# Bulk Universality for Wigner Matrices

László Erdős[1]*, José A. Ramírez[2], Benjamin Schlein[3] and Horng-Tzer Yau[4]†

Institute of Mathematics, University of Munich,
Theresienstr. 39, D-80333 Munich, Germany[1]

Department of Mathematics, Universidad de Costa Rica
San Jose 2060, Costa Rica

Department of Pure Mathematics and Mathematical Statistics
University of Cambridge
Wilberforce Rd, Cambridge CB3 0WB, UK[3]

Department of Mathematics, Harvard University
Cambridge MA 02138, USA[4]

May 26, 2009

**Abstract**

We consider $N \times N$ Hermitian Wigner random matrices $H$ where the probability density for each matrix element is given by the density $\nu(x) = e^{-U(x)}$. We prove that the eigenvalue statistics in the bulk is given by Dyson sine kernel provided that $U \in C^6(\mathbb{R})$ with at most polynomially growing derivatives and $\nu(x) \leq C\, e^{-C|x|}$ for $x$ large. The proof is based upon an approximate time reversal of the Dyson Brownian motion combined with the convergence of the eigenvalue density to the Wigner semicircle law on short scales.

**AMS Subject Classification:** 15A52, 82B44

*Running title:* Universality for Wigner matrices

*Key words:* Wigner random matrix, Dyson sine kernel.

*Partially supported by SFB-TR 12 Grant of the German Research Council
†Partially supported by NSF grants DMS-0757425, DMS-0804279

# 1 Introduction

The fundamental reason why random matrices have been used to model many large systems is based on the belief that their local eigenvalue statistics are universal. This is generally referred to as the universality of random matrices. It is well-known that the local behavior of eigenvalues near the spectral edge and in the bulk are governed by the Tracy-Widom law and by the Dyson sine kernel, respectively. Since the seminal work of Dyson [4] for the Gaussian Unitary Ensemble (GUE), the universality both for the edge and the bulk were proven for very general classes of unitary invariant ensembles in the past two decades (see, e.g. [11, 12, 3] and references therein). For non-unitary ensembles, the most natural examples are the Wigner matrix ensembles [14], i.e., random matrices with independent identically distributed entries. The edge universality for these ensembles was proved by Soshnikov [13] using the moment method; the bulk universality remained unknown due to a lack of method to analyze local spectral properties of large matrices inside the spectrum. For ensembles of the form

$$\widehat{H} + aV, \tag{1.1}$$

where $\widehat{H}$ is a Wigner matrix, $V$ is an independent standard GUE matrix and $a$ is a positive constant of order one (independent of $N$), the bulk universality was proved by Johansson [10]. (Strictly speaking, the range of the parameter $a$ in [10] depends on the energy $E$. This restriction was later removed by Ben Arous and Péché [1], who also extended this approach to Wishart ensembles).

The approach of [10] is based on the asymptotic analysis of an explicit formula by Brézin-Hikami [2] for the correlation functions of the eigenvalues of $\widehat{H} + aV$. This matrix can also be generated by a stochastic flow

$$s \to \widehat{H} + \sqrt{s}V, \quad s > 0,$$

and the evolution of the eigenvalues is given by the Dyson Brownian motion [5]. The result of [10, 1] thus states that the bulk universality holds for times of order one. The eigenvalue distribution of GUE is in fact the invariant measure of Dyson Brownian motion. (Rigorously speaking, the Brownian motion has to be replaced by an Ornstein-Uhlenbeck process, but we will neglect this subtlety.) It is thus tempting to derive the universality of $\widehat{H} + \sqrt{s}V$ via the convergence to equilibrium. We have recently carried out this approach [9] and the key observation is that the sine kernel, as a property of local statistics, depends almost exclusively on the convergence to local equilibrium. With this method we have reduced the necessary time to $N^{-1+\xi}$, for any $\xi > 1/4$ in [9]. Note that the relaxation time to local equilibrium is $N^{-1}$; the additional exponent $\xi$ is due to technical reasons.

From the stochastic calculus, one can see that the typical distance between the corresponding eigenvalues of $\widehat{H} + \sqrt{s}V$ and $\widehat{H}$ is of order $(s/N)^{1/2}$. Thus the bulk universality of $\widehat{H}$ would hold if we could prove the Dyson sine kernel for time $s \ll 1/N$. On the other hand, for time smaller than $1/N$, the eigenvalues do not move in the scale $1/N$ and the dynamical consideration seems to be pointless. In fact, the explicit formula of the correlation functions in [10] becomes very unstable and no conclusion can be obtained as $s \searrow 1/N$. In this paper, we provide an approach to address the comparison of eigenvalues between $\widehat{H} + \sqrt{s}V$ and $\widehat{H}$. To describe the idea, we now introduce the notations.

Fix $N \in \mathbb{N}$ and we consider a Hermitian matrix ensemble of $N \times N$ matrices $H = (h_{\ell k})$ with the normalization

$$h_{\ell k} = N^{-1/2} z_{\ell k}, \qquad z_{\ell k} = x_{\ell k} + i y_{\ell k}, \tag{1.2}$$

where $x_{\ell k}, y_{\ell k}$ for $\ell < k$ are independent, identically distributed random variables with distribution $\nu$ that has zero expectation and variance $\frac{1}{2}$. The diagonal elements are real, i.e. $y_{\ell\ell} = 0$ and $x_{\ell\ell}$ are also i.i.d. with distribution $\widetilde{\nu}$ that has zero expectation and variance one. The diagonal elements are independent from the off-diagonal ones.

Suppose the real and imaginary parts of the offdiagonal matrix elements evolve according to the Ornstein-Uhlenbeck (OU) process

$$\partial_t u_t = L u_t, \quad L = \frac{1}{4}\frac{\partial^2}{\partial x^2} - \frac{x}{2}\frac{\partial}{\partial x} \tag{1.3}$$

with the reversible measure $\mu(\mathrm{d}x) = e^{-x^2}\mathrm{d}x$ and initial distribution $u_0 = u$ (strictly speaking, a differently normalized OU process is used for the diagonal elements but we omit this detail here). Under this process, the matrix evolves as

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

and the expectation and variance of the matrix entries remain constant. Notice for time $t$ small, $t \approx a^2$ when compared with (1.1), after a trivial rescaling.

The initial distribution of all the matrix elements is $F\,\mathrm{d}\mu^{\otimes n} = (u\,\mathrm{d}\mu)^{\otimes n}$ with $n = N^2$. Let $\mathcal{L}$ be the generator on the product space and $e^{t\mathcal{L}} := (e^{tL})^{\otimes n}$ be the dynamics of the OU process for all the matrix elements. The joint probability distribution of the matrix elements at time $t$ is then given by

$$F_t \mathrm{d}\mu^{\otimes n} := e^{t\mathcal{L}} u^{\otimes n}\,\mathrm{d}\mu^{\otimes n} = (e^{tL}u)^{\otimes n}\,\mathrm{d}\mu^{\otimes n}.$$

Suppose that for some $t$ small, say, $t = N^{-1+\lambda}$ with $\lambda > 0$, we know the local eigenvalue correlation function w.r.t. $F_t$. Let

$$Var(F, F_t) = \int |F - F_t| \mathrm{d}\mu^{\otimes n}$$

be the total variation norm between $F_t$ to $F$. In order to approximate the correlation functions of $F$ by $F_t$ in a weak sense (tested against bounded observables), we need $Var(F, F_t) \to 0$. Heuristically, $Var(F, F_t) \sim tN^2$ and this requires that $t \ll N^{-2}$ which is far from the time scale $t \geq N^{-1+\xi}$ for which the sine kernel has been proven in [9]. For observables on short scales, an effective speed of convergence for the total variation is needed. For example, to test a local observable with two variables in scale $1/N$, as in the case of the Dyson sine kernel, one has to prove $Var(F, F_t) = o(N^{-2})$.

Although the heuristic bound $Var(F, F_t) \sim tN^2$ can be improved to $Var(F, F_t) \sim tN$, further improvement seems to be impossible. Thus we are unable to obtain even the weaker bound $Var(F, F_t) = o(1)$ for $t > 1/N$. The main observation in the current paper is that, while we cannot compare $F$ with $F_t$, it suffices to prove the existence of some function $G$ for which the correlation functions with respect to $e^{t\mathcal{L}}G$ can be computed for $t \geq N^{-1+\lambda}$ and $Var(F, e^{t\mathcal{L}}G) = o(N^{-2})$. Since the necessary input to compute the correlation functions is the validity of the semicircle law on short scales, which we have proved for a wide class of distributions $\nu$ in [6, 7, 8], the choice of $G$ is essentially dominated by the condition $Var(F, e^{t\mathcal{L}}G) = o(N^{-2})$. Note that $G$ itself may depend on $t$. Since $F = e^{t\mathcal{L}}(e^{-t\mathcal{L}}F)$, we could, in principle, choose $G = e^{-t\mathcal{L}}F = [e^{-tL}]^{\otimes n}F$. But the diffusive dynamics cannot be reversed besides a very special class of initial data $G$. However, we only have to approximately reverse the dynamics and the choice $G_t = \big[1 - tL + \frac{1}{2}t^2L^2\big]^{\otimes n}F$ turns out to be sufficient. In this case, $e^{t\mathcal{L}}G_t - F = O(N^2t^3)$ and we will show that

$$\big|Var(e^{t\mathcal{L}}G_t, F)\big|^2 \leq \int \frac{|e^{t\mathcal{L}}G_t - F|^2}{e^{t\mathcal{L}}G_t}\mathrm{d}\mu^{\otimes n} = O(t^6N^2). \tag{1.4}$$

Furthermore, under some mild regularity condition on $F$, $G_t$ is in the class for which we can establish the local semicircle law [8]. We will call this argument the *method of time reversal.*

We now summarize the assumptions on the initial distribution. Let the probability measure of the real and imaginary parts of the off-diagonal matrix elements be of the form

$$\nu(\mathrm{d}x) = e^{-U(x)}\mathrm{d}x = u(x)\mu(\mathrm{d}x) = e^{-V(x)}e^{-x^2}\mathrm{d}x$$

with the real function $V(x) = U(x) - x^2$ and similarly for the diagonal elements $\widetilde{\nu}(\mathrm{d}x) = e^{-\widetilde{U}(x)}\mathrm{d}x$, $\widetilde{V}(x) = \widetilde{U}(x) - \frac{1}{2}x^2$. Suppose that $V \in C^6$ and the derivatives satisfy

$$\sum_{j=1}^{6} |V^{(j)}(x)| \le C(1+x^2)^k \tag{1.5}$$

for some $k \in \mathbb{N}$ and

$$\nu(x) \le C' e^{-\delta|x|^2} \tag{1.6}$$

with some constants $\delta > 0$, $C$ and $C'$. In Section 5 we explain how to relax this latter condition to exponential decay,

$$\nu(x) \le C' e^{-C|x|} \tag{1.7}$$

with some constants $C, C'$ (in fact, some high power law decay is sufficient). We assume that the first moment of $\nu$ is zero and the variance is $\frac{1}{2}$

$$\int x\,\mathrm{d}\nu(x) = 0, \qquad \int x^2 \mathrm{d}\nu(x) = \frac{1}{2}. \tag{1.8}$$

We assume the conditions (1.5), (1.6) and (1.8) for $\widetilde{V}$ as well with the variance changed to 1.

Let $p_N(x_1, x_2, \ldots, x_N)$ denote the probability density of eigenvalues and for any $k = 1, 2, \ldots, N$, let

$$p_N^{(k)}(x_1, x_2, \ldots x_k) := \int_{\mathbb{R}^{N-k}} p_N(x_1, x_2, \ldots, x_N)\mathrm{d}x_{k+1}\ldots\mathrm{d}x_N \tag{1.9}$$

be the $k$-point correlation function. With our choice of the variance of $\nu$, the density $p_N^{(1)}(x)$ is supported in $[-2,2] + o(1)$ and in the $N \to \infty$ limit it converges to the Wigner semicircle law given by the density

$$\varrho_{sc}(x) = \frac{1}{2\pi}\sqrt{(4-x^2)_+}\,. \tag{1.10}$$

**Theorem 1.1** *Let the probability measure of the matrix elements satisfy conditions* (1.5)*,* (1.6) *and* (1.8)*. Then for any $u$ with $|u| < 2$ and for any compactly supported and bounded observable $O \in L_c^\infty(\mathbb{R}^2)$ we have*

$$\begin{aligned}\lim_{N\to\infty} \int_{\mathbb{R}^2} O(\alpha,\beta)\frac{1}{[\varrho_{sc}(u)]^2} p_N^{(2)}\Big(u + \frac{\alpha}{N\varrho_{sc}(u)}, u + \frac{\beta}{N\varrho_{sc}(u)}\Big)\mathrm{d}\alpha\mathrm{d}\beta \\ = \int_{\mathbb{R}^2} O(\alpha,\beta)\Big[1 - \Big(\frac{\sin\pi(\alpha-\beta)}{\pi(\alpha-\beta)}\Big)^2\Big]\mathrm{d}\alpha\mathrm{d}\beta.\end{aligned} \tag{1.11}$$

We now state our result concerning the eigenvalue gap distribution. For any $s > 0$ and $|u| < 2$ we define the density of eigenvalue pairs with distance less than $s/N\varrho_{sc}(u)$ in the vicinity of $u$ by

$$\Lambda(u; s, x) = \frac{1}{2Nt_N\varrho_{sc}(u)} \#\Big\{1 \le j \le N-1 \,:\, x_{j+1} - x_j \le \frac{s}{N\varrho_{sc}(u)},\ |x_j - u| \le t_N\Big\} \qquad (1.12)$$

where $t_N = N^{-1+\delta}$ for some $0 < \delta < 1$.

**Theorem 1.2** *Suppose the probability measure of the matrix elements satisfies conditions* (1.5), (1.6) *and* (1.8). *Let $\mathcal{K}_\alpha$ be the operator acting on $L^2((0,\alpha))$ with kernel $\frac{\sin\pi(x-y)}{\pi(x-y)}$. Then for any $u$ with $|u| < 2$ and for any $s > 0$ we have*

$$\lim_{N\to\infty} \mathbb{E}\,\Lambda(u; s, x) = \int_0^s p(\alpha)\,\mathrm{d}\alpha, \qquad p(\alpha) = \frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\det(1 - \mathcal{K}_\alpha), \qquad (1.13)$$

*where* det *denotes the Fredholm determinant of the compact operator $1 - \mathcal{K}_\alpha$.*

As a corollary of Theorem 1.2, the probability to find no eigenvalue in the interval $[0, \alpha]$ is given by $\det(1 - \mathcal{K}_\alpha)$, same as in the case of GUE (see, e.g., [3]).

The proof of Theorem 1.1 and 1.2 consists of two main parts. In Section 2 we prove the approximation (1.4) under precise conditions on the initial distribution $u = e^{-V}$. In Section 3 we combine the formula in [10] with the local semicircle law from [8] to prove the sine kernel for the distribution $e^{t\mathcal{L}}G_t$ with $t = N^{-1+\lambda}$ for any $\lambda > 0$. We complete the proofs of the main theorems in Section 4.

The method of time reversal described previously is very general and should be applicable to a wide range of models. More significantly, it explains the *origin* of the universality, i.e., the universality comes from the "time reversal". To summarize, the universality consists of the following observations: (1) The local statistics are determined by the local equilibrium measures. (2) The relaxation to local equilibria takes place in a short time. (3) The original distribution can be well-approximated by the distribution of the Dyson Brownian motion for a short time with initial data given by an approximate inverse flow. To implement this scheme, a key input is to estimate the fluctuations of the empirical density of eigenvalues in short scales.

*Conventions.* We will use the letters $C$ and $c$ to denote general constants whose precise values are irrelevant and they may change from line to line. These constants may depend on the constants in (1.5)–(1.8).

# 2 Method of Time Reversal

Recall the Ornstein-Uhlenbeck process from (1.3) with the reversible measure $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x = e^{-x^2}\mathrm{d}x$. Let $u$ be a positive density with respect to $\mu$, i.e. $\int u\mathrm{d}\mu = 1$ and we write $u(x) = \exp(-V(x))$.

**Proposition 2.1** *Let $V$ satisfy the conditions* (1.5), (1.6) *with some $k$ and* (1.8). *Let $\lambda > 0$ be sufficiently small and $t = N^{-1+\lambda}$. Define a cutoff initial density as*

$$v(x) := e^{-V_c(x)}, \qquad V_c(x) := V(x)\theta((x - c_N)N^{-\lambda/4k}) + d_N,$$

*where $\theta$ is a smooth cutoff function satisfying $\theta(x) = 1$ for $|x| \le 1$ and $\theta(x) = 0$ for $|x| \ge 2$ and $c_N$ and $d_N$ are chosen such that $v(x)\mathrm{d}\mu(x)$ is a probability density with zero expectation. Denote $\mathcal{L} = L^{\otimes n}$, $F = u^{\otimes n}$ and $F_c = v^{\otimes n}$ with $n = N^2$.*

*(i) We have*

$$\int |F_c - F| \, \mathrm{d}\mu^{\otimes n} \le C \, e^{-cN^c}. \tag{2.1}$$

*with some $c > 0$ depending on $k$ and $\lambda$.*

*(ii) $g_t := (1 - tL + \frac{1}{2}t^2L^2)v$ is a probability measure with respect to $\mathrm{d}\mu$ and for $G_t := [g_t]^{\otimes n}$ we have*

$$\int \frac{\left|e^{t\mathcal{L}}G_t - F_c\right|^2}{e^{t\mathcal{L}}G_t} \mathrm{d}\mu^{\otimes n} \le CN^2 t^{6-\lambda} \le CN^{-4+8\lambda}, \tag{2.2}$$

*where $C$ depends on $\lambda$ and on the constants in* (1.5), (1.6).

In the formulation of this proposition we have not taken into account that in our application the diagonal elements of the matrix evolve under a differently normalized OU process with generator $\widetilde{L} = \frac{1}{2}\partial_x^2 - \frac{x}{2}\partial_x$ with invariant measure $e^{-x^2/2}\mathrm{d}x$. This modification is only notational and does not affect the validity of the estimates (2.1) and (2.2).

*Proof.* From condition (1.6) the estimate (2.1) follows directly by noting that the constants $c_N$ and $d_N$ are subexponentially small in $N$. For the proof of (2.2), we first control the evolution of each matrix element under the OU process (1.3). We assume that for the initial density $v$

$$Lv(x) \le A_1 v(x), \qquad L^2 v(x) \ge -A_2 v(x), \qquad |L^3 v(x)| \le A_3 v(x) \tag{2.3}$$

hold with some constants positive $A_1, A_2$ and $A_3$. Set $g_t = (1 - tL + \frac{1}{2}t^2L^2)v$ for some $t > 0$ and note that $g_t$ is a probability density with respect to $\mu$ if

$$tA_1 + \frac{t^2}{2}A_2 \le 1. \tag{2.4}$$

Define

$$v_t = e^{tL}g_t = e^{tL}\Big(1 - tL + \frac{1}{2}t^2L^2\Big)v,$$

then

$$\partial_t v_t = \frac{1}{2}t^2 L^3 e^{tL} v.$$

Note that by the monotonicity preserving property of the Ornstein-Uhlenbeck kernel and by (2.3), we have

$$e^{sL}L^3 v \le A_3 e^{sL} v \le A_3 e^{sA_1} v, \qquad s \ge 0. \tag{2.5}$$

Here we used the fact that $e^{sL}v \le e^{sA_1}v$ under the first condition in (2.3), which follows from integrating the inequality

$$\frac{\mathrm{d}}{\mathrm{d}s} e^{sL} v = e^{sL}Lv \le A_1 e^{sL} v.$$

In particular

$$v_t = v + \frac{1}{2}\int_0^t s^2 L^3 e^{sL} v \, \mathrm{d}s \ge v\Big(1 - \frac{1}{6}t^3 A_3 e^{tA_1}\Big) \ge \frac{1}{2}v, \tag{2.6}$$

assuming (2.4) and

$$t^3 A_3 \le 1. \tag{2.7}$$

Then

$$
\begin{aligned}
\int \frac{(v-v_t)^2}{v_t}\,\mathrm{d}\mu &= \int v_t^{-1}\Big[\int_0^t \mathrm{d}s\,\frac{1}{2}s^2L^3e^{sL}v\Big]^2\mathrm{d}\mu \\
&\le \frac{t^5}{20}\int_0^t\int v_t^{-1}\big[e^{sL}L^3v\big]^2\mathrm{d}\mu\mathrm{d}s \\
&\le \frac{t^5}{10}\int_0^t\int v^{-1}\big[L^3e^{sL}v\big]^2\mathrm{d}\mu\mathrm{d}s \\
&\le \frac{1}{10}A_3^2t^6e^{2tA_1} \le e^{CA_3^2t^6}-1,
\end{aligned}
\tag{2.8}
$$

where we used (2.6), (2.5) and finally (2.4).

Now we consider the evolution of the product density $F_c = v^{\otimes n}$, note that $\int F_c\,\mathrm{d}\mu^{\otimes n} = 1$. Applying the same procedure to each variable, we have

$$
\int \frac{(e^{t\mathcal{L}}G_t - F_c)^2}{e^{t\mathcal{L}}G_t}\,\mathrm{d}\mu^{\otimes n} \le e^{CA_3^2t^6n}-1 \le CA_3^2t^6n \tag{2.9}
$$

as long as $A_3^2t^6n$ is bounded. In our application $n = N^2$, thus (2.9) will imply (2.2) provided that

$$
A_3 \le Ct^{-\lambda/2} \tag{2.10}
$$

which will also guarantee (2.7). It is straightforward to check that the density $v(x)$ satisfies (2.3) with constants $A_j$ subject to (2.4) and (2.10). This completes the proof. □

# 3 Sine kernel for the time evolved measure

We use the contour integral representation for the correlation functions of the eigenvalues of a matrix of the form $H = \widehat{H} + aV$, where $V$ is a GUE matrix [2, 10]. We will apply this result for the matrix

$$
e^{t\mathcal{L}}G_t = e^{-t/2}\big[G_t + (e^t-1)^{1/2}V\big] \tag{3.1}
$$

where, apart from a trivial prefactor $e^{-t/2}$, $G_t$ plays the role of $\widehat{H}$ and $a = (e^t-1)^{1/2} \approx t^{1/2}$. In order to be able to use the formula given in Proposition 1.1 of [10] to analyze $H = \widehat{H} + aV$, we rescale the variance of $\mathrm{d}\nu$ from $\frac{1}{2}$ to $\frac{1}{8} + \frac{1}{2}a^2$ which changes the semicircle law for $H = \widehat{H} + aV$ to

$$
\varrho(u) := \frac{2}{\pi(1+4a^2)}\sqrt{(1+4a^2-u^2)_+}. \tag{3.2}
$$

In particular, the support changes from $[-2,2]$ to $[-\sqrt{1+4a^2}, \sqrt{1+4a^2}]$. Since eventually $a$ will go to zero, the condition $|u| < 2$ in Theorem (1.1) to be away from the spectral edge changes to the condition $|u| < 1$ which we assume in the sequel. The semicircle law for $\widehat{H}$ will also change from the one given in (1.10) to

$$
\varrho_{sc}(v) := \frac{2}{\pi}\sqrt{(1-v^2)_+}. \tag{3.3}
$$

In the rest of this Section we will use (3.3). The main result of this section is

**Proposition 3.1** *Let $\widetilde{p}_N^{(m)}$ be the m-point eigenvalue correlation function for the ensemble $\widehat{H}+aV$ defined above and let $O:\mathbb{R}^m\to\mathbb{R}$ be a compactly supported bounded observable function. Then for any $|u|<1$ and $a:=N^{-1/2+\lambda/2}$ we have*

$$\begin{aligned}\lim_{N\to\infty}\int_{\mathbb{R}^m} O(\alpha_1,\ldots,\alpha_m)\frac{1}{[\varrho(u)]^m}\widetilde{p}_N^{(m)}\Big(u+\frac{\alpha_1}{N\varrho(u)},\ldots,u+\frac{\alpha_m}{N\varrho(u)}\Big)\mathrm{d}\alpha_1\ldots\mathrm{d}\alpha_m\\ =\int_{\mathbb{R}^m} O(\alpha_1,\ldots,\alpha_m)\det\Big(\frac{\sin\pi(\alpha_i-\alpha_j)}{\pi(\alpha_i-\alpha_j)}\Big)_{i,j=1}^m\mathrm{d}\alpha_1\ldots\mathrm{d}\alpha_m.\end{aligned} \tag{3.4}$$

*Proof.* Using Proposition 1.1 of [10], the (symmetrized) distribution of the eigenvalues $x=(x_1,\ldots,x_N)$ of $H=\widehat{H}+aV$ for any fixed $\widehat{H}$ is given by

$$q_S(x,y):=\frac{1}{(2\pi S)^{N/2}}\frac{\Delta_N(x)}{\Delta_N(y)}\det\big(e^{-(x_j-y_k)^2/2S}\big)_{j,k=1}^N, \tag{3.5}$$

where $y=(y_1,\ldots y_N)$ is the eigenvalues of the Wigner matrix $\widehat{H}$ with the choice of $S=a^2/N$. Note that

$$\begin{aligned}\int_{\mathbb{R}^m} O(\alpha_1,\ldots,\alpha_m)\frac{1}{[\varrho(u)]^m}\widetilde{p}_N^{(m)}\Big(u+\frac{\alpha_1}{N\varrho(u)},\ldots,u+\frac{\alpha_m}{N\varrho(u)}\Big)\mathrm{d}\alpha_1\ldots\mathrm{d}\alpha_m\\ =\widehat{\mathbb{E}}\int_{\mathbb{R}^N}\sum_{i_1,i_2,\ldots,i_m=1}^N O\Big(N\varrho(u)(x_{i_1}-u),\ldots,N\varrho(u)(x_{i_m}-u)\Big)q_S(x,y)\mathrm{d}x_1\ldots\mathrm{d}x_N,\end{aligned} \tag{3.6}$$

where $\widehat{\mathbb{E}}$ denotes the expectation is w.r.t. the $\widehat{H}$ ensemble. Since $O$ is bounded and the sum contains $N^m$ terms, we thus need to compute the limit of the correlation functions of $q_S(x,y)$ in the $x=(x_1,\ldots,x_N)$ variables for a large set $\mathcal{Y}_N\subset\mathbb{R}^N$ of fixed $y=(y_1,\ldots,y_N)$ so that probability of the complement set $\mathcal{Y}_N^c$ satisfies

$$\widehat{\mathbb{P}}(\mathcal{Y}_N^c)=o(N^{-m}).$$

We will choose $\mathcal{Y}_N$ to be the event that the eigenvalues $y(\widehat{H})=(y_1(\widehat{H}),\ldots,y_N(\widehat{H}))$ of the Wigner matrix $\widehat{H}$ follow the semicircle law (3.3). The limit of the correlation functions of $q_S(x,y)$ will be computed starting from the next section in Proposition 3.2.

More precisely, let

$$\eta:=\eta_0 t\sqrt{1-u^2} \tag{3.7}$$

with some sufficiently small $\eta_0<1$ and we set

$$\mathcal{Y}_N:=\Big\{\widehat{H}\ :\ \sup_{Imz\geq\eta}\Big|\frac{1}{N}\sum_j\frac{1}{z-y_j(\widehat{H})}-\int\frac{\varrho_{sc}(r)\mathrm{d}r}{z-r}\Big|\leq N^{-\lambda/4}\ \text{ and }\ \sup_j|y_j(\widehat{H})|\leq K\Big\} \tag{3.8}$$

for some large constant $K$.

By Theorem 3.1 of [8] we then have

$$\widehat{\mathbb{P}}(\mathcal{Y}_N^c)\leq Ce^{-cN^{\lambda/4}} \tag{3.9}$$

(after taking the supremum over all energies, which can be controlled taking energies on a grid of spacing $\eta$). Note that the variance of the matrix elements in [8] was different (see remark at the beginning of Section 3.1) but this does not change the estimates. The condition **C1)** of [8] on the Gaussian decay for the initial density $g_t\mu=(1-tL+\frac{1}{2}t^2L^2)v\mu$ is clearly satisfied by (2.3) and (1.6). Combining the estimate (3.9) with Proposition 3.2 and with the argument after (3.6), we have proved Proposition 3.1. □

## 3.1 Contour integral representation of the correlation function

We compute the correlation functions of $q_S(x;y)$ in $x$, for any fixed $y \in \mathcal{Y}_N$:

$$\widetilde{p}^{(m)}_{N,y,S}(x_1,\dots,x_m) = \int_{\mathbb{R}^{N-m}} q_S(x_1,\dots,x_N;y)\mathrm{d}x_{m+1}\dots\mathrm{d}x_N. \tag{3.10}$$

Note that this definition of the correlation functions differs from the definition of $R^N_m$ given in [10]; the relation being

$$R^N_m(x_1,\dots,x_m;y) = \frac{N!}{(N-m)!}\widetilde{p}^{(m)}_{N,y,S}(x_1,\dots,x_m).$$

By Proposition 2.3 of [10]

$$R^N_m(x_1,\dots,x_m;y) = \det\big(K^S_N(x_i,x_j;y)\big)^m_{i,j=1}, \tag{3.11}$$

where

$$\begin{aligned} K^S_N(u,v;y) =& \frac{e^{(v^2-u^2)/2S}}{(2\pi i)^2(v-u)S}\int_\gamma \mathrm{d}z\int_\Gamma \mathrm{d}w(1-e^{(v-u)z/S})\prod_{j=1}^N\frac{w-y_j}{z-y_j} \\ &\times\frac{1}{z}\Big(w+z-v-S\sum_j\frac{y_j}{(w-y_j)(z-y_j)}\Big)e^{(w^2-2vw-z^2+2uz)/2S}, \end{aligned} \tag{3.12}$$

where $\gamma=\gamma_+\cup\gamma_-$ is the union of two lines $\gamma_+ : s\to -s+i\omega$ and $\gamma_- : s\to s-i\omega$ ($s\in\mathbb{R}$) for any fixed $\omega>0$ and $\Gamma$ is $s\to is$, $s\in\mathbb{R}$. We note that $\Gamma$ can be shifted to any vertical line and since the integrand is an entire function in $w$ and has a Gaussian decay as $|Im\ w|\to\infty$. The constant $\omega>0$ (appearing in the definition of the contour $\gamma$ in $K_N$) can be arbitrary and will be specified later. We fix $|u|<1$.

Set

$$\mathcal{K}_N(u,v;y) = e^{\frac{N(u^2-v^2)}{2a^2}}K^S_N(u,v;y) \tag{3.13}$$

with $S=a^2/N$, then the correlation functions $R^N_m$ defined in (3.11) do not change if $K_N$ is replaced with $\mathcal{K}_N$.

**Proposition 3.2** *Let $\kappa>0$. For any sequence $y=y^{(N)}\in\mathcal{Y}_N$ with the choice $S=N^{-2+\lambda}$ we have*

$$\lim_{N\to\infty}\frac{1}{N\varrho(u)}\mathcal{K}^S_N\Big(u+\frac{\alpha}{N\varrho(u)},u+\frac{\beta}{N\varrho(u)};y\Big) = \frac{\sin\pi(\alpha-\beta)}{\pi(\alpha-\beta)} \tag{3.14}$$

*uniformly for $|u|\le 1-\kappa$ and for $\alpha,\beta$ in a compact set. Moreover, the correlation functions satisfy*

$$\lim_{N\to\infty}\frac{1}{[\varrho(u)]^m}\widetilde{p}^{(m)}_{N,y,S}\Big(u+\frac{\alpha_1}{N\varrho(u)},\dots,u+\frac{\alpha_m}{N\varrho(u)}\Big) = \det\Big(\frac{\sin\pi(\alpha_i-\alpha_j)}{\pi(\alpha_i-\alpha_j)}\Big)^m_{i,j=1}, \tag{3.15}$$

*uniformly for $|u|\le 1-\kappa$ and for $\alpha_1,\dots,\alpha_m$ in a compact set.*

*Proof.* The statement in (3.15) follows directly from (3.14), (3.11) and (3.13), so it is sufficient to prove (3.14). We will prove (3.14) in the form

$$\frac{1}{N\varrho(u)}\mathcal{K}^S_N\Big(u_N,u_N+\frac{\tau}{N\varrho(u)};y\Big)\to\frac{\sin\pi\tau}{\pi\tau}$$

for any sequence $u_N \to u$.

Set

$$\varrho = \varrho(u), \quad t = a^2 = N^{-1+\lambda}. \tag{3.16}$$

We have the representation (see Eq. (3.14) of [10] without the irrelevant prefactor $e^{\omega(u-v)}$)

$$\frac{1}{N\varrho}\mathcal{K}_N\Big(u_N, u_N + \frac{\tau}{N\varrho}; y\Big) = N\int_\gamma \frac{\mathrm{d}z}{2\pi i}\int_\Gamma \frac{\mathrm{d}w}{2\pi i} h_N(z,w)g_N(z,w)e^{N(f_N(w)-f_N(z))} \tag{3.17}$$

with

$$f_N(z) = \frac{1}{2t}(z^2 - 2u_N z) + \frac{1}{N}\sum_j \log(z - y_j) \tag{3.18}$$

$$g_N(z,w) = \frac{1}{tz}[w + z - u_N] - \frac{1}{Nz}\sum_j \frac{y_j}{(w-y_j)(z-y_j)} \tag{3.19}$$

$$h_N(z,w) = \frac{1}{\tau}\Big(e^{-\tau w/t\varrho} - e^{-\tau(w-z)/t\varrho}\Big) \tag{3.20}$$

with (3.16). We will need the identity (we remark that there was a typo in (2.21) of [10])

$$g_N(z,w) = \frac{1}{z}f_N'(w) + \frac{f_N'(z) - f_N'(w)}{z - w}. \tag{3.21}$$

## 3.2 Saddle points

For brevity, we will denote $u_N$ by $u$ in the sequel. We first determine the critical points of $f_N$, i.e. we solve

$$f_N'(z) = t^{-1}(z - u) + \frac{1}{N}\sum_j \frac{1}{z - y_j} = 0. \tag{3.22}$$

This is equivalent to finding the zeros of a polynomial of degree $N+1$. There are $N-1$ real roots and two complex roots, called $q_N^\pm$, that are complex conjugates of each other

$$f_N'(q_N^\pm) = 0.$$

We will work with $q_N := q_N^+$, the analysis of the other saddle is analogous. Clearly $|Re\ q_N| \le K$ for some large $K$.

We can define

$$f(z) = \frac{1}{2t}(z^2 - 2uz) + \int_{\mathbb{R}} \varrho_{sc}(y)\log(z-y)\mathrm{d}y$$

and instead of (3.22), we can solve

$$f'(z) = t^{-1}(z-u) + 2(z - \sqrt{z^2-1}) = 0. \tag{3.23}$$

The solutions of this latter equation (for small $t$) are given by

$$q^\pm = \frac{(2t+1)u \pm 2ti\sqrt{1+4t-u^2}}{1+4t} = u(1-2t) \pm 2ti\sqrt{1-u^2} + O(t^2), \tag{3.24}$$

and thus in particular

$$Im(q^\pm) = \pm O(t).$$

We have

$$f''(q) = \frac{1}{t} + 2 - \frac{2q}{\sqrt{q^2-1}}$$

and

$$f''(q^{\pm}) = \frac{1}{t} + 2 \pm \frac{2ui}{\sqrt{1-u^2}} + O(t)$$

where we also used the equation (3.23) for $q^{\pm}$. We set $q = q^+$.

We need to know that $f_N'' \neq 0$ at the $q_N$ saddle.

$$f_N''(q_N) = \frac{1}{t} - \frac{1}{N}\sum_j \frac{1}{(q_N - y_j)^2}.$$

It follows from (3.8) that for $y \in \mathcal{Y}$ we have

$$\sup_{Imz\geq\eta} |f_N^{(\ell)}(z) - f^{(\ell)}(z)| \leq \frac{C}{t^{\ell-1}N^{\lambda/4}} \tag{3.25}$$

by contour integration.

We compare $q$ and $q_N$. We have from (3.22)

$$q_N = F_N(q_N) := u - \frac{t}{N}\sum_j \frac{1}{q_N - y_j}, \qquad Im\ q_N > 0 \tag{3.26}$$

and

$$q = F(q) := u - t\int \frac{\varrho_{sc}(y)\mathrm{d}y}{q-y} = u - 2t(q - \sqrt{q^2-1}) \tag{3.27}$$

First we show that for the only solution to (3.26) with positive imaginary part we have $Im\ q_N \geq \eta$. This is a fixed point argument.

Define the compact set

$$\Xi := \Big\{ z \ : \ |Re\ z - u| \leq Ct,\ \eta \leq Im\ z \leq Ct \Big\}$$

for some large constant $C$. Since $y \in \mathcal{Y}$, we know that

$$\sup_{Im\ z\geq\eta} |F_N(z) - F(z)| \leq \frac{Ct}{N^{\lambda/4}}.$$

For $z \in \Xi$ clearly

$$Re\ F(z) = u + O(t)$$

and

$$Im\ F(z) = 2t\sqrt{1-u^2} + O(t^2)$$

thus

$$Re\ F_N(z) = u + O(t), \qquad Im\ F_N(z) = 2t\sqrt{1-u^2} + o(t)$$

so $F_N(\Xi) \subset \Xi$.

Now we compute, for $z \in \Xi$,

$$F_N'(z) = \frac{t}{N}\sum_j \frac{1}{(z-y_j)^2} = F'(z) + O(N^{-\lambda/4})$$

(here we used (3.25) with $\ell = 2$ and observed that $F_N' = tf_N''$), and

$$F'(z) = -2t\Big[1 - \frac{z}{\sqrt{z^2-1}}\Big]$$

with $F'(z) = O(t)$ if $z \in \Xi$. Thus $|F_N'(z)| \le 1/2$ for $z \in \Xi$, so $F_N$ is a contraction on $\Xi$ and thus (3.26) has a unique solution, which is $q_N$.

Comparing the two solutions, we have

$$|q_N - q| = |F_N(q_N) - F(q)| \le \sup_{z\in\Xi} |F_N'(z)||q_N - q| + |F_N(q) - F(q)|.$$

Since $y \in \mathcal{Y}$, we get

$$|F_N(q) - F(q)| \le t\left|\frac{1}{N}\sum \frac{1}{q_N - y_j} - \int \frac{\varrho_{sc}(y)\mathrm{d}y}{z-y}\right| = \frac{Ct}{N^{\lambda/4}}$$

thus

$$|q_N - q| \le \frac{Ct}{N^{\lambda/4}}. \tag{3.28}$$

## 3.3 Evaluating the integrals

Using Laplace asymptotics, we compute the integrals in (3.17). We choose the horizontal curves $\gamma_\pm$ to pass through the two saddles $q^\pm = a \pm ib$ of $f$ (see (3.24)), i.e. we set $\omega = b$ (see the definition of $\gamma^\pm$ after (3.12)). The vertical line $\Gamma$ is shifted to pass through the saddles, i.e. $Re\,\Gamma = a$. Moreover, if necessary, we deform $\Gamma$ in a $O(N^{-1})$-neighborhood of $a$ so that $\min_j \mathrm{dist}(\Gamma, y_j) \ge N^{-2}$; this is always possible.

We split the integrals as follows

$$\frac{1}{N\varrho}\mathcal{K}_N(u, u + \frac{\tau}{N\varrho}; y) = A^{++} + A^{+-} + A^{-+} + A^{--}$$

according to whether $Im\, z$ and $Im\, w$ are positive or negative, e.g.

$$A^{\pm\pm} := N\int_{\gamma^\pm}\frac{\mathrm{d}z}{2\pi i}\int_{\Gamma^\pm}\frac{\mathrm{d}w}{2\pi i} h_N(z,w)g_N(z,w)e^{N(f_N(w)-f_N(z))} \tag{3.29}$$

where $\Gamma_+ = \Gamma \cap \{w \,:\, Im\, w \ge 0\}$ and $\Gamma_- = \Gamma \cap \{w \,:\, Im\, w \le 0\}$. We will work on $A^{++}$, the other three integrals are treated similarly.

The main contribution to the integral $A^{++}$ will come from an $\varepsilon$-neighborhood in $z$ and $w$ of the saddle point $q_N = q_N^+$. The radius $\varepsilon$ will be chosen such that after a local change of variable $f$ and $f_N$ become quadratic near the saddle. We now explain the local change of variable.

Since $f(z) : \mathbb{C} \to \mathbb{C}$ is an analytic function with $f'(q) = 0$ and $f''(q) \ne 0$ for $q = q^+$, there exists an invertible analytic map $\phi : z \to \phi(z)$ in

$$D_\varepsilon := \{z \,:\, |z - q| \le \varepsilon\}$$

with $\phi(q) = 0$, $\phi'(q) = \sqrt{tf''(q)}$ such that

$$f(z) = f(q) + \frac{1}{2t}[\phi(z)]^2 \qquad z \in D_\varepsilon \tag{3.30}$$

with

$$\phi(z) = \sqrt{tf''(q)}(z-q)(1+O(z-q)), \qquad z \in D_\varepsilon. \tag{3.31}$$

Here $\varepsilon$ must satisfy

$$\varepsilon \le \frac{|f''(q)|}{2\sup_{D_\varepsilon}|f'''(z)|} \tag{3.32}$$

we also assume that $\varepsilon \le \eta$. We will choose $\varepsilon = ct$ with a small $c$, depending on $u$. We have

$$f''(q) = t^{-1} + O(1), \qquad \sup_{D_\varepsilon}|f'''(z)| \le C \tag{3.33}$$

from the explicit formula (3.23), so (3.32) is satisfied. Note that $\phi'(q) = \sqrt{tf''(q)} = 1 + O(t)$.

We have a similar change of variables for $f_N$, i.e. $\phi_N$ with the properties that

$$\phi_N(q_N) = 0, \qquad \phi_N'(q_N) = \sqrt{tf_N''(q_N)} = 1 + O(t) \tag{3.34}$$

and

$$f_N(z) = f_N(q_N) + \frac{1}{2t}[\phi_N(z)]^2 \qquad z \in D_{\varepsilon,N} = \{z \,:\, |z-q_N| \le \varepsilon\} \tag{3.35}$$

with

$$\phi_N(z) = \sqrt{tf_N''(q_N)}(z-q_N)(1+O(z-q_N)), \qquad z \in D_{\varepsilon,N}. \tag{3.36}$$

This holds if

$$\varepsilon \le \frac{c|f_N''(q_N)|}{\sup_{D_{\varepsilon,N}}|f_N'''(z)|}.$$

For $y \in \mathcal{Y}$, we have $f_N''(q_N) = t^{-1}\left[1 + O(N^{-\lambda/4})\right]$ and $|f_N'''(z)| \le Ct^{-2}N^{-\lambda/4}$ by (3.25) and (3.33), thus we can choose $\varepsilon = ct$ for some small constant $c \le \sqrt{1-u^2}$.

Moreover we have $|\phi_N(z)| \le C|z-q_N|$ for $|z-q| \le ct$, so by Cauchy formula $|\phi_N'(z)| \le C$ and $|\phi_N''(z)| \le Ct^{-1}$ for $|z-q| \le ct$ (maybe after reducing $c$). The same formulas hold for $\phi$ as well. We also have

$$|\phi'(q) - \phi_N'(q_N)| \le \left|\sqrt{tf''(q)} - \sqrt{tf_N''(q)}\right| + \left|\sqrt{tf_N''(q)} - \sqrt{tf_N''(q_N)}\right| \le CN^{-\lambda/4},$$

where in the first term we used (3.25) and in the second we used $|f_N'''(z)| \le Ct^{-2}$.

From (3.31) and (3.36) we have

$$|\phi(z) - \phi_N(z)| \le |\phi'(q) - \phi'(q_N)||z-q| + |\phi'(q_N)||q-q_N| + C|z-q|^2 \le CtN^{-\lambda/4} \tag{3.37}$$

and then by contour integration

$$|\phi'(z) - \phi_N'(z)| \le CtN^{-\lambda/4} \tag{3.38}$$

for any $z$ with $|z-q| \le ct$. Therefore the maps $\phi$ and $\phi_N$ are $C^1$-close within $D_\varepsilon$ and both of them are $C^1$-close to the shift map $z \to z - q$.

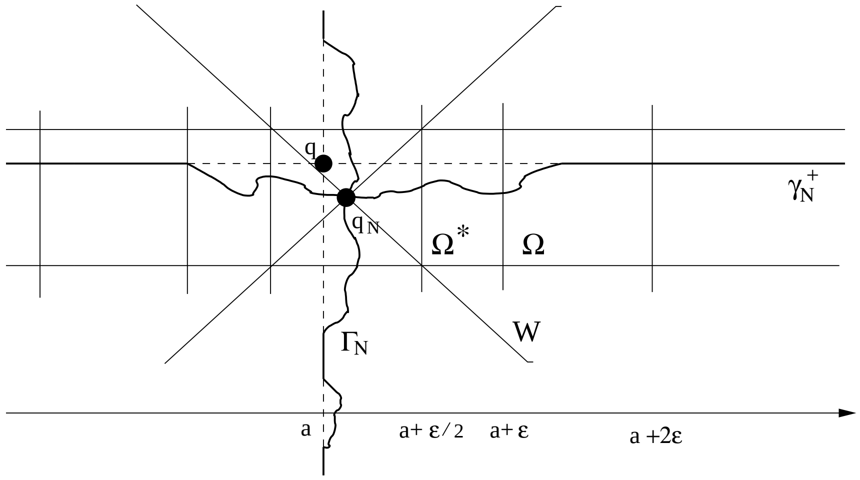


Figure 1: Integration contours around the saddle $q_N = q_N^+$

We first consider the $z$ integration. Recall that $q_N = q_N^+ = a_N + ib_N$ from (3.24). We fix a small positive constant $c_1 \ll 1$ and we define the domains

$$\Omega := \Big\{ z = x + iy \ : \ |x - a_N| \geq \varepsilon, |y - b_N| \leq c_1\varepsilon/2 \Big\}$$

$$\Omega^* := \Big\{ z = x + iy \ : \ |x - a_N| \geq \varepsilon/2, |y - b_N| \leq c_1\varepsilon/2 \Big\}$$

and

$$W := \Big\{ z = x + iy \ : \ |x - a_N| \leq 2\varepsilon \ , \ |y - b_N| \leq c_1|x - a_N| \Big\}$$

where $\varepsilon = ct$. Recall that $\gamma^+$ was the horizontal line going through $q = a + ib$, the saddle of $f$. We will deform $\gamma^+$ to $\gamma_N^+$ so that it passes through $q_N$ and it matches with $\gamma^+$ at the points $a_N \pm 2\varepsilon + ib$. Within the regime $|Re\ z - a_N| \leq \varepsilon$, we define $\gamma_N^+$ by the requirement that $Im\ \phi_N = 0$ along $\gamma_N^+$. Since $\phi_N(z)$ is close to the map $z \to z - q_N$ by (3.36), clearly $\gamma_N^+$ is almost horizontal curve in small neighborhood of $q_N$, so it remains in $W$ until it reaches the vertical lines $|Re\ z - a_N| = \varepsilon$. In the regime $\varepsilon \leq |Re\ z - a_N| \leq 2\varepsilon$, we require that $\gamma_N^+$ matches with $\gamma^+$ at the points $a_N \pm 2\varepsilon + ib$ and it remains in the wedge $W$. In the outside regime, $|Re\ z - a_N| \geq 2\varepsilon$ we set $\gamma_N^+ = \gamma^+$, in particular $\gamma_N^+ \subset W \cup \Omega$ (see Fig. 1).

**Lemma 3.1** *We have*

$$Re\big[f_N(x + iy) - f_N(q_N)\big] \geq \frac{1}{12t}(x - a)^2 \qquad \text{for } x + iy \in \Omega \tag{3.39}$$

*and*

$$Re\big[f_N(z) - f_N(q_N)\big] \geq 0 \qquad \text{for} \quad z \in W \text{ and } \quad |Re\ z - a| \leq \varepsilon \tag{3.40}$$

*Proof.* The second statement (3.40) follows from the normal form (3.35) and the fact that for $z \in W$ we have $|Im\ (z - q_N)| \leq c_1|Re(z - q_N)|$, i.e. $Re(z - q_N)^2 \geq 0$, and $\phi_N$ is close to the map $z \to z - q_N$ in $W$, so $Re[\phi_N(z)]^2 \geq 0$ for $z \in W$.

For the first statement, we assume $x \geq a$, the case $x \leq a$ is analogous. We get by explicit calculation

$$Re\, f'(x+iy) \geq \frac{1}{2t}(x-a), \quad \text{for } x+iy \in \Omega^*,\ x \geq a.$$

Using (3.25) for $\ell = 1$, we have

$$Re\, \partial_x f_N(x+iy) \geq Re\, f'(x+iy) - CN^{-\lambda/4} \geq \frac{1}{3t}(x-a) \quad \text{for } x+iy \in \Omega^*\ x \geq a \tag{3.41}$$

(the error is absorbed since $|x-a| \geq ct/2$ for $x+iy \in \Omega^*$). Since $Re[f_N(z) - f_N(q_N)] \geq 0$ on the vertical lines $|x-a| = \varepsilon/2$, $|y-b| \leq c_1\varepsilon/2$, we can integrate the inequality (3.41) to obtain (3.39).

□

In order to estimate the $w$ integration along $\Gamma^+$ parametrized as $a+is$, $s \geq 0$, we analyze the behaviour of $Re\, f$ along $\Gamma^+$. For $|x-a| \leq Ct$ and $y \in \mathbb{R}$ we first compute

$$Re\, \partial_y f_N(x+iy) = -Im\, f_N'(x+iy) = -Im\, f'(x+iy) + O(N^{-\lambda/4})$$

which holds for $|y| \geq \eta$. By explicit computation, and using $f'(a+ib) = 0$,

$$-Im\, f'(x+iy) = -(y-b)\Big(\frac{1}{t} + 2\Big) + O(t) + O(y^2)$$

if $|y| \leq \frac{1}{2}\sqrt{1-u^2}$, $|x-a| \leq Ct$ for some large $C$. Thus we have

$$Re\, \partial_y f_N(x+iy) \leq -\frac{y-b}{3t}, \qquad \eta \leq |y| \leq \frac{1}{2}\sqrt{1-u^2} \text{ and } y-b \geq \varepsilon/2,$$

where $\varepsilon = ct$ with a small $c$ as before and a similar lower bound holds for $y-b \leq -\varepsilon/2$. Defining

$$\widetilde{\Omega} := \Big\{ w = x+iy \ :\ \varepsilon \leq |y-b_N|,\ \eta \leq y \leq \frac{1}{2}\sqrt{1-u^2}, \quad |x-a_N| \leq c_1\varepsilon/2 \Big\}$$

$$\widetilde{W} := \Big\{ w = x+iy \ :\ |y-b_N| \leq 2\varepsilon,\ |x-a_N| \leq c_1|y-b_N| \Big\}$$

analogously to $W$ before, we easily obtain

$$Re\, \big[f_N(x+iy) - f_N(q_N)\big] \leq -\frac{1}{18t}(y-b)^2 \quad \text{for } x+iy \in \widetilde{\Omega} \tag{3.42}$$

and

$$Re\, \big[f_N(w) - f_N(q_N)\big] \leq 0 \ \ w \in \widetilde{W}, \text{ and } |Im\, w - b| \leq \varepsilon, \tag{3.43}$$

similarly to the proof of Lemma 3.1.

The regimes $0 \leq y \leq \eta$ and $y \geq \frac{1}{2}\sqrt{1-u^2}$ are treated directly. We use

$$\begin{aligned} Re\partial_y f_N(x+iy) &= -\, Im\left[ t^{-1}(z-u) + \frac{1}{N}\sum_j \frac{1}{z-y_j}\right] \\ &= y\left[ -t^{-1} + \frac{1}{N}\sum_j \frac{1}{(x-y_j)^2 + y^2}\right] \geq -y/t. \end{aligned} \tag{3.44}$$

Hence for $0 \le y \le \eta$ we have

$$Re\,\big[f_N(x+iy) - f_N(q_N)\big] \le Re\,\big[f_N(x+i\eta) - f_N(q_N)\big] + \frac{\eta^2}{2t} \le -\frac{1}{36t}(y-b)^2$$

from (3.42), if $\eta_0$ is sufficiently small, see (3.7).

If $y \ge \frac{1}{2}\sqrt{1-u^2}$, then

$$\frac{1}{N}\sum_j \frac{1}{(x-y_j)^2+y^2} \le \frac{4}{\sqrt{1-u^2}},$$

hence

$$Re\,\partial_y f_N(x+iy) \le -y/2t \tag{3.45}$$

and thus $Re\, f_N(x+iy) \le -y^2/4t$ in this regime. Summarizing these results, we have

$$Re\,\big[f_N(x+iy) - f_N(q_N)\big] \le -\frac{1}{36t}(y-b)^2 \tag{3.46}$$

holds for any $y \in \mathbb{R}$ and $|x-a| \le c_1\varepsilon/2$.

We can define a new contour $\Gamma_N^+$ similar to the $\gamma_N^+$. It follows the path where $\phi_N$ has zero imaginary part when $|Im\ w - b| \le \varepsilon/2$ and then it returns to $\Gamma^+$ when $|Im\ w - b| \ge \varepsilon$. We recall that $\min_j \mathrm{dist}(\Gamma_N^+, y_j) \ge N^{-2}$ by the choice of $\Gamma$.

With the paths $\gamma_N^+$ and $\Gamma_N^+$ defined, we can now do the integration

$$A^{++} := N\int_{\gamma_N^+} \frac{\mathrm{d}z}{2\pi i}\int_{\Gamma_N^+}\frac{\mathrm{d}w}{2\pi i} h_N(z,w)g_N(z,w)e^{N(f_N(w)-f_N(z))}. \tag{3.47}$$

Near the saddle we need the bounds

$$|g_N(z,w)| \le C/t, \quad |\partial_z g_N(z,w)| \le C/t^2, \quad |h_N(z,w)| \le C, \quad |\partial_z h_N(z,w)| \le Ct^{-1} \tag{3.48}$$

if $|z-(a+ib)| \le \varepsilon$, $|w-(a+ib)| \le \varepsilon$. Away from the saddle we will use

$$|h_N(z,w)| \le Ce^{Ct^{-1}|Re\,z-a|}, \quad |g_N(z,w)| \ \le CN^3 \tag{3.49}$$

that hold for $|Im\ z| \ge \eta$, $Im\ w \ge 0$. These bound follow from (3.19), (3.20) and (3.21) and when $w$ is near the real axis, we also used that $\Gamma_N$ is away from the $y_j$'s.

The integration in $A^{++}$ (see (3.47)) will be divided into regimes near the saddle $q_N$ ("inside") or away from the saddle ("outside"):

$$A^{++} = A_{ii} + A_{io} + A_{oi} + A_{oo}. \tag{3.50}$$

Recall that $|q_N - q| = o(t)$ and $q = q^+ = a+ib$ (see (3.24)). For example

$$A_{io} := N\int_{\gamma_N^+} \chi(Re\ z-a)\frac{\mathrm{d}z}{2\pi i}\int_{\Gamma_N^+}(1-\chi(Im\ w-b))\frac{\mathrm{d}w}{2\pi i}h_N(z,w)g_N(z,w)e^{N(f_N(w)-f_N(z))},$$

where $\chi$ is the characteristic function of the interval $[-\varepsilon,\varepsilon]$. The other $A$'s are defined analogously.

Using (3.39) (3.40) and (3.46), we have

$$|A_{io}| \leq N \int_{\gamma_N} \mathrm{d}z\, \chi(Re\ z - a) \int_{\Gamma_N} (1 - \chi(Im\ w - b))|g_N(z,w)||h_N(z,w)|\ e^{N Re\ [f_N(w) - f_N(q_N)]} \mathrm{d}w.$$

The integral of the exponential term is bounded by

$$\int_{|y-b| \geq \varepsilon = ct} \mathrm{d}y\ e^{-cN(y-b)^2/t} \leq e^{-cNt}.$$

Taking into account (3.48) and (3.49), we see that $|A_{io}| \leq e^{-cNt}$ since $t = N^{-1+\lambda}$. Similarly we can bound all other terms with an outside part. When $|Re\, z - a| \geq ct \gg N^{-1}$, then the exponential growth of $h_N$ in (3.49) will be controlled by the Gaussian decay of

$$e^{-N\ Re[f_N(z) - f_N(q_N)]} \leq e^{-cNt^{-1}|Re\, z - a|^2}$$

from (3.39).

Finally, we have to compute the contribution of the saddle, i.e. the term $A_{ii}$. We let $\widetilde{\gamma}$ be the part of $\gamma_N^+$ with $|Re\ \gamma_N - a| \leq \varepsilon$ and similarly defined $\widetilde{\Gamma}$. Recall that $Im\ \phi_N = 0$ on $\widetilde{\gamma}$. From standard Laplace asymptotics calculation, we have

$$\int_{\widetilde{\gamma}} e^{-N[f_N(z) - f_N(q_N)]} h_N(z,w) g_N(z,w) \mathrm{d}z = \int_{\widetilde{\gamma}} e^{-N[\phi_N(z)]^2/2t} h_N(z,w) g_N(z,w) \mathrm{d}z$$

$$= \sqrt{\frac{2\pi}{N f_N''(q_N)}} \Big[ h_N(q_N, w) g_N(q_N, w) + \Omega(w) \Big] \tag{3.51}$$

using (3.34) with

$$|\Omega(w)| \leq C \sqrt{\frac{t}{N}} \max_{z \in D_\varepsilon} \Big[ |\partial_z h_N(z,w)||g_N(z,w)| + |\partial_z g_N(z,w)||h_N(z,w)| \Big].$$

Using (3.48), we have

$$|\Omega| \leq C t^{-2} \sqrt{\frac{t}{N}} = \frac{C}{t} \frac{1}{\sqrt{Nt}}$$

while the main term in the bracket on the r.h.s. of (3.51) is of order $t^{-1}$. Analogously performing the $\mathrm{d}w$ integration, we obtain that

$$A_{ii} = \frac{-1}{2\pi f_N''(q_N)} g_N(q_N, q_N) h_N(q_N, q_N) \Big[ 1 + O\Big(\frac{1}{\sqrt{Nt}}\Big) \Big] = \frac{-h_N(q_N, q_N)}{2\pi} \Big[ 1 + O\Big(\frac{1}{\sqrt{Nt}}\Big) \Big],$$

where we also used $g_N(q_N, q_N) = f_N''(q_N)$ following from (3.21). So far we considered the saddle $q_N = q_N^+$ with positive imaginary part for both the $z$ and $w$ integrals. The same calculation can be performed at the saddle $z = w = q_N^-$. The mixed case, when $z$ is integrated near one of the saddles and $w$ is near the other one, gives zero contribution, since $g_N(q_N^-, q_N^+) = g_N(q_N^+, q_N^-) = 0$ by (3.21). Adding up the contributions of the two relevant saddles, $z = w = q_N^+$ and $z = w = q_N^-$, taking into account the opposite orientations of the two pieces of $\gamma_N$, one obtains

$$\frac{1}{2\pi} \Big[ - h_N(q_N^+, q_N^+) + h_N(q_N^-, q_N^-) \Big] = \frac{1}{2\pi\tau} \Big( - e^{-\tau q_N^+/t\varrho} + e^{-\tau q_N^-/t\varrho} \Big) = \frac{\sin \pi\tau}{\pi\tau}(1 + o(1)),$$

where we used (3.16), (3.24) and (3.28). This completes the proof of Proposition 3.2. □

# 4 Proof of the main theorems

*Proof of Theorem 1.1.* We follow the notations of Proposition 2.1. In Proposition 3.1 we have shown that the sine kernel holds for the measure $e^{t\mathcal{L}}G_t$ if $t = N^{-1+\lambda}$. More precisely, let $p_{N,t}(x)$, denote the density function of the eigenvalues $x = (x_1, \ldots, x_N)$ w.r.t. $e^{t\mathcal{L}}G_t$ and let $p_{N,t}^{(2)}$ be the two point correlation function, defined analogously to (1.9). Similarly, we define $p_{N,c}(x)$ and $p_{N,c}^{(2)}$ for the eigenvalue density and two point correlation function w.r.t. truncated measure $F_c = v^{\otimes n}$.

In Proposition 3.1 we showed that

$$\lim_{N\to\infty} \int_{\mathbb{R}^2} \frac{1}{\varrho^2} p_{N,t}^{(2)}\Big(u + \frac{\alpha}{N\varrho}, u + \frac{\beta}{N\varrho}\Big) O(\alpha,\beta)\mathrm{d}\alpha\mathrm{d}\beta = \int_{\mathbb{R}^2} O(\alpha,\beta)\Big[1 - \Big(\frac{\sin\pi(\alpha-\beta)}{\pi(\alpha-\beta)}\Big)^2\Big]\mathrm{d}\alpha\mathrm{d}\beta. \quad (4.1)$$

for any $|u| < 2$ and with the notation $\varrho = \varrho_{sc}(u)$. (We remark that $p_{N,t}^{(2)}$ was denoted by $\widetilde{p}_N^{(2)}$ in Proposition 3.1 and the condition $|u| < 2$ is translated into $|u| < 1$ after rescaling.)

To prove (1.11), we thus only need to control the difference as follows

$$\Big| \int \Big[p_N^{(2)}\Big(u + \frac{\alpha}{N\varrho}, u + \frac{\beta}{N\varrho}\Big) - p_{N,t}^{(2)}\Big(u + \frac{\alpha}{N\varrho}, u + \frac{\beta}{N\varrho}\Big)\Big] O(\alpha,\beta)\mathrm{d}\alpha\mathrm{d}\beta\Big| \le (I) + (II),$$

where

$$(I) := \Big| \int \Big[p_N^{(2)}\Big(u + \frac{\alpha}{N\varrho}, u + \frac{\beta}{N\varrho}\Big) - p_{N,c}^{(2)}\Big(u + \frac{\alpha}{N\varrho}, u + \frac{\beta}{N\varrho}\Big)\Big] O(\alpha,\beta)\mathrm{d}\alpha\mathrm{d}\beta\Big|,$$

$$(II) := \int \Big|p_{N,c}^{(2)}\Big(u + \frac{\alpha}{N\varrho}, u + \frac{\beta}{N\varrho}\Big) - p_{N,t}^{(2)}\Big(u + \frac{\alpha}{N\varrho}, u + \frac{\beta}{N\varrho}\Big)\Big| \, |O(\alpha,\beta)|\mathrm{d}\alpha\mathrm{d}\beta.$$

Using (2.1), we have

$$(I) \le N^2 \|O\|_\infty \int |F - F_c| \mathrm{d}\mu^{\otimes n} \le Ce^{-cN^c} \to 0$$

with some $c > 0$ as $N \to \infty$. To estimate $(II)$, we have

$$\begin{aligned}(II) &\le \int \Big| \frac{p_{N,c}^{(2)}}{p_{N,t}^{(2)}}\Big(u + \frac{\alpha}{N\varrho}, u + \frac{\beta}{N\varrho}\Big) - 1 \Big| p_{N,t}^{(2)}\Big(u + \frac{\alpha}{N\varrho}, u + \frac{\beta}{N\varrho}\Big)|O(\alpha,\beta)|\mathrm{d}\alpha\mathrm{d}\beta \\ &\le \Big( \int \Big[\frac{p_{N,c}^{(2)}}{p_{N,t}^{(2)}}\Big(u + \frac{\alpha}{N\varrho}, u + \frac{\beta}{N\varrho}\Big) - 1\Big]^2 p_{N,t}^{(2)}\Big(u + \frac{\alpha}{N\varrho}, u + \frac{\beta}{N\varrho}\Big)|O(\alpha,\beta)|\mathrm{d}\alpha\mathrm{d}\beta\Big]^{1/2} \\ &\quad \times \Big[ \int p_{N,t}^{(2)}\Big(u + \frac{\alpha}{N\varrho}, u + \frac{\beta}{N\varrho}\Big)|O(\alpha,\beta)|\mathrm{d}\alpha\mathrm{d}\beta\Big]^{1/2}.\end{aligned} \quad (4.2)$$

Using (4.1) for the observable $|O|$ instead of $O$, the second factor on the r.h.s. of (4.2) is bounded.

Since $O$ is bounded, the first factor is smaller than

$$
\begin{aligned}
C\left[N^2\varrho^2\int\Big[\frac{p^{(2)}_{N,c}(z,y)}{p^{(2)}_{N,t}(z,y)}-1\Big]^2 p^{(2)}_{N,t}(z,y)\mathrm{d}z\mathrm{d}y\right]^{1/2} &\le C\left[N^2\varrho^2\int\Big(\frac{p_{N,c}(x)}{p_{N,t}(x)}-1\Big)^2 p_{N,t}(x)\mathrm{d}x\right]^{1/2}\\
&\le C\left[N^2\varrho^2\int\frac{\left|e^{t\mathcal{L}}G_t-F_c\right|^2}{e^{t\mathcal{L}}G_t}\mathrm{d}\mu^{\otimes n}\right]^{1/2}\\
&\le CN^{-1+4\lambda}.
\end{aligned}
\tag{4.3}
$$

Here in the first step we used that the quantity $D(f,g)=\int|f/g-1|^2 g$ for two probability measures $f$ and $g$ decreases when taking marginals. In the second step, we used that $D(f,g)$ decreases when passing the probability laws from matrix elements to the induced probability laws for the eigenvalues. Finally, we used the estimate (2.2). This completes the proof of Theorem 1.1. $\square$

*Proof of Theorem 1.2.* We first prove Theorem 1.2 for the ensemble $\widehat{H}+aV$ with $a=N^{-1/2+\lambda/2}$ (see the beginning of Section 3 for the necessary rescaling). Let $\mathbb{E}$ denote the expectation with respect to this ensemble and let $\mathbb{E}_y$ denote the expectation with respect to the density $x\to q_S(x,y)$ for any fixed $y$ and $S=a^2/N=N^{-2+\lambda}$. Then we have

$$
\mathbb{E}\,\Lambda(u;s,\cdot)=\int\mathbb{E}_y\;\Lambda(u;s,\cdot)\mathbf{1}(y\in\mathcal{Y})\mathrm{d}\widehat{\mathbb{P}}(y)+\int\mathbb{E}_y\;\Lambda(u;s,\cdot)\mathbf{1}(y\in\mathcal{Y}^c)\mathrm{d}\widehat{\mathbb{P}}(y)
\tag{4.4}
$$

by recalling (3.5). The second term can be estimated by using $|\Lambda|\le N$ and (3.9) as

$$
\int\mathbb{E}_y\;\Lambda(u;s,\cdot)\mathbf{1}(y\in\mathcal{Y}^c)\mathrm{d}\widehat{\mathbb{P}}(y)\le CNe^{-cN^{\lambda/4}}.
\tag{4.5}
$$

For the first term in (4.4), we use the exclusion-inclusion principle to compute

$$
\begin{aligned}
\mathbb{E}_y\;\Lambda(u;s,\cdot)=\frac{1}{2Nt_N\varrho}\sum_{m=2}^{N}(-1)^m\int_{-t_N}^{t_N}\mathrm{d}v_1\ldots\int_{-t_N}^{t_N}\mathrm{d}v_m\mathbf{1}\Big\{\max|v_i-v_j|\le\frac{s}{N\varrho}\Big\}\\
\times\binom{N}{m}\widetilde{p}^{(m)}_{N,y,S}(u+v_1,u+v_2,\ldots,u+v_m)
\end{aligned}
\tag{4.6}
$$

with $\varrho=\varrho(u)$ (see (3.2)) and recall that $\widetilde{p}^{(m)}_{N,y,S}$ denote the correlation functions of $q_S(x,y)$ (see (3.10)). After a change of variables,

$$
\begin{aligned}
\mathbb{E}_y\;\Lambda(u;s,\cdot)=&\;\frac{1}{2Nt_N\varrho}\sum_{m=2}^{\infty}(-1)^m\int_{-N\varrho t_N}^{N\varrho t_N}\mathrm{d}z_1\ldots\int_{-N\varrho t_N}^{N\varrho t_N}\mathrm{d}z_m\\
&\times\binom{N}{m}\frac{1}{(N\varrho)^m}\widetilde{p}^{(m)}_{N,y,S}\Big(u+\frac{z_1}{N\rho},\ldots,u+\frac{z_m}{N\rho}\Big)\mathbf{1}\Big\{\max|z_i-z_j|\le s\Big\}\\
=&\;\frac{1}{2Nt_N\varrho}\sum_{m=2}^{\infty}(-1)^m m\int_{-N\varrho t_N}^{N\varrho t_N}\mathrm{d}z_1\int_0^s\mathrm{d}a_2\ldots\int_0^s\mathrm{d}a_m\\
&\times\binom{N}{m}\frac{1}{(N\varrho)^m}\,\widetilde{p}^{(m)}_{N,y,S}\Big(u+\frac{z_1}{N\rho},u+\frac{z_1+a_2}{N\rho},\ldots,u+\frac{z_1+a_m}{N\rho}\Big),
\end{aligned}
\tag{4.7}
$$

where the factor $m$ comes from considering the integration sector $z_1 \le z_j$, $j \ge 2$. Taking $N \to \infty$ and using Proposition 3.2, we get

$$\lim_{N\to\infty} \mathbb{E}_y\, \Lambda(u; s, \cdot) = \sum_{m=2}^{\infty} \frac{(-1)^m}{(m-1)!} \int_0^s \mathrm{d}a_2 \ldots \int_0^s \mathrm{d}a_m \, \det\left(\frac{\sin \pi(a_i - a_j)}{\pi(a_i - a_j)}\right)_{i,j=1}^m, \tag{4.8}$$

where in the last determinant term we set $a_1 = 0$. The interchange of the limit and the summation can be justified by noting that the exclusion-inclusion principle guarantees that (4.6) is an alternating series where the difference between the sum and its $M$-term truncation can be controlled by the $(M+1)$-th term for any $M$. We note that the left hand side of (4.8) is $\int_0^s p(\alpha)\mathrm{d}\alpha$, where $p(\alpha)$ is the second derivative of the Fredholm determinant $\det(1 - \mathcal{K}_\alpha)$ (see (1.13)). Combining (4.8) with the estimate (4.5), we have

$$\lim_{N\to\infty} \mathbb{E}\, \Lambda(u; s, \cdot) = \int_0^s p(\alpha)\mathrm{d}\alpha. \tag{4.9}$$

After rescaling (3.1), we also conclude that the limit of the expectation of $\Lambda$ with respect to the time evolved ensemble $e^{t\mathcal{L}}G_t$ (see Proposition 2.1) is given by right hand side of (4.9).

Finally, the difference of the expectation of $\Lambda$ with respect to the measure $e^{t\mathcal{L}}G_t$ and w.r.t. the initial ensemble $F$ vanishes since $|\Lambda| \le N$ and $\mathrm{Var}(e^{t\mathcal{L}}G_t, F) \le CN^{-2+4\lambda}$ (see (2.1) and (2.2)). This completes the proof of Theorem 1.2. □

# 5 Some extensions and comments

In this section we explain how to relax some of the conditions on the initial distribution $\nu$.

We first explain how to extend our proof to include distributions $\nu$ with compact support. Take for example a density w.r.t. the Gaussian measure $\mathrm{d}\mu(x) = e^{-x^2}$ that is given by a nice bump function $u(x)$ supported in $[-1, 1]$ decaying like $(1 \pm x)^m$ near the boundary $x = \pm 1$. Clearly, for any $m$ fixed this distribution violates the assumptions of Theorem 1.1. We now show that for $m$ large enough, it is still possible to prove the universality. Define a new distribution with density

$$q(x) = \frac{\tau^m + u(x)}{1 + \tau^m} \tag{5.1}$$

with a small parameter $\tau > 0$ to be determined later. Near the edge 1 we have $Lq(1-y) \lesssim Cy^{m-2}$ for $0 \le y \ll 1$ with some $m$-dependent constant $C$. We thus need the condition

$$Cy^{m-2} \le t^{-1}[\tau^m + y^m], \qquad 0 \le y \ll 1,$$

to guarantee that $(1 - tL)q$ is a probability density. This inequality holds if

$$\tau^2 \ge Ct.$$

The other conditions concerning $L^2$ and $L^3$ (see (2.3)) can be handled similarly. Choosing $\tau = Ct^{1/2}$, the total variation norm is bounded by

$$\int |q^{\otimes n} - u^{\otimes n}| \mathrm{d}\mu^{\otimes n} \le Cn\tau^m = Cnt^{m/2}.$$

Since $n = N^2$ and $t = N^{-1+\varepsilon}$, we have

$$\int |q^{\otimes n} - u^{\otimes n}| \mathrm{d}\mu^{\otimes n} \le C_m N^{2-m/2+m\varepsilon/2}.$$

Let, say, $m \ge 9$, then the error term will be smaller than $N^{-2-\delta}$ with some $\delta > 0$ and this will imply Theorem 1.1 for the initial distribution $u$. The modification of $u$ in (5.1) can certainly be more sophisticated to reduce the exponent $m$.

Second, we show that the Gaussian decay condition (1.6) can be replaced by the exponential decay (1.7). For any $\ell > 0$ define

$$\nu_\ell(x) = \nu(x + a_\ell)\mathbf{1}(|x| \le \ell)/Z_\ell,$$

where $a_\ell$ and $Z_\ell$ are chosen so that

$$\int x \, \mathrm{d}\nu_\ell = 0, \qquad \int \mathrm{d}\nu_\ell(x) = 1.$$

Due to the assumption (1.7), we have

$$|a_\ell| + |Z_\ell - 1| \le e^{-c\ell}.$$

Let $\widetilde{\nu}_\ell(x) = \ell\nu_\ell(x\ell)$. Clearly, the random variable $x$ distributed according to $\tilde{\nu}_\ell$ is bounded by 1, in particular it has a finite Gaussian moment. Denote the variance of $\tilde{\nu}_\ell$ by $\sigma_\ell^2$ and we have $\sigma_\ell = 1/\ell + O(e^{-c\ell})$. We will neglect all the exponential small terms $O(e^{-c\ell})$ and assume $\sigma_\ell = 1/\ell$. Similar cutoff and rescaling applies to the distribution of the diagonal elements.

Consider the random matrix generated by the measure $\nu_\ell$ and $\tilde{\nu}_\ell$ and denote the probability law of the eigenvalues by $f_\ell$ and $\tilde{f}_\ell$. Since all quantities introduced below can be defined w.r.t. to both $\nu_\ell$ and $\tilde{\nu}_\ell$, we will only give explicit definitions for $\nu_\ell$. Recall the Stieltjes transform of the eigenvalue distribution w.r.t. $\nu_\ell$ is defined as

$$m_\ell = m_\ell(z) = \int_{\mathbb{R}} \frac{\mathrm{d}F_\ell(E)}{E - z}\,, \tag{5.2}$$

where $F_\ell$ is the empirical distribution function of the eigenvalues. Then the empirical density of eigenvalues and $\tilde{m}_\ell$ converges to the rescaled semicircle law

$$\widetilde{\rho}_{sc}^{\ell}(x) = \ell\rho_{sc}(x\ell) = \frac{\ell}{2\pi}\sqrt{4 - x^2\ell^2}, \quad \widetilde{m}_{sc}^{\ell}(z) = \ell m_{sc}(z\ell).$$

We now follow the proof given in [8] to prove the local semicircle law Theorem 4.1 [8] for $\widetilde{\nu}_\ell$. The key estimate is contained in Proposition 4.3 which depends on Proposition 4.5. The random variables $b_j$ in Proposition 4.5 are now distributed according to $\tilde{\nu}_\ell$ and the only assumption of this proposition, the Gaussian bound (1.3) (i.e., the condition **C1**) is now trivially satisfied since $\widetilde{\nu}_\ell$ has compact support. Hence we can now prove Proposition 4.3 using the same strategy. Thus the equation for the probability estimate appearing after (4.6) in the paper still holds but the upper bound on the constant $A^2$ defined in (4.6) now becomes $2M\ell^2/(N\eta)$ due to the scaling. Thus the key estimate at the end of the proof of Proposition 4.3 of [8] is now changed to

$$\mathbb{E}_{\widetilde{\nu}_\ell}\Big[\mathbf{1}_{\Omega^c} \cdot \mathbb{P}_{\mathbf{b}}[|X| \ge \delta]\Big] \le 4\exp\big(-c\min\{\delta\sqrt{N\eta}/\ell,\, \delta^2 N\eta/\ell^2\}\big)\,. \tag{5.3}$$

Therefore, Theorem 4.1 of [8] holds with the estimates taking the form

$$\mathbb{P}_{\widetilde{\nu}_\ell}\Big\{\sup_{E\in[-(2-\kappa)/\ell,(2-\kappa)/\ell]}|\widetilde{m}^\ell_N(E+i\eta)-\widetilde{m}^\ell_{sc}(E+i\eta)|\geq\delta\Big\}\leq Ce^{-c\delta\sqrt{N\eta}/\ell} \tag{5.4}$$

for any $\delta\leq c_1\kappa/\ell$. Passing from $\widetilde{\nu}_\ell$ to $\nu_\ell$ via scaling, we have

$$\mathbb{P}_{\nu_\ell}\Big\{\sup_{E\in[-2+\kappa,2-\kappa]}|m^\ell_N(E+i\eta)-m^\ell_{sc}(E+i\eta)|\geq\delta\Big\}\leq Ce^{-c\delta\sqrt{N\eta/\ell}} \tag{5.5}$$

for $\delta\leq c_1\kappa$. Comparing this estimate with the original bound (4.1) in [8], note that the only change is that the $\eta$ in the exponent has deterioriated to $\eta/\ell$. This is due to fact that we applied the Proposition 4.5 of [8] without taking the advantage that the variance is now reduced to $1/\ell^2$, which should enhance the large deviation estimate (5.3). For our case, however, the estimate (5.5) is already sufficient since we are interested in the case $N\eta=N^\varepsilon$ and $\ell=(\log N)^2$.

Finally we need to pass estimates to the original measure $\nu_\ell^{\otimes n}$. We can check that

$$\mathrm{Var}\big(\nu_\ell^{\otimes n},\nu^{\otimes n}\big)\leq Cne^{-c\ell}.$$

Since in our application $n=N^2$ and $\ell=(\log N)^2$, the right hand side is smaller than any negative power of $N$, all necessary expectation values of observables w.r.t. $\nu_\ell^{\otimes n}$ can thus be passed to $\nu^{\otimes n}$. This shows that the local semicircle law holds on scales $\eta\geq N^{-1+\varepsilon}$ for any $\varepsilon>0$ assuming only exponential bound (1.7) instead of the Gaussian bound (1.6) required in **C1)** of [8]. This input is sufficient to conclude the proof in Section 3 if $t=a^2$ is changed to $N^{-1+\lambda}$ in (3.16).

# References


[1] Ben Arous, G., Péché, S.: Universality of local eigenvalue statistics for some sample covariance matrices. *Comm. Pure Appl. Math.* **LVIII.** (2005), 1–42.

[2] 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** (1997), 4067–4083.

[3] Deift, P.: Orthogonal polynomials and random matrices: a Riemann-Hilbert approach. *Courant Lecture Notes in Mathematics* **3**, American Mathematical Society, Providence, RI, 1999

[4] 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).

[5] Dyson, F.J.: A Brownian-motion model for the eigenvalues of a random matrix. *J. Math. Phys.* **3**, 1191-1198 (1962).

[6] Erdős, L., Schlein, B., Yau, H.-T.: Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices. Accepted in Ann. Probab. Preprint. arXiv.org:0711.1730

[7] Erdős, L., Schlein, B., Yau, H.-T.: Local semicircle law and complete delocalization for Wigner random matrices. *Commun. Math. Phys.* **287**, 641–655 (2009)

[8] Erdős, L., Schlein, B., Yau, H.-T.: Wegner estimate and level repulsion for Wigner random matrices. Submitted to Int. Math. Res. Notices (2008). Preprint arxiv.org/abs/0811.2591

[9] Erdős, L., Ramirez, J., Schlein, B., Yau, H.-T.: *Universality of sine-kernel for Wigner matrices with a small Gaussian perturbation.* Preprint arxiv.org/abs/0905.2089

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

[11] Mehta, M.L.: Random Matrices. Academic Press, New York, 1991.

[12] Pastur, L., Shcherbina M.: Bulk universality and related properties of Hermitian matrix models. J. Stat. Phys. **130** (2008), no.2., 205-250.

[13] Soshnikov, A.: Universality at the edge of the spectrum in Wigner random matrices. *Commun. Math. Phys.* **207** (1999), no.3. 697-733.

[14] Wigner, E.: Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math.* **62** (1955), 548-564.