Methods of Proof in Random Matrix Theory

by
Adina Roxana Feier
afeier@fas.harvard.edu

Advisor: Michael P. Brenner

Submitted to the
Harvard University Department of Mathematics
in partial fulfillment of the honors requirement for the degree of
Bachelor of Arts in Mathematics

Cambridge, Massachusetts March 2012

Contents

1 Introduction

Random matrix theory is concerned with the study of the eigenvalues, eigenvectors, and singular values of large-dimensional matrices whose entries are sampled according to known probability densities. Early interest in random matrices arose in the context of multivariate statistics with the works of Wishart [22] and Hsu [5] in the 1930s, but it was Wigner in the 1950s who introduced random matrix ensembles and derived the first asymptotic result through a series of papers motivated by nuclear physics [19, 20, 21].

As the theory developed, it was soon realized that the asymptotic behavior of random matrices is often independent of the distribution of the entries, a property called *universality*. Furthermore, the limiting distribution typically takes nonzero values only on a bounded interval, displaying *sharp edges*. For instance, Wigner's semicircle law is universal in the sense that the eigenvalue distribution of a symmetric or Hermitian matrix with i.i.d. entries, properly normalized, converges to the same density regardless of the underlying distribution of the matrix entries (figure 1). In addition, in this asymptotic limit the eigenvalues are almost surely supported on the interval $[-2, 2]$, illustrating the sharp edges behavior mentioned before.

Universality is important for theoretical as well as practical reasons. Historically, results such as Wigner's semicircle law were initially discovered for specific matrix ensembles, only to be later extended to more general classes of matrices. As another example, the circular law for the eigenvalues of a (non-symmetric) matrix with i.i.d. entries was initially established for Gaussian entries in 1965 [4], but only in 2008 was it fully expanded to arbitrary densities [10]. From a practical standpoint, the benefits of universality are clear, given that the same result can be applied to a vast class of problems.

Sharp edges are important for practical applications, where the hope is to use the behavior of random matrices to separate out a signal from noise. In such applications, the finite size of the matrices of interest poses a problem when adapting asymptotic results valid for matrices of infinite size. Nonetheless, an eigenvalue that appears significantly outside of the asymptotic range is a good indicator of non-random behavior. In contrast, trying to apply the same kind of heuristics when the asymptotic distribution is not compactly supported requires a much better understanding of the rate of convergence.

Although recently there has been increased interest in studying the eigenvectors of random matrices, a majority of the results established so far are concerned with the spectra, or eigenvalue distributions, of such matrices. Of interest are both the global regime, which refers to statistics on the entire set of eigenvalues, and the local regime, concerned with spacings between individual

eigenvalues.

In this paper, we present three classical theorems spanning both of these regimes: Wigner's semicircle law for the eigenvalues of symmetric or Hermitian matrices, the Marcenko-Pastur law for the eigenvalues of sample covariance matrices, and the Tracy-Widom distribution for the largest eigenvalue of Gaussian unitary matrices. In particular, we focus on exploring the different methods of proof used to derive these results, emphasizing the advantages and limitations of each. In doing so, we also trace the shift over the past few decades from proofs based on combinatorial arguments, seen in what is known as the moment method, to complex-analytical proofs based on the Stietljes transform, and finally to the new paradigm provided by the theory of free probability.

2 Preliminaries

2.1 Wigner matrices

In this section, we define the general Wigner matrix ensemble, and then describe several cases of particular interest. This ensemble is important for historical reasons, since it provided the first model of random matrices when introduced by Wigner, but it is still prominent in random matrix theory today because it is mathematically simple to work with, and yet has a high degree of generality.

Definition 2.1.1. Let $\{Y_i\}_{1\leq i}$ and $\{Z_{ij}\}_{1\leq i<j}$ be two real-valued families of zero mean, i.i.d. random variables. Furthermore, suppose that $\mathbf{E}Z_{12}^2 = 1$ and for each $k \in \mathbb{N}$,

$$\max(\mathbf{E}|Z_{12}|^k, \mathbf{E}|Y_1|^k) < \infty.$$

Consider a $n \times n$ symmetric matrix M_n whose entries are given by:

$$\begin{cases} M_n(i,i) = Y_i \\ M_n(i,j) = Z_{ij} = M_n(j,i), \quad \text{if } i < j \end{cases}$$

The matrix M_n is known as a *real symmetric Wigner matrix*.

Remark 2.1.2. Occasionally, the assumptions above are relaxed so that the entries of M_n don't necessarily have finite moments of all orders. Typically, the off-diagonal entries are still required to have identical second moments.

Definition 2.1.3. If the off-diagonal entries are complex-valued and we ask that M_n be Hermitian rather than symmetric, the construction above gives a *complex Hermitian Wigner matrix*.

The most important classes of Wigner matrices are presented in the examples below.

Example 2.1.4. If the Y_i and Z_{ij} are Gaussian, with Z_{ij} either real or complex, the resulting matrix M_n is called a *Gaussian Wigner matrix*. When $Y_i \sim \mathcal{N}(0,2)_{\mathbb{R}}$ and $Z_{ij} \sim \mathcal{N}(0,1)_{\mathbb{R}}$, one obtains the *Gaussian Orthogonal Ensemble*, which bears this name due to its invariance under orthogonal transformations. Similarly, the *Gaussian Unitary Ensemble*, invariant under unitary transformations, has $Y_i \sim \mathcal{N}(0,1)_{\mathbb{R}}$ and $Z_{ij} \sim \mathcal{N}(0,1)_{\mathbb{C}}$. The orthogonal and unitary ensembles are useful due to their highly symmetric nature, which makes possible direct calculations that would be infeasible in the general case.

Example 2.1.5. When Y_i and Z_{ij} are symmetric random sign random variables, the resulting matrices form the *symmetric Bernoulli ensemble*. Wigner's

semicircle law was initially proven for symmetric Bernoulli random matrices [20], before the author realized three years later that the result holds more generally [21].

Thinking of a random matrix M_n as a linear operator, we can form its *operator norm*, defined as

$$||M_n||_{\text{op}} = \min\{c \geq 0 : ||M_n v|| \leq c||v||, \forall v \in \mathbb{R}^n\}, \tag{2.1}$$

where $||\cdot||$ denotes the usual Euclidean norm. For reasons that will become more clear later on, the operator norm of a Wigner matrix of size n is typically $\mathcal{O}(\sqrt{n})$. Therefore, when studying the asymptotics of various statistics about Wigner matrices, we will often consider the normalised matrices $X_n := M_n/\sqrt{n}$ instead.

2.2 The empirical spectral distribution

Given a normalised Wigner matrix $X_n = M_n/\sqrt{n}$, consider its n eigenvalues $\lambda_1(X_n) \leq \ldots \leq \lambda_n(X_n)$. Because X_n is symmetric or Hermitian, these eigenvalues are all real. To study their distribution, we form the *empirical spectral distribution* (ESD),

$$\mu_{X_n} := \frac{1}{n}\sum_{j=1}^{n} \delta_{\lambda_j(X_n)}, \tag{2.2}$$

with $\delta_{\lambda_j(X_n)}(x)$ being the indicator function $\mathbf{1}_{\lambda_j(X_n)\leq x}$. When the matrix X_n or M_n can be inferred from the context, we will often write μ_n and λ_j^n in place of μ_{X_n} and $\lambda_j(X_n)$. Note that μ_{X_n} is a cumulative distribution function, as it is non-decreasing with $\lim_{x\to-\infty}\mu_{X_n}(x) = 0$ and $\lim_{x\to\infty}\mu_{X_n}(x) = 1$.

When M_n is random, the ESD is a probability measure on probability measures, giving the density of a random eigenvalue of a random matrix drawn from the Wigner ensemble. In particular, we can form the *deterministic* probability measure

$$\overline{\mu}_n := \mathbf{E}\frac{1}{n}\sum_{j=1}^{n} \delta_{\lambda_j(X_n)}, \tag{2.3}$$

defined by

$$\int_{\mathbb{R}} \phi \; d\mathbf{E}\mu_n = \mathbf{E}\int_{\mathbb{R}} \phi \; d\mu_n$$

for every continuous compactly supported $\phi \in \mathcal{C}_c(\mathbb{R})$.

In general, it is much easier to prove asymptotic results for the expected ESD $\mathbf{E}\mu_n$. In most cases, this turns out to be sufficient, as the difference

$$\left|\int_{\mathbb{R}} \phi d\mu_n - \int_{\mathbb{R}} \phi d\mathbf{E}\mu_n\right|$$

typically converges to 0 as $n \to \infty$ for every fixed $\phi \in \mathcal{C}_c(\mathbb{R})$.

2.3 Convergence of measures

Throughout the rest of this paper, when we say that a probability measure dependent on n (such as the ESD or expected ESD) converges to some asymptotic distribution, we mean so in the following *weak* sense:

Definition 2.3.1. A sequence ν_n of deterministic probability measures on a space Ω with associated σ-algebra $\mathcal{F}$ is said to *converge weakly* to a probability measure ν if for any bounded, continuous function f

$$\int_\Omega f \, d\nu_n \to \int_\Omega f \, d\nu$$

as $n \to \infty$. If ν_n is itself random, we are instead concerned with *weak convergence* to ν either *in probability* or *almost surely*.

3 Eigenvalue distribution of Wigner matrices: the semicircle law

3.1 Introduction

The goal of this section is to provide three different proofs of the following result:

Theorem 3.1.1. *Let $\{M_n\}_{n=1}^{\infty}$ be a sequence of Wigner matrices, and for each n denote $X_n = M_n/\sqrt{n}$. Then μ_{X_n} converges weakly, in probability to the semicircle distribution,*

$$\sigma(x)dx = \frac{1}{2\pi}\sqrt{4-x^2}\mathbf{1}_{|x|\leq 2}dx. \tag{3.1}$$

As it turns out, convergence in probability to the semicircle distribution can be updated to almost sure convergence, something which will be addressed later.

The semicircle law is as important to random matrix theory as the central limit theorem is to scalar probability theory. A first similarity between the two relies in their universality, as they are both valid for large classes of random matrices and random variables, respectively. Recently, with the development of free probability, it was realized that the connection between the two results run deeper, with the semicircle law essentially being the free analogue of the central limit theorem.

Before discussing this connection, we provide two other proofs of theorem 3.1.1, the first based on a direct calculation of the moments, and the second relying on complex-analytical methods that have been successful in proving other results as well.

3.2 The moment method

The most direct proof of the semicircle law, which is also the one advanced by Wigner in his original paper [20], uses the moment method. This approach relies on the intuition that eigenvalues of Wigner matrices are distributed according to some limiting non-random law – which, in our case, is the semicircle distribution $\sigma(x)$. The moments of the empirical distribution spectrum μ_n correspond to *sample* moments of the limiting distribution, where the number of samples is given by the size of the matrix. In the limit as this size goes to

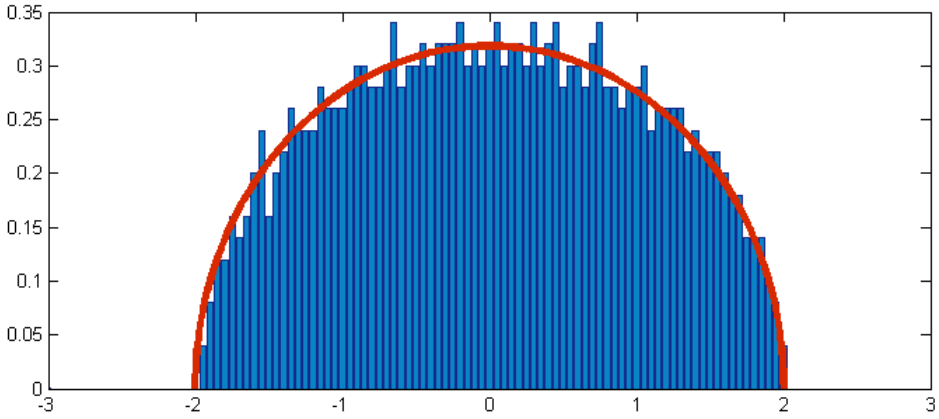

Figure 1: Simulation of the semicircle law using 1000 samples of the eigenvalues of 1000 by 1000 matrices. Bin size is 0.05.

infinity, it is therefore expected that the sample moments precisely recover the moments of the limiting distribution.

In what follows, we use the notation $\langle \mu, \phi \rangle := \int_{\mathbb{R}} \phi(x)\, d\mu(x)$ for a probability measure μ on $\mathbb{R}$. In particular, $\langle \mu, x^k \rangle$ denotes the kth moment of the law μ. The moment method proof of the semicircle law consists of the following two key steps [1]:

Lemma 3.2.1. *For any positive integer k, $\langle \overline{\mu}_n, x^k \rangle$ converges (deterministically) to $\langle \sigma, x^k \rangle$.*

Indeed, it is much easier to work with the *average* ESD $\overline{\mu}_n$ rather than the ESD μ_n corresponding to one particular matrix, and the following result shows that asymptotically, working with the former is just as accurate:

Lemma 3.2.2. *Fix $\epsilon > 0$ and k a positive integer. Then*

$$\lim_{n \to \infty} P(|\langle \mu_n, x^k \rangle - \langle \overline{\mu}_n, x^k \rangle| > \epsilon) = 0.$$

Because the law σ is symmetric about 0, all the odd moments are 0. To compute the even moments, substitute $x = 2 \sin \theta$ for $\theta \in [-\pi/2, \pi/2]$ to obtain a recurrence relation between consecutive even moments and establish the following:

Lemma 3.2.3. *The moments of the semicircle law are given by*

$$\langle \sigma, x^k \rangle = \begin{cases} C_{k/2} & \text{if } k \text{ is even} \\ 0 & \text{if } k \text{ is odd} \end{cases},$$

where C_n is the nth Catalan number, $C_n = \frac{1}{n+1}\binom{2n}{n}$.

Assuming these results, we can provide a proof to the semicircle law, based on a combination of [1], [14], and [20].

Proof of Theorem 3.1.1. To conclude $\mu_n \to \sigma$ in the weak sense, we need to show that for any bounded, continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\langle \mu_n, f \rangle \to \langle \sigma, f \rangle, \text{ in probability.}$$

The trick is to use the Weierstrass approximation theorem to replace f with a polynomial, and thus rewrite the integrals above as linear combinations of the moments. Because f needs to be compactly supported for this approximation to work, note first that for $B > 0$ we have by Markov's inequality:

$$P\left(\langle \mu_n, |x|^k \mathbf{1}_{|x|>B} \rangle > \epsilon\right) \leq \frac{1}{\epsilon} \mathbf{E}\langle \mu_n, |x|^k \mathbf{1}_{|x|>B} \rangle \leq \frac{\langle \overline{\mu}_n, x^{2k} \rangle}{\epsilon B^k},$$

where the last inequality follows by introducing the factor $|x|^k/B^k > 1$ inside the integral.

Using lemma 3.2.1 and the fact that $C_k \leq 4^k$,

$$\limsup_{n\to\infty} P\left(\langle \mu_n, |x|^k \mathbf{1}_{|x|>B} \rangle > \epsilon\right) \leq \frac{\langle \sigma, x^{2k} \rangle}{\epsilon B^k} \leq \frac{4^k}{\epsilon B^k}.$$

Now, let $B = 5$. This inequality holds true for any $k \in \mathbb{N}$. In particular, the left hand side is either zero or increasing in k, whereas the right hand side is strictly decreasing in k. Therefore, the only possibility is

$$\limsup_{n\to\infty} P\left(\langle \mu_n, |x|^k \mathbf{1}_{|x|>5} \rangle > \epsilon\right) = 0. \tag{3.2}$$

Next, consider $\delta > 0$ and $f : \mathbb{R} \to \mathbb{R}$ bounded. Because of what was done above, we can assume f to be compactly supported on $[-5, 5]$. On this interval, consider a polynomial p_δ such that $|p_\delta(x) - f(x)| \leq \delta/4$, $\forall x \in [-5, 5]$. Then, from the triangle inequality,

$$\begin{aligned} |\langle \mu_n, f \rangle - \langle \sigma, f \rangle| &\leq |\langle \mu_n, f - p_\delta \rangle - \langle \sigma, f - p_\delta \rangle| + |\langle \mu_n, p_\delta \rangle - \langle \sigma, p_\delta \rangle| \\ &\leq |\langle \mu_n, (f - p_\delta)\mathbf{1}_{|x|\leq 5} \rangle| + |\langle \sigma, (f - p_\delta)\mathbf{1}_{|x|\leq 5} \rangle| \\ &\quad + |\langle \mu_n, p_\delta \mathbf{1}_{|x|>5} \rangle| + |\langle \mu_n, p_\delta \rangle - \langle \sigma, p_\delta \rangle|, \end{aligned}$$

where we used the fact that the function f and the measure σ are 0 when $|x| > 5$. By choice of p_δ, the first two terms in the sum above are each bounded by $\delta/4$. Hence by applying the triangle inequality we get

$$\begin{aligned} P\left(|\langle \mu_n, f \rangle - \langle \sigma, f \rangle| > \delta\right) &\leq P\left(|\langle \mu_n, p_\delta \mathbf{1}_{|x|>5} \rangle| > \delta/2\right) \\ &+ P\left(|\langle \overline{\mu}_n, p_\delta \rangle - \langle \sigma, p_\delta \rangle| > \delta/2\right) \\ &+ P\left(|\langle \mu_n, p_\delta \rangle - \langle \overline{\mu}_n, p_\delta \rangle| > \delta/2\right) \end{aligned}$$

By (3.2), the first summand above goes to 0 as $n \to \infty$. The second term is *equal* to 0 when n is sufficiently large, by lemma 3.2.1. Lastly, by lemma 3.2.2, the third summand converges to 0 as $n \to \infty$.

Thus, for any $\delta > 0$ and any bounded function f, we have shown that

$$\lim_{n\to\infty} P\left(|\langle \mu_n, f\rangle - \langle \sigma, f\rangle| > \delta\right) = 0,$$

which proves that $\mu_n \to \sigma$ weakly, in probability. ∎

The philosophy behind the moment method is best seen in the proofs of the two outstanding lemmas 3.2.1 and 3.2.2.

Proof of Lemma 3.2.1. The starting point in proving the convergence of empirical spectral moments is the identity

$$\langle \mu_n, x^k\rangle = \int_{\mathbb{R}} x^k d\mu(x) = \frac{1}{n}\mathrm{tr}X_n^k, \tag{3.3}$$

which holds true because both sides are equal to $\frac{1}{n}(\lambda_1^k + \ldots + \lambda_n^k)$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of X_n. Taking expectations and writing ζ_{ij} for the (i,j) entry of X_n, we have

$$\langle \overline{\mu}_n, x^k\rangle = \frac{1}{n}\sum_{i_1,\ldots,i_k=1}^{n} \mathbf{E}\ \zeta_{i_1 i_2}\cdots\zeta_{i_{k-1}i_k}\zeta_{i_k i_1}. \tag{3.4}$$

The combinatorial analysis that follows is very close in spirit to the approach used originally by Wigner, though he initially studied the less general case of symmetric Bernoulli matrices [20]. Consider the sequence $\mathbf{i} = i_1 i_2 \cdots i_k i_1$ of length $k+1$, with $i_j \in \{1, \ldots, n\}$. Each sequence of this form corresponds uniquely to a term in the sum (3.4), and can be thought of as a *closed, connected path* on the set of vertices $\{i_1, \ldots, i_k\}$, with the edges described by pairs of consecutive indices $i_j i_{j+1}$.

Therefore, each term in the sum (3.4) corresponds bijectively to a path of length k on the set of vertices specified by $\mathbf{i}$. In particular, because the entries of X_n have mean 0 and are independent (up to the Hermitian condition), the summand corresponding to a sequence $\mathbf{i}$ will be 0 unless every edge in the corresponding path is traversed at least twice, possibly in reverse. Thus, there are at most $k/2$ unique edges, and hence at most $k/2+1$ distinct vertices $i_1, \ldots, i_k$.

Define the *weight* t of a sequence $\mathbf{i}$ to be the number of distinct indices $i_1, \ldots, i_k$. By the observation above, the nonzero terms in (3.4) have $t \leq k/2+1$. Furthermore, we say that two sequences $\mathbf{i} = i_1 i_2 \cdots i_k i_1$ and $\mathbf{i}' = i_1' i_2' \cdots i_k' i_1'$ are

equivalent if there exists a bijection on the set $\{1, \ldots, n\}$ mapping each i_j to i'_j. Note that equivalent sequences have the same weight and, more importantly, their corresponding terms in (3.4) are equal. Also, the number of distinct equivalent classes depends on k but not on n, since each class has a representative where all $i_1, \ldots, i_k$ are in $\{1, \ldots, k\}$.

We first show that terms with $t < k/2 + 1$ are negligible in the limit $n \to \infty$. Given $\mathbf{i} = i_1 i_2 \cdots i_k i_1$ of weight t, there are $n(n-1)\cdots(n-t+1) \leq n^t$ sequences equivalent to it. The contribution of each term in this equivalence class to the sum (3.4) is

$$\frac{1}{n}\mathbf{E}\ \zeta_{i_1 i_2} \cdots \zeta_{i_{k-1} i_k} \zeta_{i_k i_1} = \mathcal{O}\left(\frac{1}{n} \cdot \frac{1}{\sqrt{n}^k}\right),$$

because $X_n = M_n/\sqrt{n}$ and the entries of M_n have uniformly bounded moments for all n. Thus, for each equivalence class with weight $t < k/2 + 1$, the total contribution to (3.4) is at most $\mathcal{O}(n^t/n^{k/2+1}) \to 0$ as $n \to \infty$. Since the number of equivalence classes does not depend on n, we can ignore all terms of weight $t < k/2 + 1$.

When k is odd, $t = k/2 + 1$ is impossible, so the odd moments of $\overline{\mu}_n$ converge to 0, as expected from lemma 3.2.3.

Next, we focus on the terms with $t = k/2 + 1$ with k even, corresponding to connected graphs on $k/2 + 1$ vertices with k distinct edges. We see that such graphs are in fact trees, and the sequence $\mathbf{i} = i_1 i_2 \cdots i_k i_1$ represents a closed path on this tree which traverses each edge exactly twice, once in each direction. In particular, there are no self-loops in the graph, meaning that the corresponding term in (3.4) contains no diagonal entries of X_n. Because each off-diagonal element of M_n is assumed to have variance 1, we see that

$$\frac{1}{n}\mathbf{E}\ \zeta_{i_1 i_2} \cdots \zeta_{i_{k-1} i_k} \zeta_{i_k i_1} = \frac{1}{n} \cdot \frac{1}{\sqrt{n}^k}. \tag{3.5}$$

We have thus reduced the problem to counting the number of sequences $\mathbf{i} = i_1 i_2 \cdots i_k i_1$ with $t = k/2 + 1$ and k distinct edges, each traversed twice. Such paths are called *non-crossing*. During the traversal of $\mathbf{i}$, an edge $i_j i_{j+1}$ is called *free* if it appears for the first time in $\mathbf{i}$, and *repetitive* if it has been traversed once before. Given a non-crossing path $\mathbf{i}$ of length k, define its *type sequence*, whose jth entry is the number of free steps minus the number of repetitive steps in the path $i_1 \cdots i_{j+1}$ (with the convention $i_{k+1} = 1$). Note that a type sequence starts at 1, ends in 0, and has successive terms differing by ± 1. For example, the type sequence of $\mathbf{i} = 132524231$ is 12323210.

Now, observe that two non-crossing sequences are equivalent if and only if they have the same type sequence. Thus, the number of $\mathbf{i}$ corresponding to a given type sequence is $n(n-1)\cdots(n-t+1) = \mathcal{O}(n^{k/2+1})$. Combining this

with (3.5), and recalling that the terms with $t < k/2 + 1$ are negligible for n large, we see that

$$\langle \overline{\mu}_n, x^k \rangle = \text{number of type sequences of length } k.$$

Let m_l denote the number of type sequences of length $2l$. Denote by m'_l the number of type sequences of length $2l$ with no 0 occurring before the last term. Any such sequence corresponds bijectively to a type sequence of length $2l-2$, since we can remove the first and last terms and subtract 1 from the rest to obtain a type sequence that is still valid. Hence, $m'_l = m_{l-1}$. By considering the position of the first 0 in a type sequence of length $2l$, one can similarly deduce

$$m_l = \sum_{j=1}^{l} m_{j-1} m_{l-j},$$

with the convention $m_0 = 1$. This is precisely the recursion satisfied by the Catalan numbers, which implies that the number of type sequences of length k is $C_{k/2}$.

Therefore, for $n \to \infty$, $\langle \overline{\mu}_n, x^k \rangle \to 0$ when k is odd and $\langle \overline{\mu}_n, x^k \rangle \to C_{k/2}$ when k is even. Together with lemma 3.2.3, this shows that the moments of $\overline{\mu}_n$ converge deterministically to the moments of the semicircle law σ. ∎

Similar combinatorial arguments are used to show that the moments of an ESD μ_n are close to the average ESD moments given by $\overline{\mu}_n$.

Proof of Lemma 3.2.2. By Chebyshev's inequality,

$$P\left(|\langle \mu_n, x^k \rangle - \langle \overline{\mu}_n, x^k \rangle| > \epsilon\right) \leq \frac{1}{\epsilon^2} \left| \mathbf{E}\left(\langle \mu_n, x^k \rangle\right)^2 - \left(\mathbf{E}\langle \mu_n, x^k \rangle\right)^2 \right|,$$

so it suffices to show that the right hand side goes to 0 as $n \to \infty$.

Again, we can rewrite moments in terms of matrix traces:

$$\begin{aligned} \left| \mathbf{E}\left(\langle \mu_n, x^k \rangle\right)^2 - \left(\mathbf{E}\langle \mu_n, x^k \rangle\right)^2 \right| &= \frac{1}{n^2} \left[\mathbf{E}(\mathrm{tr} X_n^k)^2 - (\mathbf{E}\mathrm{tr} X_n^k)^2 \right] \\ &= \frac{1}{n^2} \sum_{\mathbf{i},\mathbf{i}'} [\mathbf{E}\zeta_{\mathbf{i}}\zeta_{\mathbf{i}'} - \mathbf{E}\zeta_{\mathbf{i}}\mathbf{E}\zeta_{\mathbf{i}'}], \end{aligned} \tag{3.6}$$

where $\zeta_{\mathbf{i}}$ is shorthand for the product $\zeta_{i_1 i_2} \cdots \zeta_{i_k i_1}$, with $i_1, \ldots, i_k \in \{1, \ldots, n\}$, and similarly for $\zeta_{\mathbf{i}'}$.

As before, each pair $(\mathbf{i}, \mathbf{i}')$ generates a graph with vertices $V_{\mathbf{i},\mathbf{i}'} = \{i_1, \ldots, i_k\} \cup \{i'_1, \ldots, i'_k\}$ and edges $E_{\mathbf{i},\mathbf{i}'} = \{i_1 i_2, \ldots, i_k i_1\} \cup \{i'_1 i'_2, \ldots, i'_k i'_1\}$. With pairs rather than single sequences, however, the resulting graph is not necessarily

connected. The *weight* of $(\mathbf{i}, \mathbf{i}')$ is defined as the cardinality of $V_{\mathbf{i},\mathbf{i}'}$. Two pairs $(\mathbf{i}, \mathbf{i}')$ and $(\mathbf{j}, \mathbf{j}')$ are again said to be *equivalent* if there is a bijection on $\{1, \ldots, n\}$ mapping corresponding indices to each other; equivalent pairs of sequences contribute the same amount to the sum in (3.6).

In order for the term in (3.6) corresponding to $(\mathbf{i}, \mathbf{i})$ to be nonzero, the following are necessary:

- Each edge in $E_{\mathbf{i},\mathbf{i}'}$ appears at least twice, since the entries of X_n have 0 mean.
- The graphs generated by $\mathbf{i}$ and $\mathbf{i}'$ have at least one edge in common, otherwise it follows by independence that $\mathbf{E}\zeta_{\mathbf{i}}\zeta_{\mathbf{i}'} - \mathbf{E}\zeta_{\mathbf{i}}\mathbf{E}\zeta_{\mathbf{i}'} = 0$.

Pairs $(\mathbf{i}, \mathbf{i}')$ satisfying these two conditions will be called *nonzero pairs*.

Given $(\mathbf{i}, \mathbf{i}')$ of weight $t \leq k+1$, there are $n(n-1)\cdots(n-t+1) \leq n^{k+1}$ equivalent pairs. Furthermore, the contribution of each such pair to (3.6) is $\mathcal{O}(1/n^{k+2})$, since $X_n = M_n/\sqrt{n}$ and the entries of M_n have bounded moments, uniformly in n. Thus, each equivalence class with weight $t \leq k+1$ contributes an asymptotically negligible amount to (3.6). Because the number of equivalence classes depends on k but not on n, the total contribution of terms with $t \leq k+1$ converges to 0 as $n \to \infty$.

Next, focus on the terms with $t \geq k+2$. Each equivalence class with such t has $\mathcal{O}(n^t)$ elements, which contribute at least $\mathcal{O}(1)$ to (3.6). Thus, in order for $\mathbf{E}(\langle \mu_n, x^k \rangle)^2 - (\mathbf{E}\langle \mu_n, x^k \rangle)^2$ to converge to 0, it must be the case that there are no nonzero pairs $(\mathbf{i}, \mathbf{i}')$ of weight $t \geq k+2$. In fact, this is the case for $t \geq k+1$, and because it will be useful later on, we will prove this somewhat stronger statement.

When $(\mathbf{i}, \mathbf{i}')$ is a nonzero pair, the corresponding graph is connected with at most k unique edges. This is impossible if the weight t, which equals the number of distinct vertices, is at least $k+2$.

Finally, consider $(\mathbf{i}, \mathbf{i}')$ with $t = k+1$, in which case the resulting graph is a tree and each edge gets traversed exactly twice, once in each direction. Because the path generated by $\mathbf{i}$ in this tree starts and ends at i_1, it must traverse each edge an even number of times. The equivalent statement is true for $\mathbf{i}'$. Thus, each edge in $V_{\mathbf{i},\mathbf{i}'}$ gets traversed by one of $\mathbf{i}$ or $\mathbf{i}'$, bot not both. Hence $\mathbf{i}$ and $\mathbf{i}'$ have disjoint edges, contradicting the assumption that $(\mathbf{i}, \mathbf{i}')$ is a nonzero pair.

With this, we have shown that $\mathbf{E}(\langle \mu_n, x^k \rangle)^2 - (\mathbf{E}\langle \mu_n, x^k \rangle)^2$ is $\mathcal{O}(1/n^2)$, which proves that $\langle \mu_n, x^k \rangle \to \langle \overline{\mu}_n, x^k \rangle$ as $n \to \infty$, in probability. ∎

In the course of proving the above lemma, we also showed the following:

Lemma 3.2.4. *Let X_n be a Wigner matrix with ESD μ_n. Then for every fixed k, there exists a constant C not depending on n such that*

$$\left|\mathbf{E}(\langle \mu_n, x^k\rangle)^2 - (\mathbf{E}\langle \mu_n, x^k\rangle)^2\right| \leq \frac{C}{n^2}$$

for all sufficiently large n.

With this, one can show that the convergence in Wigner's semicircle law holds almost surely:

Corollary 3.2.5. *Let $\{M_n\}_{n=1}^\infty$ be a sequence of Wigner matrices, and for each n denote $X_n = M_n/\sqrt{n}$. Then μ_{X_n} converges weakly, almost surely to the semicircle distribution.*

Proof. The convergence established by lemma 3.2.1 is already deterministic. It remains to check that the convergence in probability in lemma 3.2.2 can be updated to almost sure convergence. Using Chebyshev's inequality and lemma 3.2.4, we have that for every k

$$\begin{aligned}\sum_{n=1}^\infty P(|\langle \mu_n, x^k\rangle - \langle \overline{\mu}_n, x^k\rangle| > \epsilon) &\leq \frac{1}{\epsilon^2}\sum_{n=1}^\infty \left|\mathbf{E}(\langle \mu_n, x^k\rangle)^2 - (\mathbf{E}\langle \mu_n, x^k\rangle)^2\right| \\ &\leq C_1 + \frac{1}{\epsilon}\sum_{n=1}^\infty \frac{C}{n^2} < \infty,\end{aligned}$$

where the constant C_1 accounts for the fact that lemma 3.2.4 only becomes valid for n sufficiently large.

The Borel-Cantelli lemma then gives

$$P\left(\limsup_{n\to\infty} |\langle \mu_n, x^k\rangle - \langle \overline{\mu}_n, x^k\rangle| > \epsilon\right) = 0,$$

thus showing that $|\langle \mu_n, x^k\rangle - \langle \overline{\mu}_n, x^k\rangle| \to 0$ almost surely.

Now, with f compactly supported on $[-5,5]$ and p_δ a polynomial approximation to f as in the original proof of theorem 3.1.1, we have

$$\begin{aligned}|\langle \mu_n, f\rangle - \langle \sigma, f\rangle| &\leq |\langle \mu_n, (f-p_\delta)\mathbf{1}_{|x|\leq 5}\rangle| + |\langle \sigma, (f-p_\delta)\mathbf{1}_{|x|\leq 5}\rangle| \\ &\quad + |\langle \mu_n, p_\delta \mathbf{1}_{|x|>5}\rangle| + |\langle \overline{\mu}_n, p_\delta\rangle - \langle \sigma, p_\delta\rangle| \\ &\quad + |\langle \mu_n, p_\delta\rangle - \langle \overline{\mu}_n, p_\delta\rangle|.\end{aligned}$$

By suitable choice of p_δ, the first two terms can be made arbitrarily small. The third and fourth terms approach 0 deterministically, and the last one converges

to 0 almost surely. Overall, this shows $|\langle \mu_n, f\rangle - \langle \sigma, f\rangle| \to 0$ almost surely for every bounded f. Thus, the ESD μ_n converges almost surely, weakly to the semicircle distribution. ∎

What is interesting about the moment method is that it reduces statements about convergence of moments to simple counting arguments. However, one shortfall of this approach is that it's not constructive – while the calculation gives an explicit formula for the moments of the asymptotic distribution, deducing the law that corresponds to those moments cannot be done by such elementary means. Wigner originally arrived at the semicircle density (3.1) by setting up a differential equation having σ as a solution, before another mathematician, Feller, suggested to him a derivation which uses the characteristic function [20].

Nonetheless, the moment method provides a valuable insight into the universality of this result. Our counting argument illustrates that the moments of the matrix entries of order above two are negligible in the asymptotic limit, as long as they are uniformly bounded in n. The only significant terms correspond to second moments, which are easily dealt with if assumed to be the same for all entries. Some of these assumptions can be further relaxed with more work, but what remains striking is the fact that the same argument devised by Wigner for the symmetric Bernoulli ensemble works for such a general class of matrices with essentially no modifications.

3.3 Stieltjes transform method

With the moment method, we are proving convergence to the semicircle law one moment at a time. For scalar random variables, it is often convenient to group the moments together using constructs such as the moment generating function or the characteristic function. It is a natural question, then, whether this can be done for matrix-valued random variables.

For any Wigner matrix M_n, we can consider the *Stieltjes transform* [14] of its (normalised) ESD $\mu_n = \mu_{M_n/\sqrt{n}}$, defined for complex z outside of the support of μ_n:

$$s_n(z) = \int_{\mathbb{R}} \frac{1}{x - z} d\mu_n. \tag{3.7}$$

Keeping in mind the definition of $d\mu_n$, which is concentrated around the eigenvalues $\lambda_1, \ldots, \lambda_n$ of M_n, we have the identity

$$s_n(z) = \int_{\mathbb{R}} \frac{1}{x - z} d\mu_n = \frac{1}{n} \mathrm{tr}\left(M_n/\sqrt{n} - zI\right)^{-1}.$$

This leads to the formal equality

$$s_n(z) = -\frac{1}{n}\sum_{k=0}^{\infty}\frac{\mathrm{tr}M_n^k}{z^{k+1}},$$

which converges for large enough z. Thus, a good understanding of the Stieltjes transform also provides information about the moments. In fact, the underlying density can be directly recovered from the Stieltjes transform:

Proposition 3.3.1. *For $a \in \mathbb{R}$,*

$$\lim_{b\to 0^+}\frac{s_\mu(a+bi) - s_\mu(a-bi)}{2\pi i} = d\mu(a).$$

In addition, we have the following useful criterion for convergence:

Proposition 3.3.2. *Let μ_n be a sequence of random probability measures on $\mathbb{R}$, and μ a deterministic probability measure. Then μ_n converges weakly in probability to μ if and only is $s_{\mu_n}(z)$ converges in probability to $s_\mu(z)$ for every z in the upper-half plane.*

For the Stieltjes transform proof of the semicircle law, we will have the sequence of $n \times n$ matrices M_n represent successive top-left minors of an infinite Wigner matrix. Thus, M_n is formed by adding one independent row and one independent column to M_{n-1}. While this choice does not affect the conclusion of the semicircle law, it does make it easier to relate the Stieltjes transforms s_n and s_{n-1}, thanks to the following result:

Proposition 3.3.3. *Let A_n be an $n \times n$ matrix, and fix i between 1 and n. Let A_{n-1}^i be the $(n-1)\times(n-1)$ matrix with column i and row i eliminated. Denote by c_i the ith column of A_n with the entry $A_n(i,i)$ removed, and by r_i the ith row of A_n with the entry $A_n(i,i)$ removed. Suppose A_n and A_{n-1}^i are invertible. Then:*

$$A_n^{-1}(i,i) = \frac{1}{A_n(i,i) - r_i(A_{n-1}^i)^{-1}c_i}.$$

Because the Stieltjes transform is related to traces of matrices, which in turn are related to eigenvalues, the following standard result comes in useful later on:

Proposition 3.3.4. *Let A, B be $n \times n$ Hermitian matrices with eigenvalues $\lambda_1^A \leq \ldots \leq \lambda_n^A$ and $\lambda_1^B \leq \ldots \leq \lambda_n^B$. Then*

$$\sum_{i=1}^{n}|\lambda_i^A - \lambda_i^B| \leq \mathrm{tr}(A-B)^2.$$

This result is known as the Hoffman-Wielandt inequality. The proof can be framed as a linear optimization problem over the convex set of doubly stochastic matrices (i.e., matrices with nonnegative real entries with the entries in each row and each column summing to 1). For the specific details, we refer the reader to [1].

Before proceeding with a proof of theorem 3.1.1, we make the following reductions:

Lemma 3.3.5. *For the matrices M_n in theorem 3.1.1, it can be assumed without loss of generality that the diagonal entries are 0 and that the off-diagonal entries are bounded, i.e. $|M_n(i,j)| \leq C$ for all i, j, where C is a constant not depending on n.*

Proof. For every n, define

$$\overline{X}_n(i,j) = \begin{cases} 0 & \text{if } i = j \\ X_n(i,j)\mathbf{1}_{\sqrt{n}|X_n(i,j)|\leq C} - \mathbf{E}(X_n(i,j)\mathbf{1}_{\sqrt{n}|X_n(i,j)|\leq C}) & \text{if } i \neq j \end{cases}.$$

Thus, $\overline{X}_n$ is obtained from X_n by setting the diagonal entries to 0 and keeping just those off-diagonal entries that are bounded by C in the original matrix M_n (recall that $X_n = M_n/\sqrt{n}$). Because the distribution of these entries changes, it is necessary to recenter them so that they have mean 0.

Assume that the semicircle law holds for $\overline{X}_n$. The goal is to deduce this for X_n as well. To this end, define

$$\begin{aligned} W_n &= \frac{1}{n}\mathrm{tr}(X_n - \overline{X}_n)^2 \\ &\leq \frac{1}{n^2}\sum_{i\neq j}\left[\sqrt{n}X_n(i,j)\mathbf{1}_{\sqrt{n}|X_n(i,j)|\geq C} - \mathbf{E}(\sqrt{n}X_n(i,j)\mathbf{1}_{\sqrt{n}|X_n(i,j)|\geq C})\right]^2 \\ &\quad + \frac{1}{n}\sum_i (X_n(i,i))^2. \end{aligned}$$

By the strong law of large numbers, the second term in the sum above is almost surely 0 as $n \to \infty$. With

$$Y_n(i,j) = \sqrt{n}X_n(i,j)\mathbf{1}_{\sqrt{n}|X_n(i,j)|\geq C} - \mathbf{E}(\sqrt{n}X_n(i,j)\mathbf{1}_{\sqrt{n}|X_n(i,j)|\geq C})$$

and $\epsilon > 0$,

$$P(|W_n| > \epsilon) \leq \frac{1}{n^2}\sum_{i\neq j} P\left(Y_n(i,j)^2 > \epsilon\right).$$

By Markov's inequality,

$$\begin{aligned} P(|Y_n(i,j)|^2 > \epsilon) &\leq \frac{1}{\epsilon}\mathbf{E}|Y_n(i,j)|^2 \\ &\leq \frac{1}{\epsilon}\mathbf{E}[(\sqrt{n}X_n(i,j))^2\mathbf{1}_{\sqrt{n}|X_n(i,j)|\geq C}] \\ &\quad + \frac{1}{\epsilon}\left[\mathbf{E}(\sqrt{n}X_n(i,j)\mathbf{1}_{\sqrt{n}|X_n(i,j)|\geq C})\right]^2. \end{aligned}$$

Because the entries $\sqrt{n}X_n(i,j)$ have finite variances uniformly for all n, the right hand side of the equality above converges to 0 as $C \to \infty$. Consequently, $P(|W_n| > \epsilon) \to 0$ as $C \to \infty$ as well. Therefore, given any $\delta > 0$, we can find a large enough C so that $P(|W_n| > \epsilon) < \delta$ for all sufficiently large n.

Now, we are ready to prove that the ESD $\mu_{\overline{X}_n}$ approximates μ_{X_n} and, heuristically, because the former converges to the semicircle law, so should the latter. By the portmanteau theorem (see [1]), to show weak convergence it is sufficient to check

$$|\langle \mu_{X_n}, f\rangle - \langle \mu_{\overline{X}_n}, f\rangle| \to 0$$

in probability when f is a bounded, Lipschitz continuous function with Lipschitz constant 1. In this case, if $\lambda_1 \leq \ldots \leq \lambda_n$ and $\overline{\lambda}_1 \leq \ldots \leq \overline{\lambda}_n$ denote the eigenvalues of X_n and $\overline{X}_n$,

$$|\langle \mu_{X_n}, f\rangle - \langle \mu_{\overline{X}_n}, f\rangle| \leq \frac{1}{n}\sum_{i=1}^{n}|\lambda_i - \overline{\lambda}_i| \leq \left[\frac{1}{n}\sum_{i=1}^{n}(\lambda_i - \overline{\lambda}_i)^2\right]^{1/2},$$

which together with proposition 3.3.4 gives

$$|\langle \mu_{X_n}, f\rangle - \langle \mu_{\overline{X}_n}, f\rangle| \leq \left[\frac{1}{n}\mathrm{tr}(X_n - \overline{X}_n)^2\right]^{1/2} \leq \sqrt{\epsilon},$$

as long as $|W_n| < \epsilon$.

Putting everything together, we have that for each $\epsilon > 0$ and $\delta > 0$, there exists C large enough with the corresponding $\mu_{\overline{X}_n}$ converging in probability to the semicircle law, in which case

$$P(|\langle \mu_n, f\rangle - \langle \sigma, f\rangle| > \epsilon) \leq P(|\langle \mu_n, f\rangle - \langle \mu_{\overline{X}_n}, f\rangle| > \sqrt{\epsilon}) + P(|\langle \mu_{\overline{X}_n}, f\rangle - \langle \sigma, f\rangle| > \epsilon).$$

Recall that C was chosen so that $P(|W_n| > \epsilon) < \delta$, meaning that as $n \to \infty$ we get

$$\lim_{n\to\infty} P(|\langle \mu_n, f\rangle - \langle \sigma, f\rangle| > \epsilon) < \delta.$$

For fixed ϵ, the above must hold true for all $\delta > 0$, which implies

$$\lim_{n\to\infty} P(|\langle \mu_n, f\rangle - \langle \sigma, f\rangle| > \epsilon) = 0,$$

and thus the ESD of X_n converges weakly, in probability to the semicircle law. ∎

Remark 3.3.6. The argument in this proof can also be used to show that the semicircle law holds for Wigner matrices whose entries have mean 0 and finite variance, without making other assumptions about the moments.

With this setup, we come to a second proof of the semicircle law, which uses the Stieltjes transform.

Proof of Theorem 3.1.1. From $s_n(z) = \frac{1}{n}\text{tr}\left(\frac{1}{\sqrt{n}}M_n - zI\right)^{-1}$, linearity of trace implies

$$s_n(z) = \frac{1}{n}\sum_{i=1}^{n}(M_n/\sqrt{n} - zI)^{-1}(i,i).$$

To make notation simpler, write X_n and X_{n-1}^i for $M_n/\sqrt{n}$ and $M_{n-1}^i/\sqrt{n}$. In particular, note that the latter is normalized by $\sqrt{n}$, not $\sqrt{n-1}$. If w_i denotes the ith column of X_n excluding the entry $X_n(i,i)$, the Hermitian condition on X_n implies that the ith row of X_n, excluding the (i,i) entry, is w_i^*. Proposition 3.3.3 then gives

$$s_n(z) = -\sum_{i=1}^{n}\frac{1}{z + w_i^*\left(X_{n-1}^i - zI\right)^{-1}w_i}.$$

Define $\delta_n(z)$ by

$$s_n(z) = -\frac{1}{z + s_n(z)} - \delta_n(z),$$

so that δ_n measures the error in the two expressions for $s_n(z)$ being equal. Together with the previous equality, we get

$$\delta_n(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{s_n(z) - w_i^*(X_{n-1}^i - zI)^{-1}w}{(z + s_n(z))(z + w_i^*(X_{n-1}^i - zI)^{-1}w_i)}.$$

The goal is to show that for any fixed z in the upper-half plane, $\delta_n(z) \to 0$ in probability as $n \to \infty$. Restricting to the upper-half plane is sufficient because for any measure μ, $\overline{s_\mu(z)} = s_\mu(\overline{z})$. Let $\epsilon_n^i = s_n(z) - w_i^*(X_{n-1}^i - zI)^{-1}w_i$. Then

$$\delta_n(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{\epsilon_n^i}{(-z - s_n(z))(-z - s_n(z) + \epsilon_n^i)}.$$

A simple calculation shows that for $z = a + bi$ with $b > 0$, the imaginary part of $s_n(z)$ is positive. This implies $|z + s_n(z)| > b$, meaning that the convergence of $\delta_n(z)$ depends only on the limiting behaviour of ϵ_n^i. Specifically, because the sum above is normalized by $1/n$, it suffices to show that $\sup_i |\epsilon_n^i| \to 0$ in probability.

Let $\overline{X}_n^i$ be the matrix obtained from X_n by replacing all elements in the ith row and ith column with 0. Then $(\overline{X}_n^i - zI)^{-1}$ and $(X_{n-1}^i - zI)^{-1}$ have $n-1$

identical eigenvalues, with $(\overline{X}_n^i - zI)^{-1}$ having an additional eigenvalue equal to $-1/z$. Therefore,

$$\begin{aligned}\left| s_{\overline{X}_n^i}(z) - \frac{1}{n}\mathrm{tr}(X_{n-1}^i - zI)^{-1} \right| &= \frac{1}{n}\left|\mathrm{tr}(\overline{X}_n^i - zI)^{-1} - \mathrm{tr}(X_{n-1}^i - zI)^{-1}\right| \\ &= \frac{1}{n|z|} \leq \frac{1}{nb} = o(1). \end{aligned} \tag{3.8}$$

Next, we want to bound $|s_{X_n}(z) - s_{\overline{X}_n^i}(z)|$. Let $\lambda_1 \leq \ldots \leq \lambda_n$ denote the eigenvalues of X_n, and $\overline{\lambda}_1 \leq \ldots \leq \overline{\lambda}_n$ the eigenvalues of $\overline{X}_n^i$. Then the eigenvalues of $(X_n - zI)^{-1}$ and $(\overline{X}_n^i - zI)^{-1}$ are $(\lambda_j - z)^{-1}$ and $(\overline{\lambda}_j - z)^{-1}$, for $j = 1, \ldots, n$. This observation, together with the Cauchy-Schwartz inequality and proposition 3.3.4, yields the following bound:

$$\begin{aligned}|s_{X_n}(z) - s_{\overline{X}_n^i}(z)| &= \frac{1}{n}\sum_{j=1}^n \left|\frac{1}{\lambda_j - z} - \frac{1}{\overline{\lambda}_j - z}\right| \\ &= \frac{1}{n}\sum_{j=1}^n \frac{|\lambda_j - \overline{\lambda}_j|}{|(\lambda_j - a) - bi| \cdot |(\overline{\lambda}_j - a) - bi|} \\ &\leq \frac{1}{b^2}\left(\frac{1}{n}\sum_{j=1}^n |\lambda_j - \overline{\lambda}_j|^2\right)^{1/2} \leq \frac{1}{b^2}\left(\frac{1}{n}\mathrm{tr}(\overline{X}_n^i - X_n)^2)\right)^{1/2} \\ &\leq \frac{1}{b^2}\left(\frac{2}{n}\sum_{j=1}^n X_n(i,j)^2\right)^{1/2} \leq \frac{1}{b^2}\left(\frac{2C^2}{n}\right)^{1/2} = o(1), \end{aligned} \tag{3.9}$$

where the last inequality follows from the assumption that each entry of M_n is bounded by C, and thus $X_n(i,j) \leq C/\sqrt{n}$.

By the triangle inequality and the bounds established in (3.8) and (3.9), showing $\sup_i |\epsilon_n^i| \to 0$ in probability reduces to proving $\sup_i |\overline{\epsilon}_n^i| \to 0$ in probability, where

$$\overline{\epsilon}_n^i := w_i^*(X_{n-1}^i - zI)^{-1}w_i - \frac{1}{n}\mathrm{tr}(X_{n-1}^i - zI)^{-1} \to 0. \tag{3.10}$$

The significance of this reduction, while not at first obvious, relies in the fact that the vector w_i is now independent from the matrix X_{n-1}^i, which greatly simplifies the calculations that follow.

Let $Y_{n-1}^i := (X_{n-1}^i - zI)^{-1}$. Using $w_i(k)$ to denote the kth entry of the column vector w_i, we can write

$$\begin{aligned}\overline{\epsilon}_n^i &= w_i^* Y_{n-1}^i w_i - \frac{1}{n}\mathrm{tr}\ Y_{n-1}^i \qquad (3.11) \\ &= \frac{1}{n}\sum_{k=1}^{n-1}((\sqrt{n}w_i(k))^2 - 1)Y_{n-1}^i(k,k) + \sum_{\substack{k,k'=1 \\ k\neq k'}}^{n-1} w_i(k)w_i(k')Y_{n-1}^i(k,k'). \end{aligned}$$

First, note that $\mathbf{E}\overline{\epsilon}_n^i = \mathbf{E}(\mathbf{E}(\overline{\epsilon}_n^i|X_{n-1}))$, by the independence of w_i from X_{n-1} and the fact that each $w_i(k)$ has mean 0 and variance $1/n$. Now, denote the two terms on the right hand side of (3.11) by ϵ_1 and ϵ_2.

If $\lambda_1^i, \ldots, \lambda_{n-1}^i$ are the eigenvalues of Y_{n-1}^i, then

$$\left|\frac{1}{n}\mathrm{tr}(Y_{n-1}^i)^2\right| \leq \frac{1}{n}\sum_{k=1}^{n-1}\frac{1}{|(\lambda_k^i - z)^2|} \leq \frac{1}{b^2},$$

and together with the hypothesis that the entries $\sqrt{n}w_i(k)$ are uniformly bounded by C, we deduce

$$\mathbf{E}(\epsilon_1^2) \leq \frac{1}{n^2}\sum_{k=1}^{n-1}\mathbf{E}\left|\left((\sqrt{n}w_i(k)^2 - 1\right)^2 Y_{n-1}^i(k,k)^2\right| \leq \frac{C_1}{n^2},$$

for some constant C_1 depending on C and b.

Similarly,

$$\mathbf{E}(\epsilon_2^2) = \sum_{\substack{k,k'=1\\k\neq k'}}^{n-1} w_i(k)^2 w_i(k')^2 Y_{n-1}^i(k,k')^2 \leq \frac{C^4}{n^2}\sum_{\substack{k,k'=1\\k\neq k'}}^{n-1} Y_{n-1}^i(k,k')^2 \leq \frac{C_2}{n^2},$$

where C_2 is again a constant depending on C and b.

Now, for $\delta > 0$ fixed,

$$P\left(\sup_{i\leq n}|\overline{\epsilon}_n^i| > \delta\right) \leq \sum_{i=1}^{n} P(|\overline{\epsilon}_n^i| > \delta).$$

Since $\overline{\epsilon}_n^i$ has expectation 0 for every i, using Chebyshev's inequality followed by the standard inequality $(x+y)^2 \leq 2(x^2+y^2)$, we have

$$P\left(\sup_{i\leq n}|\overline{\epsilon}_n^i| > \delta\right) \leq \frac{1}{\delta^2}\sum_{i=1}^{n}\mathbf{E}|\overline{\epsilon}_n^i|^2 \leq \frac{2}{\delta^2}\sum_{i=1}^{n}\left(\mathbf{E}|\epsilon_1|^2 + \mathbf{E}|\epsilon_2|^2\right) \leq \frac{2n}{\delta^2}\left(\frac{C_1}{n^2} + \frac{C_2}{n^2}\right),$$

which goes to 0 as $n \to \infty$. Hence $\sup_{i\leq n}\overline{\epsilon}_n^i$ converges to 0 in probability, and consequently so does δ_n. Equivalently, we have shown that

$$s_n(z) + \frac{1}{z + s_n(z)} \to 0 \tag{3.12}$$

as $n \to \infty$, in probability.

At this point, $s_n(z)$ is still a random variable, so passing to the limit $n \to \infty$ becomes a subtle issue. To help with this, a simple application of McDiarmid's

inequality shows that $s_n(z) - \mathbf{E}s_n(z) \to 0$ in probability, as $n \to \infty$. Then, letting $\rho_n = s_n(z) - \mathbf{E}s_n(z)$, we have

$$\left| \frac{1}{z + s_n(z)} - \frac{1}{z + \mathbf{E}s_n(z)} \right| = \frac{|\rho_n|}{|z + \mathbf{E}s_n(z)| \cdot |z + s_n(z)|} \leq \frac{|\rho_n|}{b^2},$$

since the imaginary part of $s_n(z)$ (and consequently $\mathbf{E}s_n(z)$ also) is always the same as the imaginary part of z. Thus, by taking expectation in (3.12) we see that

$$\mathbf{E}s_n(z) + \frac{1}{z + \mathbf{E}s_n(z)} \to 0$$

deterministically as $n \to \infty$. Since $|s_n(z)|$ is bounded by $1/b$, we conclude that for fixed z, $\mathbf{E}s_n(z)$ has a convergent subsequence, whose limit $s(z)$ necessarily satisfies

$$s(z) + \frac{1}{z + s(z)} = 0 \Leftrightarrow s(z) = \frac{-z \pm \sqrt{z^2 - 4}}{2}.$$

To choose the correct branch of the square root, recall that the imaginary part of $s(z)$ has the same sign as the imaginary part of z, which gives

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

Therefore, $s_n(z)$ converges pointwise to $s(z)$. From the inversion formula in proposition 3.3.1, we deduce that the density corresponding to this Stieltjes transform is indeed the semicircle law, as

$$\lim_{b \to 0^+} \frac{s(x + ib) - s(x - ib)}{2\pi i} = \frac{1}{2\pi} \sqrt{4 - x^2} = \sigma(x).$$

Finally, from proposition 3.3.2 it follows that the ESD μ_n converges in probability to the semicircle law, as desired. ∎

The first thing to note about the Stieltjes transform method is that, unlike the moment method in the previous section, it provides a *constructive* proof of the semicircle law. In fact, Stieltjes transforms are also useful when proving asymptotic results via the moment method. Once the moments m_k of the limiting density are known, one can form the formal generating function $g(z) = \sum_{k=0}^{\infty} m_k z^k$, which is related to the Stieltjes transform corresponding to the same density by the equality $s(z) = -g(1/z)/z$. Using the inversion formula 3.3.1, the density can thus be inferred from the moments.

Although proving the semicircle law with the Stieltjes method led to a simple, quadratic equation in $s(z)$, for other results the situation can get more complicated and may involve, for instance, differential equations [17]. On the other hand, in cases when the asymptotic density cannot be derived in closed form, analytical methods that use the Stieltjes transform may provide numerical approximations, which are often sufficient to make the result useful in applications.

3.4 Free probability

The theory of free probability, originally developed by Voiculescu in the context of functional analysis [18], provides a justification of the semicircle law that is much deeper than the previous two proofs. Although this result holds to such universality – essentially for all symmetric or Hermitian matrices with sufficiently well-behaved moments – it is unclear from either the moments or the Stieltjes transform why the semicircular density is special in this way. In this section, we describe how Wigner's semicircle law becomes an analogue of the central limit theorem for a different kind of probability theory. Our discussion is based on [13] and, to a smaller extent, on [16].

Classical probability is built on three fundamental objects: a sample space Ω describing the set of possible outcomes of an experiment, a σ-algebra $\mathcal{F}$ of events, and a probability measure P which assigns a number between 0 and 1 to each element of $\mathcal{F}$ in a consistent manner. In this framework, one then defines random variables as functions from Ω to $\mathbb{R}$ or $\mathbb{C}$, to project what could be an unwieldy space Ω to the more accessible real and complex numbers. Finally, random variables are assigned expectations, which capture the "typical" behaviour of those random variables.

The essence of free probability relies in abstracting away the sample space, σ-algebra, and probability measure, and instead focusing on the *algebra of random variables*, along with their expectations. The main advantage of this approach is that it gives rise to the study of *non-commutative* probability theory. In particular, this enables the study of random matrices as stand-alone entities, without the need to look at the individual entries to get probabilistic-type results.

Recall that a $*$-ring over $\mathbb{C}$ is an associative ring R equipped with a conjugation operation $* : R \times R$ having the following properties:

(a) $(x^*)^* = x$, $\forall x \in R$.

(b) $(x + y)^* = x^* + y^*$, $\forall x, y \in R$.

(c) $(xy)^* = y^* x^*$, $\forall x, y, \in R$.

(d) $(cx)^* = \overline{c} x^*$, $\forall c \in \mathbb{C}, x \in R$.

Thus, $*$ is an involution which preserves addition, reverses multiplication, and is anti-homogeneous. If $X \in \mathcal{A}$ and $X^* = X$, we say that X is *self-adjoint*. If $X^* X = X X^*$, we say that X is *normal*.

A $*$-algebra is a $*$-ring A which also has the structure of an associative algebra over $\mathbb{C}$, where the restriction of $*$ to $\mathbb{C}$ is usual complex conjugation.

Definition 3.4.1. A non-commutative probability space $(\mathcal{A}, \tau)$ consists of a $*$-algebra $\mathcal{A}$ with identity 1 and a trace operator $\tau : \mathcal{A} \to \mathbb{C}$ which is $*$-linear, maps 1 to 1, and is non-negative, in the sense that $\tau(X^*X) \geq 0$, $\forall X \in \mathcal{A}$.

A first example of a noncommutative probability space is provided by classical probability theory.

Example 3.4.2. Let $\mathcal{A}$ consist of all complex-valued random variables defined on some sample space Ω with all moments finite. The ring operations on $\mathcal{A}$ correspond to the usual addition and multiplication of functions, and the unit of the ring is the deterministic function mapping everything to 1. The involution $*$ corresponds to complex conjugation, and the trace τ is just the expectation operator.

Interestingly, this new framework is also appropriate for the study of spectral theory, which deals with *deterministic* matrices. This is also a first example that is specifically *not* commutative.

Example 3.4.3. Consider the $*$-algebra $M_n(\mathbb{C})$ of $n \times n$ matrices with complex-valued entries, with unity given by the identity matrix I_n, where the $*$ operator is given by taking the conjugate transpose. The operator τ is given by taking the *normalised* trace, $\tau(X) = \mathbf{E}X/n$. Note that the normalisation factor is necessary due to the condition that τ maps 1 to 1.

Lastly, we come to our main area of interest, random matrices.

Example 3.4.4. We can represent the algebra of $n \times n$ random matrices as the tensor product space $L^{\infty-} \otimes M_n(\mathbb{C})$. This notation specifies a random matrix of size n by indicating a random variable for each of the n^2 entries. This is a $*$-algebra with identity I_n, where the conjugate X^* of some matrix X is the conjugate transpose of X. As suggested by the two previous example, a natural choice for τ is the normalised expected trace, $\tau(X) = \mathbf{E}\ \mathrm{tr}X/n$.

Note that all three examples satisfy

$$\tau(XY) = \tau(YX),\ \forall X, Y \in \mathcal{A}.$$

Equivalently, the trace operator is invariant under cyclic permutations. This additional axiom is needed for certain generalizations to the results discussed below, but it is not necessary for our discussion.

So far, the basic definition of a non-commutative probability space gives a natural generalization of the *moments* of a random variable. Specifically, for

an element X of a non-commutative probability space $(\mathcal{A}, \tau)$, the kth moment is given by $\tau(X^k)$. In particular, because $X^k \in \mathcal{A}$ and τ is defined on all of $\mathcal{A}$, these moments are all finite.

Now, we would like to expand this generalized framework to include the usual definition of the moments, given in terms of the probability density μ_X:

$$\tau(X^k) = \int_{\mathbb{C}} z^k d\mu_X(z).$$

Equivalently, by linearity of τ, we would like to find a complex-valued measure μ_X such that for any polynomial $P : \mathbb{C} \to \mathbb{C}$

$$\tau(P(X)) = \int_{\mathbb{C}} P(z) d\mu_X(z). \tag{3.13}$$

In what follows, assume that X is self-adjoint, so that the measure μ_X is supported on the real line. The key idea is to use the Stieltjes transform, defined previously as

$$s_X(z) = \mathrm{tr}((X - z)^{-1}) = \int_{\mathbb{C}} \frac{1}{x - z} d\mu_X(z),$$

to bridge the moments of X that we already know in this non-commutative setting to a measure μ_X obeying (3.13).

In the classical probability setting, the Stieltjes transform is defined everywhere *outside* of the support of the density μ. In this case, we would like to go backwards, first extending analytically the Stieltjes transform to the largest possible region of the complex plane, and then inferring the density μ to be supported on the complement of that region.

To start, define the Stieltjes transform of some element X of a non-commutative probability space $(\mathcal{A}, \tau)$ as

$$s_X(z) := \tau((X - z)^{-1}). \tag{3.14}$$

Expanding the right hand side formally as a Laurent series, we see that

$$s_X(z) = -\sum_{k=0}^{\infty} \frac{\tau(X^k)}{z^{k+1}}. \tag{3.15}$$

In order to establish where this formal series converges, we need a better understanding of the growth of the moments $\tau(X^k)$ with k. The conjugation operation $*$ and the trace τ give an elegant inner-product space structure on $\mathcal{A}$, via the positive semi-definite inner product

$$\langle X, Y \rangle := \tau(X^* Y).$$

This could be made into a positive-definite inner product with the additional *faithfulness* axiom $\tau(X^*X) = 0$ if and only if $X = 0$, though in general this is not needed. Now, each $X \in \mathcal{A}$ has an associated norm

$$||X|| = (\langle X, X\rangle)^{1/2} = \tau(X^*X)^{1/2},$$

satisfying the Cauchy-Schwartz inequality

$$|\langle X, Y\rangle| \leq ||X||\cdot||Y||. \tag{3.16}$$

One can easily show inductively using the Cauchy-Schwartz inequality that any self-adjoint element X satisfies

$$|\tau(X^{2k-1})|^{1/(2k-1)} \leq |\tau(X^{2k})|^{1/2k} \leq |\tau(X^{2k+2})|^{1/(2k+2)},$$

for all $k \geq 0$. In particular, this gives full monotonicity on the even moments, which means that the limit

$$\rho(X) := \lim_{k\to\infty} |\tau(X^{2k})|^{1/2k} \tag{3.17}$$

exists. The real number $\rho(X)$ is called the *spectral radius* of X. Furthermore, we see that

$$|\tau(X^k)| \leq \rho(X)^k \tag{3.18}$$

for any k, which by (3.15) immediately implies that the Stieltjes transform $s_X(z)$ exists for $|z| > \rho(X)$.

With some more work, the Stieltjes transform can be analytically extended to part of the region $|z| \leq \rho(X)$. To prove this, we need the following:

Lemma 3.4.5. *(a) Let X be self-adjoint and bounded. For any $R \in \mathbb{R}$,*

$$\rho(R^2 + X^2) = R^2 + \rho(X)^2. \tag{3.19}$$

(b) Let X be normal and bounded. Then:

$$|\tau(X^k)| \leq \tau((X^*X)^k)^{1/2} \leq \rho(X^*X)^{k/2}. \tag{3.20}$$

Proof. (a) Without loss of generality, let $R \geq 0$. For every $k \in \mathbb{N}$, (3.18) gives:

$$\begin{aligned}\rho(R^2 + X^2) &\geq \left|\tau((R^2 + X^2)^{2k})\right|^{1/2k} \\ &= \left|R^{4k} + \sum_{l=1}^{2k-1} \binom{2k}{l} R^{2k-l}\tau(X^{2l}) + \tau(X^{4k})\right|^{1/2k}.\end{aligned}$$

Because X is self-adjoint and τ is a nonnegative trace operator, we see that $\tau(X^{2l}) = \tau((X^l)^*X^l)) \geq 0$, and thus

$$\rho(R^2 + X^2) \geq (R^{4k})^{1/2k} + \left(\tau(X^{4k})\right)^{1/2k} = R^2 + \left[\left(\tau(X^{4k})\right)^{1/4k}\right]^2.$$

With $k \to \infty$, the above implies $\rho(R^2 + X^2) \geq R^2 + \rho(X)^2$.

For the reverse inequality, use (3.18) again:

$$\begin{aligned} \tau((R^2 + X^2)^{2k}) &= \sum_{l=0}^{2k} \binom{2k}{l} R^{2(2k-l)} \tau(X^{2l}) \\ &\leq \sum_{l=0}^{2k} \binom{2k}{l} R^{2(2k-l)} \rho(X)^{2l} = (R^2 + \rho(X)^2)^{2k}. \end{aligned}$$

Raising everything to the $1/2k$ power and letting $k \to \infty$, it follows that $\rho(R^2 + X^2) \leq R^2 + \rho(X)^2$.

Therefore, $\rho(R^2 + X^2) = R^2 + \rho(X)^2$, as desired.

(b) If X is normal, then X^*X is self-adjoint, so the second inequality follows directly from (3.18). The first inequality is a direct application of the Cauchy-Schwartz inequality (3.16) with $X := X^k$, $Y = 1$. ∎

The above lemma allows us to establish where the Stieltjes transform converges if the imaginary part of z changes. Specifically, writing

$$(X - z)^{-1} = ((X + iR) - (z + iR))^{-1},$$

$s_X(z)$ can be written as a formal Laurent series as follows:

$$s_X(z) = -\sum_{k=0}^{\infty} \frac{\tau((X + iR)^k)}{(z + iR)^{k+1}}. \tag{3.21}$$

From lemma 3.4.5, because X is self-adjoint we have $|\tau((X + iR)^k)| \leq (R^2 + \rho(X)^2)^{k/2}$. Thus, the Laurent series (3.21) converges for $|z + iR| > (R^2 + \rho(X)^2)^{1/2}$. Furthermore, because Laurent expansions are unique whenever they exist, this expansion must agree with (3.15) for z large enough so that $|z| > \rho(X)$.

Now, if $z = a + ib$, the condition $|z + iR| > (R^2 + \rho(X)^2)^{1/2}$ is equivalent to $a^2 + (b + R)^2 > R^2 + \rho(X)^2 \Leftrightarrow a^2 + b^2 + 2bR > \rho(X)^2$, which becomes valid for R large enough as long as $b > 0$. Hence, we have analytically extended the Stieltjes transform $s_X(z)$ to the entire upper-half plane. Similarly, $s_X(z)$ can be defined for any z in the lower-half plane. Recall from earlier that $s_X(z)$ also exists when z is real with $|z| > \rho(X)$.

As it turns out, this is the maximal region on which $s_X(z)$ can be defined. Indeed, suppose $\exists\, 0 < \epsilon < \rho(X)$ such that $s_X(z)$ exists on the region $\{z : |z| >$

$\rho(X) - \epsilon\}$. Let $R > \rho(X) - \epsilon$ and consider the contour $\gamma = \{z : |z| = R\}$. From the residue theorem applied at infinity,

$$\tau(X^k) = -\frac{1}{2\pi i} \int_\gamma s_X(z) z^k dz,$$

which by (3.18) implies

$$|\tau(X^k)| \leq \frac{1}{2\pi} \int_\gamma \left(\frac{\rho(X)^k}{R} + \frac{\rho(X)^{k+1}}{R^2} + \dots \right) dz.$$

On raising both sides to the $1/k$ power, in order for this to hold in the limit $k \to \infty$, it must be the case that $\rho(X) \leq R$. Since R can be chosen arbitrarily close to 0 by taking ϵ arbitrarily close to $\rho(X)$, the previous inequality implies $\rho(X) = 0$, which is an obvious contradiction assuming $X \neq 0$.

Therefore, the Stieltjes transform $s_X(z)$ is defined and analytic everywhere outside the interval $[-\rho(X), \rho(X)]$. With this, we can deduce the existence of a measure μ_X supported on $[-\rho(X), \rho(X)]$, which gives the usual way for computing the moments of X, as stated in (3.13).

Theorem 3.4.6. *Let X be a bounded, self-adjoint element of a non-commutative probability space $(\mathcal{A}, \tau)$. Then there exists a unique Borel probability measure on $[-\rho(X), \rho(X)]$ such that for any polynomial $P : \mathbb{C} \to \mathbb{C}$,*

$$\tau(P(X)) = \int_{\mathbb{C}} P(z) d\mu_X(z).$$

To prove this theorem, we rely on the following result:

Proposition 3.4.7. *Let X be bounded, self-adjoint. Then for any polynomial $P : \mathbb{C} \to \mathbb{C}$,*

$$|\tau(P(X))| \leq \sup_{x \in [-\rho(X), \rho(X)]} |P(x)|$$

Loosely speaking, the left hand side gives the average of $P(X)$, which should be smaller than the maximum value that $P(x)$ takes on the domain on which the non-commutative random variable X is distributed. The proposition is proven by first noting that the Stieltjes transform of $P(X)$ is defined outside of $[-\rho(P(X)), \rho(P(X))]$ and, furthermore, it cannot be extended inside the interval; upon showing that the Stieltjes transform of $P(X)$ exists on the region $\Omega = \{z \in \mathbb{C} : z > \sup_{x \in [-\rho(X), \rho(X)]} |P(x)|\}$, it becomes clear that Ω is necessarily contained in $\mathbb{C} - [-\rho(P(X)), \rho(P(X))]$, and the desired conclusion follows. For the specific details, we direct the reader to [13].

Proof of Theorem 3.4.6. Consider the linear functional ϕ sending polynomials $P : \mathbb{C} \to \mathbb{C}$ to $\tau(P(X))$. By the Weierstrass approximation theorem, for every continuous $f : \mathbb{C} \to \mathbb{C}$ which is compactly supported on $[-\rho(X), \rho(X)]$, and for every $\epsilon > 0$, there exists a polynomial P_ϵ such that $|f(x) - P_\epsilon(x)| < \epsilon$ $\forall x \in [-\rho(X), \rho(X)]$. Thus, we can assign a value to $\tau(f(X))$ by taking the limit of $\tau(P_\epsilon(X))$ as $\epsilon \to 0$. This means that ϕ can be extended to a continuous linear functional on the space $C([-\rho(X), \rho(X)]$. From the Riesz representation theorem, there exists a unique countably additive, regular measure μ_X on $[-\rho(X), \rho(X)]$ such that

$$\phi(f) = \int_{\mathbb{C}} f(x) d\mu_X(x).$$

Furthermore, the total variation of μ_X equals the operator norm of ϕ, which is equal to 1 by proposition 3.4.7. With the additional observation that $d\mu_X$ integrates to 1 (since $\phi(1) = 1$), we conclude that μ_X is the desired probability measure. ∎

This theorem also recovers the familiar definition of the Stieltjes transform from classical probability theory:

Corollary 3.4.8. *Let X be a bounded, self-adjoint element of a non-commutative probability space with spectral measure μ_X. With the Stieltjes transform of X defined as in (3.14), we have:*

$$s_X(z) = \int_{-\rho(X)}^{\rho(X)} \frac{1}{x - z} d\mu_X(z),$$

for all $z \in \mathbb{C} \backslash [-\rho(X), \rho(X)]$.

Proof. When $|z| > \rho(X)$, we can write $s_X(z)$ as a convergent Laurent series, and then express the moments as integrals over the spectral measure μ_X:

$$\begin{aligned} s_X(z) &= -\sum_{k=0}^{\infty} \frac{\tau(X^k)}{z^{k+1}} = -\sum_{k=0}^{\infty} \frac{1}{z^{k+1}} \int_{\mathbb{C}} x^k d\mu_X(x) = -\int_{-\rho(X)}^{\rho(X)} \frac{1}{z} \sum_{k=0}^{\infty} \left(\frac{x}{z}\right)^k d\mu_X(x) \\ &= \int_{-\rho(X)}^{\rho(X)} \frac{1}{z(x/z - 1)} d\mu_X(x) = \int_{-\rho(X)}^{\rho(X)} \frac{1}{x - z} d\mu_X(z), \end{aligned}$$

as desired. Note that we could switch the sum and integral signs because $\sum_{k=0}^{\infty} |(x/z)^k|$ is bounded by the convergent series $\sum_{k=0}^{\infty} (\rho(X)/z)^k$, which does not depend on x. ∎

To summarize, we started by defining a probability space whose basic objects are random variables and their expectations, rather than σ-algebras and

measures. In this framework, we were able to recreate much of the classical probability theory, including Stieltjes transforms and probability measures. In doing so, no commutativity was assumed between random variables, which is what ultimately keeps free probability theory separate from classical probability.

The asymptotic distribution of spectra of various classes of matrices is determined, for the most part, by the patterns of independence among the matrix entries. For instance, what determines the convergence of Wigner matrix spectra to the semicircle law is the symmetry (or Hermitian) condition and the fact that all upper-triangular entries are independent, more so than the individual moments of these random variables. It is therefore not surprising that the concept of independence in the context of free probability is what governs the asymptotic behaviour of non-commutative random variables.

Definition 3.4.9. The random variables $X_1, \ldots, X_k$ of some non-commutative probability space $(\mathcal{A}, \tau)$ are said to be *freely independent* (or *free*) if for every $m \in \mathbb{N}$ and $i_1, \ldots, i_m \in \{1, \ldots, k\}$ with no two consecutive indices equal, we have

$$\tau((P_1(X_{i_1}) - \tau(P_1(X_{i_1}))) \cdots (P_m(X_{i_m}) - \tau(P_m(X_{i_m})))) = 0,$$

where $P_1, \ldots, P_m$ are polynomials.

An equally important notion is that of *asymptotic freeness*:

Definition 3.4.10. A sequence of random variables $X_{1,n}, \ldots, X_{k,n}$, $n \geq 1$ in some non-commutative probability space $(\mathcal{A}_n, \tau_n)$ is said to be *asymptotically freely independent* (or *asymptotically free*), if

$$\tau((P_1(X_{n,i_1}) - \tau(P_1(X_{n,i_1}))) \cdots (P_m(X_{n,i_m}) - \tau(P_m(X_{n,i_m})))) \to 0$$

as $n \to \infty$, where $i_1, \ldots, i_m$ and $P_1, \ldots, P_m$ are as in the previous definition.

Conceptually, free independence is very similar to classical independence, in the sense that they both require the expectation of a product of mean zero random variables to be equal to zero. However, because one framework requires commutativity and the other does not, the two definitions tend to be rather different in practice. In particular, random variables in a non-commutative probability space that commute in the classical sense are rarely freely independent.

Example 3.4.11. Suppose X and Y are two *diagonal matrices* with mean zero, independent entries along their respective diagonals. Observe that X and Y commute as matrices, so we have, for instance, $\tau(XYXY) = \tau(X^2Y^2) = \tau(X^2)\tau(Y^2)$, which in most cases is nonzero.

Nonetheless, classical and free independence are deeply connected, as best illustrated in the case of Wigner matrices:

Theorem 3.4.12. *Consider k independent $n \times n$ Wigner matrices $M_1, \ldots, M_k$ whose entries have uniformly bounded moments of all orders. Then the rescaled matrices $X_j = M_j/\sqrt{n}$, $j = 1, \ldots, m$, are asymptotically free with respect to the trace operator $\tau = \frac{1}{n}\mathbf{E}\mathrm{tr}$.*

Proof. By linearity of τ and the definition of asymptotic freeness, it suffices to show that

$$\tau\left(\prod_{j=1}^{m}\left(X_{k_j}^{a_j} - \tau(X_{k_j}^{a_j})\right)\right) = o(1) \Leftrightarrow \tau\left(\prod_{j=1}^{m} X_{k_j}^{a_j}\right) - \prod_{j=1}^{m}\tau\left(X_{k_j}^{a_j}\right) = o(1),$$

where $a_1, \ldots, a_m$ are positive integers and $1 \leq k_1, \ldots, k_m \leq k$ such that consecutive k_js are distinct. The little-o notation indicates that the expression above goes to 0 as $n \to \infty$.

Again, this will be shown using a combinatorial approach. We have

$$\tau\left(\prod_{j=1}^{m} X_{k_j}^{a_j}\right) = \frac{1}{\sqrt{n}^{a_1+\ldots+a_m}} \sum_{\mathbf{i_1},\ldots,\mathbf{i_m}} \frac{1}{n}\mathbf{E}\prod_{j=1}^{m}\zeta(i_{1,j}, i_{2,j})\cdots\zeta(i_{a_j,j}, i_{1,j+1}), \quad (3.22)$$

with $\mathbf{i_j} = (i_{1,j}, \ldots, i_{a_j,j})$ such that each coordinate takes values in the set $\{1, \ldots, n\}$, $i_{1,m+1} := i_{1,1}$ so that the first and last indices in the product above match, and with $\zeta(i_{s,j}, i_{s+1,j})$ denoting the entry of the matrix M_{k_j} situated in row i_s and column i_{s+1}.

Now, each term in (3.22) can be thought of as a *connected, closed path* described as

$$(i_{1,1}i_{2,1}\cdots i_{a_1,1})\,(i_{1,2}i_{2,2}\cdots i_{a_2,2})\cdots(i_{1,m}i_{2,m}\cdots i_{a_m,m}i_{1,1})\,,$$

where the first a_1 edges are labelled k_1, the next a_2 are labelled k_2, and so on. The labels are important because when dealing with multiple matrices, it is necessary to know not just the row and column indices of the entries to be considered, but also which matrix those entries come from.

As before, due to independence and the fact that the entries of $M_1, \ldots, M_k$ have mean 0, we conclude that within each subpath of a_j edges, each edge needs to be traversed at least twice, and thus the corresponding number of vertices is at most $a_j/2 + 1$. If t_j denotes the *weight* of $\mathbf{i_j}$ – i.e., the number of distinct components of $\mathbf{i_j}$ – we see that $t_j \leq a_j/2 + 1$.

Now, let $t_1, \ldots, t_m$ be fixed. There are

$$\prod_{j=1}^{m} n(n-1)\cdots(n-t_j+1) = \mathcal{O}(n^{t_1+\ldots+t_m})$$

choices of $\mathbf{i_1}, \ldots, \mathbf{i_m}$ corresponding to the chosen weights. If $t_j \leq a_j/2$ for all j, the contribution of all these terms to the sum (3.22) is $\mathcal{O}(1/n) = o(1)$. Again, it is crucial that the number of choices for the weights $t_1, \ldots, t_m$ depends on $a_1, \ldots, a_m$ but not on n, so the combined contribution of all terms with $t_j \leq a_j/2$ for all j is asymptotically negligible.

Thus, suppose $t_j = a_j/2 + 1$ for some j. Note, in particular, that a_j must be even. Because each of the $a_j/2$ distinct edges needs to be traversed exactly twice, we see that the corresponding subpath of a_j edges is closed – i.e., $i_{1,j+1} = i_{1,j}$ – and the distinct edges form a tree. From the moment method proof of theorem 3.1.1, the contribution of all such terms to $\tau(X_{k_1}^{a_1} \cdots X_{k_m}^{k_m})$ is a factor of $C_{a_j/2}$.

In fact, by pulling out this factor, the remaining terms form another closed path of $(a_1 + \ldots + a_m) - a_j$ labelled edges, and the same argument can be applied to conclude that

$$\tau\left(\prod_{j=1}^{m} X_{k_j}^{a_j}\right) = C_{a_1/2} \cdots C_{a_m/2} + o(1),$$

where the first term is included only if $a_1, a_2, \ldots, a_m$ are all even.

Similar, though simpler reasoning shows that when $a_1, \ldots, a_m$ are all even,

$$\prod_{j=1}^{m} \tau\left(X_{k_j}^{a_j}\right) = \big(C_{a_1/2} + o(1)\big) \cdots \big(C_{a_m/2} + o(1)\big) = C_{a_1/2} \cdots C_{a_m/2} + o(1),$$

and thus

$$\tau\left(\prod_{j=1}^{m} X_{k_j}^{a_j}\right) - \prod_{j=1}^{m} \tau\left(X_{k_j}^{a_j}\right) = o(1),$$

implying that the rescaled Wigner matrices $X_1, \ldots, X_m$ are asymptotically free. ∎

Of interest at this point are sums of freely independent random variables. Just as the characteristic function or the moment generating function are used to calculate convolutions of commutative, scalar random variables, free probability possesses an analogous transform which is additive over free random variables. Specifically, given a random variable X consider its Stieltjes transform $s_X(z)$ with functional inverse written as $z_X(s)$. The *R-transform* of X is defined as

$$R_X(s) := z_x(-s) - s^{-1},$$

and has the property

$$R_{X+Y} = R_X + R_Y$$

whenever X and Y are freely independent.

The R-transform clarifies the special role played by the semicircular density:

Lemma 3.4.13. *If σ has the semicircular density (3.1), then $R_\sigma(s) = s$. Furthermore, if u and v are two freely independent random variables with the semicircle density, their convolution is given by $\sqrt{2}u$.*

The proof is a straightforward calculation, and for this reason it will be omitted.

Expressing the R-transform $R_X(s)$ as a power series in s reveals an analogy with the cumulant generating function from classical probablity theory:

Proposition 3.4.14 ([13])**.** *For a non-commutative random variable X, write $R_X(s) = \sum_{k=1}^{\infty} C_k(X)s^{k-1}$, with the coefficients $C_k(X)$ given recursively by*

$$C_k(X) = \tau(X^k) - \sum_{j=1}^{k-1} C_j(X) \sum_{a_1+\ldots+a_j=k-j} \tau(X^{a_1}) \cdots \tau(X^{a_j}).$$

This is sufficient to establish the *free central limit theorem*:

Theorem 3.4.15. *Let X_1, X_2, ... be free copies of a self-adjoint random variable X in a non-commutative probability space, such that $\tau(X) = 0$ and $\tau(X^2) = 1$. For each n, define $S_n := (X_1 + \ldots + X_n)/\sqrt{n}$. Then S_n converges in the sense of moments to an element having the semicircular density σ.*

Proof. Using the additivity of the R-transform, we have

$$R_{S_n}(s) = nR_{X_1/\sqrt{n}}(s).$$

Since X_1 is an element of a non-commutative probability space, it has finite moments of all orders. By proposition 3.4.14, we have

$$C_k\left(\frac{X_1}{\sqrt{n}}\right) = \tau\left(\frac{X_1^k}{n^{k/2}}\right) - \sum_{j=1}^{k-1} C_j\left(\frac{X_1}{\sqrt{n}}\right) \sum_{a_1+\ldots+a_k=k-j} \tau\left(\frac{X_1^{a_1}}{n^{a_1/2}}\right) \cdots \tau\left(\frac{X_1^{a_k}}{n^{a_k/2}}\right),$$

which shows by induction that $nC_k\left(\frac{X_1}{\sqrt{n}}\right) \to 0$ as $n \to \infty$ for $k \geq 3$. Computing C_1, C_2, C_3, we deduce that $R_{S_n}(s) \to s$ as $n \to \infty$ in the sense of moments, which leads to the desired conclusion. ∎

This result provides a quick heuristic proof of the semicircle law in the case of GUE Wigner matrices. Let M_n and M_n' be two classically independent

matrices from the GUE ensemble. Because the entries are Gaussians, we have $M_n + M'_n \sim \sqrt{2}M_n$, and passing to the limit shows that the only possible asymptotic distribution of the ESD of M_n. One can then extend this result to arbitrary Wigner matrices by using a variation on the Lindeberg replacement trick to substitute the Gaussian entries of M_n, one by one, with arbitrary distributions [13].

4 Eigenvalue distribution of Wishart matrices: the Marcenko-Pastur law

4.1 Introduction

Let $(m_n)_{n\geq 1}$ be a sequence of positive integers such that $\lim_{n\to\infty} \frac{m_n}{n} = \alpha \geq 1$. Consider the $n \times m_n$ matrix X_n whose entries are i.i.d. of mean 0 and variance 1, and with the kth moment bounded by some $r_k < \infty$ not depending on n. As before, we will actually study the normalized matrix $Y_n := X_n/\sqrt{n}$.

The Marcenko-Pastur law is concerned with the distribution of the *singular values* of Y_n, which by definition are the eigenvalues of the $n \times n$ *Wishart matrix* $W_n = Y_n Y_n^T$. As with the semicircle law, the limiting behaviour of these eigenvalues can be understood by considering the empirical spectral distribution μ_n of a Wishart matrix W_n as $n \to \infty$.

Theorem 4.1.1. *The empirical law μ_n converges weakly, in probability, to the distribution with density f_α supported on $[\lambda_-, \lambda_+]$, where $\lambda_- = (1-\sqrt{\alpha})^2$ and $\lambda_+ = (1+\sqrt{\alpha})^2$, and given by*

$$f_\alpha(x) = \frac{\sqrt{(x-\lambda_-)(\lambda_+ - x)}}{2\pi x} \mathbf{1}_{x\in[\lambda_-,\lambda_+]}. \tag{4.1}$$

4.2 The moment method

As in the case of the semicircle law, the most straightforward method to prove Marcenko-Pastur uses the simple observation that the kth empirical moment $\langle \mu_n, x^k \rangle = \int_{\mathbb{R}} x^k d\mu_n$ is equal to $\frac{1}{n}\text{tr } W_n^k$ [1]. Actually, it suffices to consider the *expected* empirical moments $\langle \overline{\mu}_n, x^k \rangle = \int_{\mathbb{R}} x^k d\overline{\mu}_n = \frac{1}{n}\mathbf{E}\text{tr } W_n^k$, due to the following result:

Lemma 4.2.1. *For every fixed $k \in \mathbb{N}$ and $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|\langle \mu_n, x^k \rangle - \langle \overline{\mu}_n, x^k \rangle| > \epsilon) = 0.$$

Therefore, it suffices to prove that the expected empirical moments of μ_n converge to the moments of the Marcenko-Pastur law (4.1). This is important because working with $\overline{\mu}_n = \mathbf{E}\mu_n$ enables us to take full advantage of the fact that the entries of the original matrix X_n are i.i.d., and reduces the computation of $\frac{1}{n}\mathbf{E}\text{tr } W_n^k$ in the limit $n \to \infty$ to a combinatorial argument.

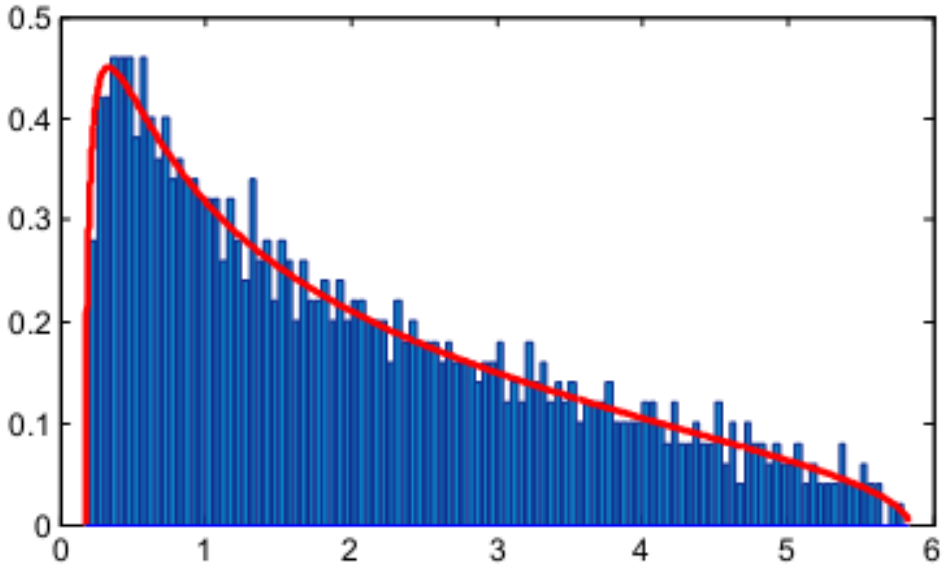

Figure 2: Simulation of the Marcenko-Pastur law with $\alpha = 2$ using 100 samples of the eigenvalues of 1000 by 1000 matrices. Bin size is 0.05.

Proof of Theorem 4.1.1. From the usual rules of matrix multiplication, we see that

$$\begin{aligned} \langle \overline{\mu}_n, x^k \rangle &= \frac{1}{n} \mathbf{E} \text{ tr } W_n^k = \frac{1}{n} \mathbf{E} \text{ tr}(Y_n Y_n^T)^k \qquad (4.2) \\ &= \frac{1}{n} \sum_{\substack{i_1,\ldots,i_k \\ j_1,\ldots,j_k}} \mathbf{E}\, Y_n(i_1,j_1) Y_n(i_2,j_1) Y_n(i_2,j_2) \cdots Y_n(i_k,j_k) Y_n(i_1,i_k), \end{aligned}$$

where the row indices $i_1, \ldots, i_k$ take values in $\{1, \ldots, n\}$ and the column indices $j_1, \ldots, j_k$ take values in $\{1, \ldots, m_n\}$.

Because the entries of Y_n are independent, each factor in the product

$$Y_n(i_1,j_1) Y_n(i_2,j_1) Y_n(i_2,j_2) Y_n(i_3,j_2) \cdots Y_n(i_k,j_k) Y_n(i_1,i_k)$$

must appear at least twice for the expectation to be nonzero. We can think of each such product as a connected *bipartite* graph on the sets of vertices $\{i_1, \ldots, i_k\}$ and $\{j_1, \ldots, j_k\}$, where the total number of edges (with repetitions) is $2k$. Suppose n_i and n_j denote the number of distinct i indices and j indices, respectively. Since each edge needs to be traversed at least twice, there are at most $k+1$ distinct vertices, so $n_i + n_j \leq k+1$. In particular, if $n_i + n_j = k+1$ there are k unique edges and the resulting graph is a tree. Such terms will become the dominant ones in the sum (4.2) in the limit $n \to \infty$.

Indeed, let us show that all the terms with $n_i + n_j \leq k$ contribute an amount that is $o(n)$ to the sum in (4.2). To this end, define the *weight vector* $\mathbf{t_i}$ corresponding to the vector $\mathbf{i} = (i_1, \ldots, i_k)$, describing which entries of $\mathbf{i}$ are equal. For example, if $\mathbf{i} = (2, 5, 2, 1, 1)$, then $\mathbf{t_i} = (1, 2, 1, 3, 3)$, indicating that the first and third entries of $\mathbf{i}$ are equal, and that the fourth and fifth entries are also equal. Similarly, associate a weight vector to $\mathbf{j}$. Because the entries of Y_n are identically distributed, it is easy to see that choices of $\mathbf{i}$ and $\mathbf{j}$ which generate the same weight vectors contribute the same amounts to the sum in (4.2).

For a fixed weight vector $\mathbf{t_i}$ with n_i distinct entries (same n_i which gives the number of distinct row indices i_l), there are $n(n-1)\cdots(n-n_i+1) < n^{n_i}$ choices of $\mathbf{i}$ with weight vector $\mathbf{t_i}$. Similarly, there are $m_n(m_n-1)\cdot(m_n-n_j+1) < m_n^{n_j}$ choices of $\mathbf{j}$ corresponding to some fixed weight vector $\mathbf{t_j}$. Therefore, there are less than $n^{n_i} \cdot m_n^{n_j} < Cn^{n_i+n_j} \leq Cn^k$ for $n_i + n_j \leq k$, where C is a constant depending on k and α, but not n.

In addition, each term $\frac{1}{n}\mathbf{E}\, Y_n(i_1, j_1)Y_n(i_2, j_1)\cdots Y_n(i_k, j_k)Y_n(i_1, i_k)$ is $\mathcal{O}(1/n^{k+1})$ because of the scaling $Y_n = X_n/\sqrt{n}$ and the assumption that the moments of each entry $X_n(i, j)$ are finite. Therefore, the sum over all $\mathbf{i}$ and $\mathbf{j}$ corresponding to fixed weight vectors $\mathbf{t_i}$, $\mathbf{t_j}$ is $o(n)$. Furthermore, the number of possible weight vectors $\mathbf{t_i}$ and $\mathbf{t_j}$ depends on k but not n, which means that the contribution of all terms in the sum (4.2) is asymptotically 0.

Therefore, we now focus on the terms corresponding to $\mathbf{i}$ and $\mathbf{j}$ with $n_i + n_j = k + 1$. This is the case where the product

$$Y_n(i_1, j_1)Y_n(i_2, j_1)Y_n(i_2, j_2)Y_n(i_3, j_2)\cdots Y_n(i_k, j_k)Y_n(i_1, i_k)$$

contains exactly two copies of each distinct entry. Correspondingly, each edge in the path $i_1 j_1 \cdots i_k j_k i_1$ gets traversed twice, once in each direction. Because each entry of Y_n has variance $1/n$, we conclude that

$$\frac{1}{n}\mathbf{E}\, Y_n(i_1, j_1)Y_n(i_2, j_1)Y_n(i_2, j_2)Y_n(i_3, j_2)\cdots Y_n(i_k, j_k)Y_n(i_1, i_k) = \frac{1}{n^{k+1}}$$

for each choice of $\mathbf{i}$ and $\mathbf{j}$ with $n_i + n_j = k + 1$.

Again, fix two weight vectors $\mathbf{t_i}$ and $\mathbf{t_j}$ with n_i and n_j distinct entries. There are $n(n-1)\cdots(n-n_i+1)\cdot m_n(m_n-1)\cdots(m_n-n_j+1)$ corresponding choices for $\mathbf{i}$ and $\mathbf{j}$. Because $m_n \approx n\alpha$ for large n and we are in the case $n_i+n_j = k+1$, the number of choices is asymptotically equal to $n^{k+1}\alpha^{n_j}$.

From the last two observations, it follows that

$$\langle \overline{\mu}_n, x^k \rangle = \sum \alpha^{n_j},$$

where the sum is taken over all pairs $(\mathbf{t_i}, \mathbf{t_j})$ with $n_i + n_j = k + 1$ and distinct weight vectors.

To continue, we proceed as in the moment method proof of the semicircle law. To each closed path $i_1 j_1 i_2 j_2 \cdots i_k j_k i_1$ we can associate a *type sequence* of length $2k$, whose jth term gives the number of free steps minus the number of repetitive steps within the first j edge traversals. As before, every type sequence corresponding to a path where each edge gets traversed exactly twice starts at 1, ends at 0, and has consecutive terms differing by ± 1. Also note that the odd terms in a type sequence correspond to edges ending at a j vertex, whereas even terms correspond to edges terminating at an i vertex.

For a given type sequence, let l be the number of times there is a decrease by 1 going from an odd to an even term. Then $l = n_j$, the number of distinct j indices. Indeed, l counts the number of paths of the form $j_s i_{s+1}$ such that i_{s+1} has been visited once before, which by the condition that each edge is traversed exactly twice gives the number of distinct js. Furthermore, pairs of weight vectors $(\mathbf{t_i}, \mathbf{t_j})$ correspond bijectively to type sequences. Thus, letting

$$\beta_k = \sum_{\substack{\text{type sequences} \\ \text{of lenght } 2k}} \alpha^l,$$

we deduce

$$\langle \overline{\mu}_n, x^k \rangle = \beta_k.$$

The goal is to establish a recurrence relation between the β_k in order to compute the general term. To do this, associate to each type sequence of even length a second parameter $\bar{l}$ which counts the number of times there is a decrease by 1 going from an *even* to an *odd* term. Denote by $\gamma_k = \sum \alpha^{\bar{l}}$, where the sum is taken over type sequences of length $2k$.

Next, consider the (necessarily even) position $2j$ of the first occurrence of a zero in a type sequence of length $2k$. Then the elements beyond this index make up an arbitrary type sequence of length $2k - 2j$, with the first $2j$ terms forming a type sequence of length $2j$ with no zero occurring before the last position. By eliminating the first and last terms and subtracting 1 from each of the remaining elements, we see that such sequences are in bijection with arbitrary type sequences of length $2j - 2$. Furthermore, if l counts the number of decreases from odd to even indices in the sequence of length $2j$, then $l - 1$ gives the number of decreases from even to odd indices in this new sequence of length $2j - 2$. Keeping in mind how β_k and γ_k were defined in terms of powers of α, we deduce

$$\beta_k = \alpha \sum_{j=1}^{k} \gamma_{j-1} \beta_{k-j}.$$

Similar reasoning gives

$$\gamma_k = \sum_{j=1}^{k} \beta_{k-j}\gamma_{j-1}.$$

Thus, $\beta_k = \alpha\gamma_k$ for $k \geq 1$, with $\beta_0 = \gamma_0 = 1$ in order for these recurrences to hold. Since we are primarily interested in the β_k, note that these identities imply

$$\beta_k = (\alpha - 1)\beta_{k-1} + \sum_{j=1}^{k} \beta_{k-j}\beta_{j-1}.$$

In particular, if $\hat{\beta}(x) := \sum_{k=0}^{\infty} \beta_k x^k$ is the generating function for the β_k, the previous identity leads to the following equality for $\hat{\beta}$:

$$\hat{\beta}(x) = 1 + x\hat{\beta}(x)^2 + (\alpha - 1)x\hat{\beta}(x).$$

The expected ESD $\overline{\mu}_n$ thus converges to a distribution whose moments are encoded by $\hat{\beta}$. This asymptotic density has a Stieltjes transform $s(z)$ which can be easily computed as $s(z) = -\hat{\beta}(1/z)/z$, a claim which follows directly from the definition of the Stieltjes transform. Upon solving the quadratic equation in $\hat{\beta}$ from earlier, we have:

$$s(z) = \frac{-z + (\alpha - 1) + \sqrt{z^2 - 2z(\alpha + 1) + (\alpha - 1)^2}}{2z}.$$

Upon inversion of the Stieltjes transform, we see that the limiting density is given by

$$f_\alpha(x) == \frac{\sqrt{(x - \lambda_-)(\lambda_+ - x)}}{2\pi x}\mathbf{1}_{x \in [\lambda_-, \lambda_+]},$$

where $\lambda_- = (1 - \sqrt{\alpha})^2$ and $\lambda_+ = (1 + \sqrt{\alpha})^2$, as before. We have thus far shown that $\mathbf{E}\mu_n \to f_\alpha$ deterministically. An argument similar to that used to prove lemma 3.2.2 shows that, in fact, the ESD of an arbitrary Wishart matrix converges to the Marcenko-Pastur distribution, and thus theorem 4.1.1 is proven. ∎

Already, we see that the combinatorial argument underlying the moment method becomes more complicated going from Wigner to Wishart matrices. The main issue, however, is the fact that this kind of argument is unable to handle (true) correlations between matrix elements other than 0 or 1. This automatically excludes a large class of problems that are relevant to many applications. With the approach in the next section, it becomes much easier to handle such cases.

4.3 Free probability

With the free probability theory developed earlier, we can give a much simpler proof of theorem 4.1.1, based on [3]. For $n, t \geq 1$, consider a sequence of matrices

$$X_n = \left(r_i^s / \sqrt{t} \right)_{\substack{1 \leq i \leq n \\ 1 \leq j \leq t}},$$

with i.i.d. entries r_i^s of mean 0 and variance 1. The Wishart matrix $W_n = X_n X_n^T$, whose (i,j) entry is $\frac{1}{t} \sum_{s=1}^{t} r_i^s r_j^s$, gives the empirical correlation matrix observed over some finite amount of time t of what are otherwise uncorrelated quantities r_i^s. As before, we are interested in the case $t/n \to \alpha$ as $n \to \infty$, where $\alpha > 1$, such that the number of data points exceeds the dimensionality of the problem. For the purpose of this section, the inverse ratio $\beta = 1/\alpha < 1$ will be more useful.

Proof of Theorem 4.1.1. Note that W_n can be written as the sum of the rank one matrices

$$W_n^s = \left(r_i^s r_j^s \right)_{1 \leq i,j \leq n}.$$

For each s, denote by $\mathbf{r^s}$ the column vector $\begin{bmatrix} r_1^s & \dots & r_n^s \end{bmatrix}^T$. By the weak law of large numbers, it follows that for n large W_n^s has one eigenvalue equal to β in the direction of $\mathbf{r^s}$. The other $n-1$ eigenvalues are 0, corresponding to eigenvectors orthogonal to $\mathbf{r^s}$.

By the spectral theorem, each matrix W_n^s can be written as $U_n^s D_n^s (U_n^s)^*$, where U_n^s is a unitary matrix whose columns are the eigenvectors of W_n^s, and D_n^s is a diagonal matrix containing the eigenvalues $\beta, 0, \dots, 0$. Now, for $s \neq s'$, the vectors $\mathbf{r^s}$ and $\mathbf{r^{s'}}$ are almost surely orthogonal as $n \to \infty$, by the strong law of large numbers. Equivalently, the eigenvectors of W_n^s and $W_n^{s'}$ are almost surely orthogonal. A standard result [18] then implies that the matrices W_n^s with $1 \leq s \leq t$ are asymptotically free. Therefore, we can compute the spectrum of W_n by using the R-transform trick developed in an earlier section.

To start, the Stieltjes transform of each matrix W_n^s can be computed as follows:

$$\begin{aligned} s_n(z) = s_{W_n^s}(z) &= \frac{1}{n} \mathrm{tr}(W_n^s - zI)^{-1} = -\frac{1}{n} \sum_{k=0}^{\infty} \frac{\mathrm{tr}(W_n^s)^t}{z^{k+1}} = -\frac{1}{n} \left(\frac{n}{z} + \sum_{k=1}^{\infty} \frac{\beta^k}{z^{k+1}} \right) \\ &= -\frac{1}{z} + \frac{1}{nz} - \frac{1}{nz} \sum_{k=0}^{\infty} \left(\frac{\beta}{z} \right)^k = -\frac{1}{n} \left(\frac{n-1}{z} + \frac{1}{z - \beta} \right). \end{aligned}$$

As before, write $z := z_n(s)$ in order to find the functional inverse of the Stieltjes transform:

$$s = -\frac{1}{n} \left(\frac{n-1}{z_n(s)} + \frac{1}{z_n(s) - \beta} \right) \quad \Leftrightarrow \quad n s z_n(s)^2 - n(s\beta - 1) z_n(s) - (n-1)\beta = 0.$$

Solving the quadratic equation yields

$$\begin{aligned} z_n(s) &= \frac{n(s\beta-1) \pm \sqrt{n^2(s\beta-1)^2 + 4n(n-1)s\beta}}{2ns} \\ &= \frac{1}{2ns}\left[n(s\beta-1) \pm \sqrt{n^2(s\beta+1)^2 - 4ns\beta + \left(\frac{2s\beta}{s\beta+1}\right)^2 - \left(\frac{2s\beta}{s\beta+1}\right)^2}\right] \\ &\approx \frac{1}{2ns}\left(n(s\beta-1) \pm \left|n(s\beta+1) - \frac{2s\beta}{s\beta+1}\right|\right), \end{aligned}$$

since for n large enough the term $(2s\beta/(s\beta+1))^2$ is negligible. Using the intuition $z_n(s) \approx -1/s$ for n large, which follows directly from the definition of the Stieltjes transform, we can pick the correct root above and deduce

$$z_n(s) = -\frac{1}{s} + \frac{\beta}{n(1+s\beta)}.$$

The R transform of each W_n^s is therefore given by

$$R_{W_n^s}(s) = z_n(-s) - \frac{1}{s} = \frac{\beta}{n(1-s\beta)}.$$

As mentioned before, W_n is the free convolution of the random matrices W_n^s for $1 \leq s \leq t$, and thus its R-transform is given by

$$R_{W_n}(s) = tR_{W_n^s}(s) = \frac{\beta t}{n(1-s\beta)} \approx \frac{1}{1-s\beta},$$

for n large, since $n/t \to \beta$. Thus, the inverse of the Stieltjes transform of the limit of W_n is

$$z(s) = -\frac{1}{s} + \frac{1}{1+s\beta}.$$

Now, invert this again to obtain s as a function of z, and thus compute the Stieltjes transform $s(z)$ of the limit of W_n:

$$\begin{aligned} z = -\frac{1}{s(z)} + \frac{1}{1+\beta s(z)} \quad &\Leftrightarrow \quad \beta z s_n(z) + (z+\beta-1)s_n(z) + 1 = 0 \qquad (4.3) \\ &\Leftrightarrow \quad s(z) = \frac{-(z+\beta-1) + \sqrt{(z+\beta-1)^2 - 4\beta z}}{2z\beta}, \end{aligned}$$

again using the fact that $s_n(z) \approx -1/z$ to pick the correct root.

Finally, the Stieltjes transform given by (4.3) can be inverted using proposition 3.3.1 to find the limiting distribution of W_n:

$$f_\beta(y) = \frac{\sqrt{4y\beta - (y+\beta-1)^2}}{2\pi y\beta}, \qquad y \in [(1-\sqrt{\beta})^2, (1+\sqrt{\beta})^2].$$

In terms of $\alpha = 1/\beta$ and $x = \alpha y$, the limiting density is

$$f_\alpha(x) = \frac{\sqrt{(x-\lambda_-)(\lambda_+ - x)}}{2\pi x}, \qquad x \in [(1-\sqrt{\alpha})^2, (1+\sqrt{\alpha})^2],$$

which is precisely the Marcenko-Pastur distribution. ■

Compared to the moment method in the previous section, the free probability approach relies on the interactions between matrices, rather than individual entries, to derive the asymptotic result. The main advantage of this technique is that it allows for generalizations to sample correlation matrices where the true correlations between entries are strictly positive. This is important for applications, which often deal with data that has intrinsic correlations that we would like separated from the random noise. To achieve this, it is first important to understand at the theoretical level how various levels of correlations perturb the random spectrum, a kind of goal that would be infeasible with the basic techniques provided by the moment method. Free probability has been used to investigate this class of problems, as for example in [2, 3].

As a second observation, note that our proof in this section provides, at least in theory, a recipe for using free probability to derive asymptotic results. Specifically, if one can find a way to break up the matrix of interest into freely independent components that are already well understood, the $\mathbb{R}$-transform trick provides a way to put this information together to find an asymptotic limit on the original matrix. The problem with this is the fact that, as of now, free independence is not very intuitive coming from the classical probability theory mindset, and so finding those freely independent pieces would be difficult. Perhaps a better grasp of what free noncommutative random variables look like would be desirable.

5 Edge asymptotics for the GUE: the Tracy-Widom law

5.1 Introduction

To complement the two global asymptotics derived so far, we take this section to describe a local result. Not surprisingly, the moment method and the Stieltjes transform, which describe the behavior of all eigenvalues at once, are no longer the main machinery for proving theorems concerning just a few of these eigenvalues. Free probability also lacks the "resolution" to handle fluctuations of individual eigenvalues, as it considers entire matrices at once. Instead, special classes of orthogonal polynomials and the technique of integrating to eliminate variables that are not of interest become more prominent.

The Tracy-Widom law gives the limiting distribution of the largest eigenvalue for specific classes of matrices. In this section, we will derive the Tracy-Widom law for the Gaussian Unitary Ensemble, consisting of Hermitian matrices invariant under conjugation by unitary matrices. However, this result holds for matrices whose entries are i.i.d. (up to the hermitian constraint) with a distribution that is symmetric and has sub-Gaussian tail [9].

Theorem 5.1.1 (Preliminary)**.** *Let M_N, $N \geq 1$, be a matrix from the GUE whose largest eigenvalue is denoted by λ_N^N. Then there exists a cumulative distribution function F_2 such that for all $-\infty \leq t \leq \infty$,*

$$\lim_{N\to\infty} P\left[N^{2/3}\left(\frac{\lambda_N^N}{\sqrt{N}} - 2\right) \leq t\right] = F_2(t).$$

Informally, this theorem is looking at the fluctuations of the largest eigenvalue around the distribution predicted by Wigner's semicircle law. Recall that as $N \to \infty$, the *empirical spectral distribution* of the rescaled matrix $M_N/\sqrt{N}$ converge to $\sigma(x) = \sqrt{4-x^2}/2\pi$. In particular, $\lambda_N^N/\sqrt{N} \to 2$ as $N \to \infty$, so $\lambda_N^N/\sqrt{N} - 2$ has a trivial distribution. The theorem above suggests that multiplying by $N^{2/3}$ and considering the distribution of $N^{2/3}(\lambda_N^N/\sqrt{N} - 2)$ instead provides a much more interesting result.

There is a simple heuristic argument that explains the $N^{2/3}$ factor. Suppose that we're looking just at fluctuations of the largest eigenvalue *below* the limiting threshold $\lambda_+ = 2$. In particular, consider the random variable $N^\alpha(2 - \lambda_N^N/\sqrt{N})$, where α is just large enough to make the fluctuations non-trivial. Then:

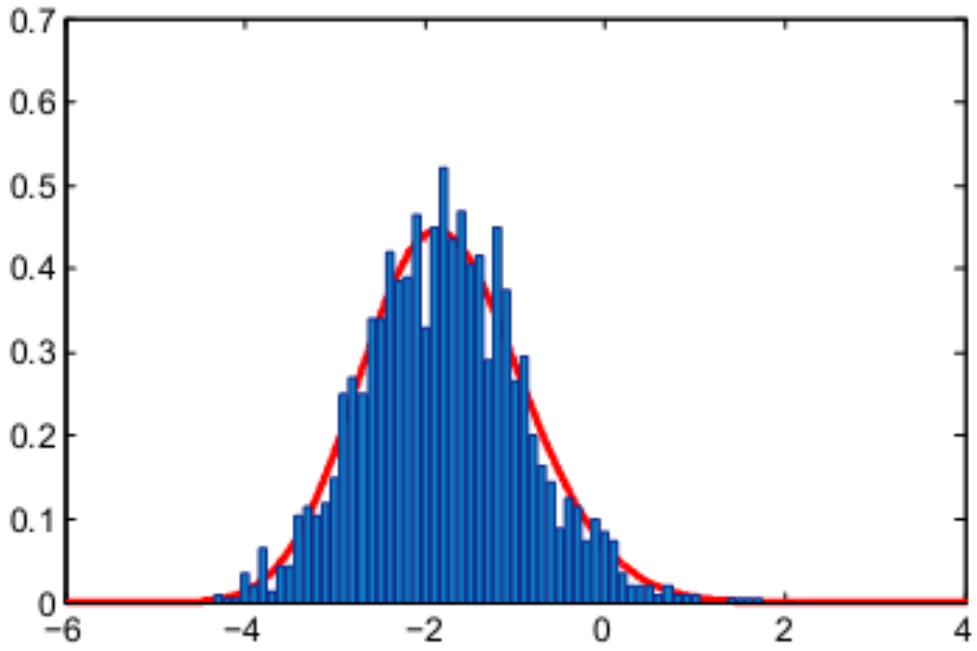

Figure 3: Simulation of the Tracy-Widom law using 2000 samples of the eigenvalues of 100 by 100 matrices. Bin size is 0.05. Matlab code based on [8].

$$P\left[N^{\alpha}\left(2-\frac{\lambda_N^N}{\sqrt{N}}\right)\leq t\right]=P\left(2-\frac{\lambda_N^N}{\sqrt{N}}\leq\frac{t}{N^{\alpha}}\right)=P\left(-\frac{t}{N^{\alpha}}\leq\frac{\lambda_N^N}{\sqrt{N}}\leq 2\right)$$
$$\approx\sqrt{2-\left(2-\frac{t}{N^{\alpha}}\right)^2}\frac{t}{N^{\alpha}}\approx\sqrt{\frac{t}{N^{\alpha}}}\frac{t}{N^{\alpha}}=\mathcal{O}\left(\frac{1}{N^{3\alpha/2}}\right).$$

Here we are using the intuition that close to $\lambda_+ = 2$, the probability distribution of the largest eigenvalue is given by the semicircle law. Then, since $\sigma(x)$ is a pdf, it follows that $P(2 - t/N^{\alpha} \leq \lambda_N^N/\sqrt{N} \leq 2) = \sigma(2 - t/N^{\alpha})t/N^{\alpha}$.

The question now is what size would be desirable for the fluctuations. Again by the semicircle law, the eigenvalues $\lambda_1^N, \ldots, \lambda_N^N$ are distributed roughly between -2 and 2, and the typical separation between two consecutive eigenvalues is $\mathcal{O}(1/N)$. For fixed N, we can think of $\mathcal{O}(1/N)$ as the "resolution" of the empirical distribution spectrum. In Figure 3.2, each bin corresponds to the number of eigenvalues that are in the small interval defined by that bin. If the bin size is decreased below the typical gap between eigenvalues, the resulting histogram would contains bins that have very few eigenvalues, or none at all.

This is the kind of degenerate behaviour that we would want to avoid in dealing with the fluctuations of the largest eigenvalue. Hence, we ask that:

$$P\left[N^{\alpha}\left(2-\frac{\lambda_N^N}{\sqrt{N}}\right)\leq t\right]\approx\mathcal{O}\left(\frac{1}{N^{3\alpha/2}}\right)=\mathcal{O}\left(\frac{1}{N}\right),$$

which gives the anticipated value $\alpha = 2/3$.

Throughout the rest of the section, we primarily follow [1].

5.2 The Gaussian Unitary Ensemble

Definition 5.2.1. Let $\{X_i\}_{1\le i\le N}$, $\{Y_{ij}\}_{1\le i<j\le N}$, and $\{Z_{ij}\}_{1\le i<j\le N}$ be independent families of i.i.d. standard normals. Consider a $N \times N$ matrix M whose entries are given by:

$$\begin{cases} M_{ii} = X_i \\ M_{ij} = \frac{Y_{ij}+iZ_{ij}}{2} = \overline{M_{ji}}, & \text{if } i < j \end{cases}$$

Matrices constructed in this manner form the *Gaussian Unitary Ensemble (GUE).*

The joint probability distribution of the matrix entries with respect to Lebesgue measure is easily determined by multiplying together the pdfs of the independent entries:

$$P_N(H)dH = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-h_{ii}^2/2} dh_{ii} \cdot \prod_{1\le i<j\le N} \frac{1}{\pi} e^{-|h_{ij}|^2} dh_{ij},$$

where $H = (h_{ij})_{i,j=1}^N$ and $dH = \prod_{1\le i\le j\le N} dh_{ij}$ is the Lebesgue measure on the space of Hermitian $N \times N$ matrices. The two products correspond to the diagonal and upper-triangular entries of H. Using the Hermitian condition, we can write $e^{-|h_{ij}|^2} = e^{-|h_{ij}|^2/2 - |h_{ji}|^2/2}$, which gives:

$$P(H)dH = \frac{1}{2^{N/2}} \frac{1}{\pi^{N^2/2}} \cdot e^{-\sum_{1\le i,j\le N} |h_{ij}|^2/2} dH = \frac{1}{2^{N/2}} \frac{1}{\pi^{N^2/2}} \cdot e^{-\mathrm{tr}H^2/2} dH. \quad (5.1)$$

This distribution is invariant under unitary conjugation:

$$\mathrm{tr}((UHU^*)^2) = \mathrm{tr}(UH^2U^*) = \mathrm{tr}(H^2U^*U) = \mathrm{tr}(H^2 I_N) = \mathrm{tr}H^2,$$

so H and its conjugate UHU^* have the same pdf. This justifies the name of the ensemble as the Gaussian Unitary Ensemble.

5.3 Joint eigenvalue distribution for the GUE

By diagonalizing H and performing a suitable change of variables, one can directly obtain the eigenvalue density of a GUE matrix from (5.1) above:

Theorem 5.3.1. *Let H be a random matrix from the GUE. The joint distribution of its eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$ is given by:*

$$\rho_N(\lambda_1, \ldots, \lambda_N)d\lambda = (2\pi)^{-N/2}\frac{1}{1!\cdot 2!\cdot \ldots \cdot N!}e^{-\mathrm{tr}H^2/2}\prod_{1\leq j<k\leq N}(\lambda_k - \lambda_j)^2 d\lambda, \tag{5.2}$$

where $d\lambda = \prod_{j=1}^{N} d\lambda_j$.

A complete proof can be found in [1] or [7].

The quantity $\prod_{1\leq j<k\leq N}(\lambda_k - \lambda_j)$ is the *Vandermonde determinant* associated to $\lambda_1, \ldots, \lambda_N$, and will be denoted $\Delta(\lambda)$.

5.4 Determinantal laws in the GUE

In this section we discuss how Hermite polynomials and wave functions arise naturally in the study of the GUE eigenvalue density, as described in [11].

Recall the joint eigenvalue distribution for the GUE from equation (5.2), rewritten in terms of the Vandermonde determinant $\Delta(\lambda) = \det_{1\leq i,j\leq N}(\lambda_i^{j-1})$:

$$P_N(\lambda_1, \ldots, \lambda_N)d\lambda = C_N e^{-\mathrm{tr}H^2/2}|\Delta(\lambda)|^2 d\lambda. \tag{5.3}$$

More generally, let $(\{P_j\})_{0\leq i\leq N-1}$ be a family of polynomials such that P_j has degree j. Consider the determinant $\det_{1\leq i,j\leq N}(P_{j-1}(\lambda_i))$. Using row operations to successively eliminate terms of degree less than $j-1$ from P_j, it follows that $\det_{1\leq i,j\leq N}(P_{j-1}(\lambda_i))$ is a constant multiple of $\Delta(\lambda)$. Furthermore, if P is the matrix $(P_{j-1}(\lambda_i))_{1\leq i,j\leq N}$, then $\det(PP^t)$ is a constant multiple of $|\Delta(\lambda)|^2$. Hence:

$$\rho_N(\lambda) = C'_N \det_{1\leq i,j\leq N}\Big(\sum_{k=0}^{N-1} P_k(\lambda_i)e^{-\lambda_i^2/4}P_k(\lambda_j)e^{-\lambda_j^2/4}\Big), \tag{5.4}$$

where C'_N is some nonzero constant dependent on N.

We have thus expressed the density ρ_N in terms of the family of polynomials (P_j), whose only constraint is that P_j has degree j, $0 \leq j \leq N-1$. Our goal is to choose these polynomials conveniently so that integrating (5.4) to obtain the density of $p \leq N$ of the eigenvalues becomes simpler.

Definition 5.4.1. (a) The *kth Hermite polynomial* is defined by:

$$\mathfrak{H}_k(x) := (-1)^k e^{x^2/2}\frac{d}{dx^k}e^{-x^2/2}. \tag{5.5}$$

(b) The *kth normalized oscillator wave function* is defined by:

$$\psi_k(x) = \frac{\mathfrak{H}_k(x)e^{-x^2/4}}{\sqrt{\sqrt{2\pi}k!}}. \tag{5.6}$$

Hermite polynomials are useful due to the following *orthogonality property*:

$$\frac{1}{\sqrt{2\pi}k!}\int_{\mathbb{R}} \mathfrak{H}_k(x)\mathfrak{H}_l(x)dx = \delta_{kl}. \tag{5.7}$$

Equivalently, the functions $(\mathfrak{H}_k)_{k\geq 0}$, properly normalized, form an orthonormal basis for the space of L^2 functions with the Gaussian measure $e^{-x^2/2}/2\pi dx$.

Orthogonality can also be expressed in terms of the oscillator wave functions ψ_k:

$$\int_{\mathbb{R}} \psi_k(x)\psi_l(x)dx = \delta_{kl}. \tag{5.8}$$

The propositions below summarize several general facts about Hermite polynomials and oscillator wave functions that we will use later on. A proof of these statements can be found in [1].

Proposition 5.4.2. *The Hermite polynomials $\{\mathfrak{h}_n(x)\}_{n=0}^{\infty}$ have the following properties:*

1. *$\mathfrak{h}_n(x)$ is monic polynomial in x of degree n.*
2. **(Orthogonality)** $\int \mathfrak{h}_k(x)\mathfrak{h}_l(x)e^{-x^2/2}dx = \sqrt{2\pi}k!\delta_{kl}$, *where δ_{kl} is the indicator function.*
3. $\mathfrak{h}_n'(x) = n\mathfrak{h}_{n-1}(x)$.
4. **(Christoffel-Darboux formula)** *For $x \neq y$,*

$$\sum_{k=0}^{n-1} \frac{\mathfrak{h}_k(x)\mathfrak{h}_k(y)}{k!} = \frac{\mathfrak{h}_n(x)\mathfrak{h}_{n-1}(y) - \mathfrak{h}_{n-1}(x)\mathfrak{h}_n(y)}{(n-1)!(x-y)}.$$

Proposition 5.4.3. *The oscillator wave functions $\{\psi_n(x)\}_{n=0}^{\infty}$ have the following properties:*

1. **(Orthogonality)** $\int \psi_k(x)\psi_l(x)dx = \delta_{kl}$.

2. **(Christoffel-Darboux formula)** *For $x \neq y$,*

$$\sum_{k=0}^{n-1} \psi_k(x)\psi_k(y) = \sqrt{n}\frac{\psi_n(x)\psi_{n-1}(y) - \psi_{n-1}(x)\psi_n(y)}{x-y}.$$

3. $\psi_n'(x) = -\frac{x}{2}\psi_n(x) + \sqrt{n}\psi_{n-1}(x)$.

4. $\psi_n''(x) + \left(n + \frac{1}{2} - \frac{x^2}{4}\right)\psi_n(x) = 0$.

The Hermite polynomials $\mathfrak{H}_k$, which are monic of degree k [1], play the role of our polynomials P_k above. Also, because $\psi_k(x)$ is a constant multiple of $\mathfrak{H}_k(x)e^{-x^2/4}$, the wave functions correspond to the terms $P_k(x)e^{-x^2/4}$ in (5.4) above. Therefore, it is natural that we introduce the notation

$$K_N(x,y) := \sum_{k=0}^{N-1} \psi_k(x)\psi_k(y), \tag{5.9}$$

which gives the following density for the eigenvalues:

$$\rho_N(\lambda) = C_N'' \det_{1\leq i,j\leq N}(K_N(\lambda_i,\lambda_j)). \tag{5.10}$$

We note the following useful property of the function K_N:

$$K_N(x,y) = \int_{\mathbb{R}} K_N(x,z)K_N(z,y)dz. \tag{5.11}$$

This is proven easily by writing K_N in terms of the ψ_k and using the orthogonality relations.

The following result tells us how to integrate the kernel with respect to one variable:

Lemma 5.4.4. *For any $k \geq 0$,*

$$\int_{\mathbb{R}} \det_{1\leq i,j\leq k+1}(K_N(\lambda_i,\lambda_j))d\lambda_{k+1} = (N-k)\det_{1\leq i,j\leq k}(K_N(\lambda_i,\lambda_j)).$$

Proof. We proceed by induction on k. For $k = 0$,

$$\int_{\mathbb{R}} K_N(\lambda,\lambda)d\lambda = \sum_{l=1}^{N-1}\int_{\mathbb{R}} \psi_l(\lambda)\psi_l(\lambda)d\lambda = N,$$

from the orthogonality of the ψ_l. This verifies the identity above.

Now, suppose the statement holds for $k-1 \geq 0$, and we wish to prove it for k. Applying cofactor expansion along the last row of $\det_{i,j=1}^{k+1}(K_N(\lambda_i, \lambda_j))$ yields:

$$\det_{i,j=1}^{k+1}(K_N(\lambda_i,\lambda_j)) = K_N(\lambda_{k+1},\lambda_{k+1}) \det_{i,j=1}^{k} K_N(\lambda_i,\lambda_j) + \sum_{l=1}^{k} (-1)^{k+1+l} K_N(\lambda_{k+1},\lambda_l) \det_{1\leq i\leq j, 1\leq j\leq k+1, j\neq l} K_N(\lambda_i,\lambda_j). \tag{5.12}$$

Integrating over λ_{k+1}, the first term on the right hand side becomes equal to $N \det_{i,j=1}^{k} K_N(\lambda_i,\lambda_j)$. For term l in the sum above, use multilinearity of the determinant to introduce the factor $K_N(\lambda_{k+1},\lambda_l)$ into the last column. By expanding on this last column, using (5.11), and swapping columns as necessary, the left hand side of (5.12) ends up being equal to $(n+1-k)\det_{1\leq i,j\leq k}(K_N(\lambda_i,\lambda_j))$, which proves the inductive step and hence the lemma. ∎

5.5 Local properties of the GUE eigenvalue distribution

We now come to the first result that speaks directly to the local fluctuations of the eigenvalues.

Lemma 5.5.1. *Let $A \subset \mathbb{R}$ be a measurable set. Then:*

$$P(\lambda_i \in A, i = 1,\ldots,N) = 1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!} \int_{A^c} \cdots \int_{A^c} \det_{i,j=1}^{k} K_N(x_i,x_j) \prod_{i=1}^{k} dx_i. \tag{5.13}$$

Before proving the statement above, we introduce a useful result that will simplify future calculations:

Proposition 5.5.2. *Given two families $f_1,\ldots,f_n$ and $g_1,\ldots,g_n$ of real-valued, square-integrable functions, the following identity holds:*

$$\frac{1}{n!}\int\cdots\int \det_{i,j=1}^{n}\left(\sum_{k=1}^{n} f_k(x_i)g_k(x_j)\right)\prod_{i=1}^{n} dx_i = \frac{1}{n!}\int\cdots\int \det_{i,j=1}^{n} f_i(x_j) \det_{i,j=1}^{n} g_i(x_j) \prod_{i=1}^{n} dx_i = \det_{i,j=1}^{n} \int f_i(x)g_j(x)dx. \tag{5.14}$$

The proof, which uses the identity $\det(AB) = \det(A)\det(B)$ and the permutation expansion of the determinant, can be found in [1].

Proof of lemma 5.5.1. The first key step is to use the joint eigenvalue distribution given by theorem 5.3.1 and integrate over the volume of space generated by A. In addition, from 5.5.2 and orthogonality of wave functions, we have:

$$\begin{aligned} P\left(\lambda_i \in A, i = 1, \ldots, N\right) &= \frac{1}{N!} \int_A \ldots \int_A \det_{i,j=1}^{N} K_N(x_i, x_j) \prod_{i=1}^{k} dx_i \\ &= \det_{i,j=0}^{N-1} \int_A \psi_i(x)\psi_j(x)dx \\ &= \det_{i,j=0}^{N-1} \left(\delta_{ij} - \int_{A^c} \psi_i(x)\psi_j(x)dx \right) \end{aligned}$$

Note that the indexing starts at 0 to be consistent with the definition of the wave functions. Now, expand the determinant above into a sum indexed over k, the number of factors in the product that are not equal to 1:

$$\begin{aligned} P \quad & (\lambda_i \in A, i = 1, \ldots, N) \\ & = 1 + \sum_{k=1}^{N} (-1)^k \sum_{0 \le v_1 \le \ldots \le v_k \le N-1} \det_{i,j=1}^{k} \left(\int_{A^c} \psi_{v_i}(x)\psi_{v_j}(x)dx \right) \end{aligned}$$

Using proposition 5.5.2 again and the identity $(\det A)^2 = \det(A^2)$ for any matrix A, we get

$$\begin{aligned} P \quad & (\lambda_i \in A, i = 1, \ldots, N) \\ & = 1 + \sum_{k=1}^{N} \frac{(-1)^k}{k!} \int_{A^c} \ldots \int_{A^c} \sum_{0 \le v_1 \le \ldots \le v_k \le N-1} \left(\det_{i,j=1}^{k} \psi_{v_i}(x_j) \right)^2 \prod_{i=1}^{k} dx_i \\ & = \sum_{k=1}^{N} \frac{(-1)^k}{k!} \int_{A^c} \ldots \int_{A^c} \det_{i,j=1}^{k} K_N(x_i, x_j) \prod_{i=1}^{k} dx_i. \end{aligned}$$

Lastly, because the rank of $\{K_N(x_i, x_j)\}_{i,j=1}^{k}$ is at most N, the sum above can be indexed from 1 to ∞, thus implying (5.13). ∎

5.6 Moments of the empirical distribution spectrum

Recall that λ_N^N denotes the largest eigenvalue of a GUE matrix of size $N \times N$. In particular, λ_N^N is a random variable. The goal of this section is to prove the following result:

Lemma 5.6.1. *There exist constants $c, C > 0$ so that*

$$P\left(\frac{\lambda_N^N}{2\sqrt{N}} \ge e^{N^{-2/3}\epsilon} \right) \le Ce^{-c\epsilon}, \tag{5.15}$$

$\forall N \ge 1$, $\epsilon > 0$.

This inequality is known as **Ledoux's bound**.

Let $\lambda_1^N \leq \lambda_2^N \leq \ldots \leq \lambda_N^N$ be the eigenvalues of the GUE matrix M_N. Recall that we can form the *empirical disribution function*,

$$\mu_{M_N/\sqrt{N}}(x) := \frac{1}{N}\left(\delta(\lambda_N^1 \leq x) + \delta(\lambda_N^2 \leq x) + \ldots + \delta(\lambda_N^N \leq x)\right). \tag{5.16}$$

Recall that $\mu_{M_N/\sqrt{N}}$ is a probability measure on probability measures, and in particular the average empirical distribution spectrum is a probability measure:

$$\overline{\mu_N} := E\mu_{M_N/\sqrt{N}}. \tag{5.17}$$

Because λ_N^N is drawn from the distribution $\overline{\mu_N}$, understanding this distribution will be helpful for proving lemma 5.6.1. In particular, we seek to bound the moments of $\overline{\mu_N}$.

Lemma 5.6.2. *Fix $N \in \mathbb{N}$. Then for any $t \in \mathbb{R}$, the moment generating function of $\overline{\mu_N}$ is given by:*

$$M_{\overline{\mu_N}}(t) := \int_{-\infty}^{\infty} e^{tx} d\overline{\mu_N}(x) = e^{t^2/2N} \sum_{k=0}^{N-1} \frac{1}{k+1}\binom{2k}{k}\frac{(N-1)\cdots(N-k)}{N^k}\frac{t^{2k}}{(2k)!} \tag{5.18}$$

Proof. By (5.11), $\frac{1}{N}K_N(x,x)dx = \rho_{1,N}$, and $\rho_{1,N}$ gives the probability density of one eigenvalue around x. Thus, $E\mu_{M_N}(x) = 1/NK_N(x,x)dx$, or equivalently $E\mu_{M_N/\sqrt{N}}(x/\sqrt{N}) = 1/NK_N(x,x)dx$. With the change of variables $x := \sqrt{N}x$, it follows that

$$\overline{\mu_N}(x) = \frac{1}{\sqrt{N}}K_N(\sqrt{N}x, \sqrt{N}x)dx.$$

In particular, the MGF of $\overline{\mu_N}$ can be written as:

$$M_{\overline{\mu_N}}(t) = \frac{1}{N}\int_{-\infty}^{\infty} e^{tx/\sqrt{N}} K_N(x,x)dx. \tag{5.19}$$

By the Christoffel-Darboux formula from proposition 5.4.3,

$$\frac{K_N(x,y)}{\sqrt{N}} = \frac{\psi_N(x)\psi_{N-1}(y) - \psi_{N-1}(x)\psi_N(y)}{x-y}.$$

By L'Hopital's rule with $y \to x$,

$$\frac{K_N(x,x)}{\sqrt{N}} = \psi_N'(x)\psi_{N-1}(x) - \psi_{N-1}'(x)\psi_N(x),$$

which also implies

$$K_N'(x,x)\sqrt{N} = \psi_N''(x)\psi_{N-1}(x) - \psi_{N-1}''(x)\psi_N(x) = -\psi_N(x)\psi_{N-1}(x),$$

where the last equality follows from proposition 5.4.3.

Integrating (5.19) by parts gives

$$M_{\overline{\mu_N}}(t) = \frac{1}{t\sqrt{N}} e^{tx/\sqrt{N}} K_N(x,x)\Big|_{-\infty}^{\infty} + \frac{1}{t}\int_{-\infty}^{\infty} e^{tx/\sqrt{N}}\psi_N(x)\psi_{N-1}(x)dx.$$

Since $K_N(x,x) \propto e^{-x^2/2}$ which goes to 0 faster than $e^{tx/\sqrt{N}}$ goes to ∞ when $x \to \infty$, and all other dependence on x is subexponential, it follows that the first term is 0. Hence

$$M_{\overline{\mu_N}}(t) = \frac{1}{t}\int_{-\infty}^{\infty} e^{tx/\sqrt{N}}\psi_N(x)\psi_{N-1}(x)dx. \tag{5.20}$$

Thus, we want to understand the integral $\int_{-\infty}^{\infty} e^{tx/\sqrt{N}}\psi_N(x)\psi_{N-1}(x)dx$. It as this point that we want to make use of the orthogonality of Hermite polynomials functions with respect to the Gaussian measure, as described by proposition 5.4.2. Specifically,

$$S_t^n = \frac{\sqrt{n}}{n!\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathfrak{h}_n(x)h_{n-1}(x)e^{-x^2/2+tx}dx.$$

With the change of variables $x := x + t$, the exponential inside the integral becomes Gaussian:

$$S_t^n = \frac{\sqrt{n}e^{t^2/2}}{n!\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathfrak{h}_n(x+t)h_{n-1}(x+t)e^{-x^2/2}dx. \tag{5.21}$$

For any $n \geq 0$, we have by Taylor's theorem

$$\mathfrak{h}_n(x+t) = \sum_{k=0}^{\infty}\frac{\mathfrak{h}_n^{(k)}(x)}{k!}t^k = \sum_{k=0}^{n}\binom{n}{k}\mathfrak{h}_{n-k}(x)t^k = \sum_{k=0}^{n}\binom{n}{k}\mathfrak{h}_k(x)t^{n-k}.$$

Note, in particular, that all derivatives of order higher than n vanish, since $\mathfrak{h}_n$ is a polynomials of degree n.

Substituting this sum into (5.21) and using the orthogonality relations, it

follows that

$$\begin{aligned}
S_t^n &= e^{t^2/2}\sqrt{n}\sum_{k=0}^{n-1}\frac{k!}{n!}\binom{n}{k}\binom{n-1}{k}t^{2n-1-2k} \\
&= e^{t^2/2}\sqrt{n}\sum_{k=0}^{n-1}\frac{(n-1-k)!}{n!}\binom{n}{n-1-k}\binom{n-1}{n-1-k}t^{2k+1} \\
&= e^{t^2/2}\sqrt{n}\sum_{k=0}^{n-1}\frac{(n-1-k)!}{n!}\frac{n!}{(k+1)!(n-1-k)!}\frac{(n-1)!}{k!(n-1-k)!}t^{2k+1} \\
&= e^{t^2/2}\sqrt{n}\sum_{k=0}^{n-1}\frac{1}{k+1}\binom{2k}{k}\frac{(n-1)\cdots(n-k)}{(2k)!}t^{2k+1}.
\end{aligned}$$

Therefore,

$$M_{\overline{\mu_N}}(t) = \frac{1}{t}S^N_{t/\sqrt{N}} = e^{t^2/2N}\sum_{k=0}^{N-1}\frac{1}{k+1}\binom{2k}{k}\frac{(N-1)\cdots(N-k)}{N^k}\frac{t^{2k}}{(2k)!},$$

which proves the lemma. ∎

Note that the identity above implies that all odd moments of $\overline{\mu_N}$ are 0. Beyond that, however, there is no information about individual moments, due to the $e^{t^2/2N}$ factor that has not been expanded as a power series.

The next lemma provides such information. To this end, for fixed N define $\{b_k\}_{k=0}^{\infty}$ such that:

$$M_{\overline{\mu_N}}(t) = \sum_{k=0}^{\infty}\frac{b_k}{k+1}\binom{2k}{k}\frac{t^{2k}}{(2k)!}.$$

Lemma 5.6.3. *For any integer k,*

$$b_{k+1} = b_k + \frac{k(k+1)}{4N^2}b_{k-1},$$

where $b_{-1} := 0$,

Proof. Establishing a second order recurrence relation between the coefficients of a power series is equivalent to finding a linear dependence between the power series and its first and second derivatives, which is precisely what this proof is doing.

Define

$$F(t) = \sum_{k=0}^{\infty}\frac{(-1)^k}{(k+1)!}\binom{N-1}{k}t^k$$

and

$$\Phi(t) = e^{-t/2}F(t).$$

Then $M_{\overline{\mu_N}}(t) = \Phi(-t^2/N)$, using lemma 5.6.2.

It is easy to verify that

$$\left(t\frac{d^2}{dt^2} + (2-t)\frac{d}{dt} + (N-1)\right)F(t) = 0,$$

and consequently

$$\left(4t\frac{d^2}{dt^2} + 8\frac{d}{dt} + 4N - t\right)\Phi(t) = 0. \tag{5.22}$$

Writing $\Phi(t) = \sum_{k=0}^{\infty} a_k t^k$, (5.22) gives

$$4(k+1)(k+2)a_{k+1} + 4Na_k - a_{k-1} = 0. \tag{5.23}$$

Bceause of how the a_k were defined, we have

$$\frac{(-1)^k a_k (2k)!}{N^k} = \frac{b_k}{k+1}\binom{2k}{k},$$

and thus (5.23) becomes

$$b_{k+1} = b_k + \frac{k(k+1)}{4N^2}b_{k-1},$$

as claimed. ∎

Proof of lemma 5.6.1. Note that $b_0 = b_1 \geq 0$ by definition, so the recursion in lemma 5.6.3 implies $b_{k-1} \leq b_k$, $\forall k \geq 1$. Then by lemma 5.6.3 again,

$$b_{k+1} \leq \left(1 + \frac{k(k+1)}{4N^2}\right)b_k.$$

Then

$$b_k \leq \prod_{l=0}^{k-1}\left(1 + \frac{l(l+1)}{4N^2}\right),$$

or equivalently

$$\log b_k \leq \sum_{l=0}^{k-1}\left(1 + \frac{l^2+l}{4N^2}\right) = k + \frac{1}{4N^2}\frac{k(k-1)(2k-1)}{6} + \frac{k(k-1)}{8N^2} \leq \frac{c'k^3}{N^2}, \tag{5.24}$$

for sufficiently large $c' > 0$ not depending on k or N.

By Stirling's approximation,

$$\frac{k^{3/2}}{2^{2k}(k+1)}\frac{(2k)!}{k!\cdot k!} \approx \frac{k^{3/2}}{2^{2k}(k+1)}\sqrt{4\pi k}\left(\frac{2k}{e}\right)^{2k}\frac{1}{2\pi k}\left(\frac{e}{k}\right)^{2k} \to 1/\sqrt{\pi} \qquad (5.25)$$

as $k \to \infty$, which means that

$$\sup_{k=0}^{\infty} \frac{k^{3/2}}{2^{2k}(k+1)}\frac{(2k)!}{k!\cdot k!} = C' < \infty.$$

Lastly, we would like to relate the moments of the random variable λ_N^N to those of the distribution $\overline{\mu_N}$, and consequently to the b_k. It is a general fact that the *sample* kth moment $(X_1^k + \ldots + X_n^k)/n$ is equal in expectation to the kth moment of the distribution that $X_1, \ldots, X_n$ come from. For our problem, the normalized eigenvalues $\lambda_N^1/\sqrt{N}, \ldots, \lambda_N^N/\sqrt{N}$ are drawn from the distribution $\overline{\mu_N}$. Since $\lambda_N^1 \leq \ldots \leq \lambda_N^N$, we get

$$\frac{E(\lambda_N^N/\sqrt{N})^{2k}}{N} \leq E\left(\frac{(\lambda_N^1/\sqrt{N})^{2k} + \ldots + (\lambda_N^N/\sqrt{N})^{2k}}{N}\right) = \int_{\mathbb{R}} x^{2k} d\overline{\mu_N}(x)dx.$$

Writing the kth moment of the law $\overline{\mu_N}$ in terms of b_{2k}, this implies:

$$E\left(\frac{\lambda_N^N}{\sqrt{N}}\right) \leq \frac{Nb_k}{k+1}\binom{2k}{k}. \qquad (5.26)$$

Now, from (5.24), (5.25), and (5.26), along with Markov's inequality, we get:

$$\begin{aligned} P\left(\frac{\lambda(N^N}{2\sqrt{N}} \geq e^{\epsilon}\right) &\leq E\left(\frac{\lambda_N^N}{2\sqrt{N}e^{\epsilon}}\right)^{2k} \leq \frac{Ne^{-2k\epsilon}}{2^{2k}}\frac{b_k}{k+1}\binom{2k}{k} \\ &\leq C'Nt^{-3/2}e^{-2\epsilon t + c't^3/N^2}, \end{aligned}$$

where $\lfloor t \rfloor = k$ and c', C' are absolute constants.

Replacing ϵ with $N^{-2/3}\epsilon$ and letting $t = N^{2/3}$, the result follows.

■

5.7 Fredholm determinants

Consider a locally compact space X such that X is homeomorphic to a complete space with a countable dense subset (this if typically known as a *Polish* space). In most applications, X is just $\mathbb{R}$. Also consider a measure ν on the Borel σ-algebra of X with $||\nu||_1 := \int_X |\nu(dx)| < \infty$.

Definition 5.7.1. A *kernel* is a Borel-measureable function $K : X \times X \to \mathbb{C}$ with

$$||K|| := \sup_{(x,y)\in X\times X} |K(x,y)| < \infty.$$

The *trace* of K with respect to the measure ν is defined as

$$\mathrm{tr}(K) = \int_X K(x,x)d\nu(x).$$

The *composition* of two kernels K and L defined on the same space X is given by

$$(K \star L)(x,y) = \int_X K(x,z)L(z,y)d\nu(z).$$

The conditions $||\nu||_1 < \infty$ and $||K|| < \infty$ ensure that the trace and the composition above are well-defined.

Proposition 5.7.2. (Hadamard's inequality) *Suppose $v_1, \ldots, v_n$ are $n \times 1$ column vectors. Then*

$$\det[v_1 \cdots v_n] \leq n^{n/2} \prod_{i=1}^{n} |v_i|_\infty.$$

Lemma 5.7.3. *Let $n > 0$. Consider two kernels $F(x,y)$ and $G(x,y)$. Then*

$$\left| \det_{i,j=1}^{n} F(x_i,y_j) - \det_{i,j=1}^{n} G(x_i,y_j) \right| \leq n^{1+n/2}||F-G|| \max(||F||,||G||)^{n-1} \quad (5.27)$$

and

$$\left| \det_{i,j=1}^{n} F(x_i,y_j) \leq n^{n/2}||F||^n \right|. \quad (5.28)$$

Proof. Let

$$H_i^k(x,y) = \begin{cases} G(x,y) & \text{if } i < k; \\ F(x,y) - G(x,y) & \text{if } i = k; \\ F(x,y) & \text{if } i > k. \end{cases}$$

By the linearity of the determinant, we have

$$\det_{i,j=1}^{n} F(x_i,y_j) - \det_{i,j=1}^{n} G(x_i,y_j) = \sum_{k=1}^{n} \det_{i,j=1}^{n} H_i^k(x_i,y_j).$$

Now, consider $\det_{i,j=1}^{n} H_i^k(x_i,y_j)$ for each k. One row of this determinant contains entries of the form $(F-G)(x_k,y_j)$, with the others rows given either

by $F(x_i, y_j)$ or $G(x_i, y_j)$. Applying Hadamard's inequality to the transpose of H_i^k implies

$$\left| \det_{i,j}^{n} H_i^k(x_i, y_j) \right| \leq n^{n/2} ||F - G|| \max(||F||, ||G||)^{n-1},$$

as desired.

Similarly, applying Hadamard's inequality to $\det_{i,j=1}^{n} F(x_i, y_j)$ yields (5.28) above.

■

Definition 5.7.4. For $n > 0$, define

$$\Delta_n = \Delta_n(K, \nu) = \int_X \cdots \int_X \det_{i,j=1}^{n} K(\xi_i, \xi_j) d\nu(\xi_1) \cdots d\nu(\xi_n).$$

Let $\Delta_0 = \Delta_0(K, \nu) = 1$.

The *Fredholm determinant* associated to $K(x, y)$ is defined as

$$\Delta(K) = \Delta(K, \nu) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \Delta_n(K, \nu). \tag{5.29}$$

Using (5.28) to obtain a uniform bound on $\det_{i,j=1}^{n} K(\xi_i, \xi_j)$, and then integrating with respect to $d\nu$ n times, we get:

$$\left| \int_X \cdots \int_X \det_{i,j=1}^{n} K(\xi_i, \xi_j) d\nu(\xi_1) \cdots d\nu(\xi_n) \right| \leq n^{n/2} ||K||^n ||\nu||_1^n. \tag{5.30}$$

In view of this bound on Δ_n, and by Stirling's approximation, we see that $\Delta(K)$ converges absolutely, and hence it is well-defined.

Although K itself was not required to be continuous, its Fredholm determinant $\Delta(K)$ satisfies certain continuity properties, as illustrated by the next lemma.

Lemma 5.7.5. *Consider two kernels K, L with respect to the same measure ν. Then:*

$$|\Delta(K) - \Delta(L)| \leq \left(\sum_{n=1}^{\infty} \frac{n^{1+n/2} ||\nu||_1^n \max(||K||, ||L||)^{n-1}}{n!} \right) ||K - L||. \tag{5.31}$$

Proof. Using the bound in (5.27), we have:

$$\begin{aligned}
&|\Delta(K) - \Delta(L)| \\
&\quad \leq \sum_{n=0}^{\infty} |\Delta_n(K,\nu) - \Delta_n(L,\nu)| \\
&\quad = \sum_{n=1}^{\infty} \frac{1}{n!} \left| \int \cdots \int (\det_{i,j=1}^{n} K(\xi_i,\xi_j) - \det_{i,j=1}^{n} L(\xi_i,\xi_j)) d\nu(\xi_1) \cdots d\nu(\xi_n) \right| \\
&\quad \leq \sum_{n=1}^{\infty} \frac{1}{n!} \int \cdots \int n^{1+n/2} ||K-L|| \max(||K||, ||L||)^{n-1} d\nu(\xi_1) \cdots d\nu(\xi_n).
\end{aligned}$$

Integrating $d\nu$ n times yields a $||\nu||_1^n$ factor, and the conclusion follows. ∎

In what follows, we will make use of the alternate notation

$$K \begin{pmatrix} x_1 & \cdots & x_n \\ y_1 & \cdots & y_n \end{pmatrix}$$

to denote $\det_{i,j=1}^{n} K(x_i, y_j)$.

As before, assume the measure ν and the kernel $K(x,y)$ are fixed.

Definition 5.7.6. For $n \geq 1$, consider

$$H_n(x,y) = \int \cdots \int K \begin{pmatrix} x & \xi_1 & \cdots & \xi_n \\ y & \xi_1 & \cdots & \xi_n \end{pmatrix} d\nu(\xi_1) \cdots d\nu(\xi_n),$$

and set $H_0(x,y) = K(x,y)$. Define the *Fredholm adjugant* of $K(x,y)$ as the function

$$H(x,y) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} H_n(x,y).$$

For $\Delta(K) \neq 0$, define the *resolvent* of the kernel $K(x,y)$ as

$$R(x,y) = \frac{H(x,y)}{\Delta(K)}.$$

Lemma 5.7.3 applied to H_n, together with Stirling's approximation, ensure that $H(x,y)$ converges absolutely and uniformly on $X \times X$, and thus the Fredholm adjugant is well-defined. In fact, both $H(x,y)$ and $R(x,y)$ (if defined) are kernels.

A kernel and its Fredholm adjugant are related through the following fundamental identity:

Lemma 5.7.7. *Let $K(x,y)$ be a kernel and $H(x,y)$ its Fredholm adjugant. Then:*

$$\int K(x,z)H(z,y)d\nu(z) = H(x,y) - \Delta(K)\cdot K(x,y) = \int H(x,z)K(z,y)d\nu(z). \tag{5.32}$$

Proof. We will prove the first equality only, as the other one follows similarly.

Using expansion by minors along the first row, we have

$$\begin{aligned}
&K\begin{pmatrix} x & \xi_1 & \cdots & \xi_n \\ y & \xi_1 & \cdots & \xi_n \end{pmatrix} \\
&\quad = K(x,y)K\begin{pmatrix} \xi_1 & \cdots & \xi_n \\ \xi_1 & \cdots & \xi_n \end{pmatrix} \\
&\qquad + \sum_{i=1}^{n} (-1)^i K(x,\xi_i) K\begin{pmatrix} \xi_1 & \cdots & \xi_{i-1} & \xi_i & \cdots & \xi_n \\ y & \xi_1 & \cdots & \xi_{i-1} & \cdots & \xi_n \end{pmatrix} \\
&\quad = K(x,y)K\begin{pmatrix} \xi_1 & \cdots & \xi_n \\ \xi_1 & \cdots & \xi_n \end{pmatrix} \\
&\qquad - \sum_{i=1}^{n} (-1)^i K(x,\xi_i) K\begin{pmatrix} \xi_i & \xi_1 & \cdots & \xi_{i-1} & \xi_{i+1} & \cdots & \xi_n \\ y & \xi_1 & \cdots & \xi_{i-1} & \xi_{i+1} & \cdots & \xi_n \end{pmatrix}.
\end{aligned}$$

Integrating this equality with respect to $\xi_1, \ldots, \xi_n$ gives

$$H_n(x,y) = \Delta_n K(x,y) - n\int K(x,z)H_{n-1}(z,y)d\nu(z). \tag{5.33}$$

Summing the relevant quantities over n, this yields

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{(n-1)!} \int K(x,z)H_{n-1}(z,y)d\nu(z) = \sum_{n=1}^{\infty} \frac{(-1)^n}{n!}(H_n(x,y) - \Delta_n K(x,y)).$$

Noting that $H_0(x,y) - \Delta_n K(x,y) = 0$ by definition, the second sum above can be indexed from $n = 0$, and the desired identity follows. ∎

Corollary 5.7.8. *For any $n \geq 0$,*

$$\frac{(-1)^n}{n!} H_n(x,y) = \sum_{k=0}^{n} \frac{(-1)^k}{k!} \Delta_k(K)\cdot \underbrace{(K \star \cdots \star K)}_{n+1-k}(x,y). \tag{5.34}$$

Additionally,

$$\frac{(-1)^n}{n!} \Delta_{n+1} = \sum_{k=0}^{n} \frac{(-1)^k}{k!} \Delta_k(K)\cdot \operatorname{tr} \underbrace{(K \star \cdots \star K)}_{n+1-k}. \tag{5.35}$$

Proof. In this proof, we use Δ_k to denote $\Delta_k(K)$.

For the first claim, we will proceed inductively. If $n = 0$, $H_0(x, y) = \Delta_0 \cdot K(x, y)$ holds by definition.

Now let $n \geq 1$. From (5.33) and the induction hypothesis, we have:

$$\begin{aligned}
&\frac{(-1)^n}{n!} H_n(x, y) \\
&= \frac{(-1)^n}{n!} \Delta_n \cdot K(x, y) + \frac{(-1)^{n-1}}{(n-1)!} \int K(x, z) H_{n-1}(z, y) d\nu(z) \\
&= \frac{(-1)^n}{n!} \Delta_n \cdot K(x, y) + \int K(x, z) \sum_{k=0}^{n-1} \frac{(-1)^k}{k!} \Delta_k \cdot \underbrace{(K \star \cdots \star K)}_{n-k}(z, y) d\nu(z) \\
&= \frac{(-1)^n}{n!} \Delta_n \cdot K(x, y) + \sum_{k=0}^{n-1} \frac{(-1)^k}{k!} \Delta_k \int K(x, z) \cdot \underbrace{(K \star \cdots \star K)}_{n-k}(z, y) d\nu(z),
\end{aligned}$$

and the conclusion follows.

Furthermore, taking $x = y = \xi$ in (5.34) and integrating with respect to ξ yields (5.35) above. ∎

5.8 The Airy kernel and the Tracy-Widom law

Definition 5.8.1. Let C be the contour in the complex plane defined by the ray joining the origin to ∞ through the point $e^{-\pi i/3}$ and the ray joining the origin to infinity through the point $e^{\pi i/3}$. The *Airy function* is defined by

$$\mathrm{Ai}(x) = \frac{1}{2\pi i} \int_C e^{\zeta^3/3 - x\zeta} d\zeta. \tag{5.36}$$

The *Airy kernel* is given by

$$K(x, y) = A(x, y) := \frac{\mathrm{Ai}(x)\mathrm{Ai}'(y) - \mathrm{Ai}'(x)\mathrm{Ai}(y)}{x - y}, \tag{5.37}$$

with the value for $x = y$ determined by continuity.

As before, let $\lambda_1^N, \lambda_2^N, \ldots, \lambda_N^N$ be the eigenvalues of a GUE matrix.

Theorem 5.8.2. *For $-\infty < t < \infty$,*

$$\begin{aligned}
&\lim_{N \to \infty} P\left[N^{2/3} \left(\frac{\lambda_N^N}{\sqrt{N}} - 2 \right) \leq t \right] \\
&\quad = 1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!} \int_t^{\infty} \cdots \int_t^{\infty} \det A(x_i, x_j)_{i,j=1}^k \prod_{j=1}^k dx_j := F_2(t).
\end{aligned} \tag{5.38}$$

Of course, this doesn't say anything about what the distribution $F_2(t)$ is. Although it cannot be computed in closed form, $F_2(t)$ can be represented as a solution to a specific differential equation as follows:

Theorem 5.8.3. *The function $F_2(t)$ above is given by*

$$F_2(t) = \exp - \int_t^\infty (x-t)q^2(x)dx, \tag{5.39}$$

where q is a solution of the Painlevé II differential equation:

$$q'' = tq + 2q^3, q(t) \to Ai(t) \text{ as } t \to \infty. \tag{5.40}$$

A proof of this fact can be found either in [1], or in the original paper by Tracy and Widom [15].

Proof of Theorem 5.8.2. Let $-\infty < t < t' < \infty$. Our first goal is to show the following:

$$\lim_{N\to\infty} P\left[N^{2/3}\left(\frac{\lambda_i^N}{\sqrt{N}} - 2\right) \notin [t,t'], i = 1,\dots,N\right] = \\ 1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!} \int_t^{t'} \dots \int_t^{t'} \det A(x_i,x_j)_{i,j=1}^k \prod_{j=1}^k dx_j \tag{5.41}$$

The idea is to let $t' \to \infty$ and thus deduce (5.38) above. We now focus on proving (5.41).

As anticipated, we will make use of lemma 5.5.1, which gives the probability that all the eigenvalues are contained in a set A in terms of a Fredholm determinant associated to the kernel K_N. In order to use this result, note that $N^{2/3}(\lambda_i^N/\sqrt{N}-2) \notin [t,t']$ is equivalent to $\lambda_i^N \notin [N^{-1/6}t+2\sqrt{N}, N^{-1/6}t'+2\sqrt{N}]$. Thus, letting A be the complement of the interval $[N^{-1/6}t + 2\sqrt{N}, N^{-1/6}t' + 2\sqrt{N}]$, lemma 5.5.1 implies:

$$P\left[N^{2/3}\left(\frac{\lambda_i^N}{\sqrt{N}} - 2\right) \notin [t,t'], i = 1,\dots,N\right] \\ = 1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!} \int_u^{u'} \dots \int_u^{u'} \det_{i,j=1}^k K_N(x_i',x_j') \prod_{i=1}^k dx_i',$$

where $u = N^{-1/6}t + 2\sqrt{N}$ and $u' = N^{-1/6}t' + 2\sqrt{N}$. With the change of

variables $x_i' := N^{-1/6}x_i + 2\sqrt{N}$, we get:

$$P\left[N^{2/3}\left(\frac{\lambda_i^N}{\sqrt{N}} - 2\right) \notin [t,t'], i = 1,\dots,N\right] \tag{5.42}$$
$$= 1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!} \int_t^{t'} \dots \int_t^{t'} \det_{i,j=1}^{k} \frac{1}{N^{1/6}} K_N\left(\frac{x_i}{N^{1/6}} + 2\sqrt{N}, \frac{x_j}{N^{1/6}} + 2\sqrt{N}\right) \prod_{i=1}^{k} dx_i.$$

It is therefore useful to introduce the notation

$$A_N(x,y) = \frac{1}{N^{1/6}} K_N\left(\frac{x_i}{N^{1/6}} + 2\sqrt{N}, \frac{x_j}{N^{1/6}} + 2\sqrt{N}\right).$$

Then, the right hand side of (5.42) is precisely $\Delta(A_N)$, the Fredholm determinant associated to the kernel $A_N(x,y)$. This means that proving (5.41) reduces to showing $\Delta(A_N) \to \Delta(A)$ as $N \to \infty$, where A is the Airy kernel.

In fact, because of the continuity property of Δ described in lemma 5.7.5, it suffices to show $A_N \to A$ as $N \to \infty$.

By the Christoffel-Darboux identity in proposition 5.4.3,

$$K_N(x,y) = \sqrt{N}\frac{\psi_N(x)\psi_{N-1}(y) - \psi_{N-1}(x)\psi_N(y)}{x-y}.$$

Furthermore, by property 4 in the same proposition, we can write

$$K_N(x,y) = \frac{\psi_N(x)\psi_N'(y) - \psi_N(y)\psi_N'(x)}{x-y} - \frac{1}{2}\psi_N(x)\psi_N(y).$$

With the notation $\Psi_n(x) := n^{1/12}\psi_n(x/n^{1/6} + 2\sqrt{n})$, it follows that

$$A_N(x,y) = \frac{\Psi_N(x)\Psi_N'(y) - \Psi_N'(x)\Psi_N(y)}{x-y} - \frac{1}{2N^{1/3}}\Psi_N(x)\Psi_N(y).$$

As $N \to \infty$, we have

$$\frac{1}{2N^{1/3}}\Psi_N(x)\Psi_N(y) \to 0.$$

Indeed, from the definition of the wave functions ψ_n and their correspondents Ψ_N,

$$\frac{1}{2N^{1/3}}\Psi_N(x)\Psi_N(y) = \mathcal{O}\left(\frac{1}{N^{1/6}} \cdot \frac{\sqrt{N}^N}{N!}\right),$$

which goes to 0 by Stirling's approximation.

Recall that the Airy kernel is defined as

$$A(x,y) = \frac{\mathrm{Ai}(x)\mathrm{Ai}'(y) - \mathrm{Ai}'(x)\mathrm{Ai}(y)}{x-y}.$$

Therefore, it suffices to show that $\Psi_N(x) \to \mathrm{Ai}(x)$ and $\Psi_N'(x) \to \mathrm{Ai}'(x)$, $\forall x$. ∎

Typically, this convergence is proven using the method of steepest descent. This method is used for computing highly oscillatory integrals by altering the contour in the complex plane to smooth out this highly fluctuating behavior. For the actual proof, we refer the reader to [1]. For a simpler, though not entirely rigorous argument based on Fourier analysis, see [12].

We see, therefore, that getting better control on the eigenvalue fluctuations requires more technical proofs. What remains striking, though, is that even local asymptotic results are highly universal.

References

[1] G.W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*. Cambridge studies in advanced mathematics. Cambridge University Press, 2009.

[2] J.-P. Bouchaud, L. Laloux, M. A. Miceli, and M. Potters. Large dimension forecasting models and random singular value spectra. *The European Physical Journal B - Condensed Matter and Complex Systems*, 55:201–207, 2007.

[3] J. P. Bouchaud and M. Potters. Financial Applications of Random Matrix Theory: a short review. *ArXiv e-prints*, 2009.

[4] J. Ginibre. Statistical Ensembles of Complex, Quaternion and Real Matrices. *Journal of Mathematical Physics*, 6:440–449, March 1965.

[5] P. L. Hsu. On the distribution of roots of certain determinantal equations. *Annals of Human Genetics*, 9(3):250–258, 1939.

[6] J.L. Kelley and T.P. Srinivasan. *Measure and integral*. Number v. 1 in Graduate texts in mathematics. Springer-Verlag, 1988.

[7] M.L. Mehta. *Random matrices*. Academic Press, 1991.

[8] N. Rao and Alan Edelman. The polynomial method for random matrices. *Foundations of Computational Mathematics*, 8:649–702, 2008.

[9] Alexander Soshnikov. Universality at the edge of the spectrum in Wigner random matrices. *Communications in Mathematical Physics*, 207:697–733, 1999.

[10] T. Tao, V. Vu, and M. Krishnapur. Random matrices: Universality of ESDs and the circular law. *ArXiv e-prints*, July 2008.

[11] Terrence Tao. 254a, Notes 6: Gaussian ensembles. `http://terrytao.wordpress.com/2010/02/23/254a-notes-6-gaussian-ensembles/`, 2010.

[12] Terrence Tao. The Dyson and Airy kernels of GUE via semiclassical analysis. `http://terrytao.wordpress.com/2010/10/23/the-dyson-and-airy-kernels-of-gue-via-semiclassical-analysis/`, 2010.

[13] Terrence Tao. Free probability. `http://terrytao.wordpress.com/2010/02/10/245a-notes-5-free-probability/`, 2010.

[14] Terrence Tao. The semi-circular law.

http://terrytao.wordpress.com/2010/02/02/254a-notes-4-the-semi-circular-law/, 2010.

[15] Craig Tracy and Harold Widom. Level-spacing distributions and the airy kernel. *Communications in Mathematical Physics*, 159:151–174, 1994. 10.1007/BF02100489.

[16] A. M. Tulino and S. Verdu. *Random matrix theory and wireless communications*. Now, Hanover, MA, 2004.

[17] L. A. Pastur V. A. Marchenko. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb.*, 1(4):457–483, 1967.

[18] Dan Voiculescu. Limit laws for random matrices and free products. *Inventiones Mathematicae*, 104:201–220, 1991. 10.1007/BF01245072.

[19] Eugene P. Wigner. On the statistical distribution of the widths and spacings of nuclear resonance levels. *Mathematical Proceedings of the Cambridge Philosophical Society*, 47(04):790–798, 1951.

[20] Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *The Annals of Mathematics*, 62(3):548–564, 1955.

[21] Eugene P. Wigner. On the distribution of the roots of certain symmetric matrices. *The Annals of Mathematics*, 67(2):325–327, 1958.

[22] John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1/2):32–52, 1928.