Durham University
Department of Mathematical Sciences

Introduction to Random Matrix Theory:

Global and Local Phenomena

Ben Rickard

Supervised by Mustazee Rahman and Kohei Suzuki

April 2025

Declaration

This piece of work is a result of my own work and I have complied with the Department's guidance on multiple submission and on the use of AI tools. Material from the work of others not involved in the project has been acknowledged, quotations and paraphrases suitably indicated, and all uses of AI tools have been declared.

The AI tool DeepSeek has been intermittently used for spelling and grammar checks, as well as for general readability in small parts.

Contents

Chapter 1

Introduction to Random Matrix Theory

1.1 Historical Motivations

Random matrices are matrices whose elements are randomly sampled from probability distributions. The study of random matrices has a rich history, beginning in the early 20th century and evolving into a multidisciplinary field with applications in statistics, physics, number theory, and beyond.

The first appearance of random matrices is in John Wishart's work in the 1920s. Wishart introduced random matrices in the context of statistics, specifically as sample covariance matrices. His 1928 paper, *The Generalized Product Moment Distribution in Samples from a Normal Multivariate Population* [16], laid the foundation for the study of random matrices in statistical inference. Wishart's work was motivated by the need to understand the distribution of sample covariance matrices in multivariate analysis.

However, the field gained significant momentum in the 1950s, largely due to the contributions of Eugene Wigner, a physicist working on nuclear physics. Wigner was interested in modelling the energy levels of heavy atomic nuclei, which are highly complex and unpredictable. He hypothesized that the spacings between energy levels could be described by the eigenvalues of large random matrices. Wigner introduced the **Gaussian Orthogonal Ensemble** (GOE), a specific class of random matrices with symmetries reflecting those of physical systems. Note, a formal definition and exploration of these matrices is given in Chapter 4. His work demonstrated that the eigenvalue spacings of these matrices followed a universal distribution, now known as the **Wigner surmise**, which matched empirical observations in nuclear physics. This marked the beginning of random matrix theory (RMT) as a tool for understanding chaotic and complex systems. [1]

Shortly after, Freeman Dyson and Madan Lal Mehta made significant contributions to the field. Dyson extended Wigner's work by introducing other matrix ensembles corresponding to different symmetry classes of physical systems. Mehta, on the other hand, focused on developing a rigorous mathematical framework for the study of random matrices. His 1967 book, *Random Matrices* [12], became a foundational text in the field, providing a comprehensive treatment of the spectral properties of random matrices and their connections to physical systems.

By the early 1960s, RMT had evolved into a formal mathematical discipline, with applications extending far beyond its origins in physics and statistics. One of the most surprising connections emerged in analytic number theory, particularly in the study of the renowned Riemann zeta function. In the 1970s, Hugh Montgomery and Freeman Dyson discovered a remarkable link between the spacings of the non-trivial zeros of the Riemann zeta function and the eigenvalue spacings of random matrices. This connection, suggested that the statistical behaviour of the zeros of the zeta function could be modelled using random matrices. [13]

In addition to analytic number theory, RMT has found applications in diverse fields such as quantum chaos, wireless communication, and finance. For example, in finance, random matrix theory is employed to analyse the correlation matrices of asset returns, and to filter out noise from financial data.

In many such applications, the focus is on the eigenvalues of large random matrices. For instance, in Principal Component Analysis (PCA), a widely used technique in statistics and data science, the largest eigenvalues of a covariance matrix are used to identify the directions of maximum variance in high-dimensional data. This allows statisticians to reduce the dimensionality of data, whilst preserving its most important features [11]. Recall also the work done by Wigner concerning the correspondence between eigenvalue spacings of certain large matrices and energy level spacings of atomic nuclei. Consequently, much of the work in RMT has been devoted to understanding the distribution and behaviour of eigenvalues, particularly in the limit as the matrix size grows infinitely large.

This report will concentrate on stating and proving established theorems in RMT that describe the behaviour of eigenvalues. We will explore theorems which correspond to both global and local behaviour of random matrices' eigenvalues. By exploring these results, we aim to provide a deeper understanding of the mathematical foundations of RMT.

1.2 Motivating Example

Lets first see an example of what a random matrix could look like, and how its eigenvalues would behave. Take the class of 2×2 matrices whose upper-triangular elements have been sampled from a Bernoulli(p) distribution, and then scaled by multiplying by 2 and subtracting 1, so that the elements are either ± 1 with equal probability. The bottom-left element is then set so that the resultant matrix is symmetric. It is easy to see that there are 2^3 such matrices, which are all shown below:

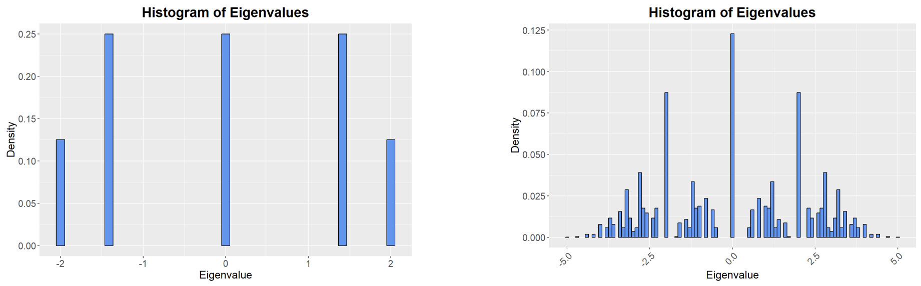
$$\begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix} \quad \begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix} \quad \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \quad \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & -1 \\ -1 & -1 \end{pmatrix} \quad \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Calculating the eigenvalues for each of these matrices, and compiling them into a frequency table gives us the following:

Eigenvalue	-2	$-\sqrt{2}$	0	$\sqrt{2}$	2
Frequency	2	4	4	4	2

If we plot this in a histogram as in Figure 1.1a, we get a visual display of where the eigenvalues of any random 2×2 matrix defined as above might be. Note that a similar example to this given in [15]. We are able to do the same thing for 5×5 matrices defined analogously as above, and plot the histogram in Figure 1.1b. In both of these plots, we notice that there is a similar overall shape, with a bulk of the eigenvalues appearing around the centre at zero, and fewer and fewer as we get towards the edge.



(a) 2×2

(b) 5×5

Figure 1.1: Histograms of Eigenvalues for 2×2 and 5×5 Matrices with Elements as Defined Above

We may question what occurs to a matrix's eigenvalues as its size $N \times N$ grows infinitely large, i.e. as $N \to \infty$. In order to attain a limit, we must scale

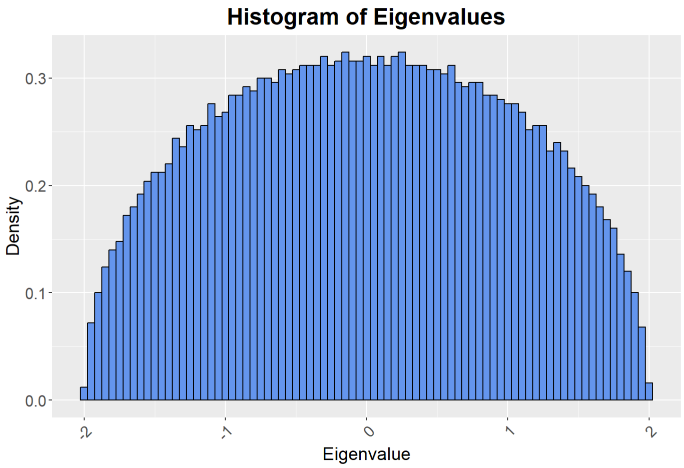


Figure 1.2: Histogram of Eigenvalues for a Scaled 5000 × 5000 Matrix with Elements as Defined Above

the matrix by a factor of N, otherwise the largest eigenvalue would also grow to infinity. The details of this scaling will be seen later. The histogram plot of the eigenvalues of a singular sample of a large 5000×5000 matrix, again defined analogously but now appropriately scaled, is seen in Figure 1.2. Maybe surprisingly, we have the same general shape as before, with the bulk around zero, and fewer eigenvalues towards the edge. This should spark some intrigue as to whether this general shape is true as $N \to \infty$, and if it has some limiting distribution. Furthermore, we have seen some motivation to continue our study of the eigenvalues of random matrices in order to formalise these statements.

1.3 Report Outline and Literature

This report is intended to firstly give a brief introduction to RMT. It will begin in Chapter 2 by formally defining a class of random matrices of which we will explore properties of. In particular, we will formalise our motivating example for the newly defined class of matrices. This involves a little amount of measure theory in order to correctly define the distributions and measures we require. In Chapter 2.2, we are then able to state one of the most renowned results in RMT known widely as **Wigner's Semicircle Theorem**, which captures the ideas seen in the motivating example.

Chapter 3 is then dedicated to proving said theorem. It is quite an involved proof which takes some time to work through in completion. The proof presented here mostly follows the proof seen in [1], with supplementation from [15]. The reference [15] proves the theorem for a different class of matrices than what we will, however, it is easily adapted for our work.

Next, Chapter 4 first discusses the key differences between **global and local phenomena** of random matrices. Chapters 2 and 3 will have focussed on global behaviour, and so we will then explore a local behaviour result. In order to do this we first define a new class of random matrices in Chapter 4.2, and then in Chapter 4.3 we state a theorem concerning local phenomena, and provide a simple proof. The theorem is presented in [3], with an outline of how to prove the statement.

Then, Chapter 5 starts to more thoroughly explore a possible application of RMT in neural networks. We will give a brief introduction to neural networks, but mostly focus on the possible appearance of our previous results to the large matrices found in neural networks. Our main reference here will be [3], but there are a multitude of sources discussing such work.

Finally, the conclusion will summarise the results we have seen, as well as providing some critique of the report, highlighting key areas which we will not have discussed. It will also give some suggestions for further exploration of RMT, and where more research could take place.

Chapter 2

Wigner Matrices and Eigenvalue Distribution

2.1 Key Definitions

We start by defining a class of random matrices known as Wigner matrices, following the conventions as in [1]. Then we will be exploring the asymptotic eigenvalue behaviour of such matrices as they grow large.

Definition 1 *An $N \times N$ matrix X_N is a **real Wigner matrix** if it has entries:*

$$X_N(i,j) = X_N(j,i) = \begin{cases} \frac{Z_{i,j}}{\sqrt{N}}, & \text{if } i < j, \\ \frac{Y_i}{\sqrt{N}}, & \text{if } i = j, \end{cases}$$

where there are two families of independent and identically distributed zero-mean real random variables $\{Z_{i,j}\}_{1 \leq i < j}$ such that $\mathbb{E}[Z_{1,2}^2] = 1$, and $\{Y_i\}_{1 \leq i}$. It is also required that $r_k := max\{\mathbb{E}[|Z_{1,2}|^k], \mathbb{E}[|Y_1|^k]\} < \infty$, i.e. that $Z_{i,j}$ and Y_i have finite moments. [1]

Note that by construction, real Wigner matrices are symmetric and hence Hermitian since we have assumed all entries are real. Through this report, we will drop 'real' and refer to them simply as **Wigner matrices** unless otherwise specified, since we will not be considering complex entries. Now, because these matrices are Hermitian, a Wigner matrix X_N will have N real eigenvalues, denoted $\{\lambda\}_{i=1}^N$, which can be ordered such that $\lambda_1^N < \cdots < \lambda_N^N$.

We can quickly prove this fact by taking a Hermitian matrix A, meaning that $A = A^\dagger$, where $A^\dagger$ denotes the conjugate transpose. Then, note that for an eigenvalue λ, with associated non-zero eigenvector $\mathbf{v}$, we have that $A\mathbf{v} = \lambda\mathbf{v}$ and so,

$$\mathbf{v}^\dagger A\mathbf{v} = \mathbf{v}^\dagger A^\dagger \mathbf{v} = (A\mathbf{v})^\dagger \mathbf{v} = (\lambda\mathbf{v})^\dagger \mathbf{v} = \lambda^* \mathbf{v}^\dagger \mathbf{v}.$$

However, we also have that,

$$\mathbf{v}^\dagger A\mathbf{v} = \mathbf{v}^\dagger (A\mathbf{v}) = \mathbf{v}^\dagger \lambda \mathbf{v} = \lambda \mathbf{v}^\dagger \mathbf{v}.$$

Hence, we get that $\lambda^* = \lambda$. Then since $\mathbf{v}^\dagger \mathbf{v} \neq 0$, the eigenvalue λ must be real. As this is true for an arbitrary eigenvalue of any Hermitian matrix, it is then true for all eigenvalues of all Hermitian matrices. Furthermore, indeed we have that the eigenvalues of a Hermitian matrix are real and so can be ordered.

Initially, we will focus on the empirical distribution of such eigenvalues which describes the spread of the eigenvalues. This formalises the histogram construction in Figure 1.2 from our motivating example.

Definition 2 *We define the* ***empirical distribution of eigenvalues****,* L_N*, for an* $N \times N$ *matrix as the probability measure on* $\mathbb{R}$ *such that,*

$$L_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i^N},$$

where δ *is the usual Dirac-delta distribution defined here as,*

$$\delta_{\lambda_i^N}(A) = \begin{cases} 1, & \lambda_i^N \in A \\ 0, & \textit{else.} \end{cases} \qquad \textit{[1]}$$

Hence, the distribution is a discrete probability measure which assigns equal measure $\frac{1}{N}$ at each eigenvalue λ_i^N of the matrix. The distribution is in particular a probability measure, since $L_N(\mathbb{R}) = 1$, because there are N eigenvalues, each of which contribute measure $\frac{1}{N}$. Note, we will interchangeably refer to this distribution as the empirical eigenvalue distribution.

Example 1 Lets consider a Wigner matrix such that the diagonal elements $Y_i \sim N(0,1)$, and the off-diagonal elements $Z_{i,j} \sim N(0,3)$. Note that since both families of random variables are Normally distributed, these matrices can be called **Gaussian Wigner matrices**. So, such a 3×3 matrix would have the general form:

$$\frac{1}{\sqrt{3}}\begin{pmatrix} y_1 & z_{1,2} & z_{1,3} \\ z_{1,2} & y_2 & z_{2,3} \\ z_{1,3} & z_{2,3} & y_3 \end{pmatrix}.$$

Note the symmetry and scaling which are vital in the analysis of such matrices. A possible realisation of this could look like:

$$\frac{1}{\sqrt{3}}\begin{pmatrix} 0.746 & 1.259 & -0.647 \\ 1.259 & -1.351 & 3.394 \\ -0.647 & 3.394 & 0.239 \end{pmatrix}.$$

The eigenvalues of this matrix can then be easily computed to be $-2.550, 0.624, 1.714$. Then the empirical distribution of eigenvalues can be plotted, as shown in Figure 2.1.

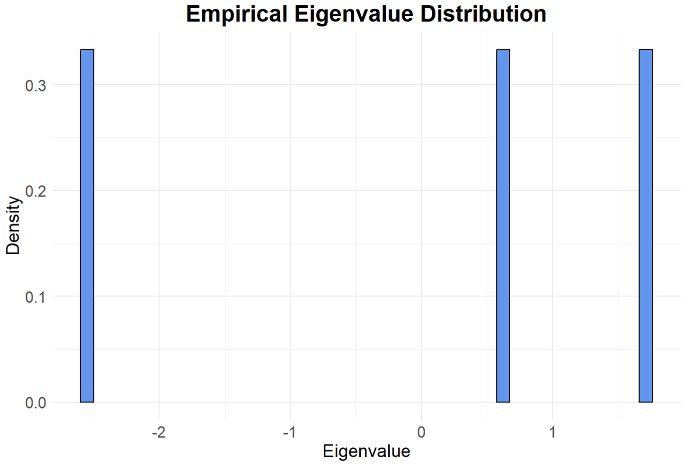


Figure 2.1: Empirical Eigenvalue Distribution for the 3 × 3 Example Gaussian Wigner Matrix

2.2 Wigner's Semicircle Theorem

Now we have formally defined the eigenvalue distribution of a specific class of random matrices, we recall our motivating example. In Figure 1.2, we saw that for large $N = 5000$, the empirical eigenvalue distribution has a curved shape, with the bulk of the concentration around zero, and much less density towards the edges at ± 2. We postulated whether as $N \to \infty$, the shape remains similar, and if it tends to some limiting distribution. We will soon see that this is in fact the case in our first major theorem, but we must first define some more mathematical objects.

Definition 3 *The* ***semicircle distribution****, denoted $\sigma(x)dx$, is defined on $\mathbb{R}$ with density,*

$$\sigma(x) = \frac{1}{2\pi}\sqrt{4-x^2}\mathbf{1}_{\{|x|\leq 2\}}. \quad \textit{[1]}$$

We are using the notation that $\mathbf{1}$ is the indicator function defined by

$$\mathbf{1}_{\{A\}}(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{else.} \end{cases}$$

The semicircle density is plotted in Figure 2.2a, where we see it is in fact a semicircular shape. Now, in order to formalise the convergence of probability distributions, we need some notion of convergence of measure.

Definition 4 *A sequence of measures $\{\mu_n\}_{n\in\mathbb{N}}$* ***converges weakly*** *to a measure μ, denoted $\mu_n \xrightarrow{w} \mu$, if for all continuous bounded functions $f \in C_b(\mathbb{R})$, we have*

$$\int_{\mathbb{R}} f d\mu_n \to \int_{\mathbb{R}} f d\mu. \qquad \text{[1]}$$

Example 2 We now give an illustrative example of a sequence of measures μ_n weakly converging to a measure μ. We start with a sequence of real measures $\{\mu_n\}_{n\in\mathbb{N}}$ where we define,

$$\mu_n(A) = \int_A \mathbf{1}_{[0,\frac{1}{n}]}(x)dx.$$

Intuitively, as $n \to \infty$, the measure shrinks to only give non-zero measure at zero. Hence, we will hypothesise that the limit under weak convergence of $\{\mu_n\}$ is the Dirac-delta distribution δ_0, where we use the previously stated definition. Testing our conjecture by using the Definition 4 of weak convergence, for any continuous bounded f, the left-hand side of the definition is,

$$\int_{\mathbb{R}} f d\mu_n(x) = \int_0^{\frac{1}{n}} f(x)dx \xrightarrow{n\to\infty} f(0).$$

Then considering our suggested limit measure,

$$\int_{\mathbb{R}} f d\delta_0(x) = f(0).$$

Hence, we do in fact get that μ_n converges weakly to measure δ_0.

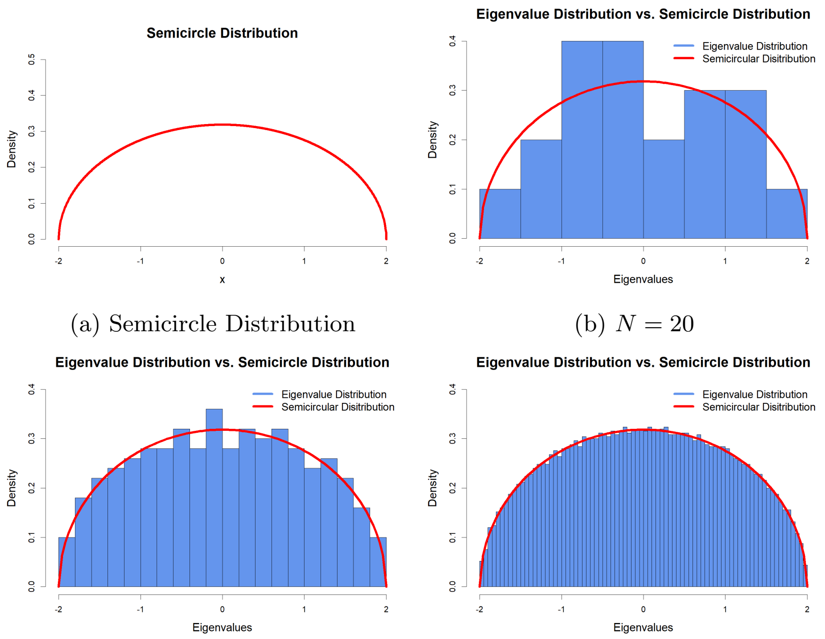
For ease of notation, going forward we use the notation $\langle \mu, f\rangle := \int_{\mathbb{R}} f d\mu$. We now have the tools to formally state the convergence theorem regarding empirical eigenvalue distributions of Wigner matrices.

Theorem 1 ***(Wigner's Semicircle Theorem)*** *For any Wigner matrix X_N, the empirical eigenvalue distribution L_N converges weakly in probability to the semicircle distribution σ, i.e. for all $f \in C_b(\mathbb{R})$ and for all $\epsilon > 0$,*

$$\lim_{N\to\infty} \mathbb{P}(|\langle L_N, f\rangle - \langle \sigma, f\rangle| > \epsilon) = 0. \quad \text{[1]}$$

Hence, as Wigner matrices grow large, their empirical eigenvalue density converges to the density of the semicircle law. Note that the convergence is in probability since, because the matrix elements and hence eigenvalues are random, the empirical eigenvalue distribution is also random. Furthermore, the convergence is described as a weak convergence in probability. This is visualised in Figure 2.2 where the convergence between the two distributions is apparent.

Note that in the definition of Wigner matrices, the elements are all scaled by a factor $\frac{1}{\sqrt{N}}$. This is the scaling required for the convergence as mentioned in the motivating examples, since otherwise the largest eigenvalue would tend to infinity with the size of the matrix. Hence, it gives us a non-trivial limit distribution as given in Theorem 1.



(a) Semicircle Distribution

(b) $N = 20$

(c) $N = 250$

(d) $N = 5000$

Figure 2.2: The Semicircle Distribution and Empirical Eigenvalue Distributions of Sampled Wigner Matrices for Increasing Values of N

Chapter 3

Proof of Wigner's Semicircle Theorem

In this chapter, we will work towards an analytical proof of Wigner's Semicircle Theorem in the case where X_N is a Gaussian Wigner Matrix. We limit ourselves to the Gaussian case here since we can exploit properties of Gaussian variables. This somewhat simplifies the proof, without limiting our understanding of the mechanics. We will take note through the proof where this is done. Also note that there are many different proofs of this result, including combinatorial approaches. However, we mostly follow the analytical proof presented in [1], with supplements from other sources where noted, as well as some independent work to provide full justifications where required. In particular, [15] will often be referred to when [1] skims over some important results.

3.1 Stieltjes Transform

3.1.1 Definition and Key Properties

Definition 5 *Let μ be a positive finite measure on the real line. Then the* ***Stieltjes transform*** *of μ is the function,*

$$S_\mu(z) := \int_{\mathbb{R}} \frac{d\mu(t)}{t-z}, \qquad \forall z \in \mathbb{C}^+. \quad \textit{[1]}$$

Here, we use the convention that $\mathbb{C}^+ := \{a + bi : a, b \in \mathbb{R}; b > 0\}$, and $\mathbb{C}^- := \{a + bi : a, b \in \mathbb{R}; b < 0\}$. Note that the Stieltjes transform is analytic on $\mathbb{C}^+$, but by analytic continuation, can be extended across $\mathbb{C}^-$, with a discontinuity along the real line. Also, if $z \in \mathbb{C} \setminus \mathbb{R}$, so we can write $z = a + bi$ where $a, b \in \mathbb{R}$

and $b \neq 0$, then it can be calculated that,

$$\begin{aligned}\frac{1}{t-z} = \frac{1}{t-(a+bi)} &= \frac{1}{(t-a)-bi} \cdot \frac{(t-a)+bi}{(t-a)+bi} \\ &= \frac{t-a}{(t-a)^2+b^2} + i\frac{b}{(t-a)^2+b^2}. \qquad (3.1)\end{aligned}$$

Hence, we have that the real and imaginary parts of $\frac{1}{t-z}$ are continuous bounded functions of $t \in \mathbb{R}$, and moreover we have that,

$$\begin{aligned}|S_\mu(z)| &\leq \int_{\mathbb{R}} \left|\frac{1}{t-z}\right| d\mu(t), && \text{by triangle inequality for integrals} \\ &\leq \int_{\mathbb{R}} \frac{1}{|\operatorname{Im}(z)|} d\mu(t), && \text{since } \max_{t\in\mathbb{R}} \left|\frac{1}{t-z}\right| = \frac{1}{|\operatorname{Im}(z)|} \text{ at } t = a \\ &= \frac{\mu(\mathbb{R})}{|\operatorname{Im}(z)|}.\end{aligned}$$

We will now claim that μ can be recovered from S_μ by the **Stieltjes inversion formula**. The statement of the formula, and proof outline comes from [15], but the finer details have been independently added to ensure full rigour.

Lemma 1 *For $a < b$, the finite measure μ can be recovered from its Stieltjes transform S_μ by the formula:*

$$\lim_{\epsilon\to 0^+} \frac{1}{\pi}\int_a^b \operatorname{Im}(S_\mu(x+i\epsilon))dx = \mu((a,b)) + \frac{1}{2}\mu(\{a,b\}).$$

Proof:
First we compute, by (3.1),

$$\begin{aligned}\operatorname{Im}(S_\mu(x+i\epsilon)) &= \int_{\mathbb{R}} \operatorname{Im}(\frac{1}{t-x-i\epsilon}) d\mu(t) \\ &= \int_{\mathbb{R}} \frac{\epsilon}{(t-x)^2+\epsilon^2} d\mu(t).\end{aligned}$$

Hence we have that,

$$\int_a^b \operatorname{Im}(S_\mu(x+i\epsilon))dx = \int_{\mathbb{R}}\int_a^b \frac{\epsilon}{(t-x)^2+\epsilon^2} dx d\mu(t), \qquad (3.2)$$

where we have been permitted to use Fubini's Theorem to swap the integrals since the integrand is integrable for our $\epsilon > 0$, and μ is a finite measure.

Now computing the inner integral by doing a u-substitution $u = \frac{x-t}{\epsilon}$,

$$\begin{aligned}
\int_a^b \frac{\epsilon}{(t-x)^2+\epsilon^2}dx &= \int_{\frac{a-t}{\epsilon}}^{\frac{b-t}{\epsilon}} \frac{\epsilon}{\epsilon^2u^2+\epsilon^2}\epsilon du \\
&= \int_{\frac{a-t}{\epsilon}}^{\frac{b-t}{\epsilon}} \frac{1}{u^2+1}du \\
&= \arctan(\frac{b-t}{\epsilon}) - \arctan(\frac{a-t}{\epsilon}) \\
&\xrightarrow{\epsilon\to 0^+} f(t) := \begin{cases} 0, & \text{if } t \notin [a,b] \\ \frac{\pi}{2}, & \text{if } t \in \{a,b\} \\ \pi, & \text{if } t \in (a,b) \end{cases}.
\end{aligned}$$

The final line is clear by considering that $\lim\limits_{\epsilon\to 0^+} \arctan(\frac{\delta}{\epsilon}) = \begin{cases} 0, & \text{if } \delta = 0 \\ \frac{\pi}{2}, & \text{if } \delta > 0 \\ -\frac{\pi}{2}, & \text{if } \delta < 0 \end{cases}.$

So now we have that,

$$\begin{aligned}
\lim_{\epsilon\to 0^+} \frac{1}{\pi}\int_a^b \mathrm{Im}(S_\mu(x+i\epsilon))dx &= \lim_{\epsilon\to 0^+} \frac{1}{\pi}\int_a^b\int_{\mathbb{R}} \frac{\epsilon}{(t-x)^2+\epsilon^2}d\mu(t)dx \\
&= \lim_{\epsilon\to 0^+} \frac{1}{\pi}\int_{\mathbb{R}}\int_a^b \frac{\epsilon}{(t-x)^2+\epsilon^2}dxd\mu(t), \quad \text{by (3.2).}
\end{aligned}$$

Now we want to be able to take the limit inside the first integral. So first note that since the integrand is positive,

$$g_\epsilon(t) := \int_a^b \frac{\epsilon}{(t-x)^2+\epsilon^2}dx < \int_{-\infty}^{\infty} \frac{\epsilon}{(t-x)^2+\epsilon^2}dx = \pi.$$

Hence, $|g_\epsilon(t)| < \pi$, and since μ is a finite measure,

$$\int_{\mathbb{R}} \pi d\mu = \pi\mu(\mathbb{R}) < \infty.$$

Furthermore, we are able to use the dominated convergence theorem to bring the limit inside the integral, and so we continue:

$$\begin{aligned}
&= \frac{1}{\pi}\int_{\mathbb{R}} \lim_{\epsilon\to 0^+}\int_a^b \frac{\epsilon}{(t-x)^2+\epsilon}dx\, d\mu(t) \\
&= \frac{1}{\pi}\int_{\mathbb{R}} f(t)d\mu(t) \\
&= \mu((a,b)) + \frac{1}{2}\mu(\{a,b\}) \quad \text{finishing the proof.}
\end{aligned}$$

□

Ideally, we would like to have that the Stieltjes transform uniquely defines a corresponding measure. The following Lemma 2 will give us this desired result.

Similarly to Lemma 1, the statement and proof outline come from [15], with more detail added in independently. In particular, the argument that there are only countably many atoms was not included in the proof from [15].

Lemma 2 *We have that for two finite measures μ and ν, if $S_\mu = S_\nu$, then necessarily $\mu = \nu$.*

Proof:
First we recall the definition of an atom. An atom of a measure μ is a measurable set A such that:

- $\mu(A) > 0$,
- for any measurable subset B, if $B \subset A$, then either $\mu(B) = 0$, or $\mu(B) = \mu(A)$.

This effectively means that A is the smallest set of positive measure, which cannot be divided into smaller sets of positive measure. [9]

Now, we can continue with the proof by first assuming $S_\mu = S_\nu$. Immediately by the Stieltjes inversion formula Lemma 1, we get that $\mu((a,b)) = \nu((a,b))$ for all open intervals (a,b), such that a and b aren't atoms of measures μ or ν. This is because since $S_\mu = S_\nu$, by the inversion formula, we get that $\mu((a,b)) + \frac{1}{2}\mu(\{a,b\}) = \nu((a,b)) + \frac{1}{2}\nu(\{a,b\})$. Then by our assumption that a and b aren't atoms of either measure, $\mu(\{a\}) = \mu(\{b\}) = \nu(\{a\}) = \nu(\{b\}) = 0$, leaving that $\mu((a,b)) = \nu((a,b))$.

Now we claim that there can only be countably many atoms of a finite measure μ. First note that any atoms A_1, A_2 must be disjoint. If they weren't disjoint, then $A_1 \cap A_2$ must also be an atom, which would contradict the minimality of atoms, unless $A_1 = A_2$. So suppose we have a collection of disjoint atoms $\{A_i\}_{i \in I}$ for some index I. We have that $\mu(A_i) > 0$, and since the atoms are disjoint we have by additivity, $\mu\left(\bigcup\limits_{i \in I} A_i\right) = \sum\limits_{i \in I} \mu(A_i)$. However, if I is uncountably big, then the sum must diverge to infinity, which would contradict that μ is a finite measure. Hence, I must be countable, and so there are countably many atoms.

Therefore, we can write any open interval (a,b) as,

$$(a,b) = \bigcap_{n=1}^{\infty} (a - \epsilon_n, b + \epsilon_n)$$

where ϵ_n is a monotonically decreasing sequence, tending to zero, such that all $a - \epsilon_n, b + \epsilon_n$ are not atoms of μ, or ν. So finally, by the monotone convergence theorem for measures, we get that,

$$\begin{aligned}\mu((a,b)) &= \lim_{n\to\infty} \mu((a-\epsilon_n, b+\epsilon_n)) \\ &= \lim_{n\to\infty} \nu((a-\epsilon_n, b+\epsilon_n)) \quad \text{since } a-\epsilon_n, b+\epsilon_n \text{ are not atoms of } \mu, \text{ or } \nu \\ &= \nu((a,b)).\end{aligned}$$

Hence we have finished the proof. □

3.1.2 Stieltjes Transform of the Semicircle Distribution

We are now going to derive the Stieltjes transform for the semicircle distribution. This will be very useful in proving Wigner's Semicircle Theorem, as we will see in the next section. First, we will prove the following Lemma 3 which gives an alternate formulation for the Stieltjes transform. This result and proof again comes from [15], but details about the uniform convergence have been added independently.

Lemma 3 *For a compactly supported measure μ, say $\mu([-r,r]) = 1$ for some $r > 0$, then S_μ has the following power series expansion about ∞:*

$$S_\mu(z) = -\sum_{k=0}^{\infty} \frac{m_k}{z^{k+1}}, \qquad \text{for } |z| > r,$$

where $m_k := \int_{\mathbb{R}} x^k d\mu(x)$ are the moments of the measure μ. [15]

Proof:
For $|z| > r$, we can expand as follows:

$$\frac{1}{t-z} = -\frac{1}{z}\left(\frac{1}{1-\frac{t}{z}}\right) = -\frac{1}{z}\sum_{k=0}^{\infty}\left(\frac{t}{z}\right)^k, \qquad \forall\ t \in [-r,r].$$

Note that since $\left|\frac{t^k}{z^{k+1}}\right| \le \frac{r^k}{|z^{k+1}|} < 1$, the expansion is a geometric series which converges absolutely. Then since we have that $\sum\limits_{k=0}^{\infty} \frac{r^k}{|z^{k+1}|}$ converges absolutely, the convergence on $[-r,r]$ is uniform, by the Weierstrass M-Test. Hence, we are permitted to pass the integral inside the infinite series as follows:

$$\begin{aligned}S_\mu(z) &= \int_{\mathbb{R}} \frac{d\mu(t)}{t-z} = \int_{-r}^{r} \frac{1}{t-z} d\mu(t) = \int_{-r}^{r} -\frac{1}{z}\sum_{k=0}^{\infty}\left(\frac{t}{z}\right)^k d\mu(t) \\ &= -\sum_{k=0}^{\infty}\int_{-r}^{r}\frac{t^k}{z^{k+1}}d\mu(t) = -\sum_{k=0}^{\infty}\frac{1}{z^{k+1}}\int_{\mathbb{R}} t^k d\mu(t) = -\sum_{k=0}^{\infty}\frac{m_k}{z^{k+1}}\end{aligned}$$

where the penultimate equality is due to the compact support of measure μ. Hence the proof is complete. □

So now we can calculate the Stieltjes transform of the semicircle distribution by computing its moments. A sparse version of this calculation can be found in [1], but here the computations are completed in more detail to ensure the argument can easily be followed. Again, we denote the moments as m_k for $k \geq 1$, and they are calculated by,

$$m_k := \langle \sigma, x^k \rangle = \frac{1}{2\pi} \int_{-2}^{2} x^k \sqrt{4 - x^2} dx.$$

First, it is clear by symmetry, $m_{2k+1} = 0$. Then in order to compute m_{2k} we must compute,

$$m_{2k} = \frac{1}{2\pi} \int_{-2}^{2} x^{2k} \sqrt{4 - x^2} dx.$$

We let $x = 2\sin(\theta)$, so $dx = 2\cos(\theta)d\theta$ and $\sqrt{4 - x^2} = 2\cos(\theta)$, so now,

$$\begin{aligned} m_{2k} &= \frac{2^{2k+1}}{\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin^{2k}(\theta) \cos^2(\theta) d\theta \\ &= \frac{2^{2k+1}}{\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin^{2k}(\theta) - \sin^{2k+2}(\theta) d\theta. \end{aligned}$$

Integrating by parts on the second term with $u = \sin^{2k+1}(\theta)$ and $dv = \sin(\theta)$ gives,

$$m_{2k} = \frac{2^{2k+1}}{\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin^{2k}(\theta) d\theta - (2k+1) m_{2k}.$$

Hence we rearrange to get that,

$$m_{2k} = \frac{2^{2k+1}}{\pi(2k+2)} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin^{2k}(\theta) d\theta. \tag{3.3}$$

Then integrating by parts, again, with $u = \sin^{2k-1}(\theta)$ and $dv = \sin(\theta)$,

$$\begin{aligned} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin^{2k}(\theta) d\theta &= (2k-1) \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin^{2k-2}(\theta) \cos^2(\theta) d\theta \\ &= (2k-1) \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin^{2k-2}(\theta) - \sin^{2k}(\theta) d\theta. \end{aligned}$$

Rearranging we get,

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin^{2k}(\theta)d\theta = \frac{2k-1}{2k} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin^{2k-2}(\theta)d\theta. \tag{3.4}$$

Then combining (3.3) and (3.4), we attain the recurrence relation $m_{2k} = \frac{4(2k-1)}{2k+2} m_{2k-2}$. Together with initial value $m_0 = 1$, we claim that $m_{2k} = C_k$, where C_k is the k^{th} Catalan value defined by,

$$C_k := \frac{\binom{2k}{k}}{k+1} = \frac{(2k)!}{(k+1)!k!}.$$

It is immediately clear that $C_0 = 1$. Then in the change of variables from $2k$ to k, we aim to show that,

$$C_k = \frac{4k-2}{k+1} C_{k-1}.$$

So, by our definition of the Catalan numbers,

$$\begin{aligned} \frac{C_k}{C_{k-1}} &= \frac{(2k)!}{(k+1)!k!} \frac{k!(k-1)!}{(2k-2)!} \\ &= \frac{2k(2k-1)}{(k+1)k} \\ &= \frac{4k-2}{k+1}. \end{aligned}$$

So now we have finished calculating the moments of the semicircle distribution: $m_{2k} = C_k$ and $m_{2k+1} = 0$. Thus, we are now ready to calculate the Stieltjes transform of the semicircle distribution. To do this, we will follow the arguments in [15], adding extra detail where necessary to ensure fully justified arguments.

Lemma 4 *The Stieltjes transform of the semicircle distribution, $S_\sigma(z)$ is for $z \in \mathbb{C}^+$*

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

Proof:
By Lemma 3, and that $m_{2k} = C_k$ and $m_{2k+1} = 0$,

$$S_\sigma(z) = -\sum_{k=0}^{\infty} \frac{C_k}{z^{2k+1}}, \qquad \text{for } |z| > 2.$$

Also, recall the well-known recursion for Catalan numbers $C_k = \sum_{i=0}^{k-1} C_i C_{k-1-i}$, which emerges frequently in combinatorics [10]. Using both of these we get that

$S_\sigma(z)^2 + zS_\sigma(z) + 1 = 0$ for $|z| > 2$. The calculation to show this can be done as follows:

$$\begin{aligned}
S_\sigma(z)^2 + zS_\sigma(z) + 1 &= \left(-\sum_{k=0}^{\infty} \frac{C_k}{z^{2k+1}}\right)^2 - \sum_{k=0}^{\infty} \frac{C_k}{z^{2k}} + 1, \qquad \text{for } |z| > 2 \\
&= \sum_{k=0}^{\infty} \left(\sum_{i=0}^{k} C_i C_{k-i}\right) \frac{1}{z^{2k+2}} - \sum_{k=1}^{\infty} \frac{C_k}{z^{2k}} - C_0 + 1 \\
&= \sum_{k=1}^{\infty} \left(\sum_{i=0}^{k-1} C_i C_{k-1-i}\right) \frac{1}{z^{2k}} - \sum_{k=1}^{\infty} \frac{C_k}{z^{2k}} - 1 + 1 \\
&= \sum_{k=1}^{\infty} \frac{C_k}{z^{2k}} - \sum_{k=1}^{\infty} \frac{C_k}{z^{2k}}, \qquad \text{by the Catalan recursion} \\
&= 0.
\end{aligned}$$

This quadratic has two solutions,

$$S_\sigma(z) = \frac{-z \ \pm \sqrt{z^2 - 4}}{2},$$

but since we know that $S_\sigma(z)$ is analytic on $\mathbb{C}^+$, we only take the solution in $\mathbb{C}^+$, and so the proof is done. □

3.2 Convergence of Measure and Transform

In this section, we will prove the equivalence in the convergence of measures and the convergence of their Stieltjes transform. This will motivate our later strategy in proving Theorem 1 for Gaussian Wigner matrices. Moreover, this also motivates why the Stieltjes transform computation was done in the previous section. The main results in this subchapter again are taken from [1]. However, as before, independently more details have been added to fully justify any arguments. Firstly, we require another notion of convergence in measure, in addition to weak convergence.

Definition 6 *A sequence of measures $\{\mu_n\}_{n\in\mathbb{N}}$* ***converges vaguely*** *to a measure μ, denoted $\mu_n \xrightarrow{v} \mu$, if for all continuous functions which vanish at infinity, $f \in C_0(\mathbb{R})$, we have*

$$\int_{\mathbb{R}} f d\mu_n \to \int_{\mathbb{R}} f d\mu. \qquad \textit{[1]}$$

Example 3 We are now going to see an example of a sequence of real measures $\{\mu\}_{n\in\mathbb{N}}$ which converge vaguely to some limit measure μ, but not weakly to such μ. We take $\mu_n = \delta_n$, a Dirac-delta distribution, defined by:

$$\delta_n(A) = \begin{cases} 1, & n \in A \\ 0 & \text{else} \end{cases}$$

for any set $A \subset \mathbb{R}$. If we let $f \in C_0(\mathbb{R})$, then there is an N large enough such that for all $n \geq N, f(n) = 0$. Therefore, for all $n \geq N$,

$$\int_{\mathbb{R}} f d\delta_n = f(n) = 0.$$

Now let $\mu = 0$, the zero measure, which satisfies,

$$\int_{\mathbb{R}} f d\mu = 0 \quad \text{for all } f \in C_0(\mathbb{R}).$$

Hence we have shown that,

$$\lim_{n\to\infty} \int_{\mathbb{R}} f d\delta_n = \int_{\mathbb{R}} f d\mu,$$

and so δ_n converges vaguely to $\mu = 0$. Now we will consider if the convergence is also weak. If we let $f(x) = 1 \in C_b(\mathbb{R})$, we have that,

$$\int_{\mathbb{R}} f(x) d\delta_n = f(n) = 1 \quad \text{for all n.}$$

However, we still get that for $\mu = 0$,

$$\int_{\mathbb{R}} f(x) d\mu = 0.$$

Hence δ_n doesn't converge weakly to $\mu = 0$. Furthermore, there is a clear difference between the two modes of convergence.

We are almost ready to prove the equivalence between convergence in the Stieltjes transform of probability measures and weak convergence of said probability measures. First, recall that a probability measure μ is such that $\mu(\mathbb{R}) = 1$, and a sub-probability measure ν is such that $\nu(\mathbb{R}) \leq 1$. Then, before the main result we recall a well-known theorem from analysis which we will use without proof. We don't prove this theorem here since it relies on some heavy analysis which we don't have the space for here. It also wouldn't add much to the discussion of RMT.

Theorem 2 ***(Helly's Selection Theorem)*** *Let $\{\mu_n\}_{n\in\mathbb{N}}$ be a sequence of tight Borel probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, recalling tight means that for all $\epsilon > 0$, there exists a compact set $K \subset \mathbb{R}$ such that $\mu_n(\mathbb{R} \setminus K) < \epsilon$ for each n. Then there exists a subsequence $\{\mu_{n_k}\}_{k\in\mathbb{N}}$ and a sub-probability measure μ such that $\mu_{n_k} \xrightarrow{v} \mu$ i.e. for every function $f \in C_0(\mathbb{R})$,*

$$\int_{\mathbb{R}} f d\mu_{n_k} \to \int_{\mathbb{R}} f d\mu. \qquad [2]$$

Finally, we can formally state the theorem concerning the relevant equivalence in modes of convergence. The statement and ideas behind the proof are from [1], but here is a more detailed proof, in particular why we can apply Theorem 2.

Theorem 3 *Let μ_n be a sequence of probability measures, then the following three statements below are true.*

i) If μ_n converges weakly to a probability measure μ, then $S_{\mu_n}(z)$ converges to $S_\mu(z)$ for all $z \in \mathbb{C} \setminus \mathbb{R}$.

ii) If $S_{\mu_n}(z)$ converges to a limit $S(z)$ for all $z \in \mathbb{C}\setminus\mathbb{R}$, then $S(z)$ is the Stieltjes transform of a sub-probability measure μ, with μ_n converging vaguely to μ.

iii) If the probability measures μ_n are random, and $S_{\mu_n}(z)$ converges in probability for all $z \in \mathbb{C} \setminus \mathbb{R}$ to a deterministic limit $S(z)$ which is the Stieltjes transform of a probability measure μ, then μ_n converges weakly in probability to μ. [1]

Proof:
i) For $z \in \mathbb{C}^+$, we will consider the function $f_z(t) = \frac{1}{t-z}$ which is continuous over our domain with $t \in \mathbb{R}$. Then since $\lim_{|t|\to\infty} f_z(t) = 0$, we have that $f_z \in C_0(\mathbb{R}) \subset C_b(\mathbb{R})$. Hence, by the assumption of weak convergence and simply applying the definition,

$$S_{\mu_n}(z) = \int_{\mathbb{R}} f_z(t) d\mu_n(t) \to \int_{\mathbb{R}} f_z(t) d\mu(t) = S_\mu(z),$$

and so we are done.

ii) We want to use Theorem 2, so first we must argue that the sequence of probability measures are tight. Take an arbitrary probability measure, ν defined on $\mathbb{R}$, meaning that $\nu(\mathbb{R}) = 1$. Then we consider a compact set $K_M = [-M, M] \subset \mathbb{R}$. Note that as M tends to infinity, the set K_M tends to $\mathbb{R}$ with $\bigcup\limits_{M>0} K_M = \mathbb{R}$. Hence, we get that $\lim\limits_{M\to\infty} \nu([-M, M]) = 1$. Furthermore, by the continuity of measure, there exists M large enough such that $\nu(\mathbb{R} \setminus K_M) = 1 - \nu(K_M) < \epsilon$, for all $\epsilon > 0$. Furthermore, we are able to use Theorem 2 on the sequence of probability measures, μ_n, implicitly assumed to be defined on $\mathbb{R}$.

So now, by Helly's Selection Theorem, Theorem 2, there exists a subsequence n_k such that μ_{n_k} converges vaguely to some probability measure μ. Then again considering the same $f_z(t) \in C_0(\mathbb{R})$ as in the proof of (i), and applying the definition of vague convergence, we get that $S_{\mu_{n_k}}(z) \to S_\mu(z)$. Hence, by our starting assumption and uniqueness of limits, we conclude that $S_\mu(z) = S(z)$.

Then recall that Lemma 2 gives us that the Stieltjes transform uniquely defines a measure. So since all subsequences of μ_n converge vaguely to the same limit μ, then μ_n converges vaguely to μ.

iii) To quantify the convergence of measures, we introduce the metric,

$$\rho(\nu_1, \nu_2) = \sum_{i=1}^{\infty} \frac{1}{2^i} \left| S_{\nu_1}(z_i) - S_{\nu_2}(z_i) \right|,$$

where $\{z_i\}_{i=1}^{\infty}$ is a countable dense subset of $\mathbb{C} \setminus \mathbb{R}$. This metric satisfies the property that if $\rho(\nu_n, \nu) \to 0$, then ν_n converges weakly to ν.

Again by Helly's Selection Theorem, Theorem 2, the sequence μ_n of probability measures has a vaguely convergent subsequence μ_{n_k} that converges to some sub-probability measure θ. Since $S_{\mu_{n_k}}(z) \to S_\mu(z)$ by assumption, it follows that,

$$S_\theta(z) = S_\mu(z), \quad \forall z \in \mathbb{C} \setminus \mathbb{R}.$$

By the uniqueness of measures from their Stieltjes transforms, Lemma 2, we conclude that $\theta = \mu$. Thus, all vaguely convergent subsequences of μ_n must converge to μ, meaning that μ_n itself converges vaguely to μ.

Finally, since $S_{\mu_n}(z)$ converges in probability to $S(z)$, it follows that,

$$\rho(\mu_n, \mu) \to 0 \quad \text{in probability.}$$

Since ρ corresponds with weak convergence, this implies that μ_n converges weakly to μ in probability. This completes the proof. □

Hence, we now have that if we can prove the convergence of the empirical eigenvalue distribution Stieltjes transform to the semicircle distribution Stieltjes transform, then we equivalently have weak convergence between the empirical eigenvalue distribution to the semicircle distribution. Recall that the empirical eigenvalue distribution is a random probability measure, whilst the semicircle distribution and its Stieltjes transform are deterministic, as is required in Theorem 3*iii)*. Consequently, we are able to use this to move forward with the proof of Wigner's Semicircle Theorem.

3.3 Convergence of the Expectation

We will continue to work on a proof of Wigner's Semicircle Theorem by considering the **expected empirical eigenvalue distribution** and its convergence to the semicircle distribution. Again, we will be following the main ideas from [1], along with some more details from [15]. However, [15] only considers Gaussian Wigner matrices whose diagonal elements have variance $\sigma^2 = 2$, so we will have to make some adjustments to the arguments presented there to make them suitable for general variance. First we must calculate the Stieltjes transform of the empirical eigenvalue distribution L_N, which is an easy computation:

$$\begin{aligned} S_{L_N}(z) &= \int_{\mathbb{R}} \frac{dL_N(t)}{t-z} \\ &= \int_{\mathbb{R}} \frac{1}{t-z} d\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i^N}(t)\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} \int_{\mathbb{R}} \frac{1}{t-z} d\delta_{\lambda_i^N}(t) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\lambda_i^N - z}. \end{aligned}$$

We will now show that $S_{L_N}(z)$ has an alternative representation which will be easier for us to use. First note that since Wigner matrices are symmetric, they are diagonalisable and can be written as,

$$X_N = ADA^T,$$

where D is a diagonal matrix with the eigenvalues of X_N, and A is an orthogonal matrix. By the orthogonality of A, in particular $AA^T = I$, we then have,

$$X_N - zI = ADA^T - zI = ADA^T - zAA^T = A(D - zI)A^T$$

$$\Rightarrow (X_N - zI)^{-1} = A(D - zI)^{-1}A^T.$$

Furthermore, taking the trace of each side, and using the cyclic property, we attain,

$$\operatorname{tr}[(X_N - zI)^{-1}] = \operatorname{tr}[A(D - zI)^{-1}A^T] = \operatorname{tr}[(D - zI)^{-1}] = \sum_{i=1}^{N} \frac{1}{\lambda_i^N - z}.$$

Hence, we have that,

$$S_{L_N}(z) = \frac{1}{N} \operatorname{tr}[(X_N - zI)^{-1}].$$

Now, we also define the probability distribution $\bar{L}_N = \mathbb{E}[L_N]$ by $\langle \bar{L}_N, f\rangle = \mathbb{E}[\langle L_N, f\rangle]$ for all $f \in C_b(\mathbb{R})$ [1]. It is clear we have a similar equality for the Stieltjes transform of $\bar{L}_N$,

$$S_{\bar{L}_N}(z) = \frac{1}{N}\mathbb{E}[\operatorname{tr}[(X - zI)^{-1}]]. \qquad [15]$$

Recall that for a matrix M, the resolvent of M, denoted $R_M(z)$, is given by,

$$R_M(z) = (X - zI)^{-1}.$$

Then note that,

$$
\begin{aligned}
(X_N - zI)R_{X_N}(z) = I \\
\Rightarrow X_N R_{X_N}(z) - zR_{X_N}(z) = I \\
\Rightarrow R_{X_N}(z) = -\frac{1}{z}I + \frac{1}{z}X_N R_{X_N}(z).
\end{aligned}
$$

Furthermore, we have that, by taking the trace, expectation and scaling by $\frac{1}{N}$,

$$
S_{\bar{L}_N}(z) = \frac{1}{N}\mathbb{E}[\mathrm{tr}[R_{X_N}(z)]] = -\frac{1}{z} + \frac{1}{zN}\mathbb{E}[\mathrm{tr}[X_N R_{X_N}(z)]]. \tag{3.5}
$$

Next, we require the following Lemma 5, which appears in [1] without proof. Hence, we have provided our own simple proof. Note that this lemma is where our assumption of Gaussian Wigner matrices is required. The proof of Theorem 1 for general Wigner matrices can be found in [1], but is more complicated, and doesn't add much to our discussion here.

Lemma 5 *If Y is a zero-mean variance σ^2 Gaussian random variable, then for all differentiable functions f, such that f and f' grow as a polynomial,*

$$
\mathbb{E}[Yf(Y)] = \mathbb{E}[f'(Y)]\mathbb{E}[Y^2]. \quad \textit{[1]}
$$

Proof:
Let $p(Y)$ be the probability density function of the Gaussian random variable Y. Then we proceed by integrating by parts,

$$
\mathbb{E}[Yf(Y)] = \int_{\mathbb{R}} Yf(Y)p(Y)dY.
$$

Then letting $u = f(Y)$ and $dv = Yp(Y)dY$ so $du = f'(Y)dY$ and $v = -\sigma^2 p(Y)$,

$$
\begin{aligned}
&= \left[-\sigma^2 f(Y)p(Y)\right]_{-\infty}^{\infty} + \sigma^2 \int_{-\infty}^{\infty} f'(Y)p(Y)dY \\
&= \sigma^2 \mathbb{E}[f'(Y)]. \qquad \square
\end{aligned}
$$

Consequently, it is clear that we must be able to differentiate the resolvent in order to use Lemma 5 with (3.5). Recall we are aiming to show that the expected empirical eigenvalue distribution converges to the semicircle distribution. The next section of computation to the end of the chapter is again based off [1], with supplement from [15], but with more detail added independently to make the arguments easier to follow.

Now to calculate the derivative of the resolvent, we first define the matrix $\Delta_N^{i,k}$ as the symmetric $N \times N$ matrix,

$$\Delta_N^{i,k}(j,l) = \begin{cases} 1, & \text{if } (i,k) = (j,l) \text{ or } (l,j), \\ 0, & \text{otherwise.} \end{cases}$$

Then we claim for a Wigner matrix X_N,

$$\frac{\partial R_{X_N}(z)}{\partial X_N(i,k)} = -R_{X_N}\Delta_N^{i,k}(z)R_{X_N}(z). \tag{3.6}$$

To show this, first we prove what is known as the inverse matrix derivative identity. For an invertible matrix A,

$$\begin{aligned} AA^{-1} &= I \\ \Rightarrow \frac{d}{dt}(AA^{-1}) &= \frac{d}{dt}I \\ \Rightarrow \frac{dA}{dt}A^{-1} + A\frac{dA^{-1}}{dt} &= 0 \\ \Rightarrow \frac{dA^{-1}}{dt} &= A^{-1}\frac{dA}{dt}A^{-1}. \end{aligned}$$

So taking $A = (X - zI)$, we get that,

$$\frac{\partial (X-zI)^{-1}}{\partial X(i,k)} = -(X-zI)^{-1}\frac{\partial (X-zI)}{\partial X(i,k)}(X-zI)^{-1}$$

$$\Rightarrow \frac{\partial R_{X_N}(z)}{\partial X_N(i,k)} = -R_{X_N}(z)\Delta_N^{i,k}(z)R_{X_N}(z)$$

where we use notation $X_N(i,j)$ to denote the element in the i^{th} row and j^{th} column of X_N.

So now we have,

$$\begin{aligned} \frac{1}{N}\mathbb{E}[\text{tr}[R_{X_N}(z)]] &= -\frac{1}{z} + \frac{1}{zN}\mathbb{E}[\text{tr}[X_N R_{X_N}(z)]] \\ &= -\frac{1}{z} + \frac{1}{zN}\sum_{i,j=1}^{N}\mathbb{E}[X_N(i,j)R_{X_N}(z)(j,i)], \qquad \text{by linearity of expectation} \\ &= -\frac{1}{z} + \frac{1}{zN}\sum_{i,j=1}^{N}\mathbb{E}[X_N(i,j)^2]\mathbb{E}\left[\frac{\partial R_{X_N}(z)(j,i)}{\partial X_N(j,i)}\right], \end{aligned}$$

where in the final equality, we have used the symmetry of X_N so that $X_N(i,j) = X_N(j,i)$ with Lemma 5.

Recall Definition 1, and in particular the variance of the diagonal and off-diagonal elements. For diagonal terms $i = j$,

$$\mathbb{E}[X_N(i,j)^2] = \frac{\mathbb{E}[Y_i^2]}{N} = \frac{\sigma^2}{N},$$

and,

$$\mathbb{E}\left[\frac{\partial R_{X_N}(z)(i,i)}{\partial X_N(i,i)}\right] = -\mathbb{E}[R_{X_N}(z)(i,i)R_{X_N}(z)(i,i)].$$

Then for the off-diagonal terms $i \neq j$,

$$\mathbb{E}[X_N(i,j)] = \frac{\mathbb{E}[Z_{i,j}^2]}{N} = \frac{1}{N},$$

and,

$$\mathbb{E}\left[\frac{\partial R_{X_N}(z)(j,i)}{\partial X_N(j,i)}\right] = -\mathbb{E}[R_{X_N}(z)(i,j)R_{X_N}(z)(i,j) + R_{X_N}(z)(i,i)R_{X_N}(z)(j,j)].$$

So summing over i, j, we get that,

$$\begin{aligned}\sum_{i,j=1}^{N} &\mathbb{E}[X_N(i,j)^2]\mathbb{E}\left[\frac{\partial R_{X_N}(z)(j,i)}{\partial X_N(j,i)}\right] \\ &= -\frac{1}{N}\sum_{i,j=1;i\neq j}^{N} \mathbb{E}[R_{X_N}(z)(i,j)R_{X_N}(z)(i,j) + R_{X_N}(z)(i,i)R_{X_N}(z)(j,j)] \\ &\quad - \frac{\sigma^2}{N}\sum_{i=1}^{N}\mathbb{E}[R_{X_N}(z)(i,i)R_{X_N}(z)(i,i)].\end{aligned}$$

Hence, we have that,

$$\begin{aligned}\frac{1}{N}\mathbb{E}[\mathrm{tr}[R_{X_N}(z)]] &= -\frac{1}{z} + \frac{1}{zN}\sum_{i,j=1}^{N}\mathbb{E}[X_N(i,j)^2]\mathbb{E}\left[\frac{\partial R_{X_N}(z)(j,i)}{\partial X_N(j,i)}\right] \\ &= -\frac{1}{z} - \frac{1}{zN^2}\sum_{i,j=1;i\neq j}^{N}\mathbb{E}[R_{X_N}(z)(i,j)R_{X_N}(z)(i,j) + R_{X_N}(z)(i,i)R_{X_N}(z)(j,j)] \\ &\quad - \frac{\sigma^2}{zN^2}\sum_{i=1}^{N}\mathbb{E}[R_{X_N}(z)(i,i)R_{X_N}(z)(i,i)].\end{aligned}$$

Noting that,

$$\sum_{i,j=1}^{N} \mathbb{E}[R_{X_N}(z)(i,j)R_{X_N}(z)(i,j)+R_{X_N}(z)(i,i)R_{X_N}(z)(j,j)] = \mathbb{E}[\mathrm{tr}(R_{X_N}(z)^2)+\mathrm{tr}(R_{X_N}(z))^2],$$

so that in the sum $\sum\limits_{i,j=1;i\neq j}$ we are over-counting, we get that,

$$\begin{aligned}\frac{1}{N}\mathbb{E}[\mathrm{tr}[R_{X_N}(z)]] &= -\frac{1}{z} - \frac{1}{zN^2}\mathbb{E}[\mathrm{tr}(R_{X_N}(z)^2) + \mathrm{tr}(R_{X_N}(z))^2] - \frac{1}{zN^2}\sum_{i=1}^{N}(\sigma^2-2)\mathbb{E}[R_{X_N}(i,i)^2] \\ &= -\frac{1}{z} - \underbrace{\frac{1}{zN}\langle \bar{L}_N, \frac{1}{(x-z)^2}\rangle}_{\text{(i)}} - \frac{1}{z}\underbrace{\langle \bar{L}_N, \frac{1}{x-z}\rangle^2}_{\text{(ii)}} - \underbrace{\frac{\sigma^2-2}{zN^2}\mathbb{E}[\mathrm{tr}(R^2_{X_N})]}_{\text{(iii)}} \\ &= S_{\bar{L}_N(z).}\end{aligned}$$

So we now want to analyse the above expression as $N \to \infty$, aiming for convergence to the Stieltjes transform of the semicircle distribution. First note that (ii) is simply $S^2_{\bar{L}_N(z)}$. Then for (i), since $\frac{1}{(x-z)^2}$ is bounded for $z \in \mathbb{C} \setminus \mathbb{R}$, as $N \to \infty$, the expression tends to zero. Finally, for (iii), note that $\mathbb{E}[\mathrm{tr}(R^2_{X_N})]$ is at most of order N, and so the full expression is order N^{-1} and tends to zero as $N \to \infty$ also. Thus we have the quadratic equation for any limit point s of $S_{\bar{L}_N}(z)$,

$$s = -\frac{1}{z} - \frac{1}{z}s^2.$$

Solving the quadratic, noting that by definition s must have positive imaginary part, we have that

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

So since $s(z)$ equals the Stieltjes transform for the semicircle distribution for $|z| > 2$, then by analyticity, it is equivalent for all $z \in \mathbb{C} \setminus \mathbb{R}$. Hence, we have that the expectation of the empirical eigenvalue distribution converges to the semicircle distribution by Theorem 3. All that is left to do is consider the variance of the empirical eigenvalue distribution in the limit $N \to \infty$. If the variance tends to zero, then by Chebyshev's Inequality, we get the convergence in probability we require.

To see this, assume we have some sequence of random variables Y_n whose expectation $\mathbb{E}[Y_n]$ converges to some deterministic limit Y. Hence, for all $\epsilon > 0$ there exists an n large enough such that $|\mathbb{E}[Y_n] - Y| < \frac{\epsilon}{2}$. Then we use the triangle inequality to get,

$$|Y_n - Y| \leq |Y_n - \mathbb{E}[Y_n]| + |\mathbb{E}[Y_n] - Y|.$$

So we get, by using Chebyshev's Inequality,

$$\mathbb{P}(|Y_n - Y| > \epsilon) \leq \mathbb{P}(|Y_n - \mathbb{E}[Y_n]| > \frac{\epsilon}{2}) < \frac{4\mathbb{V}\mathrm{ar}[Y_n]}{\epsilon^2},$$

and so this will tend to zero if the variance of Y_n tends to zero, leaving Y_n converging to Y in probability. Hence, next we will consider how to sufficiently bound the variance of S_{L_N} in order to complete the proof of Wigner's Semicircle Theorem.

3.4 Concentration Phenomena

We will now consider **concentration phenomena** in Gaussian Wigner matrices. Concentration phenomena generally refers to instances where the expectation of a random variable, is close to realisations of the random variable. As discussed in the previous subchapter, if as N grows, the variance of Stieltjes transform of the empirical eigenvalue distribution vanishes, then we will have the concentration we would like to finish the proof of Theorem 1.

In order to achieve this, we are going to mostly follow the results and proofs from [15]. However, again, adaptations have been made to proofs so that they are applicable for general variance on the diagonal elements. Note that by the end of this chapter, we will have that the variance decays at rate order N. In [1], they achieve decay at an exponential rate, however, this relies of much heavier analysis, and is more than enough for our proof here. Now, first we must define what is means for a random variable to satisfy a Poincaré inequality.

Definition 7 *A random variable $X = (X_1, ..., X_n) : \Omega \to \mathbb{R}^n$ satisfies a **Poincaré inequality** with constant $c > 0$, if for any differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ with $\mathbb{E}[f(X)^2] < \infty$, we get*

$$\mathbb{V}ar[f(X)] \leq c \cdot \mathbb{E}\left[\|\nabla f(X)\|_2^2\right] \quad \textit{where} \quad \|\nabla f(X)\|_2^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2. \qquad \textit{[15]}$$

Now, we are able to start bounding the variance of a function of a random variable. This first result, Theorem 4, the Efron-Stein Inequality, is a widely known result from probability theory. The statement and body of the proof are from [15].

Theorem 4 ***(Efron-Stein Inequality)*** *Let $X_1, \ldots, X_n$ be independent random variables, and let $f(X_1, \ldots, X_n)$ be a square-integrable function of $X = (X_1, \ldots, X_n)$. Then we have*

$$\mathbb{V}ar[f(X)] \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{V}ar_i[f(X)]\right],$$

where $\mathbb{V}ar_i$ denotes taking the variance with respect to the i-th variable, keeping all other variables fixed, and the expectation is then taken over all the other variables. [15]

Proof:
We begin by defining the distribution of each random variable X_i as μ_i, where μ_i is a probability measure on $\mathbb{R}$. Since $X_1, \ldots, X_n$ are independent, the joint distribution of the random vector $X = (X_1, \ldots, X_n)$ is given by the product measure $\mu_1 \times \cdots \times \mu_n$ on $\mathbb{R}^n$. Let $Z = f(X_1, \ldots, X_n)$, so then the expectation and variance of Z can then be expressed as:

$$\mathbb{E}[Z] = \int_{\mathbb{R}^n} f(x_1, \ldots, x_n)\, d\mu_1(x_1) \cdots d\mu_n(x_n),$$

$$\mathbb{V}\mathrm{ar}[Z] = \mathbb{E}\left[(Z - \mathbb{E}[Z])^2\right] = \int_{\mathbb{R}^n} (f(x_1, \ldots, x_n) - \mathbb{E}[Z])^2\, d\mu_1(x_1) \cdots d\mu_n(x_n).$$

To analyse the variance, we decompose $Z - \mathbb{E}[Z]$ into a sum of terms that capture the contribution of each variable X_i. Specifically, we write,

$$Z - \mathbb{E}[Z] = (Z - \mathbb{E}_1[Z]) + (\mathbb{E}_1[Z] - \mathbb{E}_{1,2}[Z]) + \cdots + (\mathbb{E}_{1,2,\ldots,n-1}[Z] - \mathbb{E}[Z]),$$

where $\mathbb{E}_{1,\ldots,i}[Z]$ denotes the conditional expectation of Z with respect to the variables $X_1, \ldots, X_i$, treating $X_{i+1}, \ldots, X_n$ as fixed. Each term in this decomposition represents the incremental effect of adding one more variable to the conditioning. Now let $\Delta_i := \mathbb{E}_{1,\ldots,i-1}[Z] - \mathbb{E}_{1,\ldots,i}[Z]$. This allows us to express the difference $Z - \mathbb{E}[Z]$ as,

$$Z - \mathbb{E}[Z] = \sum_{i=1}^{n} \Delta_i.$$

The variance of Z can now be written as,

$$\mathbb{V}\mathrm{ar}[Z] = \mathbb{E}\left[\left(\sum_{i=1}^{n} \Delta_i\right)^2\right] = \sum_{i=1}^{n} \mathbb{E}[\Delta_i^2] + \sum_{i \neq j} \mathbb{E}[\Delta_i \Delta_j].$$

We claim that the cross-terms $\mathbb{E}[\Delta_i \Delta_j]$ vanish for $i \neq j$. To see this, consider the case $i < j$. The term Δ_i depends only on $X_1, \ldots, X_i$, while Δ_j depends on $X_1, \ldots, X_j$. However, when taking the expectation $\mathbb{E}[\Delta_i \Delta_j]$, the integration over X_i (or X_j) causes the product $\Delta_i \Delta_j$ to vanish due to the orthogonality of the terms. Specifically,

$$\mathbb{E}[\Delta_i \Delta_j] = \mathbb{E}\left[\mathbb{E}[\Delta_i \Delta_j \mid X_1, \ldots, X_{i-1}]\right] = 0,$$

because Δ_i and Δ_j are uncorrelated when $i \neq j$. Thus, the variance simplifies to,

$$\mathbb{V}\mathrm{ar}[Z] = \sum_{i=1}^{n} \mathbb{E}[\Delta_i^2]. \tag{3.7}$$

Each term Δ_i^2 can be interpreted as the conditional variance of $f(X)$ with respect to the variable X_i, holding the other variables fixed. Specifically,

$$\Delta_i^2 = \mathbb{V}\mathrm{ar}_i[f(X)],$$

where $\mathbb{V}\mathrm{ar}_i[f(X)]$ denotes the variance of $f(X)$ with respect to X_i, conditioned on $X_1, \ldots, X_{i-1}$. Then by Jensen's inequality, since the variances are non-negative, it is bounded below by the function of the expectation, we have,

$$\mathbb{E}[\Delta_i^2] \leq \mathbb{E}\left[\mathbb{V}\mathrm{ar}_i[f(X)]\right]. \tag{3.8}$$

This inequality ensures that the contribution of each Δ_i^2 to the total variance is appropriately bounded. Combining (3.7) and (3.8), we obtain the desired inequality,

$$\mathbb{V}\mathrm{ar}[f(X)] \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{V}\mathrm{ar}_i[f(X)]\right]. \qquad \square$$

This next result, Lemma 6, follows quickly from the Efron-Stein Inequality, Theorem 4. It allows us to make comments on the Poincaré inequality of random multi-variables, and again follows what is said in [15].

Lemma 6 *Let $X_1, .., X_n$ be independent real random variables such that each X_i satisfies a Poincaré inequality with constant c_i. Then $X = (X_1, .., X_n)$ satisfies a Poincaré inequality in $\mathbb{R}^n$ with constant $c = \max\{c_1, ..., c_n\}$. [15]*

Proof:
Firstly, by Theorem 4, the Efron-Stein Inequality,

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[f(X)] &\leq \sum_{i=1}^{N} \mathbb{E}[\mathbb{V}\mathrm{ar}_i[f(X)]] \\
&\leq \sum_{i=1}^{N} \mathbb{E}\left[c_i \mathbb{E}_i\left[\left(\frac{\partial f}{\partial x_i}\right)^2\right]\right] \qquad \text{by Poincaré inequality} \\
&\leq c \sum_{i=1}^{N} \mathbb{E}_i\left[\left(\frac{\partial f}{\partial x_i}\right)^2\right] \\
&= c\, \mathbb{E}\left[\|\nabla f(X)\|^2\right]. \qquad \square
\end{aligned}$$

Now we are ready to state and prove the main required result, Theorem 5, the Gaussian Poincaré Inequality. This is what will allow us to finish the proof of Wigner's Semicircle Theorem, Theorem 1. Note that again this result limits our proof for Gaussian Wigner matrices, and the proof for general Wigner matrices can be found in [1], but doesn't add to our discussion here. Also note that the proof of this theorem is based on the work in [6], but with more detail added, for example, to fully justify the convergence in the variance of $f(S_n)$.

Theorem 5 ***(Gaussian Poincaré Inequality)*** *Let $X = (X_1, ..., X_n)$ be a vector of independent standard Gaussian random variables, and $f : \mathbb{R}^n \to \mathbb{R}$ be any continuously differentiable function. Then X satisfies a Poincaré inequality with constant 1, i.e.*

$$\mathbb{V}ar[f(X)] \leq \mathbb{E}\left[\|\nabla f(X)\|^2\right]. \qquad \textit{[6]}$$

Proof:
First we assume that $\mathbb{E}[\|\nabla f(X)\|^2] < \infty$, as otherwise we are immediately done. Then by Lemma 6, we only need to prove the statement for $n = 1$.

First we let $\epsilon_1, ..., \epsilon_n$ be an independent and identically distributed sequence of random variables such that,

$$\mathbb{P}(\epsilon_1 = 1) = \mathbb{P}(\epsilon_1 = -1) = \frac{1}{2},$$

and we set $S_n = n^{-\frac{1}{2}} \sum\limits_{j=1}^{n} \epsilon_j$.

Then since we have for all i,

$$\mathbb{V}\mathrm{ar}_i[f(S_n)] = \frac{1}{4}\left(f(S_n + \frac{1-\epsilon_i}{\sqrt{n}}) - f(S_n - \frac{1+\epsilon_i}{\sqrt{n}})\right)^2,$$

by the Efron-Stein inequality, Theorem 4,

$$\mathbb{V}\mathrm{ar}[f(S_n)] \leq \frac{1}{4}\sum_{i=1}^{n}\mathbb{E}\left[\left(f(S_n + \frac{1-\epsilon_i}{\sqrt{n}}) - f(S_n - \frac{1+\epsilon_i}{\sqrt{n}})\right)^2\right].$$

Next note that by the Central Limit Theorem, S_n converges in distribution to a standard Gaussian as n grows. Then by the Continuous Mapping Theorem, which states that limits of random variables can pass through continuous functions [5], since f is continuously differentiable and X is a standard Gaussian, $f(S_n)$ converges in distribution to $f(X)$. Then, because $\mathbb{E}[\|\nabla f(X)\|^2] < \infty$, f has square-integrable derivatives, and so f grows at most linearly in expectation.

This means $f(S_n)$ is uniformly integrable. Now, convergence in distribution and uniform integrability implies the convergence in moments,

$$\mathbb{E}[f(S_n)] \to \mathbb{E}[f(X)] \qquad \mathbb{E}[f(S_n)^2] \to \mathbb{E}[f(X)^2].$$

Consequently, we get that $\mathbb{V}\mathrm{ar}[f(S_n)] \to \mathbb{V}\mathrm{ar}[f(X)]$. Then letting $K = \sup_x |f''(x)|$, by Taylor's Theorem, we have that,

$$\left| f\left(S_n + \frac{1-\epsilon_i}{\sqrt{n}}\right) - f\left(S_n - \frac{1+\epsilon_i}{\sqrt{n}}\right) \right| \leq \frac{2}{\sqrt{n}}|f'(S_n)| + \frac{2K}{n}$$

$$\Rightarrow \frac{n}{4}\left(f(S_n + \frac{1-\epsilon_i}{\sqrt{n}}) - f(S_n - \frac{1+\epsilon_i}{\sqrt{n}})\right)^2 \leq f'(S_n)^2 + \frac{2K}{\sqrt{n}}|f'(S_n)| + \frac{K^2}{n}.$$

Combining this result with the Central Limit Theorem gives us that,

$$\limsup_{n\to\infty} \frac{1}{4}\sum_{i=1}^{n} \mathbb{E}\left[\left(f(S_n + \frac{1-\epsilon_i}{\sqrt{n}}) - f(S_n - \frac{1+\epsilon_i}{\sqrt{n}})\right)^2\right] = \mathbb{E}[f'(X)^2],$$

and so we are done. □

Now, we apply these results to Gaussian Wigner matrices, following the arguments in [15], adjusting for general variances. First recall that $\mathbb{E}[X_N(i,j)] = 0$ and $\mathbb{V}\mathrm{ar}[X_N(i,i)] = \frac{\sigma^2}{N}$, $\mathbb{V}\mathrm{ar}[X_N(i,j)] = \frac{1}{N}$. Then by a change of variables and the Gaussian Poincaré Inequality,

$$\mathbb{V}\mathrm{ar}[f(X_N)] \leq \frac{\sigma^2}{N}\mathbb{E}[\|\nabla f(X_N)^2\|].$$

Now we let the function $g(X_N) := \mathrm{tr}[R_{X_N}(z)]$ since we want to bound the variance of $\mathrm{tr}[R_{X_N}(z)]$. However, g is complex-valued, so we bound first as,

$$|\mathbb{V}\mathrm{ar}[g(X_N)]| \leq 2\left(\mathbb{V}\mathrm{ar}[\mathrm{Re}(g(X_N))] + \mathbb{V}\mathrm{ar}[\mathrm{Im}(g(X_N))]\right).$$

Now we want to bound $f(X_N) := \mathrm{Re}(g(X_N))$ by Gaussian Poincaré Inequality, so first recall its partial derivatives,

$$\left|\frac{\partial f(X_N)}{\partial X_N(i,j)}\right| \leq \frac{\sigma^2}{N}\left|R^2_{X_N}(i,j)\right| \leq \frac{\sigma^2}{N}\left|R^2_{X_N}(i,j)\right| \leq \frac{\sigma^2}{N(\mathrm{Im}(z))^2}.$$

Hence we have,

$$\left|\frac{\partial f(X_N)}{\partial X(i,j)}\right|^2 \leq \frac{\sigma^4}{N^2(\mathrm{Im}(z))^4}.$$

Then by Gaussian Poincaré Inequality,

$$\mathbb{V}\mathrm{ar}[f(X_N)] \leq \frac{\sigma^2}{N}\sum_{i,j}\left|\frac{\partial f(X_N)}{\partial X_N(i,j)}\right|^2 \leq \frac{\sigma^6}{N(\mathrm{Im}(z))^4}.$$

We can bound $\mathrm{Im}(g(X_N))$ in a symmetrical way and so we get that

$$\mathbb{V}\mathrm{ar}[\mathrm{tr}[R_{X_N}(z)]] \leq \frac{2\sigma^6}{N(\mathrm{Im}(z))^4} \xrightarrow{N\to\infty} 0,$$

and so we have the bound required which finally finishes the proof of Wigner's Semicircle Theorem for Gaussian Wigner matrices. As mentioned throughout this chapter, a proof for more general Wigner matrices can be found in [1]. However, this is surplus to what we want to discuss here.

Chapter 4

Eigenvalue Spacing

4.1 Global vs. Local Phenomena

So far, our discussion has focused on **global phenomena** in RMT. Global phenomena refers to properties that describe the entire spectrum of a random matrix or the bulk behaviour of its eigenvalues. These global results provide a macroscopic view of the eigenvalue distribution and are fundamental to understanding the overall structure of random matrices. For example, we have seen that the eigenvalue distribution of Wigner matrices follows the semicircle distribution, in Theorem 1, which describes the density of eigenvalues over the full range of the spectrum. In fact, Wigner's Semicircle Theorem, Theorem 1, is one of the most renowned results in RMT, and is often the first result in introductory texts such as [1] and [15].

However, RMT also studies **local phenomena**, which focus on microscopic details of the eigenvalue distribution. Local phenomena include the analysis of spacings between neighbouring eigenvalues and the distribution of individual eigenvalues, in particular the largest or smallest. Local eigenvalue statistics play a key role in many practical applications. In Principal Component Analysis (PCA), for instance, the largest eigenvalue of a covariance matrix often indicates the presence of a dominant factor or signal in high-dimensional data [11]. Similarly, in statistical inference and machine learning, extreme eigenvalues help distinguish meaningful structure from noise. An example of a local phenomenon result concerning the largest eigenvalue of a Wigner matrix is the Tracy-Widom Theorem, Theorem 6. We state the theorem as an illustrative example, without proof. Note that the full statement of the theorem requires some more complicated asymptotic properties that we don't state as it would require a lot more prerequisite material, without adding much to the general discussion.

Theorem 6 *Consider a Wigner matrix X_N whose largest eigenvalue is λ_N^N. Then for all $-\infty < t \leq \infty$,*

$$\lim_{N\to\infty} \mathbb{P}(N^{\frac{2}{3}}(\lambda_N^N - 2) \leq t) = \exp\left(-\int_t^\infty (x-t)q(x)^2 dx\right) =: \mathscr{F}(t),$$

where:

- *q satisfies $q'' = tq + 2q^3$ and also satisfies some asymptotic properties,*
- *$\mathscr{F}(t)$ is called the Tracy-Widom distribution. [1]*

A fundamental principle in RMT is the **universality** of local fluctuations, meaning that the results stand for a large variety of random matrices. Remarkably, local eigenvalue statistics are often independent of the specific details of the matrix ensemble, depending only on broad symmetry classes (e.g., orthogonal, unitary). To isolate local fluctuations from the influence of the global eigenvalue density, we use a technique called **unfolding** [3]. Unfolding involves transforming eigenvalues so that their mean spacing is uniform across the spectrum. This step is crucial when studying eigenvalue spacing distributions or eigenvalue correlations, as it ensures that the observed fluctuations reflect intrinsic properties of the system, rather than variations in the global eigenvalue density. Without unfolding, the local statistics would be distorted by the macroscopic shape of the eigenvalue distribution, making meaningful comparisons very difficult.

We will now investigate the distribution of spacings between adjacent eigenvalues as an example of local phenomena. Here, in particular, the unfolding process is important to normalise the mean spacing. However, first we will define a new class of random matrices.

4.2 Gaussian Orthogonal Ensemble

For our exploration of eigenvalue spacings, we will define a new ensemble of random matrices, less general than Wigner matrices. There are three Wigner ensembles of random matrices regularly used across the literature called the Gaussian Orthogonal Ensemble (GOE), Gaussian Unitary Ensemble (GUE), and the Gaussian Symplectic Ensemble (GSE). We will focus on the GOE defined below. Our definition here follows the convention used in [1], although in a slightly different way. In particular, [1], samples all elements from the standard Normal distribution, but scales the diagonal elements by $\sqrt{2}$. Hence our definition uses the same distributions, just mathematically expressed slightly differently.

Definition 8 *Let H be an $N \times N$ matrix with independent elements $(H_{i,j})_{i,j=1}^{N}$, up to symmetry, which follow Gaussian distributions as follows:*

- $H_{i,i} \sim N(0,2)$
- $H_{i,j} \sim N(0,1)$.

Then H is a member of the Gaussian orthogonal ensemble (GOE).

Note that if we scale the elements by $N^{-\frac{1}{2}}$, then we get Gaussian Wigner matrices, which hence will abide Wigner's Semicircle Theorem for their empirical eigenvalue distribution, Theorem 1. Also note that, again our class of matrices is symmetric, hence we have guaranteed ourself real eigenvalues which can be ordered. We are going to build to a theory on how the eigenvalue spacing for GOE matrices behaves, and so first we must compute the joint probability distributions for the matrix elements. Theorem 7 is an easily proved result which is presented without proof in [1] and [15], but we show a full computation here for completeness.

Theorem 7 *The joint probability distribution for a $N \times N$ GOE matrix H with elements $H_{i,j}$ is,*

$$C_N e^{-\frac{1}{4}\,\mathrm{Tr(H^2)}} dH,$$

where C_N is a constant dependent on N, and $dH = \prod_{1 \le i < j \le N} dH_{i,j}$.

Proof:
Since the matrix elements are independent up to symmetry, Normally distributed random variables, we immediately get that the joint pdf is,

$$\begin{aligned}
\prod_{i=1}^{N} \frac{1}{\sqrt{4\pi}} e^{-\frac{1}{4}H_{i,i}^2} \prod_{1 \le i<j \le N} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}H_{i,j}^2} &= C_N \exp\left(-\frac{1}{4}\left(\sum_{i=1}^{N} H_{i,i}^2 + 2\sum_{1 \le i<j \le N} H_{i,j}^2\right)\right) \\
&= C_N \exp\left(-\frac{1}{4}\left(\sum_{i=1}^{N} H_{i,i}^2 + \sum_{i \ne j} H_{i,j}^2\right)\right), \qquad \text{by symmetry} \\
&= C_N \exp\left(-\frac{1}{4}\left(\sum_{i,j=1}^{N} H_{i,j}^2\right)\right) \\
&= C_N \exp\left(-\frac{1}{4}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} H_{i,j}^2\right)\right) \\
&= C_N \exp\left(-\frac{1}{4}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} H_{i,j}H_{j,i}\right)\right), \qquad \text{by symmetry} \\
&= C_N \exp\left(-\frac{1}{4}\sum_{i=1}^{N} \mathrm{Tr}(H^2)\right).
\end{aligned}$$

Hence, it is now clear that the joint probability distribution is,

$$C_N e^{-\frac{1}{4}\,\mathrm{Tr(H^2)}} \prod_{1\le i<j\le N} dH_{i,j}. \qquad \square$$

This means that for any continuous bounded function $f \in C_b(\mathbb{R})$, we can compute the expectation as,

$$\mathbb{E}[f(H)] = \int_{\mathbb{R}} C_N f(H) e^{-\frac{1}{4}\,\mathrm{Tr(H^2)}} dH.$$

4.3 Wigner Surmise

We now want to explore the behaviour of spacings between adjacent eigenvalues for GOE matrices. In order to do this, we scale the eigenvalues to have mean level spacing of one first. This removes any global effects from the limiting spectral density and so only microscopic correlations remain. This is the unfolding process we discussed in Chapter 4.1. This allows us to make meaningful comparisons of eigenvalue spacings across the entire eigenvalue distribution. We will give a quick example of how this process works for a 4 × 4 GOE matrix.

Example 4 Let's sample a 4 × 4 GOE matrix H:

$$H = \begin{pmatrix} 0.123 & -0.568 & 0.346 & -0.235 \\ -0.568 & 0.988 & -0.457 & 0.679 \\ 0.346 & -0.457 & -0.654 & 0.123 \\ -0.235 & 0.679 & 0.123 & 0.543 \end{pmatrix}.$$

Now the eigenvalues are computed to be $-0.988, -0.654, 0.543, 1.235$, and so the eigenvalue spacings are $0.334, 1.197, 0.692$. Now to normalise these to having mean 1, we first must calculate the mean eigenvalue spacing which is 0.741. Dividing our spacings by this gives our normalise eigenvalue spacings $0.451, 1.615, 0.934$.

The next result we prove, Theorem 8, gives an explicit formula for the probability density of the eigenvalue spacing in 2 × 2 GOE matrices. The statement of the result, and idea behind the proof is from [3], but there is a lack of detail in the proof there. Again we show the full computation for completeness.

Theorem 8 ***The Wigner Surmise*** *For a 2 × 2 GOE matrix H with eigenvalues λ_1, λ_2, the Wigner surmise is the probability density of the eigenvalue spacing $s = |\lambda_1 - \lambda_2|$ and is,*

$$P(s) = \frac{\pi}{2} s e^{-\frac{\pi}{4}s^2}.$$

Proof:
To start the proof, we take some GOE matrix,

$$\tilde{H} = \begin{pmatrix} \tilde{h}_{11} & \tilde{h}_{12} \\ \tilde{h}_{12} & \tilde{h}_{22} \end{pmatrix},$$

where $\tilde{h}_{11}, \tilde{h}_{22} \sim N(0,2)$ and $\tilde{h}_{12} \sim N(0,1)$. Then by Theorem 7, we know that the joint probability density function of $\tilde{H}$ is proportional to,

$$\exp\left\{-\frac{1}{2}\operatorname{Tr}(\tilde{H}^2)\right\} = \exp\left\{-\frac{1}{2}(\tilde{h}_{11}^2 + \tilde{h}_{22}^2 + 2\tilde{h}_{12}^2)\right\}.$$

Now, in order to achieve a standardised meal eigenvalue spacing, we will later have to scale our matrix. Hence, we define a matrix $H := K\tilde{H}$ for some real positive constant K. Then simple eigenvalue calculations give that the eigenvalues of $\tilde{H}$ are,

$$\tilde{\lambda}_1, \tilde{\lambda}_1 = \frac{\tilde{h}_{11} + \tilde{h}_{22}}{2} \pm \frac{\sqrt{(\tilde{h}_{11} - \tilde{h}_{22})^2 + 4\tilde{h}_{12}^2}}{2},$$

and the eigenvalues λ_1, λ_2 of H are simply scaled by K. The eigenvalue spacing for $\tilde{H}$, denoted $\tilde{s} = |\tilde{\lambda}_1 - \tilde{\lambda}_2|$ is then given by,

$$\tilde{s} = \sqrt{(\tilde{h}_{11} - \tilde{h}_{22})^2 + 4\tilde{h}_{12}^2},$$

and again, for H we scale by K.

If we let $v = \tilde{h}_{11} - \tilde{h}_{22} \sim N(0,4)$, and $w = 2\tilde{h}_{12} \sim N(0,4)$, then the average spacing $\tilde{s} = \sqrt{v^2 + w^2}$. Now since v and w are independent Gaussian distributions, their joint probability density function is,

$$P(v,w) \propto e^{-\frac{v^2+w^2}{8}}.$$

If we then transform v, w to polar coordinates as $v = r\cos(\theta), w = r\sin(\theta)$, the joint probability density function of r and θ, recalling the Jacobian of the transformation is r, is proportional to $re^{-\frac{r^2}{8}}$. If we then note that $\tilde{s} = r$, and marginalise out θ, we get the pdf for $\tilde{s}$ is,

$$P(\tilde{s}) \propto \tilde{s}e^{-\frac{\tilde{s}^2}{8}}.$$

Normalising this, by integrating over zero to infinity, since the spacings are by definition positive, we get that the final pdf is,

$$P(\tilde{s}) = \frac{1}{4}\tilde{s}e^{-\frac{-\tilde{s}^2}{8}}.$$

Now, to calculate the pdf of $s = K\tilde{s}$, we note that $|\frac{d\tilde{s}}{ds}| = \frac{1}{K}$ and $\tilde{s} = \frac{s}{K}$. This means that

$$P(s) = \frac{s}{4K^2}e^{-\frac{s^2}{8K^2}}.$$

To finish the computation, we now calculate the value of K by normalising the mean of s to be 1.

$$\mathbb{E}[s] = \int_0^\infty sP(s) = \int_0^\infty \frac{s^2}{4K^2} e^{-\frac{s^2}{8K^2}} ds.$$

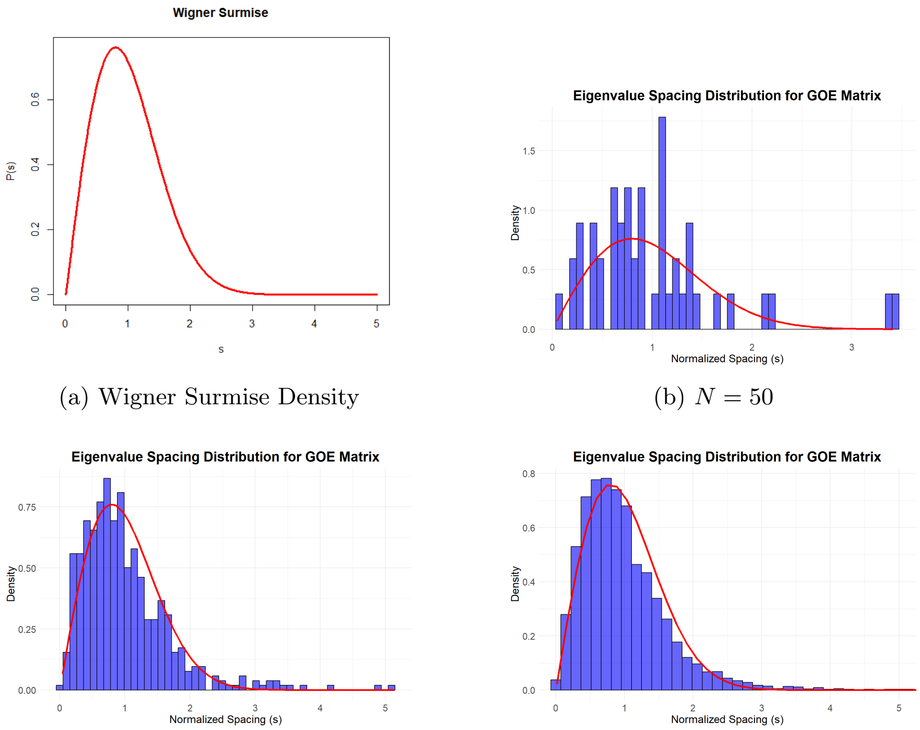
Make the substitution $x = \frac{s^2}{8K^2}$ so that $dx = \frac{2s}{8K^2} ds = \frac{\sqrt{2}K}{\sqrt{x}}$,

$$\begin{aligned}\mathbb{E}[s] &= \int_0^\infty 2xe^{-x} \frac{\sqrt{2}K}{\sqrt{x}} dx \\ &= 2\sqrt{2}K \int_0^\infty x^{\frac{1}{2}} e^{-x} dx \\ &= \sqrt{2\pi}K = 1.\end{aligned}$$

Hence, $K = \frac{1}{\sqrt{2\pi}}$, which gives our desired pdf,

$$\begin{aligned}P(s) &= \frac{s}{4\frac{1}{2\pi}} \exp(-\frac{s^2}{8\frac{1}{2\pi}}) \\ &= \frac{\pi s}{2} e^{-\frac{\pi s^2}{4}}.\end{aligned}$$ □

The Wigner surmise distribution has been plotted in Figure 4.1a. Note that there is zero density when the spacing $s = 0$, so there is effectively a repulsion between neighbouring eigenvalues. This is also clear from the definition of $P(s)$, that $P(0) = 0$, and is characteristic of random matrix statistics. This behaviour directly mirrors Wigner's original observations when modelling nuclear energy levels. He noted that these levels rarely cross, making the repulsive property a natural fit. There is also very little density as $s > 3$, and a much larger concentration of density around the mean spacing, $s = 1$.



(a) Wigner Surmise Density

(b) $N = 50$

(c) $N = 500$

(d) $N = 5000$

Figure 4.1: Comparison of Eigenvalue Spacing Distributions for Different Matrix Sizes N with the Wigner Surmise Density Shown for Reference

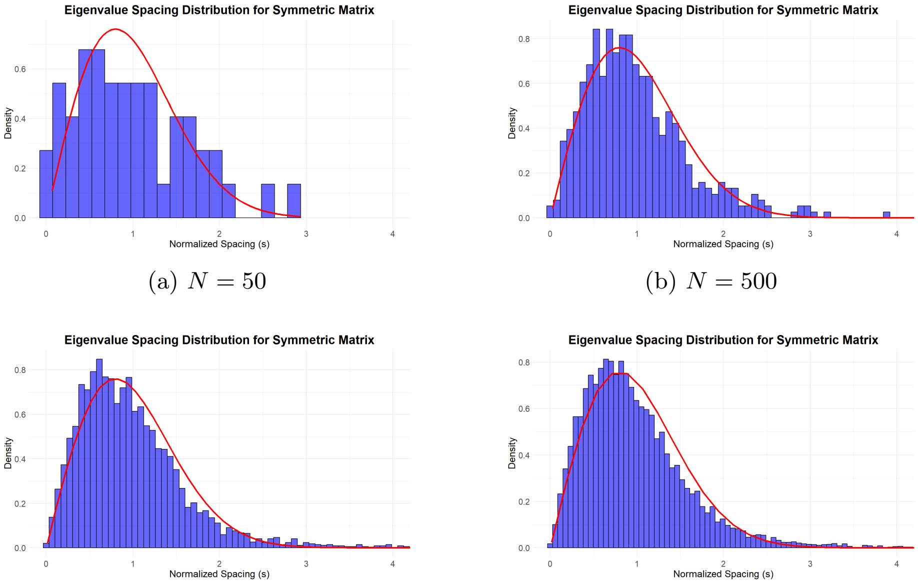
Note it is only possible to easily prove the closed form Wigner Surmise for 2×2 GOE matrices. However, for larger GOE matrices, their eigenvalue spacing distribution closely follow the same distribution to a high accuracy, provided they have been scaled to have a mean spacing of one. Figure 4.1 shows the eigenvalue spacing distribution for sampled GOE matrices of size $N = 50, 500, 5000$. It is clear that for the large N, the Wigner surmise gives a good approximation for our sampled matrices.

However, note that there is no closed form for of the eigenvalue spacing distribution for larger N, due to the increased complexity of the joint eigenvalue distribution. To illustrate this point, note the joint eigenvalue distribution of GOE matrices for general N is:

$$P(\lambda_1, ..., \lambda_N) \propto \prod_{i<j} |\lambda_j - \lambda_i| \exp\left(-\sum_{i=1}^{N} \frac{\lambda_i^2}{4}\right). \qquad [1]$$

It is clear that this becomes much more complicated as $N > 2$, since it becomes intractable to integrate. The eigenvalues are also highly correlated due to the inherent symmetry of GOE matrices, making the calculations difficult.

Another point of interest, is that in our computation of the Wigner surmise, the variance was general through the real constant K, and is then normalised to set the mean spacing to one. This means we might think the Wigner surmise should also be approximately true for symmetric matrices where individual elements can be taken from any zero-mean Gaussian distribution. There is evidence of this being true in Figure 4.2, where every element of the sampled matrices (up to symmetry) has come from a different zero-mean Normal distribution. It is evident that for large N, again the Wigner surmise is a good approximation for the eigenvalue spacing distribution of the such sampled matrices.



(a) $N = 50$

(b) $N = 500$

(c) $N = 5,000$

(d) $N = 10,000$

Figure 4.2: Comparison of Eigenvalue Spacing Distributions for Symmetric Gaussian Matrices of Different Sizes whose Elements Have Different Variances, with the Wigner Surmise Density for Reference

4.4 Universality

Through Chapter 4, we have vaguely discussed the idea of universality in random matrices, but we will now state what this means a little more formally, following the discussion in [3]. Universality refers to the phenomenon that certain properties of special random matrix ensembles, e.g. the GOE, remain true for more general random matrices, provided they share key features. For example, the Wigner Semicircle Theorem, Theorem 1, is true for GOE matrices, but also a more general group the Wigner matrices. However, the convergence to the semicircle law relies on independence, up to symmetry, of the elements in the random matrix. This is unlikely to be seen in matrices derived from real systems, which will be discussed further in Chapter 5. This property of a random matrix ensemble is referred to as a **macroscopic** property since the matrix is normalised by $\frac{1}{\sqrt{N}}$. In particular, this means that the average eigenvalue distance will be of order $\frac{1}{\sqrt{N}}$.

In comparison, the **microscopic** scale is when the normalisation of the matrix ensemble is such that the average eigenvalue distance is of order 1. On this scale, random matrices display strong universality. Wigner conjectured that on this scale, certain properties of GOE, for example the Wigner surmise, should hold for general symmetric matrices. In this Chapter, we have seen evidence of this somewhat being true in our plots. However, we have still only discussed the instances where elements have been drawn from a zero-mean Gaussian. The key takeaway from this short formalised introduction to universality, and Chapter 4 as a whole, is that local results display much more universality than global results, since the normalisation on microscopic scale removes the influence from the global eigenvalue density.

Chapter 5

Applications to Neural Networks

5.1 Introduction to Neural Networks

We are now going to explore one of the many applications of RMT: neural networks. So first, we must give a brief introduction into what neural networks are, and where large matrices appear. We will follow the neural network notation and convention as in [3], with supplementing information from [14].

5.1.1 Statistical Learning

First, we start by recalling the general **statistical learning** process. Suppose we have a dataset $\{x_i, y_i\}_{i=1}^n$, such that $x_i \in \mathbb{R}^d$, and $y_i \in \mathbb{R}^c$. Here, $\mathbb{P}$ is the underlying distribution from which the pairs (x_i, y_i) are sampled. The goal of statistical learning is to infer a function $f : \mathbb{R}^d \to \mathbb{R}^c$ that captures the relationship between the inputs x_i and the outputs y_i. This function f is often referred to as a **model**, and the process of learning f is called **training**.

In statistical learning, we assume that the data is generated by an unknown but fixed distribution $\mathbb{P}(x, y)$. The objective is to find a model f that minimizes the **expected loss**,

$$R_f = \mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\mathcal{L}(f(x), y)\right],$$

where $\mathcal{L}$ is a **loss function** that quantifies the discrepancy between the predicted output $f(x)$ and the true output y.

However, since the true distribution $\mathbb{P}$ is unknown, we cannot directly compute

$R(f)$. Instead, we approximate it using the **empirical loss**,

$$\hat{R}_f = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x_i), y_i),$$

where we evaluate the loss function at a randomly selected batch of n sample data points. Minimizing the empirical loss is the foundation of many machine learning algorithms.

5.1.2 Machine Learning

Machine learning extends the statistical learning framework by introducing computational methods to learn f from data. Unlike traditional statistical methods, which often rely on explicit assumptions about the form of f, machine learning algorithms are designed to automatically discover patterns and relationships in the data. This flexibility makes machine learning particularly powerful for high-dimensional and complex datasets.

A key distinction in machine learning is between supervised learning and unsupervised learning. In **supervised learning**, the dataset $\{x_i, y_i\}_{i=1}^{n}$ includes both inputs x_i and corresponding outputs y_i. The goal is to learn a mapping f that predicts y from x. Whereas, in **unsupervised learning**, only the inputs $\{x_i\}_{i=1}^{n}$ are available. Now the goal is to discover hidden structures or patterns in the data, such as clusters or low-dimensional representations.

The choice of model f is central to the success of a machine learning algorithm. Simple models, such as linear regression or logistic regression, are interpretable and computationally efficient, but may lack the ability to capture complex relationships in the data. On the other hand, more expressive models, such as decision trees, support vector machines, and neural networks, can model non-linear relationships effectively, but may require more computational resources to train well. Furthermore, there is a trade-off when choosing what model to utilise in analysing a dataset.

5.1.3 Neural Networks

Among the most powerful and flexible models in machine learning are **neural networks**. A neural network is a composition of simple, non-linear functions (called neurons) that can approximate any continuous function, given sufficient capacity. This property, known as the **Universal Approximation Theorem**, makes neural networks particularly well-suited for tasks where the underlying relationship between inputs and outputs is complex and unknown.

More formally, neural networks can model non-linear functions $\mathbb{R}^d \to \mathbb{R}^c$ parameterised by weights $w \in \mathbb{R}^N$. The function is formed by the composition of affine linear maps and non-linear functions in a layered architecture. Heuristically, for a larger N, we can model more complicated patterns effectively. Neural networks allow easy scaling of the model to be arbitrarily large. This leads to neural networks typically being thought to be over-parameterised due to large N. However, they are surprisingly still robust to **overfitting**, making them, in general, very effective machine learning models. Neural networks are defined by both their weights, and their architecture. In this report, we will mainly consider **multi-layer perceptrons**.

Multi-layer perceptrons have the simplest architecture of all the types of neural networks. We begin by assuming the neural network has positive integer L layers of neurons, with $d = n_0, n_1, ..., n_L = c$ neurons in each layer. We also require for each layer a weights matrix $W^{(i)} \in \mathbb{R}^{n_{i-1} \times n_i}$, a bias vector $b^{(i)} \in \mathbb{R}^{n_i}$, and non-linear activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Common activation functions include:

- RELU activation function: $\sigma(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{else,} \end{cases}$
- Sigmoid activation function: $\sigma(x) = \frac{1}{1+e^{-x}}$.

We then define the output from each layer as:

$$z^{(0)} = x; \qquad z^{(l)} = W^{(l)} \sigma(z^{(l-1)}) + b^{(l)} \text{ for } l = 1, ..., L; \qquad f_w(x) = z^{(L)}.$$

The process of training the neural network to reduce the cost function, is non-trivial, but surplus to our discussion here. All that is key is that the training process is iterative in its nature, repetitively adjusting the weight matrix until we have satisfied some termination criteria. Then, the final concept we require for our discussion is the **loss surface** of a neural network, defined by,

$$\mathbb{E}[\mathcal{L}(y, f(x))]$$

where $f(x)$ is the output from our trained neural network. Hence, it is essentially the expected loss as defined previously.

Example 5 In order to clearly illustrate how a multi-layer perceptron neural network works, we will see a quick example. Let $L = 3$ so that we have three layers in our network, including one hidden layer. Then also assume that the first layer has 2 neurons, the middle hidden layer has 3 neurons, and the final layer has 1 output neuron. Assuming that all neurons are fully connected, with a bias and with activation function σ, then the neural network architecture is shown in Figure 5.1.

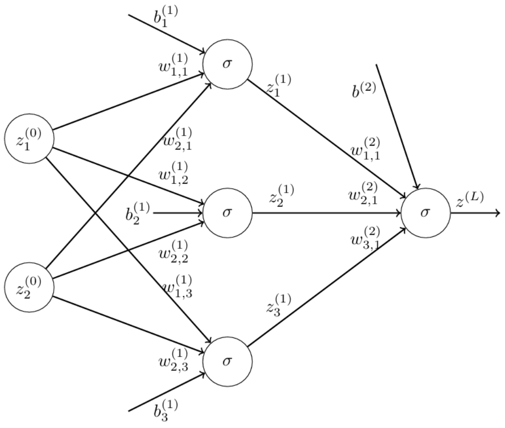


Figure 5.1: Multi-Layer Perceptron Neural Network Architecture

5.2 Appearance of Random Matrix Phenomena

In this final subchapter, we will discuss some of the results shown in [3], specifically in Chapter 7, where the appearance of random matrix phenomena in the **empirical Hessian** of loss surface of neural networks is explored. The empirical Hessian is defined as,

$$H_{emp}(w) = \nabla^2 \hat{R}_f(w),$$

recalling that $\hat{R}_f$ is the empirical loss function. The empirical Hessian thus describes the loss curvature at a point w in weight space. The argument w represents the weights and biases that the trained neural network has in its architecture after training.

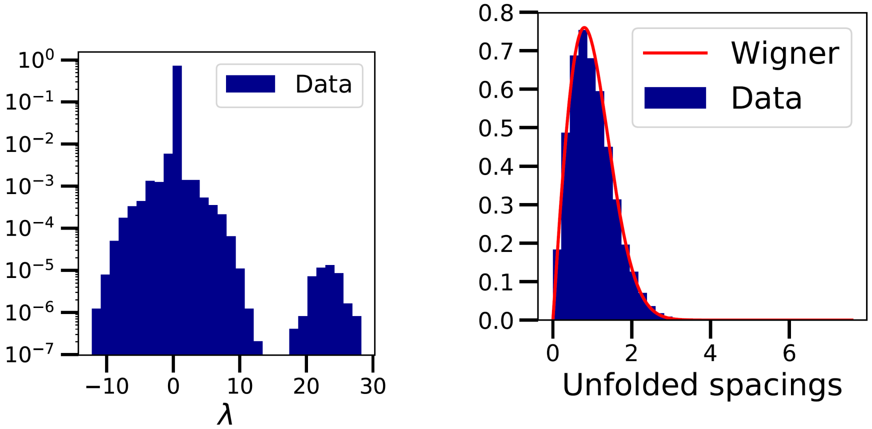
The eigenvalues of the Hessian are particularly of interest since they determine the characterisation of stationary points, and guide second order optimisation methods. The appearance of local minima in the empirical loss is especially of concern, as these are regions where the training process of a neural network model can falter. Getting stuck in shallow local minima is a practical challenge in the usage of neural networks, as can prevent the user from achieving a well-optimised model, achieving sub-par loss. Hence, understanding the Hessian may lead to insights into optimising neural network training, and designing more robust algorithms. Also, note that the loss Hessians are necessarily symmetric, which again means the eigenvalues are real and so can be ordered. We are interested in whether the eigenvalue distributions of these matrices reflect the random matrix statistics we have previously discussed.

Through Chapter 7 in [3], Baskerville analyses the loss Hessians of many different neural network architectures. We will discuss the results from small MLP networks, since we have described how these are built previously. However, the text also analyses convolutional neural networks, which we haven't defined here. Note that only relatively small networks are considered due to the rapidly increasing computation costs of eigenvalue decompositions. We won't explicitly describe the training process of the networks used since it don't add value to our discussion of universality results, however full details are given in the original text.

The first result we will discuss is depicted in Figure 5.2a which is copied from [3] (page 212). The plot shows the empirical eigenvalue distribution of the loss Hessian calculated over the test set of a 3-hidden layer MLP which was trained on the renowned dataset MNIST. This large dataset contains 70,000 images of handwritten digits 0-9 with their corresponding label [8]. Again, the details of how the Hessian calculation was performed is surplus to the discussion here, but full details can be found in [3]. The matrix had not been scaled appropriately, yet it is still visually clear that there is not the semicircular shape as in Wigner's Semicircle Theorem, Theorem 1. There is a large peak at the origin, seemingly with outliers at the right edge. Hence, it is clear that the global phenomena random matrix statistics have not appeared here. Note that this might be expected, since the Hessian matrix does not necessarily contain elements that could have come from appropriate distributions.

However, the next result is far more surprising and interesting for us, and is shown in Figure 5.2b, which is copied from [3] (page 208). This plot shows the unfolded eigenvalue spacing distribution of the loss Hessian calculated over the test set of a 3-hidden layer MLP, which was this time trained on the bike dataset, along with the Wigner surmise. The bike dataset contains the hourly and daily count of rental bikes used from a certain bike-share company, as well as other predictor variables such as the day of the week and the weather [7]. It is clear in Figure 5.2b that the unfolded eigenvalue spacing distribution is very similar to the Wigner surmise. Therefore this gives numerical evidence in favour of our discussions of universality in local phenomena. Again, we have no reason to expect that the loss Hessian has elements that could have come from Normal distributions, yet we still see the Wigner surmise arising.

Recall that this can be explained by the unfolding of the eigenvalue spacings, so they are normalised to have mean one. This averaging removes the effect of the global eigenvalue density and the effect of the average spectral density on pair correlations between eigenvalues. This leaves just the effects of the general symmetry class of the matrix. As stated before, the Hessians are symmetric and so should vaguely resemble a GOE matrix, however, the elements are not necessarily representative of being sampled from Gaussian. Yet we still get the results that reflect this. The work done in [3] is the first analysis on evaluating the eigenvalue spacing distribution of an artificial neural network, and hence it shows promising numerical results in terms of the appearance of local random matrix statistics in such neural networks.



(a) Empirical Eigenvalue Distribution of the Loss Hessian Over the Test Set of a 3-Hidden Layer MLP

(b) Unfolded Eigenvalue Spacings for the Loss Hessian Over the Test Set of a 3-Hidden Layer MLP with the Wigner Surmise

Figure 5.2: Analysis of the Eigenvalues of Loss Hessian for 2 Different 3-Hidden Layer MLPs

Chapter 6

Conclusion

This dissertation report has acted as an introduction to the field of RMT, focussing on both global and local phenomena, describing the difference between the two. In detail, we have proved one of the most renowned results in RMT: Wigner's Semicircle Theorem, Theorem 1. This also acted as our main example of global phenomena, since it describes the full eigenvalue spectrum of certain random matrices. Through the long proof for the case of Gaussian Wigner matrices, we covered ideas from measure theory and probability, as well as bounding statistics.

Then we began the comparison between global and local phenomena and what they mean. We discussed that local phenomena describe both the behaviour of certain eigenvalues, as well as the distances between neighbouring eigenvalues. We focussed on the latter by exploring the Wigner surmise, Theorem 8. This also probed our discussion into universality through the unfolding process. We saw numerical evidence that the Wigner surmise roughly holds for many symmetric matrices due to unfolding which removes the global effects, leaving behind only the effects due to the overarching class of matrices. In our case, this was the orthogonality in the GOE.

Then finally, we explored whether some of these random matrix statistics appear within neural networks. This was greatly linked to our universality discussion, since it was only the local phenomena we saw in the Hessian of loss functions from a MLP. The full eigenvalue spectrum didn't reflect the theoretical results we had seen, but again, the unfolding process meant that we saw the Wigner surmise appearing. This was surprising since there is no reason for the loss Hessian to have Gaussian-distributed values. This gave an idea into how applicable and far-reaching the field of RMT can be.

It is important to note that we have only scratched the surface of the field RMT. We have only considered a few different theorems. Hence, there are a number of ways that this report can be extended. One of the main aims was to prove the main theorems in full detail, and hence we laboured over the proof of Wigner's Semicircle Theorem. However, this means that we haven't discusses many other important theorems we could have. We also only proved Theorem 1 for Gaussian Wigner matrices. Hence there are endless ways that the exploration of RMT could have been extended, including but not limited to: considering random matrices beyond real Wigner matrices and the GOE, and exploring other global and local results in more detail such as Theorem 6.

There is also much more room for further exploration in the applications of RMT. We discussed some of the work in [3] which is the first work of evaluating the eigenvalue spacing distribution of loss Hessians in some neural networks. Hence, further work can be done to numerically check other forms of neural networks not discussed here, to see if local random matrix statistics appear in their loss Hessians. Recall, understanding eigenvalues of the loss surface is important in effective training of neural networks, making this work important. Furthermore, a more mathematically rigorous approach is vital in continuing this research, to go beyond just experimental success.

Bibliography

[1] G W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, 2009.

[2] K. B. Athreya and S. N. Lahiri. *Measure Theory and Probability Theory*. Springer Science+Business Media, 2006.

[3] N. P. Baskerville. "Random Matrix Theory and the Loss Surfaces of Neural Networks". PhD Thesis. Bristol: University of Bristol, 2023. URL: `https://arxiv.org/pdf/2306.02108`.

[4] S. Bauman. *The Tracy-Widom Distribution and its Application to Statistical Physics*. MIT Department of Physics, 2017.

[5] P. Billingsley. *Convergence of Probability Measures*. John Wiley Sons, 1968.

[6] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[7] H. Fanaee-T. *Bike Sharing*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5W894. 2013.

[8] Geeks for Geeks. *MNIST Dataset : Practical Applications Using Keras and PyTorch*. Accessed: 24/04/2025. 2024.

[9] P. Halmos. *Measure Theory*. D. Van Nostrand Company, Inc., 1950.

[10] J. M. Harris, J. L. Hirst, and M. J. Mossinghoff. *Combinatorics and Graph Theory*. Springer Science+Business Media, 2000.

[11] I. T. Jolliffe. *Principal Component Analysis*. Springer Science+Business Media, 2002.

[12] M. L. Mehta. *Random Matrices*. Academic Press Inc., 1967.

[13] M. L. Mehta. *Random Matrices: Third Edition*. Elsevier, 2004.

[14] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[15] R. Speicher. *Random Matrices*. Lecture Notes. Saarland University, 2020. URL: `https://arxiv.org/abs/2009.05157`.

[16] J. Wishart. "The Generalised Product Moment Dsitribution in Samples From a Normal Multivariate Population". In: *Biometrika* 20A (1-2 1928), pp. 32–52.