited for our purposes.

**DEFINITION 2.2** (Stochastic domination). Let

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

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

$$\sup_{u \in U^{(N)}} \mathbb{P} \left[ X^{(N)}(u) > N^{\varepsilon} Y^{(N)}(u) \right] \leq N^{-D}$$

for large enough  $N \geq N_0(\varepsilon, D)$ . Unless stated otherwise, throughout this paper the stochastic domination will always be uniform in all parameters apart from the parameter  $\delta$  in (2.2) and the sequence of constants  $\mu_p$  in (2.6); thus,  $N_0(\varepsilon, D)$  also depends on  $\delta$  and  $\mu_p$ . If  $X$  is stochastically dominated by  $Y$ , uniformly in  $u$ , we use the notation  $X \prec Y$ . Moreover, if for some complex family  $X$  we have  $|X| \prec Y$  we also write  $X = O_{\prec}(Y)$ .

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

$$(2.16) \quad h_{ij} \prec (s_{ij})^{1/2} \prec M^{-1/2},$$

uniformly in  $i$  and  $j$ , so that we may also write  $h_{ij} = O_{\prec}((s_{ij})^{1/2})$ . An easy exercise shows that the relation  $\prec$  satisfies the familiar algebraic rules of order relations, e.g. such relations can be added and multiplied. The definition of  $\prec$  with the polynomial factors  $N^{-\varepsilon}$  and  $N^{-D}$  are tailored for the assumption (2.6). We remark that if (1.17) is assumed, a stronger form of stochastic domination can be introduced but we will not pursue this direction here.

Since

$$\lim_{\eta \rightarrow 0^+} \operatorname{Im} m(E + i\eta) = \pi \varrho(E),$$

the convergence of the Stieltjes transform  $m_N(z)$  to  $m(z)$  as  $N \rightarrow \infty$  will show that the empirical local density of the eigenvalues around the energy  $E$  in a window of size  $\eta$  converges to the semicircle law  $\varrho(E)$ . Therefore the key task is to control  $m_N(z)$  for small  $\eta = \operatorname{Im} z$ .

We now define the lower threshold for  $\eta$  that depends on the energy  $E \in [-10, 10]$

$$(2.17) \quad \tilde{\eta}_E := \min \left\{ \eta : \frac{1}{M\eta} \leq \min \left\{ \frac{M^{-\gamma}}{\tilde{\Gamma}(z)^3}, \frac{M^{-2\gamma}}{\tilde{\Gamma}(z)^4 \operatorname{Im} m(z)} \right\} \text{ for all } z \in [E + i\eta, E + 10i] \right\}.$$

Here  $\gamma > 0$  is a parameter that can be chosen arbitrarily small; for all practical purposes the reader can neglect it. For generalized Wigner matrices,  $M \asymp N$ , from (2.15) we have

$$\tilde{\eta}_E \leq CN^{-1+2\gamma},$$

i.e. we will get the local semicircle law on the smallest possible scale  $\eta \gg N^{-1}$ , modulo a polynomial correction with an arbitrary small exponent. We remark that if we assume subexponential decay (1.17) instead of the polynomial decay (2.6), then the small polynomial correction can be replaced with a logarithmic correction factor.

Finally we define our fundamental control parameter

$$(2.18) \quad \Pi(z) := \sqrt{\frac{\operatorname{Im} m(z)}{M\eta}} + \frac{1}{M\eta}.$$

We can now state the main result of this section, which in this form appeared in [31].

**THEOREM 2.3** (Local semicircle law). *Uniformly in the energy  $|E| \leq 10$  and  $\eta \in [\tilde{\eta}_E, 10]$  we have the bounds*

$$(2.19) \quad |G_{ij}(z) - \delta_{ij}m(z)| \prec \Pi(z) = \sqrt{\frac{\operatorname{Im} m(z)}{M\eta}} + \frac{1}{M\eta}, \quad z = E + i\eta$$

uniformly in  $i, j$ , as well as

$$(2.20) \quad |m_N(z) - m(z)| \prec \frac{1}{M\eta}.$$

We point out two remarkable features of these bounds. The error term for the resolvent entries behaves essentially as  $(M\eta)^{-1/2}$ , with an improvement near the edges where  $\text{Im } m$  vanishes. The error bound for the Stieltjes transform, i.e. for the average of the diagonal resolvent entries, is one order better,  $(M\eta)^{-1}$ , but without improvement near the edge.

Various local semicircle laws have a long history. For standard Wigner matrices (i.e.  $M = N$  and  $\tilde{\Gamma} = 1$ ), the optimal threshold for the smallest possible  $\eta \gg 1/N$  has first been achieved in [39] in the bulk after an intermediate result on scale  $\eta \gg N^{-2/3}$  in [38]. The first effective result near the edge was given in [40]. The optimal power  $(N\eta)^{-1}$  for  $m_N - m$  in the bulk has first been obtained in [47] where the first version of the *fluctuation averaging mechanism* has appeared. The optimal behavior near the edge was first derived in [48]. The case of  $M \ll N$  has first been studied in [46] where the threshold  $\eta \gg 1/M$  in the bulk spectrum,  $|E| < 2$ , has been achieved. The optimal power  $(M\eta)^{-1}$  is proved in [31], where the technique of [46] was combined with the fluctuation averaging mechanism. The edge behavior, i.e. the deterioration of the threshold  $\tilde{\eta}_E$  near the edge has also been extensively studied in [31] and it led to the power  $-3$  of  $\tilde{\Gamma}$  in the definition of  $\tilde{\eta}_E$ , but it is yet unclear whether this power is optimal.

In Section 2.3 we demonstrate that all proofs of the local semicircle law rely on some version of a self-consistent equation. At the beginning this was a scalar equation for  $m_N$ . The self-consistent vector equation for  $v_i = G_{ii} - m$  (see Section 2.3.2) first appeared in [46]. This allowed us to deviate from the identical distributions for  $h_{ij}$  and opened up the route to estimates on individual resolvent matrix elements. Finally, the self-consistent matrix equation for  $\mathbb{E}|G_{xy}|^2$  first appeared in [34] and it yielded diffusion profile for the resolvent (Section 2.3.4).

We now list a few consequences of the local semicircle law. It is an elementary property of the Stieltjes transform that once  $m_N \approx m$  is established for all spectral parameter  $z$  with  $\text{Im } z \geq \tilde{\eta}_E$ , then  $\varrho_N$  and  $\varrho$  coincide on scales larger than  $\tilde{\eta}_E$  near the energy  $E$ . This means that  $\int f \varrho_N \rightarrow \int f \varrho$  for test functions on scale  $\tilde{\eta}_E$  i.e.  $|f'| \ll 1/\tilde{\eta}_E$ . In particular, if  $m_N(z) \rightarrow m(z)$  uniformly on the half planes  $\{z : \text{Im } z \geq \varepsilon\}$  for any fixed  $\varepsilon$ , then  $\varrho_N$  converges to  $\varrho$  weakly. The bound (2.20) asserts much more: it identifies  $\varrho_N$  with  $\varrho$  on scales of order  $1/M$ . In the standard Wigner case,  $M = N$ , it is basically the optimal scale since below scale  $1/N$  the empirical measure  $\varrho_N$  strongly fluctuates due to individual eigenvalues.

Once the local density is identified, we can deduce results on the location of individual eigenvalues and on the counting function. Here we formulate the corresponding statements only for the simpler case when  $s_{ij}$  is comparable with  $1/N$  as they were stated in [48] (apart from the fact that in [48] a subexponential decay was assumed). The precise results in the general case are somewhat more complicated and they can be found in [31].

Let  $\gamma_\alpha = \gamma_{\alpha,N}$  denote the location of the  $\alpha$ -th point under the semicircle law, i.e.,  $\gamma_\alpha$  is defined by

$$(2.21) \quad N \int_{-\infty}^{\gamma_\alpha} \varrho(x) dx = \alpha, \quad 1 \leq \alpha \leq N.$$

We will call  $\gamma_\alpha$  the *classical location* of the  $\alpha$ -th point. Furthermore, for any real energy  $E$ , let

$$\mathbf{n}_N(E) := \frac{1}{N} \# \{ \lambda_\alpha \leq E \}, \quad n(E) := \int_{-\infty}^E \varrho(x) dx$$

be the *empirical counting function* of the eigenvalues and its classical counterpart.

**COROLLARY 2.4** (Rigidity of eigenvalues and limit of the counting function). *[48, Theorem 2.2] For generalized Wigner matrices (Definition 1.1) we have*

$$(2.22) \quad |\lambda_\alpha - \gamma_\alpha| \prec \frac{1}{N^{2/3} \widehat{\alpha}^{1/3}}, \quad \widehat{\alpha} := \min\{\alpha, N + 1 - \alpha\},$$

*uniformly in  $\alpha \in \{1, 2, \dots, N\}$ . Furthermore,*

$$(2.23) \quad |\mathbf{n}_N(E) - n(E)| \prec \frac{1}{N}.$$

*uniformly in  $E \in \mathbb{R}$ .*

We remark that under the stronger decay condition (1.17) instead of (2.6), the  $N^\varepsilon$  factors implicitly present in the notation  $\prec$  can be improved to logarithmic factors, see [48].

Corollary 2.4 is a simple consequence of the Helffer-Sjöstrand formula which translates information on the Stieltjes transform of the empirical measure first to the counting function and then to the locations of eigenvalues. The formula yields the representation

$$(2.24) \quad \begin{aligned} f(\lambda) &= \frac{1}{2\pi} \int_{\mathbb{R}^2} \frac{\partial_{\bar{z}} \tilde{f}(x + iy)}{\lambda - x - iy} dx dy \\ &= \frac{1}{2\pi} \int_{\mathbb{R}^2} \frac{iy f''(x) \chi(y) + i(f(x) + iy f'(x)) \chi'(y)}{\lambda - x - iy} dx dy \end{aligned}$$

for any real valued  $C^2$  function  $f$  on  $\mathbb{R}$ , where  $\chi(y)$  is any smooth cutoff function with bounded derivatives and supported in  $[-1, 1]$  with  $\chi(y) = 1$  for  $|y| \leq 1/2$ . In the applications,  $f$  will be a smoothed version of the characteristic functions of spectral intervals so that  $\sum_j f(\lambda_j)$  counts eigenvalues in that interval. From (2.24) we have

$$\frac{1}{N} \sum_j f(\lambda_j) = \frac{1}{2\pi} \int_{\mathbb{R}^2} \left( iy f''(x) \chi(y) + i(f(x) + iy f'(x)) \chi'(y) \right) m_N(x + iy) dx dy,$$

and then  $m_N$  can be approximated by  $m(x + iy)$ . The details of the argument can be found in [37]. Once (2.23) is established, it is an elementary argument to translate it into the rigidity of the eigenvalues (2.22). The powers  $2/3$  and  $1/3$  in (2.22) stem from the fact that  $\varrho(x)$  has a square root singularity near the spectral edges  $x \approx \pm 2$ , therefore  $n(x) \sim (x + 2)_+^{3/2}$  for  $x$  near  $-2$ .

Although Wigner's semicircle law and its local version only concerns  $m_N = \frac{1}{N} \text{Tr } G$ , we remark that the resolvent matrix elements,  $G_{ii}$  and  $G_{ij}$  also carry important information. For example a good bound on  $G_{ii}$  implies delocalization of the eigenvectors. Indeed, by the spectral decomposition, we have

$$\text{Im } G_{ii}(z) = \eta \sum_{\alpha} \frac{|u_{\alpha}(i)|^2}{(\lambda_{\alpha} - E)^2 + \eta^2},$$

where  $\mathbf{u}_\alpha = (u_\alpha(1), \dots, u_\alpha(N))$  is the (normalized) eigenvector belonging to the eigenvalue  $\lambda_\alpha$ . Choosing the energy  $E$  in the  $\eta$ -vicinity of  $\lambda_\alpha$ , we obtain  $|u_\alpha(i)|^2 \leq \eta \operatorname{Im} G_{ii}(z)$ . Therefore, if  $|G_{ii}(z)| \leq C$  can be shown uniformly for any  $z$  with  $\operatorname{Im} z \geq \eta(N)$  for some  $N$ -dependent threshold  $\eta(N)$ , then we conclude

$$\max_{\alpha} \|\mathbf{u}_\alpha\|_\infty^2 \leq C\eta(N).$$

In the Wigner case, the threshold  $\eta(N)$  is almost  $1/N$ , thus we obtain the complete delocalization of the eigenvectors:

**COROLLARY 2.5.** *For the  $\ell^2$ -normalized eigenvectors  $\mathbf{u}_\alpha$ ,  $\alpha = 1, 2, \dots, N$ , of the standard Wigner matrix, we have*

$$(2.25) \quad \|\mathbf{u}_\alpha\|_\infty \prec N^{-1/2}.$$

This result was first proven in [39] without bounding  $G_{ii}$ . In contrast to the argument in [39], the proof via  $G_{ii}$  can also be easily extended to a general class of Wigner-type matrices. For example, an elementary argument shows that if there are two positive constants  $c$  and  $C$  such that  $c \leq N s_{ij} \leq C$  for all  $i, j$ , then  $\tilde{\Gamma}(z) \leq C$  uniformly in  $z$ , thus  $\tilde{\eta}_E \leq CN^{-1+\gamma}$  and (2.25) holds.

**2.2. Tools.** In this subsection we collect some basic definitions and facts.

**DEFINITION 2.6 (Minors).** For  $\mathbb{T} \subset \{1, \dots, N\}$  we define  $H^{(\mathbb{T})}$  by

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

Moreover, we define the resolvent of  $H^{(\mathbb{T})}$  and its normalized trace through

$$G_{ij}^{(\mathbb{T})}(z) := (H^{(\mathbb{T})} - z)^{-1}_{ij}, \quad m^{(\mathbb{T})}(z) := \frac{1}{N} \operatorname{Tr} G^{(\mathbb{T})}(z).$$

We also set

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

**DEFINITION 2.7 (Partial expectation and independence).** Let  $X \equiv X(H)$  be a random variable. For  $i \in \{1, \dots, N\}$  define the operations  $P_i$  and  $Q_i$  through

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

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

We shall frequently make use of Schur's well-known complement formula, which we write as

$$(2.26) \quad \frac{1}{G_{ii}^{(\mathbb{T})}} = h_{ii} - z - \sum_{k,l}^{(\mathbb{T}i)} h_{ik} G_{kl}^{(\mathbb{T}i)} h_{li},$$

where  $i \notin \mathbb{T} \subset \{1, \dots, N\}$ .

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

column of  $H$ . We added a third identity that relates sums of off-diagonal resolvent entries with a diagonal one. The proofs are elementary.

LEMMA 2.8 (Resolvent identities). *For any real or complex Hermitian matrix  $H$  and  $\mathbb{T} \subset \{1, \dots, N\}$  the following identities hold. If  $i, j, k \notin \mathbb{T}$  and  $i, j \neq k$  then*

$$(2.27) \quad G_{ij}^{(\mathbb{T})} = G_{ij}^{(\mathbb{T}k)} + \frac{G_{ik}^{(\mathbb{T})} G_{kj}^{(\mathbb{T})}}{G_{kk}^{(\mathbb{T})}}, \quad \frac{1}{G_{ii}^{(\mathbb{T})}} = \frac{1}{G_{ii}^{(\mathbb{T}k)}} - \frac{G_{ik}^{(\mathbb{T})} G_{ki}^{(\mathbb{T})}}{G_{ii}^{(\mathbb{T})} G_{ii}^{(\mathbb{T}k)} G_{kk}^{(\mathbb{T})}}.$$

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

$$(2.28) \quad G_{ij}^{(\mathbb{T})} = -G_{ii}^{(\mathbb{T})} \sum_k^{(\mathbb{T}i)} h_{ik} G_{kj}^{(\mathbb{T}i)} = -G_{jj}^{(\mathbb{T})} \sum_k^{(\mathbb{T}j)} G_{ik}^{(\mathbb{T}j)} h_{kj}.$$

Moreover, we have

$$(2.29) \quad \sum_j |G_{ij}^{(\mathbb{T})}|^2 = \frac{1}{\eta} \operatorname{Im} G_{ii}^{(\mathbb{T})},$$

which is sometimes called the Ward identity.

Finally, in order to estimate large sums of independent random variables as in (2.26) and (2.28), we will need a large deviation estimate for linear and quadratic functionals of independent random variables:

THEOREM 2.9 (Large deviation bounds). *Let  $(X_i^{(N)})$  and  $(Y_i^{(N)})$  be independent families of random variables and  $(a_{ij}^{(N)})$  and  $(b_i^{(N)})$  be deterministic; here  $N \in \mathbb{N}$  and  $i, j = 1, \dots, N$ . Suppose that all entries  $X_i^{(N)}$  and  $Y_i^{(N)}$  are independent and satisfy*

$$(2.30) \quad \mathbb{E}X = 0, \quad \mathbb{E}|X|^2 = 1, \quad (\mathbb{E}|X|^p)^{1/p} \leq \mu_p$$

for all  $p \in \mathbb{N}$  and some constants  $\mu_p$ . Then we have the bounds

$$(2.31) \quad \sum_i b_i X_i \prec \left( \sum_i |b_i|^2 \right)^{1/2},$$

$$(2.32) \quad \sum_{i,j} a_{ij} X_i Y_j \prec \left( \sum_{i,j} |a_{ij}|^2 \right)^{1/2},$$

$$(2.33) \quad \sum_{i \neq j} a_{ij} X_i X_j \prec \left( \sum_{i \neq j} |a_{ij}|^2 \right)^{1/2}.$$

*Sketch of the proof.* The estimates (2.31), (2.32), and (2.33) follow from estimating high moments of the left hand sides combined with Chebyshev's inequality. The high moments of (2.31) directly follow from the Marcinkiewicz-Zygmund martingale inequality. High moments of (2.32), and (2.33) are computed by reducing them to (2.31) with a decoupling argument. The details are found in Lemmas B.2, B.3, and B.4 of [30].  $\square$



**2.3. Self-consistent equations on three levels.** By the Schur complement formula (2.26), we have

$$(2.34) \quad G_{ii} = \frac{1}{h_{ii} - z - \sum_{k,l}^{(i)} h_{ik} G_{kl}^{(i)} h_{li}}.$$

The partial expectation with respect to the index  $i$  gives

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

where in the second step we used (2.27). Introducing the notation

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

and recalling (2.3), we get the following self-consistent equation for  $v_i$ :

$$(2.35) \quad v_i = \frac{1}{-z - m - (\sum_k s_{ik} v_k - \Upsilon_i)} - m,$$

where

(2.36)

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

All these quantities depend on  $z$ , which fact is suppressed in the notation. We will show that  $\Upsilon$  is a lower order error term. This is clear about  $h_{ii}$  by (2.16). The term  $A_i$  will be small since off-diagonal resolvent entries are small. Finally,  $Z_i$  will be small by a large deviation estimate Theorem 2.9. Before we present more details, we heuristically show the power of the self-consistent equations.

**2.3.1. A scalar self-consistent equation.** Introduce the notation

$$(2.37) \quad [a] = \frac{1}{N} \sum_i a_i$$

for the *average* of a vector  $(a_i)_{i=1}^N$ . Consider the standard Wigner case,  $s_{ij} = 1/N$ . Then

$$\sum_k s_{ik} v_k = \frac{1}{N} \sum_k v_k = [v] \quad (= m_N - m).$$

Neglecting  $\Upsilon_i$  in (2.35) and taking the average of this relation for each  $i$ , we get

$$(2.38) \quad [v] \approx \frac{1}{-z - m - [v]} - m.$$

Since

$$m = \frac{1}{-z - m}$$

by the defining equation of  $m$ , see (2.9), and this equation is stable under small perturbations, at least away from the spectral edges  $z = \pm 2$ , we obtain from (2.38) that  $[v] \approx 0$ . This means that  $m_N \approx m$  and hence  $\varrho_N \approx \varrho$ , i.e. we obtained the Wigner's original semicircle law.

Historically the semicircle law was first found via the moment method [87] by computing  $\frac{1}{N} \mathbb{E} \text{Tr } H^k$ ,  $k = 1, 2, \dots$  in the  $N \rightarrow \infty$  limit, and identifying them with the moments of the semicircle measure  $\varrho(x)dx$ . In this approach the semicircle law emerges as a result of a somewhat tedious, albeit elementary calculation. A more

direct approach is to take the average in  $i$  of the Schur's formula (2.34) which immediately gives

$$m_N \approx \frac{1}{-z - m_N}$$

after neglecting the error terms  $\Upsilon_i$ . This identifies the limit of  $m_N$  immediately with  $m$ , the (unique) solution to (2.9). Taking the inverse Stieltjes transform then yields the semicircle law in a very direct way.

**2.3.2. A vector self-consistent equation.** If the variances  $s_{ij}$  are not constant or we are interested in individual resolvent matrix elements  $G_{ii}$  instead of their average,  $\frac{1}{N} \text{Tr } G$ , then the scalar equation (2.38) discussed in the previous section is not sufficient. We have to consider (2.35) as a system of equations for the components of the vector  $\mathbf{v} = (v_1, \dots, v_N)$ .

From the explicit formula for  $m$  (2.10), we know that  $|m + z| \geq 1$ . Assuming temporarily that

$$(2.39) \quad \left| \sum_k s_{ik} v_k - \Upsilon_i \right| \leq \frac{1}{2},$$

we can expand the right-hand side of (2.35) around  $-z - m$  up to second order and using the identity (2.9) we obtain

$$(2.40) \quad v_i = m^2 \left( \sum_k s_{ik} v_k - \Upsilon_i \right) + O \left[ \left( \sum_k s_{ik} v_k - \Upsilon_i \right)^2 \right].$$

This is the key equation to study  $v_i$ . Considering all higher order terms and  $\Upsilon_i$  as errors we get a self-consistent equation

$$v_i = m^2 \sum_k s_{ik} v_k + \text{Error}$$

or, with matrix notation for the vector  $\mathbf{v}$ :

$$(2.41) \quad \mathbf{v} = m^2 S \mathbf{v} + \mathcal{E},$$

where  $\mathcal{E}$  represent the vector of error terms. Thus

$$\mathbf{v} = \frac{1}{1 - m^2 S} \mathcal{E}, \quad \text{hence} \quad \|\mathbf{v}\|_\infty \leq \left\| \frac{1}{1 - m^2 S} \right\|_{\infty \rightarrow \infty} \|\mathcal{E}\|_\infty = \Gamma \|\mathcal{E}\|_\infty,$$

and this relation shows how the quantity  $\Gamma$  emerges. If the error term is indeed small and  $\Gamma$  is bounded, then we obtain that  $\|\mathbf{v}\|_\infty = \max |G_{ii} - m|$  is small.

**2.3.3. A matrix self-consistent equation.** The off-diagonal resolvent matrix elements,  $G_{ij}$ , are strongly oscillating quantities and they are not expected to have a deterministic limit. However the local averages of their squares,

$$T_{xy} := \sum_i s_{xi} |G_{iy}|^2,$$

are expected to behave regularly. Note that in the Wigner case, using the identity (2.29), the quantity

$$T_{xy} = \sum_i s_{xi} |G_{iy}|^2 = \frac{1}{N} \sum_i G_{yi} G_{iy}^* = \frac{1}{N} (|G|^2)_{yy} = \frac{1}{N\eta} \text{Im } G_{yy}$$

is independent of  $x$ , but in the general case  $T_{xy}$  carries information on the localization length of the eigenfunctions. In particular, if  $T_{xy}$  decays only beyond a scale

$\ell$ , i.e.  $T_{xy}$  remains comparable with  $T_{xx}$  for  $|x - y| \ll \ell$ , then most eigenfunctions have a localization length at least  $\ell$ .

The self-consistent equation for  $T_{xy}$  can be derived from (2.28):

$$G_{iy} = -G_{ii} \sum_k^{(i)} h_{ik} G_{ky}^{(i)}.$$

Replacing  $G_{ii}$  with  $m$  and taking the square, we have

$$|G_{iy}|^2 \approx |m|^2 \sum_{m,k}^{(i)} h_{ik} G_{ky}^{(i)} \bar{h}_{im} \overline{G_{my}^{(i)}}.$$

Taking partial expectation yields

$$P_i |G_{iy}|^2 \approx |m|^2 \sum_k^{(i)} s_{ik} |G_{ky}^{(i)}|^2 \approx |m|^2 \sum_k s_{ik} |G_{ky}|^2 = |m|^2 T_{iy},$$

where in the second step we used (2.27) to remove the upper index  $i$  and we used that the off-diagonal elements are of smaller order. This formula holds for  $i \neq y$ , for the special case  $i = y$  we have a diagonal element, i.e.

$$P_i |G_{iy}|^2 \approx |m|^2 T_{iy} + |m|^2 \delta_{iy}.$$

Thus we have

$$(2.42) \quad T_{xy} = \sum_i s_{xi} |G_{iy}|^2 = \sum_i s_{xi} P_i |G_{iy}|^2 + \tilde{T}_{xy}, \quad \tilde{T}_{xy} := \sum_i s_{xi} Q_i |G_{iy}|^2.$$

The term  $\tilde{T}_{xy}$  is lower order by a *fluctuation averaging mechanism*, see the explanation after Theorem 2.16. Hence, neglecting this term we have

$$T_{xy} \approx \sum_i s_{xi} P_i |G_{iy}|^2 = |m|^2 \sum_i s_{xi} T_{iy} + |m|^2 s_{xy},$$

i.e. the matrix  $T$  satisfies the self-consistent matrix equation

$$(2.43) \quad T = |m|^2 S T + |m|^2 S + \mathcal{E},$$

where  $\mathcal{E}$  is an error matrix. The solution is

$$(2.44) \quad T = \frac{|m|^2 S}{1 - |m|^2 S} + \frac{1}{1 - |m|^2 S} \mathcal{E},$$

where the first term, given explicitly in terms of the variance matrix  $S$ , gives the leading order behavior for  $T$ :

$$T_{xy} \approx \Theta_{xy}, \quad \Theta := \frac{|m|^2 S}{1 - |m|^2 S}.$$

**2.3.4. Application: diffusion profile for random band matrices.** Depending on the structure of  $S$ , in some cases  $\Theta$  can be computed. Consider, for example, case of the random band matrices, (2.4). Here  $s_{xy}$  and hence  $\Theta_{xy}$  are translation invariant,  $\Theta_{xy} = \theta_{x-y}$ , and the Fourier transform of  $\theta$  is approximately given by

$$(2.45) \quad \theta(p) \approx \frac{1}{\alpha \eta + W^2 D p^2}, \quad D := \frac{1}{2} \int x^2 f(x) dx, \quad \alpha := \frac{2}{\sqrt{4 - E^2}}.$$

This means that the profile of  $\Theta$  is approximately given by a diffusion profile on scale  $W$  with diffusion constant  $D$ .

The analysis of the error term in (2.44) requires estimating the norm of  $(1 - |m|^2 S)^{-1}$ . Note that unlike in the analysis of (2.41), here  $1 - |m|^2 S$  and not  $1 - m^2 S$  has to be inverted. Since  $|m|^2 \sim 1 - C\eta$  for small  $\eta$  and  $S$  has eigenvalue 1, the inverse of  $1 - |m|^2 S$  is very unstable. Fortunately, one can subtract the constant mode in (2.43) before solving the equation, thus eventually only the norm of  $(1 - |m|^2 S)^{-1}$  on the subspace orthogonal to the constants is relevant. Thus the spectral gap of  $S$  plays an important role and we use that for band matrices the gap is of order  $(W/N)^2$ . The details are found in [34], where, among other results, the following theorem was shown:

**THEOREM 2.10** (Diffusion profile). *[34, Theorem 2.4] Let  $H$  be a random band matrix with band width  $W$ , i.e. the variances are given by (2.4). Suppose that  $N \ll W^{5/4}$  and  $(W/N)^2 \leq \eta \leq 1$ . Then*

$$(2.46) \quad |T_{xy} - \Theta_{xy}| \prec \frac{1}{N\eta}, \quad \left| P_x |G_{xy}|^2 - \delta_{xy} |m|^2 - |m|^2 \Theta_{xy} \right| \prec \frac{1}{N\eta} + \frac{\delta_{xy}}{\sqrt{W}}.$$

*All estimates are uniform in  $x, y \in \mathbb{T}$  and in the spectral parameter  $z = E + i\eta$  for  $|E| \leq 2 - \kappa$  and for any fixed  $\kappa > 0$ .*

This theorem identifies  $|G_{xy}|^2$  in two different senses of averaging;  $T_{xy}$  averages in one of the indices, while  $P_x$  takes partial expectation. In both cases the result is essentially  $\Theta_{xy}$ .

The behavior of  $\theta_{x-y} = \Theta_{xy}$  can be analyzed by inverse Fourier transform from (2.45). If  $\eta \ll (W/N)^2$  then  $\theta$  is essentially a constant, i.e. the profile is flat. Conversely, if  $\eta \gg (W/N)^2$  then we get an exponentially decaying profile on the scale  $|x| \sim W\eta^{-1/2}$ . The shape of the profile is therefore nontrivial if and only if  $\eta \gg (W/N)^2$ . The total mass of the profile

$$(2.47) \quad \sum_{x \in \mathbb{T}} \theta_x = \frac{\operatorname{Im} m}{\eta} (1 + O(\eta)) = O(\eta^{-1}),$$

and the average height of the profile is of order  $(N\eta)^{-1}$ . The peak of the exponential profile has height of order  $(W\sqrt{\eta})^{-1}$ , which dominates over the average height if and only if  $\eta \gg (W/N)^2$ . The regime  $\eta \gg (W/N)^2$  corresponds to the regime where  $\eta$  is sufficiently large that the complete delocalization has not taken place, and the profile is mostly concentrated in the region  $|x - y| \leq W\eta^{-1/2} \ll N$ .

These scenarios are best understood in a dynamical picture in which  $\eta$  is decreased down from 1. The ensuing dynamics of  $\theta$  corresponds to the diffusion approximation, where the quantum problem is replaced with a random walk of step-size of order  $W$ . On a configuration space consisting of  $N$  sites, such a random walk will reach an equilibrium beyond time scales  $(N/W)^2$ . Here  $\eta^{-1}$  plays essentially the role of time  $t$ , so that in this dynamical picture equilibrium is reached for  $t \sim \eta^{-1} \gg (N/W)^2$ . Figure 2.1 illustrates this diffusive spreading of the profile for different values of  $\eta$ .

One important consequence of Theorem 2.10 is that it proves delocalization for band matrices with band width  $W \gg N^{4/5}$  (see Corollary 2.3 of [34] for the precise statement). This improves the earlier result from [28, 29] where delocalization for  $W \gg N^{6/7}$  was proved with very different methods. We remark that from the other side it is known that narrow band matrices with  $W \ll N^{1/8}$  are in the localized regime [78]. The conjectured threshold for the phase transition is  $W \sim \sqrt{N}$ , see [52].

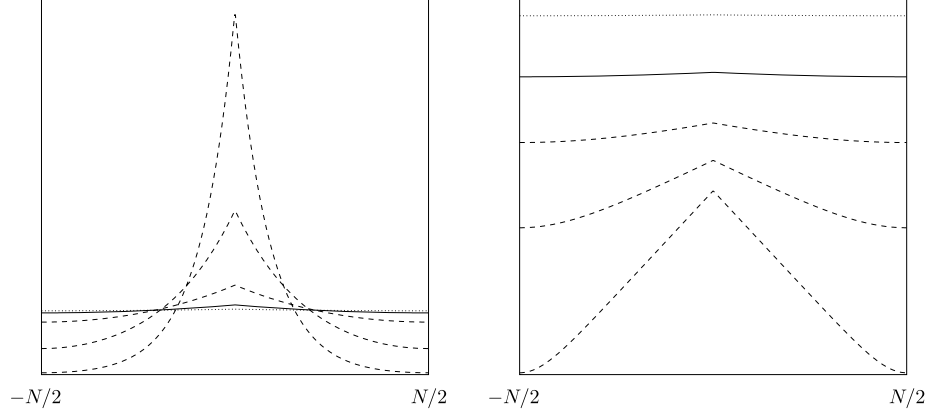


FIGURE 2.1. A plot of the diffusion profile function at five different values of  $\eta$ , where the argument  $x$  ranges over the torus  $\mathbb{T}$ . Left: the graph  $x \mapsto \eta\theta_x$  (see (2.47) for the choice of normalization). Right: the graph  $x \mapsto \log \theta_x$ . Here we chose  $N = 25W$  and  $\eta = 5^{-k}$  for  $k = 1, 2, 3, 4, 5$ . The cases  $k = 1, 2, 3$  (where  $\eta > (W/N)^2$ ) are drawn using dashed lines, the case  $k = 4$  (where  $\eta = (W/N)^2$ ) using solid lines, and the case  $k = 5$  (where  $\eta < (W/N)^2$ ) using dotted lines.

**2.4. Proof of the local semicircle law without using the spectral gap.** In this section we sketch the proof of a weaker version of Theorem 2.3, namely we replace threshold  $\tilde{\eta}_E$  with a larger threshold  $\eta_E$  defined as

$$(2.48) \quad \eta_E := \min \left\{ \eta : \frac{1}{M\eta} \leq \min \left\{ \frac{M^{-\gamma}}{\Gamma(z)^3}, \frac{M^{-2\gamma}}{\Gamma(z)^4 \operatorname{Im} m(z)} \right\} \text{ for all } z \in [E + i\eta, E + 10i] \right\}.$$

This definition is exactly the same as (2.17), but  $\tilde{\Gamma}$  is replaced with the larger quantity  $\Gamma$ , in other words we do not make use of the spectral gap in  $S$ . This will pedagogically simplify the presentation and in Section 2.5 we will comment on the modifications for the stronger result.

We recall that here is no difference between  $\Gamma$  and  $\tilde{\Gamma}$  away from the edges (both are of order 1), so readers interested in the local semicircle law only in the bulk should be content with the simpler proof. Near the spectral edges, however, there is a substantial difference. Note that even in the Wigner case (see (2.14)),  $\eta_E$  is much larger near the spectral edges than the optimal threshold  $\tilde{\eta}_E \sim 1/N$ .

**DEFINITION 2.11.** We call a deterministic nonnegative function  $\Psi \equiv \Psi^{(N)}(z)$  an *admissible control parameter* if we have

$$cM^{-1/2} \leq \Psi \leq M^{-c}$$

for some constant  $c > 0$  and large enough  $N$ . Moreover, we call any (possibly  $N$ -dependent) subset

$$\mathbf{D} = \mathbf{D}^{(N)} \subset \left\{ z : |E| \leq 10, \eta \geq M^{-1+\gamma} \right\}$$

a *spectral domain*.

In this section we will mostly use the spectral domain

$$\mathbf{S} := \left\{ z : |E| \leq 10, \eta \in [\eta_E, 10] \right\}.$$

Define the random control parameters

(2.49)

$$\Lambda_o := \max_{i \neq j} |G_{ij}|, \quad \Lambda_d := \max_i |G_{ii} - m|, \quad \Lambda := \max(\Lambda_o, \Lambda_d), \quad \Theta := |m_N - m|.$$

In the typical regime we will work, all these quantities are small. The key quantity is  $\Lambda$  and we will develop an iterative argument to control it. The first step is an apriori bound:

PROPOSITION 2.12. *We have  $\Lambda \prec M^{-\gamma/3} \Gamma^{-1}$  uniformly in  $\mathbf{S}$ .*

The main estimate behind the proof of Theorem 2.3 for  $\eta \geq \eta_E$  is the following iteration statement:

PROPOSITION 2.13. *Let  $\Psi$  be a control parameter satisfying*

$$(2.50) \quad cM^{-1/2} \leq \Psi \leq M^{-\gamma/3} \Gamma^{-1}.$$

*and fix  $\varepsilon \in (0, \gamma/3)$ . Then on the domain  $\mathbf{S}$  we have the implication*

$$(2.51) \quad \Lambda \prec \Psi \quad \implies \quad \Lambda \prec F(\Psi),$$

*where we defined*

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

The proofs of these two propositions are postponed, we first complete the proof of the local semicircle law.

It is easy to check that, on the domain  $\mathbf{S}$ , if  $\Psi$  satisfies (2.50) then so does  $F(\Psi)$ . We may therefore iterate (2.51). This yields a bound on  $\Lambda$  that is essentially the fixed point of the map  $\Psi \mapsto F(\Psi)$ , which is given by  $\Pi$ , defined in (2.18), (up to the factor  $M^\varepsilon$ ). More precisely, the iteration is started with  $\Psi_0 := M^{-\gamma/3} \Gamma^{-1}$ ; the initial hypothesis  $\Lambda \prec \Psi_0$  is provided by Proposition 2.12. For  $k \geq 1$  we set  $\Psi_{k+1} := F(\Psi_k)$ . Hence from (2.51) we conclude that  $\Lambda \prec \Psi_k$  for all  $k$ . Choosing  $k := \lceil \varepsilon^{-1} \rceil$  yields

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

Since  $\varepsilon$  was arbitrary, we have proved that

$$(2.52) \quad \Lambda \prec \Pi,$$

which is (2.19).

To prove (2.20), i.e. to estimate  $\Theta$ , we rewrite (2.35) as

$$(2.53) \quad -\sum_k s_{ik} v_k + \Upsilon_i = \frac{1}{m + v_i} - \frac{1}{m},$$

and expand the right hand side. Since  $|m| \geq c$  and  $|v_i| \leq \Lambda$ , the expansion is possible on the event where  $\Lambda \ll 1$ , which occurs with very high probability by Proposition 2.12. On this event we get

$$(2.54) \quad m^2 \left( -\sum_k s_{ik} v_k + \Upsilon_i \right) = -v_i + O(\Lambda^2).$$

Averaging in (2.54) yields

$$(2.55) \quad m^2(-[v] + [\Upsilon]) = -[v] + O_{\prec}(\Lambda^2).$$

We will show in Lemma 2.15 in the next section that  $|\Upsilon_i| \prec \Pi$ , but in fact the average  $[\Upsilon]$  is one order better. This is due to the fluctuation averaging phenomenon, and we have

**PROPOSITION 2.14.** *Suppose that  $\Lambda_o \prec \Psi_o$  for some deterministic control parameter  $\Psi_o$  satisfying  $M^{-1/2} \leq \Psi \leq M^{-c}$ . Then  $[\Upsilon] = O_{\prec}(\Psi_o^2)$ .*

We will explain the proof in Section 2.4.4. Using this proposition and (2.52), we get

$$[v] = m^2[v] + O_{\prec}(\Pi^2).$$

Therefore

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

Here in the third step we used the elementary explicit bound  $\operatorname{Im} m \leq C|1 - m^2|$ , and the bound  $\Gamma \geq |1 - m^2|^{-1}$  which follows from the definition of  $\Gamma$  by applying the matrix  $(1 - m^2 S)^{-1}$  to the constant vector. The last step follows from the definition of  $\mathbf{S}$ . Since  $\Theta = |[v]|$ , this concludes the proof of (2.20), and hence of Theorem 2.3 in the regime  $\mathbf{S}$ , i.e. for  $\eta \geq \eta_E$ .  $\square$

In the next sections we explain the proofs of the three propositions used in this argument. We first control the off-diagonal elements, i.e.  $\Lambda_o$ , then we turn to the proof of Propositions 2.12, 2.13 and 2.14.

#### 2.4.1. Basic estimates for $\Lambda_o$ and $\Upsilon_i$ .

**LEMMA 2.15.** *The following statements hold for any spectral domain  $\mathbf{D}$  and admissible control parameter  $\Psi$ . If  $\Lambda \prec \Psi$  then*

$$(2.56) \quad \Lambda_o + |\Upsilon_i| \prec \sqrt{\frac{\operatorname{Im} m + \Psi}{M\eta}}.$$

Moreover, for any fixed ( $N$ -independent)  $\eta > 0$  we have

$$(2.57) \quad \Lambda_o + |\Upsilon_i| \prec M^{-1/2},$$

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

We remark that we could have written (2.56) as

$$(2.58) \quad \Lambda_o + |\Upsilon_i| \prec \sqrt{\frac{\operatorname{Im} m + \Lambda}{M\eta}},$$

but this formulation, while it carries the essence, is literally incorrect since it holds only if  $\Lambda \prec M^{-c}$  has been apriori established.

*Proof of Lemma 2.15.* We first observe that  $\Lambda \prec \Psi \ll 1$  and the positive lower bound  $|m(z)| \geq c$  implies that

$$(2.59) \quad \frac{1}{|G_{ii}|} \prec 1.$$

A simple iteration of the expansion formulas (2.27) concludes that

$$(2.60) \quad |G_{ij}^{(\mathbb{T})}| \prec \Psi, \quad \text{for } i \neq j, \quad |G_{ii}^{(\mathbb{T})}| \prec 1, \quad \frac{1}{|G_{ii}^{(\mathbb{T})}|} \prec 1$$

for any subset  $\mathbb{T}$  of fixed cardinality.

We begin with the first statement in Lemma 2.15. First we estimate  $Z_i$ , which we split as

$$(2.61) \quad |Z_i| \leq \left| \sum_k^{(i)} (|h_{ik}|^2 - s_{ik}) G_{kk}^{(i)} \right| + \left| \sum_{k \neq l}^{(i)} h_{ik} G_{kl}^{(i)} h_{li} \right|.$$

We estimate each term using Theorem 2.9 by conditioning on  $G^{(i)}$  and using the fact that the family  $(h_{ik})_{k=1}^N$  is independent of  $G^{(i)}$ . By (2.31) the first term of (2.61) is stochastically dominated by

$$\left[ \sum_k^{(i)} s_{ik}^2 |G_{kk}^{(i)}|^2 \right]^{1/2} \prec M^{-1/2},$$

where (2.60), (2.1) and (2.3) were used. For the second term of (2.61) we apply Theorem 2.9 (ii) with  $a_{kl} = s_{ik}^{1/2} G_{kl}^{(i)} s_{li}^{1/2}$  and  $X_k = \zeta_{ik}$  (see (2.5)). We find

$$(2.62) \quad \sum_{k,l}^{(i)} s_{ik} |G_{kl}^{(i)}|^2 s_{li} \leq \frac{1}{M} \sum_{k,l} s_{ik} |G_{kl}^{(i)}|^2 = \frac{1}{M\eta} \sum_k^{(i)} s_{ik} \operatorname{Im} G_{kk}^{(i)} \prec \frac{\operatorname{Im} m + \Psi}{M\eta},$$

where in the last step we used (2.27) and the estimate  $1/G_{ii} \prec 1$ . Thus we get

$$(2.63) \quad |Z_i| \prec \sqrt{\frac{\operatorname{Im} m + \Psi}{M\eta}},$$

where we absorbed the bound  $M^{-1/2}$  on the first term of (2.61) into the right-hand side of (2.63), using  $\operatorname{Im} m \geq c\eta$  as follows from an explicit estimate.

Next, we estimate  $\Lambda_o$ . We can iterate (2.28) once to get, for  $i \neq j$ ,

$$(2.64) \quad G_{ij} = -G_{ii} \sum_k^{(i)} h_{ik} G_{kj}^{(i)} = -G_{ii} G_{jj}^{(i)} \left( h_{ij} - \sum_{k,l}^{(ij)} h_{ik} G_{kl}^{(ij)} h_{lj} \right).$$

The term  $h_{ij}$  is trivially  $O_{\prec}(M^{-1/2})$ . In order to estimate the other term, we invoke Theorem 2.9 (iii) with  $a_{kl} = s_{ik}^{1/2} G_{kl}^{(ij)} s_{lj}^{1/2}$ ,  $X_k = \zeta_{ik}$ , and  $Y_l = \zeta_{lj}$ . As in (2.62), we find

$$\sum_{k,l}^{(i)} s_{ik} |G_{kl}^{(ij)}|^2 s_{lj} \prec \frac{\operatorname{Im} m + \Psi}{M\eta},$$

and thus

$$(2.65) \quad \Lambda_o \prec \sqrt{\frac{\operatorname{Im} m + \Psi}{M\eta}},$$

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



In order to estimate  $A_i$  and  $h_{ii}$  in the definition of  $\Upsilon_i$ , we use (2.60) to estimate

$$|A_i| + |h_{ii}| \prec \Lambda_o^2 + M^{-1/2} \leq \Lambda_o + C \sqrt{\frac{\operatorname{Im} m}{M\eta}} \prec \sqrt{\frac{\operatorname{Im} m + \Lambda}{M\eta}},$$

where the second step follows from  $\operatorname{Im} m \geq c\eta$ . Collecting (2.63), (2.65), this completes the proof of (2.56).

The proof of (2.57) is almost identical to that of (2.56). The quantities  $|G_{kk}^{(i)}|$  and  $|G_{kk}^{(ij)}|$  are estimated by the trivial deterministic bound  $\eta^{-1} = O(1)$ . We omit the details.  $\square$

**2.4.2. Sketch of the proof of Proposition 2.12.** The core of the proof is a *continuity argument*. Its basic idea is to establish a *gap* in the range of  $\Lambda$  by proving

**Claim 1.** On the event  $\Lambda \leq M^{-\gamma/4}\Gamma^{-1}$  we actually have the stronger bound  $\Lambda \prec M^{-\gamma/2}\Gamma^{-1}$ .

In other words, for all  $z \in \mathbf{S}$ , with high probability either  $\Lambda \leq M^{-\gamma/2}\Gamma^{-1}$  or  $\Lambda \geq M^{-\gamma/4}\Gamma^{-1}$ . The second step is to show that  $\Lambda \leq M^{-\gamma/2}\Gamma^{-1}$  holds for  $z$  with a large imaginary part  $\eta$ :

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

Thus, for large  $\eta$  the parameter  $\Lambda$  is below the gap. Using the fact that  $\Lambda$  is continuous in  $\eta = \operatorname{Im} z$  and hence cannot jump from one side of the gap to the other, we then conclude that  $\Lambda$  is below the gap for all  $z \in \mathbf{S}$  and this is Proposition 2.12. See Figure 2.2 for an illustration of this argument.

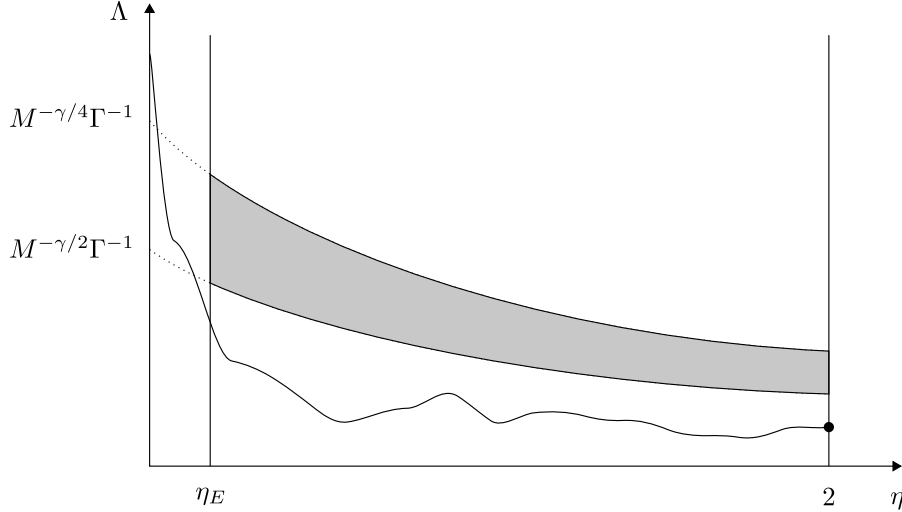


FIGURE 2.2. The  $(\eta, \Lambda)$ -plane for a fixed  $E$  with the graph of  $\eta \rightarrow \Lambda(E + i\eta)$ . The shaded region is forbidden with high probability by Claim 1. The initial estimate, Claim 2, is marked with a black dot. The graph of  $\Lambda = \Lambda(E + i\eta)$  is continuous and lies beneath the shaded region. Note that this method does not control  $\Lambda(E + i\eta)$  in the regime  $\eta \leq \eta_E$ .

Now we explain Claim 1. We will work on the event  $\Lambda \leq M^{-\gamma/4}\Gamma^{-1} \leq M^{-c}$ , where we may invoke (2.56) to estimate  $\Lambda_o$  and  $\Upsilon_i$ . In order to estimate  $\Lambda_d$ , we expand the right-hand side of (2.53) in  $v_i$  to get (2.54). Using (2.56) to estimate  $\Upsilon_i$ , we therefore have

$$(2.66) \quad v_i - m^2 \sum_k s_{ik} v_k = O_{\prec} \left( \Lambda^2 + \sqrt{\frac{\operatorname{Im} m + \Lambda}{M\eta}} \right).$$

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

$$(2.67) \quad \Lambda_d = \max_i |v_i| \prec \Gamma \left( \Lambda^2 + \sqrt{\frac{\operatorname{Im} m + \Lambda}{M\eta}} \right).$$

Together with (2.56) and  $\Gamma \geq c$ , we therefore get

$$(2.68) \quad \Lambda \prec \Gamma \left( \Lambda^2 + \sqrt{\frac{\operatorname{Im} m + \Lambda}{M\eta}} \right).$$

On the event  $\Lambda \leq M^{-\gamma/4}\Gamma^{-1}$  we may estimate

$$\Gamma \Lambda^2 \leq M^{-\gamma/2}\Gamma^{-1}.$$

Moreover, by definition of  $\mathbf{S}$ , we have  $\Gamma^3 \Pi \leq M^{-\gamma}$ . Using the definition of  $\Pi$  (2.18), we therefore get

$$\Gamma \sqrt{\frac{\operatorname{Im} m + \Lambda}{M\eta}} \leq \Gamma \Pi + \Gamma \sqrt{\Gamma^{-1} \Pi} \leq C M^{-\gamma/2} \Gamma^{-1}.$$

Plugging this into (2.68) yields  $\Lambda \prec M^{-\gamma/2}\Gamma^{-1}$ , which is Claim 1.

Finally, we explain Claim 2. We write (2.53) as

$$(2.69) \quad v_i = \frac{m(\sum_k s_{ik} v_k - \Upsilon_i)}{(m^{-1} - \sum_k s_{ik} v_k + \Upsilon_i)}.$$

In the regime  $\eta = 2$ , all resolvent entries are bounded by  $1/2$ , thus  $|v_i| \leq 1$ . Since  $|\Upsilon_i| \prec M^{-1/2}$  by (2.57) and  $|m^{-1}| \geq 2$  we find

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

Using  $|m| \leq 1/2$  we therefore conclude from (2.69) that

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

i.e.  $\Lambda_d = O_{\prec}(M^{-1/2})$ . Together with the estimate on  $\Lambda_o$  from (2.57), we obtained Claim 2.  $\square$

2.4.3. *Sketch of the proof of Proposition 2.13.* This argument is very similar to the proof of Claim 1 in Section 2.4.2. We can work on the event  $\Lambda \leq M^{-\gamma/4}$ , then the bound (2.56) is available to estimate  $\Lambda_o$  and  $\Upsilon_i$ . Next, we estimate  $\Lambda_d$ . We expand the right-hand side of (2.53), we get

$$\psi|v_i| \leq C\psi \left| \sum_k s_{ik} v_k - \Upsilon_i \right| + C\psi \Lambda^2.$$

Using the fluctuation averaging estimate (2.75) explained in the next section, as well as (2.56), we find

$$(2.70) \quad \Lambda_d \prec \Gamma \Psi^2 + \sqrt{\frac{\operatorname{Im} m + \Psi}{M\eta}},$$

which, combined with (2.56), yields

$$(2.71) \quad \Lambda \prec \Gamma \Psi^2 + \sqrt{\frac{\operatorname{Im} m + \Psi}{M\eta}} \leq M^{-\varepsilon} \Psi + \sqrt{\frac{\operatorname{Im} m}{M\eta}} + \frac{M^\varepsilon}{M\eta} = F(\Psi),$$

where in the last step we used the assumption  $\Psi \leq M^{-\gamma/3} \Gamma^{-1}$  and  $\varepsilon \leq \gamma/3$ .  $\square$

2.4.4. *Fluctuation averaging: proof of Proposition 2.14.* The leading error in the self-consistent equation (2.54) for  $v_i$  is  $\Upsilon_i$ . Among the three summands of  $\Upsilon_i$ , see (2.36), typically  $Z_i$  is the largest, thus we typically have

$$(2.72) \quad v_i \prec Z_i$$

in the regime where  $\Gamma$  is bounded. The large deviation bounds in Theorem 2.9 show that  $Z_i \prec \Lambda_o$  and a simple second moment calculation shows that this bound is essentially optimal. On the other hand, (2.29) shows that the typical size of the off-diagonal resolvent matrix elements is at least of order  $(N\eta)^{-1/2}$ , thus the estimate

$$Z_i \prec \Lambda_o \prec \frac{1}{\sqrt{N\eta}}$$

is essentially optimal in the standard Wigner case ( $M = N$ ). Together with (2.72) this shows that the natural bound for  $\Lambda$  is  $(N\eta)^{-1/2}$ , which is also reflected in the bound (2.19).

However, the bound (2.20) for the average,  $m_N - m = [v]$ , is of order  $\Lambda^2 \sim (N\eta)^{-1}$ , i.e. it is one order better than the bound for  $v_i$ . For the purpose of  $[v]$ , it is the average  $[\Upsilon]$  of the leading errors  $\Upsilon_i$  that matters, see (2.55). Since  $Z_i$ , the leading term in  $\Upsilon_i$ , is a fluctuating quantity with zero expectation, the improvement comes from the fact that fluctuations cancel out in the average. The basic example of this phenomenon is the central limit theorem. In our case, however,  $Z_i$  are not independent. In fact, their correlations do not decay, at least in the Wigner case where all indices  $i$  play symmetric role. Thus standard results on central limit theorems for weakly correlated random variables do not apply.

Here we formulate a version of the fluctuation averaging mechanism, taken from [31], that is the most useful for this discussion and we comment on the history afterwards.

We shall perform the averaging with respect to a family of weights  $T = (t_{ik})$  satisfying

$$(2.73) \quad 0 \leq t_{ik} \leq M^{-1}, \quad \sum_k t_{ik} = 1.$$

Typical example weights are  $t_{ik} = s_{ik}$  and  $t_{ik} = N^{-1}$ . Note that in both of these cases  $T$  commutes with  $S$ .

**THEOREM 2.16** (Fluctuation averaging). *Fix a spectral domain  $\mathbf{D}$  and deterministic admissible control parameters  $\Psi, \Psi_o \leq M^{-c}$ . Suppose that  $\Lambda \prec \Psi$ ,  $\Lambda_o \prec \Psi_o$  and the weight  $T = (t_{ik})$  satisfies (2.73). Then we have*

$$(2.74) \quad \sum_k t_{ik} Q_k \frac{1}{G_{kk}} = O_{\prec}(\Psi_o^2), \quad \sum_k t_{ik} Q_k G_{kk} = O_{\prec}(\Psi^2).$$

If in addition  $T$  commutes with  $S$  then

$$(2.75) \quad \sum_k t_{ik} v_k = O_{\prec}(\Gamma \Psi^2), \quad \sum_k t_{ik} (v_k - [v]) = O_{\prec}(\tilde{\Gamma} \Psi^2).$$

The estimates (2.74) and (2.75) are uniform in the index  $i$ .

The first version of the fluctuation averaging mechanism appeared in [47] for the Wigner case, where  $[Z] = N^{-1} \sum_k Z_k$  was bounded by  $\Lambda_o^2$ . Since  $Q_k [G_{kk}]^{-1}$  is essentially  $Z_k$ , see (2.26), this corresponds to the first bound in (2.74). A different proof (with a better bound on the constants) was given in [48]. A conceptually streamlined version of the original proof was extended to sparse matrices [32] and to sample covariance matrices [74]. Finally, an extensive analysis in [30] treated the fluctuation averaging of general polynomials of resolvent entries and identified the order of cancellations depending on the algebraic structure of the polynomial. Moreover, in [30] an additional cancellation effect was found for the quantity  $Q_i |G_{ij}|^2$ . This improvement plays a key role in the proof of Theorem 2.10, see (2.42).

All proofs of the fluctuation averaging theorems rely on computing expectations of high moments of the averages and carefully estimating various terms of different combinatorial structure. In [30] we have developed a Feynman diagrammatic representation for bookkeeping the terms, but this is necessary only for the case of general polynomials. For the special cases stated in Theorem 2.16, the proof is relatively simple and it is presented in Appendix B of [31]. Here we will not repeat the proof, we only indicate the main mechanism by estimating the second moment of the first term in (2.74). The actual proof requires estimating all moments and then use Chebysev inequality to translate the moment estimates to probabilistic estimates standing behind the notation  $O_{\prec}(\Psi_o^2)$  in (2.74).

*Second moment calculation.* First we claim that

$$(2.76) \quad \left| Q_k \frac{1}{G_{kk}} \right| \prec \Psi_o.$$

Indeed, from Schur's complement formula (2.26) we get  $|Q_k (G_{kk})^{-1}| \leq |h_{kk}| + |Z_k|$ . The first term is estimated by  $|h_{kk}| \prec M^{-1/2} \leq \Psi_o$ . The second term is estimated exactly as in (2.61) and (2.62), giving  $|Z_k| \prec \Psi_o$ . In fact, the same bound (2.76) holds if  $G_{kk}$  is replaced with  $G_{kk}^{(\mathbb{T})}$  as long as  $|\mathbb{T}|$  is bounded.

Abbreviate  $X_k := Q_k (G_{kk})^{-1}$  and compute the variance

$$(2.77) \quad \mathbb{E} \left| \sum_k t_{ik} X_k \right|^2 = \sum_{k,l} t_{ik} t_{il} \mathbb{E} X_k \overline{X_l} = \sum_k t_{ik}^2 \mathbb{E} X_k \overline{X_k} + \sum_{k \neq l} t_{ik} t_{il} \mathbb{E} X_k \overline{X_l}.$$

Using the bounds (2.73) on  $t_{ik}$  and (2.76), we find that the first term on the right-hand side of (2.77) is  $O_{\prec}(M^{-1} \Psi_o^2) = O_{\prec}(\Psi_o^4)$ , where we used that  $\Psi_o$  is admissible.

Let us therefore focus on the second term of (2.77). Using the fact that  $k \neq l$ , we apply (2.27) to  $X_k$  and  $X_l$  to get

$$(2.78) \quad \begin{aligned} \mathbb{E}X_k \overline{X_l} &= \mathbb{E}Q_k \left( \frac{1}{G_{kk}} \right) Q_l \left( \overline{\frac{1}{G_{ll}}} \right) \\ &= \mathbb{E}Q_k \left( \frac{1}{G_{kk}^{(l)}} - \frac{G_{kl}G_{lk}}{G_{kk}G_{kk}^{(l)}G_{ll}} \right) Q_l \left( \overline{\frac{1}{G_{ll}^{(k)}} - \frac{G_{lk}G_{kl}}{G_{ll}G_{ll}^{(k)}G_{kk}}} \right). \end{aligned}$$

We multiply out the parentheses on the right-hand side. The crucial observation is that if the random variable  $Y$  is independent of  $i$  (see Definition 2.7) then  $\mathbb{E}Q_i(X)Y = \mathbb{E}Q_i(XY) = 0$ . Hence out of the four terms obtained from the right-hand side of (2.78), the only nonvanishing one is

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

where we used that the denominators are harmless, see (2.59). Together with (2.73), this concludes the proof of  $\mathbb{E}|\sum_k t_{ik}X_k|^2 \prec \Psi_o^4$ , which means that  $\sum_k t_{ik}X_k$  is bounded by  $\Psi_o^2$  in second moment sense.  $\square$

Finally, Proposition 2.14 directly follows from the first estimate in (2.74) with  $t_{ik} = 1/N$ , since from (2.26) we have

$$Q_k \frac{1}{G_{kk}} = h_{kk} - Z_k = \Upsilon_k - A_k = \Upsilon_k + O_{\prec}(\Psi_o^2).$$

Taking the average over  $k$ , we get  $[\Upsilon] = O_{\prec}(\Psi_o^2)$ .  $\square$

**2.5. Remark on the proof of the local semicircle law with using the spectral gap.** In Section 2.4 we proved the local semicircle law, Theorem 2.3, uniformly for  $\eta \geq \eta_E$  instead of the larger regime  $\eta \geq \tilde{\eta}_E$ . The difference between these two thresholds stems from the difference between  $\Gamma$  and  $\tilde{\Gamma}$ , see (2.17) and (2.48).

The bound  $\Gamma$  on the norm of  $(1 - m^2 S)^{-1}$  entered the proof when the self-consistent equation (2.66) was solved. The key idea is to solve the self-consistent equation (2.66) separately on the subspace of constants (the span of the vector  $\mathbf{e}$ ) and on its orthogonal complement  $\mathbf{e}^\perp$ . Roughly speaking, we obtain

$$(2.79) \quad |v_i - [v]| \prec \tilde{\Gamma}(\Lambda^2 + r(\Lambda)), \quad r(\Lambda) := \sqrt{\frac{\text{Im } m + \Lambda}{M\eta}},$$

instead of (2.67). In fact, we can improve this by using the fluctuation averaging. Subtracting the average over  $i$  from the self-consistent equation (2.54) and estimating  $|m|^2 \leq C$ , we have

$$(2.80) \quad |v_i - [v]| \leq C \left| \sum_k s_{ik}(v_k - [v]) - (\Upsilon_i - [\Upsilon]) \right| + O_{\prec}(\Lambda^2) \prec \tilde{\Gamma}\Lambda^2 + r(\Lambda),$$

where in the last step we used the fluctuation averaging estimate (2.75) with  $s_{ik} = t_{ik}$ , and  $|\Upsilon_i| \prec r(\Lambda)$  from (2.56).

On the space of constant vectors, (2.53) becomes a scalar equation for the average  $[v]$ , which can be expanded up to second order. More precisely, assuming

$|v_i| \ll 1$ , we can expand (2.53) up to second order:

$$(2.81) \quad -\sum_k s_{ik} v_k + \Upsilon_i = -\frac{1}{m^2} v_i + \frac{1}{m^3} v_i^2 + O(\Lambda^3).$$

In order to get a closed equation for  $[v]$ , we take the average over  $i$ :

$$(2.82) \quad (1 - m^2)[v] - m^{-1} \frac{1}{N} \sum_i v_i^2 = -m^2 [\Upsilon] + O(\Lambda^3).$$

The nonlinear term is estimated by

$$\frac{1}{N} \sum_i v_i^2 = [v]^2 + \frac{1}{N} \sum_i (v_i - [v])^2 = [v]^2 + O_{\prec} \left( (\tilde{\Gamma} \Lambda^2 + r(\Lambda))^2 \right),$$

where (2.80) was used. The average  $[\Upsilon]$  can be estimated by  $O_{\prec}(\Lambda_o^2)$  as in Proposition 2.14. Moreover,  $\Lambda_o$  is estimated by  $r(\Lambda)$  as in Lemma 2.15. We thus obtain a quadratic equation for the scalar quantity  $[v]$ :

$$(2.83) \quad (1 - m^2)[v] - m^{-1}[v]^2 = O_{\prec} \left( \Lambda^3 + (\tilde{\Gamma} \Lambda^2 + r(\Lambda))^2 \right).$$

The main control parameter in this proof is  $\Theta = |[v]|$ , and the key iterative scheme is formulated in terms of  $\Theta$ . However, many intermediate estimates still involve  $\Lambda$ . In particular, the self-consistent equation (2.53) is effective only in the regime where  $v_i$  is already small and in the calculation above we tacitly used that  $|v_i| \ll 1$ . Hence we need a preparatory step to prove an apriori bound on  $\Lambda$ , essentially showing that  $\Lambda \ll 1$ , in fact we will need  $\Lambda \ll \tilde{\Gamma}^{-1}$  (compare with Proposition 2.12). This proof itself is a continuity argument similar to the proof of Proposition 2.12; now, however, we have to follow  $\Lambda$  and  $\Theta$  in tandem. The main reason why  $\Theta$  is already involved in this part is that we work in larger spectral domain  $\eta \geq \tilde{\eta}_E$ , defined by using  $\tilde{\Gamma}$ . Thus, already in this preparatory step, the self-consistent equation has to be solved separately on the subspace of constants and its orthogonal complement. We will omit here these details.

The second preparatory step is to control  $\Lambda$  in terms of  $\Theta$ , which allows us to express the error terms in the self-consistent equation (2.83) in terms of  $\Theta = |[v]|$ . We first notice that (2.80) implies

$$|v_i| \prec \Theta + \tilde{\Gamma} \Lambda^2 + r(\Lambda),$$

i.e.

$$\Lambda \prec \Theta + \tilde{\Gamma} \Lambda^2 + r(\Lambda).$$

Using the apriori bound  $\Lambda \ll \tilde{\Gamma}^{-1}$ , we get

$$\Lambda \prec \Theta + r(\Lambda).$$

This equation is analogous to (2.51) and can be iterated as in the application of in Proposition 2.13 leading to (2.52) to obtain

$$(2.84) \quad \Lambda \prec \Theta + \sqrt{\frac{\operatorname{Im} m}{M\eta}} + \frac{1}{M\eta}.$$

Plugging this bound into (2.83), we have a self-consistent equation for the scalar quantity  $[v]$  since  $\Theta = |[v]|$ . An elementary calculation, using the apriori bound

$\Lambda \ll \tilde{\Gamma}^{-1}$ , yields

(2.85)

$$(1 - m^2)[v] - m^{-1}[v]^2 = O_{\prec}\left(p(\Theta)^2 + M^{-\gamma/4}\Theta^2\right), \quad p(\Theta) := \sqrt{\frac{\operatorname{Im} m + \Theta}{M\eta}} + \frac{1}{M\eta}.$$

Finally, in the main step we solve the quadratic inequality (2.85) for  $\Theta$ . If we neglect the error term in (2.85), then the equation reduces to

$$(2.86) \quad (1 - m^2)[v] = m^3[v]^2,$$

which has two solutions: either  $[v] = 0$  or  $[v] = (1 - m^2)/m^3$ . Away from the spectral edge we have  $|1 - m^2| \geq c$  with some positive constant  $c$ , so the two solutions are separated and they both are stable under small perturbations. The second solution would mean that  $[v]$  is strictly separated away from zero. But this can be excluded by a continuity argument: for large  $\eta$ , say  $\eta = 2$ , it is easy to prove that  $[v]$  is small. Since  $[v]$  is a continuous function of  $\eta$ , we find that as  $\eta$  decreases continuously,  $[v]$  cannot suddenly jump from a value near zero to a value near  $(1 - m^2)/m^3$ . Thus  $[v]$  must remain in the vicinity of the zero solution to (2.86). This completes the sketch of the proof of Theorem 2.3 for the more general case  $\eta \geq \tilde{\eta}_E$ .  $\square$

### 3. Universality of the correlation functions for Wigner matrices

In this section we explain the sketch of the proof of Theorem 1.2.

#### 3.1. Dyson Brownian motion and the local relaxation flow.

3.1.1. *Concept and results.* The Dyson Brownian motion (DBM) describes the evolution of the eigenvalues of a Wigner matrix as an interacting point process if each matrix element  $h_{ij}$  evolves according to independent (up to symmetry restriction) Brownian motions. We will slightly alter this definition by generating the dynamics of the matrix elements by an Ornstein-Uhlenbeck (OU) process which leaves the standard Gaussian distribution invariant. In the Hermitian case, the OU process for the rescaled matrix elements  $\zeta_{ij} := N^{1/2}h_{ij}$  is given by the stochastic differential equation

$$(3.1) \quad d\zeta_{ij} = d\beta_{ij} - \frac{1}{2}\zeta_{ij}dt, \quad i, j = 1, 2, \dots, N,$$

where  $\beta_{ij}$ ,  $i < j$ , are independent complex Brownian motions with variance one and  $\beta_{ii}$  are real Brownian motions of the same variance. Denote the distribution of the eigenvalues  $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_N)$  of  $H_t$  at time  $t$  by  $f_t(\boldsymbol{\lambda})\mu(d\boldsymbol{\lambda})$  where  $\mu$  is given by (1.10) with the Gaussian potential  $V(x) = x^2/2$ .

Then  $f_t = f_{t,N}$  satisfies [25]

$$(3.2) \quad \partial_t f_t = \mathcal{L} f_t,$$

where

$$(3.3) \quad \mathcal{L} = \mathcal{L}_N := \sum_{i=1}^N \frac{1}{2N} \partial_i^2 + \sum_{i=1}^N \left( -\frac{\beta}{4} \lambda_i + \frac{\beta}{2N} \sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} \right) \partial_i, \quad \partial_i = \frac{\partial}{\partial \lambda_i}.$$

The parameter  $\beta$  is chosen as follows:  $\beta = 2$  for complex Hermitian matrices and  $\beta = 1$  for symmetric real matrices. Our formulation of the problem has already taken into account Dyson's observation that the invariant measure for this dynamics is  $\mu$ . A natural question regarding the DBM is how fast the dynamics reaches equilibrium. Dyson had already posed this question in 1962:

**Dyson's conjecture [25]:** The global equilibrium of DBM is reached in time of order one and the local equilibrium (in the bulk) is reached in time of order  $1/N$ . Dyson further remarked,

The picture of the gas coming into equilibrium in two well-separated stages, with microscopic and macroscopic time scales, is suggested with the help of physical intuition. A rigorous proof that this picture is accurate would require a much deeper mathematical analysis.

We will prove that Dyson's conjecture is correct if the initial data of the flow is a Wigner ensemble, which was Dyson's original interest. Our result in fact is valid for DBM with much more general initial data that we now survey. Briefly, it will turn out that the *global* equilibrium is indeed reached within a time of order one, but *local* equilibrium is achieved much faster if an a-priori estimate on the location of the eigenvalues (also called points) is satisfied. To formulate this estimate, let  $\gamma_j = \gamma_{j,N}$  denote the location of the  $j$ -th point under the semicircle law, i.e.,  $\gamma_j$  is defined by (2.21).

**A-priori Estimate:** There exists an  $\alpha > 0$  such that

$$(3.4) \quad Q = Q_\alpha := \sup_{t \geq N^{-2\alpha}} \frac{1}{N} \int \sum_{j=1}^N (\lambda_j - \gamma_j)^2 f_t(\lambda) \mu(d\lambda) \leq CN^{-1-2\alpha}$$

with a constant  $C$  uniformly in  $N$ . This condition first appeared in [41].

The main result on the local ergodicity of Dyson Brownian motion states that if the a-priori estimate (3.4) is satisfied then the local correlation functions of the measure  $f_t \mu$  are the same as the corresponding ones for the Gaussian measure,  $\mu = f_\infty \mu$ , provided that  $t$  is larger than  $N^{-2\alpha}$ . The  $n$ -point correlation functions of the probability measure  $f_t d\mu$  are defined, similarly to (1.4), by

$$(3.5) \quad p_{t,N}^{(n)}(x_1, x_2, \dots, x_n) = \int_{\mathbb{R}^{N-n}} f_t(\mathbf{x}) \mu(\mathbf{x}) dx_{n+1} \dots dx_N, \quad \mathbf{x} = (x_1, x_2, \dots, x_N).$$

Due to the convention that one can view the locations of eigenvalues as the coordinates of particles, we have used  $\mathbf{x}$ , instead of  $\boldsymbol{\lambda}$ , in the last equation. From now on, we will use both conventions depending on which viewpoint we wish to emphasize. Notice that the probability distribution of the eigenvalues at the time  $t$ ,  $f_t \mu$ , is the same as that of the Gaussian divisible matrix:

$$(3.6) \quad H_t = e^{-t/2} H_0 + (1 - e^{-t})^{1/2} U,$$

where  $H_0$  is the initial generalized Wigner matrix and  $U$  is an independent standard GUE (or GOE) matrix. This establishes the universality of the Gaussian divisible ensembles. The precise statement is the following theorem:

**THEOREM 3.1.** [42, Theorem 2.1] *Suppose that the a-priori estimate (3.4) holds for the solution  $f_t$  of the forward equation (3.2) with some exponent  $\alpha > 0$ . Let  $E \in (-2, 2)$  and  $b > 0$  such that  $[E - b, E + b] \subset (-2, 2)$ . Then for any  $s > 0$ , for any integer  $n \geq 1$  and for any compactly supported continuous test function*



$O : \mathbb{R}^n \rightarrow \mathbb{R}$ , we have

$$(3.7) \quad \lim_{N \rightarrow \infty} \sup_{t \geq N^{-2\alpha+s}} \int_{E-b}^{E+b} \frac{dE'}{2b} \int_{\mathbb{R}^n} d\alpha_1 \dots d\alpha_n O(\alpha_1, \dots, \alpha_n) \\ \times \frac{1}{\varrho_{sc}(E)^n} \left( p_{t,N}^{(n)} - p_{G,N}^{(n)} \right) \\ \times \left( E' + \frac{\alpha_1}{N \varrho_{sc}(E)}, \dots, E' + \frac{\alpha_n}{N \varrho_{sc}(E)} \right) = 0.$$

We can choose  $b = b_N$  depending on  $N$ . In [42] explicit bounds on the speed of convergence and the optimal range of  $b$  were also established. In particular, thanks to the optimal rigidity estimate (2.22) which implies that (3.4) holds with any  $\alpha < 1/2$ , the range of the energy averaging in (3.7) can be reduced to  $b_N \geq N^{-1+\xi}$ ,  $\xi > 0$ , but only for  $t \geq N^{-\xi/8}$  (Theorem 2.3 of [48]).

Theorem 3.1 is a consequence of the following theorem which identifies the averaged gap distribution of the eigenvalues.

**THEOREM 3.2** (Universality of the Dyson Brownian motion for short time).

[42, Theorem 4.1]

Suppose  $\beta \geq 1$  and let  $O : \mathbb{R} \rightarrow \mathbb{R}$  be a smooth function with compact support. Then for any sufficiently small  $\varepsilon > 0$ , independent of  $N$ , there exist constants  $C, c > 0$ , depending only on  $\varepsilon$  and  $O$  such that for any  $J \subset \{1, 2, \dots, N-1\}$  we have

$$(3.8) \quad \left| \int \frac{1}{|J|} \sum_{i \in J} O(N(x_i - x_{i+1})) f_t d\mu - \int \frac{1}{|J|} \sum_{i \in J} O(N(x_i - x_{i+1})) d\mu \right| \\ \leq C N^\varepsilon \sqrt{\frac{N^2 Q}{|J| t}} + C e^{-c N^\varepsilon}.$$

In particular, if the a-priori estimate (3.4) holds with some  $\alpha > 0$  and  $|J|$  is of order  $N$ , then for any  $t > N^{-2\alpha+3\varepsilon}$  the right hand side converges to zero as  $N \rightarrow \infty$ , i.e. the gap distributions for  $f_t d\mu$  and  $d\mu$  coincide.

The test functions can be generalized to

$$(3.9) \quad O\left(N(x_i - x_{i+1}), N(x_{i+1} - x_{i+2}), \dots, N(x_{i+n-1} - x_{i+n})\right)$$

for any  $n$  fixed which is needed to identify higher order correlation functions. In applications,  $J$  is chosen to be the indices of the eigenvalues in the interval  $[E-b, E+b]$  and thus  $|J| \sim Nb$ . This identifies the averaged gap distributions of eigenvalues and thus also identifies the correlation functions after energy averaging. We will not explain here in detail how to pass information from gap distribution to correlation functions (see Section 7 in [42]), but we note that this transfer is relatively easy if both statistics are averaged on a scale larger than the typical fluctuation of a single eigenvalue (which is smaller than  $N^{-1+\varepsilon}$  in the bulk by (2.22)). This concludes Theorem 3.1. Note that the input of this theorem, the apriori estimate (3.4), identifies the location of the eigenvalues only on a scale  $N^{-1/2-\alpha}$  which is much weaker than the  $1/N$  precision for the eigenvalue differences in (3.8).

By the rigidity estimate (2.22), the a-priori estimate (3.4) holds for any  $\alpha < 1/2$  if the initial data of the DBM is a generalized Wigner ensemble. Therefore, Theorem 3.2 holds for any  $t \geq N^{-1+\varepsilon}$  for any  $\varepsilon > 0$  and this establishes Dyson's conjecture for any generalized Wigner matrices.

3.1.2. *Main ideas behind the proof of Theorem 3.2.* The key method is to analyze the relaxation to equilibrium of the dynamics (3.2). This approach was first introduced in Section 5.1 of [41]; the presentation here follows [42].

We start with a short review of the logarithmic Sobolev inequality for a general measure. Let the probability measure  $\mu$  on  $\mathbb{R}^N$  be given by a general Hamiltonian  $\mathcal{H}$ :

$$(3.10) \quad d\mu(\mathbf{x}) = \frac{e^{-N\mathcal{H}(\mathbf{x})}}{Z} d\mathbf{x},$$

In applications  $\mu$  will be the Gaussian equilibrium measure, (1.10) with  $V(x) = x^2/2$ , so we use the same notation  $\mu$ , but the statements in the beginning of this section hold for a general measure. Let  $\mathcal{L}$  be the generator, symmetric with respect to the measure  $d\mu$ , defined by the associated Dirichlet form

$$(3.11) \quad D(f) = D_\mu(f) = - \int f \mathcal{L} f d\mu := \frac{1}{2N} \sum_j \int (\partial_j f)^2 d\mu, \quad \partial_j = \partial_{x_j}.$$

Recall the relative entropy of two probability measures:

$$S(\nu|\mu) := \int \frac{d\nu}{d\mu} \log \left( \frac{d\nu}{d\mu} \right) d\mu.$$

If  $d\nu = f d\mu$ , then we will sometimes use the notation  $S_\mu(f) := S(f\mu|\mu)$ . The entropy can be used to control the total variation norm via the well known inequality

$$(3.12) \quad \int |f - 1| d\mu \leq \sqrt{2S_\mu(f)}.$$

Let  $f_t$  be the solution to the evolution equation

$$(3.13) \quad \partial_t f_t = \mathcal{L} f_t, \quad t > 0,$$

with a given initial condition  $f_0$ . The evolution of the entropy  $S_\mu(f_t) = S(f_t\mu|\mu)$  satisfies

$$(3.14) \quad \partial_t S_\mu(f_t) = -4D_\mu(\sqrt{f_t}).$$

Following Bakry and Émery [6], the evolution of the Dirichlet form satisfies the inequality

$$(3.15) \quad \partial_t D_\mu(\sqrt{f_t}) \leq -\frac{1}{2N} \int (\nabla \sqrt{f_t})(\nabla^2 \mathcal{H}) \nabla \sqrt{f_t} d\mu.$$

If the Hamiltonian is convex, i.e.,

$$(3.16) \quad \nabla^2 \mathcal{H}(\mathbf{x}) = \text{Hess } \mathcal{H}(\mathbf{x}) \geq \varpi \quad \text{for all } \mathbf{x} \in \mathbb{R}^N$$

with some constant  $\varpi > 0$ , then we have

$$(3.17) \quad \partial_t D_\mu(\sqrt{f_t}) \leq -\varpi D_\mu(\sqrt{f_t}).$$

Integrating (3.14) and (3.17) back from infinity to 0, we obtain the *logarithmic Sobolev inequality (LSI)*

$$(3.18) \quad S_\mu(f) \leq \frac{4}{\varpi} D_\mu(\sqrt{f}), \quad f = f_0$$

and the *exponential relaxation of the entropy and Dirichlet form on time scale*  $t \sim 1/\varpi$

$$(3.19) \quad S_\mu(f_t) \leq e^{-t\varpi} S_\mu(f_0), \quad D_\mu(\sqrt{f_t}) \leq e^{-t\varpi} D_\mu(\sqrt{f_0}).$$

As a consequence of the logarithmic Sobolev inequality, we also have the concentration inequality for any  $k$  and  $a > 0$

$$(3.20) \quad \mathbb{P}^\mu (|x_k - \mathbb{E}_\mu(x_k)| > a) \leq 2e^{-\varpi N a^2/2}.$$

We will not use this inequality in this section, but it will become important in Section 4.1.

Returning to the classical ensembles, we assume from now on that  $\mathcal{H}$  is given by (1.10) with  $V(x) = x^2/2$  and the equilibrium measure  $\mu$  is the Gaussian one. We then have the convexity inequality

$$(3.21) \quad \left\langle \mathbf{v}, \nabla^2 \mathcal{H}(\mathbf{x}) \mathbf{v} \right\rangle \geq \frac{1}{2} \|\mathbf{v}\|^2 + \frac{1}{N} \sum_{i < j} \frac{(v_i - v_j)^2}{(x_i - x_j)^2} \geq \frac{1}{2} \|\mathbf{v}\|^2, \quad \mathbf{v} \in \mathbb{R}^N.$$

This guarantees that  $\mu$  satisfies the LSI with  $\varpi = 1/2$  and the relaxation time to equilibrium is of order one.

The key idea is that the relaxation time is in fact much shorter than order one for local observables that depend only on the eigenvalue *differences*. Equation (3.21) shows that the relaxation in the direction  $v_i - v_j$  is much faster than order one provided that  $x_i - x_j$  are close. However, this effect is hard to exploit directly due to that all modes of different wavelengths are coupled. Our idea is to add an auxiliary strongly convex potential  $W(\mathbf{x})$  to the Hamiltonian to “speed up” the convergence to local equilibrium. On the other hand, we will also show that the cost of this speeding up can be effectively controlled if the a-priori estimate (3.4) holds.

The auxiliary potential  $W(\mathbf{x})$  is defined by

$$(3.22) \quad W(\mathbf{x}) := \sum_{j=1}^N W_j(x_j), \quad W_j(x) := \frac{1}{2\tau} (x_j - \gamma_j)^2,$$

i.e. it is a quadratic confinement on scale  $\sqrt{\tau}$  for each eigenvalue near its classical location, where the parameter  $\tau > 0$  will be chosen later. The total Hamiltonian is given by

$$(3.23) \quad \tilde{\mathcal{H}} := \mathcal{H} + W,$$

where  $\mathcal{H}$  is the Gaussian Hamiltonian given by (1.10). The measure with Hamiltonian  $\tilde{\mathcal{H}}$ ,

$$(3.24) \quad d\omega := \omega(\mathbf{x}) d\mathbf{x}, \quad \omega := e^{-N\tilde{\mathcal{H}}} / \tilde{Z},$$

will be called the *local relaxation measure*.

The *local relaxation flow* is defined to be the flow with the generator characterized by the natural Dirichlet form w.r.t.  $\omega$ , explicitly,  $\tilde{\mathcal{L}}$ :

$$(3.25) \quad \tilde{\mathcal{L}} = \mathcal{L} - \sum_j b_j \partial_j, \quad b_j = W'_j(x_j) = \frac{x_j - \gamma_j}{\tau}.$$

We will choose  $\tau \ll 1$  so that the additional term  $W$  substantially increases the lower bound (3.16) on the Hessian, hence speeding up the dynamics so that the relaxation time is at most  $\tau$ .

The idea of adding an artificial potential  $W$  to speed up the convergence appears to be unnatural here. The current formulation is a streamlined version of a much more complicated approach that appeared in [41] and which took ideas from the

earlier work [37]. Roughly speaking, in hydrodynamical limit, the short wavelength modes always have shorter relaxation times than the long wavelength modes. A direct implementation of this idea is extremely complicated due to the logarithmic interaction that couples short and long wavelength modes. Adding a strongly convex auxiliary potential  $W(\mathbf{x})$  shortens the relaxation time of the long wavelength modes, but it does not affect the short modes, i.e. the local statistics, which are our main interest. The analysis of the new system is much simpler since now the relaxation is faster, uniform for all modes. Finally, we need to compare the local statistics of the original system with those of the modified one. It turns out that the difference is governed by  $(\nabla W)^2$  which can be directly controlled by the a-priori estimate (3.4).

Our method for enhancing the convexity of  $\mathcal{H}$  is reminiscent of a standard convexification idea concerning metastable states. To explain the similarity, consider a particle near one of the local minima of a double well potential separated by a local maximum, or energy barrier. Although the potential is not convex globally, one may still study a reference problem defined by convexifying the potential along with the well in which the particle initially resides. Before the particle reaches the energy barrier, there is no difference between these two problems. Thus questions concerning time scales shorter than the typical escape time can be conveniently answered by considering the convexified problem; in particular the escape time in the metastability problem itself can be estimated by using convex analysis. Our DBM problem is already convex, but not sufficiently convex. The modification by adding  $W$  enhances convexity without altering the local statistics. This is similar to the convexification in the metastability problem which does not alter events before the escape time.

3.1.3. *Some details on the proof of Theorem 3.2.* The core of the proof is divided into three theorems. For the flow with generator  $\widetilde{\mathcal{L}}$ , we have the following estimates on the entropy and Dirichlet form.

THEOREM 3.3. *Consider the forward equation*

$$(3.26) \quad \partial_t q_t = \widetilde{\mathcal{L}} q_t, \quad t \geq 0,$$

*with the reversible measure  $\omega$  defined in (3.24). The initial condition  $q_0$  satisfies  $\int q_0 d\omega = 1$ . Then we have the following estimates*

$$(3.27) \quad \partial_t D_\omega(\sqrt{q_t}) \leq -\frac{1}{2\tau} D_\omega(\sqrt{q_t}) - \frac{1}{2N^2} \int \sum_{i,j=1}^N \frac{(\partial_i \sqrt{q_t} - \partial_j \sqrt{q_t})^2}{(x_i - x_j)^2} d\omega,$$

$$(3.28) \quad \frac{1}{2N^2} \int_0^\infty ds \int \sum_{i,j=1}^N \frac{(\partial_i \sqrt{q_s} - \partial_j \sqrt{q_s})^2}{(x_i - x_j)^2} d\omega \leq D_\omega(\sqrt{q_0})$$

*and the logarithmic Sobolev inequality*

$$(3.29) \quad S_\omega(q) \leq C\tau D_\omega(\sqrt{q_0})$$

*with a universal constant  $C$ . Thus the relaxation time to equilibrium is of order  $\tau$ :*

$$(3.30) \quad S_\omega(q_t) \leq e^{-Ct/\tau} S_\omega(q_0), \quad D_\omega(q_t) \leq e^{-Ct/\tau} D_\omega(q_0).$$

PROOF. Denote by  $h = \sqrt{q}$  and we have the equation

$$(3.31) \quad \partial_t D_\omega(h_t) = \partial_t \frac{1}{2N} \int (\nabla h)^2 e^{-N\widetilde{\mathcal{H}}} d\mathbf{x} \leq -\frac{1}{2N} \int \nabla h (\nabla^2 \widetilde{\mathcal{H}}) \nabla h e^{-N\widetilde{\mathcal{H}}} d\mathbf{x}.$$

In our case, (3.21) and (3.22) imply that the Hessian of  $\tilde{\mathcal{H}}$  is bounded from below as

$$(3.32) \quad \nabla h(\nabla^2 \tilde{\mathcal{H}}) \nabla h \geq \frac{C}{\tau} \sum_j (\partial_j h)^2 + \frac{1}{2N} \sum_{i,j} \frac{1}{(x_i - x_j)^2} (\partial_i h - \partial_j h)^2$$

with some positive constant  $C$ . This proves (3.27) and (3.28). The rest can be proved by straightforward arguments analogously to (3.14)–(3.19).  $\square$

Notice that the estimate (3.28) is an additional information that we extracted from the Bakry-Émery argument by using the second term in the Hessian estimate (3.21). It plays a key role in the next theorem.

**THEOREM 3.4** (Dirichlet form inequality). *Let  $q$  be a probability density  $\int q d\omega = 1$  and let  $O : \mathbb{R} \rightarrow \mathbb{R}$  be a smooth function with compact support. Then for any  $J \subset \{1, 2, \dots, N-1\}$  and any  $t > 0$  we have*

$$(3.33) \quad \left| \int \frac{1}{|J|} \sum_{i \in J} O(N(x_i - x_{i+1})) q d\omega - \int \frac{1}{|J|} \sum_{i \in J} O(N(x_i - x_{i+1})) d\omega \right| \leq C \left( t \frac{D_\omega(\sqrt{q})}{|J|} \right)^{1/2} + C \sqrt{S_\omega(q)} e^{-ct/\tau}.$$

*Proof.* For simplicity, we assume that  $J = \{1, 2, \dots, N-1\}$ . Let  $q_t$  satisfy

$$\partial_t q_t = \tilde{\mathcal{L}} q_t, \quad t \geq 0,$$

with an initial condition  $q_0 = q$ . We write

$$(3.34) \quad \begin{aligned} & \int \left[ \frac{1}{|J|} \sum_{i \in J} O(N(x_i - x_{i+1})) \right] (q - 1) d\omega \\ &= \int \left[ \frac{1}{|J|} \sum_{i \in J} O(N(x_i - x_{i+1})) \right] (q - q_t) d\omega + \int \left[ \frac{1}{|J|} \sum_{i \in J} O(N(x_i - x_{i+1})) \right] (q_t - 1) d\omega. \end{aligned}$$

The second term in (3.34) can be estimated by (3.12), the decay of the entropy (3.30) and the boundedness of  $O$ ; this gives the second term in (3.33).

To estimate the first term in (3.34), by the evolution equation  $\partial_t q_t = \tilde{\mathcal{L}} q_t$  and the definition of  $\tilde{\mathcal{L}}$  we have

$$\begin{aligned} & \int \frac{1}{|J|} \sum_{i \in J} O(N(x_i - x_{i+1})) q_t d\omega - \int \frac{1}{|J|} \sum_{i \in J} O(N(x_i - x_{i+1})) q_0 d\omega \\ &= \int_0^t ds \int \frac{1}{|J|} \sum_{i \in J} O'(N(x_i - x_{i+1})) [\partial_i q_s - \partial_{i+1} q_s] d\omega. \end{aligned}$$

From the Schwarz inequality and  $\partial q = 2\sqrt{q}\partial\sqrt{q}$ , the last term is bounded by

$$(3.35) \quad 2 \left[ \int_0^t ds \int_{\mathbb{R}^N} \frac{N^2}{|J|^2} \sum_{i \in J} \left[ O'(N(x_i - x_{i+1})) \right]^2 (x_i - x_{i+1})^2 q_s d\omega \right]^{1/2} \\ \times \left[ \int_0^t ds \int_{\mathbb{R}^N} \frac{1}{N^2} \sum_i \frac{1}{(x_i - x_{i+1})^2} [\partial_i \sqrt{q_s} - \partial_{i+1} \sqrt{q_s}]^2 d\omega \right]^{1/2} \\ \leq C \left( \frac{D_\omega(\sqrt{q_0})t}{|J|} \right)^{1/2},$$

where we have used (3.28) and that  $\left[ O'(N(x_i - x_{i+1})) \right]^2 (x_i - x_{i+1})^2 \leq CN^{-2}$  due to  $O$  being smooth and compactly supported.  $\square$

Alternatively, we could have directly estimated the left hand side of (3.33) by using the total variation norm between  $q\omega$  and  $\omega$ , which in turn could be estimated by the entropy (3.12) and the Dirichlet form using the logarithmic Sobolev inequality, i.e., by

$$(3.36) \quad C \int |q - 1| d\omega \leq C \sqrt{S_\omega(q)} \leq C \sqrt{\tau D_\omega(\sqrt{q})}.$$

However, compared with this simple bound, the estimate (3.33) gains an extra factor  $|J| \sim N$  in the denominator, i.e. it is in terms of Dirichlet form *per particle*. The improvement is due to the fact that the observable in (3.33) depends only on the gap, i.e. *difference* of points. This allows us to exploit the additional term (3.28) gained in the Bakry-Émery argument. This is a manifestation of the general observation that gap-observables behave much better than point-observables.

The final ingredient in proving Theorem 3.2 is the following entropy and Dirichlet form estimates.

**THEOREM 3.5.** *Suppose that (3.21) holds. Let  $\mathfrak{a} > 0$  be fixed and recall the definition of  $Q = Q_{\mathfrak{a}}$  from (3.4). Fix a constant  $\tau \geq N^{-2\mathfrak{a}}$  and consider the local relaxation measure  $\omega$  with this  $\tau$ . Set  $\psi := \omega/\mu$  and let  $g_t := f_t/\psi$ . Suppose there is a constant  $m$  such that*

$$(3.37) \quad S(f_\tau \omega | \omega) \leq CN^m.$$

*Then for any  $t \geq \tau N^\varepsilon$  the entropy and the Dirichlet form satisfy the estimates:*

$$(3.38) \quad S(g_t \omega | \omega) \leq CN^2 Q \tau^{-1}, \quad D_\omega(\sqrt{g_t}) \leq CN^2 Q \tau^{-2}$$

*where the constants depend on  $\varepsilon$  and  $m$ .*

*Proof.* The evolution of the entropy  $S(f_t \mu | \omega) = S_\omega(g_t)$  can be computed explicitly by the formula [88]

$$\partial_t S(f_t \mu | \omega) = -\frac{2}{N} \sum_j \int (\partial_j \sqrt{g_t})^2 \psi d\mu + \int g_t \mathcal{L} \psi d\mu.$$

Hence we have, by using (3.25),

$$\partial_t S(f_t \mu | \omega) = -\frac{2}{N} \sum_j \int (\partial_j \sqrt{g_t})^2 d\omega + \int \widetilde{\mathcal{L}} g_t d\omega + \sum_j \int b_j \partial_j g_t d\omega.$$

Since  $\omega$  is  $\widetilde{\mathcal{L}}$ -invariant and time independent, the middle term on the right hand side vanishes, and from the Schwarz inequality

$$(3.39) \quad \partial_t S(f_t \mu | \omega) \leq -D_\omega(\sqrt{g_t}) + CN \sum_j \int b_j^2 g_t d\omega \leq -D_\omega(\sqrt{g_t}) + CN^2 Q \tau^{-2}.$$

Notice that (3.39) is reminiscent to (3.14) for the derivative of the entropy of the measure  $g_t \omega = f_t \mu$  with respect to  $\omega$ . The difference is, however, that  $g_t$  does not satisfy the evolution equation with the generator  $\widetilde{\mathcal{L}}$ . The last term in (3.39) expresses the error.

Together with the logarithmic Sobolev inequality (3.29), we have

$$(3.40) \quad \partial_t S(f_t \mu | \omega) \leq -D_\omega(\sqrt{g_t}) + CN^2 Q \tau^{-2} \leq -C\tau^{-1} S(f_t \mu | \omega) + CN^2 Q \tau^{-2}.$$

Integrating the last inequality from  $\tau$  to  $t$  and using the assumption (3.37) and  $t \geq \tau N^\varepsilon$ , we have proved the first inequality of (3.38). Using this result and integrating (3.39), we have

$$\int_\tau^t D_\omega(\sqrt{g_s}) ds \leq CN^2 Q \tau^{-1}.$$

By the convexity of the Hamiltonian,  $D_\mu(\sqrt{f_t})$  is decreasing in  $t$ . Since  $D_\omega(\sqrt{g_s}) \leq CD_\mu(\sqrt{f_s}) + CN^2 Q \tau^{-2}$ , this proves the second inequality of (3.38).  $\square$

Finally, we complete the proof of Theorem 3.2. For any given  $t > 0$  we now choose  $\tau := tN^{-\varepsilon}$  and we construct the local relaxation measure  $\omega$  with this  $\tau$ . Set  $\psi = \omega/\mu$  and let  $q := g_t = f_t/\psi$  be the density  $q$  in Theorem 3.4. Then Theorem 3.5, Theorem 3.4 and an easy bound on the entropy  $S_\omega(q) \leq CN^m$  imply that

$$(3.41) \quad \left| \int \frac{1}{N} \sum_{i \in J} O(N(x_i - x_{i+1}))(f_t d\mu - d\omega) \right| \leq C \left( t \frac{D_\omega(\sqrt{q})}{|J|} \right)^{1/2} + C \sqrt{S_\omega(q)} e^{-cN^\varepsilon}.$$

$$\leq C \left( t \frac{N^2 Q}{|J| \tau^2} \right)^{1/2} + C e^{-cN^\varepsilon} \leq CN^\varepsilon \sqrt{\frac{N^2 Q}{|J| t}} + C e^{-cN^\varepsilon},$$

i.e., the local statistics of  $f_t \mu$  and  $\omega$  are the same for any initial data  $f_\tau$  for which (3.37) is satisfied. Applying the same argument to the Gaussian initial data,  $f_0 = f_\tau = 1$ , we can also compare  $\mu$  and  $\omega$ . We have thus proved (3.8) and hence the universality.  $\square$

### 3.2. The Green function comparison theorems and four moment matching.

We now state the Green function comparison theorem, Theorem 3.6. It will quickly lead to Theorem 3.7 stating that the correlation functions of eigenvalues of two matrix ensembles are identical on a scale smaller than  $1/N$  provided that the first four moments of all matrix elements of these two ensembles are almost the same. We will state a limited version for real Wigner matrices for simplicity of presentation.

**THEOREM 3.6** (Green function comparison). *[46, Theorem 2.3] Suppose that we have two  $N \times N$  Wigner matrices,  $H^{(v)}$  and  $H^{(w)}$ , with matrix elements  $h_{ij}$  given by the random variables  $N^{-1/2}v_{ij}$  and  $N^{-1/2}w_{ij}$ , respectively, with  $v_{ij}$  and  $w_{ij}$  satisfying the uniform subexponential decay condition (1.17). We assume that the first four moments of  $v_{ij}$  and  $w_{ij}$  are close to each other in the sense that*

$$(3.42) \quad |\mathbb{E} v_{ij}^s - \mathbb{E} w_{ij}^s| \leq N^{-\delta-2+s/2}, \quad 1 \leq s \leq 4,$$

holds for some  $\delta > 0$ . Then there are positive constants  $C_1$  and  $\varepsilon$ , depending on  $\vartheta$  and  $C_0$  from (1.17) such that for any  $\eta$  with  $N^{-1-\varepsilon} \leq \eta \leq N^{-1}$  and for any  $z_1, z_2$  with  $\text{Im } z_j = \pm\eta$ ,  $j = 1, 2$ , we have

$$(3.43) \quad \lim_{N \rightarrow \infty} \left[ \mathbb{E} \text{Tr } G^{(v)}(z_1) \text{Tr } G^{(v)}(z_2) - \mathbb{E} \text{Tr } G^{(w)}(z_1) \text{Tr } G^{(w)}(z_2) \right] = 0,$$

where  $G^{(v)}$  and  $G^{(w)}$  denotes the Green functions of  $H^{(v)}$  and  $H^{(w)}$ .

Here we formulated Theorem 3.6 for a product of two traces of the Green function, but the result holds for a large class of smooth functions depending on several individual matrix elements of the Green functions as well, see [46] for the precise statement. (The matching condition (3.42) is slightly weaker than in [46], but the proof in [46] without any change yields this slightly stronger version.) This general version of Theorem 3.6 implies the correlation functions of the two ensembles at the scale  $1/N$  are identical:

**THEOREM 3.7** (Correlation function comparison). [46, Theorem 6.4] Suppose the assumptions of Theorem 3.6 hold. Let  $p_{v,N}^{(n)}$  and  $p_{w,N}^{(n)}$  be the  $n$ -point functions of the eigenvalues w.r.t. the probability law of the matrix  $H^{(v)}$  and  $H^{(w)}$ , respectively. Then for any  $|E| < 2$ , any  $n \geq 1$  and any compactly supported continuous test function  $O : \mathbb{R}^n \rightarrow \mathbb{R}$  we have

$$(3.44) \quad \lim_{N \rightarrow \infty} \int_{\mathbb{R}^n} d\alpha_1 \dots d\alpha_n O(\alpha_1, \dots, \alpha_n) \left( p_{v,N}^{(n)} - p_{w,N}^{(n)} \right) \left( E + \frac{\alpha_1}{N}, \dots, E + \frac{\alpha_n}{N} \right) = 0.$$

Notice that these comparison theorems hold for any fixed energy  $E$ , i.e. no averaging in energy is necessary.

The basic idea for proving Theorem 3.6 is similar to Lindeberg's proof of the central limit theorem, where the random variables are replaced one by one with a Gaussian one. We will replace the matrix elements  $v_{ij}$  with  $w_{ij}$  one by one and estimate the effect of this change on the resolvent by a resolvent expansion. The idea of applying Lindeberg's method in random matrices was recently used by Chatterjee [16] for comparing the traces of the Green functions; the idea was also used by Tao and Vu [82] in the context of comparing individual eigenvalue distributions.

The four moment matching condition (3.42) with  $\delta = 0$  first appeared in [82]. For comparison with Theorem 3.6, we state here the main result of [82]. Let  $\lambda_1 < \lambda_2 < \dots < \lambda_N$  and  $\lambda'_1 < \lambda'_2 < \dots < \lambda'_N$  denote the eigenvalues of  $H$  and  $H'$ , respectively. Then the joint distribution of any  $k$ -tuple of eigenvalues on scale  $1/N$  is very close to each other in the following sense:

**THEOREM 3.8** (Four moment theorem for eigenvalues). [82, Theorem 15] Let  $H$  and  $H'$  be two Wigner matrices. Assume that the first four moments of  $h_{ij}$  and  $h'_{ij}$  exactly match and the subexponential decay (1.17) holds for the single entry distributions. Then for any sufficiently small positive  $\varepsilon$  and  $\varepsilon'$  and for any function  $F : \mathbb{R}^k \rightarrow \mathbb{R}$  satisfying  $|\nabla^j F| \leq N^\varepsilon$ ,  $j \leq 5$ , and for any selection of  $k$ -tuple of indices  $i_1, i_2, \dots, i_k \in [\varepsilon N, (1 - \varepsilon)N]$  away from the edge, we have

$$(3.45) \quad \left| \mathbb{E} F\left(N\lambda_{i_1}, N\lambda_{i_2}, \dots, N\lambda_{i_k}\right) - \mathbb{E} F\left(N\lambda'_{i_1}, N\lambda'_{i_2}, \dots, N\lambda'_{i_k}\right) \right| \leq N^{-c_0}$$

with some  $c_0 > 0$ . The exact moment matching condition can be relaxed to (3.42), but  $c_0$  will depend on  $\delta$ .



Note that the arguments in (3.45) are magnified by a factor  $N$  so the result is sufficiently precise to detect individual eigenvalue correlations. Therefore Theorem 3.6 or 3.8 can prove bulk universality for a Wigner matrix  $H$  if another  $H'$  is found, with matching four moments, for which universality is already proved. This will be explained in Section 3.3.

Both Theorem 3.6 and Theorem 3.8 rely on some version of the local semicircle law on the shortest possible scale. There are, however, three main differences between them.

- (i) Theorem 3.6 compares the statistics of eigenvalues of two different ensembles near *fixed energies* while Theorem 3.8 compares the statistics of the  $j_1, j_2, \dots, j_k$ -th eigenvalues for *fixed labels*  $j_1, j_2, \dots, j_k$ .
- (ii) Both theorems are of perturbative nature that require some apriori information. Theorem 3.6 uses a bound on the resolvent matrix entries,  $|G_{ij}(z)|$ , that has already been obtained in the local semicircle law (see, e.g. (2.19)). Theorem 3.8 needs an apriori lower bound on the gaps to exclude possible eigenvalue resonances that may render the expansion unstable. This is achieved by a level repulsion estimate that is the most complicated technical part of [82]. Previously, even more precise level repulsion estimates were obtained in [40] but only for smooth distributions.
- (iii) Theorem 3.6 also compares off-diagonal Green function elements, an information that cannot be obtained from Theorem 3.8. Hence it directly provides information on the eigenvectors as well, see [59] for the development. In fact, once Theorem 3.6 is proved for all energies, it also implies Theorem 3.8. The reason is that we can integrate correlation functions in energy with a precision smaller than the typical size of the gap, hence eigenvalues with a *fixed label* can be identified. This was first done near the edge in [33] and later in the bulk in [59].

*Sketch of the proof of Theorem 3.6.* We fix a bijective ordering map on the index set of the independent matrix elements,

$$\phi : \{(i, j) : 1 \leq i \leq j \leq N\} \rightarrow \{1, \dots, \gamma(N)\}, \quad \gamma(N) := \frac{N(N+1)}{2},$$

and denote by  $H_\gamma$  the Wigner matrix whose matrix elements  $h_{ij}$  follow the  $v$ -distribution if  $\phi(i, j) \leq \gamma$  and they follow the  $w$ -distribution otherwise; in particular  $H^{(v)} = H_0$  and  $H^{(w)} = H_{\gamma(N)}$ .

Consider the telescopic sum of differences of expectations (we present only one resolvent for simplicity of the presentation):

(3.46)

$$\begin{aligned} \mathbb{E} \left( \frac{1}{N} \operatorname{Tr} \frac{1}{H^{(w)} - z} \right) - \mathbb{E} \left( \frac{1}{N} \operatorname{Tr} \frac{1}{H^{(v)} - z} \right) \\ = \sum_{\gamma=1}^{\gamma(N)} \left[ \mathbb{E} \left( \frac{1}{N} \operatorname{Tr} \frac{1}{H_\gamma - z} \right) - \mathbb{E} \left( \frac{1}{N} \operatorname{Tr} \frac{1}{H_{\gamma-1} - z} \right) \right]. \end{aligned}$$

Let  $E^{(ij)}$  denote the matrix whose matrix elements are zero everywhere except at the  $(i, j)$  position, where it is 1, i.e.,  $E_{k\ell}^{(ij)} = \delta_{ik}\delta_{j\ell}$ . Fix a  $\gamma \geq 1$  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

$$\begin{aligned} H_{\gamma-1} &= Q + \frac{1}{\sqrt{N}}V, & V &:= v_{ij}E^{(ij)} + v_{ji}E^{(ji)}, & v_{ji} &:= \bar{v}_{ij}, \\ H_{\gamma} &= Q + \frac{1}{\sqrt{N}}W, & W &:= w_{ij}E^{(ij)} + w_{ji}E^{(ji)}, & w_{ji} &:= \bar{w}_{ij}, \end{aligned}$$

with a matrix  $Q$  that has zero matrix element at the  $(i, j)$  and  $(j, i)$  positions.

By the resolvent expansion,

$$\begin{aligned} S_{\gamma-1} &= R - N^{-1/2}RV R + \dots + N^{-2}(RV)^4 R - N^{-5/2}(RV)^5 S, \\ R &:= \frac{1}{Q - z}, \quad S_{\gamma-1} := \frac{1}{H_{\gamma-1} - z}, \end{aligned}$$

and a similar expression holds for the resolvent  $S_{\gamma}$  of  $H_{\gamma}$ . From the local semi-circle law for individual matrix elements (2.19), the matrix elements of all Green functions  $R, S_{\gamma-1}, S_{\gamma}$  are bounded by  $CN^{2\varepsilon}$ . Although (2.19) is not directly applicable to  $\eta \geq N^{-1+\varepsilon}$ , it is easy to show that

$$|G_{ij}(E + i\eta)| \leq \max_i |G_{ii}(E + i\eta)| \leq \frac{\eta'}{\eta} \max_i |G_{ii}(E + i\eta')|$$

so choosing  $\eta' \sim N^{-1+\varepsilon}$  we can prove a bound for  $\eta$  slightly below  $1/N$  at the expense of a factor  $\eta/\eta'$ . The estimates of the related resolvents  $R, S_{\gamma-1}, S_{\gamma}$  are similar.

By assumption (3.42), the difference between the expectation of matrix elements of  $S_{\gamma-1}$  and  $S_{\gamma}$  is of order  $N^{-2-\delta+C\varepsilon}$ . Since the number of steps,  $\gamma(N)$  is of order  $N^2$ , the difference in (3.46) is of order  $N^2 N^{-2-\delta+C\varepsilon} \ll 1$ , and this proves Theorem 3.6 for a single resolvent. It is very simple to turn this heuristic argument into a rigorous proof and to generalize it to the product of several resolvents. The real difficulty is the input that the resolvent entries can be bounded for a general class of Wigner matrices down to the almost optimal scale  $\eta \sim 1/N$ .

**3.3. Universality for generalized Wigner matrices: putting it together.** In this short section we put the previous information together to prove Theorem 1.2. We first focus on the case when  $b_N$  is independent of  $N$ . Recall that Theorem 3.1 states that the correlation functions of the Gaussian divisible ensemble,

$$(3.47) \quad H_t = e^{-t/2}H_0 + (1 - e^{-t})^{1/2}U,$$

where  $H_0$  is the initial Wigner matrix and  $U$  is an independent standard GUE (or GOE) matrix, are given by the corresponding GUE (or GOE) for  $t \geq N^{-2\alpha+\varepsilon}$  provided that the a-priori estimate (3.4) holds for the solution  $f_t$  of the forward equation (3.2) with some exponent  $\alpha > 0$ . Since the rigidity of eigenvalues (2.22) holds uniformly for all generalized Wigner matrices, we have proved (3.4) for  $\alpha = 1/2 - \varepsilon$  with any  $\varepsilon > 0$ .

From the evolution of the OU process (3.1) for  $v_{ij} = N^{1/2}h_{ij}$  we have

$$(3.48) \quad |\mathbb{E}v_{ij}^s(t) - \mathbb{E}v_{ij}^s(0)| \leq Ct = CN^{-1+3\varepsilon}$$

for  $s = 3, 4$  and with the choice of  $t = N^{-1+3\varepsilon}$ . Furthermore,  $\mathbb{E}h_{ij}^s(t)$  are independent of  $t$  for  $s = 1, 2$  due to  $\mathbb{E}v_{ij}(0) = 0$  and  $\mathbb{E}v_{ij}^2(t) = 1$ . Hence (3.42) is satisfied for the matrix elements of  $H_t$  and  $H_0$  and we can thus use Theorem 3.7 to conclude that the correlation functions of  $H_t$  and  $H_0$  are identical at the scale  $1/N$ . Since the correlation functions of  $H_t$  are given by the corresponding Gaussian case,

we have proved Theorem 1.2 under the condition that the probability distribution of the matrix elements decay subexponentially. Finally, we need a technical cutoff argument to relax the decay condition to (1.13) which we omit here (see Section 7 in [33]).

The argument for  $N$ -dependent  $b = b_N$  in the range  $b_N \geq N^{-1+\xi}$ ,  $\xi > 0$ , is slightly different. For such a small  $b_N$ , (3.7) could be established only for relatively large times,  $t \geq N^{-\xi/8}$ . We cannot therefore compare  $H_0$  with  $H_t$  directly, since the deviation of the third moments of  $v_{ij}(0)$  and  $v_{ij}(t)$  in (3.48) would not satisfy (3.42). Instead, we construct an auxiliary Wigner matrix  $\hat{H}_0$  such that up to the third moment its time evolution  $\hat{H}_t$  under the OU flow (3.47) matches exactly the *original* matrix  $H_0$  and the fourth moments are close even for  $t$  of order  $N^{-\xi/8}$  (see Lemma 3.4 of [47]). Theorem 3.1 will then be applied for  $\hat{H}_t$ , and Theorem 3.6 can be used to compare  $\hat{H}_t$  and  $H_0$ . This completes the sketch of the proof of Theorem 1.2.

#### 4. Universality of the correlation functions for $\beta$ -ensembles

In this section we outline the proof of Theorem 1.4. We will use the notation  $\mu = \mu_{\beta,V}^{(N)}$  for the probability measure defining the general  $\beta$ -ensemble with a potential  $V$  on  $N$  ordered real points  $\lambda_1 \leq \dots \leq \lambda_N$ , see (1.10). We let  $\mathbb{P}^\mu$  and  $\mathbb{E}^\mu$  denote the probability and the expectation with respect to  $\mu$ . The equilibrium density is denoted by  $\varrho = \varrho_V$  and its Stieltjes transform by

$$m(z) = m_V(z) = \int_{\mathbb{R}} \frac{\varrho_V(x)}{x - z} dx.$$

The classical location of the  $k$ -th point will be denoted by  $\gamma_k = \gamma_{k,V}$ , see (1.22). Note that in this section  $\mu$ ,  $\varrho$ ,  $m$  and  $\gamma_k$  refer to the quantities related to the general  $V$  and not the Gaussian one as in the previous sections. This avoids carrying the  $V$  subscripts all the time as in Section 2 we dropped the subscripts  $sc$  referring to the semicircle law.

**4.1. Rigidity estimates.** For simplicity of presentation we assume that the potential  $V$  is convex, i.e.,

$$(4.1) \quad \varpi := \frac{1}{2} \inf_{x \in \mathbb{R}} V''(x) > 0,$$

the equilibrium density  $\varrho(s)$  is supported on a single interval  $[A, B] \subset \mathbb{R}$  and satisfies (1.20) (for the general case, see [12]). The Gaussian case corresponds to  $V(x) = x^2/2$ , in which case the equilibrium density is the semicircle law,  $\varrho_{sc}$ , given by (1.3). Our main result concerning the universality is Theorem 1.4 and similar statement holds for the universality of the averaged gap distributions directly. In fact, the proof of Theorem 1.4 goes via the averaged gap distribution as we now explain.

The first step to prove Theorem 1.4 is the following theorem which provides a rigidity estimate on the location of each individual point. The precision in the bulk is almost down to the optimal scale  $1/N$ , the estimate is weaker near the edges. In the following, we will denote  $\llbracket x, y \rrbracket = \mathbb{N} \cap [x, y]$ .

**THEOREM 4.1.** [11, Theorem 3.1 and Lemma 3.6] *Fix any  $\alpha, \varepsilon > 0$  and assume that (4.1) holds. Then there are constants  $\delta, c_1, c_2 > 0$  such that for any  $N \geq 1$  and*

$$k \in \llbracket \alpha N, (1 - \alpha)N \rrbracket,$$

$$(4.2) \quad \mathbb{P}^\mu (|\lambda_k - \gamma_k| > N^{-1+\varepsilon}) \leq c_1 e^{-c_2 N^\delta}.$$

The following weaker bound is valid close to the spectral edges:

$$(4.3) \quad \mathbb{P}^\mu (|\lambda_k - \gamma_k| \geq N^{-4/15+\varepsilon}) \leq c_1 e^{-c_2 N^\delta}.$$

for any  $k \in \llbracket N^{3/5+\varepsilon}, N - N^{3/5+\varepsilon} \rrbracket$ . Finally, the bound

$$(4.4) \quad \mathbb{P}^\mu (|\lambda_k - \gamma_k| \geq \varepsilon) \leq c_1 e^{-c_2 N^\delta}.$$

holds for any  $k \in \llbracket 1, N \rrbracket$ .

We explain some ideas of the proof of (4.2), the arguments for (4.3) are similar and (4.4) follows from an easy large deviation bound, see [11]. The first ingredient to prove (4.2) is an analysis of the loop equation following Johansson [58] and Shcherbina [79]. The equilibrium density  $\varrho$ , for a convex potential  $V$ , is given by

$$(4.5) \quad \varrho(t) = \frac{1}{\pi} r(t) \sqrt{(t - A)(B - t)} \mathbf{1}_{[A, B]}(t),$$

where  $r$  is a real function that can be extended to an analytic function in  $\mathbb{C}$  and  $r$  has no zero in  $\mathbb{R}$ . Denote by  $s(z) := -2r(z) \sqrt{(A - z)(B - z)}$  where the square root is defined such that its asymptotic value is  $z$  as  $z \rightarrow \infty$ . Recall that the density is the one-point correlation function which is characterized by

$$(4.6) \quad \int_{\mathbb{R}} d\lambda_1 O(\lambda_1) p_N^{(1)}(\lambda_1) = \int_{\mathbb{R}^N} O(\lambda_1) d\mu_{\beta, V}^{(N)}(\lambda), \quad \lambda = (\lambda_1, \lambda_2, \dots, \lambda_N).$$

Let  $\bar{m}_N$  and  $m$  be the Stieltjes transforms of the density  $p_N^{(1)}$  and the equilibrium density  $\varrho$ , respectively. Notice that in Section 2 we have used  $m_N$  to denote the Stieltjes transform of the empirical measure (2.11); here  $\bar{m}_N$  denotes the ensemble average of the analogous quantity.

Define the analytic functions

$$b_N(z) := \int_{\mathbb{R}} \frac{V'(z) - V'(t)}{z - t} (p_1^{(N)} - \varrho)(t) \mathbf{1}(t)$$

and

$$c_N(z) := \frac{1}{N^2} k_N(z) + \frac{1}{N} \left( \frac{2}{\beta} - 1 \right) \bar{m}'_N(z), \quad \text{with} \quad k_N(z) := \text{Var}_\mu \left( \sum_{k=1}^N \frac{1}{z - \lambda_k} \right).$$

Here for complex random variables  $X$  we use the definition that  $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ .

The equation used by Johansson (which can be obtained by a change of variables in (4.6) [58] or by integration by parts [79]), is a variation of the loop equation (see, e.g., [49]) used in the physics literature and it takes the form

$$(4.7) \quad (\bar{m}_N - m)^2 + s(\bar{m}_N - m) + b_N = c_N.$$

Equation (4.7) can be used to express the difference  $\bar{m}_N - m$  in terms of  $(\bar{m}_N - m)^2$ ,  $b_N$  and  $c_N$ . In the regime where  $|\bar{m}_N - m|$  is small, we can neglect the quadratic term. The term  $b_N$  is of the same order as  $|\bar{m}_N - m|$  and is difficult to treat. As

observed in [5, 79], for analytic  $V$ , this term vanishes when we perform a contour integration. So we have roughly the relation

$$(4.8) \quad (\bar{m}_N - m) \sim \frac{1}{N^2} \text{Var}_\mu \left( \sum_{k=1}^N \frac{1}{z - \lambda_k} \right),$$

where we dropped the less important error involving  $\bar{m}'_N(z)/N$  due to the extra  $1/N$  factor. In the convex setting, the variance can be estimated by the logarithmic Sobolev inequality and we immediately obtain an estimate on  $\bar{m}_N - m$ . We then use the Helffer-Sjöstrand formula, see (2.24), to estimate the locations of the particles. This will provide us with an accuracy of order  $N^{-1/2}$  for  $\mathbb{E}^\mu \lambda_k - \gamma_k$ . This argument gives only an estimate on the expectation of the locations of the particles since we only have information on the averaged quantity,  $\bar{m}_N$ . Although it is tempting to use this new accuracy information on the particles to estimate the variance again in (4.8), the information on the expectation on  $\lambda_k$  alone is very difficult to use in a bootstrap argument. To estimate the variance of a non-trivial function of  $\lambda_k$  we need high probability estimates on  $\lambda_k$ .

The key idea in this section is the observation that the accuracy information on the  $\lambda$ 's can be used to improve the local convexity of the measure  $\mu$  in the direction involving the *differences* of  $\lambda$ 's. To explain this idea, we compute the Hessian of the Hamiltonian of  $\mu$ :

$$\langle \mathbf{v}, \nabla^2 \mathcal{H}(\lambda) \mathbf{v} \rangle \geq \varpi \|\mathbf{v}\|^2 + \frac{1}{N} \sum_{i < j} \frac{(v_i - v_j)^2}{(\lambda_i - \lambda_j)^2}.$$

The naive lower bound on  $\nabla^2 \mathcal{H}$  is  $\varpi$ , but for a typical  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_N)$  it is in fact much better in most directions. To see this effect, suppose we know  $|\lambda_i - \lambda_j| \lesssim M/N$  with some  $M$  for any  $i, j \in I_k^M$ , where  $I_k^M := \llbracket k - M, k + M \rrbracket$ . Then for  $\mathbf{v} = (v_{k-M}, \dots, v_{k+M})$  with  $\sum_j v_j = 0$  we have

$$(4.9) \quad \langle \mathbf{v}, \nabla^2 \mathcal{H}(\lambda) \mathbf{v} \rangle \geq \frac{N}{M^2} \sum_{i, j \in I_k^M} (v_i - v_j)^2 \geq C \frac{N}{M} \sum_j v_j^2.$$

This improves the convexity of the Hessian to  $N/M$  on the hyperplane  $\sum_j v_j = 0$ . Let

$$\lambda_k^{[M]} := \frac{1}{|I_k^M|} \sum_{j \in I_k^M} \lambda_j = \frac{1}{2M+1} \sum_{j \in I_k^M} \lambda_j$$

denote the block average of the locations of particles and rewrite

$$\lambda_k - \lambda_k^{[N^{1-\varepsilon}]} = \sum_j \left( \lambda_k^{[M_j]} - \lambda_k^{[M_{j+1}]} \right)$$

as a telescopic sum with an appropriate sequence of  $M_1 = 0, M_2, \dots$  with the property that  $M_j/M_{j-1} \leq N^\varepsilon$ . We can now use the improved concentration on the hyperplane  $\sum_j v_j = 0$  to the variables  $\lambda_k^{[M_j]} - \lambda_k^{[M_{j+1}]}$  to control the fluctuation of  $\lambda_k - \lambda_k^{[N^{1-\varepsilon}]}$ . Since the fluctuation of  $\lambda_k^{[N^{1-\varepsilon}]}$  is very small for small  $\varepsilon$ , we finally arrive at the estimate

$$(4.10) \quad \mathbb{P}^\mu (|\lambda_k - \mathbb{E}^\mu \lambda_k| > a) \leq C e^{-CN^2 a^2/M}.$$

From (4.10) we thus have that  $|\lambda_k - \mathbb{E}^\mu \lambda_k| \lesssim \sqrt{M}/N$  with high probability. This improves the starting accuracy  $|\lambda_i - \lambda_j| \lesssim M/N$  for  $i, j \in I_k^M$  to  $|\lambda_i - \lambda_j| \lesssim M'/N$

with some  $M' \ll M$ , provided that we can prove that  $|\mathbb{E}^\mu(\lambda_i - \lambda_j)| \ll M'/N$ . But the last inequality involves only expectations and it will follow from the analysis of the loop equation (4.7) we just mentioned above. Starting from  $M = N$ , this procedure can be repeated by decreasing  $M$  step by step until we get the optimal accuracy,  $M \sim O(1)$ . The implementation of this argument in [11] is somewhat different from this sketch due to various technical issues, but it follows the same basic idea.

**4.2. The local equilibrium measure.** Having completed the first step, the rigidity estimate, we now focus on the second step, i.e. on the uniqueness of the local Gibbs measure. Let  $0 < \kappa < 1/2$ . Choose  $q \in [\kappa, 1 - \kappa]$  and set  $L = [Nq]$  (the integer part). Fix an integer  $K = N^k$  with  $k < 1$ . We will study the local spacing statistics of  $K$  consecutive particles

$$\{\lambda_j : j \in I\}, \quad I = I_L := \llbracket L + 1, L + K \rrbracket.$$

These particles are typically located near  $E_q$  determined by the relation

$$\int_{-\infty}^{E_q} \varrho(t) dt = q.$$

Note that  $|\gamma_L - E_q| \leq C/N$ .

We will distinguish the inside and outside particles by renaming them as

$$(4.11) \quad (\lambda_1, \lambda_2, \dots, \lambda_N) := (y_1, \dots, y_L, x_{L+1}, \dots, x_{L+K}, y_{L+K+1}, \dots, y_N) \in \Xi^{(N)},$$

but note that they keep their original indices. The notation  $\Xi^{(N)}$  refers to the simplex  $\{\mathbf{z} : z_1 < z_2 < \dots < z_N\}$  in  $\mathbb{R}^N$ . In short we will write

$$\mathbf{x} = (x_{L+1}, \dots, x_{L+K}), \quad \text{and} \quad \mathbf{y} = (y_1, \dots, y_L, y_{L+K+1}, \dots, y_N),$$

all in increasing order, i.e.  $\mathbf{x} \in \Xi^{(K)}$  and  $\mathbf{y} \in \Xi^{(N-K)}$ . We will refer to the  $y$ 's as *external points* and to the  $x$ 's as *internal points*.

We will fix the external points (also called as boundary conditions) and study conditional measures on the internal points. We define the *local equilibrium measure* on  $\mathbf{x}$  with fixed boundary condition  $\mathbf{y}$  by

$$(4.12) \quad \mu_{\mathbf{y}}(d\mathbf{x}) = \mu_{\mathbf{y}}(\mathbf{x}) d\mathbf{x}, \quad \mu_{\mathbf{y}}(\mathbf{x}) := \mu(\mathbf{y}, \mathbf{x}) \left[ \int \mu(\mathbf{y}, \mathbf{x}) d\mathbf{x} \right]^{-1}.$$

Note that for any fixed  $\mathbf{y} \in \Xi^{(N-K)}$ , the measure  $\mu_{\mathbf{y}}$  is supported on configurations of  $K$  points  $\mathbf{x} = \{x_j\}_{j \in I}$  located in the interval  $[y_L, y_{L+K+1}]$ .

The Hamiltonian  $\mathcal{H}_{\mathbf{y}}$  of the measure  $\mu_{\mathbf{y}}(d\mathbf{x}) \sim \exp(-\beta N \mathcal{H}_{\mathbf{y}}(\mathbf{x})) d\mathbf{x}$  is given by

$$(4.13) \quad \mathcal{H}_{\mathbf{y}}(\mathbf{x}) := \sum_{i \in I} \frac{1}{2} V_{\mathbf{y}}(x_i) - \frac{1}{N} \sum_{\substack{i, j \in I \\ i < j}} \log |x_j - x_i| \quad \text{with} \quad V_{\mathbf{y}}(x) := V(x) - \frac{2}{N} \sum_{j \notin I} \log |x - y_j|.$$

We now define the set of *good boundary configurations* with a parameter  $\delta = \delta(N) > 0$

$$(4.14) \quad \mathcal{G}_\delta = \mathcal{G} := \left\{ \mathbf{y} \in \Xi^{(N-K)} : |y_j - \gamma_j| \leq \delta, \forall j \in \llbracket N\kappa/2, L \rrbracket \cup \llbracket L + K + 1, N(1 - \kappa/2) \rrbracket \right\},$$

where  $\kappa$  is a small constant to cutoff points near the spectral edges. Some rather weak additional conditions for  $\mathbf{y}$  near the spectral edges will also be needed. They

can be built in the definition of  $\mathcal{G}$  based upon the bounds (4.3) and (4.4) but we will neglect this issue here.

Let  $\sigma$  and  $\mu$  be two measures of the form (1.10) with potentials  $W$  and  $V$  and densities  $\varrho = \varrho_W$  and  $\varrho_V$ , respectively. For our purpose  $W(x) = x^2/2$ , i.e.,  $\sigma$  is the Gaussian  $\beta$ -ensemble and its density  $\varrho_W(t) = \frac{1}{2\pi}(4-t^2)_+^{1/2}$  is the Wigner semicircle law. Let the sequence  $\gamma_j$  be the classical locations for  $\mu$  and the sequence  $\theta_j$  be the classical locations for  $\sigma$ . Similarly to the construction of the measure  $\mu_{\mathbf{y}}$ , for any positive integer  $L' \in \llbracket 1, N-K \rrbracket$  we can construct the measure  $\sigma_{\theta}$  conditioned that the particles outside are given by the classical locations  $\theta_j$  for  $j \notin \llbracket L', L'+K \rrbracket$ . More precisely, we define a *reference local Gaussian measure*  $\sigma_{\theta} \sim \exp(-\beta N \mathcal{H}_{\theta}(\mathbf{x})) d\mathbf{x}$  on the set  $[\theta_{L'}, \theta_{L'+K+1}]$  via the Hamiltonian

$$(4.15) \quad \mathcal{H}_{\theta}(\mathbf{x}) = \sum_{i \in I'} \left[ \frac{1}{4} x_i^2 - \frac{1}{N} \sum_{j \notin I'} \log |x_i - \theta_j| \right] - \frac{1}{N} \sum_{\substack{i, j \in I' \\ i < j}} \log |x_j - x_i|,$$

where  $I' := \llbracket L' + 1, L' + K \rrbracket$ . Since  $L'$  will not play an active role, we will abuse the notation and set  $L' = L$ .

The measure  $\mu_{\mathbf{y}}$  lives on the interval  $[y_L, y_{L+K+1}]$  while the measure  $\sigma_{\theta}$  lives on the interval  $[\theta_L, \theta_{L+K+1}]$  and it is difficult to compare them. But after an appropriate translation and dilation, they will live on the same interval and from now on we assume that  $[y_L, y_{L+K+1}] = [\theta_L, \theta_{L+K+1}]$ . The parameter  $K = N^k$  has to be sufficiently small since  $\varrho_V$  and  $\varrho_W$  are not constant functions and we have to match these two densities quite precisely in the whole interval. There are some other subtle issues related to the rescaling, but we will neglect them here to concentrate on the main ideas. Our main result is the following theorem which is essentially a combination of Proposition 4.2 and Theorem 4.4 from [11].

**THEOREM 4.2.** *Let  $0 < \varphi \leq \frac{1}{38}$ . Fix  $K = N^k$ ,  $\delta = N^{-1+\varphi}$  and  $k = \frac{39}{2}\varphi$ . Then for  $\mathbf{y} \in \mathcal{G}_{\delta}$  we have*

$$(4.16) \quad \left| \mathbb{E}^{\mu_{\mathbf{y}}} \frac{1}{K} \sum_{i \in I} O(N(x_i - x_{i+1})) - \mathbb{E}^{\sigma_{\theta}} \frac{1}{K} \sum_{i \in I} O(N(x_i - x_{i+1})) \right| \rightarrow 0$$

as  $N \rightarrow \infty$  for any smooth and compactly supported test function  $O$ . A similar formula holds for more complicated observables of the form (3.9).

From the rigidity estimate, Theorem 4.1, it follows that  $\mathcal{G}_{\delta}$  has an overwhelming probability, so the expectation  $\mathbb{E}^{\mu_{\mathbf{y}}}$  can be changed to  $\mathbb{E}^{\mu}$  and similarly for the reference measure. Once (4.16) is proven for all observables of the form (3.9), we get that the locally averaged gap statistics for  $\mu$  coincide with those of the reference Gaussian case, hence they are universal. Since averaged gap statistics identifies locally averaged correlation functions, we obtain Theorem 1.4.

It remains to prove Theorem 4.2. The basic idea is to use the Dirichlet form inequality (3.33). Although (3.33) was stated for an infinite volume measure, it holds for any measure with repulsive logarithmic interactions in a finite volume and with the parameter  $\tau^{-1}$  being the lower bound on the Hessian of the Hamiltonian. In our setting, we denote by  $\tau_{\sigma}^{-1}$  the lower bound for  $\nabla^2 \mathcal{H}_{\theta}$ , and the Dirichlet form

inequality becomes

$$(4.17) \quad \left| [\mathbb{E}^{\mu_y} - \mathbb{E}^{\sigma_\theta}] \frac{1}{K} \sum_{i \in I} O(N(x_i - x_{i+1})) \right| \leq C \left( \frac{\tau_\sigma N^\varepsilon}{K} D(\mu_y | \sigma_\theta) \right)^{1/2} + C e^{-cN^\varepsilon} \sqrt{S(\mu_y | \sigma_\theta)},$$

where

$$(4.18) \quad D(\mu_y | \sigma_\theta) := \frac{1}{2N} \int \left| \nabla \sqrt{\frac{d\mu_y}{d\sigma_\theta}} \right|^2 d\sigma_\theta.$$

Thus our task is to prove that

$$(4.19) \quad \tau_\sigma N^\varepsilon \frac{D(\mu_y | \sigma_\theta)}{K} \rightarrow 0.$$

By definition,

$$\frac{\tau_\sigma}{K} D(\mu_y | \sigma_\theta) \leq \frac{\tau_\sigma N}{K} \int \sum_{L+1 \leq j \leq L+K} Z_j^2 d\mu_y,$$

where  $Z_j$  is defined as

$$(4.20) \quad Z_j := \frac{\beta}{2} V'(x_j) - \frac{\beta}{N} \sum_{\substack{k < L \\ k > L+K}} \frac{1}{x_j - y_k} - \frac{\beta}{2} W'(x_j) + \frac{\beta}{N} \sum_{\substack{k < L \\ k > L+K}} \frac{1}{x_j - \theta_k}.$$

Using the equilibrium relation (1.20) between the potentials  $V$ ,  $W$  and the densities  $\varrho_V$ ,  $\varrho_W$ , we have

$$Z_j = \beta \int_{\mathbb{R}} \frac{\varrho_V(y)}{x_j - y} dy - \frac{\beta}{N} \sum_{\substack{k < L \\ k > L+K}} \frac{1}{x_j - y_k} - \beta \int_{\mathbb{R}} \frac{\varrho_W(y)}{x_j - y} dy + \frac{\beta}{N} \sum_{\substack{k < L \\ k > L+K}} \frac{1}{x_j - \theta_k}.$$

Hence  $Z_j$  is the sum of the error terms,

$$(4.21) \quad A_j := \int_{y \notin [y_L, y_{L+K+1}]} \frac{\varrho_V(y)}{x_j - y} dy - \frac{1}{N} \sum_{\substack{k < L \\ k > L+K}} \frac{1}{x_j - y_k},$$

$$(4.22) \quad B_j := \int_{y_L}^{y_{L+K+1}} \frac{\varrho_V(y) - \varrho_W(y)}{x_j - y} dy,$$

and there is a term similar to  $A_j$  with  $y_j$  replaced by  $\theta_j$  and  $\varrho_V$  replaced by  $\varrho_W$ .

With our convention, the total numbers of particles in the interval  $[y_{L+K+1}, y_L]$  are equal and thus

$$\int_{y_L}^{y_{L+K+1}} \varrho_V(y) dy = \int_{y_L}^{y_{L+K+1}} \varrho_W(y) dy.$$

Since the densities  $\rho_V$  and  $\rho_W$  are  $C^1$  functions away from the endpoints  $A$  and  $B$  and  $y_{L+K+1} - y_L$  is small,  $|\rho_V - \rho_W|$  is small in the interval  $[y_{L+K+1}, y_L]$  and thus  $B_j$  is small. For estimating  $A_j$ , we can replace the integral

$$\int_{-\infty}^{y_L} \frac{\varrho_V(y)}{x_j - y} dy \quad \text{by} \quad \frac{1}{N} \sum_{k < L} \frac{1}{x_j - \gamma_k}$$

with negligible errors, at least for  $j$ 's away from the edges,  $j \in \llbracket L + N^\varepsilon, L + K - N^\varepsilon \rrbracket$ .

Thus

$$(4.23) \quad |A_j| \leq \frac{C}{N} \left| \sum_{\substack{k < L \\ k > L+K}} T_j^k \right|, \quad T_j^k := \frac{1}{x_j - y_k} - \frac{1}{x_j - \gamma_k},$$



and  $T_j^k$  can be estimated by the assumption  $|y_k - \gamma_k| \leq \delta$  from  $\mathbf{y} \in \mathcal{G}_\delta$ . The same argument works if  $j$  is close to the edge, but  $k$  is away from the edges, i.e.  $k \leq L - N^\varepsilon$  or  $k \geq L + K + N^\varepsilon$ . The edge terms,  $T_j^k$  for  $|j - k| \leq N^\varepsilon$ , are difficult to estimate due to the singularity in the denominator and the event that many  $y_k$ 's with  $k < L$  may pile up near  $y_L$ . To resolve this difficulty, we show that the averaged local statistics of the measure  $\mu_{\mathbf{y}}$  are insensitive to the change of the boundary conditions for  $\mathbf{y}$  near the edges. This can be achieved by the simple inequality

$$(4.24) \quad \left| \frac{1}{K} \sum_{i \in I} \int O(N(x_i - x_{i+1})) [d\mu_{\mathbf{y}'} - d\mu_{\mathbf{y}}] \right| \leq C \int |d\mu_{\mathbf{y}'} - d\mu_{\mathbf{y}}| \leq C \sqrt{S(\mu_{\mathbf{y}'} | \mu_{\mathbf{y}})}$$

for any two boundary conditions  $\mathbf{y}$  and  $\mathbf{y}'$ . Although we still have to estimate the entropy that includes a logarithmic singularity, this can be done much more easily since entropy is less sensitive to singularities than Dirichlet form. Therefore, we can replace the boundary condition  $y_k$  with  $y'_k = \theta_k$  for  $|j - k| \leq N^\varepsilon$  and then the most singular edge terms in (4.20) cancel out.

We note that we can perform this replacement only for a small number of index pairs  $(j, k)$ , since estimating the gap distribution by the total entropy, as noted in (3.36) in Section 3.1, is not as efficient as the estimate using the Dirichlet form per particle. Thus we can afford to use this argument only for the edge terms,  $|j - k| \leq N^\varepsilon$ . For all other index pairs  $(j, k)$  we still have to estimate  $T_j^k$  by exploiting that  $\mathbf{y}$  is a good configuration, i.e.  $y_k - \gamma_k$  is small.

Unfortunately, even with the optimal accuracy  $\delta \sim N^{-1+\varepsilon'}$  in (4.14) as an input, the relation (4.19) still cannot be satisfied for any choice of  $N^{c\varepsilon'} \leq K \leq N^{1-c\varepsilon'}$ . To understand this problem, we remark that while the edge terms become a smaller percentage of the total terms in (4.24) as  $K$  gets bigger, the relaxation time to equilibrium for  $\sigma_\theta$ , determined by the convexity of  $\mathcal{H}_\theta''$ , increases at the same time. At the end of our calculation, there is no good regime for the choice of  $K$ . Fortunately, this can be resolved by using the idea of the local relaxation measure as in (3.22), i.e., we add a quadratic term  $\frac{1}{2\tau}(x_j - \gamma_j)^2$  to the Hamiltonian of the measure  $\mu_{\mathbf{y}}$  and  $\frac{1}{2\tau_\sigma}(x_j - \theta_j)^2$  for the measure  $\sigma_\theta$ . With these ideas, we can complete the proof of Theorem 4.2.

## 5. Single gap universality

In this section we outline the proofs of Theorem 1.3 and 1.5 following closely [45]. Both proofs rely on the single gap universality for the locally conditioned measure  $\mu_{\mathbf{y}}$  introduced already in (4.12). This will be stated in Theorem 5.1, whose proof takes up most of this section. At the end, in Section 5.4 we complete the proofs of Theorem 1.3 and 1.5.

### 5.1. Statements on local equilibrium measures.

**5.1.1. Definition.** We work in the bulk spectrum and we consider the local equilibrium measure on  $\mathcal{K} := 2K + 1$  points which is the conditional measure after fixing all other points. To define it precisely, we fix two small positive numbers,  $\alpha, \delta > 0$  and choose two positive integer parameters  $L, K$  such that

$$(5.1) \quad L \in \llbracket \alpha N, (1 - \alpha)N \rrbracket, \quad N^\delta \leq K \leq N^{1/4}.$$

All results will hold for any sufficiently small  $\alpha, \delta$  and for any sufficiently large  $N \geq N_0$ , where the threshold  $N_0$  depends on  $\alpha, \delta$  and maybe on other parameters of the model.

Denote  $I = I_{L,K} := \llbracket L - K, L + K \rrbracket$  the set of  $\mathcal{K}$  consecutive indices in the bulk. As in Section 4.2, we will distinguish external and internal points by renaming them as

$$(5.2) \quad (\lambda_1, \lambda_2, \dots, \lambda_N) := (y_1, \dots, y_{L-K-1}, x_{L-K}, \dots, x_{L+K}, y_{L+K+1}, \dots, y_N) \in \Xi^{(N)},$$

the only difference is that here the internal particles are labelled symmetrically to  $L$ . This discrepancy is only notational, but we prefer to follow the notations of the original papers. In short we will write

$$\mathbf{x} = (x_{L-K}, \dots, x_{L+K}) \in \Xi^{(\mathcal{K})}, \text{ and } \mathbf{y} = (y_1, \dots, y_{L-K-1}, y_{L+K+1}, \dots, y_N) \in \Xi^{(N-\mathcal{K})}.$$

As in (4.12) we again define the local equilibrium measure on  $\mathbf{x}$  with boundary condition  $\mathbf{y}$  by

$$(5.3) \quad \mu_{\mathbf{y}}(d\mathbf{x}) := \mu_{\mathbf{y}}(\mathbf{x})d\mathbf{x}, \quad \mu_{\mathbf{y}}(\mathbf{x}) := \mu(\mathbf{y}, \mathbf{x}) \left[ \int \mu(\mathbf{y}, \mathbf{x})d\mathbf{x} \right]^{-1},$$

where  $\mu = \mu(\mathbf{y}, \mathbf{x})$  is the (global) equilibrium measure (1.10). For a fixed  $\mathbf{y}$ , this measure can also be written as a Gibbs measure,

$$(5.4) \quad \mu_{\mathbf{y}} = \mu_{\mathbf{y}, \beta, V} = Z_{\mathbf{y}}^{-1} e^{-N\beta \mathcal{H}_{\mathbf{y}}},$$

with Hamiltonian

$$(5.5) \quad \mathcal{H}_{\mathbf{y}}(\mathbf{x}) := \sum_{i \in I} \frac{1}{2} V_{\mathbf{y}}(x_i) - \frac{1}{N} \sum_{\substack{i, j \in I \\ i < j}} \log |x_j - x_i|, \quad V_{\mathbf{y}}(x) := V(x) - \frac{2}{N} \sum_{k \notin I} \log |x - y_k|.$$

Here  $V_{\mathbf{y}}(x)$  can be viewed as the external potential of a  $\beta$ -log-gas of the points  $\{x_i : i \in I\}$  in the configuration interval  $J = J_{\mathbf{y}} := (y_{L-K-1}, y_{L+K+1})$ .

**5.1.2. Universality of the local gap statistics for  $\mu_{\mathbf{y}}$ .** Our main technical result, Theorem 5.1 below, asserts that the local gap statistics is essentially independent of  $V$  and  $\mathbf{y}$  as long as the boundary conditions  $\mathbf{y}$  are regular. This property is expressed by defining the following set of “good” boundary conditions with some given positive parameters  $\xi, \alpha$  (the set  $\mathcal{G}$  in (4.14) played exactly the same role)

$$(5.6) \quad \begin{aligned} \mathcal{R} = \mathcal{R}_{L,K}(\xi, \alpha) := & \{ \mathbf{y} : |y_k - \gamma_k| \leq N^{-1} K^\xi, \quad k \in \llbracket \alpha N, (1 - \alpha)N \rrbracket \setminus I_{L,K} \} \\ & \cap \{ \mathbf{y} : |y_k - \gamma_k| \leq N^{-4/15} K^\xi, \quad k \in \llbracket N^{3/5} K^\xi, N - N^{3/5} K^\xi \rrbracket \} \\ & \cap \{ \mathbf{y} : |y_k - \gamma_k| \leq 1, \quad k \in \llbracket 1, N \rrbracket \setminus I_{L,K} \}. \end{aligned}$$

This definition is tailored to the rigidity bounds for the  $\beta$ -ensemble, see Theorem 4.1. Note that  $\mathcal{R}$  has a key parameter, the exponent  $\xi$ , which will be chosen as an arbitrary small positive number in the applications. We will not follow its dependence precisely and we will often neglect it from the notation, i.e. we will talk about “good” boundary conditions  $\mathbf{y} \in \mathcal{R}$ .

Good boundary conditions give rise to a regular potential  $V_{\mathbf{y}}$ . More precisely, if  $\mathbf{y} \in \mathcal{R}$ , then

$$(5.7) \quad |J_{\mathbf{y}}| = \frac{\mathcal{K}}{N\varrho(\bar{y})} + O\left(\frac{K^\xi}{N}\right),$$

$$(5.8) \quad V'_{\mathbf{y}}(x) = \varrho(\bar{y}) \log \frac{d_+(x)}{d_-(x)} + O\left(\frac{K^\xi}{Nd(x)}\right), \quad x \in J_{\mathbf{y}},$$

$$(5.9) \quad V''_{\mathbf{y}}(x) \geq \inf V'' + \frac{c}{d(x)}, \quad x \in J_{\mathbf{y}}.$$

Here

$$\bar{y} := \frac{1}{2}(y_{L-K-1} + y_{L+K+1})$$

denotes the midpoint of the configuration interval,  $d(x) := \min\{|x - y_{L-K-1}|, |x - y_{L+K+1}|\}$  is the distance of  $x$  to the boundary of the configuration interval  $J = (y_{L-K-1}, y_{L+K+1})$  and  $d_-(x)$  and  $d_+(x)$  are regularized versions of the distances of  $x$  to the closest and to the farthest endpoints of  $J$ , respectively. The key point is that the leading term of  $V'_{\mathbf{y}}(x)$  depends on the boundary conditions only through the density in the center,  $\varrho(\bar{y})$ . We also introduce

$$(5.10) \quad \alpha_j := \bar{y} + \frac{j-L}{\mathcal{K}+1}|J|, \quad j \in I_{L,K},$$

to denote the  $\mathcal{K}$  equidistant points within the interval  $J$ .

**THEOREM 5.1** (Gap universality for local measures). *Fix  $L, \tilde{L}$  and  $\mathcal{K} = 2K+1$  satisfying (5.1) with an exponent  $\delta > 0$ . Consider two boundary conditions  $\mathbf{y}, \tilde{\mathbf{y}}$  such that the configuration intervals coincide,*

$$(5.11) \quad J = (y_{L-K-1}, y_{L+K+1}) = (\tilde{y}_{\tilde{L}-K-1}, \tilde{y}_{\tilde{L}+K+1}).$$

*We consider the measures  $\mu = \mu_{\mathbf{y}, \beta, V}$  and  $\tilde{\mu} = \mu_{\tilde{\mathbf{y}}, \beta, \tilde{V}}$  defined as in (5.4), with possibly two different external potentials  $V$  and  $\tilde{V}$ . Let  $\xi > 0$  be a small constant. Assume that  $|J|$  satisfies*

$$(5.12) \quad |J| = \frac{\mathcal{K}}{N\varrho(\bar{y})} + O\left(\frac{K^\xi}{N}\right).$$

*Suppose that  $\mathbf{y}, \tilde{\mathbf{y}} \in \mathcal{R}$  and that*

$$(5.13) \quad \max_{j \in I_{L,K}} \left| \mathbb{E}^{\mu_{\mathbf{y}}} x_j - \alpha_j \right| + \max_{j \in I_{\tilde{L},K}} \left| \mathbb{E}^{\tilde{\mu}_{\tilde{\mathbf{y}}}} x_j - \alpha_j \right| \leq CN^{-1}K^\xi$$

*holds. Let the integer number  $p$  satisfy  $|p| \leq K - K^{1-\xi^*}$  for some small  $\xi^* > 0$ . Then there exists  $\xi_0 > 0$ , depending on  $\delta$ , such that if  $\xi, \xi^* \leq \xi_0$  then for any  $n$  fixed and any bounded smooth observable  $O : \mathbb{R}^n \rightarrow \mathbb{R}$  with compact support we have*

$$(5.14) \quad \left| \mathbb{E}^{\mu_{\mathbf{y}}} O(N(x_{L+p} - x_{L+p+1}), \dots, N(x_{L+p} - x_{L+p+n})) \right. \\ \left. - \mathbb{E}^{\tilde{\mu}_{\tilde{\mathbf{y}}}} O(N(x_{\tilde{L}+p} - x_{\tilde{L}+p+1}), \dots, N(x_{\tilde{L}+p} - x_{\tilde{L}+p+n})) \right| \leq CK^{-\varepsilon}$$

*for some  $\varepsilon > 0$  depending on  $\delta, \alpha$  and for some  $C$  depending on  $O$ . This holds for any  $N \geq N_0$  sufficiently large, where  $N_0$  depends on the parameters  $\xi, \xi^*, \alpha$ , and  $C$  in (5.13).*

5.1.3. *Rigidity and level repulsion of  $\mu_{\mathbf{y}}$ .* In the following two theorems we establish rigidity and level repulsion estimates for the local log-gas  $\mu_{\mathbf{y}}$  with good boundary conditions  $\mathbf{y}$ . While both rigidity and level repulsion are basic questions for log gases and are interesting in themselves, our main motivation to prove these theorems is to use them in the proof of Theorem 5.1.

We remark that an almost optimal rigidity estimate in the bulk was given in Theorem 4.1 and some level repulsion bound was given in (4.11) of [11], these results hold with respect to the global measure  $\mu$ . For the proof of Theorem 5.1 we need their local versions with respect to  $\mu_{\mathbf{y}}$ , at least for most  $\mathbf{y}$ . Naively, this looks as a simple conditioning argument, but there is a subtle point. From the estimates w.r.t.  $\mu$ , one can conclude that  $\mu_{\mathbf{y}}$  has a good rigidity bound for a set of boundary conditions with high probability w.r.t. the global measure  $\mu$ . This will be sufficient for the proof of Theorem 1.5, but not for Theorem 1.3. In the proof for the gap universality of Wigner matrices we will need a rigidity estimate for  $\mu_{\mathbf{y}}$  for a set of  $\mathbf{y}$ 's with high probability with respect to by the time evolved measure  $f_t\mu$  which may be asymptotically singular to  $\mu$  for large  $N$ . The following result asserts that a rigidity estimate holds for  $\mu_{\mathbf{y}}$  provided that  $\mathbf{y}$  itself satisfies a rigidity bound and an extra condition, (5.15), holds. This condition will have to be verified with different methods in the Wigner case.

**THEOREM 5.2** (Rigidity estimate for local measures). *For  $\mathbf{y} \in \mathcal{R}$  consider the local equilibrium measure  $\mu_{\mathbf{y}}$  defined in (5.4) and assume that*

$$(5.15) \quad \left| \mathbb{E}^{\mu_{\mathbf{y}}} x_k - \alpha_k \right| \leq CN^{-1}K^\xi, \quad k \in I = I_{L,K},$$

*is satisfied. Then there are positive constants  $C, c$ , depending on  $\xi$ , such that for any  $k \in I$  and  $u > 0$ ,*

$$(5.16) \quad \mathbb{P}^{\mu_{\mathbf{y}}} \left( N |x_k - \alpha_k| \geq uK^\xi \right) \leq Ce^{-cu^2}.$$

The proof of this result is similar to that of the concentration estimate (4.10) in Theorem 4.1. To estimate  $x_k - \mathbb{E}^{\mu_{\mathbf{y}}} x_k$ , we again use a multiscale argument of local averages for which stronger convexity bounds are available. The analogue of the accuracy estimate controlling  $\mathbb{E}^{\mu_{\mathbf{y}}} x_k - \gamma_k$  in Theorem 4.1 is replaced by the assumption (5.15). Notice that, unlike for the global measure  $\mu$ , a direct accuracy control via the loop equation is not available for  $\mu_{\mathbf{y}}$  since the potential  $V_{\mathbf{y}}$  is not analytic.

Now we state the level repulsion estimates.

**THEOREM 5.3** (Level repulsion estimate for local measures). *For  $\mathbf{y} \in \mathcal{R}$  we have the following estimates:*

*i) [Weak form of level repulsion] For any  $s > 0$  we have*

$$(5.17) \quad \mathbb{P}^{\mu_{\mathbf{y}}} [N(x_{i+1} - x_i) \leq s] \leq C (Ns)^{\beta+1}, \quad i \in \llbracket L - K - 1, L + K \rrbracket.$$

*ii) [Strong form of level repulsion] Suppose that there exist positive constants  $C, c$  such that the following rigidity estimate holds for any  $k \in I$ :*

$$(5.18) \quad \mathbb{P}^{\mu_{\mathbf{y}}} \left( N |x_k - \alpha_k| \geq CK^{\xi^2} \right) \leq C \exp(-K^c).$$

*Then there exists small a constant  $\theta$ , depending on  $C, c$  in (5.18), such that for any  $s \geq \exp(-K^\theta)$ , we have*

$$(5.19) \quad \mathbb{P}^{\mu_{\mathbf{y}}} [N(x_{i+1} - x_i) \leq s] \leq C (K^\xi s \log N)^{\beta+1}, \quad i \in \llbracket L - K - 1, L + K \rrbracket.$$

The level repulsion bounds will mostly be used in the following estimate which trivially follows from Theorem 5.3

COROLLARY 5.4. *Let  $\mathbf{y} \in \mathcal{R}$ , then for any  $p < \beta + 1$  we have*

$$(5.20) \quad \mathbb{E}^{\mu_{\mathbf{y}}} \frac{1}{[N|x_i - x_{i+1}|]^p} \leq C_p K^{C_3 \xi}, \quad i \in \llbracket L - K - 1, L + K \rrbracket.$$

□

5.1.4. *Sketch of the proof of the level repulsion.* The proof of part (ii) of Theorem 5.3 goes in three steps. For simplicity, we consider (5.19) only for the first gap, i.e.  $i = L - K - 1$  and we also assume that  $\bar{y} = 0$  by a simple shift.

*Step 1.* In this step we prove

$$(5.21) \quad \mathbb{P}^{\mu_{\mathbf{y}}} (x_{L-K} - y_{L-K-1} \leq s/N) \leq CKs \log N,$$

which is essentially (5.19) but with factor  $K$  instead of  $K^\xi$  and with the exponent  $\beta + 1$  replaced with one. The proof of (5.21) is dilation argument. For a nonnegative parameter  $\varphi$ , we define

$$(5.22) \quad \begin{aligned} Z_\varphi &:= \int \dots \int_{-a+a\varphi}^{a-a\varphi} d\mathbf{x} \prod_{\substack{i,j \in I \\ i < j}} (x_i - x_j)^\beta e^{-N \frac{\beta}{2} \sum_j V_{\mathbf{y}}(x_j)} \\ &= (1 - \varphi)^{K+\beta K(K-1)/2} \int \dots \int_{-a}^a d\mathbf{w} \prod_{i < j} (w_i - w_j)^\beta e^{-N \frac{\beta}{2} \sum_j V_{\mathbf{y}}((1-\varphi)w_j)}, \end{aligned}$$

where we set

$$a := -y_{L-K-1}, \quad w_j := (1 - \varphi)^{-1} x_{L+j}, \quad d\mathbf{x} = \prod_{|j| \leq K} dx_{L+j} \quad d\mathbf{w} = \prod_{|j| \leq K} dw_j.$$

Clearly  $Z_{\varphi=0}$  is the normalization constant of the measure  $\mu_{\mathbf{y}}$  and we have

$$(5.23) \quad \mathbb{P}^{\mu_{\mathbf{y}}} (x_{L-K} - (-a) \geq a\varphi) \geq \frac{Z_\varphi}{Z_0}.$$

The multiple integral on the r.h.s of (5.22) is almost the same as  $Z_0$ , except that the argument of  $V_{\mathbf{y}}$  is rescaled by  $1 - \varphi$ . This effect can be estimated from the explicit formula (5.5) for  $V_{\mathbf{y}}$ . The external potential  $V$  in (5.5) is unproblematic since it is smooth. Due to  $\mathbf{y} \in \mathcal{R}$ , the points  $y_j$  are regularly spaced on scales at least  $K^\xi/N$ , thus the sum of the interaction terms  $\log|x - y_k|$  for  $k$ 's away from the edges of  $I^c$ , i.e.  $k \leq L - 2K$  or  $k \geq L + 2K$ , is a regular function of  $x$  and the effect of dilation can be well approximated by Taylor expansion. For nearby  $k$ 's right below the lower edge, i.e.  $L - 2K \leq k \leq L - K$ , we use the trivial bound  $(1 - \varphi)x - y_k \geq (1 - \varphi)(x - y_k)$ . From these estimates it follows that

$$(5.24) \quad \frac{Z_\varphi}{Z_0} \geq 1 - CK^2 \varphi \log N.$$

Since  $a \sim K/N$ , together with (5.23) it implies (5.21).

*Step 2.* Now we consider an auxiliary measure which are slightly modified version of the local equilibrium measures:

$$(5.25) \quad \mu^{(0)} := Z^{(0)}(x_{L-K} - y_{L-K-1})^{-\beta} \mu_{\mathbf{y}};$$

where  $Z^{(0)}$  are chosen for normalization. In other words, we drop the term  $(x_{L-K} - y_{L-K-1})^\beta$  from the measure  $\mu_{\mathbf{y}}$ . Setting  $X := x_{L-K} - y_{L-K-1}$  for brevity, we have

$$(5.26) \quad \mathbb{P}^{\mu_{\mathbf{y}}} [X \leq s/N] = \frac{\mathbb{E}^{\mu^{(0)}} [1(X \leq s/N) X^\beta]}{\mathbb{E}^{\mu^{(0)}} [X^\beta]}.$$

The estimate (5.21) also holds for  $\mu^{(0)}$  and thus

$$\mathbb{E}^{\mu^{(0)}} [1(X \leq s/N) X^\beta] \leq C(s/N)^\beta K s \log N$$

and with the choice  $s = cK^{-1}(\log N)^{-1}$  in (5.21) we also have

$$\mathbb{P}^{\mu^{(0)}} \left( X \geq \frac{c}{NK \log N} \right) \geq 1/2$$

with some positive constant  $c$ . This implies that

$$\mathbb{E}^{\mu^{(0)}} [X^\beta] \geq \frac{1}{2} \left( \frac{c}{NK \log N} \right)^\beta.$$

Combining with (5.26), we have thus proved that

$$(5.27) \quad \mathbb{P}^{\mu_{\mathbf{y}}} [X \leq s/N] \leq C (K s \log N)^{\beta+1},$$

i.e. we obtained (5.21) but with an exponent  $\beta + 1$  in the r.h.s.

*Step 3.* We now improve the constant  $K$  to  $K^\xi$  in the r.h.s of (5.21). The factor  $K$  originated from the number of particles in  $\mu_{\mathbf{y}}$ . We can further condition the measure  $\mu_{\mathbf{y}}$  on the points

$$z_j := x_j \quad j \geq L - K + K^\xi,$$

and we let  $\mu_{\mathbf{y}, \mathbf{z}}$  denote the conditional measure on the remaining  $x$  variables  $\{x_j : L - K \leq j \leq L - K + K^\xi\}$ . From the rigidity estimate (5.18) we have  $(\mathbf{y}, \mathbf{z}) \in \mathcal{R}$  with a very high probability w.r.t.  $\mu_{\mathbf{y}}$ . We will now apply (5.27) to the measure  $\mu_{\mathbf{y}, \mathbf{z}}$  to obtain

$$(5.28) \quad \mathbb{P}^{\mu_{\mathbf{y}, \mathbf{z}}} [X \leq s/N] \leq C (K^\xi s \log N)^{\beta+1}.$$

This holds for all  $z$  with a high  $\mu_{\mathbf{y}}$ -probability. The subexponential lower bound on  $s$ , assumed in part ii) of Theorem 5.3, allows us to include the probability of the complement of  $\mathcal{R}$  in the estimate, we thus have proved (5.19).

The proofs of the weaker bound (5.17) for any  $s > 0$  use similar arguments that have led to (5.21), but without assuming  $\mathbf{y} \in \mathcal{R}$  which yields that one factor of  $K$  has to be replaced with  $N$  in (5.24). The assumption that the boundary conditions are good needs to be dropped since in Step 3 of the above argument, (5.21) is also used after additional conditioning on  $\mathbf{z}$ , distributed according to  $\mu_{\mathbf{y}}$ , and without (5.18) there is no rigidity result available for  $\mu_{\mathbf{y}}$ .

**5.2. Proof of Theorem 5.1.** In this section, we start to compare gap distributions of two local log-gases on the same configuration interval but with different external potential and boundary conditions. For simplicity, we consider only an observable of a single gap; a few consecutive gaps can be handled similarly. From now on, we use microscopic coordinates, i.e. we replace  $x_j$  with  $x_j/N$ , and we also relabel the indices so that the coordinates of  $x_j$  are  $j \in I = \{-K, \dots, 0, 1, \dots, K\}$ . This will have the advantage that  $K$  remains the only large parameter;  $N$  disappears.

The local equilibrium measures and their Hamiltonians will be denoted by the same symbols,  $\mu_{\mathbf{y}}$  and  $\mathcal{H}_{\mathbf{y}}$ , as before, but with a slight abuse of notations we redefine them now to the microscopic scaling, i.e.

(5.29)

$$\mathcal{H}_{\mathbf{y}}(\mathbf{x}) := \sum_{i \in I} \frac{1}{2} V_{\mathbf{y}}(x_i) - \sum_{\substack{i, j \in I \\ i < j}} \log |x_j - x_i|, \quad V_{\mathbf{y}}(x) := NV(x/N) - 2 \sum_{j \notin I} \log |x - y_j|,$$

The other Hamiltonian  $\tilde{H}_{\tilde{\mathbf{y}}}$  is defined in a similar way with  $V$  in (5.29) replaced with another external potential  $\tilde{V}$ . We also rewrite (5.13) in the microscopic coordinate as

$$(5.30) \quad |\mathbb{E}^{\mu_{\mathbf{y}}} x_j - \alpha_j| + |\mathbb{E}^{\tilde{\mu}_{\tilde{\mathbf{y}}}} x_j - \alpha_j| \leq CK^{\xi},$$

where  $\alpha_j := \frac{j}{\kappa+1}|J|$  is the rescaled version of the definition given in (5.10), but we keep the same notation. The concept of “good” set  $\mathcal{R}$  is also rescaled accordingly.

Suppose that  $\mathbf{y}, \tilde{\mathbf{y}} \in \mathcal{R}$  and define the interpolating measures

$$(5.31) \quad \omega_{\mathbf{y}, \tilde{\mathbf{y}}}^r = Z_r e^{-\beta r(\tilde{V}_{\tilde{\mathbf{y}}}(\mathbf{x}) - V_{\mathbf{y}}(\mathbf{x}))} \mu_{\mathbf{y}}, \quad r \in [0, 1],$$

so that  $\omega_{\mathbf{y}, \tilde{\mathbf{y}}}^1 = \tilde{\mu}_{\tilde{\mathbf{y}}}$  and  $\omega_{\mathbf{y}, \tilde{\mathbf{y}}}^0 = \mu_{\mathbf{y}}$  ( $Z_r$  is a normalization constant). This is again a local log-gas with Hamiltonian

(5.32)

$$\mathcal{H}_{\mathbf{y}, \tilde{\mathbf{y}}}^r(\mathbf{x}) = \frac{1}{2} \sum_{i \in I} V_{\mathbf{y}, \tilde{\mathbf{y}}}^r(x_i) - \frac{1}{N} \sum_{i < j} \log |x_i - x_j|, \quad V_{\mathbf{y}, \tilde{\mathbf{y}}}^r(x) := (1-r)V_{\mathbf{y}}(x) + r\tilde{V}_{\tilde{\mathbf{y}}}(x).$$

For any fixed  $r$ , the measure  $\omega_{\mathbf{y}, \tilde{\mathbf{y}}}^r$  inherits all relevant properties of  $\mu_{\mathbf{y}}$ . In particular the rigidity bound in the form

$$(5.33) \quad \mathbb{P}^{\omega}(|x_i - \alpha_i| \geq CK^{C\xi}) \leq Ce^{-K^{\theta}}, \quad i \in I,$$

the level repulsion bounds (5.17)–(5.19) and their consequence in (5.20) hold w.r.t. the measure  $\omega = \omega_{\mathbf{y}, \tilde{\mathbf{y}}}^r$  as well (in the new microscopic coordinates there are no  $N$  factors in the left hand sides of these inequalities). The proofs are basically parallel with the arguments for  $\mu_{\mathbf{y}}$ ; the only nontrivial step is to show that (5.30) implies the analogous bound

$$|\mathbb{E}^{\omega} x_k - \alpha_k| \leq CK^{\xi}$$

w.r.t.  $\omega = \omega_{\mathbf{y}, \tilde{\mathbf{y}}}^r$  as well. Although  $\omega$  appears to be some easy combination of  $\mu_{\mathbf{y}}$  and  $\tilde{\mu}_{\tilde{\mathbf{y}}}$ , this conclusion is nontrivial. It requires comparing  $\omega$  and  $\mu_{\mathbf{y}}$  via the entropy inequality, which involves controlling the exponential moment of  $|x_k - \alpha_k|$  w.r.t.  $\mu_{\mathbf{y}}$ . At this point the Gaussian tail proven in (5.16) is necessary.

The right hand side of (5.14) with  $n = 1$ , in the rescaled coordinates and with  $L = \tilde{L} = 0$ , is estimated by

$$(5.34) \quad \left| [\mathbb{E}^{\mu_{\mathbf{y}}} - \mathbb{E}^{\tilde{\mu}_{\tilde{\mathbf{y}}}}] O(x_p - x_{p+1}) \right| \leq \int_0^1 dr \frac{d}{dr} \mathbb{E}^{\omega_{\mathbf{y}, \tilde{\mathbf{y}}}^r} O(x_p - x_{p+1}).$$

For any bounded smooth function  $O$  with compact support

$$(5.35) \quad \frac{d}{dr} \mathbb{E}^{\omega_{\mathbf{y}, \tilde{\mathbf{y}}}^r} O(x_p - x_{p+1}) = \beta \langle h_0; O(x_p - x_{p+1}) \rangle_{\omega_{\mathbf{y}, \tilde{\mathbf{y}}}^r},$$

where

$$(5.36) \quad h_0 = h_0(\mathbf{x}) = \sum_{i \in I} (V_{\mathbf{y}}(x_i) - \tilde{V}_{\tilde{\mathbf{y}}}(x_i))$$

and  $\langle f; g \rangle_\omega := \mathbb{E}^\omega fg - (\mathbb{E}^\omega f)(\mathbb{E}^\omega g)$  denotes the covariance. Thus Theorem 5.1 follows immediately from the following estimate on the gap covariance function.  $\square$

**THEOREM 5.5.** *Consider two smooth potentials  $V, \tilde{V}$  and two good boundary conditions,  $\mathbf{y}, \tilde{\mathbf{y}} \in \mathcal{R}$ , such that the configuration intervals coincide,  $J_{\mathbf{y}} = J_{\tilde{\mathbf{y}}}$ . For any  $r \in [0, 1]$  let  $\omega = \omega_{\mathbf{y}, \tilde{\mathbf{y}}}^r$  be the interpolating measure defined in (5.31). Assume that (5.30) holds for both boundary conditions  $\mathbf{y}, \tilde{\mathbf{y}}$ . Fix  $\xi^* > 0$ . Then there exist  $\varepsilon > 0$  and  $C > 0$ , depending on  $\xi^*$ , such that for any sufficiently small  $\xi$ , for  $|p| \leq K^{1-\xi^*}$  we have*

$$(5.37) \quad |\langle h_0(\mathbf{x}); O(x_p - x_{p+1}) \rangle_\omega| \leq K^{C\xi} K^{-\varepsilon}$$

for any smooth function  $O : \mathbb{R} \rightarrow \mathbb{R}$  with compact support provided that  $K$  is large enough.

Theorem 5.5 is our key technical result. The main difficulty behind it is due to the fact that the covariance function of two points,  $\langle x_i; x_j \rangle_\omega$ , decays only logarithmically. In fact, for the GUE, Gustavsson proved that (Theorem 1.3 in [55])

$$(5.38) \quad \langle x_i; x_j \rangle_{GUE} \sim \log \frac{N}{|i - j| + 1},$$

and a similar formula is expected for  $\omega$ . Although  $h_0(\mathbf{x})$  depends strongly only on points near the boundary and  $x_p$  is away from the boundary, it is still very difficult to prove Theorem 5.5 based on this slow logarithmic decay. However, the covariance function of the type

$$(5.39) \quad \langle g_1(x_i); g_2(x_j - x_{j+1}) \rangle_\omega$$

decays much faster in  $|i - j|$ . Since the second factor  $g_2(x_j - x_{j+1})$  depends only on the difference of two neighboring points, it is expected that the decay is the (discrete) derivative in  $j$  of the covariance (5.38), i.e. it is  $|i - j|^{-1}$ . The actual result (5.37) is much weaker, but it still provides a power-law decay in  $K$  instead of a logarithmic decay. Covariances of the form  $\langle g_1(x_i - x_{i+1}); g_2(x_j - x_{j+1}) \rangle_\omega$  are expected to decay even faster but we have not pursued this direction further.

We point out that the fact that observables of differences of particles behave much nicer was a basic observation in DBM analysis (Theorem 3.4), see the explanation around (3.36).

**5.3. Decay of correlation functions: Proof of Theorem 5.5.** We will express the difference of gap distributions between two measures in terms of random walks in time dependent random environments. The decay of correlation functions will be translated into a partial regularity property of the corresponding parabolic equation. This partial regularity is a discrete version of the De Giorgi-Nash-Moser theory but with a long range elliptic part.

**5.3.1. Random Walk Representation.** In this section we derive a random walk representation for the gap correlation function on the left hand side of (5.37). We will apply it for the interpolating measure  $\omega = \omega_{\mathbf{y}, \tilde{\mathbf{y}}}^r$  (5.31) and for the function  $h_0$  given in (5.36), but the representation formula (Proposition 5.6 below) is valid for any  $\omega$  and  $h_0$ .

Let  $\mathcal{L}^\omega$  be the reversible generator given by the Dirichlet form

$$(5.40) \quad D^\omega(f) = - \int f \mathcal{L}^\omega f d\omega = \sum_{|j| \leq K} \int (\partial_j f)^2 d\omega.$$



This process can also be characterized by the following SDE

$$(5.41) \quad dx_i = dB_i + \beta \left[ -\frac{1}{2} (V_{\mathbf{y}, \tilde{\mathbf{y}}}^r)'(x_i) + \frac{1}{2} \sum_{j \neq i} \frac{1}{(x_i - x_j)} \right] dt,$$

where  $\{B_i : |i| \leq K\}$  is a family of independent standard real Brownian motions. Let  $\mathbb{E}_{\mathbf{x}}$  denote the expectation for this process with initial point  $\mathbf{x}(0) = \mathbf{x}$ . The expectation with respect to the process starting from equilibrium is  $\mathbb{E}^\omega[\cdot] = \int \mathbb{E}_{\mathbf{x}}[\cdot] \omega(d\mathbf{x})$ . With a slight abuse of notations, when we talk about the process, we will use  $\mathbb{P}^\omega$  and  $\mathbb{E}^\omega$  also to denote the probability and expectation w.r.t. this dynamics with initial data distributed w.r.t.  $\omega$ , i.e., in equilibrium.

Suppose  $h(t) = h(t, \mathbf{x})$  is the solution of the equation  $\partial_t h = \mathcal{L}^\omega h$  with an initial condition  $h_0$ . Introduce the notation

$$(5.42) \quad \mathbf{v}(t, \mathbf{x}) = \nabla_{\mathbf{x}} h(t, \mathbf{x}), \quad \text{i.e.} \quad v_j(t, \mathbf{x}) := \partial_{x_j} h(t, \mathbf{x}).$$

By integrating the time derivative of  $\langle h(t, \mathbf{x}); O(x_p - x_{p+1}) \rangle_\omega$  and using the equation  $h(t, \mathbf{x})$  satisfies, we have

$$(5.43) \quad \langle h_0(\mathbf{x}); O(x_p - x_{p+1}) \rangle_\omega = \int_0^\infty d\sigma \int O'(x_p - x_{p+1}) [v_p(\sigma, \mathbf{x}) - v_{p+1}(\sigma, \mathbf{x})] d\omega(\mathbf{x}).$$

For any fixed  $\sigma$ , the inner integral on the right hand side can be expressed by a random walk representation. Fix a path  $\{\mathbf{x}(s) : s \in [0, \sigma]\}$ . Define the following operators on  $\mathbb{R}^K$

$$(5.44) \quad \mathcal{A}(s) := \mathcal{B}(s) + \mathcal{W}(s)$$

$$(5.45) \quad [\mathcal{B}(s)\mathbf{v}]_j = - \sum_k B_{jk}(\mathbf{x}(\sigma - s))(v_k - v_j), \quad B_{jk}(\mathbf{x}) = \frac{1}{(x_j - x_k)^2} \geq 0$$

$$[\mathcal{W}(s)\mathbf{v}]_j = \mathcal{W}_j(s)v_j, \quad \mathcal{W}_j(s) := [V_{\mathbf{y}, \tilde{\mathbf{y}}}^r]''(x_j(\sigma - s)).$$

Clearly  $\mathcal{B}(s)$  is diffusion operator with random rates and  $\mathcal{W}(s)$  is a potential representing a random environment. These operators depend on the whole path  $\mathbf{x}(s)$ , but we omit this fact from the notation.

With these notations we have the following representation:

**PROPOSITION 5.6.** *For any smooth function  $h_0 : J^K \rightarrow \mathbb{R}$ , for any  $p \in I$ ,  $-K \leq p \leq K - 1$ , we have*

$$(5.46) \quad \langle h_0; O(x_p - x_{p+1}) \rangle_\omega = \int_0^\infty d\sigma \int O'(x_p - x_{p+1}) \mathbb{E}_{\mathbf{x}}[w_p(\sigma, \mathbf{x}(\cdot); \sigma) - w_{p+1}(\sigma, \mathbf{x}(\cdot); \sigma)] \omega(d\mathbf{x}).$$

Here, for any  $\sigma > 0$  and for any fixed path  $\{\mathbf{x}(s) : s \in [0, \sigma]\}$  we let  $\mathbf{w}$  denote the solution of the evolution equation

$$(5.47) \quad \partial_s \mathbf{w}(s; \mathbf{x}(\cdot), \sigma) = -\mathcal{A}(s)\mathbf{w}(s; \mathbf{x}(\cdot), \sigma),$$

with initial data  $\mathbf{w}(0; \mathbf{x}(\cdot), \sigma) := \nabla h_0(\mathbf{x}(\sigma))$ .

This representation in a slightly different setting already appeared in Proposition 2.2 of [17] (see also Proposition 3.1 in [54]), which was a probabilistic formulation of the idea of Helffer and Sjöstrand [56] and Naddaf and Spencer [70]. The proof

relies on taking the gradient of the equation  $\partial_t h = \mathcal{L}^\omega h$ . A direct computation of the commutator  $[\nabla, \mathcal{L}^\omega]$  yields that

$$(5.48) \quad \partial_t \mathbf{v}(t, \mathbf{x}) = \mathcal{L}^\omega \mathbf{v}(t, \mathbf{x}) - \tilde{\mathcal{A}}(\mathbf{x}) \mathbf{v}(t, \mathbf{x}),$$

with initial condition  $\mathbf{v}_0(\mathbf{x}) = \mathbf{v}(0, \mathbf{x}) = \nabla h_0(\mathbf{x})$ . Here  $\tilde{\mathcal{A}}(\mathbf{x}) = \mathcal{B}(\mathbf{x}) + \mathcal{W}(\mathbf{x})$ , where  $\mathcal{B}(\mathbf{x})$  is the operator given by the matrix  $B_{jk}$  in (5.45) and  $\mathcal{W}(\mathbf{x})$  is the diagonal multiplication operator by  $[V_{\mathbf{y}, \tilde{\mathbf{y}}}^r]''(x_j)$ . Since  $\mathcal{L}^\omega$  generates the process  $\mathbf{x}(t)$ , we can represent the solution to (5.48) by the Feynman-Kac formula which can be written in the form (5.46).

When applying this proposition to our case, we will choose the initial condition  $h_0$  be given by (5.36). The initial condition for the random walk (5.47) is given by  $\nabla h_0$ . Notice that the leading term in  $\partial_j h_0(\mathbf{x}) = V_{\mathbf{y}}'(x_j) - [\tilde{V}_{\tilde{\mathbf{y}}}]'(x_j)$  cancel; this is because the leading term in (5.8) depends only on the density  $\varrho(\bar{y})$  which is matched for  $\mathbf{y}$  and  $\tilde{\mathbf{y}}$  by  $J_{\mathbf{y}} = J_{\tilde{\mathbf{y}}}$ , see (5.7). We thus have

$$(5.49) \quad |\partial_j h_0(x)| \leq \frac{CK^\xi}{d(x_j)},$$

i.e. initially  $\mathbf{w}$  is small away from the boundary and for the small  $\sigma$  regime the inner integral in the r.h.s. of (5.46) is small. After very long time  $\mathbf{w}$  becomes constant, but then the right hand side of (5.46) is zero. The analysis of (5.46) requires to monitor what happens to  $\mathbf{w}$  for coordinates  $p$  away from the boundary at intermediate times.

In the following sections we make a few preparations that exclude irrelevant regimes. First, it is easy to see that the regular spacing of  $\mathbf{y}, \tilde{\mathbf{y}} \in \mathcal{R}$  implies that  $W_j(s) \geq cK^{-1}$ , which means that the  $L^1$ -norm of the solution to (5.47) decays at a rate of order  $K$ . Thus the integral in (5.46) can be truncated at  $\sigma \leq CK \log K$ .

**5.3.2. Preparation for the De Giorgi-Nash-Moser bound: Restriction to the good paths.** The representation (5.46) expresses the covariance function in terms of the discrete spatial derivative of the solution to (5.47). To estimate  $w_p(\sigma, \mathbf{x}(\cdot); \sigma) - w_{p+1}(\sigma, \mathbf{x}(\cdot); \sigma)$  in (5.46), we will now study the Hölder continuity of the solution  $\mathbf{w}(s, \mathbf{x}(\cdot); \sigma)$  to (5.47) at time  $s = \sigma$  and at the spatial point  $p$ . We will do it for each fixed path  $\mathbf{x}(\cdot)$ , with the exception of a set of “bad” paths that will have a small probability.

Notice that if all points  $x_i$  were approximately regularly spaced in the interval  $J$ , then the operator  $\mathcal{B}$  had a kernel  $\mathcal{B}_{ij} \sim (i-j)^{-2}$ , i.e. it were essentially a discrete version of the operator  $|p| = \sqrt{-\Delta}$  (in one dimension). Hölder continuity will thus be the consequence of the De Giorgi-Nash-Moser bound for the parabolic equation (5.47). However, we need to control the coefficients in this equation, which depend on the random walk  $\mathbf{x}(\cdot)$ .

For the De Giorgi-Nash-Moser theory we need both upper and lower bounds on the time dependent kernel  $\mathcal{B}_{ij}(s)$ . The rigidity bound (5.33) guarantees a lower bound on  $\mathcal{B}_{ij}$ , up to a factor  $K^{-C\xi}$ . Since the subexponential probabilistic estimate in (5.33) is very strong, one can easily guarantee a very similar estimate uniformly in time, i.e.

$$(5.50) \quad \mathbb{P}^\omega \left\{ \mathbf{x}(s) : \sup_{0 \leq s \leq CK \log K} \sup_{|j| \leq K} |x_j(s) - \alpha_j| \leq K^{C\xi} \right\} \geq 1 - e^{-K^\theta}$$

(maybe after reducing  $\theta$  from (5.33)). This follows from the fact that  $\omega$  is invariant under the dynamics and  $\mathbf{x}(t)$  has some stochastic continuity.

The level repulsion estimate implies certain upper bounds on  $\mathcal{B}_{ij}$ , but these estimates not particularly strong. Even in the  $\beta > 1$  case, the bound (5.20) implies only that

$$\mathbb{E}^\omega \mathcal{B}_{i,i+1}(s) = \mathbb{E}^\omega \frac{1}{(x_{i+1} - x_i)^2} \leq K^{C\xi}$$

is finite. In the  $\beta = 1$  borderline case even the expectation of  $\mathcal{B}_{i,i+1}$  is infinite. Such a weak control does not allow us to guarantee an effective simultaneous bound on  $\mathcal{B}_{i,i+1}$  for all  $i$  and for all time. Instead of supremum bounds, we control these coefficients only in an average sense and we can show that for any fixed index  $Z \in I$ , time  $s$  and parameter  $M$ , we have

$$(5.51) \quad \mathbb{P}^\omega \left\{ \mathbf{x}(s) : \frac{1}{1+s} \int_0^s da \frac{1}{M} \sum_{|i-Z| \leq M} \sum_j B_{ij}(\mathbf{x}(\sigma-a)) \leq K^\rho \right\} \geq 1 - K^{C\xi-\rho}.$$

Here  $\rho$  will be chosen as large constant times  $\xi$ . The summation over  $j$  is harmless since for  $|i-j| \geq K^\xi$  the rigidity estimate can be used to bound  $\mathcal{B}_{ij}$ . By a dyadic choice of the parameters  $s, M$ , it is easy to upgrade (5.51) to hold for any  $M \leq K$  and  $s \leq CK \log K$ . But it is essential that a reference point  $Z$  be fixed, one cannot guarantee that none of the gaps closes.

The expectation over the paths,  $\int \mathbb{E}_\mathbf{x}[\cdot] \omega(d\mathbf{x})$ , in (5.46) will be restricted to the sets given in (5.50) and (5.51). Due to the strong subexponential bound, the restriction to the set in (5.50) is unproblematic. However, the estimate (5.51) is quite weak; the probability of the "bad" paths is bounded only by a small negative power of  $K$ . This is not sufficient to compensate the time integration in (5.46) even after the upper cutoff  $\sigma \leq CK \log K$ . We will need to use that the heat kernel of the equation (5.47) has an  $L^1 \rightarrow L^\infty$  decay of order  $1/s$  after time  $s$ . Thus the solution  $w_p(\sigma, \mathbf{x}(\cdot); \sigma)$  decays as  $1/\sigma$  which renders the  $d\sigma$  integration in (5.46) harmless.

For completeness, we state the  $L^p \rightarrow L^q$  heat kernel decay estimate in a general form. Notice that we only assume a lower bound in  $\mathcal{B}_{ij}$  to guarantee sufficient ellipticity; there is no upper bound required for these bounds.

PROPOSITION 5.7. *Consider the evolution equation*

$$(5.52) \quad \partial_s \mathbf{u}(s) = -\mathcal{A}(s) \mathbf{u}(s), \quad \mathbf{u}(s) \in \mathbb{R}^\mathcal{K}$$

and fix  $\sigma > 0$ . Suppose that for some constant  $b$  we have

$$(5.53) \quad \mathcal{B}_{jk}(s) \geq \frac{b}{(j-k)^2}, \quad 0 \leq s \leq \sigma, \quad j \neq k,$$

and

$$(5.54) \quad \mathcal{W}_j(s) \geq \frac{b}{d_j}, \quad d_j := ||j| - K| + 1, \quad 0 \leq s \leq \sigma.$$

Then for any  $1 \leq p \leq q \leq \infty$  we have the decay estimate

$$(5.55) \quad \|\mathbf{u}(s)\|_q \leq (sb)^{-(\frac{1}{p}-\frac{1}{q})} \|\mathbf{u}(0)\|_p, \quad 0 < s \leq \sigma.$$

The proof relies on the usual Nash argument and uses the following critical Gagliardo-Nirenberg-type inequality for the discrete version of the operator  $\sqrt{-\Delta}$ :

PROPOSITION 5.8. *There exists a positive constant  $C$  such that*

$$(5.56) \quad \|f\|_{L^4(\mathbb{Z})}^4 \leq C \|f\|_{L^2(\mathbb{Z})}^2 \sum_{i \neq j \in \mathbb{Z}} \frac{|f_i - f_j|^2}{|i - j|^2}$$

holds for any function  $f : \mathbb{Z} \rightarrow \mathbb{R}$ .

The continuous version of this inequality,  $\|\phi\|_4^4 \leq C\|\phi\|_2^2\langle\phi, |p|\phi\rangle$ , was first proven in [71].

5.3.3. *Preparation for the De Giorgi-Nash-Moser bound: Finite speed of propagation.* The Hölder continuity of the parabolic equation (5.47) emerges only after a certain time, thus for the small  $\sigma$  regime in the integral (5.46) we need a different argument. Since we are interested in the Hölder continuity around the middle of the interval  $I$  (note that  $|p| \leq K^{1-\xi^*}$  in Theorem 5.5), and the initial condition  $\nabla h_0$  is small in this region (see (5.49)), a finite speed of propagation estimate guarantees that  $w_p(\sigma; \mathbf{x}(0), \sigma)$  is small if  $\sigma$  is not too large.

Since (5.47) is linear, for the finite speed of propagation it is sufficient to consider the fundamental solution. For  $a$  fixed, let  $\mathbf{u}^a(s)$  denote the solution

$$(5.57) \quad \partial_s \mathbf{u}^a(s) = -\mathcal{A}(s)\mathbf{u}^a(s), \quad \mathbf{u}^a(0) = \delta_{aj}.$$

with a delta function as initial data. We will assume that the coefficients of  $\mathcal{A}$  satisfy, for some fixed  $|Z| \leq K/2$  and  $\rho > 0$ , the bound

$$(5.58) \quad \sup_{0 \leq s \leq \sigma} \sup_{0 \leq M \leq K} \frac{1}{1+s} \int_0^s \frac{1}{M} \sum_{i \in I: |i-Z| \leq M} \sum_{j \in I: |j-Z| \leq M} \mathcal{B}_{ij}(\sigma - a) da \leq CK^\rho.$$

Notice that (5.58) is satisfied on the set of good path given by (5.50) and (5.51). The following lemma provides a finite speed of propagation estimate for the equation (5.57) under the condition of (5.58). This estimate is not optimal, but it is sufficient for our purpose.

LEMMA 5.9. [*Finite Speed of Propagation Estimate*] Fix  $a \in I$  and  $\sigma \leq CK \log K$ . We assume that the coefficients of  $\mathcal{A}$  satisfy (5.53) and (5.54) with  $b = K^{-\xi}$ . Assume that (5.58) is satisfied for some fixed  $Z$ ,  $|Z| \leq K/2$ . Then for the fundamental solution (5.57) we have the estimate for any  $s \leq \sigma$  and  $p \in I$

$$(5.59) \quad |u_p^a(s)| \leq \frac{CK^{\rho+2\xi+1/2}\sqrt{s+1}}{|p-a|}.$$

For the proof, we split the operator  $\mathcal{A} = \mathcal{S} + \mathcal{R}$  into a short range and a long range part, where the short range part  $\mathcal{S}(s)$  is defined by

$$(5.60) \quad (\mathcal{S}(s)\mathbf{v})_j := - \sum_{k: |j-k| \leq \ell} \mathcal{B}_{jk}(s)(v_k - v_j) + \mathcal{W}_j(s)v_j$$

with some cutoff parameter  $\ell$ . The norm of the long range part in any  $L^p$  is bounded by  $\ell^{-1}$  and it is treated as a perturbation via Duhamel formula. For the short range part, we control the exponentially weighted norm of the solution of  $\partial_s \mathbf{r}(s) = -\mathcal{S}(s)\mathbf{r}(s)$ , i.e. we derive a Gronwall bound for

$$f(s) = \sum_{j \in I} e^{|j-a|/\theta} r_j^2(s).$$

The result is

$$f(s) \leq \exp \left[ C\theta^{-2}\ell^2 \int_0^s \sum_{k,j: |j-k| \leq \ell} \mathcal{B}_{kj}(s') ds' \right] f(0).$$

The exponent is estimated by (5.58) with  $M = K$ . The optimization of the length-scale  $\theta$  together with the cutoff parameter  $\ell$  yields (5.59).  $\square$

Inserting the estimate (5.59) into (5.46) and using the estimate (5.49) on the initial data  $\nabla h_0$ , we obtain that the contribution of the short time regime,  $\sigma \leq K^{1/4}$ , is negligible if  $p$  is away from edge,  $|p| \leq K^{1-\xi^*}$  for some  $\xi^* > 0$ . This allows us to disregard the  $\sigma \leq K^{1/4}$  regime in (5.46) and focus on  $\sigma \in [K^{1/4}, CK \log K]$ .

5.3.4. *A discrete De Giorgi-Nash-Moser bound.* We will now treat the main part of the integral (5.46) by parabolic regularity. The preparations in the previous sections ensure that it is sufficient to consider the integration regime  $\sigma \in [K^{1/4}, CK \log K]$  and we can assume that the path  $\mathbf{x}(\cdot)$  is good in the sense of the estimates (5.50) and (5.51). In particular, the rigidity estimate implies not only lower bounds but also upper bounds for distant indices; more precisely we have

$$(5.61) \quad \mathcal{B}_{ij}(s) \leq \frac{C}{(i-j)^2}$$

for any  $|i-j| \geq CK^\xi$  and  $0 \leq s \leq CK \log K$ ; and similarly

$$(5.62) \quad \mathcal{W}_i(s) \leq \frac{K^\xi}{d_i}, \quad \text{if } d_i \geq K^{C\xi}.$$

The following regularity theorem combined with (5.49) completes the estimate of (5.46) and completes the proof of Theorem 5.5.  $\square$

**THEOREM 5.10** (Parabolic partial regularity with singular coefficients). *Let  $\mathbf{u}$  be a solution to (5.57), where  $\mathbf{u} = \mathbf{u}^a$  for any choice of  $a$ . Suppose that the coefficients of  $\mathcal{A}$  satisfy the lower bounds (5.53) and (5.54) with  $b = K^{-\xi}$ , the upper bounds (5.61), (5.62) for distant indices and the upper bound in (5.58) in average sense for all indices. Let  $\sigma \in [K^{c_1}, C_1 K \log K]$  be fixed, where  $c_1 > 0$  is an arbitrary positive constant. Then for any  $0 < q' < 1$  there exists  $q > 0$  so that for any  $|Z| \leq K/2$*

$$(5.63) \quad \sup_{\max(|j-Z|, |j'-Z|) \leq \sigma^{1-q'}} |u_j(\sigma) - u_{j'}(\sigma)| \leq C\sigma^{-1-q},$$

where  $\mathbf{u} = \mathbf{u}^a$  for any choice of  $a$ .

Notice that this result is deterministic, all probability estimates are comprised in verifying the conditions. We also remark that if we define the rescaled function  $v(j/K, t) := tu_j(t)$ , then (5.63) can be interpreted as a type of Hölder regularity of  $v$  on scale  $\sigma^{1-q'}K^{-1} \ll 1$  at the point  $Z/K$ :

$$(5.64) \quad |v(x, \sigma) - v(y, \sigma)| \leq \sigma^{-q} \leq |x - y|^{c_1 q}$$

for  $1/K \leq |x - y| \leq \sigma^{1-q'}/K$  and  $x, y$  near  $Z/K$ . The Hölder exponent is thus at least  $c_1 q$ .

Although the statement of Theorem 5.10 seems to be complicated, the underlying mechanism is that there is a positive exponent  $q$  in (5.63), which to a great degree is an universal constant. This exponent provides an extra smallness factor in addition to the natural size of  $u_j(\sigma)$ , which is  $\sigma^{-1}$  from the  $L^1 \rightarrow L^\infty$  decay. As (5.64) indicates, this gain comes from a Hölder regularity on the relevant scale.

Our equation (5.57) is of the type considered in [15], but it is discrete and in a finite interval. The key difference, however, is that the coefficient  $\mathcal{B}_{ij} = (x_i - x_j)^{-2}$  in the elliptic part of (5.57) can be singular if gaps close, even temporarily, while [15] assumed the uniform bound  $\mathcal{B}_{ij} \leq C/|i - j|^2$ . The only control we have for the singular behavior of  $\mathcal{B}_{ij}$  is the estimate (5.58) which is very weak. This estimate essentially says that the space-time maximum function of  $\mathcal{B}_{i,i+1}(t)$  at a fixed space-time point  $(Z, 0)$  is bounded by  $K^\rho$ . Our main task is to show that this condition is

sufficient for proving Hölder continuity at the same point. Our strategy follows the approach of Caffarelli-Chan-Vasseur [15]. The main new feature of our argument is the derivation of a local energy dissipation estimate for parabolic equation with singular coefficients satisfying (5.53), (5.54) as lower bounds and only (5.58) as an upper bound. The analogous result in [15], called the *first De Giorgi lemma*, is proved under uniform bounds on the coefficients. For our proof, roughly, we have to run the argument of the first De Giorgi lemma twice; first we get a bound only in  $L^2(\mathbb{Z})$  then using this information we upgrade it to an  $L^\infty(\mathbb{Z})$  bound. This concludes the sketch of the proof of Theorem 5.10.  $\square$

**5.4. From local measures to Wigner matrices and  $\beta$ -ensembles.** Given Theorem 5.1, the proofs of Theorem 1.3 and 1.5 follow relatively standard ideas from previous results, some of them were reviewed in Sections 3 and 4. The key inputs are to verify the condition (5.13) and to ensure that the configuration intervals coincide (5.11).

For the  $\beta$ -ensemble, (5.13) simply follows from conditioning the global rigidity estimate in Theorem 4.1. For matching the configuration intervals, first we match the local density by scaling and translation that guarantees that  $|J_{\mathbf{y}}| \sim |J_{\tilde{\mathbf{y}}}|$ , see (5.12). Then, with a second scaling, we fine tune the slight discrepancy between the lengths of  $J_{\mathbf{y}}$  and  $J_{\tilde{\mathbf{y}}}$ . This finishes the proof of Theorem 1.5.  $\square$

In the Wigner case, we always work on the same configuration interval, so matching of  $J$  is automatic. The proof of (5.13), however, requires a bit more effort than for the  $\beta$ -ensemble, but it will relatively easily follow from other information we already collected along the three step strategy described in Section 1.5. As we explained in the proof of Theorem 1.2, the averaging over the energy was really needed only in the second step, where the closeness of the local statistics of  $f_t\mu$  and  $\mu$  was shown for small  $t$ , where  $f_t$  is the evolution of the DBM. As a byproduct of this step, we obtain bounds on the global entropy and Dirichlet form, see (3.38). In particular, the local Dirichlet form w.r.t  $\mu_{\mathbf{y}}$  can be estimated by the global one, which then can be used to compare expectations w.r.t. the conditional measures  $f_{t,\mathbf{y}}\mu_{\mathbf{y}}$  and  $\mu_{\mathbf{y}}$ ;

$$(5.65) \quad \mathbb{E}^{f_{t,\mathbf{y}}\mu_{\mathbf{y}}} O(\mathbf{x}) - \mathbb{E}^{\mu_{\mathbf{y}}} O(\mathbf{x}).$$

We are especially interested in controlling the difference

$$(5.66) \quad |\mathbb{E}^{f_{t,\mathbf{y}}\mu_{\mathbf{y}}} x_j - \mathbb{E}^{\mu_{\mathbf{y}}} x_j| \leq CK^\xi N^{-1}.$$

Since  $\mathbb{E}^{f_t\mu} x_j$  is close to its classical location  $\gamma_j$  by rigidity (5.16) for Wigner matrices, after conditioning, we obtain that  $\mathbb{E}^{f_{t,\mathbf{y}}\mu_{\mathbf{y}}} x_j$  is also close to  $\gamma_j$ , at least for most  $\mathbf{y}$  w.r.t.  $f_t\mu$ . Combining this information with (5.66) yields (5.13). Therefore Theorem 5.1 applies and we will use it for a Gaussian case,  $V(x) = \tilde{V}(x) = x^2/2$  but with two different boundary conditions  $\mathbf{y}, \tilde{\mathbf{y}} \in \mathbb{R}$ . Since this holds for most  $\tilde{\mathbf{y}}$  w.r.t the measure  $\mu$ , it also holds for  $\mu$  itself, i.e. the gap statistics of  $\mu_{\mathbf{y}}$  and  $\mu$  coincide. On the other hand, the estimate (5.65) applied to the observable  $O(x_j - x_{j+1})$  implies directly that the single gap distribution w.r.t.  $f_{t,\mathbf{y}}\mu_{\mathbf{y}}$  and  $\mu_{\mathbf{y}}$  coincide for most of the  $\mathbf{y}$  w.r.t.  $f_t\mu$ . Finally, the gap statistics of  $f_{t,\mathbf{y}}\mu_{\mathbf{y}}$  and  $f_t\mu$  coincide for most  $\mathbf{y}$  w.r.t.  $f_t\mu$  by conditioning. Putting these relations together we obtain that the gap statistics of  $f_t\mu$  and  $\mu$  coincide, i.e. the local measures, that played an important auxiliary role, are eliminated.

Finally, the small Gaussian component present in  $f_t\mu$  for small but non-zero  $t$  can be removed by the Green function comparison theorem, Theorem 3.6. Although the direct application of the Green functions give information only on eigenvalues around a fixed energy and not on an eigenvalue with a fixed label, the estimates are strong enough to transfer fixed energy information to fixed label. The main reason for this flexibility is that Theorem 3.6 allows for very small  $\eta \sim N^{-1-\varepsilon}$ , i.e. well below the typical spacing. Indeed, Theorem 1.10 from [59] implies that if the first four moments of two generalized Wigner ensembles,  $H^{\mathbf{v}}$  and  $H^{\mathbf{w}}$ , are the same, then we have

$$(5.67) \quad \lim_{N \rightarrow \infty} [\mathbb{E}^{\mathbf{v}} - \mathbb{E}^{\mathbf{w}}] O(N(x_j - x_{j+1}), N(x_j - x_{j+2}), \dots, N(x_j - x_{j+n})) = 0.$$

Roughly speaking, the proof of (5.67) in [59] was based on Theorem 3.6. In order to convert fixed energy to a fixed eigenvalue index, one needs to know that the total number of eigenvalues up to a fixed energy is the same for the two ensembles. The total number of eigenvalues up to a fixed energy  $E$  can be expressed in terms of integration of imaginary part of the trace of Green functions, i.e.,

$$\int_{-\infty}^E dy \operatorname{Im} \operatorname{Tr} \frac{1}{H - (y + i\eta)}$$

with an  $\eta$  slightly smaller than  $1/N$ . Thus the basic idea of the Green function comparison theorem can be employed and this leads to (5.67). This completes the proof of Theorem 1.3.  $\square$

## References

- [1] Aizenman, M., and Molchanov, S.: Localization at large disorder and at extreme energies: an elementary derivation, *Commun. Math. Phys.* **157**, 245–278 (1993)
- [2] Anantharaman, N., Nonnenmacher, S.: Half-delocalization of eigenfunctions for the Laplacian on an Anosov manifold. *Annales de l'Institut Fourier* **57**, no. 7, 2465–2523 (2007)
- [3] Anderson, G., Guionnet, A., Zeitouni, O.: An Introduction to Random Matrices. Studies in advanced mathematics, **118**, Cambridge University Press, 2009.
- [4] Anderson, P.: Absences of diffusion in certain random lattices, *Phys. Rev.* **109**, 1492–1505 (1958)
- [5] Albeverio, S., Pastur, L., Shcherbina, M.: On the  $1/n$  expansion for some unitary invariant ensembles of random matrices, *Commun. Math. Phys.* **224**, 271–305 (2001).
- [6] Bakry, D., Émery, M.: Diffusions hypercontractives. In: *Séminaire de probabilités, XIX, 1983/84, vol. 1123 of Lecture Notes in Math.* Springer, Berlin, 1985, pp. 177–206.
- [7] Ben Arous, G., Pécché, S.: Universality of local eigenvalue statistics for some sample covariance matrices. *Comm. Pure Appl. Math.* **LVIII**, (2005), 1–42.
- [8] Berry, M.V., Tabor, M.: Level clustering in the regular spectrum, *Proc. Roy. Soc. A* **356** (1977) 375–394
- [9] Bleher, P., Its, A.: Semiclassical asymptotics of orthogonal polynomials, Riemann-Hilbert problem, and universality in the matrix model. *Ann. of Math.* **150** (1999), 185–266.
- [10] Bohigas, O.; Giannoni, M.-J.; Schmit, C.: Characterization of chaotic quantum spectra and universality of level fluctuation laws. *Phys. Rev. Lett.* **52**, no. 1, 1–4, (1984)
- [11] Bourgade, P., Erdős, Yau, H.-T.: Universality of General  $\beta$ -Ensembles, arXiv:1104.2272
- [12] Bourgade, P., Erdős, Yau, H.-T.: Bulk Universality of General  $\beta$ -Ensembles with Non-convex Potential, *J. Math. Phys.* **53**, 095221 (2012)
- [13] Bourgade, P., Erdős, Yau, H.-T.: Edge Universality of General  $\beta$ -Ensembles. In preparation.
- [14] Brézin, E., Hikami, S.: Correlations of nearby levels induced by a random potential. *Nucl. Phys. B* **479** (1996), 697–706, and Spectral form factor in a random matrix theory. *Phys. Rev. E* **55**, 4067–4083 (1997)
- [15] Caffarelli, L., Chan, C.H., Vasseur, A.: Regularity theory for parabolic nonlinear integral operators, *J. Amer. Math. Soc.* **24**, no. 3, 849–889 (2011)

- [16] Chatterjee, S.: A generalization of the Lindeberg principle. *Ann. Probab.* **34**, no. 6, 2061–2076 (2006)
- [17] Deuschel, J.-D., Giacomin, G., Ioffe, D.: Large deviations and concentration properties for  $\nabla\varphi$  interface models. *Probab. Theor. Relat. Fields.* **117**, 49–111 (2000)
- [18] Deift, P.: Orthogonal polynomials and random matrices: a Riemann-Hilbert approach. *Courant Lecture Notes in Mathematics* **3**, American Mathematical Society, Providence, RI, 1999
- [19] Deift, P., Gioev, D.: Universality in random matrix theory for orthogonal and symplectic ensembles. *Int. Math. Res. Pap. IMRP* 2007, no. 2, Art. ID rpm004, 116 pp
- [20] Deift, P., Gioev, D.: Random Matrix Theory: Invariant Ensembles and Universality. *Courant Lecture Notes in Mathematics* **18**, American Mathematical Society, Providence, RI, 2009
- [21] Deift, P., Kriecherbauer, T., McLaughlin, K.T-R, Venakides, S., Zhou, X.: 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**, 1335–1425 (1999)
- [22] Deift, P., Kriecherbauer, T., McLaughlin, K.T-R, Venakides, S., Zhou, X.: Strong asymptotics of orthogonal polynomials with respect to exponential weights. *Comm. Pure Appl. Math.* **52**, 1491–1552 (1999)
- [23] Dumitriu, I., Edelman, A.: Matrix Models for Beta Ensembles, *Journal of Mathematical Physics* **43** (11), 5830–5847 (2002)
- [24] Dyson, F.J.: Statistical theory of energy levels of complex systems, I, II, and III. *J. Math. Phys.* **3**, 140–156, 157–165, 166–175 (1962)
- [25] Dyson, F.J.: A Brownian-motion model for the eigenvalues of a random matrix. *J. Math. Phys.* **3**, 1191–1198 (1962)
- [26] Dyson, F.J.: Correlations between eigenvalues of a random matrix. *Commun. Math. Phys.* **19**, 235–250 (1970)
- [27] Erdős, L.: Universality of Wigner Random Matrices: a Survey of Recent Results. *Russian Math. Surveys* **66** (3) 67–198.
- [28] Erdős, L., Knowles, A.: Quantum Diffusion and Eigenfunction Delocalization in a Random Band Matrix Model. *Commun. Math. Phys.* **303** no. 2, 509–554 (2011)
- [29] Erdős, L., A. Knowles, A.: Quantum Diffusion and Delocalization for Band Matrices with General Distribution. *Annales Inst. H. Poincaré*, **12** (7), 1227–1319 (2011)
- [30] Erdős, L., A. Knowles, A., Yau, H.-T.: Averaging Fluctuations in Resolvents of Random Band Matrices. Preprint. arXiv:1205.5664.
- [31] Erdős, L., A. Knowles, A., Yau, H.-T., J. Yin.: The local semicircle law for a general class of random matrices. Preprint. arXiv:1212.0164.
- [32] Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Spectral Statistics of Erdős-Rényi Graphs I: Local Semicircle Law. To appear in *Annals Probab.* Preprint: arXiv:1103.1919
- [33] Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Spectral Statistics of Erdős-Rényi Graphs II: Eigenvalue Spacing and the Extreme Eigenvalues. *Comm. Math. Phys.* **314** no. 3. 587–640 (2012)
- [34] Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Delocalization and Diffusion Profile for Random Band Matrices. Preprint: arXiv:1205.5669
- [35] Erdős, L., Péché, G., Ramírez, J., Schlein, B., and Yau, H.-T., Bulk universality for Wigner matrices. *Commun. Pure Appl. Math.* **63**, No. 7, 895–925 (2010)
- [36] Erdős, L., Ramirez, J., Schlein, B., Tao, T., Vu, V., Yau, H.-T.: Bulk Universality for Wigner Hermitian matrices with subexponential decay. *Math. Res. Lett.* **17** (2010), no. 4, 667–674.
- [37] Erdős, L., Ramirez, J., Schlein, B., Yau, H.-T.: Universality of sine-kernel for Wigner matrices with a small Gaussian perturbation. *Electron. J. Probab.* **15**, Paper 18, 526–604 (2010)
- [38] Erdős, L., Schlein, B., Yau, H.-T.: Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices. *Ann. Probab.* **37**, No. 3, 815–852 (2009)
- [39] Erdős, L., Schlein, B., Yau, H.-T.: Local semicircle law and complete delocalization for Wigner random matrices. *Commun. Math. Phys.* **287**, 641–655 (2009)
- [40] Erdős, L., Schlein, B., Yau, H.-T.: Wegner estimate and level repulsion for Wigner random matrices. *Int. Math. Res. Notices.* **2010**, No. 3, 436–479 (2010)
- [41] Erdős, L., Schlein, B., Yau, H.-T.: Universality of random matrices and local relaxation flow. *Invent. Math.* **185** (2011), no.1, 75–119.



- [42] Erdős, L., Schlein, B., Yau, H.-T., Yin, J.: The local relaxation flow approach to universality of the local statistics for random matrices. *Annales Inst. H. Poincaré (B), Probability and Statistics* **48**, no. 1, 1–46 (2012)
- [43] Erdős, L., Yau, H.-T.: Universality of local spectral statistics of random matrices. *Bull. Amer. Math. Soc.* **49**, no.3 (2012), 377–414.
- [44] Erdős, L., Yau, H.-T.: A comment on the Wigner-Dyson-Mehta bulk universality conjecture for Wigner matrices. *Electron. J. Probab.* **17**, no 28, 1–5 (2012)
- [45] Erdős, L., Yau, H.-T.: Gap universality of generalized Wigner and  $\beta$ -ensembles. Preprint arXiv:1211.3786
- [46] Erdős, L., Yau, H.-T., Yin, J.: Bulk universality for generalized Wigner matrices. To appear in *Prob. Theor. Rel. Fields*. Preprint arXiv:1001.3453
- [47] Erdős, L., Yau, H.-T., Yin, J.: Universality for generalized Wigner matrices with Bernoulli distribution. *J. of Combinatorics*, **1**, no. 2, 15–85 (2011)
- [48] Erdős, L., Yau, H.-T., Yin, J.: Rigidity of Eigenvalues of Generalized Wigner Matrices. *Adv. Math.* **229**, no. 3, 1435–1515 (2012)
- [49] Eynard, B.: Master loop equations, free energy and correlations for the chain of matrices. *J. High Energy Phys.* **11**, 018 (2003)
- [50] Fokas, A. S., Its, A. R., Kitaev, A. V.: The isomonodromy approach to matrix models in 2D quantum gravity. *Comm. Math. Phys.* **147**, 395–430 (1992)
- [51] Fröhlich, J., Spencer, T.: Absence of diffusion in the Anderson tight binding model for large disorder or low energy, *Commun. Math. Phys.* **88**, 151–184 (1983)
- [52] Fyodorov, Y.V. and Mirlin, A.D.: Scaling properties of localization in random band matrices: a  $\sigma$ -model approach. *Phys. Rev. Lett.* **67**, 2405–2409 (1991)
- [53] Gaudin, M.: Sur la loi limit de l’espacement des valeurs propres d’une matrice aléatoire. *Nucl. Phys.* **25**, 447–458.
- [54] Giacomin, G., Olla, S., Spohn, H.: Equilibrium fluctuations for  $\nabla\varphi$  interface model. *Ann. Probab.* **29**, no.3., 1138–1172 (2001)
- [55] Gustavsson, J.: Gaussian fluctuations of eigenvalues in the GUE. *Ann. Inst. H. Poincaré Probab. Statist.* **41**, no. 2, 151–178 (2005)
- [56] Helffer, B., Sjöstrand, J.: On the correlation for Kac-like models in the convex case. *J. Statist. Phys.* **74**, no.1-2, 349–409 (1994)
- [57] Johansson, K.: Universality of the local spacing distribution in certain ensembles of Hermitian Wigner matrices. *Comm. Math. Phys.* **215**, no.3. 683–705 (2001)
- [58] Johansson, K.: On the fluctuations of eigenvalues of random Hermitian matrices. *Duke Math. J.* **91**, 151–204 (1998)
- [59] Knowles, A., Yin, J.: Eigenvector distribution of Wigner matrices. Preprint arXiv:1102.0057.
- [60] Kriecherbauer, T., Shcherbina, M.: Fluctuations of eigenvalues of matrix models and their applications. Preprint arXiv:1003.6121
- [61] Lee, J. O., Yin, J.: A Necessary and Sufficient Condition for Edge Universality of Wigner matrices. Preprint. arXiv:1206.2251
- [62] Lindenstrauss, E.: Invariant measures and arithmetic quantum ergodicity *Ann. Math.* **163**, 165–219 (2006)
- [63] Lubinsky, D.S.: A New Approach to Universality Limits Involving Orthogonal Polynomials, *Ann. Math.*, **170**, 915–939 (2009)
- [64] Marklof, J.: Energy level statistics, lattice point problems and almost modular functions, in Cartier, Julia, Moussa, Vanhove (Herausgeber): *Frontiers in Number Theory, Physics and Geometry (Les Houches Lectures 2003)*, Band 1, Springer Verlag 2006, S. 163–181
- [65] Mehta, M.L.: *Random Matrices*. Third Edition, Academic Press, New York, 1991.
- [66] Mehta, M.L.: A note on correlations between eigenvalues of a random matrix. *Commun. Math. Phys.* **20** no.3. 245–250 (1971)
- [67] Mehta, M.L., Gaudin, M.: On the density of eigenvalues of a random matrix. *Nuclear Phys.* **18**, 420–427 (1960).
- [68] Minami, N.: Local fluctuation of the spectrum of a multidimensional Anderson tight binding model. *Commun. Math. Phys.* **177**, 709–725 (1996)
- [69] Montgomery, H.L.: The pair correlation of zeros of the zeta function. *Analytic number theory, Proc. of Sympos. in Pure Math.* **24**, Amer. Math. Soc. Providence, R.I., 181–193 (1973).
- [70] Naddaf, A., Spencer, T.: On homogenization and scaling limit of some gradient perturbations of a massless free field, *Commun. Math. Phys.* **183**, no.1., 55–84 (1997)

- [71] Ogawa, T.; Ozawa, T.: Trudinger type inequalities and uniqueness of weak solutions for the nonlinear Schrödinger mixed problem. *J. Math. Anal. Appl.* **155**, no. 2, 531–540 (1991)
- [72] Pastur, L., Shcherbina, M.: Universality of the local eigenvalue statistics for a class of unitary invariant random matrix ensembles. *J. Stat. Phys.* **86**, 109–147 (1997)
- [73] Pastur, L., Shcherbina M.: Bulk universality and related properties of Hermitian matrix models. *J. Stat. Phys.* **130**, no.2., 205–250 (2008)
- [74] Pillai, N.S. and Yin, J.: Universality of covariance matrices. Preprint arXiv:1110.2501
- [75] Pillai, N.S. and Yin, J.: Edge universality of covariance matrices. Preprint arXiv:1112.2381
- [76] Ramirez, J., Rider, B., Virág, B.: Beta ensembles, stochastic Airy spectrum, and a diffusion. *J. Amer. Math. Soc.* **24**, 919–944 (2011)
- [77] Rudnick, Z. and Sarnak, P.: The pair correlation function of fractional parts of polynomials, *Comm. Math. Phys.* **194**, 61–70 (1998)
- [78] Schenker, J.: Eigenvector localization for random band matrices with power law band width. *Commun. Math. Phys.* **290**, 1065–1097 (2009)
- [79] Shcherbina, M.: Orthogonal and symplectic matrix models: universality and other properties. *Comm. Math. Phys.* **307**, no.3., 761–790 (2011)
- [80] Sinai, Y.: Poisson distribution in a geometrical problem, *Adv. Sov. Math.* AMS Publ. **3**, 199–215 (1991)
- [81] Spencer, T.: Random banded and sparse matrices (Chapter 23), to appear in “Oxford Handbook of Random Matrix Theory” edited by G. Akemann, J. Baik, and P. Di Francesco
- [82] Tao, T. and Vu, V.: Random matrices: Universality of the local eigenvalue statistics. *Acta Math.*, **206**, no. 1, 127–204 (2011)
- [83] Tao, T. and Vu, V.: The Wigner-Dyson-Mehta bulk universality conjecture for Wigner matrices. *Electron. J. Probab.* **16**, no 77, 2104–2121 (2011)
- [84] Tao, T.: The asymptotic distribution of a single eigenvalue gap of a Wigner matrix. Preprint. arxiv:1203.1605
- [85] Valkó, B.; Virág, B.: Continuum limits of random matrices and the Brownian carousel. *Invent. Math.* **177**, no. 3, 463–508 (2009)
- [86] Widom H.: On the relation between orthogonal, symplectic and unitary matrix ensembles. *J. Statist. Phys.* **94**, no. 3-4, 347–363 (1999)
- [87] Wigner, E.: Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math.* **62**, 548–564 (1955)
- [88] Yau, H. T.: Relative entropy and the hydrodynamics of Ginzburg-Landau models, *Lett. Math. Phys.* **22**, 63–80 (1991)

INSTITUTE OF MATHEMATICS, UNIVERSITY OF MUNICH, THERESIENSTR. 39, D-80333 MUNICH, GERMANY

*E-mail address:* lerdos@math.lmu.de