Random Matrix Theory

J. P. Keating and L.-P. Arguin

January 8, 2025

Contents

Preface

These notes accompany the Part C course on Random Matrix Theory, C7.7. It is a pleasure to thank Jad Hamdan, Akshay Hegde, Nathan Creighton, Michael Curran, Johannes Forkel, Bhargavi Jonnadula, Isao Sauzedde, and Mo Dick Wong for their help in preparing them. We are grateful for comments, suggestions, and for reports of typos and errors.

L.-P. Arguin and J. P. Keating
December 2024

Part 0
Introduction

1 Why Random Matrices?

Matrices play a central role in mathematics and its applications. Just as one considers the properties of probability measures defined on a space of real or complex variables, it is also natural to associate probability measures with spaces of matrices. We shall be interested in the properties of such *random matrices*, and in applications of *Random Matrix Theory* to predict what one should expect in 'typical' situations involving linear algebra. Random matrices is a rich topic in contemporary mathematics that intersect with probability, analysis, mathematical physics, combinatorics, number theory and data sciences. This course is designed as an introduction to the foundational principles of random matrix theory.

Why should one expect random matrices to be so ubiquitous in mathematics and in nature? A linear operator T from $\mathbb{R}^m \to \mathbb{R}^n$ is a function with the property that $T(a\mathbf{v}_1 + b\mathbf{v}_2) = aT\mathbf{v}_1 + bT\mathbf{v}_2$, where $a, b \in \mathbb{R}$ and $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^m$. Linear operators are arguably the simplest functions from $\mathbb{R}^m$ to $\mathbb{R}^n$. The linear operator T can be represented as a matrix A for a given choice of basis of $\mathbb{R}^m$, say $(\mathbf{e}_i, i \leq m)$ and $(\mathbf{f}_j, j \leq n)$. The entry A_{ij} of A is then the component of $T\mathbf{e}_i$ in the direction of $\mathbf{f}_j$. [1]

With this point of view, it is natural to consider random matrices. Indeed, one should think of a random matrix as a *generic linear operator*. We will often think of the dimensions m and n as large, tending to infinity in fact. It is then in practice impossible to define exactly an operator entry by entry. A better approach is to look at the problem *statistically*. One would like to sample a linear operator among a large set of operators with perhaps some *a priori properties and symmetries*. It is thus necessary to construct probability distributions on matrices.

1.1 What is a random matrix?

There are many ways to construct probability distributions on matrices. Given the standard bases on $\mathbb{R}^m$ and $\mathbb{R}^n$, the simplest way is perhaps to consider each entry $\mathbf{M}_{ij}$ with $i \leq m$ and $i \leq n$ as random variables. There might be correlations between the entries. The distribution of the random matrix is then given by the joint distribution of the entries. A distribution on random matrices is often called a *random matrix ensemble*.[2]

Throughout these notes, we will usually denote a random matrix by $\mathbf{M}$.

We will use the notation M for a given deterministic matrix, for a example for a given realization of $\mathbf{M}$.

Example 1.1. Let $\mathbf{M}_{ij}$ be independent, identically distributed (IID) uniform random variables on $[0, 1]$. We can define a random matrix $\mathbf{M}$ by taking

$$\begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}.$$

Any event pertaining to M can be computed using the joint probability density function (PDF) of the M_{ij}. More precisely, if B is a (measurable) subset of $\mathbb{R}^{2\times 2}$ then

$$\mathbf{P}(\mathbf{M} \in B) = \int_B \mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}x_3 \mathrm{d}x_4.$$

One can then compute for example the probability that $\mathbf{M}$ is singular or symmetric.

[1] Of course, the above setup generalizes when T is a linear operator on other vector spaces. In this course, we will focus on $\mathbb{R}^n$ and sometimes $\mathbb{C}^n$.

[2] The word *ensemble* is often used in physics for a probability distribution.

To sample only symmetric matrices, we can force the symmetry by taking instead

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{12} & \mathbf{M}_{22} \end{pmatrix}.$$

Note that the off-diagonal are now completely correlated! The subspace of 2×2 symmetric matrices has dimension 3. In particular, the distribution of $\mathbf{M}$ is determined by

$$\mathbf{P}(\mathbf{M} \in B) = \int_B \mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}x_3,$$

where B is now a subset of $\mathbb{R}^3$.

Another way to sample is to directly define the probability on the space of matrices.

Example 1.2. *Consider the set of rotation matrices in $\mathbb{R}^2$. For a rotation by an angle θ, we have*

$$\mathbf{R}_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

To sample a rotation uniformly, it suffices to take θ as a random variable uniform on $[0, 2\pi)$.

We will see in Section 2 many examples of random matrix ensembles defined in similar ways.

1.2 What is our goal?

As it is often the case in probability, there are fascinating phenomena that arise in the limit where the number of random variables tends to infinity [3]. Random matrix theory is no exception. Our main objective is to understand the properties of very large matrices, when the dimensions $m, n \to \infty$, sometimes at the same rate, sometimes at different rate. Of course, one cannot keep track of the whole operator then, and it is convenient to reduce our attention to certain *statistics* or *observables* in the limit. One might object that defining random matrices by describing the distribution of each entry is too basis-specific. Thus, good observables should be independent of the basis.

An obvious choice of observables is the *eigenvalues*, if the matrix is square. If the random matrix $\mathbf{M}$ is not square, one could look at the *singular values*, i.e., the square root of the eigenvalues of the matrix $\mathbf{M}^{\mathrm{T}}\mathbf{M}$. (The eigenvalues of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ must then be real and non-negative. Why?) The eigenvalues of a matrix $\mathbf{M}$ is a function of the entries, therefore if $\mathbf{M}$ is a $n \times n$ random matrix, then

$$\text{its eigenvalues } (\Lambda_1, \ldots, \Lambda_n) \text{ are random variables taking values in } \mathbb{C} \text{ or } \mathbb{R}$$

When n is large, this collection is again large. In this course, we will study different statistics of this set:

- **Part I: Empirical Spectral Measure (ESM)**. We can define a discrete probability measure by associating to each Λ_i a point mass (also known as Dirac Delta) each with a weight of $1/n$:

$$\mu_{\mathbf{M}} = \frac{1}{n} \sum_{i=1}^{n} \delta_{\Lambda_i}.$$

  This is called the *empirical spectral measure* (ESM). It gives a macroscopic measure of how the eigenvalues are distributed. We will see that μ_M converges to some universal distributions depending on the matrix type: *the semicircle law*, *the Marchenko-Pastur law*, and *the Girko law*.

[3] This is also the case in statistical physics when we take the limit of particles going to infinity, the so-called *thermodynamic limit*.

- **Part II: Eigenvalue Statistics** In the second part, we will look at the unnormalized spectral measure $\sum_{i=1}^{n} \delta_{\Lambda_i}$. This can be seen as a *point process* that is a random configuration of points in $\mathbb{C}$ and $\mathbb{R}$. One can then ask questions about the statistics of these points such as the gap between consecutive eigenvalues or the distance between pairs. As for the ESM, we will see in examples that one recover similar statistics in the limit $n \to \infty$ for seemingly different ensembles. This gives us a glimpse of the *universality of random matrix statistics*.
- **Part III: Dyson Brownian Motion and Universality** We will end the journey by defining the Dyson Brownian motion. This is a dynamic on the eigenvalues that turns out to be extremely useful to prove universality results. We will also explore the appearances of the random matrix statistics in other fields of mathematics, most of which is still misunderstood.

2 Examples of Random Matrix Ensembles

We start by defining explicitly the random matrix ensembles that will be studied in this course. There are many important ensembles that we will not have the time to investigate. Those defined here are a representative set and are amongst the most widely studied. Here an ensemble is a space of matrices endowed with a probability measure.

2.1 The Ginibre Ensemble

A $n \times n$ real *Ginibre random matrix* $\mathbf{M}$ is arguably the simplest random matrix. It is simply defined by taking

$$\mathbf{M}_{ij} \text{ IID standard Normal random variables, } 1 \leq i, j \leq n.$$

The eigenvalues of a Ginibre matrix are not necessarily real.

2.2 Wigner Matrices

A *Wigner matrix* is a random matrix where some symmetries are imposed. More precisely, a *real Wigner matrix* $\mathbf{M}$ is such that above the diagonal

$$\mathbf{M}_{ij}, 1 \leq i < j \leq n, \quad \text{are IID real random variables with } \mathbf{E}[\mathbf{M}_{ij}] = 0 \text{ and } \mathbf{E}[\mathbf{M}_{ij}^2] = 1. \tag{2.1}$$

We then impose that $\mathbf{M}$ is *symmetric* so that $\mathbf{M}_{ji} = \mathbf{M}_{ij}$. The distribution of the diagonal may be different from the off-diagonal entries. We assume $\mathbf{M}_{ii}$ is IID with $\mathbf{E}[\mathbf{M}_{ii}] = 0$ and $\mathbf{E}[\mathbf{M}_{ii}^2] < \infty$ (uniformly in n). When taking the limit $n \to \infty$, we will sometimes rescale the matrix elements by a power of n.

A *complex Wigner matrix* is defined similarly as the real case by imposing the Hermitian property. Namely, we take

$$\mathbf{M}_{ij}, 1 \leq i < j \leq n, \text{are IID complex random variables with } \mathbf{E}[\mathbf{M}_{ij}] = 0 \text{ and } \mathbf{E}[|\mathbf{M}_{ij}|^2] = 1, \tag{2.2}$$

with $\mathbf{M}_{ji} = \overline{\mathbf{M}_{ij}}$ is the complex conjugate of $\mathbf{M}_{ij}$. Again, the diagonal may have different distributions but the entries remain IID with mean 0 and finite second moment.

2.3 The Gaussian Orthogonal Ensembles (GOE)

The *Gaussian Orthogonal Ensemble* (GOE) is the particular class of real Wigner random matrices for which the matrix elements are independent normal random variables. Specifically, $\mathbf{M}_{ij}$ with $1 \leq i < j \leq n$ are IID random variables with distribution $\mathcal{N}(0, 1)$, and $\mathbf{M}_{ii}$, $i \leq n$, are IID real random variables with distribution $\mathcal{N}(0, 2)$.

Hence, the joint probability density function (PDF) of the entries of $\mathbf{M}$ is

$$\prod_{i=1}^{n} \frac{1}{\sqrt{4\pi}} e^{-\frac{1}{4}M_{ii}^2} \prod_{1\le i<j\le n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}M_{ij}^2} = \frac{1}{2^{n/2}(2\pi)^{n(n+1)/4}} e^{-\frac{1}{4}\mathrm{Tr}(M^2)}. \tag{2.3}$$

The following is one of the most important properties of the GOE (and the reason for its name).

Lemma 2.1. *Let O be a non-random $n \times n$ orthogonal matrix (i.e. $OO^{\mathrm{T}} = \mathrm{I}$) and let $\mathbf{M}$ be an $n \times n$ GOE matrix. Then the distribution of $\mathbf{M}$ is the same as that of $O\mathbf{M}O^{\mathrm{T}}$; that is, the GOE is invariant under conjugation by all orthogonal matrices.*

Proof. We need to show that the PDF of $O\mathbf{M}O^{\mathrm{T}}$ is the same as the one of $\mathbf{M}$. Note first that

$$\mathrm{Tr}\Big((OMO^{\mathrm{T}})^2\Big) = \mathrm{Tr}\Big(OM^2O^{\mathrm{T}}\Big) = \mathrm{Tr}(M^2O^{\mathrm{T}}O) = \mathrm{Tr}(M^2). \tag{2.4}$$

It remains to show that the Jacobian of the transformation is 1. One can compute the Jacobian by hand and confirm this by explicit calculation. To see this directly, note that the Hilbert-Schmidt norm of M is precisely $\mathrm{Tr}(M^2)$. But we have just seen that this is invariant under conjugation by O [4]. The map $M \mapsto OMO^{\mathrm{T}}$ is therefore an isometry, so its Jacobian determinant is 1. □

It turns out that the only ensemble of symmetric random matrices whose entries are independent (up to the symmetry constraint) and whose distribution is invariant under conjugation by all orthogonal matrices is the GOE – so the orthogonal invariance is a special property of the Gaussian nature of the matrices. One can easily construct matrix ensembles that are invariant under conjugation by all orthogonal matrices but in which the matrix elements are not statistically independent.

2.4 The Gaussian Unitary Ensembles (GUE)

The *Gaussian Unitary Ensemble* (GUE) is the particular class of complex Wigner random matrices for which the matrix elements are independent complex normal random variables. We take $\mathbf{M}_{ij}$, $1 \le i < j \le n$ IID random variables with real and imaginary parts that are independently Gaussian $\mathcal{N}(0, 1/2)$, and $\mathbf{M}_{ii}$, $1 \le i \le n$, IID real standard Gaussian random variables. Note that the matrix $\mathbf{M}$ is Hermitian (i.e., $\mathbf{M} = \mathbf{M}^\dagger = \overline{\mathbf{M}}^{\mathrm{T}}$) by construction.

In this case the joint PDF of the entries is

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}M_{ii}^2} \prod_{1\le i<j\le n} \frac{1}{\sqrt{\pi}} e^{-(\mathrm{Re}M_{ij})^2} \frac{1}{\sqrt{\pi}} e^{-(\mathrm{Im}M_{ij})^2} = \frac{1}{2^{n/2}\pi^{n^2/2}} e^{-\frac{1}{2}\mathrm{Tr}(M^2)}. \tag{2.5}$$

Not surprisingly, we have:

Lemma 2.2. *Let U be a $n \times n$ unitary matrix (i.e. $UU^\dagger = I$) and $\mathbf{M}$ be a GUE matrix. Then $U\mathbf{M}U^\dagger$ is also a GUE matrix.*

Proof. This is done again by computing the PDF of $UMU^\dagger$. We have

$$\mathrm{Tr}\Big((UMU^\dagger)^2\Big) = \mathrm{Tr}(UM^2U^\dagger) = \mathrm{Tr}(M^2U^\dagger U) = \mathrm{Tr}(M^2). \tag{2.6}$$

Again, the Jacobian of the transformation $M \mapsto UMU^\dagger$ is 1 since $\mathrm{Tr}(M^2)$ is the Hilbert-Schmidt norm of M, and so conjugation by a unitary matrix is an isometry. □

[4] The Hilbert-Schmidt norm of a matrix is defined generally by $\mathrm{Tr}(M^{\mathrm{T}}M)$.

The only ensemble of Hermitian random matrices whose entries are independent (up to the symmetry constraint) and whose distribution is invariant under unitary conjugation is the GUE. One can easily construct matrix ensembles that are invariant under unitary conjugation but in which the matrix elements are not statistically independent.

The PDF of GOE and GUE matrices may, up to a normalization constant, be written in the form

$$e^{-\frac{\beta}{4}\mathrm{Tr}(M^2)} \tag{2.7}$$

where for the GOE $\beta = 1$ and for the GUE $\beta = 2$. There is a third Gaussian ensemble with good invariance properties for which the matrix elements are quaternions. It is called the Gaussian Symplectic Ensemble (GSE), and its distribution is invariant under conjugation by a symplectic matrix. It corresponds to $\beta = 4$. What about other β's? We will see in Part II that the distribution makes sense for any $\beta > 0$, however there is no obvious invariance under conjugation in this case.

2.5 The Circular Unitary Ensembles (CUE)

The *circular unitary ensemble* (CUE) is the space of $n \times n$ unitary matrices endowed with a probability measure that is invariant under all unitary transformations (i.e., under left and right multiplication by all unitary matrices). In this case, the matrices represent elements of the compact Lie group $U(n)$ and the invariant measure is called the *Haar measure*. This measure does not have a simple expression in terms of the matrix elements, as in the case of Wigner matrices.

Similarly, one can consider the space of $n \times n$ orthogonal matrices endowed with a probability measure that is invariant under all orthogonal transformations (i.e., under left and right multiplication by all orthogonal matrices). In this case the matrices represent elements of the compact Lie group $O(n)$ and the invariant measure is the *Haar measure* on this group. Again, this measure does not have a simple expression in terms of the matrix elements.

For a careful introduction to Haar measure on compact Lie groups, see, for example, [Mec19] and [Mez07]. The latter reference discusses how to generate random unitary and orthogonal matrices numerically.

2.6 Wishart Matrices

Let $\mathbf{X}$ be a p-dimensional Gaussian vector of mean 0 and covariance $\mathcal{C}$ (see Example B.3). Consider the $p \times n$ matrix $\mathbf{Y}$ constructed by taking n independent copies $\mathbf{X}^{(i)}$ of $\mathbf{X}$ as its columns. One can think of each the columns as independent measurements of p quantities with possible correlations between them. It is natural then to consider $p \leq n$. A *real Wishart matrix* with covariance matrix $\mathcal{C}$ is a p-dimensional random symmetric positive definite matrix M of the form

$$\mathbf{M} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}} \tag{2.8}$$

This is an example of *sample covariance matrix* since if we divide by n,

$$\frac{1}{n}\mathbf{M}_{ij} = \frac{1}{n}\sum_{k=1}^{n} X_i^{(k)} X_j^{(k)}.$$

The PDF of the entries of $\mathbf{M}$ was computed by Wishart in 1928 [Wis28] to be

$$f(M) = \frac{1}{2^{np/2}\Gamma_p(\frac{n}{2})(\det \mathcal{C})^{n/2}} (\det M)^{(n-p-1)/2} \exp\left[-\frac{1}{2}\mathrm{Tr}(\mathcal{C}^{-1}M)\right] \tag{2.9}$$

Here

$$\Gamma_p(\frac{n}{2}) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(\frac{n}{2} - \frac{j-1}{2}\right). \tag{2.10}$$

3 Occurences of Random Matrices in Mathematics and Beyond

Historically, questions in Random Matrix Theory (RMT) have arisen in a number of different mathematical contexts. We review a selection of these, by way of motivation and a gentle introduction to some of the key themes that will be explored later in more detail.

3.1 Random rotations

Arguably, the first substantial RMT calculation appeared in work of Adolf Hurwitz concerning orthogonal transformations. Rotations of objects in three dimensions can be parametrized by the Euler angles. One can therefore consider random rotations in three dimensions by defining a probability measure on these angles. In the absence of any preferred directions, it is natural to use a probability measure that is itself invariant under all rotations.

Another way of phrasing this is in terms of orthogonal matrices. Consider $r \in \mathbb{R}^3$. Rotating r corresponds to multiplying it by a 3×3 orthogonal matrix[5], i.e. a matrix O satisfying $OO^{\mathrm{T}} = \mathrm{I}$. Random rotations can therefore be thought of in terms of a probability measure on the space of orthogonal matrices. A rotationally invariant probability measure corresponds to a probability measure that is itself invariant under all orthogonal transformations.

Hurwitz was the first to investigate invariant probability measures on the orthogonal and unitary groups. See [DF17] for an overview of his work and subsequent developments, and [Mec19] for an introduction to the general theory in wider contexts.

3.2 Systems of linear equations

One of the basic problems of numerical linear algebra involves solving the system of linear equations

$$Ax = b \tag{3.1}$$

for the n-dimensional vector x, given an $m \times n$ matrix A and an m-dimensional vector b, when m and n are large. One important consideration is: if b is not specified precisely, how does this imprecision affect the accuracy with which x can be determined? The *condition number* is the maximum ratio of the relative error in x to the relative error in b. A natural way to measure the size of these errors equates the condition number to the ratio of the largest and smallest singular values of A, or in the case when A is *normal*[6], the ratio of the maximal and minimal absolute values of the eigenvalues of A.

In 1947 John von Neumann and Herman Goldstine [vNG47] asked what the condition number would be for a 'typical' large matrix and initiated the study of condition numbers of random matrices, taking the elements of A to be IID random variables. This subject has a long and interesting history, which is beautifully reviewed in [ER05]. In particular, the distribution of values taken by the condition number is related to the distribution of the largest and smallest eigenvalues of the associated random matrices.

3.3 Complex quantum systems

In quantum mechanics, the allowed values of the energy in a closed system – the *energy levels* – are, in general, the eigenvalues of a complex Hermitian matrix $H = H^\dagger$, the *quantum Hamiltonian*. In many settings this matrix is in fact real and symmetric, so $H = H^{\mathrm{T}}$. It was suggested by Eugene Wigner in the 1950s that in complex quantum systems the independent entries in the matrix H (H_{ij} with $i \geq j$, say) should be modelled as random variables. This gave rise to the Wigner matrices defined in the last section, The question then is: how are the eigenvalues of random complex Hermitian or random real symmetric matrices distributed? It turns out that for matrices whose entries are independent random variables, the eigenvalues are strongly correlated in a distinctive and mathematically interesting

[5] More generally, orthogonal matrices generate linear isometries; the orthogonality of O preserves the dot product between any two vectors

[6] i.e. if $AA^\dagger = A^\dagger A$.

way. These characteristic correlations are indeed seen in the energy level statistics of typical quantum systems, ranging from atomic nuclei to superconducting systems. The fact that large perfectly ordered systems conduct electricity while disordered systems do not is because in the latter case the quantum Hamiltonians behave like random matrices.

In fact this is not an exclusively quantum phenomenon: it is observed in all wave theories, including acoustics, optics, elasticity, electromagnetism, etc. Moreover, it is observed on the widest range of scales, ranging from the description of sub-nuclear physics in terms of quantum chromodynamics to the structure of the cosmic microwave background.

3.4 Stability of high-dimensional dynamical systems

Let $x(t)$ be an n-dimensional vector satisfying

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\lambda \mathrm{I} x \tag{3.2}$$

where λ is a positive constant and I is the $n \times n$ identity matrix. This system has a stable fixed point at $x = 0$. Clearly in (3.2) the different components of x are uncoupled (because I is diagonal). The question is: if one now introduces random coupling between these components, does this make the fixed point more or less stable? Specifically, what is the stability of the system

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\lambda \mathrm{I} x + Ax, \tag{3.3}$$

where A is a random matrix?

This question was first raised by Robert May (Lord May of Oxford) in 1972 in a famous paper [May72] on population dynamics. In that context, it is called the *May model*. Then the components of x represent the differences of the populations of various species from some equilibrium values. If the species do not interact, then it is assumed that the equilibrium values are stable. It had long been assumed that adding random interactions would make the equilibrium populations more stable when n is taken to be large. May asked whether this was indeed the case and analysed the situation using the simple model system (3.3).

Obviously the question of stability is related to knowing the expected size of the eigenvalues of A when n is large. This question arises in many other contexts as well, for example in the dynamics of neural models and on complex networks.

3.5 Principal Component Analysis

Let $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ be $p \times 1$ vectors, where p is to be treated as large with respect to the number of vectors, n. Let Y be another $p \times 1$ vector. Which of $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ is closest to Y? Obviously one can check each of the distances $\|Y - X^{(i)}\|_2$, but doing so is costly. This is an important problem in high-dimensional data analysis.

The idea of Principal Component Analysis is to project onto a subspace in which the vectors $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ show maximum variability. It is obviously natural to seek to do this in the subspace where the vectors exhibit maximum variability, because this is where their differences are largest. This can be thought of as enacting 'feature selection' in many applications. The reduction in the dimension of the space where the comparison is made increases the efficiency of the search considerably. For example, for a given unit vector $u \in \mathbb{R}^p$, which of the $u^{\mathrm{T}}X_1, u^{\mathrm{T}}X_2, \ldots, u^{\mathrm{T}}X_n$ have maximum variability. For simplicity, let us assume that $X_1 + X_2 + \cdots + X_n = 0$. Then the variance of the set $u^{\mathrm{T}}X_1, u^{\mathrm{T}}X_2, \ldots, u^{\mathrm{T}}X_n$ is

$$\frac{1}{n}\sum_{j=1}^{n}(u^{\mathrm{T}}X^{(j)})^2 = u^{\mathrm{T}}\left(\frac{1}{n}\sum_{j=1}^{n}X^{(j)}{X^{(j)}}^{\mathrm{T}}\right)u = u^{\mathrm{T}}\sigma_n u. \tag{3.4}$$

The maximum value $u^{\mathrm{T}}\sigma_n u$ can take is the largest eigenvalue of σ_n, and this occurs when u is the corresponding eigenvector.

Clearly this can be extended so that one can project onto the subspace spanned by the eigenvectors corresponding to the m largest eigenvalues. How large should m be taken? The idea is to compare the eigenvalues to those of the random matrix defined by (2.8), taken to be a null model, and to keep those that differ significantly and so represent non-random features of the data. One therefore needs to know how the eigenvalues of a random Wishart matrix are distributed.

For an overview of this and similar applications (e.g. to mathematical finance, image analysis, etc) see [Joh07] and [BS10].

3.6 Complex networks

Many problems in data science and mathematics are related to properties of networks. These can be modelled by graphs, that is by ordered pairs $G = (V, E)$ comprising a set of vertices V and edges E. The structure of the graph may be represented in terms of the *adjacency matrix* A. For a simple graph this is a square matrix of dimension $|V|$ such that A_{ij} is equal to 1 when there is an edge from vertex i to vertex j, and zero otherwise[7].

The statistical properties of complex graphs are then modelled by random square matrices where the elements are taken from $\{0, 1\}$, each with probability p, in some cases with constraints (e.g. in the case of regular graphs on the total number of ones in each row and column).

In the physics literature, these network models, where often the non-zero entries of the adjacency matrix are augmented with a 'hopping probability', are usually called *tight binding models*. One well known example is the Anderson tight binding model used to explore quantum localization and its implications for electrical conductivity.

3.7 Machine learning

In machine learning one frequently wishes to minimise functions, which are often highly complex, in an extremely high-dimensional space. One technique is to use stochastic gradient descent. This raises the important question as to how easy it is to explore effectively random surfaces in high dimensional problems; are there many local minima and saddles where one can get stuck? And at a saddle, how many downward directions are there typically? This is the subject of intense study. One way of modelling the problem is to take the Hessian at a point on the surface to be a random symmetric matrix. Understanding the structure of the critical points then reduces to understanding the distribution of the eigenvalues of this random matrix, for example how many eigenvalues are expected to be positive, and how many negative. See, for example, [CHM$^+$15].

3.8 Connections with other areas of mathematics

In the above examples it is hopefully clear that matrices play an important role, and that it is reasonable that one might wish to use random matrices as a statistical model. What is more surprising is that the mathematical structures one finds in random matrix theory are more general than one might expect, in that they arise in contexts with no obvious connection with linear algebra. We will go back to this in Part III.

One example concerns the length of the longest increasing subsequence in random permutations. Let $\mathcal{S}_n$ be the group of permutations of $1, 2, \ldots, n$. If $\pi \in \mathcal{S}_n$, $\pi(i_1), \ldots, \pi(i_k)$ is an increasing subsequence in π if $i_1 < i_2 < \cdots < i_k$ and $\pi(i_1) < \pi(i_2) < \cdots < \pi(i_k)$. Let $l_n(\pi)$ be the length of the longest increasing subsequence. For example, if $n = 5$ and π is the permutation 5 1 3 2 4, then the longest increasing subsequences are 1 2 4 and 1 3 4, and $l_n(\pi) = 3$. Equip $\mathcal{S}_n$ with uniform distribution,

$$\mathbf{P}(l_n \leq m) = \frac{\#\{\pi \in \mathcal{S}_n : l_n \leq m\}}{n!}. \tag{3.5}$$

[7] This matrix is related to the discrete Laplacian acting on the vertex set.

What is the asymptotics of this distribution as $n \to \infty$? Remarkably, it is the same as the asymptotics of the distribution of the largest eigenvalue of a random complex Hermitian $n \times n$ matrix, despite the fact that there is no matrix in the problem.

A second example concerns the Riemann zeta-function

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}. \tag{3.6}$$

The sum converges in Re $s > 1$, and the zeta function has an analytic continuation to all s (except for $s = 1$, where it has a pole). The *Riemann Hypothesis* asserts that all of the zeros of $\zeta(s)$ with non-zero imaginary parts lie on the line Re $s = 1/2$. How are the zeros distributed along the line? Like random points, or are their positions correlated? It turns out that they are correlated, and all of the evidence we have suggests that the correlations between them are the same as those between the eigenvalues of large complex Hermitian random matrices, despite the fact that again there is no obvious matrix in the problem.

Moreover, the same correlations have been found in the distances between parked cars in London, and the gaps between bus arrivals in the Mexican city of Cuernavaca [BBDS06], and subway arrivals in the New York City subway system [JT17].

Finally, many of the calculations in random matrix theory have deep connections with ideas and results in other areas of mathematics, including combinatorics and topology.

3.9 Further reading and remarks

In addition to these notes, students may wish to consult the following for further details and extensions of the material covered.

- M. Potters & J.-P. Bouchaud, A First Course in Random Matrix Theory for Physicists, Engineers and Data Scientists (Cambridge University Press)
- G. Livan, M. Novaes & P. Vivo, Introduction to Random Matrices (Springer Briefs in Mathematical Physics)
- E.S. Meckes, The Random Matrix Theory of the Classical Compact Groups (Cambridge University Press)
- G.W. Anderson, A. Guionnet & O. Zeitouni, An Introduction to Random Matrices (Cambridge Studies in Advanced Mathematics)
- M.L. Mehta, Random Matrices (Elsevier, Pure and Applied Mathematics Series)
- G. Akemann, J. Baik & P. Di Francesco, The Oxford Handbook of Random Matrix Theory (Oxford University Press)
- Z. Bai & J.W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices (Springer).
- T. Tao, Topics in Random Matrix Theory, available online from https://terrytao.files.wordpress.com/2011/02/matrix-book.pdf

It is worth remarking that random matrix theory is mathematically an extremely broad subject. Papers and books on the subject vary considerably in style and intended readership, from analysis and probability theory to engineering, theoretical physics, and data science. They therefore assume different background knowledge and differ in levels of abstraction and notions of rigour. These notes will attempt to chart a middle course, focusing on the main ideas and using a mixture of techniques, hopefully to give a sense of the various viewpoints that have influenced the development of the subject.

Part I
Empirical Spectral Measure

For a fixed $n \times n$ matrix M with eigenvalues $(\lambda_1, \ldots, \lambda_n)$, the empirical spectral measure is defined by

$$\mu_M = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i}.$$

In particular, for any subset $A \subset \mathbb{R}$, $\mu_M(A) = \int_A \mathrm{d}\mu_M$ is the proportion of eigenvalues falling in A. Note that μ_M is a probability measure on $\mathbb{R}$ if the eigenvalues are real or $\mathbb{C}$ if they are complex.

If $\mathbf{M}$ is a random matrix, then $\mu_{\mathbf{M}}(A)$ is a random variable, and $\mu_{\mathbf{M}}$ is a *random* probability measure. The distribution of a random probability measure $\mu_{\mathbf{M}}$ is determined by the collection of random variables $\int_{\mathbb{R}} f(x)\mathrm{d}\mu_{\mathbf{M}}$, for f in a suitable set of test functions. In this part, we study the ESM of different random matrix ensembles. In Section 4, we prove convergence of the ESM of Wigner matrices to the semicircle law. The proof is based on the method moments and is a beautiful piece of mathematics involving important concepts of combinatorics. For GUE matrices, the computation involved in the proof has some nice connections to topology. This is explained in Section 5. In Section 6, similar techniques are used to prove the limit of the ESM in the case of Wishart Matrices. This is the celebrated Marchenko-Pastur distribution. A different approach, and perhaps more robust technique of proof for convergence of the ESM, involves the Stieltjes transform of a probability measure. This important tool of analysis is studied in Section 7. Finally, in Section 8, we briefly discussed the ESM in the non-Hermitian case (for example, the Ginibre ensemble) where the limit of the ESM is Girko's circular law.

4 The Semicircle Law for Wigner Random Matrices

Our goal in this section is to study the ESM of Wigner matrices as defined in Section 2.2 with moment assumptions given in (2.1) and (2.2). Such matrices are Hermitian or symmetric, and therefore their eigenvalues are real. The ESM is thus a random probability measure on $\mathbb{R}$.

The following result due to Wigner is a formidable limit theorem of probability. Note that as it is often the case in limit theorems, one needs to rescale quantities to get a meaningful limit. Here, the eigenvalues needs to be rescaled by $1/\sqrt{n}$. They correspond to the eigenvalues of $\mathbf{M}/\sqrt{n}$.

Theorem 4.1 (Wigner's Semicircle Law)**.** *Let $\mathbf{M}$ be a real of complex $n \times n$ Wigner matrix as defined in Section 2.2. Denote by $\Lambda_1, \ldots, \Lambda_n$ its eigenvalues. For every bounded continuous functions $f : \mathbb{R} \to \mathbb{R}$, we have*

$$\lim_{n\to\infty} \int f \mathrm{d}\mu_{\mathbf{M}/\sqrt{n}} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(\Lambda_i/\sqrt{n}) = \int f(x)\mathrm{d}\sigma(x), \tag{4.1}$$

where $\mathrm{d}\sigma(x)$ is the semicircle distribution given by

$$\mathrm{d}\sigma(x) = \begin{cases} \frac{1}{2\pi}\sqrt{4-x^2}\mathrm{d}x & \text{if } |x| \le 2 \\ 0 & \text{if } |x| > 2 \end{cases}. \tag{4.2}$$

The convergence holds in expectation and almost surely.

Remark 4.2.

- Note the limit on the right is deterministic, yet the sequence on the left is random.
- Various versions of the proof of this result yield different notions of convergence to the limit: Wigner's original approach [Wig58] gave convergence in expectation (Not to be confused with convergence in mean or convergence in L^1!), i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}[f(\Lambda_i)] = \int f \mathrm{d}\sigma, \quad f \text{ bounded continuous.} \tag{4.3}$$

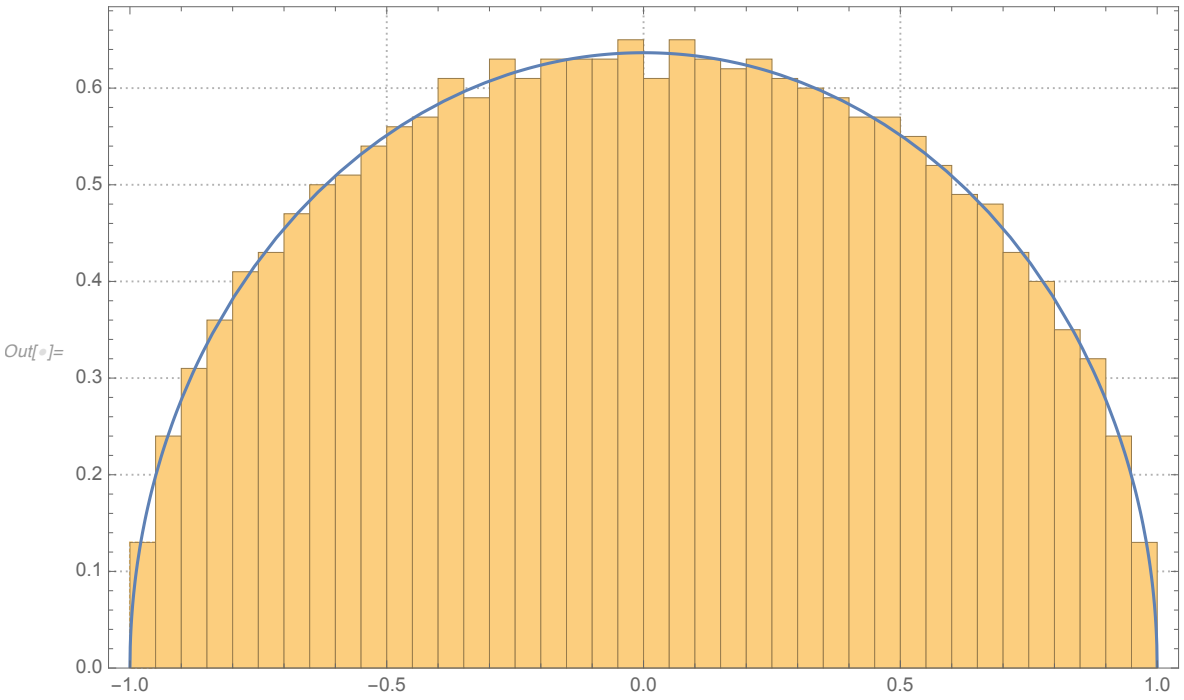

Figure 1: The spectral density of a single GOE matrix of dimension 2000 compared to the Wigner semicircle law.

- The assumption of finite second moment of M_{ij} is crucial. Without this, the theorem does not hold, see [BAG08].

Before proceeding with the proof, we first illustrate the semicircle law by showing the results of some numerical experiments. These experiments involved generating random matrices from the GOE and the GUE, plotting histograms of the eigenvalues, and comparing with the semicircle law (4.2). For ease of visualisation, the eigenvalues have each been divided by 2, so now the support of the rescaled semicircle is $|x| \leq 1$ and it has area $\pi/2$. Figure 1 shows the result for a single GOE matrix of dimension 2000, and Figure 2 shows the result of averaging over 100 GOE matrices of the same dimension. Similarly Figure 3 shows the result for a single GUE matrix of dimension 2000, and Figure 4 shows the result of averaging over 100 GUE matrices of the same dimension.

The strategy for proving the semicircle law is based on the method of moments (cf. Theorem B.10). (See also Section B.2 for a refresher on convergence of random variables.) This is done in two steps.

1. We first prove convergence in mean. Note that for each n, the collection of expectations

$$\mathbf{E}\left[\int f \mathrm{d}\mu_{\mathbf{M}/\sqrt{n}}\right], \quad f \text{ bounded continuous},$$

defines a deterministic probability measure, which is the *mean* of the ESM for a given n. Let's denote this by $\overline{\mu}_n$. In Theorem 4.6, we first show that the k-th moment $\overline{\mu}_n$ converges to the k-th moment of the semicircle law. The moments of the semicircle law are computed in the next section. The computation of the moments of $\overline{\mu}_n$ maps onto a problem in combinatorics, leading to a proof that they coincide with the moments of (4.2) in the limit when $n \to \infty$.

2. A second part of the proof of Theorem 4.6 shows that the moments of $\mu_{M/\sqrt{n}}$ converges to the moments of the semicircle distribution almost surely. This is done by computing the variance of the k-th moment $\int x^k \mathrm{d}\mu_{\mathbf{M}/\sqrt{n}}$ and applying the Borel-Cantelli lemma.

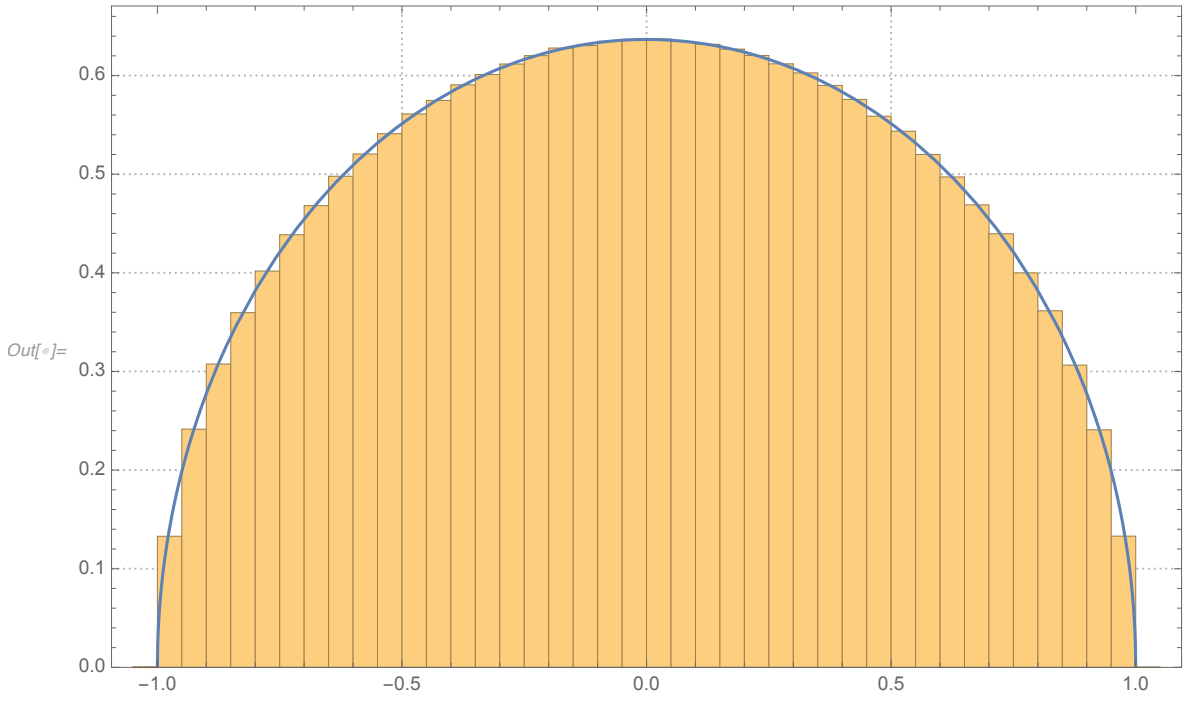

Figure 2: The spectral density obtained by averaging over 100 GOE matrices of dimension 2000 compared to the Wigner semicircle law.

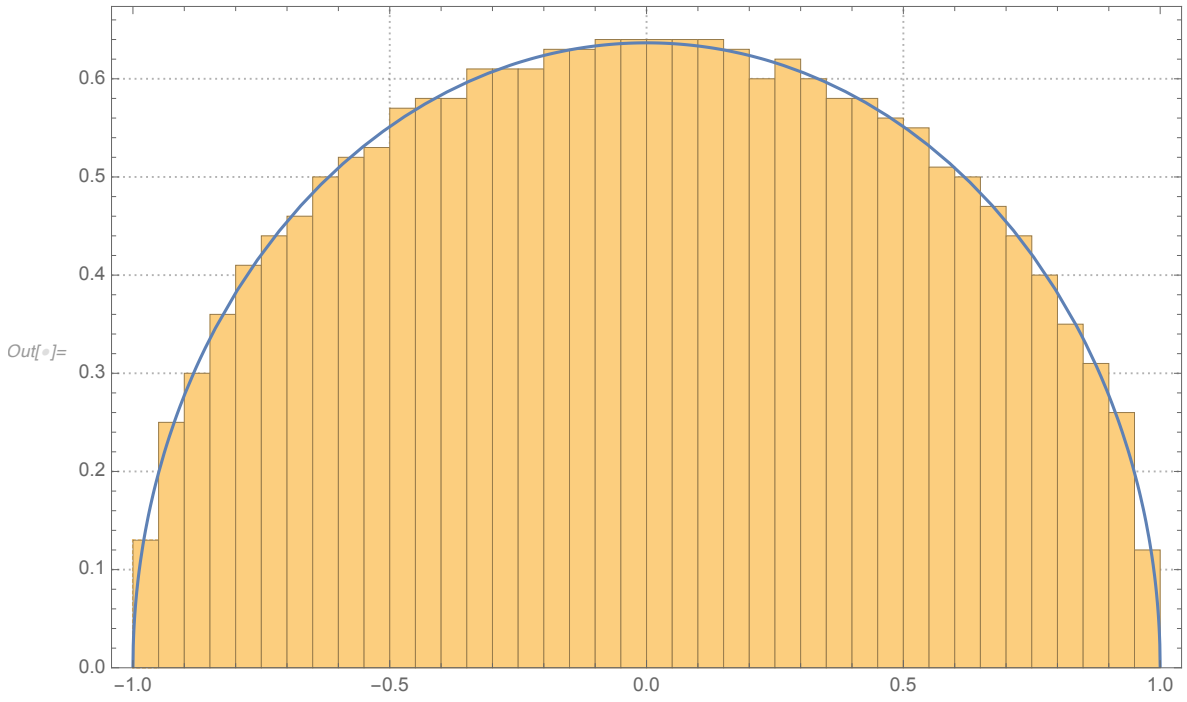

Figure 3: The spectral density of a single GUE matrix of dimension 2000 compared to the Wigner semicircle law.

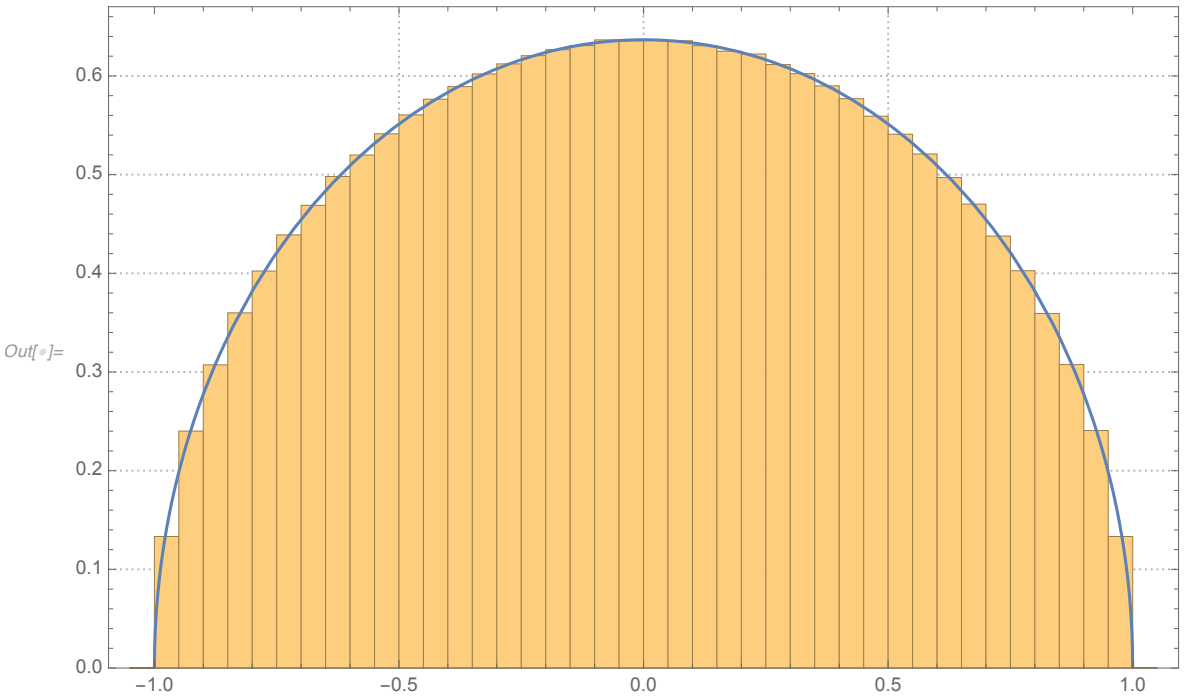

Figure 4: The spectral density obtained by averaging over 100 GUE matrices of dimension 2000 compared to the Wigner semicircle law.

4.1 Moments of the semicircle law and Catalan numbers

The moments of the semicircle law (4.2) are

Lemma 4.3. *Let $k \in \mathbb{N}$. The moments of the semicircle law are*

$$\alpha_k = \frac{1}{2\pi}\int_{-2}^{2} x^k \sqrt{4-x^2}\mathrm{d}x = \begin{cases} 0 & \text{if } k \text{ is odd;} \\ \frac{1}{m+1}\binom{2m}{m} & \text{if } k = 2m \text{ is even.} \end{cases} \tag{4.4}$$

The numbers

$$C_m = \frac{1}{m+1}\binom{2m}{m} \tag{4.5}$$

are known as the Catalan numbers.

Proof. Clearly, by symmetry, $\alpha_k = 0$ when k is odd. Setting $k = 2m$ we then have

$$\begin{aligned}
\alpha_{2m} &= \frac{1}{2\pi}\int_{-2}^{2} x^{2m}\sqrt{4-x^2}\mathrm{d}x \\
&= \frac{2^{2m+1}}{\pi}\int_0^{\pi} (\cos\theta)^{2m}\sin^2\theta \mathrm{d}\theta \\
&= \frac{2^{2m+1}}{\pi}\left[\int_0^{\pi}(\cos\theta)^{2m}\mathrm{d}\theta - \int_0^{\pi}(\cos\theta)^{2m+2}\mathrm{d}\theta\right] \\
&= \frac{2}{\pi}\left[\int_0^{\pi}(e^{i\theta}+e^{-i\theta})^{2m}\mathrm{d}\theta - \frac{1}{4}\int_0^{\pi}(e^{i\theta}+e^{-i\theta})^{2m+2}\mathrm{d}\theta\right] \\
&= 2\left[\binom{2m}{m} - \frac{1}{4}\binom{2m+2}{m+1}\right] \\
&= \frac{1}{m+1}\binom{2m}{m}, \qquad (4.6)
\end{aligned}$$

where, in passing from the fourth to the fifth line, we used the binomial expansion and the fact that only one term in this expansion gives a non-zero contribution when integrated. □

The Catalan numbers satisfy (among many other identities)

$$C_m = \binom{2m}{m} - \binom{2m}{m+1}, \qquad (4.7)$$

and the recurrence relation

$$C_{m+1} = \sum_{i=0}^{m} C_i C_{m-i}, \qquad (4.8)$$

with $C_0 = 1$. They are all integers (Why?). The sequence of Catalan numbers begins 1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, 208012, 742900, 2674440, 9694845, ... Note that it follows from (4.5) that $C_m \leq 4^m$. In particular, this implies that the semicircle law is determined by its moments, since by the dominated convergence theorem, we have

$$\int e^{\mathrm{i}\lambda x}\mathrm{d}\sigma(x) = \sum_{k\geq 0}\frac{\mathrm{i}^k\lambda^k\alpha_k}{k!},\ \lambda\in\mathbb{R}.$$

The Catalan numbers play an important role in combinatorics because they count various interesting objects, ranging from lattice paths to geometric constructions in polygons. See, for example, [Kos09, Sta15]. We now describe briefly one such connection.

Definition 4.4. For $k \in \mathbb{N}$, A *Dyck path* π of length $2k$ is a function on $\{0, 1\ldots, 2k\}$ such that

- the path starts and end at 0, i.e., $\pi(0) = \pi(2k) = 0$;
- for all $0 \leq j \leq 2k-1$, the increments $\pi(j+1) - \pi(j)$ are -1 or $+1$.
- the path stays non-negative, i.e., $\pi(j) \geq 0$ for all $0 \leq j \leq 2k$.

We denote the set of Dyck paths of length $2k$ by $\mathcal{D}_k$.

Call a simple walk a path that starts at 0 with increments ± 1. The Dyck paths of length $2k$ are exactly the simple walks from time 0 to $2k$ that start and end 0 at 0 while remaining positive. There are 2^{2k} simple walks. The number of Dyck paths is:

Proposition 4.5. *For every $k \in \mathbb{N}$, $\#\mathcal{D}_k = C_k$.*

This is an important combinatorial fact. We give two proofs.

Proof. To prove this, let $\mathcal{A}$ denote the set of walks of length $2k$ with increments ± 1 which start at 0 and end at -2, and similarly let $\mathcal{B}$ denote the set of all walks of length $2k$ with increments ± 1 which start at 0 and end at 0. Then $|\mathcal{D}_k| = |\mathcal{B}| - |\mathcal{A}|$, which may be seen by the *reflection principle*. More precisely, let $\mathcal{C}$ denote the set of walks of length $2k$ with increments ± 1 which start at 0 and end at 0, but which hit -1 at some intermediate point. Obviously $|\mathcal{D}_k| = |\mathcal{B}| - |\mathcal{C}|$. However, $|\mathcal{C}| = |\mathcal{A}|$ because the respective paths are in bijection: for a walk in $\mathcal{C}$, let j be the last visit to -1; reflecting the portion of the path after j about -1 gives a unique path terminating at -2 with the same set of visits to -1. Clearly $|\mathcal{B}| = \binom{2k}{k}$, because $\mathcal{B}$ consists of k steps $+1$ and k steps -1 in some order, and $|\mathcal{A}| = \binom{2k}{k+1}$, because $\mathcal{A}$ consists of $k-1$ steps $+1$ and $k+1$ steps -1. Hence

$$|\mathcal{D}_k| = \binom{2k}{k} - \binom{2k}{k+1} = C_k. \tag{4.9}$$

Alternatively, consider a Dyck path of length $2(k+1)$. Let $2(m+1)$ denote the first 'time' this path hits 0. We can break the path into two pieces: the section up to time $2(m+1)$ and the section after this time. The path up to time m is itself a Dyck path, as is the path afterwards. If we denote $|\mathcal{D}_n| = d_n$, then clearly

$$d_{k+1} = \sum_{m=0}^{k} d_m d_{k-m}, \tag{4.10}$$

and $d_0 = 1$. Hence d_k satisfies the recurrence relation (4.8). □

4.2 Spectral moments

In this section, we compute the k-th moments of the ESM, or *spectral moments*, of a real Wigner matrix $\mathbf{M}$ with eigenvalues $(\Lambda_1, \ldots, \Lambda_n)$ properly rescaled:

$$\frac{1}{n}\int x^k \mathrm{d}\mu_{\mathbf{M}/\sqrt{n}}(x) = \frac{1}{n}\sum_{i=1}^{n}(\Lambda_i/\sqrt{n})^k = \frac{1}{n^{k/2+1}}\mathrm{Tr}(\mathbf{M}^k). \tag{4.11}$$

Our goal now is to prove the following theorem.

Theorem 4.6 (Wigner [Wig58] [8])**.** *Let M be a $n \times n$ real Wigner matrix. Suppose that for all $k \in \mathbb{N}$ the following moment assumption is satisfied:*

$$B_k := \sup_{n\in\mathbb{N}} \sup_{(i,j)\in\{1,\ldots,n\}^2} \mathbf{E}[|\mathbf{M}_{ij}|^k] < \infty. \tag{4.12}$$

For any $k \in \mathbb{N}$, we have

$$\lim_{n\to\infty} \frac{1}{n^{k/2+1}}\mathrm{Tr}\mathbf{M}^k = \begin{cases} 0 & \text{if } k \text{ is odd} \\ C_{\frac{k}{2}} & \text{otherwise,} \end{cases} \tag{4.13}$$

where the convergence holds in expectation and almost surely.

It turns out that the moment assumption (4.12) is not necessary. It can be worked around using a truncation procedure [Tao12]. However, the finiteness of the second moment is crucial.

The main idea underlying the proof of this theorem is to expand $\mathrm{Tr}(\mathbf{M}^k)$ in terms of products of elements of $\mathbf{M}$, to use the independence of these to select those products that make a non-zero contribution to the expectation value of the trace, and then from these to identify the products of elements that make the largest contribution when $n \to \infty$.

We begin by proving Theorem 4.6 for convergence in expectation. When this is done we explain how to extend the proof to almost sure convergence.

[8] The paper is four pages long and ends with the sentence "The heuristic proof given for the special case considered before applies equally under the more general conditions here specified".

Proof of (4.13) *in expectation.* Taking the expectation of (4.13) without the limit gives

$$\mathbf{E}\left[\frac{1}{n}\mathrm{Tr}\Big((\frac{\mathbf{M}}{\sqrt{n}})^k\Big)\right] = \sum_{i_1,i_2,\ldots,i_k=1}^{n} \frac{1}{n^{1+k/2}}\mathbf{E}[\mathbf{M}_{i_1i_2}\mathbf{M}_{i_2i_3}\ldots\mathbf{M}_{i_ki_1}]. \tag{4.14}$$

For a given k, note that for all sets of indices, $\mathbf{E}[\mathbf{M}_{i_1i_2}\mathbf{M}_{i_2i_3}\ldots\mathbf{M}_{i_ki_1}]$ is uniformly bounded by B_k under the assumption (4.12) (Why?).

To evaluate the sum, we make an important connection with graphs. The sum is over ordered k-tuple $\mathbf{i} = (i_1, \ldots, i_k)$ where each i_j takes values in $\{1, \ldots, n\}$. The values taken by i_j can repeat. Now consider the set $V(\mathbf{i})$ of distinct values in contained in $\mathbf{i}$. Clearly $1 \leq \#V(\mathbf{i}) \leq k$. We now think of $V(\mathbf{i})$ as vertices of a graph, each of which labeled by their respective value. We construct an edge set $E(\mathbf{i})$ for $V(\mathbf{i})$ by putting an edge between two vertices if their labels appear consecutively in $\mathbf{i}$, so the two vertices appear as i_j, i_{j+1} (with the convention that $i_{k+1} = i_1$ in the sequence given by $\mathbf{i}$). By construction, the graph $(V(\mathbf{i}), E(\mathbf{i}))$ is connected. We now think of $\mathbf{i} = (i_1, i_2, \ldots, i_k)$ as a closed walk on the graph $(V(\mathbf{i}), E(\mathbf{i}))$ starting and ending at the vertex i_1. Note that, *a priori*, edges and vertices can be visited twice in this walk. **The evaluation of the sum then consists in understanding which types of walk have a dominant contribution to the sum.**

Some walks can be easily discarded. Indeed, since the matrix entries are independent, centred random variables, we have $\mathbf{E}[\mathbf{M}_{i_1i_2}\mathbf{M}_{i_2i_3}\ldots\mathbf{M}_{i_ki_1}] = 0$ whenever an edge is traversed only once. Therefore we can restrict to walks that traverse each edge at least twice, possibly in reverse (since the matrix is symmetric). This means that we can restrict our attention to underlying graphs with $\#E(\mathbf{i}) \leq k/2$, and hence $\#V(\mathbf{i}) \leq k/2 + 1$ distinct vertices.

Clearly, two k-tuple $\mathbf{i}$ and $\mathbf{i}'$ will carry the same weight in the sum if there exists a bijection on the set $\{1, 2, \ldots, n\}$ mapping each i_j to i'_j keeping the same edge structure. These two walks then only differ by the labeling of their vertices. We will say these walks are *equivalent*. The next step is to show that the walks with $\#V(\mathbf{i}) < k/2 + 1$ make a negligible contribution to the sum. As noted above, $\mathbf{E}[\mathbf{M}_{i_1i_2}\mathbf{M}_{i_2i_3}\ldots\mathbf{M}_{i_ki_1}]$ is uniformly bounded. It only remains to count the number of such walks For a given index sequence $i_1i_2\ldots i_ki_1$ with $\#V(\mathbf{i}) = w$, the number of equivalent walks to it is $n(n-1)(n-2)\ldots(n-w+1) \leq n^w$. The contribution of the terms in this equivalence class to the sum is therefore $\leq B_k n^{w-k/2-1}$. This tends to zero as $n \to \infty$ if $w < k/2 + 1$. Now, note that the number of distinct equivalent classes is independent of n (though it depends on k). Thus, the total contribution from walks with $w < k/2 + 1$ tends to zero in the limit. When k is even, it does so at least as fast as $1/n$.

It follows from the above that the only sequences making a non-zero contribution to (4.14) in the limit $n \to \infty$ are those for which $\#V(\mathbf{i}) = k/2 + 1$. When k is odd this equation has no solution, because w is an integer. Hence when k is odd the left-hand side of (4.14) tends to zero in the limit $n \to \infty$.

Consider now the case when k is even and $w = k/2 + 1$. In this case the graphs we need to consider are connected, have $k/2 + 1$ vertices and $k/2$ distinct edges. They are therefore *trees*, i.e., they have no cycles. To see this note that for any finite connected planar graph $G = (V, E)$ consisting of a set of vertices V and a set of edges E, we have the Euler relation

$$\#V - \#E + \#F = 2, \tag{4.15}$$

where $\#F$ is the number of faces defined by the graph. There is always one face (the one containing ∞). In particular, we have $\#V - \#E \leq 1$. There is equality if and only if $\#F = 1$, in which case the graph has no cycle. In other words $\#V - \#E = 1$ if and only if (V, E) is a tree.

When k is even and $w = k/2 + 1$, the sequence $\mathbf{i}$ therefore corresponds to a closed path on a tree that traverses each edge exactly twice, once in each direction. This means that for any $1 \leq j \leq k$ $i_{j+1} \neq i_j$. Hence in (4.14) only off-diagonal matrix entries contribute, and these are precisely paired, so each appears squared. Using the fact that $\mathbf{E}|M_{ij}|^2 = 1$ for $i \neq j$, and that the off-diagonal matrix

entries are independent, we have [9] for such $\mathbf{i}$'s

$$\mathbf{E}[\mathbf{M}_{i_1 i_2}\mathbf{M}_{i_2 i_3}\dots\mathbf{M}_{i_k i_1}] = 1. \tag{4.16}$$

The number of different labeling of the vertices for such trees is $n(n-1)(n-2)\dots(n-w+1) \sim n^{k/2+1}$, exactly canceling the normalization.

The problem has been reduced to that of counting the number of k-tuples corresponding to walks of length k on trees with $k/2+1$ vertices and $k/2$ distinct edges, each traversed twice. Such paths are called *non-crossing*. We can perform this count in the following way. As a given path is traversed, record at each edge on it whether that edge has been traversed before, or not. If it is being traversed for the first time, term it *open*, if it is being traversed for the second (and last) time, term in *closed*. For each non-crossing path of length k we associate a sequence, called the *path sequence*, whose jth entry is the number of open edges minus the number of closed edges in the path $i_1 i_2 \dots i_{j+1}$. Now we reach the main point: this sequence starts with a 1 and ends with a 0, and successive terms differ by ± 1. For example, for the path corresponding to the set of edges $(\{1,2\},\{2,3\},\{3,2\},\{2,4\},\{4,5\},\{5,4\},\{4,6\},\{6,4\},\{4,2\},\{2,1\})$, the edges are respectively open, open, closed, open, open, closed, open, closed, closed, closed, and the associated sequence is 1, 2, 1, 2, 3, 2, 3, 2, 1, 0. If we attach a label 0 at the start, then these sequences are clearly in bijection with the set of Dyck paths of length k.

Collecting together what we have shown so far, when k is even

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathbf{E}\mathrm{Tr}(\mathbf{M}^k) &= \#\{\text{path sequences of length } k\}\\ &= \#\{\text{Dyck paths of length } k\}\\ &= C_{k/2},\end{aligned} \tag{4.17}$$

and when k is odd, the limit is zero. This proves Theorem 4.6 for convergence in expectation. □

Proof of (4.13) *almost surely.* We next indicate how to prove almost sure convergence to the same limit. The strategy here is to compute the variance. We will see that the variance vanishes as $n \to \infty$ sufficiently fast that a sequence of matrices with increasing size will converge to the mean almost surely by the Borel-Cantelli lemma, Lemma B.8. It is an example of a phenomenon known as *concentration of measure* that plays an important role in random matrix theory.

The variance is

$$\mathrm{Var}\left(\frac{1}{n}\mathrm{Tr}(M^k)\right) = \mathbf{E}\left[\left(\frac{1}{n}\mathrm{Tr}(\mathbf{M}^k)\right)^2\right] - \mathbf{E}\left[\frac{1}{n}\mathrm{Tr}(\mathbf{M}^k)\right]^2 \tag{4.18}$$

and so by expanding the moments, we get

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{n}\mathrm{Tr}(\mathbf{M}^k)\right) = \frac{1}{n^{2+k}}\sum_{\mathbf{i}}\sum_{\mathbf{j}} &\mathbf{E}[\mathbf{M}_{i_1 i_2}\mathbf{M}_{i_2 i_3}\dots\mathbf{M}_{i_k i_1}\mathbf{M}_{j_1 j_2}\mathbf{M}_{j_2 j_3}\dots\mathbf{M}_{j_k j_1}]\\ &- \mathbf{E}[\mathbf{M}_{i_1 i_2}\mathbf{M}_{i_2 i_3}\dots\mathbf{M}_{i_k i_1}]\mathbf{E}[\mathbf{M}_{j_1 j_2}\mathbf{M}_{j_2 j_3}\dots\mathbf{M}_{j_k j_1}],\end{aligned} \tag{4.19}$$

where the sums are over k-tuples $\mathbf{i} = (i_1,\dots,i_k)$ and $\mathbf{j} = (j_1,\dots,j_k)$ with coordinates taking values in $\{1,\dots,n\}$. As in the calculation of the expectation value of the moments above, one can analyse this sum in terms of walks on a graph. We write $G(\mathbf{i},\mathbf{j}) = (V(\mathbf{i},\mathbf{j}), E(\mathbf{i},\mathbf{j}))$ for the graph constructed from $\mathbf{i}$ and $\mathbf{j}$. In the case of the first term on the right-hand side of (4.19), the vertices of the graph in question are now labelled by the distinct values of $\{i_1, i_2, \dots, i_k, j_1, j_2, \dots, j_k\}$.

For this graph to contribute to the sum, it must be connected, so $i_1 = j_1$, otherwise the first term factorizes and is exactly cancelled by the second term. As before, each edge must appear at least twice and must be traversed an even number of times. From the same argument as in the proof for the

[9] When $f(n)/g(n) \to 1$ as $n \to \infty$ we write $f(n) \sim g(n)$.

expectation, the terms making a non-zero contribution to (4.19) have $\#V(\mathbf{i},\mathbf{j}) \leq k+1$. The fact that the summand is uniformly bounded by (4.12) directly implies that

$$\mathrm{Var}\left(\frac{1}{n}\mathrm{Tr}(\mathbf{M}^k)\right) \leq \frac{c_k}{n}, \tag{4.20}$$

for some positive constant c_k depending on k. This proves convergence in L^2 of the moments. This also proves convergence in probability by Chebyshev's inequality.

We need a bound summable in n to lift this convergence to convergence almost sure by applying the Borel-Cantelli lemma B.8. With analogy with the convergence in expectation, the largest contribution is expected to come from paths with $\#V(\mathbf{i},\mathbf{j}) = k+1$. The number of edges $\#E(\mathbf{i},\mathbf{j})$ is smaller or equal to k since each edge must be traversed twice. The Euler relation implies that the number of edges is exactly k, and the graph is a tree. We now argue that such a graph $G(\mathbf{i},\mathbf{j})$ cannot exist. To see this, note that $G(\mathbf{i})$ and $G(\mathbf{i})$ are two subgraphs of $G(\mathbf{i},\mathbf{j})$. But $G(\mathbf{i},\mathbf{j})$ is connected! So there must be an edge linking $G(\mathbf{i})$ to $G(\mathbf{j})$. Such an edge must be traversed exactly twice by the walk $(\mathbf{i},\mathbf{j})$. But this is impossible since the walk starts at i_1 and comes back to i_1 in the $\mathbf{i}$-portion of the walk visiting the common edge only once. This means the graph would contain a cycle, which is a contradiction. We conclude that $\#V(\mathbf{i},\mathbf{j}) < k+1$, and the inequality (4.20) is sharpened to a $1/n^2$ decay.

□

We can now easily complete the proof of Theorem 4.1 under the extra assumption of uniform boundedness of the moments (4.12).

Proof of Theorem 4.1. For any $k \in \mathbb{N}$, we know from Theorem 4.6 that there exists an event A_k of probability 1 such that on A_k, $\int x^k \mathrm{d}\mu_{\mathbf{M}/\sqrt{n}} \to \int x^k \mathrm{d}\sigma$. Now consider $A = \bigcap_{k\in\mathbb{N}} A_k$. This event has probability 1. Moreover, on this event A,

$$\int x^k \mathrm{d}\mu_{\mathbf{M}/\sqrt{n}} \to \int x^k \mathrm{d}\sigma, \text{ for all } k \in \mathbb{N}.$$

By the method of moments (Theorem B.10), we get that $\mu_{\mathbf{M}/\sqrt{n}}$ converges in distribution to σ. In particular, this implies convergence of the integral of any bounded continuous function. □

4.3 Applications

The semicircle law has applications in settings where real-symmetric or complex-Hermitian matrices arise and can be modelled by random matrices. As noted in the introduction, many problems in population dynamics, modelling random neural networks, etc, reduce to the following mathematical question. The seminal paper on this subject is [May72].

Let $x(t)$ be an n-dimensional vector satisfying

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\lambda \mathrm{I} x \tag{4.21}$$

where λ is a positive constant and I is the $n \times n$ identity matrix. This system has a stable fixed point at $x = 0$. Clearly in (4.21) the different components of x are uncoupled (because I is diagonal). The question is: if one now introduces random couplings between these components, does this make the fixed point more or less stable? Specifically, what is the stability of the system

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\lambda \mathrm{I} x + \mathbf{A} x \tag{4.22}$$

where $\mathbf{A}$ is a random matrix. What, for example, should one expect if $|\lambda|$ is large compared to the typical size of the elements of $\mathbf{A}$?

In population dynamics, this is called the *May model*. Then the components of x represent the differences of the populations of various species from some equilibrium values and $\mathbf{A}$ represents interactions between the different species.

Now, if $\mathbf{A}$ is a real symmetric random matrix in which the mean-square size of the matrix entries (i.e. the interactions) is 1, it is a consequence of the semicircle law that the eigenvalues of $\mathbf{A}$ lie between $-2\sqrt{n}$ and $2\sqrt{n}$ with increasing probability as n grows. We expect to find eigenvalues over the whole of this range, but very few outside it. And any outside it will not be far from it. So when $n > \lambda^2/4$ we expect to see a transition in the dynamics from exponential stability to exponential instability. In large random neural networks, this question of stability is related to a transition in the dynamics from ordered to 'chaotic'.

If $\mathbf{A}$ is not symmetric, so the matrix entries are independent random variables, the circular law (briefly discuss in Section 8) implies a similar transition occurs for large enough systems. In this case the stability also depends on the eigenvectors, which no longer need be (and typically are not) orthogonal, so the details are more complicated, but the qualitative picture is the same.

Now, one can question whether it is reasonable to model the interactions as IID random variables with the same variance. In population dynamics, some species pairs interact far more strongly than others. For example, lions interact strongly with zebras, wildebeests, and impalas (and the interaction is clearly not symmetric!), but barely at all with butterflies or fish. There are food chains and these give rise to complex networks. Building this topological structure into population models is a significant area of research involving random matrices with structured correlations between the matrix entries and where the matrices may be rather sparse. It is also an important area of research in neuroscience, where again the connectivity matrix for neural pathways is both highly structured and also stochastic.

Finally, we remark that this discussion relates to *linear stability*. Extending it to nonlinear stability analysis is currently an active area of research.

5 The Empirical Spectral Measure of GUE

In the previous section, we calculated the spectral moments for general Wigner random matrices. We now look at the special case when the matrix elements have a Gaussian distribution. In this case we can go more deeply into the formula embodied in Theorem 4.6 and compute the moments exactly for finite n. Recall that Wigner's theorem relates the moments to the Catalan numbers in the limit as $n \to \infty$, and that this is proved by linking them to counts of certain non-crossing paths on graphs. In this limit, they therefore have an interpretation in terms of counting topologically defined objects. We shall see that for Gaussian random matrices this may be viewed as part of a bigger picture.

To start with, we need to explain a formula that will prove useful in analysing the expectation value on the right-hand side of (4.14).

5.1 Wick's theorem

Let X be a real standard Gaussian random variable. The Gaussian moments are given by

$$\mathbf{E}[X^n] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^n e^{-y^2/2} \mathrm{d}y = \begin{cases} 0 & \text{if } n \text{ is odd} \\ \frac{(2m)!}{2^m m!} & \text{if } n = 2m \text{ is even.} \end{cases} \tag{5.1}$$

Therefore $\mathbf{E}[X^n]$ is equal to the number of ways of splitting n objects into disjoint pairs. For example $\mathbf{E}[X^4] = 3$, corresponding to the 3 ways to split the numbers $1, 2, 3, 4$ into disjoint (unordered) pairs, i.e., $\{\{1,2\},\{3,4\}\}, \{\{1,3\},\{2,4\}\}$, and $\{\{1,4\},\{2,3\}\}$.

The following is then a straightforward application of (5.1) due to Isserlis in 1918 and Wick in 1950 (and many others). In probability, this is a simple application of *Gaussian integration by parts*, cf. Lemma B.2.

Theorem 5.1. *Let $Y_1, \ldots, Y_p$ be independent standard Gaussian random variables. Consider a collection $X_1, \ldots, X_n \in \{Y_1, \ldots, Y_p\}$ with possible repetition (so we can have $n > p$ for example). Then*

$$\mathbf{E}[X_1 X_2 \ldots X_n] = \sum_{\pi \in \mathcal{P}_2(n)} \prod_{\{i,j\} \in \pi} \mathbf{E}[X_i X_j], \tag{5.2}$$

where $\mathcal{P}_2(n)$ stands for the collection of pairings of n objects into disjoint unordered pairs.

Importantly, the covariances in the Wick formula can be computed straightforwardly:

$$\mathbf{E}[X_i X_j] = \begin{cases} 1 & \text{if } X_i = X_j \\ 0 & \text{if } X_i \neq X_j. \end{cases} \tag{5.3}$$

Note that because it is linear in the variables $X_1, \ldots, X_n$, this formula applies in exactly the same way to standard complex Gaussian random variables $Z = (X + iY)/\sqrt{2}$, where X and Y are independent real standard Gaussian random variables. In this case $\mathbf{E}[Z] = \mathbf{E}[Z^2] = 0$ and $\mathbf{E}[|Z|^2] = 1$. If $Z_1, \ldots, Z_p$ are independent complex standard Gaussian random variables then for $z_1, \ldots, z_n \in \{Z_1, \ldots, Z_p\}$, $\mathbf{E}[z_1 z_2 \ldots z_k]$ is equal to the number of pairings which connect z_i with its complex conjugate. In general, it is not hard to derive a formula for Gaussian vectors whose coordinates are not necessarily IID standard.

Historically, Isserlis introduced this formula in the context of statistical analysis. It was later re-introduced by Wick to count Feynman diagrams in Quantum Field Theory and it tends now to be associated with his name.

5.2 The genus expansion for Gaussian random matrices

Let $\mathbf{M}$ be a $n \times n$ complex Wigner matrix. The moments of the ESM of $\mathbf{M}/\sqrt{n}$ can be expressed as before as

$$\mathbf{E}\left[\frac{1}{n^{1+k/2}} \mathrm{Tr}(\mathbf{M}^k)\right] = \sum_{i_1, i_2, \ldots, i_k = 1}^{n} \frac{1}{n^{1+k/2}} \mathbf{E}[\mathbf{M}_{i_1 i_2} \mathbf{M}_{i_2 i_3} \ldots \mathbf{M}_{i_k i_1}]. \tag{5.4}$$

In the previous section, we analysed this for general Wigner matrices and found that when $n \to \infty$ the leading order asymptotics may be calculated by counting non-crossing paths. If we now specialize to GUE matrices (i.e., complex Hermitian Gausssian random matrices), we can apply Wick's theorem to compute this expression exactly.

We have from Wick's theorem that

$$\mathbf{E}[\mathbf{M}_{i_1 i_2}\mathbf{M}_{i_2 i_3}\dots\mathbf{M}_{i_k i_1}] = \sum_{\pi\in\mathcal{P}_2(k)}\prod_{(a,b)\in\pi}\mathbf{E}[\mathbf{M}_{i_a,i_{a+1}}\mathbf{M}_{i_b,i_{b+1}}], \tag{5.5}$$

where $i_{k+1} = i_1$. Hence, we get

$$\mathbf{E}\left[\mathrm{Tr}(\mathbf{M}^k)\right] = \sum_{\pi\in\mathcal{P}_2(k)}\sum_{i_1,i_2,\dots,i_k=1}^{n}\prod_{(a,b)\in\pi}\delta_{i_a,i_{b+1}}\delta_{i_{a+1},i_b}, \tag{5.6}$$

where $\delta_{i,j}$ denotes the Kronecker δ-symbol.[10]

We can think of π as a particular kind of permutation on k elements, namely one that factorizes into permutations between the elements being paired, so permutations made up of cycles of length 2. Hence $(a,b)\in\pi$ means that $\pi(a) = b$, or equivalently $\pi(b) = a$. This implies (double-check this)

$$\prod_{(a,b)\in\pi}\delta_{i_a,i_{b+1}}\delta_{i_{a+1},i_b} = \prod_{a=1}^{k}\delta_{i_a,i_{\pi(a)+1}}. \tag{5.7}$$

If we introduce the shift notation $\gamma(a) = a+1 \bmod k$ on the set $\{1,2,\dots,k\}$, so that $\gamma\in S_k$ is a permutation with cycle $(1,2,\dots,k)$, then we have that

$$\mathbf{E}\left[\mathrm{Tr}(\mathbf{M}^k)\right] = \sum_{\pi\in\mathcal{P}_2(k)}\sum_{i_1,i_2,\dots,i_k=1}^{n}\prod_{a=1}^{k}\delta_{i_a,i_{\gamma\pi(a)}}. \tag{5.8}$$

One can think of the k-tuple $\mathbf{i} = (i_1,i_2,\dots,i_k)$ as a function $\mathbf{i}:\{1,\dots,k\}\to\{1,\dots,n\}$. Then

$$\prod_{a=1}^{k}\delta_{i_a,i_{\gamma\pi(a)}} = \begin{cases}1 & \text{if } \mathbf{i} \text{ is constant on the cycles of } \gamma\pi\\ 0 & \text{otherwise.}\end{cases} \tag{5.9}$$

We therefore have that

$$\mathbf{E}\left[\mathrm{Tr}(\mathbf{M}^k)\right] = \sum_{\pi\in\mathcal{P}_2(k)}\#\Big\{\mathbf{i}:\{1,\dots,k\}\to\{1,\dots,n\} : \mathbf{i} \text{ is constant on the cycles of } \gamma\pi\Big\}. \tag{5.10}$$

The count in the summand is straightforward: we simply need to choose one value for each of the cycles in $\gamma\pi$ (with repeats allowed). For any permutation $\sigma\in S_k$, let $\#(\sigma)$ denote the number of cycles in σ. Then we have

$$\frac{1}{n^{k/2+1}}\mathbf{E}\left[\mathrm{Tr}M^k\right] = \sum_{\pi\in\mathcal{P}_2(k)} n^{\#(\gamma\pi)-k/2-1}. \tag{5.11}$$

Equation (5.11) is an exact formula for the spectral moments of $\mathbf{M}$. To understand its structure, note first that if k is odd then $\mathcal{P}_2(k)$ is empty and so the corresponding spectral moment is identically zero. We therefore now set $k = 2m$, so that

$$\frac{1}{n^{m+1}}\mathbf{E}\left[\mathrm{Tr}(\mathbf{M}^{2m})\right] = \sum_{\pi\in\mathcal{P}_2(2m)} n^{\#(\gamma\pi)-m-1}. \tag{5.12}$$

This sum is known as the *genus expansion*, because there is an interpretation of $\#(\gamma\pi)$ that combines geometric and topological ideas.

[10] Recall that for standard complex normal random variables, we have to pair a random variable in Wick's theorem with its complex conjugate, and that the matrices we are dealing with are Hermitian.

Theorem 5.2. *Let* $\mathbf{M}$ *be a* $n \times n$ *GUE matrix. Then for any* $m \in \mathbb{N}$*, we have*

$$\frac{1}{n^{m+1}} \mathbf{E}\left[\mathrm{Tr}(\mathbf{M}^{2m})\right] = \sum_{g \geq 0} \tau_g(m) n^{-2g}, \tag{5.13}$$

where

$$\tau_g(m) = \#\{\textit{genus-}g\textit{ surfaces obtained by gluing together pairs of edges of a } 2m\textit{-gon}\}. \tag{5.14}$$

This remarkable formula therefore implies that averaging $\mathrm{Tr}(\mathbf{M}^{2m})$ over the GUE leads to a method for counting the number of genus-g surfaces obtained by gluing together pairs of edges of a $2m$-gon!

Proof. We start by drawing a $2m$-gon and labelling its vertices in cyclic order $v_1, v_2, \ldots, v_{2m}$. We can then label its edges by the vertices they connect; so for $i = 1, \ldots, 2m-1$, $e_i = v_i v_{i+1}$, and $e_{2m} = v_{2m} v_1$. A pairing $\pi \in \mathcal{P}_2(2m)$ corresponds to an association between pairs of edges, which can then be glued together to form a compact surface. If this is done so that when $v_i v_{i+1}$ is glued to $v_j v_{j+1}$, v_i is glued to v_{j+1}, and v_{i+1} to v_j, then the surface that results is orientable. So, for example, if $\pi(1) = 3$ one identifies e_1 and e_3 by gluing v_1 to v_4 and v_2 to v_3. We term this the *tail-to-head* convention.

Consider now the surface $\mathscr{S}_\pi$ obtained by this gluing procedure. The number of distinct vertices in the graph G_π of the $2m$-gon in $\mathscr{S}_\pi$ is precisely $\#(\gamma\pi)$. To see this, note that e_i is glued to $e_{\pi(i)}$, and so v_i is glued to $v_{\gamma\pi(i)}$ for each $i \in \{1, \ldots, 2m\}$. The edge $e_{\gamma\pi(i)}$ is glued to $e_{\pi\gamma\pi(i)}$ and so $v_{\gamma\pi(i)}$, which is now the tail of the edge in question, gets glued to $v_{\gamma\pi\gamma\pi(i)}$ etc. Continuing on this way, we see that v_i ends up being identified with precisely those v_j for which $j = (\gamma\pi)^l(i)$ for some $l \in \mathbb{N}$. Therefore the cycles of $\gamma\pi$ count the number of distinct vertices after gluing.

The *Euler characteristic* of $\mathscr{S}_\pi$, $\chi(\mathscr{S}_\pi)$, is an even integer which may be defined as follows: if $G = (V, E)$ is *any* embedded polygonal complex in $\mathscr{S}_\pi$, then

$$\chi(\mathscr{S}_\pi) = \#V - \#E + \#F, \tag{5.15}$$

where F is the number of faces of G. This is a generalization of the relation (4.15) that holds for planar graphs where $\chi = 2$. Now, any orientable compact surface is homeomorphic to a g-holed torus for some $g \geq 0$ (the $g = 0$ case is the sphere). The topological invariant g is known as the *genus* of the surface. It is a theorem of Cauchy that the Euler characteristic is 2 minus twice the genus, and so

$$\chi(\mathscr{S}_\pi) = 2 - 2g(\mathscr{S}_\pi). \tag{5.16}$$

The embedded complex G_π in $\mathscr{S}_\pi$ that we constructed above has one face, m edges, and $\#(\gamma\pi)$ vertices, thus

$$2 - 2g(\mathscr{S}_\pi) = \chi(\mathscr{S}_\pi) = \#(\gamma\pi) - m + 1. \tag{5.17}$$

Putting all these observations together, we get from (5.12) that

$$\frac{1}{n^{m+1}} \mathbf{E}\left[\mathrm{Tr}(\mathbf{M}^{2m})\right] = \sum_{\pi \in \mathcal{P}_2(2m)} n^{\#(\gamma\pi) - m - 1} = \sum_{\pi \in \mathcal{P}_2(2m)} n^{-2g(\mathscr{S}_\pi)} = \sum_{g \geq 0} \tau_g(m) n^{-2g}, \tag{5.18}$$

as claimed. □

As a consequence of the theorem, we see immediately that since $g \geq 0$,

$$\lim_{n \to \infty} \frac{1}{n^{m+1}} \mathbf{E}\left[\mathrm{Tr}(\mathbf{M}^{2m})\right] = \tau_0(m) = \#\{\text{spheres obtained by gluing together pairs of edges of a } 2m\text{-gon}\}. \tag{5.19}$$

In particular, Wigner's theorem 4.6 implies

$$\#\{\text{spheres obtained by gluing together pairs of edges of a } 2m\text{-gon}\} = C_m \tag{5.20}$$

where C_m is the m-th Catalan number. This is the simplest of a wide and important variety of examples where random matrix theory connects with enumerative topology.

6 The Marchenko-Pastur law for Wishart random matrices

The Marchenko-Pastur[11] law is the analogue for Wishart random matrices of the Wigner semicircle law for Hermitian matrices.

6.1 The Marchenko-Pastur Theorem

Let $\mathbf{X}$ be a $p \times n$ matrix with entries $\mathbf{X}_{ij}$ that are IID real random variables with $\mathbf{E}[X_{ij}] = 0$ and $\mathbf{E}[\mathbf{X}_{ij}^2] = 1$. Denote by Σ the $p \times p$ matrix given by

$$\Sigma = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathrm{T}} \in \mathbb{R}^{p\times p}. \tag{6.1}$$

(As for Wigner matrices, we suppress the dependence on p and n in the notation of the matrix.) Denote by $\Lambda_1, \Lambda_2, \ldots, \Lambda_p$, the eigenvalues of Σ. We consider the empirical spectral measure of Σ

$$\mu_p = \frac{1}{p}\sum_{j=1}^{p} \delta_{\Lambda_j}. \tag{6.2}$$

The Marchenko-Pastur law asserts that:

Theorem 6.1. *Let Σ be a $p \times p$ random matrix constructed as in (6.1), and let μ_p be its corresponding ESM. Suppose also that for all $k \in \mathbb{N}$*

$$B_k := \sup_{n\in\mathbb{N}} \sup_{(i,j)\in\{1,\ldots,n\}^2} \mathbf{E}[|\mathbf{X}_{ij}|^k] < \infty. \tag{6.3}$$

When $p \to \infty$ and $n \to \infty$ such that $p/n \to \gamma$ for some $\gamma \in (0, 1]$, then μ_p converges to μ in expectation and almost surely, where μ is the deterministic measure satisfying

$$\frac{\mathrm{d}\mu}{\mathrm{d}x} = \begin{cases} \frac{1}{2\pi\gamma x}\sqrt{(a_+ - x)(x - a_-)} & \text{if } a_- \le x \le a_+ \\ 0 & \text{otherwise,} \end{cases} \tag{6.4}$$

where $a_- = (1-\sqrt{\gamma})^2$ and $a_+ = (1+\sqrt{\gamma})^2$. When $\gamma > 1$, one needs to add $(1-\gamma^{-1})\delta_0$ to the right-hand side of (6.4), where δ_0 is a Dirac delta at the origin.

Remark 6.2. The additional factor $(1 - \gamma^{-1})\delta_0$ which needs to be added to the right-hand side of (6.4) when $\gamma > 1$ is explained by the fact that, since the rank of σ_n is the smaller of p and n, there are then approximately $p - n$ zero eigenvalues which will contribute a mass of $(1 - \gamma^{-1})$ at 0 in the limiting measure.

Note that when $\gamma = 1$, $a_- = 0$ and $a_+ = 4$, and that then (6.4) coincides with the semicircle law under the mapping $x \mapsto x^2$.

We illustrate the Marchenko-Pastur law by showing the results of numerical experiments. These involved generating random matrices Σ, plotting histograms of the eigenvalues, and comparing with (6.4). Figure 5 shows the result when $p = 10^3$ and $n = 10^4$, compared to (6.4) with $\gamma = 0.1$. Similarly Figure 6 shows the result when $p = 1000$ and $n = 2000$, compared to (6.4) with $\gamma = 0.5$.

One can prove the Marchenko-Pastur formula using the method of moments as we did for the semicircle law. The main steps are:

1. Prove that the spectral moments converge in expectation to the moments of the limiting deterministic Marchenko-Pastur distribution when $n \to \infty$.

[11] Sometimes written Marčenko-Pastur.

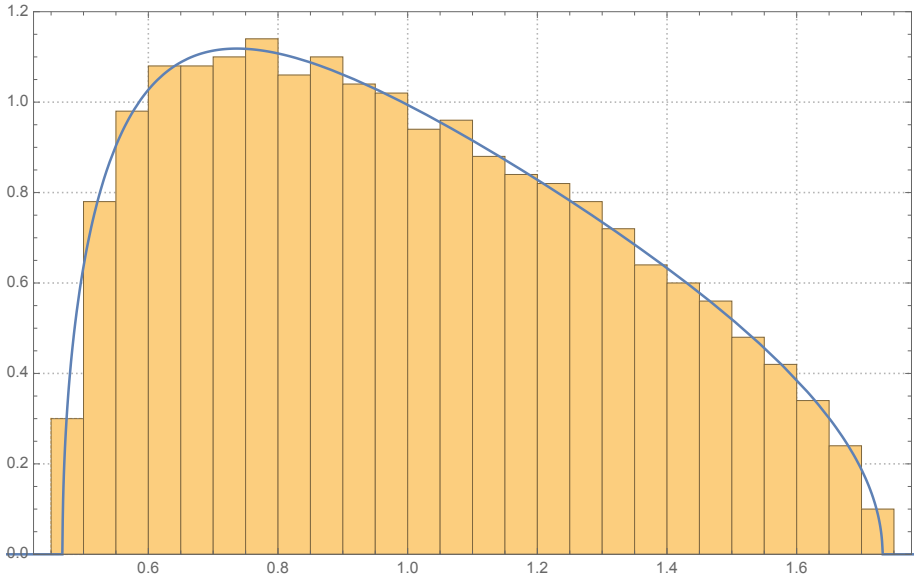

Figure 5: Eigenvalue density when $p = 10^3$ and $n = 10^4$, compared to (6.4) with $\gamma = 0.1$.

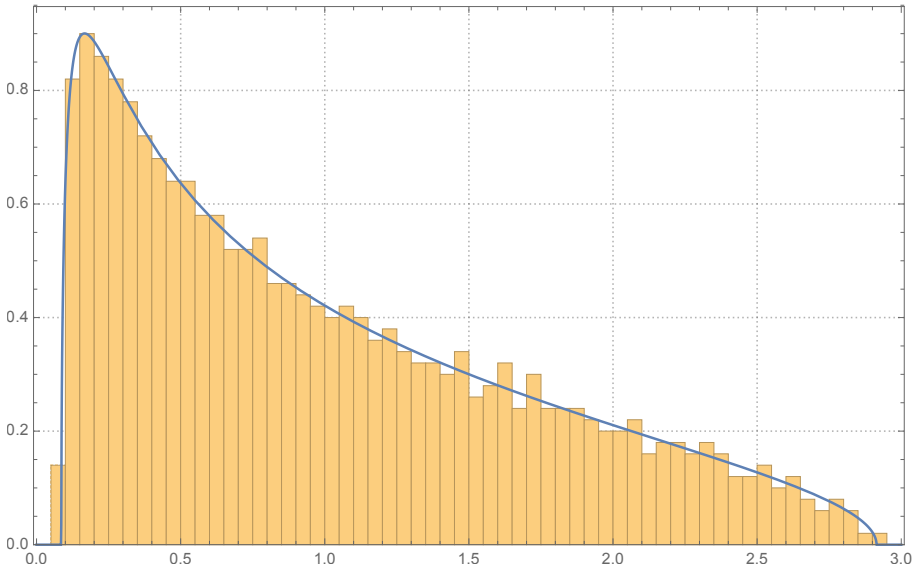

Figure 6: Eigenvalue density when $p = 1000$ and $n = 2000$, compared to (6.4) with $\gamma = 0.5$.

2. Lift this convergence to almost sure convergence by showing that the variance of the moments vanishes quickly in the limit when $n \to \infty$, and then by using the Borel-Cantelli lemma to conclude that the moments converge almost surely.

We focus on the first step. The second follows very similar lines to the corresponding ones in the proof of the semicircle law and so we shall not repeat it here.

6.2 Moments of the Marchenko-Pastur distribution

We start by computing the moments of the Marchenko-Pastur and show that the distribution is determined by its moments. We focus on the case when $\gamma \leq 1$.

Lemma 6.3. *Let $\gamma \leq 1$ and $a_\pm = (1 \pm \sqrt{\gamma})^2$. Consider the Marchenko-Pastur distribution given in (6.4). Its k-th moment, $k \in \mathbb{N}$, is given by*

$$\int_{a_-}^{a_+} x^k \frac{1}{2\pi\gamma x}\sqrt{(a_+ - x)(x - a_-)}\mathrm{d}x = \sum_{r=0}^{k-1} \frac{\gamma^r}{r+1}\binom{k}{r}\binom{k-1}{r}. \tag{6.5}$$

Proof. Note that $a_- + a_+ = 2(1+\gamma)$ and that $a_- a_+ = (1-\gamma)^2$, and hence the k-th moment is

$$\frac{1}{2\pi}\int_{-2}^{2} (\sqrt{\gamma}y + 1 + \gamma)^{k-1}\sqrt{4 - y^2}\mathrm{d}y. \tag{6.6}$$

Expanding $(\sqrt{\gamma}y + 1 + \gamma)^{k-1}$ binomially in powers of y, and using the formula for the moments of the semicircle law and Vandermonde's identity

$$\binom{m+n}{k} = \sum_{r=0}^{k}\binom{m}{r}\binom{n}{k-r} \tag{6.7}$$

yields the result (Exercise!). □

The k-th moments is trivially bounded by a_+^k, which is enough for the moments to determine the distribution.

6.3 Proof of the Marchenko-Pastur law (Theorem 6.1)

The goal is to show that

$$\mathbf{E}\left[\frac{1}{p}\sum_{i=1}^{p}\Lambda_i^k\right] = \mathbf{E}\left[\frac{1}{p}\mathrm{Tr}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^k\right] \tag{6.8}$$

converges to (6.5) in the appropriate limit.

As in the case of Wigner's semicircle theorem, we proceed by expanding the trace in terms of the matrix entries. We then use the independence of these to identify the products that make the largest contributions. We have

$$\begin{aligned}\mathbf{E}\left[\frac{1}{p}\mathrm{Tr}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^k\right] &= \frac{1}{pn^k}\sum_{\mathbf{i},\mathbf{j}}\mathbf{E}\left[\mathbf{X}_{i_1 j_1}\mathbf{X}^{\mathrm{T}}_{j_1 i_2}\mathbf{X}_{i_2 j_2}\mathbf{X}^{\mathrm{T}}_{j_2 i_3}\dots\mathbf{X}_{i_k j_k}\mathbf{X}^{\mathrm{T}}_{j_k i_1}\right] \\ &= \frac{1}{pn^k}\sum_{\mathbf{i},\mathbf{j}}\mathbf{E}\left[\mathbf{X}_{i_1 j_1}\mathbf{X}_{i_2 j_1}\mathbf{X}_{i_2 j_2}\mathbf{X}_{i_3 j_2}\dots\mathbf{X}_{i_k j_k}\mathbf{X}_{i_1 j_k}\right],\end{aligned} \tag{6.9}$$

where the sum is over k-tuples $\mathbf{i} = (i_1, \dots, i_k) \in \{1, \dots, p\}^k$ and k-tuples $\mathbf{j} = (j_1, \dots, j_k) \in \{1, \dots, n\}^k$.

As for Wigner matrices, we represent the sums in terms of walks on a graph. Here, to each pair $\mathbf{i}\,\mathbf{j}$, we associated a *directed bipartite graph* $G(\mathbf{i},\mathbf{j})$. The vertices are labeled by the distinct values of $i_1, i_2, \ldots, i_k$ on one side and the distinct values of $j_1, j_2, \ldots, j_k$ in the other. The edges are directed and go from i_1 to j_1, j_1 to i_2, i_2 to j_2, j_2 to i_3, etc, and with finally an edge from j_k to i_1. The walk represented by the two k-tuples $\mathbf{i}$ and $\mathbf{j}$ is of length $2k$ and starts and ends at the vertex i_1.

Since $\mathbf{E}\mathbf{X}_{ij} = 0$ and the $\mathbf{X}_{ij}$ are independent, the summand in (6.9) vanishes unless each edge appears at least twice in a walk. Hence there can be at most k edges and therefore $k+1$ vertices in $G(\mathbf{i},\mathbf{j})$.

We consider first the case when $\#V(\mathbf{i},\mathbf{j}) < k+1$. Say the walk visits $\ell_{\mathbf{i}}$ $\mathbf{i}$-vertices and $\ell_{\mathbf{j}}$ $\mathbf{j}$-vertices (so $\#V = \ell_{\mathbf{i}} + \ell_{\mathbf{j}}$). The total number of ways of choosing these sites is bounded from above by a constant (i.e. a factor independent of n) times $p^{\ell_{\mathbf{i}}} n^{\ell_{\mathbf{j}}}$. Hence the contribution of all such terms in (6.9) is bounded from above by a constant (that might depend on k using the assumption (6.3)) times $p^{\ell_{\mathbf{i}}} n^{\ell_{\mathbf{j}}}/pn^k$, which tends to zero as $n \to \infty$ if $p \sim \gamma n$.

As in the proof of the semicircle law, we therefore need to focus on the case when there are exactly k edges and $k+1$ vertices in the graph $G(\mathbf{i},\mathbf{j})$. The graph then is a *double tree*, namely a tree where vertices can be labeled either by values in $\mathbf{i}$ or values in $\mathbf{j}$ and these are visited alternately on the walk, which passes through each edge exactly twice. So, if we start at i_1, other $\mathbf{i}$-vertices can be reached if and only if the number of steps taken is even, and $\mathbf{j}$-vertices reached if and only if the number of steps taken is odd.

As before, the summands with different k-tuples $\mathbf{i}$ and $\mathbf{j}$ might contribute the same weight to the sum if they correspond to the same walk on the double tree with a different labeling of the vertices. We pick the representative of such equivalent walks to be the labeling corresponding to the order of appearance in the walk, i.e., i_1 is vertex 1, j_1 is vertex 2, etc. The question then is: how k-tuples $\mathbf{i}$, $\mathbf{j}$ are equivalent to this walk up to labeling? Let $r+1$ be the number of distinct $\mathbf{i}$-vertices visited. Note that $0 \le r \le k-1$. We then need to choose $r+1$ distinct labels for the $\mathbf{i}$-vertices from $\{1,\ldots,p\}$ and $k-r$ distinct labels for the $\mathbf{j}$-vertices from $\{1,\ldots,n\}$. There are

$$p(p-1)\ldots(p-r) \times n(n-1)\ldots(n-k+r+1) \tag{6.10}$$

ways of doing this. Since $p \sim \gamma n$, then as $n \to \infty$, we have

$$p(p-1)\ldots(p-r)n(n-1)\ldots(n-k+r+1) = pn^k\gamma^r(1+O(1/n)), \tag{6.11}$$

where $f(x) = O(g(x))$, with $g(x)$ strictly positive for sufficiently large values of x, means $|f(x)| \le Cg(x)$ for some constant $C > 0$ and for all sufficiently large x.

Hence, if we denote by $\mathcal{G}(r,k)$ the collection walks on double trees (with vertices labeled by order of appearance) with exactly $r+1$ $\mathbf{i}$-vertices and $k-r$ $\mathbf{j}$-vertices, we have

$$\mathbf{E}\left[\frac{1}{p}\sum_{i=1}^{p}\Lambda_i^k\right] = \sum_{r=0}^{k-1}\gamma^r \#\mathcal{G}(r,k)(1+O(1/n)). \tag{6.12}$$

In light of Lemma 6.3, it remains to show that

$$\#\mathcal{G}(r,k) = \frac{1}{r+1}\binom{k}{r}\binom{k-1}{r}. \tag{6.13}$$

This is done in the combinatorial lemma below and concludes the proof of Theorem 6.1.

Lemma 6.4.

$$\mathcal{G}(r,k) = \frac{1}{r+1}\binom{k}{r}\binom{k-1}{r} \tag{6.14}$$

Proof. This is done by establishing a bijection between $\mathcal{G}(r,k)$ with a specific collection of paths π with $2k$ steps, much as we did in the case of the semicircle law with Dyck paths. In this case the paths are

generated as follows. Let $\pi : \{0, 1, 2 \ldots, 2k\}$ with $\pi(0) = \pi(2k) = 0$. The increments are denoted by $s_m = \pi_m - \pi_{m-1}$ for $m = 1, \ldots, 2k$. We take the increments are defined as follows: $s_1 = 0$, then for for $j = 1, \ldots, k-1$

$$(s_{2j}, s_{2j+1}) = \begin{cases} (1, -1) \\ (0, 0) \\ (0, -1) \\ (1, 0). \end{cases}$$

In other words, the increment s_m is -1 or 0 if m is odd, and 0 or 1 if it is even. We restrict to paths π that stay positive, that is $\sum_{m=1}^{t} s_m \geq 0$ for all $1 \leq t \leq 2k$. Moreover, we consider paths π with exactly r increments $+1$ (and thus also r increments -1), $0 \leq r \leq k-1$. We denote this set of paths by $\mathcal{P}(r, k)$.

The bijection between $\mathcal{G}(r, k)$ and $\mathcal{P}(r, k)$ is given by the following rules. For m odd: if $s_m = 0$ go from the left vertex to an unvisited right vertex, and if $s_m = -1$ go back to the adjacent, previously visited j-vertex. For m odd: if $s_m = 1$, go from a right vertex to a new left vertex; if $s_m = 0$ go back to an adjacent, previously visited left vertex. (Working out some specific examples would help here.)

In the reverse direction, consider a path on the double tree as described above. Starting at a left (that is $\mathbf{i}$-vertex in the context above). After an even number $2m$ of edge crossings one sits at a left vertex. If this vertex is new, that is, if it is being visited for the first time, then set $s_{2m} = +1$, and otherwise set $s_{2m} = 0$. After an odd number $2m-1$ of edge crossings one must sit at a right or j-vertex. If this edge crossing is the last exit from the i-vertex in question, set $s_{2m-1} = -1$, and otherwise set $s_{2m-1} = 0$. We need to check that the paths constructed with these increments stay positive. We do this by establishing a contradiction. Let us suppose we have a first t such that $\sum_{m=1}^{2t-1} s_m = -1$. This requires $\sum_{m=1}^{2t-2} s_m = 0$ and $s_{2t-1} = -1$. Now, we could use the sequence up to $2t-2$ to construct a double tree, and having $s_{2t-1} = -1$ would then mean not establishing a new vertex, but instead going back to one previously created. This however, contradicts already having built a double tree.

Having established the bijection, we now compute $\#\mathcal{P}(r, k)$. Without the positivity constraint, the number of paths with the above increments with exactly r $+1$ increments is $\binom{k-1}{r}\binom{k}{r}$. We now subtract number of paths that become negative yet $\pi(0) = \pi(2k) = 0$. To count such paths, we use a variation on the *reflection principle*. For each of these paths, there must exist a first time $2t-1$ such that $\sum_{m=1}^{2t} s_m = -1$. Note that this must happen at an odd time. Therefore, by *reflecting* the increments (s_{2m}, s_{2m+1}) that are $(1, 0)$ by replacing them with $(0, -1)$ and vice-versa (leaving all other pairs of increments unchanged), we see that these paths are in bijection with paths that start at 0 but end at -2. There are $\binom{k-1}{r-1}\binom{k}{r+1}$ such paths (Why?). Hence, we showed that

$$\#\mathcal{P}(r, k) = \binom{k-1}{r}\binom{k}{r} - \binom{k-1}{r-1}\binom{k}{r+1} = \frac{1}{r+1}\binom{k}{r}\binom{k-1}{r}. \tag{6.15}$$

□

6.4 Applications

Many applications of mathematics involve matrices that are rectangular, not square. In these cases one is often interested in the singular values. The Marchenko-Pastur law describes the distribution of singular values when the matrix entries are taken at random.

In data science, we have already seen that the empirical correlations between data points can be represented as a matrix. These can be analysed through the spectrum of the sample covariance matrix. The Marchenko-Pastur law gives a *null hypothesis* against which to test the spectrum, enabling one to identify the directions (eigenvectors) associated with statistically significant correlations, in the same way that one might test one-dimensional data against a normal distribution.

An example of this is in Principal Component Analysis, where one wants to identify the effective dimensionality of the subspace representing the main variation in the data. A rule of thumb for how many of the largest eigenvalues should be retained is to focus on those lying outside the support of the

Marchenko-Pastur distribution, or those near the edge if these is an unexpected density of them, as these are most likely to represent non-random features.

For an overview covering a wide range of applications see [Joh07], and for specific applications to mathematical finance (where this is as an important technique), see [PB20, LCPB00].

Finally, in many algorithms in numerical linear algebra, the rate of convergence is determined by the largest singular values, and the Marchenko-Pastur law determines where these are expected to lie (at the edge of the support).

7 The Stieltjes Transform

In this section, we present another approach to prove limit theorems for the spectral measure of a random matrix, the *Stieltjes transform*. The method of moments has some drawbacks. For example, to use it, moments of the spectral measure must obviously exist, which might be a problem when dealing with entries whose distributions are heavy-tailed. The Stieltjes transform is much more powerful and has been applied in different settings, see for example [BAG08, BGK18].

7.1 A quick introduction to the Stieltjes transform

At this stage, you have probably encountered different objects that encode the information of a probability distribution on $\mathbb{R}$. There are for example the cumulative distribution function, the characteristic function (which is the Fourier transform), the moment generating function (if it exists), etc. The Stieltjes transform is yet another way to do this.

Definition 7.1. *Let μ be a probability measure on $\mathbb{R}$. The Stieltjes transform of μ is the function from $\mathbb{C} \setminus \mathbb{R}$ defined by*

$$g_\mu(z) = \int_{\mathbb{R}} \frac{1}{x - z} \mathrm{d}\mu(x). \tag{7.1}$$

In some areas of mathematics and physics, $-g_\mu(z)$ is referred to as the *Green function*. We have the following basic properties.

Lemma 7.2. *Let μ be a probability measure on $\mathbb{R}$ and g_μ its Stieltjes transform. We have*

1. *For any $z \in \mathbb{C} \setminus \mathbb{R}$, $|g_\mu(z)| \leq 1/|\mathrm{Im} z|$.*
2. *More generally, for any $z \in \mathbb{C} \setminus \mathbb{R}$ and integers $j \geq 0$, we have*

$$\left| \frac{\mathrm{d}^j g_\mu}{\mathrm{d}z^j}(z) \right| \leq \frac{C_j}{|\mathrm{Im} z|^{j+1}}, \tag{7.2}$$

   for some constant $C_j > 0$ depending on j.

3. *The function $z \mapsto g_\mu(z)$ maps the upper half plane $\{z : \mathrm{Im} z > 0\}$ to itself, and is analytic on that domain.*
4. *Suppose that the support of μ is bounded. Then for sufficiently large z, we have*

$$g_\mu(z) = -\frac{1}{z} \sum_{k=0}^{\infty} \frac{m_k}{z^k}, \tag{7.3}$$

   where m_k is the k-th moment of μ. Thus, $g_\mu(z)$ can be thought of as a moment generating function of μ.

Proof. Let's write $z = u + \mathrm{i}v$. The Stieltjes transform at z is then

$$g_\mu(z) = \int \frac{1}{x - \mu - \mathrm{i}v} \mathrm{d}\mu(x).$$

The modulus of the integrand is clearly smaller then $1/|\mathrm{Im} z|$. This proves 1). In particular, this proves also that the integral is well-defined for any $z \notin \mathbb{R}$. By linearity of the integral, we get

$$g_\mu(z) = \int \frac{x - u}{(x - u)^2 + v^2} \mathrm{d}\mu(x) + \mathrm{i} \int \frac{v}{(x - u)^2 + v^2} \mathrm{d}\mu(x).$$

It is clear that $\mathrm{Im} z > 0$ implies $\mathrm{Im} g_\mu(z) > 0$. We now prove 3). Note that this implies analyticity for z large enough in the case where μ has bounded support. The general case and part 2) are left as an

exercise. Let X be a random variable that is μ-distributed. In particular, we must have $|X| \leq C$ for some $C > 1$ with probability 1. We have for some fixed z with $|z| > C$,

$$g_\mu(z) = \mathbf{E}\left[\frac{1}{X-z}\right] = \frac{-1}{z}\mathbf{E}\left[\sum_{k=0}^{\infty}\frac{X^k}{z^k}\right],$$

by expanding $(1-x)^{-1}$. We can now apply the dominated convergence theorem (Theorem B.7) to get

$$g_\mu(z) = \mathbf{E}\left[\frac{1}{X-z}\right] = \frac{-1}{z}\mathbf{E}\left[\sum_{k=0}^{\infty}\frac{X^k}{z^k}\right] = \frac{-1}{z}\sum_{k=0}^{\infty}\frac{m^k}{z^k}.$$

□

Note that part 3) of the lemma ensures that the Stieltjes transform determines the distribution whenever the distribution has bounded support. This is also true in the general case:

Proposition 7.3. *Let μ be a probability measure on $\mathbb{R}$. Then, μ is determined by its Stieltjes transform g_μ. In fact, for any interval $[a,b]$ with $\mu(\{a\}) = \mu(\{b\}) = 0$, we have*

$$\mu([a,b]) = \lim_{\eta\to 0}\int_a^b \frac{1}{\pi}\mathrm{Im} g_\mu(x+\mathrm{i}\eta)\mathrm{d}x. \tag{7.4}$$

Proof. It suffices to prove the inversion formula. It follows from interchanging the order of integration in

$$\int_a^b \frac{1}{\pi}\mathrm{Im} g_\mu(x+\mathrm{i}\eta)\mathrm{d}x = \int_a^b \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\eta}{(y-x)^2+\eta^2}\mathrm{d}\mu(y)\mathrm{d}x. \tag{7.5}$$

This gives

$$\int_a^b \frac{1}{\pi}\mathrm{Im} g_\mu(x+\mathrm{i}\eta)dx = \int_{-\infty}^{\infty} V(y)\mathrm{d}\mu(y), \tag{7.6}$$

where

$$V(y) = \frac{1}{\pi}\left[\arctan\left(\frac{b-y}{\eta}\right) - \arctan\left(\frac{a-y}{\eta}\right)\right] \tag{7.7}$$

As $\eta \to 0$, we have $V(y) \to 1$ for $y \in (a,b)$, and $V(a), V(b) \to 1/2$. If $y \in [a,b]^c$, then $V(y) \to 0$. Since $V(y)$ is uniformly bounded and, for any η, $V(y) \sim \eta/y^2$ as $y \to \infty$, it can be majorized uniformly in η by a positive integrable function. Hence it follows from the dominated convergence theorem that

$$\int_{-\infty}^{\infty} V(y)\mathrm{d}\mu(y) \to \mu([a,b]), \text{ as } \eta \to 0. \tag{7.8}$$

□

7.2 The Stieltjes transform and the semicircle law

The Stieltjes transform provides an alternative route to proving both the Wigner semicircle law and the Marchenko-Pastur law that avoids the combinatorics and path-counting of the proofs described earlier in these notes. We sketch here the proof for GOE matrices. The proof can be extended to Wigner random matrices in general, but this is an important class of examples and specialising will allow us to introduce new techniques that provide some shortcuts.

We start by computing the Stieltjes transform of the semicircle distribution.

Lemma 7.4. *Let σ be the semicircle distribution given in* (4.2)*, then we have*

$$g_\sigma(z) = \frac{-z+\sqrt{z^2-4}}{2}, \quad \mathrm{Im} z > 0. \tag{7.9}$$

Proof. A direct computation gives

$$g_\sigma(z) = \frac{1}{2\pi}\int_{-2}^{2} \frac{1}{x-z}\sqrt{4-x^2}\mathrm{d}x = \frac{1}{\pi}\int_0^\pi \frac{2}{2\cos y - z}(\sin y)^2 \mathrm{d}y = \frac{1}{2\pi}\int_0^{2\pi} \frac{2}{2\cos y - z}(\sin y)^2\mathrm{d}y. \tag{7.10}$$

Setting $w = e^{\mathrm{i}y}$ (and so $w^{-1} = e^{-iy}$), this is

$$g_\sigma(z) = -\frac{1}{4\pi i}\oint_{|w|=1} \frac{(w^2-1)^2}{w^2(w^2+1-zw)}\mathrm{d}w, \tag{7.11}$$

The integrand here has poles at $w = 0, (z \pm \sqrt{z^2-4})/2$ (but only two lie inside the unit circle). Evaluating the integral using Cauchy's theorem yields

$$g_\sigma(z) = \frac{-z+\sqrt{z^2-4}}{2} \tag{7.12}$$

where the branch cut associated with the square root is chosen so that $\mathrm{Im} g_\sigma(z) > 0$ when $\mathrm{Im} z > 0$. □

Observe that $g_\sigma(z)$ satisfies the quadratic functional equation

$$g_\sigma(z) = -\frac{1}{z} - \frac{1}{z}g_\sigma(z)^2, \tag{7.13}$$

with (7.12) being the solution satisfying $\mathrm{Im} g_\sigma(z) > 0$ when $\mathrm{Im} z > 0$. We remark in passing that for the Marchenko-Pastur distribution, the formula analogous to (7.13) is

$$\gamma z g(z)^2 + (z - 1 + \gamma)g(z) + 1 = 0, \tag{7.14}$$

and the Stieltjes transform of the Marchenko-Pastur distribution is the solution of this equation satisfying $\mathrm{Im} g(z) > 0$ when $\mathrm{Im} z > 0$.

The goal of the section is to prove:

Theorem 7.5. *Let $\mathbf{M}$ be a $n \times n$ GOE matrix. Let μ_n be the empirical spectral measure of $\mathbf{M}/\sqrt{n}$. Then for any $z \in \mathbb{C}$ with $\mathrm{Im} z > 0$, we have*

$$\lim_{n\to\infty} g_{\mu_n}(z) = g_\sigma(z) \text{ in expectation and almost surely.}$$

We start by making some important observations. Let M be some $n \times n$ matrix (not necessarily random) with eigenvalues $\lambda_1, \ldots, \lambda_n$. The Stieltjes transform of its empirical spectral measure (ESM) μ_M is given by

$$g_{\mu_M}(z) = \frac{1}{n}\sum_j \frac{1}{\lambda_j - z} = \frac{1}{n}\mathrm{Tr}\left(\frac{1}{M - z\mathrm{I}}\right) \tag{7.15}$$

This is a consequence of the spectral theorem. The matrix $G_M(z) = (M - z\mathrm{I})^{-1}$ is called the *resolvent*. If M has real eigenvalues, then the operator is well-defined for any $z \notin \mathbb{R}$. The resolvent is an important tool in the study of Markov processes, and more generally of unbounded and bounded operators. From the definition, we get:

Lemma 7.6. *Let M and A be $n \times n$ matrices such that M and $M + A$ have real eigenvalues. Then for any $z \notin \mathbb{R}$, we have*

$$G_{M+A}(z) - G_M(z) = -G_{M+A}(z)AG_M(z). \tag{7.16}$$

In particular, for $A = -M$, we get

$$G_M(z) = -\frac{1}{z}\mathrm{I} + \frac{1}{z}MG_M(z). \tag{7.17}$$

Moreover, the derivative of the entry $(G_M)_{kl}$ with respect to M_{ij} is

$$\frac{\partial (G_M)_{kl}}{\partial M_{ij}} = -(G_M)_{ki}(G_M)_{jl} - (G_M)_{kj}(G_M)_{il}. \tag{7.18}$$

Proof. The identity (7.16) is proved by direct computation, and (7.17) by substitution. The last relation is obtained by taking $A_{ij} = A_{ji} = \epsilon$ and $A_{uv} = 0$ otherwise, and then taking the limit as $\epsilon \to 0$. $\square$

Proof of Theorem 7.5. Let μ_n be the ESM of the matrix $\mathbf{M}/\sqrt{n}$ where $\mathbf{M}$ is a $n \times n$ GOE matrix. We write $G_n(z) = (\mathbf{M}/\sqrt{n} - z\mathrm{I})^{-1}$ for the resolvent of $M/\sqrt{n}$. We compute the expectation of $g_{\mu_n}(z)$ using (7.15) and(7.17):

$$\begin{aligned} g_{\mu_n}(z) = \frac{1}{n}\mathbf{E}[\mathrm{Tr}G_n(z)] &= -\frac{1}{z} + \frac{1}{n^{3/2}z}\mathbf{E}\left[\mathrm{Tr}\big(\mathbf{M}G_n(z)\big)\right] \\ &= -\frac{1}{z} + \frac{1}{n^{3/2}z}\sum_{i,j}\mathbf{E}\left[\mathbf{M}_{ij}\big(G_n(z)\big)_{ji}\right]. \end{aligned} \tag{7.19}$$

We now use the fact that the entries of $\mathbf{M}$ are Gaussian. By Lemma B.2 (Gaussian integration by parts), the above equals

$$-\frac{1}{z} + \frac{1}{n^{3/2}z}\sum_{i,j}\mathbf{E}\left[\frac{\partial(G_n)_{ji}}{\partial M_{ij}}\right].$$

(Here the derivative is with respect to the the ij-entry of M in the map $G_n(z) = (M/\sqrt{n} - z\mathrm{I})^{-1}$.) The derivative is evaluated using (7.17). We finally get

$$\begin{aligned} \mathbf{E}[g_{\mu_n}(z)] &= -\frac{1}{z} + \frac{1}{n^2 z}\sum_{i,j}\mathbf{E}\left[-(G_n)_{ji}(G_n)_{ij} - (G_n)_{jj}(G_n)_{ii}\right] \\ &= -\frac{1}{z} - \frac{1}{n^2 z}\mathbf{E}\left[\mathrm{Tr}\big(G_n^2\big)\right] - \frac{1}{z}\mathbf{E}\left[\Big(\frac{1}{n}\mathrm{Tr}G_n\Big)^2\right]. \end{aligned}$$

The second term is negligible in the limit. Indeed, we have by the spectral theorem

$$\left|\frac{1}{n^2 z}\mathrm{Tr}(G_n(z)^2)\right| = \frac{1}{n|z|}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|\Lambda_i - z|^2}\right| \le \frac{1}{n|z||\mathrm{Im}z|^2} \to 0,$$

when $n \to \infty$. Hence we have that

$$\mathbf{E}[g_n(z)] = -\frac{1}{z} - \frac{1}{z}\mathbf{E}\left[g_n(z)^2\right] + \mathcal{E}(z,n) \tag{7.20}$$

where for any z, $\mathcal{E}(z,n) \to 0$ as $n \to \infty$.

We now want to establish that we can replace $\mathbf{E}\left[g_n(z)^2\right]$ with $(\mathbf{E}[g_n(z)])^2$, with an error that tends to zero as $n \to \infty$. This is justified by the concentration of the Stieltjes transform around its mean:

Lemma 7.7. *Let $g_n(z)$ be the Stieltjes transform of $M/\sqrt{n}$ where M is a $n \times n$ GOE matrix. Then for any $\kappa > 0$, we have*

$$\mathbf{P}\Big(\big|g_n(z) - \mathbf{E}[g_n(z)]\big| \ge \kappa/\sqrt{n}\Big) \le C\exp(-c\kappa^2), \tag{7.21}$$

for some absolute constants $c, C > 0$.

The lemma is proved at the end of the section. Note that this implies (exercise)

$$\mathbf{E}\left[g_n(z)^2\right] = \left(\mathbf{E}[g_n(z)]\right)^2 + \mathcal{E}'(z,n) \tag{7.22}$$

where $\mathcal{E}'(z,n) \to 0$ as $n \to \infty$. Using this in (7.20), we get

$$\mathbf{E}g_n(z) = -\frac{1}{z} - \frac{1}{z}\left(\mathbf{E}g_n(z)\right)^2 + \mathcal{E}''(z,n) \tag{7.23}$$

where $\mathcal{E}''(z,n) \to 0$ as $n \to \infty$. Hence, the limit of $\mathbf{E}g_n(z)$ of the Stieltjes measure of the ESM of $M/\sqrt{n}$ (if it exists) must be the solution of the self-consistency equation

$$s(z) = -\frac{1}{z} - \frac{1}{z}s(z)^2. \tag{7.24}$$

The solution is the Stieltjes transform of the semicircle distribution, as seen in (7.13), once one specifies that $\mathrm{Im}s(z) > 0$ when $\mathrm{Im}z > 0$.

The convergence of$\mathbf{E}g_n(z)$ is established by a compactness argument. Equation (7.2) implies that the sequence of functions $(\mathbf{E}g_n, n \geq 1)$ is uniformly bounded and uniformly equicontinuous. The Arzelá-Ascoli theorem then implies that there is a subsequence $(\mathbf{E}g_{n_k}, k \geq 1)$ that converges uniformly on compact sets of the upper-half plane. The limit must then satisfy (7.24) and must be the Stieltjes transform of the semicircle distribution. Since the limit is unique, we conclude that $(\mathbf{E}g_n, n \geq 1)$ converges uniformly on compact sets, hence also pointwise, to $s(z)$.

To see that the convergence holds also for $(g_n(z), n \geq 1)$ almost surely, take $\kappa = n^{1/3}$ in (7.28) and apply the Borel-Cantelli lemma (Lemma B.8). We then see that

$$\left| g_n(z) - \mathbf{E}[g_n(z)] \right| = O(n^{-1/6}) \tag{7.25}$$

for all large n. Therefore $g_n(z) - \mathbf{E}[g_n(z)] \to 0$ almost surely for all z with $\mathrm{Im}z > 0$. This proves convergence almost surely of the ESM to the semicircle distribution. □

Proof of Lemma 7.7. For the proof, we make the dependence on n explicit in the dimension of the matrix $\mathbf{M}$, and write $\mathbf{M}_n$ for $\mathbf{M}$. The strategy will be to show that replacing $\mathbf{M}_n$ by one of its $(n-1) \times (n-1)$ minor, denoted by $\mathcal{M}_{n-1}$, does not change the Stieltjes transform appreciably when n is large – i.e. that $g_{\mu_{\mathcal{M}_{n-1}/\sqrt{n}}}(z)$ is close to $g_{\mu_{\mathbf{M}_n/\sqrt{n}}}(z)$. Without loss of generality, we take the top left minor. This will allow us to use an estimate to establish that $g_{\mu_{\mathbf{M}_n/\sqrt{n}}}(z)$ lies sufficiently close to $\mathbf{E}\left(g_{\mu_{\mathbf{M}_n}}(z)\right)$ for us to prove what we need.

Let $z = a + \mathrm{i}b$, with $b > 0$. We denote the eigenvalues of $\mathbf{M}_n/\sqrt{n}$ by $\{\lambda_j^{(n)}\}_{j=1}^n$ and the eigenvalues of $\mathcal{M}_{n-1}/\sqrt{n}$ by $\{\lambda_j^{(n-1)}\}_{j=1}^{n-1}$. Now, Cauchy's Interlacement Theorem (Theorem A.2) implies that the eigenvalues of $\mathbf{M}_n/\sqrt{n}$ and those of $\mathcal{M}_{n-1}/\sqrt{n}$ interlace, that is

$$\lambda_j^{(n)} \leq \lambda_j^{(n-1)} \leq \lambda_{j+1}^{(n)} \text{ for } j = 1, \ldots, n-1.$$

Hence the difference

$$\sum_{j=1}^{n-1} \frac{b}{(\lambda_j^{(n-1)} - a)^2 + b^2} - \sum_{j=1}^{n} \frac{b}{(\lambda_j^{(n)} - a)^2 + b^2} \tag{7.26}$$

is bounded in n. This is because $\frac{b}{(x-a)^2+b^2}$ has *finite total variation* and $\lambda_1^{(n)}, \lambda_1^{(n-1)}, \lambda_2^{(n)}, \lambda_2^{(n-1)}, \ldots, \lambda_{n-1}^{(n-1)}, \lambda_n^{(n)}$ forms a partition of $\mathbb{R}$. The same conclusion holds if one replaces the numerators in the summands with $\lambda_j^{(n-1)} - a$ and $\lambda_j^{(n)} - a$ respectively. It follows that

$$g_{\mu_{\mathbf{M}_n/\sqrt{n}}}(z) = g_{\mu_{\mathcal{M}_{n-1}/\sqrt{n}}}(z) + O\left(\frac{1}{n}\right). \tag{7.27}$$

Note as well that the right-hand side of this equation depends only on the top left $(n-1) \times (n-1)$ minor of $\mathbf{M}_n$ and is independent of its nth row and nth column. We can therefore apply the concentration inequality (Theorem C.2) in the following way. First note that Equation (7.27) still applies if we resample the nth row and nth column. Denoting the Stieltjes transform of the resampled matrix by $g_{\mu_{\mathbf{M}'_n\sqrt{n}}}(z)$, it follows from (7.27) that $g_{\mu_{\mathbf{M}_n}/\sqrt{n}}(z) = g_{\mu_{\mathbf{M}'_n}/\sqrt{n}}(z) + O(1/n)$. Moreover, this applies to resampling any row and column with the same label. Therefore, applying McDiarmid's inequality with x_j being the vector of entries $(\mathbf{M}_{jj}, \mathbf{M}_{j,j+1}, \ldots, \mathbf{M}_{jn})$ gives

$$\mathbf{P}(|g_{\mu_{\mathbf{M}_n}}(z) - \mathbf{E}[g_{\mu_{\mathbf{M}_n}}(z)]| \geq \kappa/\sqrt{n}) \leq C\exp(-c\kappa^2), \textit{asclaimed.} \tag{7.28}$$

□

7.3 The $\mathcal{R}$-transform

As noted previously, $-g_\mu(z)$ is sometimes referred to as the Green function. The function $B_\mu(z)$ that is the formal inverse of the Green function, i.e., the function which satisfies

$$B_\mu(-g_\mu(z)) = z, \tag{7.29}$$

is sometimes referred to as the *Blue function*.

The $\mathcal{R}$-transform of μ is then defined to be

$$R_\mu(z) = B_\mu(z) - \frac{1}{z}. \tag{7.30}$$

The $\mathcal{R}$-transform of the semicircle distribution is easily seen (e.g. from the quadratic equation satisfied by the Stieltjes transform of the semicircle distribution) to be

$$R_{sc}(z) = z \tag{7.31}$$

and similarly for the Marchenko-Pastur distribution

$$R_{MP}(z) = \frac{1}{1-\gamma z}. \tag{7.32}$$

The fact that the $\mathcal{R}$-transforms of the semicircle distribution and the Marchenko-Pastur distributions take such elementary forms is in indication that mathematically they are rather natural and simple, in a way that may be obscured by other representations. Indeed, the semicircle distribution can be viewed as playing a similar role to that played by the normal distribution for commuting random variables.

The $\mathcal{R}$-transform plays an important role in analysing the spectral density of sums of random matrices which satisfy certain natural conditions, but we will not examine this further here.

8 Girko's Circular Law

The semicircle law holds for either real symmetric or complex Hermitian matrices, and in both cases the eigenvalues are real. It is natural to seek to extend it to non-symmetric matrices, when the eigenvalues are complex numbers. This turns out to be harder, so we simply state the result.

Theorem 8.1 (Girko's Circular Law)**.** *Let* $\mathbf{M}$ *be a* $n \times n$ *matrix with IID entries such that* $\mathbf{E}[\mathbf{M}_{ij}] = 0$ *and* $\mathbf{E}[|\mathbf{M}_{ij}|^2] = 1$*. Consider the empirical spectral measure* μ_n *of* $\mathbf{M}/\sqrt{n}$ *given by*

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{\Lambda_i/\sqrt{n}}.$$

(Note that this is now a probability measure on $\mathbb{C}$*.) Then* μ_n *converges almost surely to the uniform probability measure on the unit disc* $\{z \in \mathbb{C} : |z| \leq 1\}$*.*

The proof of this is much more involved than for the case of Wigner matrices. We refer to [Tao12] for a proof. One reason for this is that the method of moments fails for complex probability measures. (This is essentially because polynomials are no longer dense in the space of bounded continuous function in $\mathbb{C}$.) One can still work with the Stieltjes measure but it is more complicated since z must avoid the spectrum, and this spectrum can be *a priori* anywhere in the complex plane. Last but not least, the spectrum in the non-Hermitian case is very sensitive to perturbation of an entries and therefore any truncation procedure of the entries is bound to failure. [12]

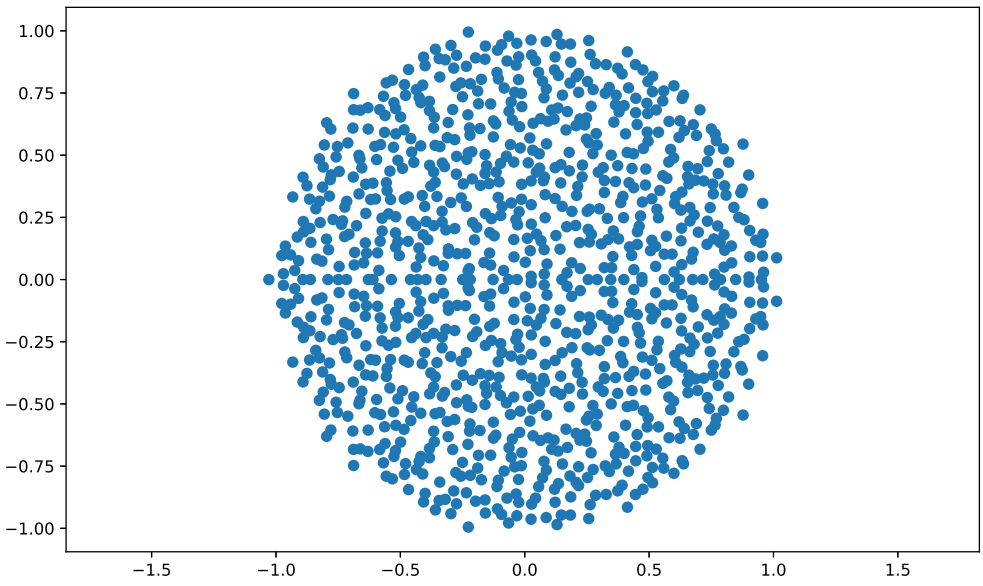

Figure 7: The empirical spectral measure of a $n \times n$ matrix (rescaled by $1/\sqrt{n}$) with IID standard normal entries for $n = 1000$.

[12] A very nice discussion on the circular law and its proof can be found here `https://terrytao.wordpress.com/2010/03/14/254a-notes-8-the-circular-law/`.

Part II
Eigenvalue Statistics

9 A Microscopic Point of View

In this section, we take a different point of view from Part I. The empirical spectral measure of a random matrix captures the spectrum *macroscopically*. In particular, we have seen that Wigner matrices satisfy the semicircle law, so that if we take the eigenvalues of a GOE or a GUE matrix and scale them by $1/\sqrt{n}$ then the support of the spectrum is almost surely $[-2, 2]$ in the limit $n \to \infty$ with a density in that interval given by the semicircle distribution. The question we wish to address now is: how are the eigenvalues distributed? We now zoom in on the spectrum and try to understand how the eigenvalues are distributed *microscopically*, i.e., on the scale of their mean spacing. This is a much smaller scale than that for which the semicircle law holds.

More specifically, the main object is now the unnormalized spectral measure, that is

$$\nu_n = \sum_{j=1}^{n} \delta_{\Lambda_j}, \tag{9.1}$$

where Λ_j are the eigenvalues of a $n \times n$ random matrix $\mathbf{M}$ with spectrum in $\mathbb{R}$. Now, μ_n is a *measure* but not a probability measure, since $\nu_n(\mathbb{R}) = n$. For a subset $A \subseteq \mathbb{R}$, we have that $\nu_n(A)$ counts the number of eigenvalues of $\mathbf{M}$ in A:

$$\nu_n(A) = \#\{i \leq n : \Lambda_i \in A\}. \tag{9.2}$$

We can see ν_n as a random counting measure. Another equivalent point of view is to think of ν_n as a random collection of points in $\mathbb{R}$. Then, we can view $\{\Lambda_j\}_{j \leq n}$ as a *point process* in $\mathbb{R}$. We are interested in the asymptotic $n \to \infty$. The eigenvalues might have to be rescaled to get some sensible limits for different observables, such as the distance between eigenvalues, the position of the largest eigenvalues, etc. We will see that the scaling will depend on where in the spectrum we zoom in.

The main point of comparison to the eigenvalue process will be the *Poisson point process*. The Poisson point process are obtained when points are thrown on the line independently. We summarize its properties in Section 9.1. We shall see that the eigenvalue process is in fact very different from the Poisson process. This is due to the important fact that the eigenvalues are strongly correlated in an interesting and significant way. This is an important feature of random matrices. We will focus mainly on matrices sampled from the GOE, the GUE, and the CUE.

9.1 Poisson point processes

You might have encountered the simplest Poisson process defined as follows.

Definition 9.1. *A Poisson point process N on $\mathbb{R}$ with intensity $\lambda > 0$ corresponds to a set of random points such that the distribution of the number of points in an interval $(a, b]$, denoted $N(a, b)$, is given by*

$$\mathbf{P}(N(a, b) = k) = \frac{(\lambda(b-a))^k}{k!} e^{-\lambda(b-a)}, \quad k = 0, 1, 2, \ldots \tag{9.3}$$

Moreover, the numbers of points in any finite collection of disjoint intervals are independent of each other.

Note that mean number of points in an interval $(a, b]$ is $\lambda(b - a)$, so it is proportional to the length of the interval. More generally, for any measurable set A, $\mathbf{E}[N(A)]$ is λ times the Lebesgue measure of A. This mean measure is sometimes called the intensity measure. It is then easy to generalize the above

definition by replacing the intensity measure by any other measure on $\mathbb{R}$. The examples where the intensity measure is a multiple of the Lebesgue measure corresponds exactly to the Poisson processes that are *stationary*, that is whose distribution is invariant under translation.

It is not hard to compute the statistics of the Poisson processes. We do two examples that serve later as a reference for the eigenvalue process.

Example 9.2. We compute the distribution of the spacing between two nearest-neighbouring points of a Poisson process N with intensity 1. Note that since the process is stationary, this will be independent on the location of the real line. Let Δ be spacing between two nearest-neighbouring points of the process. By stationarity, we can without loss of generality conditioned on the event that there is a point at 0. Therefore, we have that

$$\mathbf{P}(\Delta > x) = \lim_{\varepsilon \to 0} \frac{\mathbf{P}\Big(N((\varepsilon, \varepsilon + x]) = 0, N([0, \varepsilon]) = 1\Big)}{\mathbf{P}(N\Big([0, \varepsilon]) = 1\Big)} = e^{-x}$$

Therefore, the spacing between points is exponentially distributed with parameter 1 and the mean spacing is 1. (When the intensity is $\lambda > 0$, then the mean spacing is $1/\lambda$.) In particular, the PDF of the spacing is maximal at $x = 0$.

9.2 Joint eigenvalue density

We have seen earlier in (2.3) and (2.5) that the joint PDF of the entries of GOE and GUE, up to a normalization constant, are of the form

$$e^{-\frac{\beta}{4}\mathrm{Tr}(M^2)}, \tag{9.4}$$

where for the GOE $\beta = 1$, for the GUE $\beta = 2$. Of course, the eigenvalues are a function of the entries. Therefore, we should be able to extract the joint PDF of the eigenvalues $\Lambda_1, \ldots, \Lambda_n$ from the above. At this point, the ordering does not reflect the ordering of the eigenvalues. And thus the joint PDF will be invariant under permutations of its arguments.

Theorem 9.3. *Let $(\Lambda_j, j \leq n)$ denote the eigenvalues of the a $n \times n$ GOE or GUE matrix. Then their (symmetric) joint PDF is*

$$\rho_n(\lambda_1, \ldots, \lambda_n) = c_n^{(\beta)} \prod_{i<j} |\lambda_i - \lambda_j|^\beta e^{-\frac{\beta}{4}\sum_i \lambda_i^2}. \tag{9.5}$$

In other words, for any symmetric function $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\mathbf{E}[f(\Lambda_1, \ldots, \Lambda_n)] = \int_{\mathbb{R}^n} f(\lambda_1, \ldots, \lambda_n)\rho_n(\lambda_1, \ldots, \lambda_n)\mathrm{d}\lambda_1 \ldots \lambda_n.$$

Here $\beta = 1$ stands for the GOE, $\beta = 2$ for the GUE, and $c_n^{(1)}$ and $c_n^{(2)}$ are normalization constants.

We remark that the normalization constants can be evaluated to be:

$$\frac{1}{c_n^{(\beta)}} = (2\pi)^{n/2}\beta^{-n/2-\beta n(n-1)/4}[\Gamma(1+\beta/2)]^{-n}\prod_{j=1}^{n}\Gamma(1+\beta j/2). \tag{9.6}$$

We will not make explicit use of this fact and do not include the proof here.

The idea behind the proof of Theorem 9.3 is the spectral theorem (Theorem A.1). Any Hermitian matrix M can be diagonalized by a unitary transformation; that is, we can write $M = UDU^\dagger$, where D is diagonal and U is unitary. The idea is then to change variables from the matrix elements of M to the matrix elements of U and D. Because the GUE measure is invariant under unitary transformations, the matrix elements of U can be integrated out trivially, leaving just the matrix elements of D, which are the eigenvalues.

Proof. We first note that the eigenvalues of $\mathbf{M}$ are distinct with probability 1. To see this, observe that the joint distribution of the entries of $\mathbf{M}$ has a smooth density. Since the eigenvalues are continuous functions of the matrix entries, and because the event of two eigenvalues coinciding has codimension ≥ 1 in the n-dimensional variety of eigenvalues, it follows that the eigenvalues are almost surely distinct [13].

Now a Hermitian matrix M is obviously determined by its entries. Taking account the symmetries, it can be parametrized by

$$M = (x_{11}, \ldots, x_{nn}; x_{12}, y_{12}, \ldots, x_{n-1,n}, y_{n-1,n}) \in \mathbb{R}^{n^2}, \tag{9.7}$$

where x_{ij} stands for the real part of M_{ij} and y_{ij} stands for the imaginary part. Note that there are n^2 real parameters in this representation. If the eigenvalues of a Hermitian matrix M are distinct, then the decomposition $M = UDU^\dagger$ is unique up to (i) permuting the eigenvalues, and (ii) multiplying U by a diagonal matrix with entries $e^{\mathrm{i}\theta_1}, e^{\mathrm{i}\theta_2}, \ldots, e^{\mathrm{i}\theta_n}$. This means that we can also represent M in terms of the following coordinates:

$$(\lambda_1, \ldots, \lambda_n; w_1, \ldots, w_{n^2-n}) \in (\mathbb{R}^n/\mathcal{S}_n) \times (U(n)/\mathbb{T}^n).$$

The λ's are the eigenvalues (now ordered) and the w's are the real parameters coming from the eigenvectors. One amazing fact from the proof is that we will not need details about the w's. Our goal is to compute the determinant of the Jacobian matrix of the map expressing the entries in terms of the λ's and the w's. We denote this $n^2 \times n^2$ matrix by $\left(\frac{\partial M}{\partial \lambda_\alpha}, \frac{\partial M}{\partial w_\beta}\right)$.

Recall that for any fixed unitary matrix V, the conjugation by V, that is the mapping $M \mapsto V^\dagger M V$, is an isometry, see (2.6). We write $\tau_V : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ for the transformation induced on the entries by this conjugation. This transformation is linear (double-check!), and hence can be represented as a $n^2 \times n^2$ matrix. Moreover, since it is an isometry, we must have $\det \tau_V = 1$. This implies that

$$\det\left(\frac{\partial M}{\partial \lambda_\alpha}, \frac{\partial M}{\partial w_\beta}\right) = \det\left(\tau_V\left(\frac{\partial M}{\partial \lambda_\alpha}, \frac{\partial M}{\partial w_\beta}\right)\right), \tag{9.8}$$

for any unitary matrix V. Here comes the magic. For the unitary U that diagonalizes M, we have that

$$\tau_U\left(\frac{\partial M}{\partial \lambda_\alpha}\right) = \tau_U\left(U \frac{\partial D}{\partial \lambda_\alpha} U^\dagger\right) = \frac{\partial D}{\partial \lambda_\alpha}. \tag{9.9}$$

Note that the right-hand side gives a vector with all n^2 coordinates except the ith equal to zero (in the representation (9.7). We also have that

$$\tau_U\left(\frac{\partial M}{\partial w_\beta}\right) = \tau_U\left(\frac{\partial U}{\partial w_\beta} D U^\dagger + U D \frac{\partial U^\dagger}{\partial w_\beta}\right) = U^\dagger \frac{\partial U}{\partial w_\beta} D + D \frac{\partial U^\dagger}{\partial w_\beta} U. \tag{9.10}$$

Since $U^\dagger U = \mathrm{I}$, we have by differentiating with respect to w_β

$$U^\dagger \frac{\partial U}{\partial w_\beta} + \frac{\partial U^\dagger}{\partial w_\beta} U = 0. \tag{9.11}$$

Therefore, setting $T_\beta = U^\dagger \frac{\partial U}{\partial w_\beta}$, we get

$$\tau_U\left(\frac{\partial M}{\partial w_\beta}\right) = T_\beta D - D T_\beta. \tag{9.12}$$

[13] An alternative proof of this uses the fact that the zero set of any polynomial in k variables has zero Lebesgue measure in $\mathbb{R}^k$. There are degenerate eigenvalues if and only if the characteristic polynomial of M, $\det(M - x\mathrm{I})$, and its derivative have a common zero. A necessary and sufficient condition for this is that the discriminant $\prod_{i<j}(\lambda_i - \lambda_j)^2$ vanishes. The result then follows from the fact that discriminant is itself a polynomial in the entries of M.

This shows that the matrix $\tau_U\left(\frac{\partial M}{\partial \lambda_\alpha}, \frac{\partial M}{\partial w_\beta}\right)$ has the following form

$$\begin{pmatrix} \mathrm{I}_n & 0_n & 0_n & 0_n & \dots \\ 0_n & \mathrm{Re}(T_1)_{12}(\lambda_2-\lambda_1) & \mathrm{Im}(T_1)_{12}(\lambda_2-\lambda_1) & \mathrm{Re}(T_1)_{13}(\lambda_3-\lambda_1) & \dots \\ 0_n & \mathrm{Re}(T_2)_{12}(\lambda_2-\lambda_1) & \mathrm{Im}(T_2)_{12}(\lambda_2-\lambda_1) & \mathrm{Re}(T_2)_{13}(\lambda_3-\lambda_1) & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}, \tag{9.13}$$

where 0_n is the $n \times n$ zero matrix. Hence, computing the determinant, we have that

$$\det\left(\frac{\partial M}{\partial \lambda_\alpha}, \frac{\partial M}{\partial w_\beta}\right) = \prod_{i<j}(\lambda_j-\lambda_i)^2 \det \begin{pmatrix} \mathrm{Re}(T_1)_{12} & \mathrm{Im}(T_1)_{12} & \dots & \\ \mathrm{Re}(T_2)_{12} & \mathrm{Im}(T_2)_{12} & \dots & \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}. \tag{9.14}$$

The problem has now factorized! We are left with a product of $\prod_{i<j}(\lambda_j-\lambda_i)^2$ and the determinant, which is a function only of the w-variables. Hence in computing $\mathbf{E}[f(\Lambda_1, \Lambda_2, \dots, \Lambda_n)]$, where the expectation is computed with respect to the GUE, we can integrate with respect to all of the w-variables, obtaining some constant, and we are left only with the integrals over $\lambda_1, \dots, \lambda_n$, picking up a factor $\prod_{i<j}(\lambda_j-\lambda_i)^2$. This proves the theorem for the GUE. □

The proof for the GOE follows an identical path, except that in this case the matrix M is diagonalized by an orthogonal transformation, so $M = ODO^{\mathrm{T}}$. The number of independent random variables in M (e.g. the matrix elements on and above the diagonal) is then $n(n+1)/2$. These can be replaced by the diagonal elements of D (the eigenvalues), $\lambda_1, \dots, \lambda_n$ and $n(n-1)/2$ other parameters $w_1, \dots, w_{n(n-1)/2}$. One again needs to compute the Jacobian of the transformation $\det\left(\frac{\partial M}{\partial \lambda_\alpha}, \frac{\partial M}{\partial w_\beta}\right)$. The calculation goes through essentially as above (but with orthogonal matrices replacing the unitary matrices), and one ends up with a Jabobian which is similar to (9.13) but missing the columns involving $\mathrm{Im}(T_\beta)$ because the orthogonal matrices can be taken to be real valued. Therefore each factor $(\lambda_j-\lambda_i)$ appears only once and the Jacobian in this case factorizes as the product of $\prod_{i<j}|\lambda_j-\lambda_i|$ and a function only of the w-variables. Hence again in computing $\mathbf{E}[f(\Lambda_1, \dots, \Lambda_n)]$, where the expectation is computed now with respect to the GOE, we can integrate with respect to all of the w-variables, obtaining some constant, and we are left only with the integrals over $\lambda_1, \dots, \lambda_n$, picking up a factor $\prod_{i<j}|\lambda_j-\lambda_i|$. This proves the theorem for the GOE.

We turn next to the CUE. Let $\mathbf{M}$ be a $n \times n$ CUE matrix. In this case, the eigenvalues lie on the unit circle thus we can denote them by $e^{\mathrm{i}\Theta_1}, \dots, e^{\mathrm{i}\Theta_n}$. The theorem is proved similarly. By the spectral theorem (Theorem A.1), any unitary matrix M is diagonalizable and can be written as $M = UDU^\dagger$, where D is diagonal with entries $e^{i\theta_1}, e^{i\theta_2}, \dots, e^{i\theta_n}$.

Theorem 9.4. *Let $(e^{\mathrm{i}\Theta_j}, j \le n)$ denote the eigenvalues of a $n \times n$ CUE matrix. Then the (symmetric) joint PDF is*

$$\rho_n(\theta_1, \dots, \theta_n) = \frac{1}{(2\pi)^n n!}\int_0^{2\pi}\cdots\int_0^{2\pi} f(\theta_1, \dots, \theta_n)\prod_{j<k}|e^{i\theta_j}-e^{i\theta_k}|^2 \mathrm{d}\theta_1 \dots \mathrm{d}\theta_n. \tag{9.15}$$

In other words, for any symmetric function $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\mathbf{E}[f(\Theta_1, \dots, \Theta_n)] = \int_{\mathbb{R}^n} f(\theta_1, \dots, \theta_n)\rho_n(\theta_1, \dots, \theta_n)\mathrm{d}\theta_1 \dots \mathrm{d}\theta_n.$$

This is known as the Weyl integration formula.

Finally, for the Wishart Ensemble, the corresponding result is[14]

[14]This is the formula for real Wishart matrices. For complex Wishart matrices the factor $\prod_{i<j}|\lambda_i-\lambda_j|$ is squared.

Theorem 9.5. *Let $\Lambda_1, \ldots, \Lambda_p$ denote the eigenvalues of a $p \times p$ Wishart matrix. For any symmetric function $f : \mathbb{R}^n \to \mathbb{R}$, we have*

$$\mathbf{E}[f(\Lambda_1, \ldots, \Lambda_p)] = c_{n,p}^{(\text{Wishart})} \int_0^\infty \cdots \int_0^\infty f(\lambda_1, \ldots, \lambda_p) \prod_{i<j} |\lambda_i - \lambda_j| \prod_k \lambda_k^{(n-p-1)/2} e^{-\frac{1}{2}\lambda_k} \mathrm{d}\lambda_1 \ldots \mathrm{d}\lambda_p, \tag{9.16}$$

where $c_{n,p}^{(\text{Wishart})}$ is a normalization constant.

One of the most important consequences of Theorem 9.3, Theorem 9.4 and Theorem 9.5 is that the eigenvalues of random matrices repel each other. It is clear from Theorem 9.3 that the probability that $\lambda_i - \lambda_j \in (s, s + \mathrm{d}s)$ vanishes like s^β as $s \to 0$; so it vanishes quadratically for the GUE and linearly for the GOE. The same repulsion phenomenon may be seen in Theorem 9.4 and Theorem 9.5. Note that this is unlike the Poisson point process (Example 9.2) where the probability to find nearest-neighbouring random points a distance s apart is maximal at $s = 0$.

9.3 Eigenvalue statistics and Correlation functions

Now that we have the joint PDF of the eigenvalues for some important examples, we look at how we translate this information in terms of eigenvalue process. The main tool for this task is the correlation function.

Definition 9.6. Let $\rho_n(x_1, \ldots, x_n)$ be the (symmetric) joint PDF of the eigenvalues of a $n \times n$ random matrix with real spectrum. The *k-point correlation function* $R_n^{(k)} : \mathbb{R}^k \to [0, \infty)$ is defined to be

$$R_n^{(k)}(x_1, \ldots, x_k) = \frac{n!}{(n-k)!} \int_{-\infty}^\infty \cdots \int_{-\infty}^\infty \rho_n(x_1, \ldots, x_n) \mathrm{d}x_{k+1} \ldots \mathrm{d}x_n. \tag{9.17}$$

These correlation functions have the following interpretation. Consider a measurable set $B \subseteq R$. For $k = 1$, we have that

$$\int_B R_n^{(1)}(x) \mathrm{d}x = \mathbf{E}[\#\{i : \Lambda_i \in B\}] = \mathbf{E}[\nu_N(B)], \tag{9.18}$$

where ν_N is the counting measure defined in (9.1). This is because to count the expected number of eigenvalues in B, we can integrate out the position of Λ_2 up to Λ_n, then integrate the position of Λ_1 over B. By symmetry of ρ_n, we can then multiply by n to get all possible eigenvalues falling in B. This reasoning can be generalized to give for $k = 2$

$$\int_B \int_B R_n^{(2)}(x, y) \mathrm{d}x \mathrm{d}y = \mathbf{E}\Big[\#\Big\{(j_1, j_2) \text{ distinct} : (\Lambda_{j_1}, \Lambda_{j_2}) \in B \times B\Big\}\Big]. \tag{9.19}$$

This is the expected number of (ordered) pairs of eigenvalues falling in B. And for $k \in \mathbb{N}$, we have

$$\int_{B^k} R_n^{(k)}(x_1, \ldots, x_k) \mathrm{d}x_1 \ldots \mathrm{d}x_k = \mathbf{E}\Big[\#\Big\{(j_1 \ldots, j_k) \text{ distinct} : (\Lambda_{j_1}, \ldots, \Lambda_{j_k}) \in B\Big\}\Big].$$

10 The method of Orthogonal Polynomials

The problem of obtaining eigenvalue is now reduced to integrating out the joint PDF. The key here is to unravel a hidden determinantal structure. We illustrate the method using the case of the GUE.

In Theorems 9.3, 9.4 and 9.5, the PDF involves $\prod_{i<j} |\lambda_i - \lambda_j|$. The important observation is that this product is the *Vandermonde determinant*:

$$\prod_{i<j} |\lambda_i - \lambda_j| = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ \lambda_1 & \lambda_2 & \ldots & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & \ldots & \lambda_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \ldots & \lambda_n^{n-1} \end{vmatrix}. \tag{10.1}$$

To see this, note that the determinant is a homogeneous polynomial in the variables $\lambda_1, \lambda_2, \ldots, \lambda_n$ of degree $n(n-1)/2$. It vanishes if any pair of the variables are set equal, so is divisible by $\prod_{i<j}(\lambda_j - \lambda_i)$. This is also a homogeneous polynomial of degree $n(n-1)/2$. Therefore the two polynomials are equal up to a constant multiplier. That this constant has to be one may be seen by comparing any corresponding terms on the two sides; for example the product of terms on the diagonal of the matrix corresponds to multiplying the first terms in $\prod_{i<j}(\lambda_j - \lambda_i)$.

The determinant is unchanged under row operations. Therefore, we can write

$$\prod_{i<j} |\lambda_i - \lambda_j| = \begin{vmatrix} p_0(\lambda_1) & p_0(\lambda_2) & \ldots & p_0(\lambda_n) \\ p_1(\lambda_1) & p_1(\lambda_2) & \ldots & p_1(\lambda_n) \\ p_2(\lambda_1) & p_2(\lambda_2) & \ldots & p_2(\lambda_n) \\ \vdots & \vdots & \ddots & \vdots \\ p_{n-1}(\lambda_1) & p_{n-1}(\lambda_2) & \ldots & p_{n-1}(\lambda_n) \end{vmatrix} \tag{10.2}$$

where p_k is a monic polynomial of degree k. This is true for any monic polynomials p_k. There is, however, one choice that considerably simplifies the calculations:

Take $(p_k(\lambda), k \geq 0)$ to be orthogonal polynomials of the reference measure of the ensemble.

For GUE, the reference measure is $e^{-\lambda^2/2}\mathrm{d}\lambda$ as seen from (9.5). And for CUE in (9.15), this is the Lebesgue measure $\mathrm{d}\theta$. (We do not need to worry about the normalization for now.) As we shall see, the orthogonality of the polynomials allows for an exact computation of the correlation functions, even at finite n (!!!). The key ingredient is Gaudin's lemma (Lemma 10.2).

The exponent 2 on $|\lambda_i - \lambda_j|$ helps the analysis considerably in the case of GUE and CUE. We focus on these examples. The method of orthogonal polynomials works as well for the GOE, but is a little more complicated in that case. It works also for complex Wishart matrices, when the orthogonal polynomials are the generalized Laguerre polynomials.

10.1 Correlation functions of GUE

For the GUE, we want the polynomials to satisfy

$$\int_{-\infty}^{\infty} p_i(x) p_j(x) e^{-x^2/2} \mathrm{d}x = \kappa_i^2 \delta_{ij}, \tag{10.3}$$

for some normalization constant κ_i^2. This clearly holds for all $j < i$ if

$$\int_{-\infty}^{\infty} x^k p_i(x) e^{-x^2/2} \mathrm{d}x = 0, \quad \text{for all } k < i. \tag{10.4}$$

We can arrange this by taking $p_i(x)e^{-x^2/2}$ to be an exact ith derivative of some function that vanishes as $|x| \to \infty$, and the obvious choice is to set

$$p_i(x) = (-1)^i e^{x^2/2} \frac{\mathrm{d}^i}{\mathrm{d}x^i} e^{-x^2/2}. \tag{10.5}$$

These are known as the *Hermite polynomials* and are usually denoted $H_i(x)$ (We shall use $p_k(x)$ and $H_k(x)$ interchangeably [15].) We then also have

$$\kappa_i^2 = \int_{-\infty}^{\infty} p_i^2(x) e^{-x^2/2} \mathrm{d}x = \int_{-\infty}^{\infty} x^i p_i(x) e^{-x^2/2} \mathrm{d}x = i! \sqrt{2\pi}. \tag{10.6}$$

[15] In some references the Hermite polynomials are defined to be orthogonal with respect to the measure $e^{-x^2/2}\mathrm{d}x$, and in others with respect to the measure $e^{-x^2}\mathrm{d}x$. We choose the former convention. The latter convention is more common in the physics literature.

It turns out that the normalized Hermite polynomials $(\kappa_k^{-1} p_k(x), k \geq 0)$ form an orthonormal basis of the (complex) Hilbert space $L^2(e^{-x^2/2}\mathrm{d}x)$ of square-integrable functions, i.e., the space of function f such that

$$\int_{-\infty}^{\infty} |f(x)|^2 e^{-x^2/2}\mathrm{d}x < \infty. \tag{10.7}$$

The inner product on $L^2(e^{-x^2/2}\mathrm{d}x)$ is

$$\langle f, g\rangle = \int_{-\infty}^{\infty} f(x)\overline{g(x)}e^{-x^2/2}\mathrm{d}x. \tag{10.8}$$

Since the product $\prod_{i<j} |\lambda_i - \lambda_j|$ is squared for GUE, we can write

$$\prod_{i<j} |\lambda_i - \lambda_j|^2 = \det(A^{\mathrm{T}} A),$$

where A is the $n \times n$ matrix with elements

$$A_{ij} = p_{i-1}(\lambda_j) \tag{10.9}$$

With this notation we have that $A^{\mathrm{T}} A$ is of the form

$$[A^{\mathrm{T}} A]_{ij} = \sum_{k=0}^{n-1} p_k(\lambda_i) p_k(\lambda_j).$$

This has the form of a *projection kernel*. More precisely, when integrating a function f in L^2 over this kernel over one variable, it projects the function f on the first n polynomials.

We note for later use the following general properties of the Hermite polynomials.

Lemma 10.1. *The Hermite polynomials satisfy*

$$p_{k+1}(x) = x p_k(x) - p_k'(x), \tag{10.10}$$

$$x p_k(x) = p_{k+1}(x) + k p_{k-1}(x), \tag{10.11}$$

$$p_k''(x) - x p_k'(x) = -k p_k(x), \tag{10.12}$$

and for $x \neq y$

$$\sum_{k=0}^{n-1} \frac{p_k(x)p_k(y)}{k!} = \frac{p_n(x)p_{n-1}(y) - p_{n-1}(x)p_n(y)}{(n-1)!(x-y)}. \tag{10.13}$$

Proof. Equation (10.10) follows from differentiating (10.5). (10.11) follows from expanding $x p_k(x)$ in terms of the basis formed by all of the Hermite polynomials and using (10.4). Equation (10.12) follows by differentiating (10.10) and then using (10.11). (10.13) follows from multiplying K_n by $x - y$ and then using (10.11); it is a special case of a general formula known as the Christoffel-Darboux Theorem, which goes back to Christoffel in 1858 (for more on this see, for example, [So75]).

Note that, applying L'Hôpital's rule, it follows from (10.13) that

$$\sum_{k=0}^{n-1} \frac{p_k^2(x)}{k!} = \frac{p_n'(x)p_{n-1}(x) - p_{n-1}'(x)p_n(x)}{(n-1)!}. \tag{10.14}$$

□

It is convenient for the analysis to incorporate the reference density $e^{-x^2/2}$ as well as the normalization constants in the definition of the kernel. By the properties of the determinant, we have

$$\prod_{i<j}|\lambda_i-\lambda_j|^2 e^{-\sum_{j=1}^n \lambda_j^2/2} = \begin{vmatrix} p_0(\lambda_1)e^{-\frac{1}{4}\lambda_1^2} & p_0(\lambda_2)e^{-\frac{1}{4}\lambda_2^2} & \dots & p_0(\lambda_n)e^{-\frac{1}{4}\lambda_n^2} \\ p_1(\lambda_1)e^{-\frac{1}{4}\lambda_1^2} & p_1(\lambda_2)e^{-\frac{1}{4}\lambda_2^2} & \dots & p_1(\lambda_n)e^{-\frac{1}{4}\lambda_n^2} \\ p_2(\lambda_1)e^{-\frac{1}{4}\lambda_1^2} & p_2(\lambda_2)e^{-\frac{1}{4}\lambda_2^2} & \dots & p_2(\lambda_n)e^{-\frac{1}{4}\lambda_n^2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n-1}(\lambda_1)e^{-\frac{1}{4}\lambda_1^2} & p_{n-1}(\lambda_2)e^{-\frac{1}{4}\lambda_2^2} & \dots & p_{n-1}(\lambda_n)e^{-\frac{1}{4}\lambda_n^2} \end{vmatrix}^2. \tag{10.15}$$

Therefore, adding the normalization constants $c_n^{(2)}$ and κ_i's, we get that the joint PDF of GUE given in Equation (9.5) is

$$\rho_n(\lambda_1,\dots,\lambda_n) = c_n^{(2)}\left(\prod_{i=1}^{n-1}\kappa_i^2\right)\left(\det \mathcal{A}^{\mathrm{T}}\mathcal{A}\right) \tag{10.16}$$

where $\mathcal{A}$ is now the $n\times n$ matrix with elements

$$\mathcal{A}_{ij} = \frac{1}{\kappa_{j-1}} p_{j-1}(\lambda_i)e^{-\frac{1}{4}\lambda_i^2}. \tag{10.17}$$

The functions $\phi_k(x) := \kappa_k^{-1}p_k(x)e^{-x^2/4}$ are called the *Hermite functions*. They satisfy the differential equation

$$-\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} + \frac{1}{4}x^2\Psi = (k+\frac{1}{2})\Psi \tag{10.18}$$

which is the Schrödinger equation of the simple harmonic oscillator.

In the above notation, we now have

$$\left[\mathcal{A}^{\mathrm{T}}\mathcal{A}\right]_{ij} = \sum_{k=1}^{n}\mathcal{A}_{ik}\mathcal{A}_{jk} = \sum_{k=0}^{n-1}\phi_k(\lambda_i)\phi_k(\lambda_j). \tag{10.19}$$

The function

$$K_n(x,y) = \sum_{k=0}^{n-1}\phi_k(x)\phi_k(y) = \frac{1}{\sqrt{2\pi}}\sum_{k=0}^{n-1}\frac{1}{k!}p_k(x)p_k(y)e^{-(x^2+y^2)/4} \tag{10.20}$$

is known as the *Christoffel-Darboux kernel* in the theory of orthogonal polynomials. It is the kernel of the projection onto the span of $\phi_0,\dots,\phi_{n-1}$.

In terms of the Christoffel-Darboux kernel, we have that the right-hand side of (10.16) can be written as

$$\rho_n(\lambda_1,\dots,\lambda_n) = C_n\cdot\det\left([K_n(\lambda_i,\lambda_j)]\right), \tag{10.21}$$

for some normalization constant C_n to be determined.

We have from (10.13) that for $x\neq y$

$$K_n(x,y) = \sqrt{n}\frac{\phi_n(x)\phi_{n-1}(y)-\phi_{n-1}(x)\phi_n(y)}{x-y} \tag{10.22}$$

and that

$$K_n(x,x) = \sqrt{n}\left(\phi_n'(x)\phi_{n-1}(x)-\phi_{n-1}'(x)\phi_n(x)\right). \tag{10.23}$$

It follows from the orthonormality of the Hermite polynomials that

$$\int_{-\infty}^{\infty}K_n(x,y)K_n(y,z)\mathrm{d}y = K_n(x,z). \tag{10.24}$$

This allows us to apply the following general lemma

Lemma 10.2 (Gaudin's lemma)**.** *Consider a $n \times n$ matrix J_n of the form*

$$J_n = (J_{ij})_{1\le i,j\le n} = \Big(k(x_i, x_j)\Big)_{1\le i,j\le n}, \tag{10.25}$$

where $k(x, y)$ is a kernel satisfying

$$\int_{-\infty}^{\infty} k(x, y)k(y, z)\mathrm{d}y = k(x, z). \tag{10.26}$$

Then for $r = \int_{-\infty}^{\infty} k(x, x)\mathrm{d}x$, we get

$$\int_{-\infty}^{\infty} \det(J_n)\mathrm{d}x_n = (r - n + 1)\det(J_{n-1}). \tag{10.27}$$

Proof. We note first that expanding the determinant gives

$$\begin{aligned}\int_{-\infty}^{\infty} \det(J_n)\mathrm{d}x_n &= \int_{-\infty}^{\infty} \sum_{\sigma\in\mathcal{S}_n} \mathrm{sgn}(\sigma)k(x_1, x_{\sigma(1)}) \dots k(x_n, x_{\sigma(n)})\mathrm{d}x_n \\ &= \int_{-\infty}^{\infty} \sum_{m=1}^{n} \sum_{\sigma:\sigma(n)=m} \mathrm{sgn}(\sigma)k(x_1, x_{\sigma(1)}) \dots k(x_n, x_m)\mathrm{d}x_n. \end{aligned} \tag{10.28}$$

In the term $m = n$, σ runs over all permutations in S_{n-1} and so the integral evaluates to $r\det(J_{n-1})$, because the sign of σ as a permutation in S_{n-1} is the same as it is in $\mathcal{S}_n$.

When $m < n$, let $j = \sigma^{-1}(n)$, and take $\hat{\sigma} \in S_{n-1}$ to be given by

$$\hat{\sigma}(i) = \begin{cases} \sigma(i), & \text{if } i \neq j \\ m, & \text{if } i = j. \end{cases} \tag{10.29}$$

The map $\{\sigma \in \mathcal{S}_n : \sigma(n) = m\} \longrightarrow S_{n-1}$, $\sigma \mapsto \hat{\sigma}$, is straightforwardly seen to be a bijection, and $\mathrm{sgn}(\hat{\sigma}) = -\mathrm{sgn}(\sigma)$, because the two permutations differ by a transposition (mn). Therefore

$$\begin{aligned}&\int_{-\infty}^{\infty} \sum_{\sigma:\sigma(n)=m} \mathrm{sgn}(\sigma)k(x_1, x_{\sigma(1)}) \dots k(x_n, x_m)\mathrm{d}x_n \\ &\quad= \int_{-\infty}^{\infty} \sum_{\sigma:\sigma(n)=m} \mathrm{sgn}(\sigma)k(x_1, x_{\sigma(1)}) \dots k(x_{n-1}, x_{\sigma(n-1)})k(x_j, x_n)k(x_n, x_m)\mathrm{d}x_n \\ &\quad\quad= \sum_{\hat{\sigma}\in S_{n-1}} -\mathrm{sgn}(\hat{\sigma})k(x_1, x_{\hat{\sigma}(1)}) \dots k(x_{n-1}, x_{\hat{\sigma}(n-1)}) \\ &\quad\quad\quad= -\det J_{n-1}. \end{aligned} \tag{10.30}$$

Finally, using the fact that this last equality holds independently of the value of m, the equation in the statement of the theorem is established. □

We now return to our goal of computing the correlations functions for GUE using (10.21). From Definition (9.6), this means integrating out $n - k$ variables, to get the k-point correlation function. Lemma 10.2 is thus obviously the right tool

Theorem 10.3. *Let $k \in \mathbb{N}$. Then the k-point correlation function of a $n \times n$ GUE matrix is*

$$R_n^{(k)}(\lambda_1, \dots, \lambda_k) = \det\Big(\big[K_n(\lambda_i, \lambda_j)\big]_{1\le i,j\le k}\Big), \tag{10.31}$$

where $K_n(x, y)$ is the projection kernel defined in (10.20) *in terms of the Hermite functions.*

This is remarkable!

It asserts that integrating an $n \times n$ determinant, involving a kernel evaluated at all pairs drawn from n variables, over $n-k$ of those variables, gives a $k \times k$ determinant involving the same kernel. Point processes with this property are called *determinantal point processes*.

Proof. Applying Gaudin's lemma, and using the fact that $r = n$ for this kernel, because the Hermite functions are orthonormal, we have immediately that

$$\int_{-\infty}^{\infty} \rho_n(\lambda_1, \dots, \lambda_n) \mathrm{d}\lambda_n = C_n \det\left(K_n(\lambda_i, \lambda_j)_{1 \le i,j \le n-1}\right). \tag{10.32}$$

Applying Gaudin's lemma a second time now gives

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \rho_n(\lambda_1, \dots, \lambda_n) \mathrm{d}\lambda_{n-1} \mathrm{d}\lambda_n = 2C_n \det\left(K_n(\lambda_i, \lambda_j)_{1 \le i,j \le n-2}\right). \tag{10.33}$$

Repeated applications of Gaudin's lemma therefore give

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \rho(\lambda_1, \dots, \lambda_n) \mathrm{d}\lambda_{k+1} \dots \mathrm{d}\lambda_n = (n-k)! C_n \det\left(K_n(\lambda_i, \lambda_j)_{1 \le i,j \le k}\right). \tag{10.34}$$

In the case when $k = 0$, this reduces to

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \rho(\lambda_1, \dots, \lambda_n) \mathrm{d}\lambda_1 \dots \mathrm{d}\lambda_n = n! C_n \tag{10.35}$$

and so we see that $C_n = 1/n!$. □

10.2 Asymptotics of the correlation functions of GUE in the bulk

We are now ready to compute statistics of the eigenvalues of GUE as $n \to \infty$.

We start with the 1-point correlation function. From (9.18), we expect to be able to recover the semicircle law by rescaling the eigenvalues properly. If we rescale the eigenvalues by $1/\sqrt{n}$ and normalise the 1-point correlation function by $1/n$ as before, we should get for $x \in [-2, 2]$,

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} R_1(\sqrt{n}x) = \lim_{n\to\infty} \frac{1}{\sqrt{n}} K_n(\sqrt{n}x, \sqrt{n}x) = \frac{1}{2\pi}\sqrt{4 - x^2}, \tag{10.36}$$

and for the limit to give 0 if $|x| > 2$. This can be shown to be true using asymptotic formulae for the Hermite polynomials when $n \to \infty$, leading to yet another proof of the semicircle law for the GUE.

Specifically, let us define

$$\Xi_{n+p}(x) = n^{\frac{1}{4}} \phi_{n+p}(x\sqrt{n}) \tag{10.37}$$

for $p = -2, -1, 0$. It follows from the properties of the Hermite polynomials (Lemma 10.10) that

$$\frac{1}{\sqrt{n}} K_n(\sqrt{n}x, \sqrt{n}x) = \Xi_{n-1}^2(x) - \sqrt{\frac{n-1}{n}} \Xi_{n-2}(x) \Xi_n(x). \tag{10.38}$$

The asymptotic formulae we need[16] are as follows. First, setting $x = 2\cos\phi$ with $0 < \phi < \pi$, when $n \to \infty$

$$\Xi_{n+p}(x) \sim \frac{1}{\sqrt{\pi \sin\phi}} \cos\left[n\left(\phi - \frac{1}{2}\sin(2\phi)\right) + \left(p + \frac{1}{2}\right)\phi - \frac{\pi}{4}\right] \tag{10.39}$$

[16] These go under the general name *Plancherel-Rotach asymptotic formule* and can be established via a WKB analysis of (10.18).

uniformly for ϕ in a compact subset of $(0,\pi)$. Second, setting $|x| = 2\cosh\phi$ with $0 < \phi$,

$$\Xi_{n+p}(x) \sim \frac{e^{(p+1/2)\phi}}{\sqrt{2\pi\sinh\phi}} e^{-\frac{n}{2}(e^{2\phi}+1-2\phi)} \tag{10.40}$$

uniformly for ϕ in a compact subset of $(0,\infty)$[17]. Substituting (10.39) into (10.38) gives the semicircle law in $|x| < 2$. Since the semicircle law has a total mass of one in $(-2,2)$, there can be no limiting mass at $x = 2$ or in $|x| > 2$. When $|x| > 2$ this also follows by substituting (10.40) into (10.38).

Note that (10.23) gives that

$$\frac{1}{\sqrt{n}} R_1(\sqrt{n}x) = \phi_n'(\sqrt{n}x)\phi_{n-1}(\sqrt{n}x) - \phi_{n-1}'(\sqrt{n}x)\phi_n(\sqrt{n}x) \tag{10.41}$$

which gives an exact formula for the expectation of the empirical spectral density of a GUE matrix in terms of the Hermite functions. We illustrate this by showing in Figure 8 the results of a numerical experiment. This involved generating 10,000 random matrices from the GUE with $n = 10$, plotting a histogram of the eigenvalues, and comparing with the exact formula (10.41). For ease of visualisation, the eigenvalues have each been divided by 2, so the support of the rescaled semicircle that emerges when $n \to \infty$ is $|x| \le 1$.

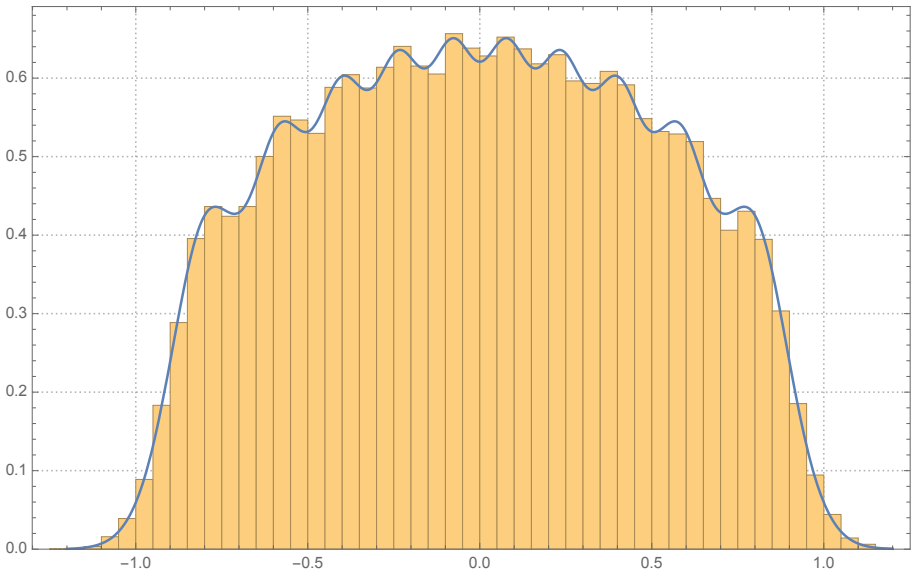

Figure 8: The spectral density of 10,000 GUE matrices of dimension 10 compared to (10.41).

The semicircle law describes the limiting eigenvalue distribution when the eigenvalues are scaled by $1/\sqrt{n}$. There are then n scaled eigenvalues lying between -2 and 2, and so their mean separation is of order $1/n$. We now want to look at correlations on the scale of the mean eigenvalue spacing. We therefore need to scale the eigenvalues so that they have constant mean spacing. There are n eigenvalues lying roughly between $-2\sqrt{n}$ and $2\sqrt{n}$. They therefore have a mean separation that scales like $1/\sqrt{n}$ and so we need to rescale them by $\sqrt{n}$ to achieve a constant mean spacing. This is the correct rescaling in the *bulk of the spectrum*. We will say more about other rescaling in Section 11.2.

More precisely, setting $y_i = \sqrt{n}\lambda_i$, we need to consider the $n \to \infty$ asymptotics of

$$\det\left[n^{-1/2}K_n(y_i/\sqrt{n}, y_j/\sqrt{n})\right]_{i,j=1}^{r} := \det\left[\hat{K}_n(y_i,y_j)\right]_{i,j=1}^{r} \tag{10.42}$$

[17]Note that when $\phi > 0$, $e^{2\phi} + 1 - 2\phi > 0$.

where, when $y_i \neq y_j$,

$$\hat{K}_n(y_i, y_j) = \frac{\phi_n(y_i/\sqrt{n})\phi_{n-1}(y_j/\sqrt{n}) - \phi_{n-1}(y_i/\sqrt{n})\phi_n(y_j/\sqrt{n})}{y_i - y_j}, \tag{10.43}$$

and

$$\hat{K}_n(y, y) = \phi_n'(y/\sqrt{n})\phi_{n-1}(y/\sqrt{n}) - \phi_{n-1}'(y/\sqrt{n})\phi_n(y/\sqrt{n}). \tag{10.44}$$

The important result that we now establish is the following one.

Theorem 10.4. *For any fixed x, y, we have*

$$\lim_{n\to\infty} \hat{K}_n(x, y) = \frac{\sin(\pi(x-y))}{\pi(x-y)}. \tag{10.45}$$

The theorem means that in the limit when $n \to \infty$ the k-point correlation function of the scaled eigenvalues at the centre of the spectrum is simply given by the $k \times k$ determinant involving the kernel

$$\frac{\sin(\pi(x-y))}{\pi(x-y)}. \tag{10.46}$$

This is known as the sine kernel. It is a fundamental object in random matrix theory (and number theory...). The theorem is an immediate consequence of the following lemma.

Lemma 10.5. *For any fixed x and any fixed l*

$$\lim_{n\to\infty} \left| n^{1/4}\phi_{n-l}(x/\sqrt{n}) - \pi^{-1/2}\cos\left(x - (n-l)\pi/2\right) \right| = 0. \tag{10.47}$$

Proof. Let us set $m = n - l$. Note that

$$\begin{aligned}(-1)^m e^{x^2/2}\frac{\mathrm{d}^m}{\mathrm{d}x^m}e^{-x^2/2} &= (-1)^m e^{x^2/2}\frac{\mathrm{d}^m}{\mathrm{d}x^m}\int_{-\infty}^{\infty} e^{-ixz-z^2/2}\frac{\mathrm{d}z}{\sqrt{2\pi}} \\ &= \frac{1}{\sqrt{2\pi}}e^{x^2/2}\int_{-\infty}^{\infty}(iz)^m e^{-ixz-z^2/2}\mathrm{d}z \end{aligned} \tag{10.48}$$

and so

$$n^{1/4}\phi_m(x/\sqrt{n}) = (2\pi)^{-3/4}(m!)^{-1/2}e^{x^2/4n}n^{1/4}\int_{-\infty}^{\infty}(iz)^m e^{-ixz/\sqrt{n}-z^2/2}\mathrm{d}z. \tag{10.49}$$

Using Stirling's formula to evaluate the prefactor and Laplace's method[18] to evaluate the integral when $n \to \infty$ proves the lemma. (See, for example, Chapter VIII in [So75].) □

Substituting the asymptotic estimate in the lemma into the formula for $\hat{K}_n$, and using (10.10) and (10.11) to evaluate the asymptotics of the derivative of ϕ, then leads to the formula in the theorem.

We can now compute the asymptotics of the 2-point correlation functions. Applying Theorem 10.4, we get

$$\lim_{n\to\infty} R_n^{(2)}(\sqrt{n}x, \sqrt{n}y) = 1 - \left(\frac{\sin(\pi(x-y))}{\pi(x-y)}\right)^2. \tag{10.50}$$

So, for example, for $f(x, y)$ such that the sum and integral converge, we get (as in (9.19))

$$\lim_{n\to\infty}\sum_{i\neq j} f(\sqrt{n}\lambda_i, \sqrt{n}\lambda_j) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\left(1 - \left(\frac{\sin(\pi(x-y))}{\pi(x-y)}\right)^2\right)\mathrm{d}x\mathrm{d}y. \tag{10.51}$$

[18] One can also use the method of steepest descent; there are two saddle points and the contributions from each have to be summed.

Note that the quadratic repulsion manifest in Theorem 9.3 is clear here in (10.50). If we set $x - y = w$, the integrand vanishes like w^2 as $w \to 0$. We also remark that for the Poisson point process, the corresponding formula is obtained by replacing the two-point correlation function (10.50) by 1, because the points are, by definition, uncorrelated.

We have shown that at the centre of the spectrum (i.e., at the centre of the semicircle) the limiting correlations are determined by the sine kernel. In fact, the sine kernel can be shown to describe the local correlations at any point in $(-2, 2)$ provided one normalizes the mean spacing to be unity using the semicircle density.

10.3 Correlation functions of CUE

The method of orthogonal polynomials extends straightforwardly to the CUE. Indeed, in this case it is considerably simpler, because the functions that are orthogonal with respect to the uniform weight on the unit circle are simply the Fourier functions $e^{\mathrm{i}m\theta}$, $m \in \mathbb{N}$. All that is involved is the manipulation of Fourier series and the asymptotic analysis is elementary.

Theorem 10.6. *Let $k \in \mathbb{N}$. Then the k-point correlation function of a $n \times n$ CUE matrix is*

$$R_n^{(k)}(e^{\mathrm{i}\theta_1}, \dots, e^{\mathrm{i}\theta_k}) = \det\Big(\big[S_n(\theta_i, \theta_j)\big]_{1 \le i,j \le k}\Big), \tag{10.52}$$

where $S_n(x, y)$ is the projection kernel defined by

$$S_n(\theta, \theta') = \frac{1}{2\pi} \sum_{p=-\frac{1}{2}(n-1)}^{\frac{1}{2}(n-1)} e^{\mathrm{i}p(\theta-\theta')} = \frac{1}{2\pi} \frac{\sin(n(\theta-\theta')/2)}{\sin((\theta-\theta')/2)}. \tag{10.53}$$

Proof. The Vandermonde factor in the Weyl integration formula in Theorem 9.4 can be written as

$$\begin{aligned} \prod_{j<k} |e^{\mathrm{i}\theta_j} - e^{\mathrm{i}\theta_k}|^2 &= \prod_{j<k} \big(e^{\mathrm{i}\theta_j} - e^{\mathrm{i}\theta_k}\big) \prod_{j<k} \big(e^{-\mathrm{i}\theta_j} - e^{-\mathrm{i}\theta_k}\big) \\ &= \det\Big([e^{\mathrm{i}m\theta_j}]\Big)\overline{\det\Big([e^{\mathrm{i}m\theta_j}]\Big)}. \end{aligned} \tag{10.54}$$

In these determinants, the rows are indexed by $m = 0, 1, \dots, n-1$, and the columns by $j = 1, 2, \dots, n$. Hence, we get

$$\prod_{j<k} |e^{i\theta_j} - e^{i\theta_k}|^2 = (2\pi)^n \det\Big(\big[S_n(\theta_j, \theta_k)\big]_{j,k=1,\dots,n}\Big). \tag{10.55}$$

We have by direct computation (this is just Fourier series) that

$$\int_0^{2\pi} S_n(\theta_j, \theta_k) S_n(\theta_k, \theta_l) \mathrm{d}\theta_k = S_n(\theta_j, \theta_l). \tag{10.56}$$

We can therefore apply Gaudin's Lemma 10.2, but now with integrals over $[0, 2\pi)$. We have trivially

$$\int_0^{2\pi} \tilde{S}_n(\theta, \theta) \mathrm{d}\theta = n \tag{10.57}$$

and so the lemma yields that the k-point correlation function is $\det\Big(\big[S_n(\theta_i, \theta_j)\big]_{1 \le i,j \le k}\Big)$. □

The asymptotics of the kernel are now almost trivial. What is the right scaling for CUE? Since there are n eigenvalues on the circle, we expect that the mean spacing is $2\pi/n$. To make this precise, note that the 1-point correlation function is

$$R_1(\theta) = S_n(0) = \frac{n}{2\pi}. \tag{10.58}$$

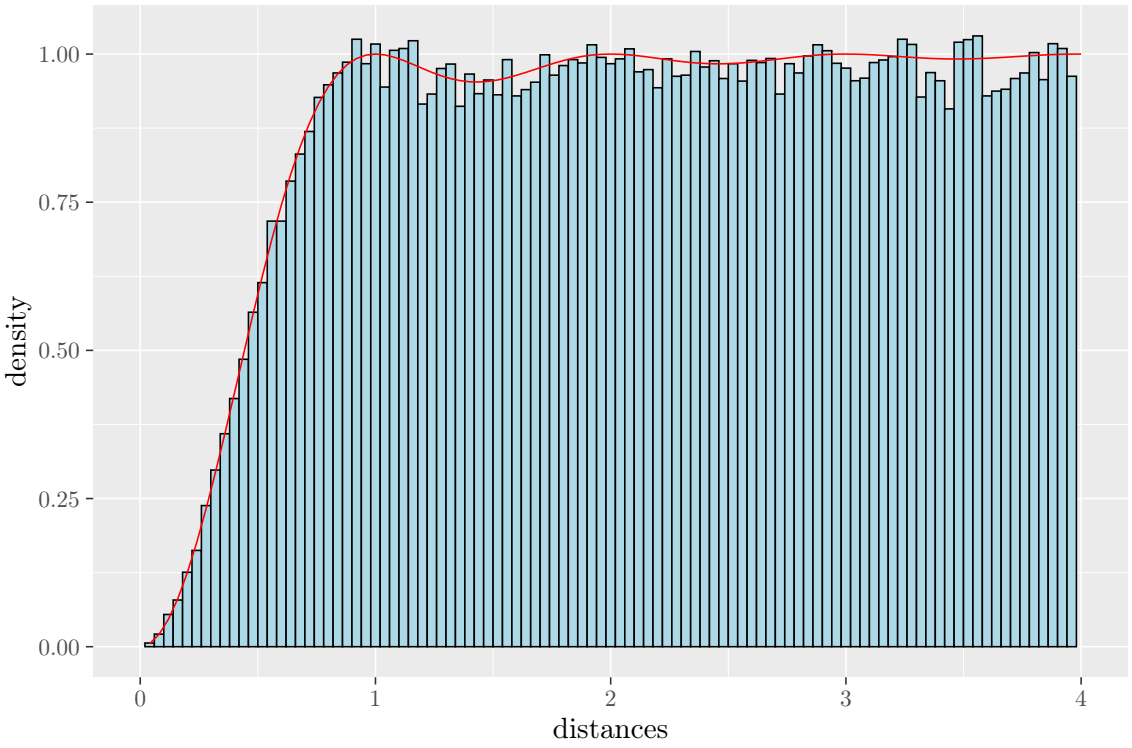

Figure 9: The two-point correlation function of the eigenvalues of 200 CUE matrices of dimension 200 compared to the 2×2 determinant involving the sine kernel. (Plot kindly provided by Johannes Forkel.)

It is the analogue of the semicircle law in this setting after normalizing by $1/n$. The eigenvalues have a constant average density around the unit circle so the ESM is the uniform probability on $[0, 2\pi]$. This was to be expected (Why?).

The 2-point correlation function is more interesting. We get

$$R_2(\theta_1, \theta_2) = \left(\frac{n}{2\pi}\right)^2 - \left(S_n(\theta_1, \theta_2)\right)^2. \tag{10.59}$$

Now crucially, observe that if we scale the eigenvalues to have unit mean spacing, this corresponds to setting $\varphi_k = \frac{n}{2\pi}\theta_k$. Making this substitution in S_n and rescaling by $2\pi/n$, we see that

$$\lim_{n\to\infty} \frac{1}{n} S_n(2\pi\varphi_j/n, 2\pi\varphi_k/n) = \lim_{n\to\infty} \frac{\sin(\pi(\varphi_j - \varphi_k)}{n \sin(\pi(\varphi_j - \varphi_k)/n)} = \frac{\sin(\pi(\varphi_j - \varphi_k)}{\pi(\varphi_j - \varphi_k))}. \tag{10.60}$$

The limit gives precisely the sine kernel we found for the GUE! Hence, even though the statistics are different when n is finite, in the limit when $n \to \infty$ the local statistics of the GUE and CUE are the same. We illustrate this by showing in Figure 9 the results of a numerical experiment. This involved generating 200 random matrices from the CUE with $n = 200$, plotting a histogram of the distances between all pairs of eigenvalues, and comparing with the 2×2 determinant involving the sine kernel.

The above is (yet) another instance of universality. Given that the GUE and the CUE exhibit the same local statistics on the scale of the mean eigenvalue spacing in the limit $n \to \infty$ one might ask whether other random matrix ensembles also fall into the same class. Does the result depend on having simple measures, for example? Or does it rely on having nice formulae for the measure in terms of the eigenvalues, expressed via Vandermonde determinants? The answer is that in the large-matrix limit

the results are independent of the measure, provided we have matrices that are essentially complex and Hermitian (unitary matrices can be thought of as the exponentials of complex Hermitian matrices). So complex Hermitian Wigner random matrices all have the same limit, once some mild conditions are satisfied, no matter what the measure one applies to the matrix elements, and for most measures one doesn't have nice expressions in terms of the eigenvalues.

Similarly, real-symmetric matrices form a different *universality class*, characterised by linear repulsion between the eigenvalues as opposed to quadratic repulsion. The GOE falls into this class, along with other real-symmetric Wigner matrices, irrespective of the measure.

Proving universality has been a major theme of research in Random Matrix Theory over the past 15 years. We will say a bit more about this in Part III.

11 A General Method

In this section, we explain how the method we have developed can be applied to get statistics beyond the correlation function and at the edge of the spectrum.

11.1 Counting statistics

The definition of the Poisson point process 9.1 gives the complete joint distribution of the random variables $N(A)$ for countably many subsets A's. This completely characterizes the counting measure in the Poisson case. We want to achieve the same level of precision for the eigenvalue process defined by the counting measure

$$\nu_n = \sum_{i=1}^{n} \delta_{\Lambda_i}, \tag{11.1}$$

where $(\Lambda_1, \ldots, \Lambda_n)$ are the eigenvalues of a $n \times n$ eigenvalues with real spectrum.

Fix a measurable subset $B \subseteq \mathbb{R}$. Recall that $\nu_n(B)$ is the number of eigenvalues lying in B. We want to determine

$$\mathbf{P}(\nu_n(B) = m), \quad m = 0, 1, 2, \ldots \tag{11.2}$$

Let $\mathbb{1}_B(x)$ denote the indicator function on B, so

$$\mathbb{1}_B(x) = \begin{cases} 1 \text{ if } x \in B \\ 0 \text{ if } x \notin B. \end{cases} \tag{11.3}$$

Then, we have

$$\mathbf{P}(\nu_n(B) = 0) = \mathbf{E}\left[\prod_{i=1}^{n}(1 - \mathbb{1}_B(\lambda_i))\right]. \tag{11.4}$$

We define for $t \in \mathbb{R}$ the generating function

$$G_B(t) = \mathbf{E}\left[\prod_{i=1}^{n}(1 - t\mathbb{1}_B(\Lambda_i))\right], \tag{11.5}$$

so that $\mathbf{P}(\nu_n(B) = 0) = G_B(1)$. It is easy to see that $-G_B'(1) = \mathbf{P}(\nu_n(B) = 1)$, and that more generally

$$\mathbf{P}(\nu_n(B) = m) = \frac{(-1)^m}{m!} G_B^{(m)}(1). \tag{11.6}$$

$G_B(t)$ can therefore be seen the generating function for $\mathbf{P}(\nu_n(B) = m)$.

Proposition 11.1. *Let ν_n be the counting measure* (11.1)*. Then for any measurable subset $B \subseteq \mathbb{R}$, we have*

$$G_B(t) = 1 + \sum_{k=1}^{\infty} \frac{(-t)^k}{k!} \int_B \cdots \int_B R_n^{(k)} \mathrm{d}x_1 \ldots \mathrm{d}x_k.$$

Proof. We expand the product over the eigenvalues to get

$$\begin{aligned} G_B(t) &= 1 - t\mathbf{E}\sum_i \mathbb{1}_B(\Lambda_i) + t^2\mathbf{E}\sum_{i\le j}\mathbb{1}_B(\Lambda_i)\mathbb{1}_B(\Lambda_j) - \dots \\ &= 1 - t\int_I R_1(x_1)\mathrm{d}x_1 + \frac{t^2}{2}\int_B\int_B R_2(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2 - \frac{t^3}{3!}\int_I\int_I\int_I R_3(x_1,x_2,x_3)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3 + \dots \\ &= 1 + \sum_{k=1}^{\infty}\frac{(-t)^k}{k!}\int_B\cdots\int_B R_n^{(k)}\mathrm{d}x_1\dots\mathrm{d}x_k. \end{aligned} \tag{11.7}$$

□

In particular, if the point process is determinantal, that is, its corrrelation functions are given in terms of the determinant of a kernel K_n, we get

$$G_B(t) = 1 + \sum_{k=1}^{\infty}\frac{(-t)^k}{k!}\int_B\cdots\int_B \det\left([K_n(x_i,x_j)]_{1\le i,j\le k}\right)\mathrm{d}x_1\dots\mathrm{d}x_k. \tag{11.8}$$

The last equation can be viewed as defining the *Fredholm determinant* of the operator with kernel $K_n(x,y)$ acting on $L^2(B,\mathrm{d}x)$:

$$\det\left(\mathrm{I} - tK_n(x,y)\right) := 1 + \sum_{k=1}^{\infty}\frac{(-t)^k}{k!}\int_B\cdots\int_B \det\left(K_n(x_i,x_j)\right)_{1\le i,j\le k}\mathrm{d}x_1\dots\mathrm{d}x_k. \tag{11.9}$$

In the theory of Fredholm determinants[19] an operator with kernel $K(x,y)$ acts on a function f according to $(Kf)(x) = \int K(x,y)f(y)\mathrm{d}y$. When $\mathrm{Tr}(K) = \int K(x,x)\mathrm{d}x < \infty$, then K is called *trace class* and one can calculate the Fredholm determinant in a number of (equivalent) ways:

$$\det\left(\mathrm{I} - tK_n(x,y)\right) = \exp\left(-\sum_{m=1}^{\infty}\frac{t^m}{m}\mathrm{Tr}K^m\right) \tag{11.10}$$

where K^m is to be understood as an m-fold convolution of K. Alternatively, denoting the eigenvalues of K by β_j,

$$\det\left(\mathrm{I} - tK_n(x,y)\right) = \prod(1 - t\beta_j). \tag{11.11}$$

We also have that then $\int K(x,x)\mathrm{d}x = \sum_j \beta_j$.

Many quantities can be computed from the Fredholm determinant. An important example is the PDF P_n of the spacing distribution between nearest-neighbouring eigenvalues. We claim that for the GUE in the bulk we have

$$P_n(s) = \frac{\mathrm{d}^2 G_{B_s}(t)}{\mathrm{d}s^2}\bigg|_{t=1}, \tag{11.12}$$

where B_s is an interval of length s. To see this, recall that the probability that this interval contains no eigenvalues is $\mathbf{P}(\nu_n(B_s) = 0)$. Let's denote this probability by $F(s)$. Now extend the interval by $\mathrm{d}s$ at one end. The probability that this extended interval has no eigenvalues in it is $F(s+\mathrm{d}s)$. Hence $F(s) - F(s+\mathrm{d}s)$ is the probability that there is no eigenvalue in the original interval, but at least one in the extended interval. Extending the interval at the other end one gets (11.12) When the eigenvalues are scaled by $\sqrt{n}$, then when $n\to\infty$ the limiting spacing distribution is the second derivative of

[19] We limit ourselves to stating results in the theory of Fredholm determinants without proof, and without setting out the theory systematically. For an introductory exposition see, for example, [McK11].

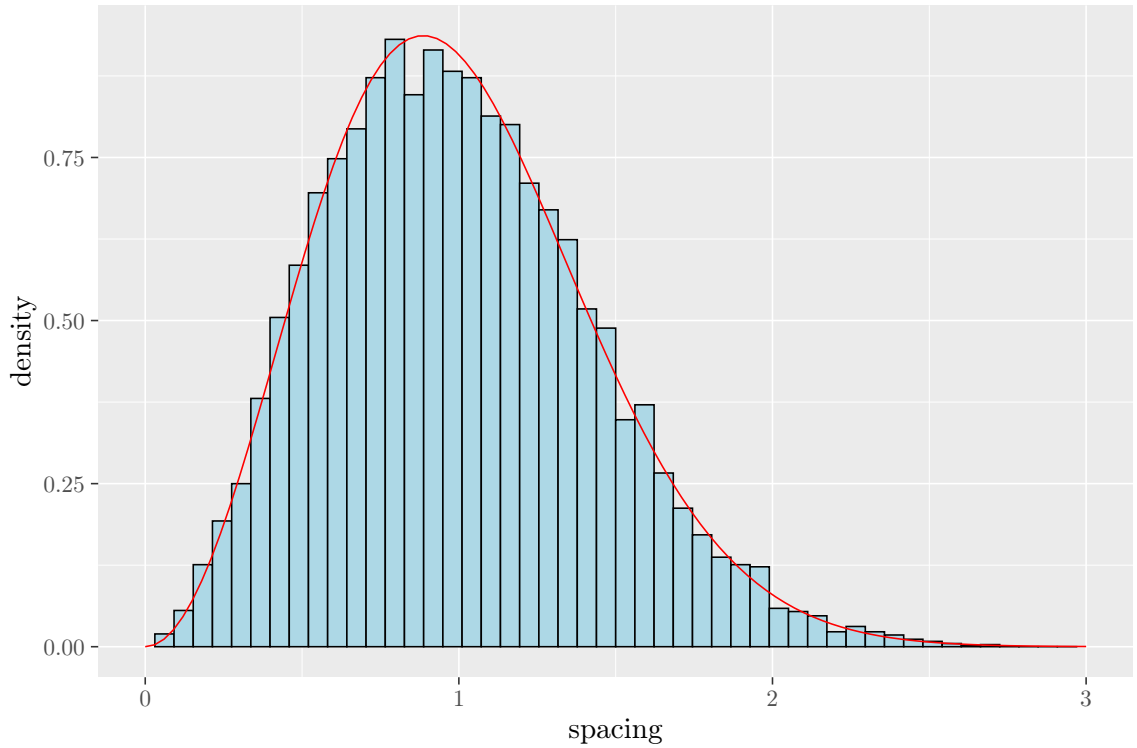

Figure 10: The spacing distribution between neighbouring eigenvalues of 10,000 GUE matrices of dimension $n = 10$ compared to the the Fredholm determinant of the sine kernel. (Plot kindly provided by Johannes Forkel.)

the Fredholm determinant of the sine kernel[20]. This distribution is often called the *Gaudin-Mehta* distribution

We illustrate this by showing in Figure 10 the results of a numerical experiment. This involved generating 10,000 random matrices from the GUE with $n = 10$, plotting a histogram of the normalized distances between neighbouring pairs of eigenvalues, and comparing with the Fredholm determinant of the sine kernel.

11.2 Scaling at the edge of the spectrum

Having seen how the sine kernel describes correlations in the *bulk* of the spectrum for GUE, we now consider how the eigenvalues are distributed near the edges.

By the semicircle law, we should have, when ϵ is small, that

$$\#\{\Lambda_i : \Lambda_i/\sqrt{n} \geq 2 - \epsilon\} \approx \frac{n}{2\pi} \int_{2-\epsilon}^{2} \sqrt{4 - x^2} \mathrm{d}x \approx \frac{2}{3\pi} n \epsilon^{3/2}. \tag{11.13}$$

We might therefore expect to get a nice limit if we set $\epsilon \propto n^{-2/3}$, i.e., if we rescale the distance between the eigenvalues and the edge by $n^{-2/3}$. If we want to find a limiting distribution for the eigenvalues close to the upper edge of the spectrum, this suggests setting $\lambda_i/\sqrt{n} = 2 + \alpha_i n^{-2/3}$, or

[20]The Fredholm determinant of the sine kernel is equal to $\exp\left(\int_0^{\pi s} \frac{\sigma(x;t)}{x} \mathrm{d}x\right)$ where σ is the solution of the Painlevé V equation $(x\sigma'')^2 + 4(x\sigma' - \sigma)(x\sigma' - \sigma + (\sigma')^2) = 0$ with $\sigma(x;t) \sim -\frac{t}{\pi} x$ as $x \to 0$, but showing this is beyond the scope of the present course.

$\lambda_i = 2\sqrt{n} + \alpha_i n^{-1/6}$ and looking at correlations between the values taken by $\alpha_1, \ldots, \alpha_n$. This involves studying the asymptotics of

$$\det\left(\left[n^{-1/6}K_n(2\sqrt{n}+\alpha_i n^{-1/6}, 2\sqrt{n}+\alpha_j n^{-1/6})\right]_{i,j\le n}\right), \tag{11.14}$$

where K_n is the GUE kernel given in (10.20). Defining

$$\tilde{K}_n(x,y) = n^{-1/6}K_n(2\sqrt{n}+xn^{-1/6}, 2\sqrt{n}+yn^{-1/6}), \tag{11.15}$$

we have the following theorem describing the $n\to\infty$ asymptotics.

Theorem 11.2. *For any fixed x, y*

$$\lim_{n\to\infty}\tilde{K}_n(x,y) = \frac{\mathrm{Ai}(x)\mathrm{Ai}'(y) - \mathrm{Ai}'(x)\mathrm{Ai}(y)}{(x-y)} \tag{11.16}$$

where

$$\mathrm{Ai}(x) = \frac{1}{\pi}\int_0^\infty \cos(t^3/3 + xt)\mathrm{d}t \tag{11.17}$$

is the Airy function.

This is an immediate consequence of the following lemma

Lemma 11.3. *For any fixed x*

$$\lim_{n\to\infty}\left|n^{1/12}\phi_n(2\sqrt{n}+x/n^{1/6}) - \mathrm{Ai}(x)\right| = 0. \tag{11.18}$$

We only sketch the proof of this lemma. This starts with (10.49). One again seeks to evaluate the integral by the method of steepest descent. The difference is that in this case the two saddle points lie so close together that one cannot treat them separately. Instead of expanding to quadratic order, one has to expand to cubic order. This leads to the Airy function, rather than the usual Gaussian integral. The Airy function is plotted in Figure 11.

Substituting the asymptotic estimate in the lemma into the formula for $\tilde{K}_n$ in terms of ϕ, and using (10.10) and (10.11) to evaluate the asymptotics of the derivative of ϕ, then leads to the formula in the theorem.

The kernel representing the $n\to\infty$ limit of $\tilde{K}_n$ is known as the *Airy kernel*. Far away from the edge on the inside of the semicircle, it reduces to the sine kernel. Away from the edge on the outside of the semicircle, it decays exponentially.

The analysis of the counting statistics of Section 11.1 extends immediately to the edge of the spectrum, with the kernel in $G_B(t)$, see (11.9), being the Airy kernel. Rather than the spacing distribution, the natural question in this case is: what is the distribution of the largest eigenvalue of a GUE matrix? Let $\lambda_{\max}$ denote the largest eigenvalue and set

$$F_n^{(\max)}(t) = \mathbf{P}(\Lambda_{\max} \le t). \tag{11.19}$$

Then from the scaling at the edge we have that

$$F^{(\max)}(s) = \lim_{n\to\infty} F_n^{(\max)}(2\sqrt{n} + s/n^{1/6}) \tag{11.20}$$

exists. Clearly $\mathbf{P}(\lambda_{\max}\le t)$ is simply the probability that the interval (t,∞) contains no eigenvalues, and so

$$F^{(\max)}(s) = \det(\mathrm{I} - K_{\mathrm{Airy}}), \tag{11.21}$$

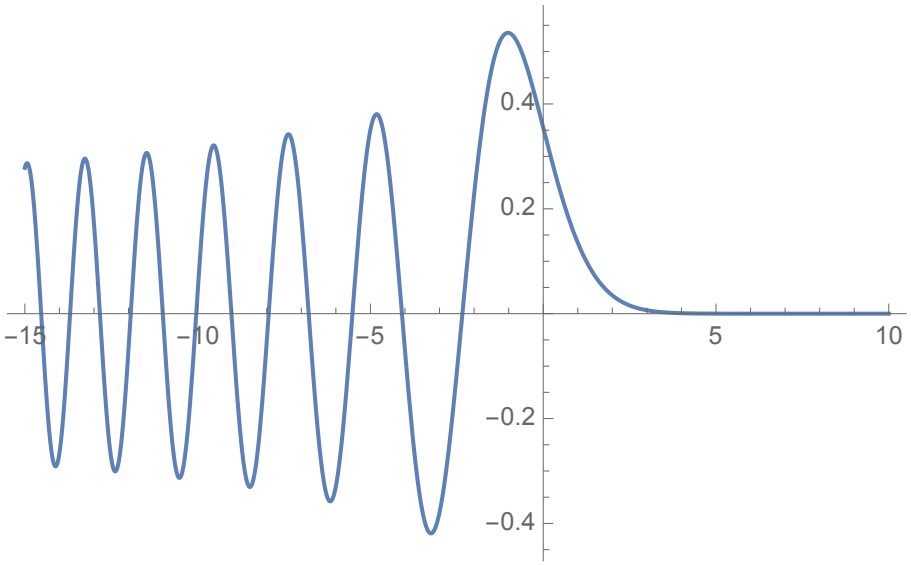

Figure 11: The Airy function $\mathrm{Ai}(x)$.

where

$$K_{\mathrm{Airy}} = \frac{\mathrm{Ai}(x)\mathrm{Ai}'(y) - \mathrm{Ai}'(x)\mathrm{Ai}(y)}{(x-y)}. \tag{11.22}$$

This kernel now acts on $L^2((s,\infty))$.

In this case, the CDF $F^{(\max)}(s)$ can be expressed in terms of a solution of a second order nonlinear ODE, which is a Painlevé equation. Specifically,

$$F^{(\max)}(s) = \exp\left(-\int_s^\infty (x-s)q(x)^2 \mathrm{d}x\right) \tag{11.23}$$

where q is a solution of the Painlevé II equation $q''(x) - xq(x) + 2q(x)^3 = 0$ with $q(x) \sim \mathrm{Ai}(x)$ as $x \to \infty$. This is called the *Tracy-Widom distribution.*

We illustrate this by showing in Figure 12 the results of a numerical experiment, which involved generating 10,000 random matrices from the GUE with $n = 1000$, plotting a histogram of $(\lambda_{\max} - 2\sqrt{n})n^{1/6}$, and comparing with the Tracy-Widom distribution.

11.3 Applications

The local spectral statistics of random matrices in the bulk of the spectrum have many applications. For example, they describe fluctuation statistics in complex quantum systems, including quantum chaotic systems, atomic spectra and nuclear spectra, and in other complex wave problems, such as lasers, elastic vibrations, and acoustics etc. They also provide the main method for modelling statistical properties of quantum dynamics in complex systems, such as in conductivity through disordered media.

Remarkably, GUE spectral statistics also appear to describe correlations between the positions of parked cars in London, and arrival times of buses in the Mexican city of Cuernavaca.

Knowing the distribution of largest and smallest eigenvalues is important in many contexts in numerical linear algebra in determining bounds on the efficiency and convergence of algorithms. In Wishart random matrices the ratio of the extremal eigenvalues determines how errors in the input data get magnified in solving systems of linear equations, as measured by the condition numbers of the matrices appearing.

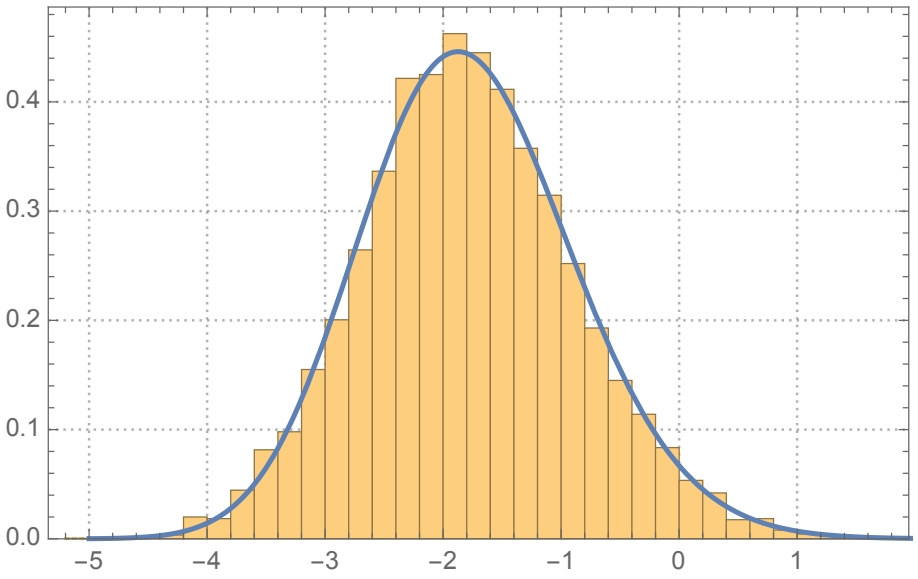

Figure 12: $(\lambda_{\max} - 2\sqrt{n})n^{1/6}$ for 10,000 GUE matrices of size $n = 1000$.

For more on these applications, see the respective chapters in The Oxford Handbook of Random Matrix Theory [ABDF11].

Part III
Towards Universality

We saw in Part III that the some ensembles (for example GUE and CUE) share the same eigenvalue statistics when looking at the correct scale of the spectrum. This *universality phenomenon* is very intriguing and has kept mathematicians and physicists busy in trying to prove it.

One can surmise for example that the correlation functions in the bulk of every complex Wigner matrix should be given in the limit $n \to \infty$ by the sine kernel, as for the GUE and the CUE. And in a similar fashion, correlation functions at the edge should be controlled by the Airy kernel. There has been tremendous progress in addressing these questions in the last twenty years. We refer to [EoY17] for a survey of the results and the techniques. The key to universality turns out to be a dynamical approach known as *Dyson Brownian motion*. We explain the basics of the theory in Section 12.

It turns out that random matrix statistics appear in other seemingly unrelated mathematical and physical problems. This is a fascinating area of research! We give some puzzling examples in Section 13.

12 A Dynamical Approach

We start by definition the Dyson Brownian motion. We then explain its relation with the Stieltjes measure.

12.1 Dyson Brownian motion

We can express the GOE and GUE measures in terms of the eigenvalues, as embodied in Theorem 9.3, in the form

$$c_n^{(\beta)} e^{-\beta H(\lambda_1,\dots,\lambda_n)}, \tag{12.1}$$

where

$$H(\lambda_1,\dots,\lambda_n) = \frac{1}{4}\sum_{j=1}^{n} \lambda_j^2 - \sum_{i<j} \ln|\lambda_i - \lambda_j|. \tag{12.2}$$

One can think of this as a Boltzmann weight (in statistical mechanics), with $H(\lambda_1,\dots,\lambda_n)$ representing the potential energy associated with particles located at positions $\lambda_1,\dots,\lambda_n$, and with β playing the role of the inverse temperature. Hence computing expectations with respect to the RMT measures in this case is the same as determining the equilibrium thermodynamics of particles moving in one dimension with interactions described by this energy function.

We now construct the dynamics on the eigenvalues with (12.1) as its equilibrium (or stationary) measure. This means that we need to define a matrix-valued stochastic process $(\mathbf{M}(t), t \geq 0)$. We start with GOE. For $t \geq 0$, consider the $n \times n$ symmetric matrix $\mathbf{B}(t)$ where $\mathbf{B}_{ij}$ are IID standard Brownian motions, see Definition D.1. The dynamics of the entry $\mathbf{M}_{ij}(t)$ is given the following stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{M}_{ij}(t) = \sigma_{ij}\mathrm{d}\mathbf{B}_{ij}(t) - k_{ij}\mathbf{M}_{ij}\mathrm{d}t, \quad i \leq j, \tag{12.3}$$

with $k_{ij} = 1/2$ for all i, j and

$$\sigma_{ij} = \begin{cases} 1 & \text{if } i < j, \\ \sqrt{2} & \text{if } i = j, \end{cases}$$

and initial condition $\mathbf{M}(0) = M_0$. This is the SDE of an *Ornstein-Uhlenbeck process*, see Definition D.4. As $t \to \infty$, we have that the distribution of $M_{ij}(t)$ approaches $\mathcal{N}(0,1)$ if $i \neq j$ and $\mathcal{N}(0,2)$ if $i = j$. For GUE, we can define a similar dynamics on the real part and imaginary part of $\mathbf{M}_{ij}(t)$. We then have $\sigma_{ij} = 1/\sqrt{2}$ for the real and imaginary part or $i \neq j$, and $\sigma_{ii} = 1/\sqrt{2}$ with $k_{ij} = 1/2$ in all cases.

The question we seek to address is: if the matrix M evolves in time according to (12.3), what is the induced dynamics on the eigenvalues $\Lambda_1(t), \Lambda_2(t), \ldots, \Lambda_n(t)$? First, one might wonder about the labelling of the eigenvalues if they become degenerate or exchange positions. In fact, if we assume that the eigenvalues at $t = 0$ are non-degenerate and if we label them in order of increasing size, $\Lambda_1(0) < \Lambda_2 < (0) < \cdots < \Lambda_n(0)$, then as time evolves it can be shown that the eigenvalues remain simple and are continuous functions of t, and so the labelling makes sense and is preserved for all $t > 0$.

The equation governing the time evolution of the eigenvalues might be expected to involve the eigenvalues and eigenvectors $\mathbf{v}_j(t) \in \mathbb{R}^n$. However, remarkably, this turns out not to be the case: the eigenvalues satisfy an autonomous system of stochastic differential equations that do not involve the eigenvectors. The solution of this equation is known as *Dyson Brownian motion* with parameter β, after Freeman Dyson, who introduced it in a seminal paper in 1962.

Theorem 12.1 ([Dys62]). *Let $\mathbf{M}(t)$, $t \geq 0$ be a $n \times n$ random matrix evolving with the dynamics (12.3). Then its eigenvalue process satisfies the SDE*

$$\mathrm{d}\Lambda_i(t) = \sqrt{2}\mathrm{d}\widetilde{B}_i(t) + \left(-\frac{\Lambda_i(t)}{2} + \sum_{j \neq i} \frac{1}{\Lambda_i(t) - \Lambda_j(t)} \right) \mathrm{d}t, \quad 1 \leq i \leq n, \tag{12.4}$$

where $\widetilde{B}(t) = (\widetilde{B}_1(t), \ldots, \widetilde{B}_n(t))$ is a n-dimensional standard Brownian motion. The same holds for GUE and the other Gaussian β-ensembles by replacing the volatility $\sqrt{2}$ by $\sqrt{2/\beta}$.

As noted above, it can be proved (see, for example, section 4.3 in [AGZ10]) that there exists a unique solution (in the strong sense) of this equation in the space of continuous functions. Moreover, if the initial conditions are such that all of the eigenvalues are simple at $t = 0$, then this remains true for all $t > 0$, i.e., none of the eigenvalues intersect.

Proof. The proof is an application of Itô's formula (Theorem D.2). We focus on the GOE case. We denote the eigenvalues of the $n \times n$ matrix $\mathbf{M}(t)$ by $(\Lambda_m(t))_{1 \leq m \leq n}$ and the corresponding eigenvectors by $(\mathbf{v}^{(m)}(t))_{1 \leq m \leq n}$. We will sometimes omit the dependence on t for conciseness. The starting point is the equation

$$\mathbf{D} = \mathbf{O}^{\mathrm{T}}\mathbf{M}\mathbf{O},$$

where the m-th column of $\mathbf{O}$ is $\mathbf{v}^{(\mathbf{m})}$, and the m-th diagonal entry of $\mathbf{D}$ is Λ_m. This is an equation between random variables. And clearly, the eigenvalues on the right are a function of the Ornstein-Uhlenbeck processes on the left. We write this explicitly as

$$\Lambda_i(t) = f_i(\mathbf{M}(t)) = {\mathbf{v}^{(\mathbf{i})}}^{\mathrm{T}}\mathbf{M}\mathbf{v}^{(\mathbf{i})}, \quad i \leq n. \tag{12.5}$$

It is convenient to index the entries of the matrix $\mathbf{M}$ by $\alpha = (\alpha_1, \alpha_2)$ with $\alpha_1 \leq \alpha_2$ to lighten the notation. This index α only runs over entries with $i \leq j$ since $i > j$ is determined by symmetry. We write $\partial_\alpha f(\mathbf{M})$ for the derivative of the function f with respect to the α-entry of $\mathbf{M}$. Itô's formula (D.3) gives

$$\mathrm{d}\Lambda_i(t) = \sum_\alpha \partial_\alpha f_i(\mathbf{M}(t))\ \mathrm{d}\mathbf{M}_\alpha(t) + \frac{1}{2}\sum_{\alpha,\beta} \partial_\alpha\partial_\beta f_i(\mathbf{M}(t))\ \mathrm{d}\mathbf{M}_\alpha(t)\mathrm{d}\mathbf{M}_\beta(t). \tag{12.6}$$

We compute the first and second derivatives. Differentiating (12.5) with respect to M_α yields

$$\partial_\alpha\Lambda_i(t) = {\mathbf{v}^{(\mathbf{i})}}^{\mathrm{T}}(\partial_\alpha\mathbf{M})\mathbf{v}^{(\mathbf{i})}.$$

This is because the relation ${\mathbf{v}^{(i)}}^{\mathrm{T}}\mathbf{v}^{(i)} = 1$ implies ${\mathbf{v}^{(i)}}^{\mathrm{T}}(\partial_\alpha\mathbf{v}^{(i)}) = 0$. The differentiated matrix $\partial_\alpha\mathbf{M}$ is for $k \neq l$, $[\partial_\alpha\mathbf{M}]_{kl} = \delta_{\alpha_1 k}\delta_{\alpha_2 l} + \delta_{\alpha_2 k}\delta_{\alpha_1 l}$ (remember the symmetry!). This is simply a matrix a 0 with 1 at the entry $\alpha_1\alpha_2$ and $\alpha_2\alpha_1$. For the diagonal, we get $[\partial_\alpha\mathbf{M}]_{kk} = \delta_{\alpha_1 k}\delta_{\alpha_2 k}$. Thus, this leads to

$$\partial_\alpha\Lambda_i(t) = (2 - \delta_{\alpha_1\alpha_2})\mathbf{v}^{(i)}_{\alpha_1}\mathbf{v}^{(i)}_{\alpha_2} = \begin{cases} \mathbf{v}^{(i)}_{\alpha_1}\mathbf{v}^{(i)}_{\alpha_2} & \text{if } \alpha_1 = \alpha_2 \\ 2\mathbf{v}^{(i)}_{\alpha_1}\mathbf{v}^{(i)}_{\alpha_2} & \text{if } \alpha_1 < \alpha_2. \end{cases} \tag{12.7}$$

The first term of (12.6) becomes

$$\sum_{\alpha} \partial_\alpha f_i \ \mathrm{d}\mathbf{M}_\alpha(t) = \sum_{\alpha_1 \le \alpha_2} (2 - \delta_{\alpha_1\alpha_2}) \mathbf{v}^{(i)}_{\alpha_1} \mathbf{v}^{(i)}_{\alpha_2} \Big(\sigma_\alpha \mathrm{d}\mathbf{B}_\alpha(t) - \frac{1}{2} \mathbf{M}_\alpha(t) \mathrm{d}t \Big).$$

Summing the second term in the parenthesis yields

$$-\frac{1}{2} \sum_{\alpha_1 \le \alpha_2} (2 - \delta_{\alpha_1\alpha_2}) \mathbf{v}^{(i)}_{\alpha_1} \mathbf{v}^{(i)}_{\alpha_2} \mathbf{M}_\alpha(t) \mathrm{d}t = \frac{-1}{2} \Lambda_i(t) \mathrm{d}t. \tag{12.8}$$

For the first term, we define

$$\sqrt{2} \ \mathrm{d}\widetilde{B}_i(t) := \sum_{\alpha_1 \le \alpha_2} (2 - \delta_{\alpha_1\alpha_2}) \mathbf{v}^{(i)}_{\alpha_1} \mathbf{v}^{(i)}_{\alpha_2} \sigma_\alpha \mathrm{d}\mathbf{B}_\alpha(t).$$

The process $(\widetilde{B}_i(t), t \le n)$ is actually a n-dimensional standard Brownian motion. This can be seen by calculating $\mathrm{d}\widetilde{B}_i(t)\mathrm{d}\widetilde{B}_j(t)$ using the rules of stochastic calculus (Equation (D.1)). It is equal to $\delta_{ij}\mathrm{d}t$. Thus it must be a Brownian motion by Lévy's characterization theorem of Brownian motion (Theorem D.5).

It remains to compute the term second-order derivatives in (12.6). The good news is that $\mathrm{d}\mathbf{M}_\alpha(t)\mathrm{d}\mathbf{M}_\beta(t) = 0$ whenever $\alpha \neq \beta$ by the rules of stochastic calculus applied with the SDE (12.3). Therefore we are left with computing

$$\partial^2_\alpha f_i(\mathbf{M}(t)) = \partial_\alpha \Big((2 - \delta_{\alpha_1\alpha_2}) \mathbf{v}^{(i)}_{\alpha_1} \mathbf{v}^{(i)}_{\alpha_2} \Big) = (2 - \delta_{\alpha_1\alpha_2}) \Big\{ (\partial_\alpha \mathbf{v}^{(i)}_{\alpha_1}) \mathbf{v}^{(i)}_{\alpha_2} + \mathbf{v}^{(i)}_{\alpha_1} (\partial_\alpha \mathbf{v}^{(i)}_{\alpha_2}) \Big\}. \tag{12.9}$$

We need to evaluate $\partial_\alpha \mathbf{v}^{(i)}$, that is the derivative with respect to $\mathbf{M}_\alpha$ for each of the component of $\mathbf{v}^{(i)}$. Since the eigenvectors form an orthonormal basis of $\mathbb{R}^n$, we compute ${\mathbf{v}^{(j)}}^{\mathrm{T}} (\partial_\alpha \mathbf{v}^{(i)})$ for every $1 \le j \le n$. We apply ∂_α to the eigenvalue equation $\mathbf{M}\mathbf{v}^{(i)} = \Lambda_i \mathbf{v}^{(i)}$:

$$(\partial_\alpha \mathbf{M}) \mathbf{v}^{(i)} + \mathbf{M}(\partial_\alpha \mathbf{v}^{(i)}) = (\partial_\alpha \Lambda_i) \mathbf{v}^{(i)} + \Lambda_i (\partial_\alpha \mathbf{v}^{(i)}).$$

Projecting on $\mathbf{v}^{(j)}$, $i \neq j$ gives

$${\mathbf{v}^{(j)}}^{\mathrm{T}} (\partial_\alpha \mathbf{M}) \mathbf{v}^{(i)} + \Lambda_j {\mathbf{v}^{(j)}}^{\mathrm{T}} (\partial_\alpha \mathbf{v}^{(i)}) = \Lambda_i \ {\mathbf{v}^{(j)}}^{\mathrm{T}} (\partial_\alpha \mathbf{v}^{(i)}),$$

and thus

$${\mathbf{v}^{(j)}}^{\mathrm{T}} (\partial_\alpha \mathbf{v}^{(i)}) = \frac{1}{\Lambda_i - \Lambda_j} \ {\mathbf{v}^{(j)}}^{\mathrm{T}} (\partial_\alpha \mathbf{M}) \mathbf{v}^{(i)}.$$

The relation ${\mathbf{v}^{(i)}}^{\mathrm{T}} \mathbf{v}^{(i)} = 1$ implies ${\mathbf{v}^{(i)}}^{\mathrm{T}} (\partial_\alpha \mathbf{v}^{(i)}) = 0$. The above then gives the decomposition

$$\partial_\alpha \mathbf{v}^{(i)} = \sum_{j \neq i} \Big(\frac{1}{\Lambda_i - \Lambda_j} \ {\mathbf{v}^{(j)}}^{\mathrm{T}} (\partial_\alpha \mathbf{M}) \mathbf{v}^{(i)} \Big) \ \mathbf{v}^{(j)}.$$

Proceeding as in (12.7) yields

$$\partial_\alpha \mathbf{v}^{(i)} = \sum_{j \neq i} \Big(\frac{\mathbf{v}^{(j)}_{\alpha_1} \mathbf{v}^{(i)}_{\alpha_2} + \mathbf{v}^{(i)}_{\alpha_1} \mathbf{v}^{(j)}_{\alpha_2} (1 - \delta_{\alpha_1\alpha_2})}{\Lambda_i - \Lambda_j} \Big) \mathbf{v}^{(j)}.$$

Finally, putting this back in (12.9) leads us to

$$\partial^2_\alpha f_i(\mathbf{M}(t)) = (2 - \delta_{\alpha_1\alpha_2}) \sum_{j \neq i} \Big(\frac{\mathbf{v}^{(j)}_{\alpha_1} \mathbf{v}^{(i)}_{\alpha_2} + \mathbf{v}^{(i)}_{\alpha_1} \mathbf{v}^{(j)}_{\alpha_2} (1 - \delta_{\alpha_1\alpha_2})}{\Lambda_i - \Lambda_j} \Big) (\mathbf{v}^{(i)}_{\alpha_2} \mathbf{v}^{(j)}_{\alpha_1} + \mathbf{v}^{(i)}_{\alpha_1} \mathbf{v}^{(j)}_{\alpha_2}).$$

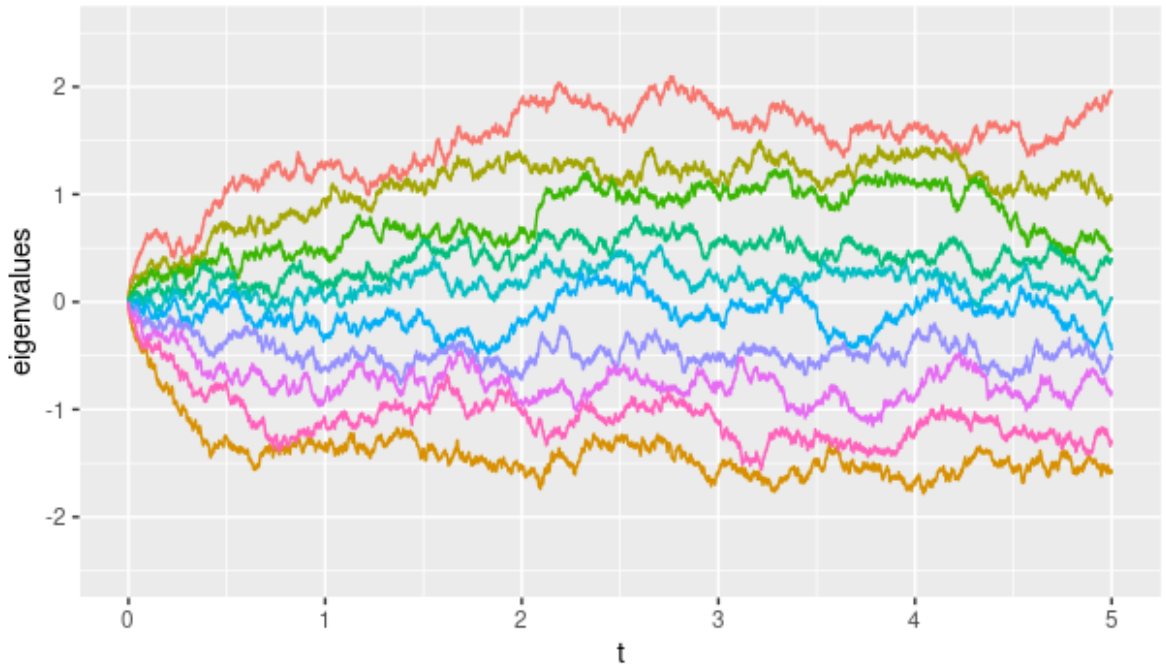

Figure 13: A simulation of GOE Dyson Brownian motion when $n = 10$. (Plot kindly provided by Johannes Forkel.)

Note that the expression is symmetric in α_1, α_2. In particular when summing over α, we can remove the restriction $\alpha_1 \leq \alpha_2$. Also, the rules of stochastic calculus imply $\mathrm{d}\mathbf{M}_\alpha(t)\mathrm{d}\mathbf{M}_\alpha(t) = \sigma_\alpha^2 \mathrm{d}t$. This is $\mathrm{d}t$ for $\alpha_1 < \alpha_2$ and $2\mathrm{d}t$ for $\alpha_1 = \alpha_2$. All in all, we get

$$\frac{1}{2} \sum_{\alpha_1 \leq \alpha_2} \partial_\alpha^2 f_i(\mathbf{M}(t)) \sigma_\alpha^2 \mathrm{d}t = \frac{1}{2} \sum_{\alpha_1, \alpha_2} \sum_{j \neq i} \frac{1}{\Lambda_i - \Lambda_j} (|\mathbf{v}_{\alpha_1}^{(j)}|^2 |\mathbf{v}_{\alpha_2}^{(i)}|^2 + |\mathbf{v}_{\alpha_1}^{(i)}|^2 |\mathbf{v}_{\alpha_2}^{(j)}|^2) \mathrm{d}t.$$

It remains to use the orthonormality of the eigenvectors to conclude the proof of the theorem. □

The properties of the Ornstein-Uhlenbeck process implies that if we start at $t = 0$ from some fixed matrix M_0, then at a later time t, $\mathbf{M}(t)$ is equal in distribution to the sum of $e^{-t/2} M_0$, and $\sqrt{1 - e^{-t}} G$, where G is a GOE matrix when $\beta = 1$ or a GUE matrix when $\beta = 2$. Hence for sufficiently large t, the GOE/GUE is the equilibrium solution which the dynamics reaches, no matter what M_0 is taken to be. Put another way, the GOE/GUE measure is invariant under Dyson Brownian motion, and is the attractor for the dynamics. It was conjectured by Dyson that this equilibrium is actually reached very quickly, on time scales of the order of $1/n$, and this has subsequently been proved. This fact plays a key role in the proof of the universality of the spectral statistics. We illustrate this by showing in Figures 13, 14 and 15 the results of a numerical simulation of the Ornstein-Uhlenbeck process with, respectively, $n = 10$, $n = 20$ and $n = 50$ for GOE matrices. The invariance of the GOE under this process is illustrated in Figures 16, 17 and 18, where M_0 is taken to be a GOE matrix in each case, again with, respectively, $n = 10$, $n = 20$ and $n = 50$.

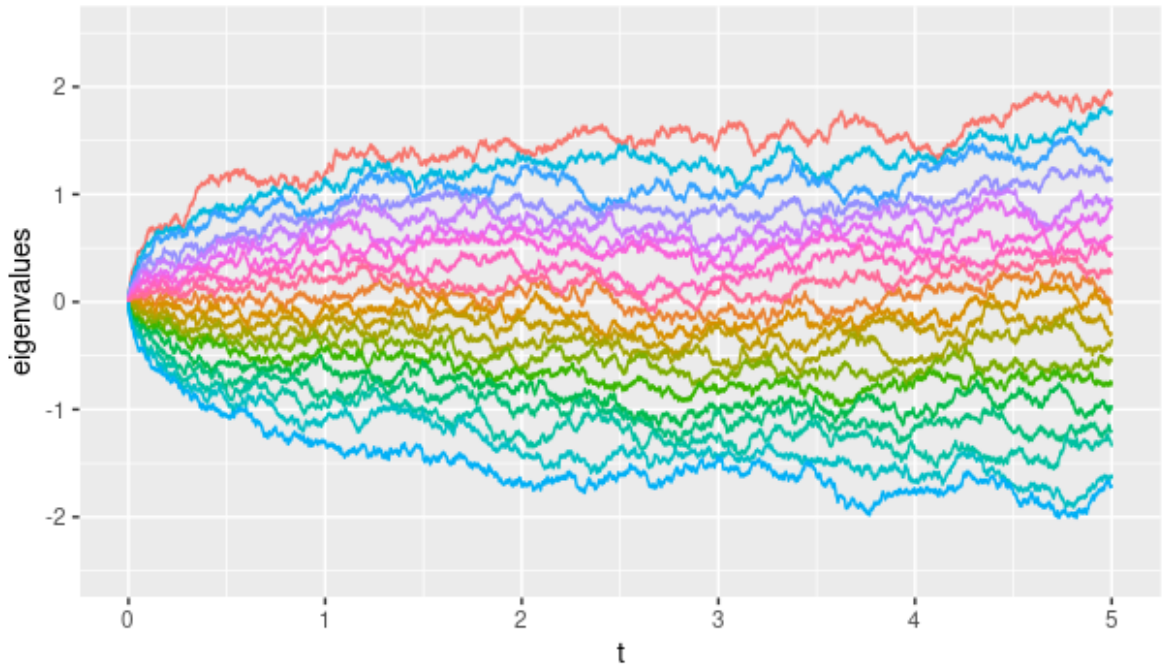

Figure 14: A simulation of GOE Dyson Brownian motion when $n = 20$. (Plot kindly provided by Johannes Forkel.)

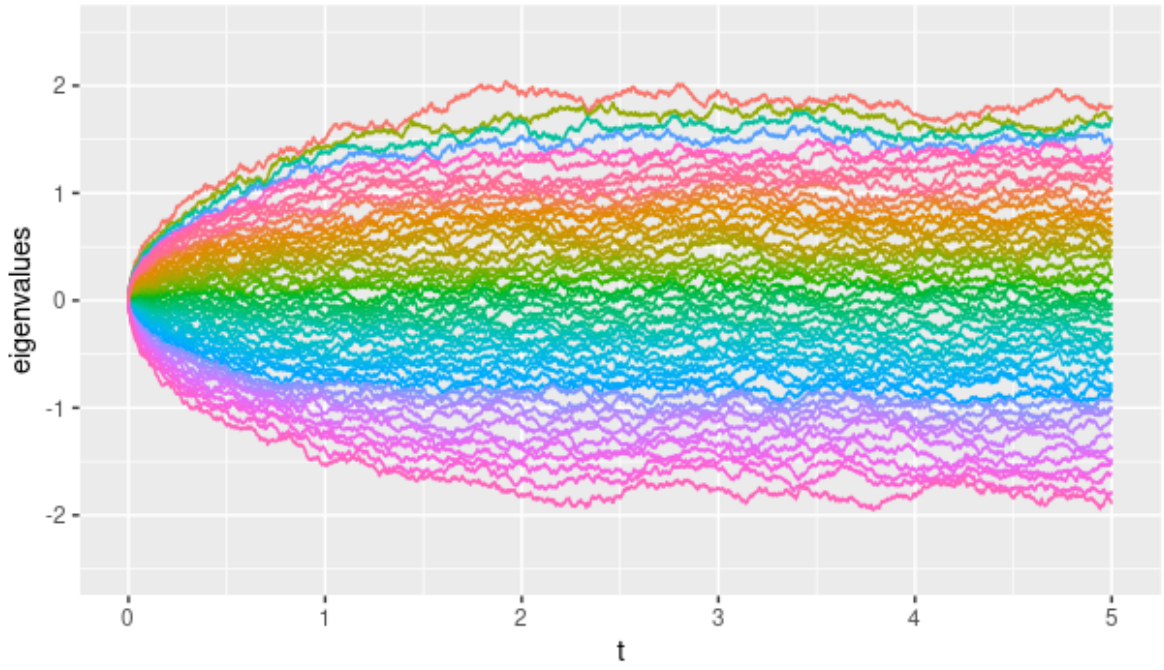

Figure 15: A simulation of GOE Dyson Brownian motion when $n = 50$. (Plot kindly provided by Johannes Forkel.)

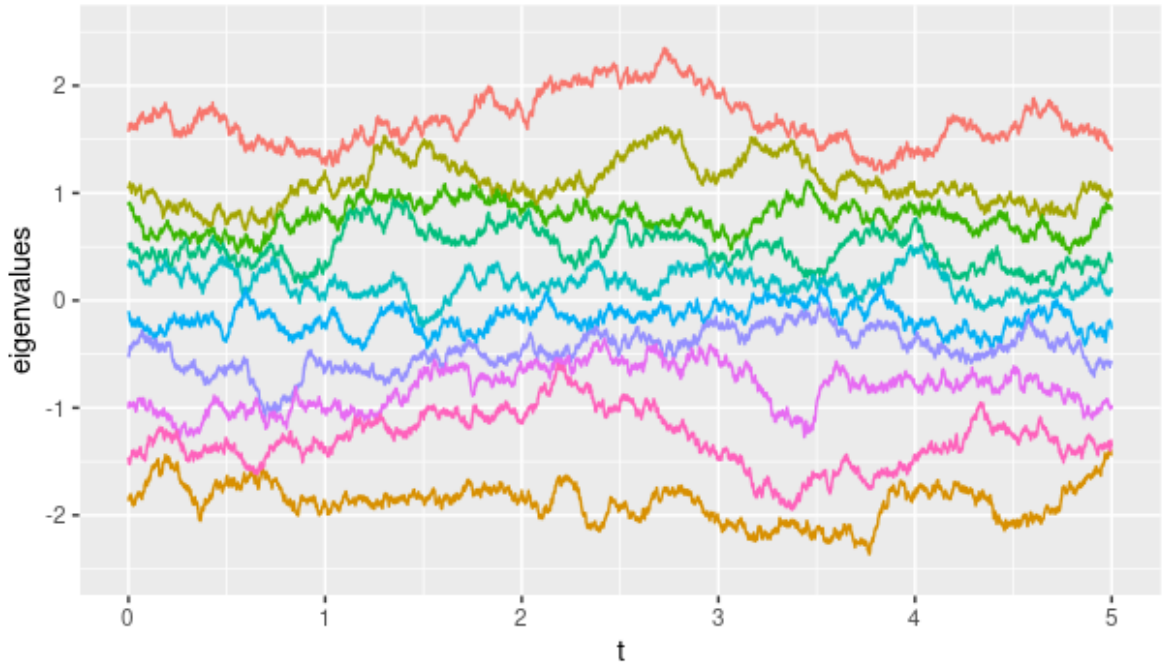

Figure 16: A simulation of GOE Dyson Brownian motion when $n = 10$, starting from a GOE matrix. (Plot kindly provided by Johannes Forkel.)

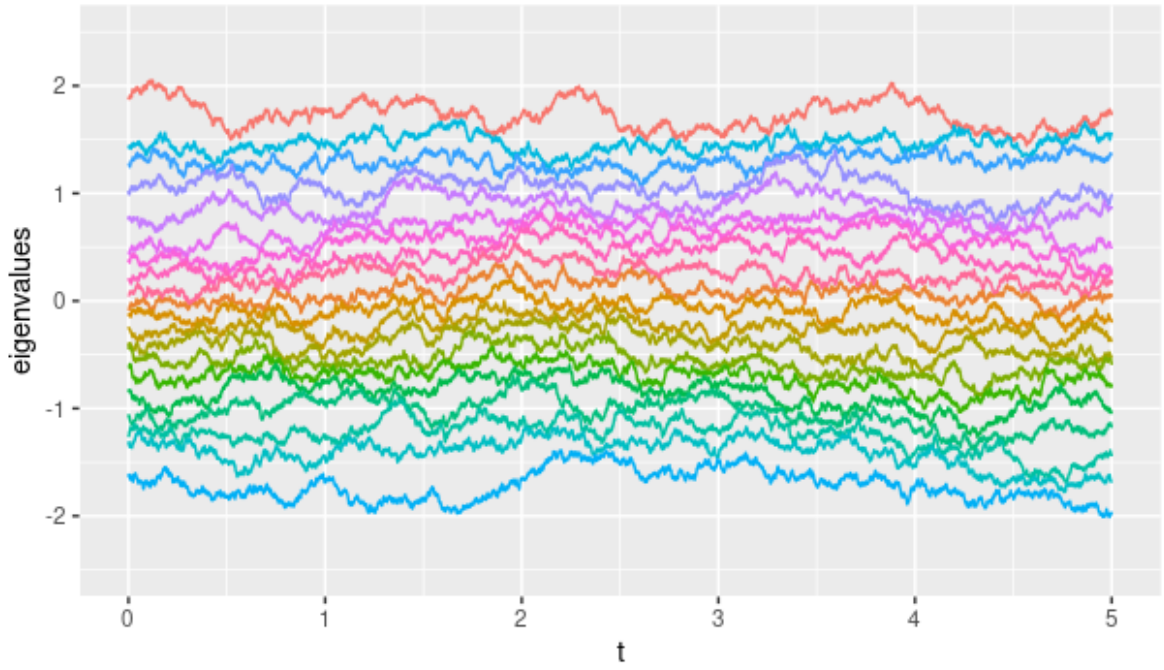

Figure 17: A simulation of GOE Dyson Brownian motion when $n = 20$, starting from a GOE matrix. (Plot kindly provided by Johannes Forkel.)

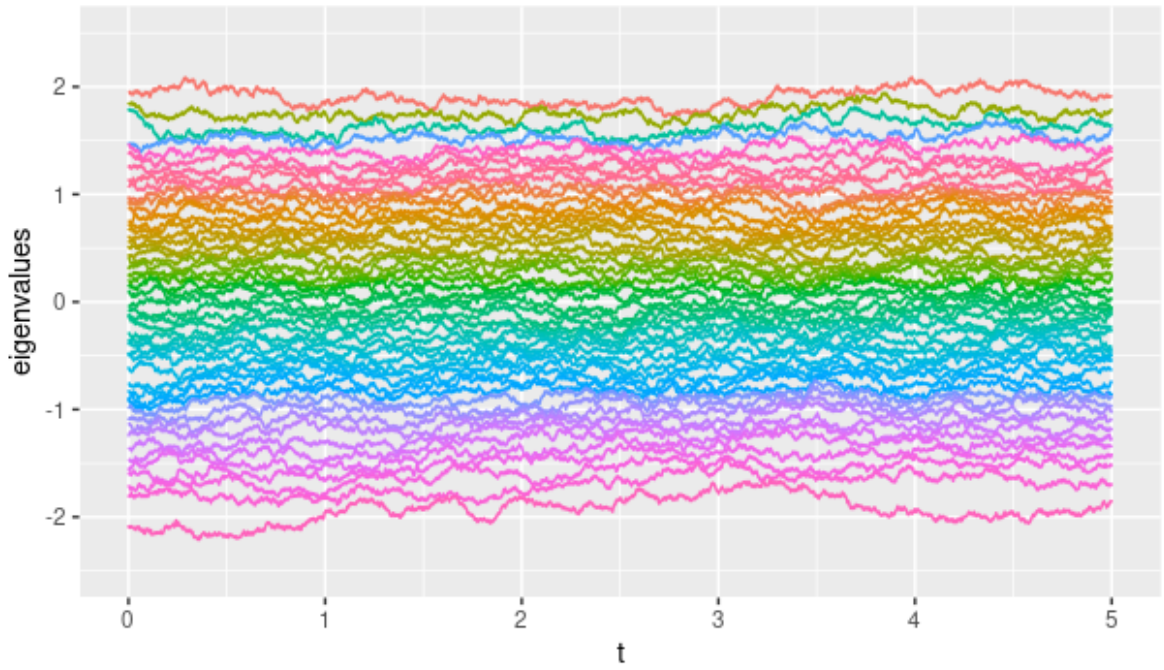

Figure 18: A simulation of GOE Dyson Brownian motion when $n = 50$, starting from a GOE matrix. (Plot kindly provided by Johannes Forkel.)

13 Connections with other areas of mathematics

We have already mentioned some of the applications of Random Matrix Theory, for example to data science, quantum mechanics, mathematical finance, population dynamics etc. There are also many beautiful and deep connections with other problems in mathematics, and we mention a few examples here.

13.1 Longest increasing subsequences

As discussed in the introduction, let $\mathcal{S}_n$ be the group of permutations of $1, 2, \ldots, n$. If $\pi \in \mathcal{S}_n$, $\pi(i_1), \ldots, \pi(i_k)$ is an increasing subsequence in π if $i_1 < i_2 < \cdots < i_k$ and $\pi(i_1) < \pi(i_2) < \cdots < \pi(i_k)$. Let $l_n(\pi)$ be the length of the longest increasing subsequence. For example, if $n = 5$ and π is the permutation 5 1 3 2 4, then the longest increasing subsequences are 1 2 4 and 1 3 4, and $l_n(\pi) = 3$. Equip $\mathcal{S}_n$ with uniform distribution,

$$\mathbf{P}(l_n \leq m) = \frac{\#\{\pi \in \mathcal{S}_n : l_n(\pi) \leq m\}}{n!}. \tag{13.1}$$

Then it was proved by Baik, Deift and Johansson in 1998 [BDJ99] that

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{l_n - 2\sqrt{n}}{n^{1/6}} \leq s\right) = F^{(\max)}(s) = \det(\mathrm{I} - K_{\mathrm{Airy}}) \tag{13.2}$$

where

$$K_{\mathrm{Airy}} = \frac{\mathrm{Ai}(x)\mathrm{Ai}'(y) - \mathrm{Ai}'(x)\mathrm{Ai}(y)}{(x-y)} \tag{13.3}$$

acting on $L^2((s,\infty))$. Or equivalently,

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{l_n - 2\sqrt{n}}{n^{1/6}} \le s\right) = \exp\left(-\int_s^\infty (x-s)q(x)^2 \mathrm{d}x\right) \tag{13.4}$$

where q is a solution of the Painlevé II equation $q''(x) - xq(x) + 2q(x)^3 = 0$ with $q(x) \sim \mathrm{Ai}(x)$ as $x \to \infty$. That is, l_n has the same limiting distribution as the largest eigenvalue of a random GUE matrix.

13.2 Stochastic growth models, random tilings, and random paths

There has been a considerable focus in the past few years on the geometrical properties of structures grown or generated by simple random processes. One example involves random tilings of Aztec diamonds, and a second involves randomly grown surfaces, such as by a random deposition of atoms. These are generated by simple probabilistic rules. We do not give the details, but show an example of a random tiling in Figure 19. The outside can be thought of as a 'frozen' phase, and the interior a 'liquid' phase. The boundary of the liquid phase is a random curve whose fluctuations are again the same as those found in the largest eigenvalue of a GUE matrix. This models laboratory experiments involving growing interfaces of liquid-crystal turbulence, where the fluctuations of the interface are found to match those of the largest eigenvalue of a GUE matrix rather well. For more details (and a movie) see [TSSS11]. For more about random growth models and random tilings, see the article by P. Ferrari and H. Spohn [FS11].

One finds the same again in models of random paths. For example, the Hammersley process refers to the following stochastic model. In the unit square mark in points uniformly at random according to a Poisson point process with intensity α. We call a path from (0, 0) to (1, 1) through these random points *up/right* if the points it passes through have coordinates $x_k \le x_{k+1}$ and $y_k \le y_{k+1}$ for each k. Let $L(\alpha)$ denote the maximum number of points on such a path. Then for all $s \in \mathbb{R}$,

$$\lim_{\alpha\to\infty} \mathbf{P}\left(\frac{L(\alpha) - 2\sqrt{\alpha}}{\alpha^{1/6}} \le s\right) = F^{(\max)}(s) = \det(\mathrm{I} - K_{\mathrm{Airy}}). \tag{13.5}$$

13.3 Zeros of the Riemann zeta-function

The Riemann zeta-function, $\zeta(s)$ is defined when Re $s > 1$ by

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}, \tag{13.6}$$

where the product runs over all primes p. It has an analytic continuation to the rest of the complex plane, except for a pole at $s = 1$. It is important because it encodes information about the distribution of the primes. The zeta function has *trivial zeros* at $s = -2n$ for $n \in \mathbb{N}$, and infinitely many other *nontrivial zeros*. The *Riemann Hypothesis* asserts that the nontrivial zeros all lie on the line Re$s = 1/2$; that is, they are all of the form $1/2 + \mathrm{i}t_n$ with $t_n \in \mathbb{R}$.

Let us assume that the Riemann Hypothesis is true and so the numbers t_n are all real. The theory of the zeta function then implies that

$$\#\{n : 0 < t_n \le T\} = \frac{T}{2\pi} \log\left(\frac{T}{2\pi}\right) + O(\log T). \tag{13.7}$$

Therefore, setting

$$w_n = \frac{t_n}{2\pi} \log\left(\frac{|t_n|}{2\pi}\right), \tag{13.8}$$

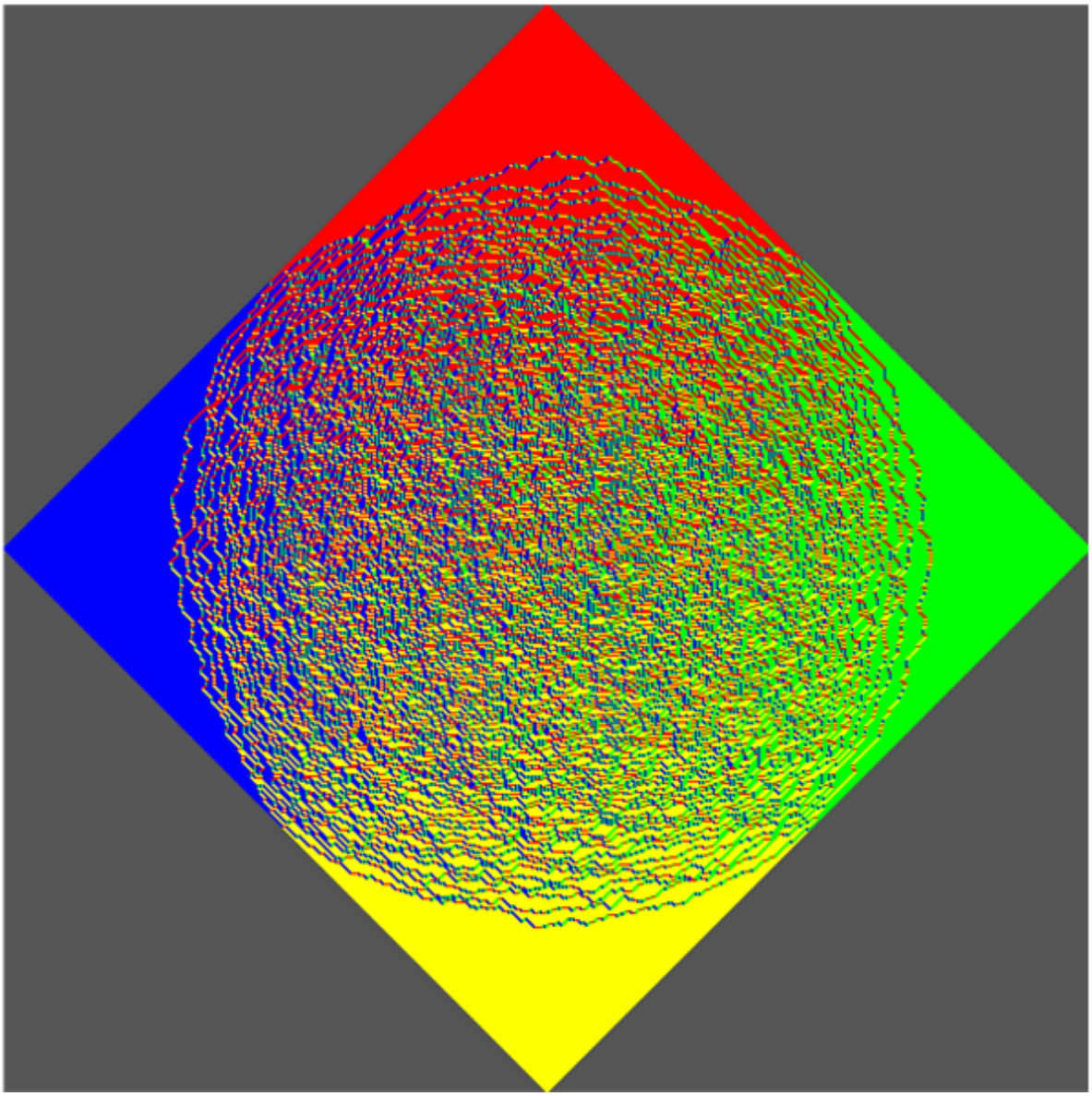

Figure 19: A random tiling, generated by P. Ferrari.

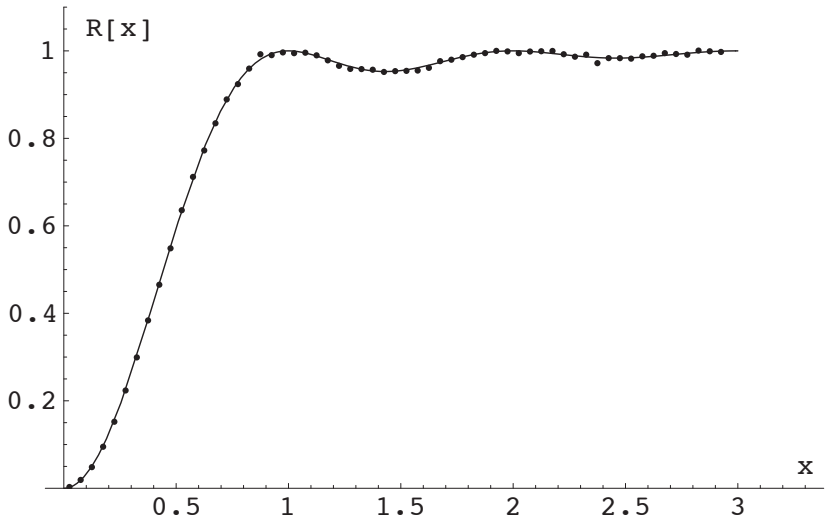

Figure 20: The 2-point correlation function for the Riemann zeros, computed by A. Odlyzko for zeros near to the 10^{20}th, compared to the GUE 2-point correlation function (13.11).

we have that the mean density of the numbers w_n is asymptotically 1, in that

$$\lim_{W\to\infty} \frac{1}{W} \#\{n : 0 < w_n \leq W\} = 1. \tag{13.9}$$

Consider now the pair correlation function of the scaled zeros w_n, defined, assuming the limit exists, by

$$\lim_{N\to\infty} \frac{1}{N} \sum_{1\leq n<m\leq N} f(w_n - w_m) = \int_{-\infty}^{\infty} f(x) R_2(x) \mathrm{d}x \tag{13.10}$$

It is a theorem due to Montgomery in 1973 [Mon73] that, assuming the Riemann Hypothesis, for functions $f(x)$ whose Fourier transform has support in $(-1, 1)$ the limit in the left hand-side exists and then

$$R_2(x) = 1 - \left(\frac{\sin \pi x}{\pi x}\right)^2 \tag{13.11}$$

which is precisely the GUE/CUE 2-point correlation function. Montgomery conjectured that this remains true for all functions f for which the sums converge. This has since been extended to all k-point correlation functions, where the result and conjecture lead to the $k \times k$ determinant of the sine kernel. It is therefore currently the belief that that all local statistics of the zeros coincide with those of the GUE/CUE, and this is supported by extensive numerical computations – see for example Figure 20. Assuming this is true, many interesting properties of the zeta function and the primes can be calculated using random matrix theory. For more on connections between number theory and random matrix theory, see [KS11]. This remains a highly active area of research.

A Linear Algebra

The *spectral theorem* is a central result in linear algebra. A common example is given by real symmetric matrices. If M is symmetric, then M is *diagonalizable* in the sense that we can write

$$M = ODO^{\mathrm{T}},$$

where D is a diagonal matrix with the eigenvalues of M as its entries. The matrix O is orthogonal and is constructed by putting the eigenvectors of M as its columns. The j-th column of O is the eigenvector of the eigenvalues appearing in the j-th entry of D. Note that the decomposition is unique up to permutation of the eigenvectors. In words, the decomposition means that we can find an orthonormal basis of $\mathbb{R}^n$ where the operator M is diagonal (so it acts as a multiplication operator). It turns out that the spectral theorem applies more generally to matrices that commute with their conjugate transpose.

Theorem A.1 (The Spectral Theorem. See for example Theorem 2.5.3 in [HJ85])**.** *Let M be normal matrix, that is*

$$M^\dagger M = MM^\dagger.$$

Then M is diagonalizable, that is, there exists a unitary matrix U and a diagonal matrix D such that

$$M = UDU^\dagger.$$

Note that the theorem applies in particular to real symmetric matrices, Hermitian matrices and unitary matrices. A simple consequence of the spectral theorem is the so-called *functional calculus*. Indeed, if M is diagonalizable, we have that $M^k = UD^kU^\dagger$ for any integer k. And therefore applying M^k to an eigenvector v with eigenvalue λ simply gives $\lambda^k v$. We can therefore make sense of the operator $f(M)$ by taking that $f(M)$ on v gives $f(\lambda)v$ for any reasonable f. This point of view generalizes very nicely to infinite-dimensional space. For an introduction to this, see [Hal63].

The eigenvalues of a Hermitian matrices and the ones of its minors have this important relation.

Theorem A.2 (Cauchy's Interlacing Theorem)**.** *Let A be an $n \times n$ Hermitian matrix with eigenvalues $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$, and let B be an $(n-1) \times (n-1)$ principal submatrix of A with eigenvalues $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_{n-1}$. Then $\alpha_1 \leq \beta_1 \leq \alpha_2 \leq \beta_2 \leq \cdots \leq \alpha_{n-1} \leq \beta_{n-1} \leq \alpha_n$.*

See for example Theorem 4.3.17 in [HJ85].

B Probability Framework

B.1 Basic Concepts

We refer to [Dur10] for more information on probability theory.

In probability theory, a random experiment is modelled by a *probability space* $(\Omega, \mathcal{F}, \mathbf{P})$. The set Ω is the *sample space*, i.e., the set of outcomes. Events are subsets of Ω. Here, $\mathcal{F}$ is called a *sigma-field* or *sigma-algebra*. It is a collection of events that contains Ω, and is closed under complement and countable union. A *probability* or *probability measure* $\mathbf{P}$ is a function that associate to an event $A \in \mathcal{F}$ a number in $[0, 1]$. This function is such that $\mathbf{P}(\Omega) = 1$ and $\mathbf{P}(\cup_{n\geq 1} A_n) = \sum_{n\geq 1} \mathbf{P}(A_n)$ whenever A_n is a countable sequence of disjoint events.

A random variable X is a function from $\Omega \to \mathbb{R}$. To ensure that the probability of the events $\{\omega \in \Omega : X(\omega) \in B\} = \{X \in B\}$ for any open set B is well defined, we ask that X is a *measurable* function meaning that $\{X \in B\} \in \mathcal{F}$ for any open set B. The *distribution* of a random variable X is a probability μ_X on $\mathbb{R}$ defined by

$$\mu_X(B) = \mathbf{P}(X \in B).$$

The distribution is well-defined on any $B \subset \mathbb{R}$ contained in $\mathcal{B}(\mathbb{R})$, the smallest sigma-algebra of $\mathbb{R}$ containing the open sets. The distribution of X is determined by its *Cumulative Distribution Function* (CDF) defined by

$$\begin{aligned} F_X : \mathbb{R} &\to \mathbb{R} \\ x &\mapsto F_X(x) = \mu_X((-\infty, x]) = \mathbf{P}(X \leq x). \end{aligned}$$

The CDF is by definition non-decreasing and right-continuous. If it is continuous at all points, we say that X has a continuous distribution. In the case where F_X is of the form

$$F_X(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y,$$

for some non-negative function f, we say that X has a density and the function f is called the *probability density function* (PDF) or *density* for short. In this case, we have that

$$\mathbf{P}(X \in B) = \int_B f(y)\mathrm{d}y \qquad \mathbf{E}[g(X)] = \int_{\mathbb{R}} g(y)f(y)\mathrm{d}y,$$

for any measurable function $g : \mathbb{R} \to \mathbb{R}$ such that the expectation of g is well-defined.

Example B.1. We say that X has a *normal* or *Gaussian* distribution with mean m and variance σ^2 if it has a PDF f given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-m)^2/(2\sigma^2)}.$$

We then sometimes write $X \sim \mathcal{N}(m, \sigma^2)$. The distribution is said to be standard if $m = 0$ and $\sigma^2 = 1$. Note that if X is $\mathcal{N}(m, \sigma^2)$ then $(X - m)/\sigma$ is a standard Gaussian random variable.

A simple yet important property (it is in fact a defining property) is the following

Lemma B.2. *: Let Z by a standard Gaussian random variable. Consider a function differentiable function $g : \mathbb{R} \to \mathbb{R}$ such that $\mathbf{E}[g'(Z)]$ and $\mathbf{E}[Zg(Z)]$ is well-defined. Then*

$$\mathbf{E}[Zg(Z)] = \mathbf{E}[g'(Z)].$$

The above property is easily generalized to multivariate Gaussians.

A *d-dimensional random vector* $\mathbf{X} = (X_1, \ldots, X_d)$ is a random point in $\mathbb{R}^d$. More precisely, it is a measurable function from $\Omega \to \mathbb{R}^d$. One can think of the random variables X_i as the i-th coordinate of $\mathbf{X}$. The distribution of $\mathbf{X}$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}^d)$ defined as in the one-dimensional case by $\mu_{\mathbf{X}}(B) = \mathbf{P}(\mathbf{X} \in B)$. We say that the random vector has a PDF f if

$$\mu_{\mathbf{X}}(B) = \int_B f(x)\mathrm{d}x_1 \ldots \mathrm{d}x_d,$$

where $B \in \mathcal{B}(\mathbb{R}^d)$ and $x = (x_1, \ldots, x_d)$.

Example B.3. A d-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)$ is said to be Gaussian or have a multivariate normal distribution with mean vector m and covariance matrix $\mathcal{C}$ if it has a PDF of the form

$$f(x) = \frac{1}{(2\pi)^{d/2}(\det C)^{1/2}} e^{-\frac{1}{2}(x-\mathrm{m})^{\mathrm{T}}\mathcal{C}^{-1}(x-\mathrm{m})}, \quad x = (x_1, \ldots x_d).$$

The covariance matrix is $\mathcal{C}_{ij} = \mathrm{Cov}(X_i, X_j)$ and the mean vector is $m_i = \mathbf{E}[X_i]$. Note that the PDF is defined only if $\det \mathcal{C} \neq 0$. We then say that the vector is non-degenerate. It is not hard to prove that a Gaussian vector is non-degenerate if and only if its coordinates are linearly independent (i.e., some rigid linear relations between coordinates).

A useful result is the following IID decomposition of Gaussian vectors. This is the generalization of the standardization in the $d = 1$ case.

Lemma B.4. *Let $\mathbf{X}$ be a non-degenerate Gaussian random vector with mean 0 and covariance matrix $\mathcal{C}$. Then there exists an invertible matrix $A : \mathbb{R}^d \to \mathbb{R}^d$ and IID standard Gaussians $\mathbf{Z} = (Z_1, \dots Z_d)^{\mathrm{T}}$ such that*

$$\mathbf{X} = A\mathbf{Z}.$$

Moreover, $\mathcal{C} = AA^{\mathrm{T}}$. The identity $\mathcal{C} = AA^{\mathrm{T}}$ is called the Cholesky decomposition. Note that the matrix A is not unique.

Let $\mathbf{X}$ be some d-dimensional random vector with PDF $f_{\mathbf{X}}$. Suppose we construct another random vector $\mathbf{Y}$ by taking $\mathbf{Y} = g(\mathbf{X})$ for some measurable function $g : \mathbb{R}^d \to \mathbb{R}^d$. Then the expectation of $\mathbf{Y}$ is given by

$$\mathbf{E}[g(\mathbf{X})] = \int_{\mathbb{R}^d} g(x) f(x) \mathrm{d}x_1 \dots \mathrm{d}x_n.$$

If g is smooth and with a well-defined inverse g^{-1}, then $\mathbf{Y}$ also has a density. Indeed, we have with the change of variable $y = g(x)$ that

$$\mu_{\mathbf{Y}}(B) = \int_{\{x \in \mathbb{R}^d : g(X) \in B\}} f_X(x) \mathrm{d}x = \int_B f_{\mathbf{X}}(g^{-1}(y)) \left| \frac{\partial x}{\partial y} \right| \mathrm{d}y,$$

where we write the Jacobian matrix of the transformation g as $\left[\frac{\partial x}{\partial y}\right]_{ij} = \frac{\partial x_i}{\partial y_j}$. We deduce that the PDF of $\mathbf{Y}$ is

$$f_{\mathbf{Y}}(y) = f_{\mathbf{X}}(g^{-1}(y)) \left| \frac{\partial x}{\partial y} \right|. \tag{B.1}$$

Note that we need to be careful when using this formula as g might be invertible only locally.

B.2 Limit Theorems

We are often interested in sequences of random variables defined on the same probability space. There are several notions of convergence in this case.

Definition B.5. Let $(X_n, n \geq)$ be a sequence of random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

- X_n converges *almost surely* (a.s., for short) to a random variable X if and only if

  $$\mathbf{P}(\{\omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\}) = 1.$$

  In other words, this is pointwise convergence on a set of probability one.

- X_n converges *in probability* if and only if for any $\varepsilon > 0$, we have

  $$\lim_{n \to \infty} \mathbf{P}(|X_n - X| > \varepsilon) = 0.$$

- X_n converges *in L^2* (or in mean-square) to a random variable X if and only if

  $$\lim_{n \to \infty} \mathbf{E}[(X_n - X)^2] = 0.$$

- X_n converges *in distribution* to a random variable X if and only if for any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$

  $$\lim_{n \to \infty} \mathbf{E}[f(X_n)] = \mathbf{E}[f(X)].$$

Note that for convergence of distribution the requirement that the random variables are defined on the same space is not necessary. There is also equivalent definition of this convergence in terms of the CDF and the characteristic function. A fundamental example of convergence in distribution is:

Theorem B.6 (Central Limit Theorem)**.** *Let $(X_n, n \geq 1)$ be IID random variables with $\mathbf{E}[X_1] = 0$ and $\mathbf{E}[X_1^2] = 1$. We have*

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j = Z \text{ in distribution,}$$

where Z is a standard Gaussian random variable.

The Central Limit Theorem is a perfect example of a universality result. It is a limit theorem of a functional of random variables where the limiting random variables depend minimally on the specificity of their distribution.

The notions of convergence are obviously not equivalent. There are some implications though. Convergence in L^2 is easily seen to imply convergence in probability by Chebyshev's inequality. Moreover, convergence almost surely implies convergence in probability by the following important theorem

Theorem B.7 (Dominated Convergence Theorem)**.** *Let $(X_n, n \geq 1)$ be a sequence of random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$ such that $\lim_{n\to\infty} X_n = X$ almost surely for some random variable X. Suppose that there exists a random variable Y with $\mathbf{E}[|Y|] < \infty$ such that $|X_n| \leq Y$ (with probability one). Then*

$$\lim_{n\to\infty} \mathbf{E}[X_n] = \mathbf{E}[X].$$

Convergence in probability does not imply convergence almost surely in general. However, it can be the case with some extra conditions. The main tool here is:

Lemma B.8 (Borel-Cantelli Lemma I)**.** *Let $(E_n, n \geq 1)$ be a sequence of (measurable) events in Ω such that $\sum_{n\geq 1} \mathbf{P}(E_n) < \infty$. Then*

$$\mathbf{P}\Big(\{\omega : \omega \in E_n \text{ for infinitely many } n\}\Big) = 1.$$

As a corollary, we get

Corollary B.9. *Consider a sequence of random variables $(X_n, n \geq 1)$ that converges to X in probability so fast that for any $\varepsilon > 0$,*

$$\sum_{n\geq 1} \mathbf{P}(|X_n - X| > \varepsilon) < \infty.$$

Then $X_n \to X$ almost surely.

B.3 Method of Moments

We say that the distribution of a random variable X is determined by its moments if we can write its characteristic function (or its moment generating function (MGF)) as a sum over its moments, that is,

$$\mathbf{E}[e^{\mathrm{i}\lambda X}] = \sum_{k\geq 0} \frac{\mathrm{i}^k \mathbf{E}[X^k]}{k!}.$$

Note that this equality is not obvious since we are interchanging an infinite sum and an expectation! However, if the moments grow not too fast, then one can invoke the dominated convergence theorem to get the inequality.

To prove convergence in distribution of a sequence of random variables, the following is very useful.

Theorem B.10 (Method of Moments, see for example Section 30.2 in [Bil95]). *Let $(X_n, n \geq 1)$ be a squence of random variables, and consider X a random variable whose distribution is determined by its moments. Then, if for every $k \in \mathbb{N}$ we have*

$$\lim_{n\to\infty} \mathbf{E}[X_n^k] = \mathbf{E}[X^k],$$

then X_n converges to X in distribution.

The Gaussian distribution is determined by its moments (check!). Therefore, a common way (robust but not the most straightforward) to prove the Central Limit Theorem is by the method of moments.

C Concentration of Measure

The phenomenon of concentration of measure is now a fundamental concept in probability, analysis, and mathematical physics. It pertains to probability measures in high dimensions. Michel Talagrand was awarded the 2024 Abel Prize for major advances in this topic. For more on the subject, we suggest Ledoux's fascinating monograph [Led01]. We cite only two important concentration inequalities.

Theorem C.1. *Let $Z = (Z_1, \ldots, Z_n)$ be IID standarc Gaussian random variables. Consider $h : \mathbb{R}^n \to \mathbb{R}$ a Lipshitz function with constant $\sigma := \|h\|_{Lip}$. Then*

$$\mathbf{P}(|h(Z) - \mathbf{E}[h(Z)]| > u) \leq 2e^{-u^2/4\sigma^2} \ .$$

Note that the smaller the Lipshitz constant is, the better is the concentration. In some sense, the smaller the Lipshitz constant, the more h *contracts* its values.

Theorem C.2 (McDiarmid's inequality). *Let $X_1, \ldots, X_n$ be independent random variables taking values in ranges $R_1, \ldots, R_n$ and let $F : R_1 \times \cdots \times R_n \to \mathbb{C}$ be a function having bounded differences, so that, there exist constants $c_1, \ldots, c_n$ such that for all i,*

$$|F(x_1, \ldots, x_n) - F(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i \tag{C.1}$$

for all $x_i, x_i' \in R_i$. Then for any $\kappa > 0$

$$\mathbf{P}\Big(|F(X_1, \ldots, X_n) - \mathbf{E}F(X_1, \ldots, X_n)| \geq \kappa\sigma\Big) \leq C\exp(-c\kappa^2) \tag{C.2}$$

for some $C > 0$ and $c > 0$, independent of F and n, where $\sigma^2 = \sum_{i=1}^n c_i^2$.

D Brownian Motion

We refer to [Arg22] for more on Brownian motion and stochastic calculus in general. For the fascinating history of Brownian motion in mathematics and physics, see [Nel67].

Definition D.1. A *standard Brownian motion* $(B(t), t \geq 0)$ is a $\mathbb{R}$-valued stochastic process such that

- $B(0) = 0$ with probability 1;
- for any $n \in \mathbb{N}$ and choices of $0 \leq t_1 < t_2 < \cdots < t_n$, the increments $B_{t_{j+1}} - B_{t_j}$, $0 \leq j \leq n-1$, are independent with Gaussian distribution of mean 0 and variance $t_{j+1} - t_j$.
- the function $t \mapsto B_t$ is continuous with probability 1.

A d-dimensional standard Brownian motion $(\mathbf{B}(t), t \geq 0)$ is the $\mathbb{R}^d$-valued process $\mathbf{B}(t) = (B_1(t), \ldots, B_d(t))$ where $B_j(t)$, $1 \leq j \leq d$, are IID standard Brownian motions.

It is not a trivial fact that such a process exists! The first proof was given by Wiener, and sometimes Brownian motion is sometimes called the *Wiener process*. The Brownian paths are continuous functions with probability one, but they are quite rough. It is possible to prove that they are nowhere differentiable. It is therefore hard to make sense of integration the usual way using Riemann sums

$$\lim_{n\to\infty}\sum_{j=1}^{n} f(t_j)(B_{t_{j+1}} - B_{t_j}) = \int_0^T f(t)\mathrm{d}B(t).$$

If the Brownian paths were differentiable, we would have $B_{t_{j+1}} - B_{t_j} \approx B'_{t_j}(t_{j+1} - t_j)$. Wiener gave the first construction of the integral over Brownian motion, and Itô generalized it substantially to include random integrands. This theory is now known as *Itô stochastic calculus*. The fundamental theorem of classical calculus states that

$$\int_a^b f'(t)\mathrm{d}t = f(b) - f(a),$$

or $\mathrm{d}f(t) = f'(t)\mathrm{d}t$ in differential notation. For Itô calculus, the theorem takes the following form:

Theorem D.2 (Itô's Formula)**.** *Let $f \in \mathcal{C}^2(R)$ and $(B(t), t \geq 0)$, a standard Brownian motion. Then the process $(f(B_t), t \geq 0)$ satisfies*

$$\mathrm{d}f(B(t)) = f'(B(t))\mathrm{d}B(t) + \frac{1}{2}f''(B(t))\mathrm{d}t.$$

The proof follows from a Taylor expansion to second order. It is necessary to go to second order since the Brownian paths have infinite variation but finite *quadratic variation*. In integral form, the theorem says that

$$f(B(t)) - f(B(0)) = \int_0^t f'(B(s))\mathrm{d}B(s) + \frac{1}{2}\int_0^t f''(B(s))\mathrm{d}s.$$

The good way to remember Itô's formula is to apply the differential to $f(B(t))$ as we would do in the classical case, but do it to the second order. This explains the appearance of the $1/2$. We then apply the rules of stochastic calculus:

$$\mathrm{d}B(t)\mathrm{d}B(t) = \mathrm{d}t \qquad \mathrm{d}t\mathrm{d}B(t) = 0 \qquad \mathrm{d}t\mathrm{d}t = 0. \tag{D.1}$$

(This makes sense in terms of quadratic variation.)

Itô's formula has a nice interpretation. The process $(f(B(t)), t \geq 0)$ constructed from a Brownian motion varies in time as follows

$$f(B(t+\Delta t)) - f(B(t)) \approx f'(B(t))(B(t+\Delta t)) - B(t)) + \frac{1}{2}f''(B(t))\Delta t.$$

The local variation is due to a *drift* given by $\frac{1}{2}f''(B(t))$ and a noisy term with normal distribution of mean 0 and variance Δt times the *volatility* $f'(B(t))$.

We can generalize this by constructing processes whose drift and volatility (or diffusion coefficient) is a function of the position of the process itself at time t.

Definition D.3. A process $(X(t), t \geq 0)$ constructed from a Brownian motion is called a (time-homogeneous) *diffusion* if it satisfies the following *stochastic differential equation* (SDE)

$$\mathrm{d}X(t) = \sigma(X(t))\mathrm{d}B(t) + \mu(X(t))\mathrm{d}t.$$

The functions $\sigma : \mathbb{R} \to (0, \infty)$ and $\mu : \mathbb{R} \to \mathbb{R}$ are called the diffusion/volatility and the drift, respectively.

Itô's formula can be generalized to diffusions (and Itô processes more generally) by taking for a smooth function f

$$\mathrm{d}f((X(t)) = f'(X(t))\mathrm{d}X(t) + \frac{1}{2}f''(X(t))\ \mathrm{d}X(t)\mathrm{d}X(t), \tag{D.2}$$

and applying the rules of stochastic calculus to compute $\mathrm{d}X(t)\mathrm{d}X(t)$.

An important example of diffusion is the *Ornstein-Uhlenbeck process*.

Definition D.4. An *Ornstein-Uhlenbeck process* $(O(T), t \geq 0)$ with constant parameter $\sigma > 0$ and $k > 0$ is a stochastic process with SDE

$$\mathrm{d}O(t) = \sigma \mathrm{d}B(t) - kO(t)\mathrm{d}t,$$

If $O(0) = x_0$, it is not hard to show that the distribution of $O(t)$ at a fixed t is Gaussian with mean $x_0 e^{-t}$ and variance $\frac{\sigma^2}{2k}(1 - e^{-2kt})$. More generally, the process $(O(t), t \geq 0)$ is a Gaussian process of mean $x_0 e^{-t}$ and covariance $\mathrm{Cov}(O(s), O(t)) = \frac{\sigma^2}{2k}(e^{-k(t-s)} - e^{-k(t+s)})$. We see two important properties. First, in the limit of large times, the process converges to a Gaussian process of mean 0 and covariance $\frac{\sigma^2}{2k}e^{-k(t-s)}$. This is the invariant distribution of the process. Moreover, when starting at time 0 from the invariant distribution, the process is then *stationary* since the whole process distribution is invariant under a time-shift.

Itô's formula easily generalizes to the case of d-dimensional diffusions constructed from Brownian motion $\mathbf{X}(t) = (X_1(t), \ldots, X_d(t))$ where X_j, $j \leq d$, are diffusions (not necessarily independent) constructed from a d-dimensional standard Brownian motion $\mathbf{B}(t) = (B_1(t), \ldots, B_d(t))$. We get for a function $f \in \mathcal{C}^2(\mathbb{R})^d$ that the process $(f(\mathbf{X}(t)), t \geq 0)$ satisfies

$$\mathrm{d}f((\mathbf{X}(t)) = \sum_{1 \leq i \leq d} \partial_i f(\mathbf{X}(t))\mathrm{d}B_i(t) + \frac{1}{2}\sum_{1 \leq i,j \leq d} \partial_i \partial_j f(\mathbf{X}(t))\ \mathrm{d}X_i(t)\mathrm{d}X_j(t), \tag{D.3}$$

where ∂_i stands for the partial derivative with respect to the i-th coordinate.

We have the following important characterization of Brownian motion.

Theorem D.5 (Lévy's Characterization Theorem)**.** *Let $(X(t), t \geq 0)$ be a process constructed from a d-dimensional Brownian motion $(\mathbf{B}(t), t \geq 0)$ of the form*

$$\mathrm{d}X(t) = \sum_{j \leq d} Y_j(s)\mathrm{d}B_j(s),$$

where the $Y(s)$'s may depend on the Brownian motion before time s. If $\mathrm{d}X(t)\mathrm{d}X(t) = \mathrm{d}t$, then $X(t)$ is also a Brownian motion.

References

[ABDF11] Gernot Akemann, Jinho Baik, and Philippe Di Francesco, editors. *The Oxford handbook of random matrix theory*. Oxford University Press, Oxford, 2011.

[AGZ10] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.

[Arg22] Louis-Pierre Arguin. *A first course in stochastic calculus*, volume 53 of *Pure and Applied Undergraduate Texts*. American Mathematical Society, Providence, RI, [2022] ©2022. With a foreword by Jim Gatheral.

[BAG08] Gérard Ben Arous and Alice Guionnet. The spectrum of heavy tailed random matrices. *Comm. Math. Phys.*, 278(3):715–751, 2008.

[BBDS06] Jinho Baik, Alexei Borodin, Percy Deift, and Toufic Suidan. A model for the bus system in Cuernavaca (Mexico). *J. Phys. A*, 39(28):8965–8975, 2006.

[BDJ99] Jinho Baik, Percy Deift, and Kurt Johansson. On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.*, 12(4):1119–1178, 1999.

[BGK18] F. Benaych-Georges and A. Knowles. *Lectures on the local semicircle law for Wigner matrices*. 2018.

[Bil95] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[BS10] Zhidong Bai and Jack W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, second edition, 2010.

[CHM+15] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. *Journal of Machine Learning Research*, 38:192–204, 2015. 18th International Conference on Artificial Intelligence and Statistics, AISTATS 2015 ; Conference date: 09-05-2015 Through 12-05-2015.

[DF17] Persi Diaconis and Peter J. Forrester. Hurwitz and the origins of random matrix theory in mathematics. *Random Matrices Theory Appl.*, 6(1):1730001, 26, 2017.

[Dur10] Rick Durrett. *Probability: theory and examples*, volume 31 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fourth edition, 2010.

[Dys62] Freeman J. Dyson. A Brownian-motion model for the eigenvalues of a random matrix. *J. Mathematical Phys.*, 3:1191–1198, 1962.

[EoY17] László Erd˝ os and Horng-Tzer Yau. *A dynamical approach to random matrix theory*, volume 28 of *Courant Lecture Notes in Mathematics*. Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 2017.

[ER05] Alan Edelman and N. Raj Rao. Random matrix theory. *Acta Numer.*, 14:233–297, 2005.

[FS11] P. L. Ferrari and H. Spohn. Random growth models. In *The Oxford handbook of random matrix theory*, pages 782–801. Oxford Univ. Press, Oxford, 2011.

[Hal63] P. R. Halmos. What does the spectral theorem say? *Amer. Math. Monthly*, 70:241–247, 1963.

[HJ85] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1985.

[Joh07] Iain M. Johnstone. High dimensional statistical inference and random matrices. In *International Congress of Mathematicians. Vol. I*, pages 307–333. Eur. Math. Soc., Zürich, 2007.

[JT17] Aukosh Jagannath and Thomas Trogdon. Random matrices and the new york city subway system. *Phys. Rev. E*, 96:030101, Sep 2017.

[Kos09] Thomas Koshy. *Catalan numbers with applications*. Oxford University Press, Oxford, 2009.

[KS11] J. P. Keating and N. C. Snaith. Number theory. In *The Oxford handbook of random matrix theory*, pages 491–509. Oxford Univ. Press, Oxford, 2011.

[LCPB00] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 03(03):391–397, 2000.

[Led01] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[May72] Robert M. May. Will a Large Complex System be Stable? *Nature*, 238(5364):413–414, August 1972.

[McK11] Henry P. McKean. Fredholm determinants. *Cent. Eur. J. Math.*, 9(2):205–243, 2011.

[Mec19] Elizabeth S. Meckes. *The random matrix theory of the classical compact groups*, volume 218 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2019.

[Mez07] Francesco Mezzadri. How to generate random matrices from the classical compact groups. *Notices Amer. Math. Soc.*, 54(5):592–604, 2007.

[Mon73] H. L. Montgomery. The pair correlation of zeros of the zeta function. In *Analytic number theory (Proc. Sympos. Pure Math., Vol. XXIV, St. Louis Univ., St. Louis, Mo., 1972)*, volume Vol. XXIV of *Proc. Sympos. Pure Math.*, pages 181–193. Amer. Math. Soc., Providence, RI, 1973.

[Nel67] Edward Nelson. *Dynamical theories of Brownian motion*. Princeton University Press, Princeton, NJ, 1967.

[PB20] Marc Potters and Jean-Philippe Bouchaud. *A First Course in Random Matrix Theory: for Physicists, Engineers and Data Scientists*. Cambridge University Press, 2020.

[So75] Gábor Szeg˝ o. *Orthogonal polynomials*, volume Vol. XXIII of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, fourth edition, 1975.

[Sta15] Richard P. Stanley. *Catalan numbers*. Cambridge University Press, New York, 2015.

[Tao12] Terence Tao. *Topics in random matrix theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[TSSS11] Kazumasa A. Takeuchi, Masaki Sano, Tomohiro Sasamoto, and Herbert Spohn. Growing interfaces uncover universal fluctuations behind scale invariance. *Scientific Reports*, 1(1):34, 2011.

[vNG47] John von Neumann and H. H. Goldstine. Numerical inverting of matrices of high order. *Bull. Amer. Math. Soc.*, 53:1021–1099, 1947.

[Wig58] Eugene P. Wigner. On the distribution of the roots of certain symmetric matrices. *Ann. of Math. (2)*, 67:325–327, 1958.

[Wis28] John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1/2):32–52, 1928.