# Random Matrix Theory

J. P. Keating

February 10, 2021

# Contents

# Preface

These notes accompany the section C course on Random Matrix Theory, C7.7. This is the first time the course has been offered and these notes are new. It is a pleasure to thank Michael Curran, Johannes Forkel, Bhargavi Jonnadula, and Mo Dick Wong for their help in preparing them. I would be grateful for comments, suggestions, and for reports of typos and errors.

J. P. Keating
December 2020

# 1 Introduction and overview

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 therefore also natural to seek 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.

Historically, questions in Random Matrix Theory (RMT) have arisen in a number of different mathematical contexts. We shall start by reviewing briefly 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.

## 1.1 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 $3 \times 3$ orthogonal matrix[1], i.e. a matrix $O$ satisfying $OO^{\mathrm{T}} = I$, where $I$ denotes the $3 \times 3$ identity matrix and, here and hereafter, $M^{\mathrm{T}}$ denotes the transpose of a matrix $M$. 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.

It is an obvious question then how this extends to orthogonal matrices of dimension $n$; that is, how one can construct an invariant probability measure on the space of $n \times n$ orthogonal matrices. Moreover, orthogonal transformations form a compact group, so one can ask how this story generalizes to other related groups, for example to the group of unitary transformations in $n$ dimensions, represented by $n \times n$ unitary matrices $U$, i.e. matrices satisfying $UU^{\dagger} = I$, where $M^{\dagger}$ denotes the complex conjugate of $M^{\mathrm{T}}$. Unitary transformations play a fundamental role in quantum mechanics, particularly in quantum information theory, and random unitary matrices have important applications to modelling 'typical' quantum transformations.

Hurwitz was the first to investigate invariant probability measures on the orthogonal and unitary groups. See [6] for an overview of his work and subsequent developments, and [12] for an introduction to the general theory in wider contexts.

## 1.2 Sample covariance matrices

Denote by $N_p(\mu, \Sigma)$ the multivariate normal distribution of a $p$-dimensional random vector $X = (X^{(1)}, X^{(2)}, \ldots, X^{(p)})^{\mathrm{T}}$, with

$$\mu = \mathbb{E}X = (\mathbb{E}X^{(1)}, \mathbb{E}X^{(2)}, \ldots, \mathbb{E}X^{(p)})^{\mathrm{T}} \tag{1.1}$$

and

$$\Sigma_{ij} = \mathbb{E}(X^{(i)} - \mu_i)(X^{(j)} - \mu_j) \tag{1.2}$$

for $1 \leq i, j \leq p$.

Now let $X_1, X_2, \ldots, X_n$ be $n$ i.i.d. $p \times 1$ random vectors with distribution $N_p(0, \Sigma)$. In many situations in data science, mathematical finance, and statistics one encounters problems of this kind

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

where $\Sigma$ is not known and where one wants to find it. From the data given, the *sample covariance matrix*

$$\sigma_n = \frac{1}{n}\sum_{k=1}^{n} X_k X_k^{\mathrm{T}} \tag{1.3}$$

is a natural estimator for $\Sigma$.

Note that defining the $n \times p$ matrix $\tilde{X}_{ij} = X_i^{(j)}$, one also has that

$$\sigma_n = \frac{1}{n}\tilde{X}\tilde{X}^{\mathrm{T}}. \tag{1.4}$$

Random matrices having this form are known as *Wishart* random matrices, after John Wishart, who introduced them in 1928 [18].

## 1.3 Systems of linear equations

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

$$Ax = b \tag{1.5}$$

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*[2], the ratio of the maximal and minimal absolute values of the eigenvalues of $A$.

In 1947 John von Neumann and Herman Goldstine [13] 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 i.i.d. normal[3] random variables. This subject has a long and interesting history, which is beautifully reviewed in [7]. 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.

## 1.4 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. 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 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 quantum Hamiltonians behave like random matrices explains why large perfectly ordered systems conduct electricity while disordered systems do not.

In fact this is not an exclusively quantum phenomenon: it is observed in all wave theories, including acoustics, optics, elasticity, 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.

[2] i.e. if $AA^{\dagger} = A^{\dagger}A$.

[3] i.e. Gaussian – not to be confused with the use of 'normal' in the sense of the previous footnote.

## 1.5 Stability of high-dimensional dynamical systems

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

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\lambda I x \tag{1.6}$$

where $\lambda$ 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 (1.6) 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

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\lambda I x + Ax \tag{1.7}$$

where $A$ is a random matrix?

This question was first raised by Robert May (Lord May of Oxford) in 1972 in a famous paper [11] 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 (1.6).

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 networks.

## 1.6 Principal Component Analysis

Let $X_1, X_2, \ldots, X_n$ be $n$ $p \times 1$ vectors, where $p$ is to be treated as large with respect to $n$. 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 this 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. So one could look for the unit vector $u \in \mathbb{R}^p$ such that $u^{\mathrm{T}}X_1, u^{\mathrm{T}}X_2, \ldots, u^{\mathrm{T}}X_n$ 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

$$\frac{1}{n}\sum_{j=1}^{n}(u^{\mathrm{T}}X_j)^2 = u^{\mathrm{T}}\left(\frac{1}{n}\sum_{j=1}^{n}X_jX_j^{\mathrm{T}}\right)u = u^{\mathrm{T}}\sigma_n u. \tag{1.8}$$

The maximum value $u^{\mathrm{T}}\sigma_n u$ can take is the largest eigenvalue[4] of $\sigma_n$, 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 $\Sigma$ defined by (1.2), 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.

Returning to the problem of comparing $Y$ with $X_1, X_2, \ldots, X_n$, it is obviously natural to seek to do this in the subspace where the vectors $X_1, X_2, \ldots, X_n$ exhibit maximum variability, because this is where their differences are largest. This can be thought of as enacting 'feature selection' in many

[4]The eigenvalues of $\sigma_n$ are real because the matrix is clearly real and symmetric.

applications. The reduction in the dimension of the space where the comparison is made increases the efficiency of the search considerably.

For an overview of this and similar applications (e.g. to mathematical finance, image analysis, etc) see [8] and [4].

## 1.7 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[5].

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.

## 1.8 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, [5].

## 1.9 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.

One example concerns the length of the longest increasing subsequence in random permutations. Let $S_n$ be the group of permutations of $1, 2, \ldots, n$. If $\pi \in S_n$, $\pi(i_1), \ldots, \pi(i_k)$ is an increasing subsequence in $\pi$ 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 $\pi$ 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 $S_n$ with uniform distribution,

$$\mathbb{P}(l_n \leq m) = \frac{\#\{\pi \in S_n : l_n \leq m\}}{n!}. \tag{1.9}$$

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.

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

A second example concerns the Riemann zeta-function

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}. \tag{1.10}$$

The sum converges in $\mathrm{Re}s > 1$, 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 $\mathrm{Re}s = 1/2$. 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.

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.

## 1.10 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.L. Mehta, Random Matrices (Elsevier, Pure and Applied Mathematics Series)
- G.W. Anderson, A. Guionnet & O. Zeitouni, An Introduction to Random Matrices (Cambridge Studies in Advanced Mathematics)
- E.S. Meckes, The Random Matrix Theory of the Classical Compact Groups (Cambridge University Press)
- G. Akemann, J. Baik & P. Di Francesco, The Oxford Handbook of Random Matrix Theory (Oxford University Press)
- G. Livan, M. Novaes & P. Vivo, Introduction to Random Matrices (Springer Briefs in Mathematical Physics)
- Z. Bai & J.W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices (Springer).
- T. Tao, Topics in Random Matrix Theory, available 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 range 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.

# 2 Examples of random matrix ensembles

Motivated by the examples described in the previous section, we start by defining explicitly the random matrix ensembles that will be studied in this course[6]. Here an ensemble is a space of matrices endowed with a probability measure.

We start with some general notation. We say that an $n \times n$ matrix $M$ is symmetric if $M_{ij}$ is real and $M_{ij} = M_{ji}$, and that it is Hermitian if $M_{ij}$ is complex and $M_{ij} = M_{ji}^*$ , where $^*$ denotes complex conjugation. We denote $\mathcal{S}_n = \{n \times n \text{ symmetric matrices}\}$ and $\mathcal{H}_n = \{n \times n \text{ Hermitian matrices}\}$. The Lebesgue measure on $\mathcal{S}_n$ is by definition the product of the Lebesgue measures on the linearly independent entries of $M$:

$$\mathrm{d}M = \prod_{1 \le i < j \le n} \mathrm{d}M_{ij} \prod_{i=1}^{n} \mathrm{d}M_{ii}. \tag{2.1}$$

Similarly on $\mathcal{H}_n$:

$$\mathrm{d}M = \prod_{1 \le i < j \le n} \mathrm{d}(\mathrm{Re}M_{ij})\mathrm{d}(\mathrm{Im}M_{ij}) \prod_{i=1}^{n} \mathrm{d}M_{ii}. \tag{2.2}$$

## 2.1 Wishart random matrices

Let $p \le n$. A *real Wishart matrix* with covariance matrix $\Sigma \in \mathcal{S}_p$ is a $p$-dimensional random symmetric positive definite matrix $M$ of the form

$$M = XX^{\mathrm{T}} \tag{2.3}$$

where $X$ is an $p \times n$ matrix with columns that are zero-mean, independent, Gaussian vectors with covariance matrix $\Sigma \in \mathcal{S}_n$.

The probability density function of $M$ was computed by Wishart in 1928 to be[7]

$$f(M) = \frac{1}{2^{np/2}\Gamma_p(\frac{n}{2})(\det \Sigma)^{n/2}} (\det M)^{(n-p-1)/2} \exp\left[-\frac{1}{2}\mathrm{Tr}(\Sigma^{-1}M)\right] \tag{2.4}$$

with respect to Lebesque measure on the cone of symmetric positive definite matrices. Here

$$\Gamma_p(\frac{n}{2}) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(\frac{n}{2} - \frac{j-1}{2}\right). \tag{2.5}$$

## 2.2 Wigner random matrices

A *Wigner matrix* is a random matrix $M \in \mathcal{S}_n$ or $M \in \mathcal{H}_n$ such that (i) $M_{ij}$ with $i < j$ are i.i.d. random variables (real or complex) with $\mathbb{E}M_{ij} = 0$ and $\mathbb{E}|M_{ij}|^2 = 1$ , (ii) $M_{ii}$ are i.i.d. real random variables (with a distribution not necessarily the same as for the off-diagonal elements) with $\mathbb{E}M_{ii} = 0$ and $\mathbb{E}M_{ii}^2 < \infty$. In some circumstances one rescales the matrix elements by a power of $n$. This then rescales the eigenvalues. For example, we shall see that some important results emerge when one rescales the matrix elements, and hence the eigenvalues, by $n^{-1/2}$, and others emerge when one scales them by $n^{1/2}$

## 2.3 Gaussian orthogonal random matrices

The *Gaussian Orthogonal Ensemble*, or *GOE*, is the particular class of Wigner random matrices for which the matrix elements are independent real normal random variables. Specifically, (i) $M_{ij}$ with $i < j$ are i.i.d. random variables with distribution $N(0, 1)$ (i.e. Gaussian with zero mean and variance

[6] There are many important ensembles that we will not have the time to investigate here. Those defined here are a representative set and are amongst the most widely studied

[7] The precise form of the density is rarely used.

1), and (ii) $M_{ii}$ are i.i.d. real random variables with distribution $N(0,2)$ (i.e. Gaussian with zero mean and variance 2).

Hence, the probability measure on the entries of $M \in \mathcal{S}_n$ is

$$\prod_{i=1}^{n} \frac{1}{\sqrt{4\pi}} e^{-\frac{1}{4}M_{ii}^2} \prod_{1\le i<j\le n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}M_{ij}^2} \mathrm{d}M = \frac{1}{2^{n/2}(2\pi)^{n(n+1)/4}} e^{-\frac{1}{4}\mathrm{Tr}M^2} \mathrm{d}M. \tag{2.6}$$

The following is one of the most important properties of the GOE. Let $O$ be a non-random $n \times n$ orthogonal matrix (i.e. $OO^{\mathrm{T}} = I$) and let $M$ be an $n \times n$ GOE matrix. Then the distribution of $M$ is the same as that of $OMO^{\mathrm{T}}$; that is, the GOE is invariant under conjugation by all orthogonal matrices[8]. To prove this, note first that

$$\mathrm{Tr}[OMO^{\mathrm{T}}]^2 = \mathrm{Tr}[OM^2O^{\mathrm{T}}] = \mathrm{Tr}[M^2O^{\mathrm{T}}O] = \mathrm{Tr}M^2; \tag{2.7}$$

second, it follows from calculating the Jacobian of the transformation that

$$\mathrm{d}[OMO^{\mathrm{T}}] = \mathrm{d}M. \tag{2.8}$$

To see this directly, note that the Hilbert-Schmidt norm of $M$ is precisely $\mathrm{Tr}M^2$, which we have just seen is invariant under conjugation by $O$. The map $M \mapsto OMO^{\mathrm{T}}$ is therefore an isometry and so its Jacobian determinant is 1. Alternatively one can compute the Jacobian by hand and confirm this by explicit calculation.

The only ensemble of symmetric random matrix 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 Gaussian unitary random matrices

The *Gaussian Unitary Ensemble*, or *GUE*, is the particular class of Wigner random matrices for which the matrix elements are independent complex normal random variables. Specifically, (i) $M_{ij}$ with $i < j$ are i.i.d. random variables with real and imaginary parts which are are independently Gaussian, each with zero mean and variance $1/2$, and (ii) $M_{ii}$ are i.i.d. real random variables with distribution $N(0,1)$ (i.e. Gaussian with zero mean and variance 1).

In this case the probability measure on the entries of $M \in \mathcal{H}_n$ is

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}M_{ii}^2} \prod_{1\le i<j\le n} \frac{1}{\sqrt{\pi}} e^{-(\mathrm{Re}M_{ij})^2} \frac{1}{\sqrt{\pi}} e^{-(\mathrm{Im}M_{ij})^2} \mathrm{d}M = \frac{1}{2^{n/2}\pi^{n^2/2}} e^{-\frac{1}{2}\mathrm{Tr}M^2} \mathrm{d}M \tag{2.9}$$

and it is invariant under conjugation by all non-random unitary matruces $U$; i.e. for $U$ an $n \times n$ unitary matrix (i.e. $UU^\dagger = I$), $M$ and $UMU^\dagger$ have the same distribution[9]. As in the case of the GOE, to prove this note that

$$\mathrm{Tr}[UMU^\dagger]^2 = \mathrm{Tr}[UM^2U^\dagger] = \mathrm{Tr}[M^2U^\dagger U] = \mathrm{Tr}M^2; \tag{2.10}$$

and that it follows from calculating the Jacobian of the transformation that

$$\mathrm{d}[UMU^\dagger] = \mathrm{d}M. \tag{2.11}$$

Again, this may be viewed as a consequence of the fact that $\mathrm{Tr}M^2$ is the Hilbert-Schmidt norm of $M$, and so conjugation by a unitary matrix is an isometry.

[8] This is the reason for the name Gaussian Orthogonal Ensemble.

[9] This is the reason for the name Gaussian Unitary Ensemble.

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.

Note that the GOE and GUE measures may, up to a normalization constant, be written in the form

$$e^{-\frac{\beta}{4}\mathrm{Tr}M^2}\mathrm{d}M \tag{2.12}$$

where for the GOE $\beta = 1$ and for the GUE $\beta = 2$.

It is also worth noting that the GOE matrix elements are real and the GUE matrix entries are complex. There is a third Gaussian ensemble, which we will not have time to discuss, for which the matrix elements are quaternions. This is called the *Gaussian Symplectic Ensemble* and is invariant under conjugation by all symplectic matrices.

## 2.5 Random unitary and orthogonal matrices

The *circular unitary ensemble*, or *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 *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.

# 3 The semicircle law for Wigner random matrices

Our goal in this section is to take the first steps in understanding how the eigenvalues of Wigner random matrices are distributed. As before, let $M$ denote an $n \times n$ matrix with real or complex entries such that for $1 \leq i \leq j \leq n$ the entries $M_{ij}$ are independent, $M = M^\dagger$ in the case when the entries are complex, and $M = M^{\mathrm{T}}$ when the entries are real. Clearly $M$ depends on $n$, but we will not record this explicitly, unless it is necessary to do so. We shall assume[10] that $\mathbb{E}M_{ij} = 0$ and

$$\lim_{n\to\infty} \max_{1\leq i,j\leq n} |n\mathbb{E}[|M_{ij}|^2] - 1| = 0. \tag{3.1}$$

We denote the eigenvalues of $M$ by $\lambda_1, \lambda_2, \ldots, \lambda_n$. It is important to note that in this case we are taking the mean-square size of the matrix entries of $M$ to be asymptotically $1/n$, rather than 1. This rescaling means that the eigenvalues are effectively rescaled by $n^{-1/2}$ compared to those of matrices where the mean-square size of the matrix elements is independent of $n$.

**Exercise** Recall the proof that for complex Hermitian matrices (and so as a special case real symmetric matrices) the eigenvalues are all real.

We will show here the following.

**Theorem 1.** *For bounded continuous functions $f(x)$ and for $M$ satisfying $\mathbb{E}M_{ij} = 0$ and* (3.1)

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(\lambda_i) = \int f(x) \mathrm{d}\sigma(x) \tag{3.2}$$

*where $\mathrm{d}\sigma(x)$ is given by*

$$\mathrm{d}\sigma(x) = \begin{cases} \frac{1}{2\pi}\sqrt{4-x^2}\mathrm{d}x & \text{if } |x| \leq 2 \\ 0 & \text{if } |x| > 2 \end{cases}. \tag{3.3}$$

*and where the convergence holds in expectation and almost surely.*

This is known as Wigner's semicircle law. Various versions of the proof of this result yield different notions of convergence to the limit: Wigner's original approach [17] gave convergence in expectation, but this can be extended to convergence in probability and almost sure convergence, etc. The different proofs rely on different assumptions about the moments of the matrix elements $M_{ij}$. We will explain the proof of convergence in expectation first, and then the extension to almost sure convergence.

Before proceeding 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 (3.3). 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 as follows.

- The main idea is to compare the moments of the semicircle law (3.3) with the corresponding spectral moments, i.e. to consider the case when $f(x) = x^k$.
- The computation of the spectral moments maps onto a problem in combinatorics, leading to a proof that they coincide with the moments of (3.3) in the limit when $n \to \infty$.

[10] These assumptions can be relaxed in various ways but are sufficient for our purposes here. Following the steps in the proof may be easier if one has in mind the weaker condition that $\mathbb{E}[|M_{ij}|^2] = 1/n$; the difference will not be important to us.

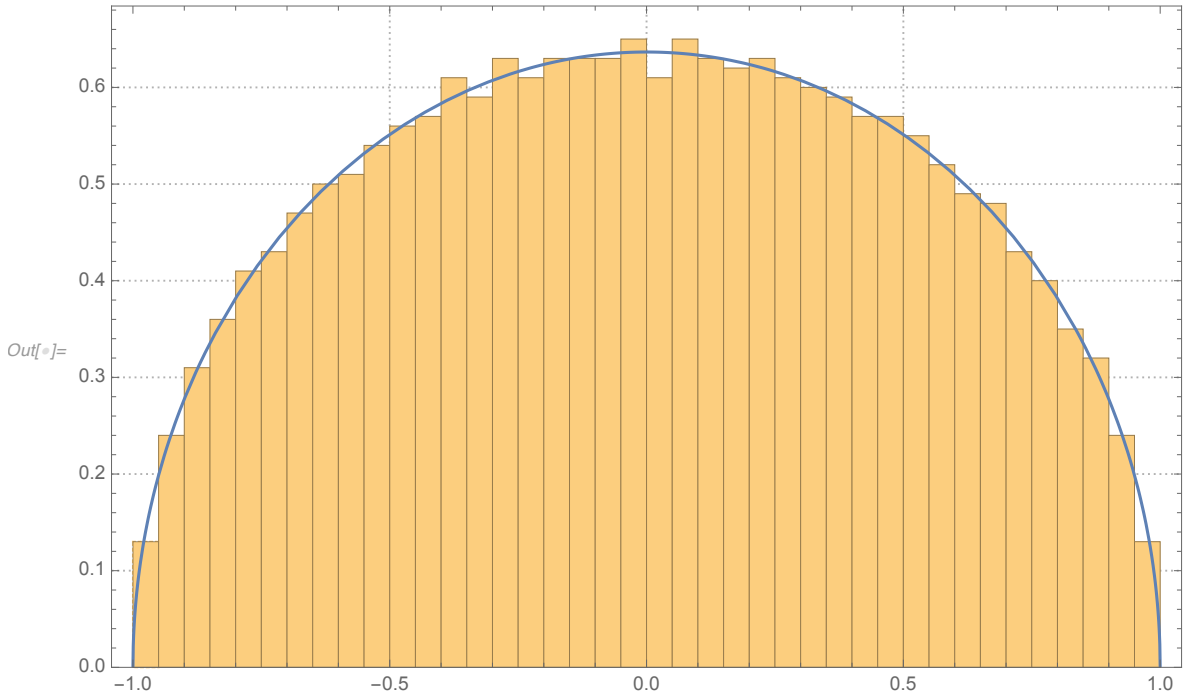


Figure 1: The spectral density of a single GOE matrix of dimension 2000 compared to the Wigner semicircle law.

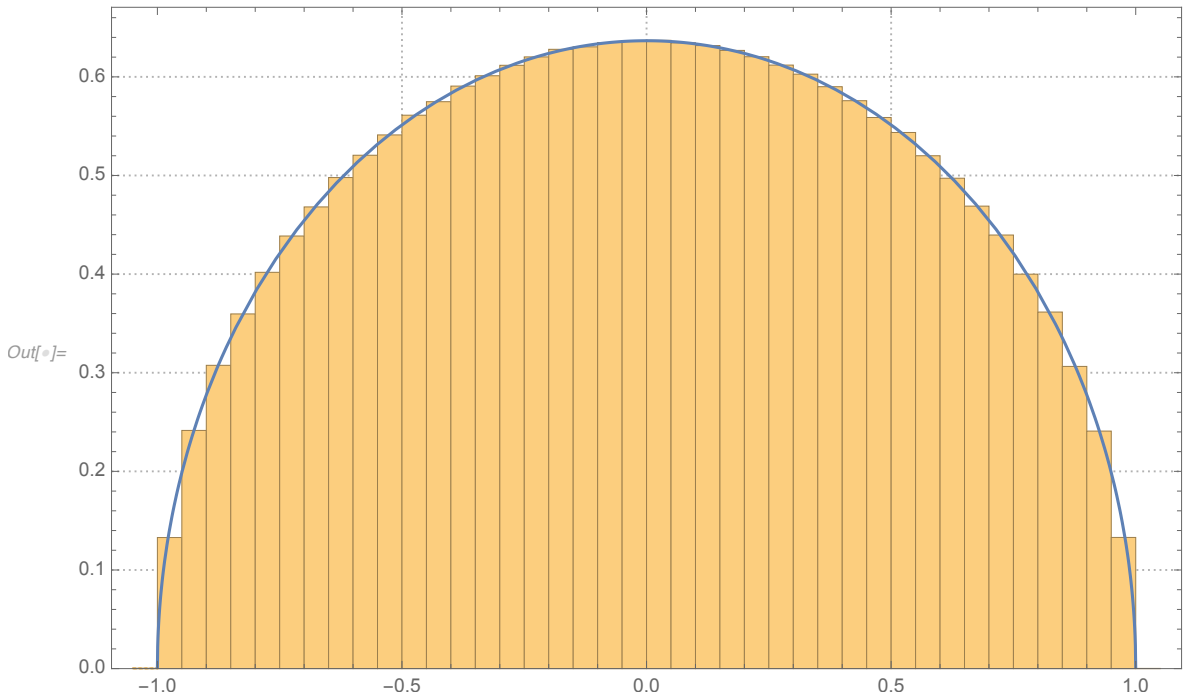


Figure 2: The spectral density obtained by averaging over 100 GOE matrices of dimension 2000 compared to the Wigner semicircle law.

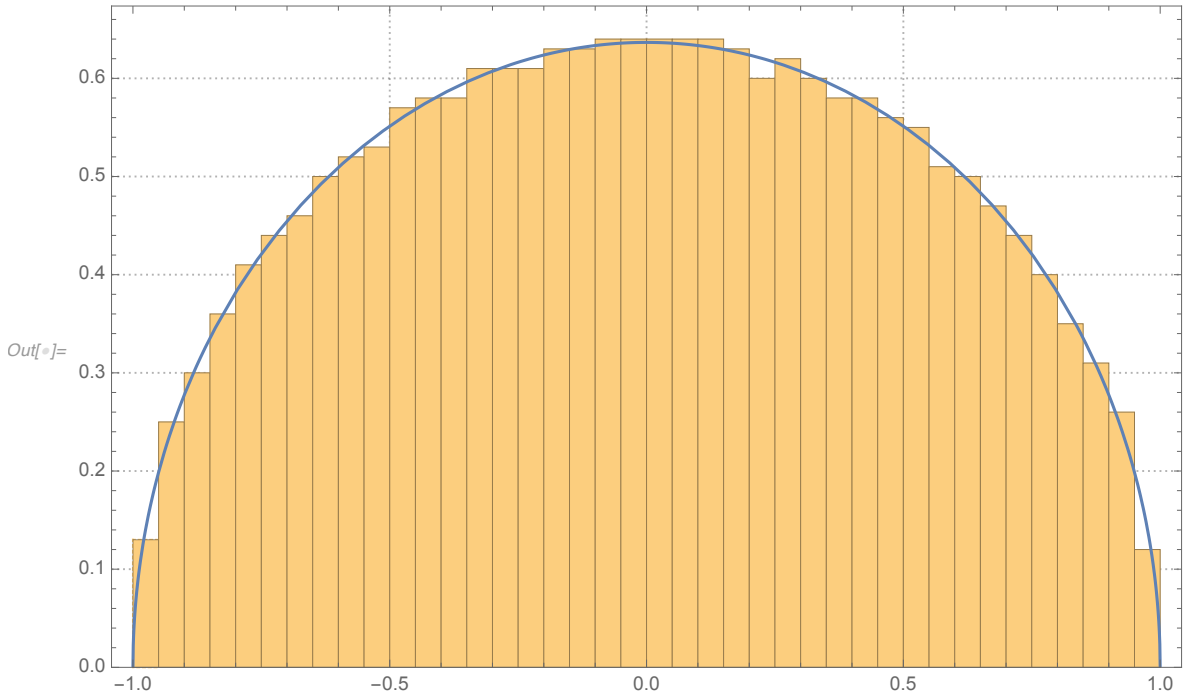


Figure 3: The spectral density of a single GUE matrix of dimension 2000 compared to the Wigner semicircle law.

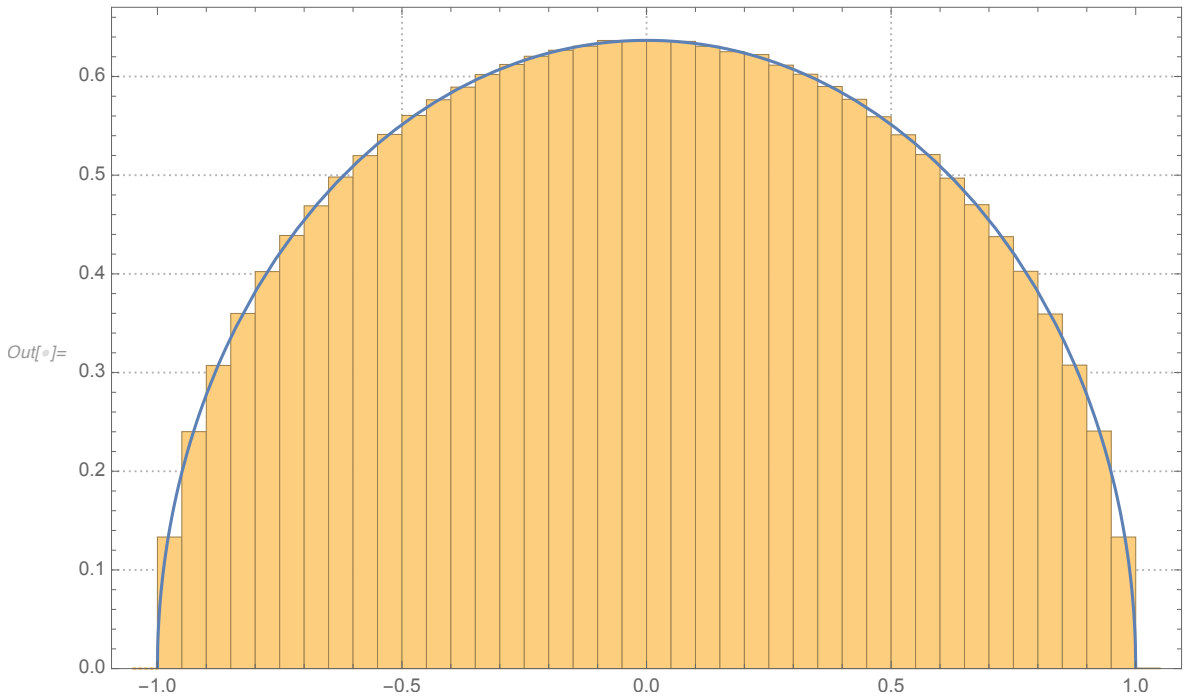


Figure 4: The spectral density obtained by averaging over 100 GUE matrices of dimension 2000 compared to the Wigner semicircle law.

- We first show this for convergence in expectation.
- A variance calculation then implies almost sure convergence.
- This proves (3.3) when $f(x)$ is a polynomial function of $x$.
- The proof of (3.3) for bounded continuous functions follows from the fact that such functions can be approximated by polynomials. We shall discuss this at the end.

## 3.1 Moments of the semicircle law and Catalan numbers

The moments of the semicircle law (3.3) are

$$\alpha_k = \frac{1}{2\pi}\int_{-2}^{2} x^k \sqrt{4-x^2}\mathrm{d}x \tag{3.4}$$

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

$$\begin{aligned}
\alpha_{2m} &= \frac{1}{2\pi}\int_{-2}^{2} x^{2m}\sqrt{4-x^2}\mathrm{d}x \\
&= \frac{2^{2m+1}}{\pi}\int_0^{\pi} (\cos\theta)^{2m}\sin^2\theta \mathrm{d}\theta \\
&= \frac{2^{2m+1}}{\pi}\left[\int_0^{\pi}(\cos\theta)^{2m}\mathrm{d}\theta - \int_0^{\pi}(\cos\theta)^{2m+2}\mathrm{d}\theta\right] \\
&= \frac{2}{\pi}\left[\int_0^{\pi}(e^{i\theta}+e^{-i\theta})^{2m}\mathrm{d}\theta - \frac{1}{4}\int_0^{\pi}(e^{i\theta}+e^{-i\theta})^{2m+2}\mathrm{d}\theta\right] \\
&= 2\left[\binom{2m}{m} - \frac{1}{4}\binom{2m+2}{m+1}\right] \\
&= \frac{1}{m+1}\binom{2m}{m},
\end{aligned} \tag{3.5}$$

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 numbers

$$C_m = \frac{1}{m+1}\binom{2m}{m} \tag{3.6}$$

are known as the *Catalan numbers*. They satisfy

$$C_m = \binom{2m}{m} - \binom{2m}{m+1}, \tag{3.7}$$

the recurrence relation

$$C_{m+1} = \sum_{i=1}^{m} C_i C_{m-i}, \tag{3.8}$$

with $C_0 = 1$, and they are all integers[11]. 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 (3.6) that $C_m \leq 4^m$.

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, [10, 15]. We now describe briefly one such connection.

[11] This is straightforward to see from (3.7) or from (3.8).

For $k \in \mathbb{N}$, the set $\mathcal{D}_k$ of *Dyck paths* is the set of paths of length $2k$ starting at 0 and ending at 0 with increments $\pm 1$ and which remain non-negative[12]. Then $|\mathcal{D}_k| = C_k$.

To prove this, let $\mathcal{A}$ denote the set of walks of length $2k$ with increments $\pm 1$ which start at 0 and end at -2, and similarly let $\mathcal{B}$ denote the set of all walks of length $2k$ with increments $\pm 1$ which start at 0 and end at 0. Then $|\mathcal{D}_k| = |\mathcal{B}| - |\mathcal{A}|$, which may be seen in the following way. Let $\mathcal{C}$ denote the set of walks of length $2k$ with increments $\pm 1$ 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

$$|\mathcal{D}_k| = \binom{2k}{k} - \binom{2k}{k+1} = C_k. \tag{3.9}$$

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

$$d_{k+1} = \sum_{m=1}^{k} d_m d_{k-m}, \tag{3.10}$$

because there are $d_m d_{k-m}$ paths which first hit 0 at time $2(m+1)$, and $d_0 = 1$. Hence $d_k$ satisfies the recurrence relation (3.8).

## 3.2 Spectral moments

We begin by focusing on polynomial functions $f(x)$ in (3.2). This means that we need only consider $f(x) = x^k$. The sum on the left hand side of (3.2), which we refer to as the $k$th *spectral moment*, is then simply

$$\frac{1}{n}\sum_{i=1}^{n} \lambda_i^k = \frac{1}{n}\mathrm{Tr}M^k. \tag{3.11}$$

Our goal now is to prove the following theorem.

**Theorem 2.** *(Wigner 1958) Assume that for all $k \in \mathbb{N}$*

$$B_k := \sup_{n\in\mathbb{N}} \sup_{(i,j)\in\{1,\dots,n\}^2} \mathbb{E}[|\sqrt{n}M_{ij}|^k] < \infty. \tag{3.12}$$

*Then*

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Tr}M^k = \begin{cases} 0 & \text{if } k \text{ is odd} \\ C_{\frac{k}{2}} & \text{otherwise} \end{cases} \tag{3.13}$$

*where the convergence holds in expectation and almost surely.*

The main idea underlying the proof of this theorem is to expand $\mathrm{Tr}M^k$ in terms of products of elements of $M$, 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 2 for convergence in expectation. When this is done we explain how to extend the proof to almost sure convergence.

[12]Note that $\mathcal{D}_k$ is in obvious bijection with the "latex bracketing problem", that is with the number of ways of writing $2k$ brackets { and } starting with { and so the brackets correctly match, if we identify { with a step of $+1$ and } with a step of $-1$.

First, we set $\mu_{ij} = \sqrt{n}M_{ij}$, and denote $\mu = \sqrt{n}M$. Then

$$\mathbb{E}\left[\frac{1}{n}\mathrm{Tr}M^k\right] = \sum_{i_1,i_2,\ldots,i_k=1}^{n} \frac{1}{n^{1+k/2}}\mathbb{E}[\mu_{i_1i_2}\mu_{i_2i_3}\ldots\mu_{i_ki_1}]. \tag{3.14}$$

Note that for all sets of indices $\mathbb{E}[\mu_{i_1i_2}\mu_{i_2i_3}\ldots\mu_{i_ki_1}]$ is uniformly bounded by $B_k$. Each set of indices corresponds to a sequence $i_1i_2\ldots i_ki_1$, which can be thought of as a closed connected path on the set of vertices $\{i_1, i_2, \ldots, i_k\}$ with edges corresponding to pairs of consecutive indices $i_ji_{j+1}$ and with the convention that $i_{k+1} = i_1$. We denote the graph with this set of vertices and edges $G_{(i_1,i_2,\ldots,i_k)}$. Clearly the graph is connected. One can think of the paths as closed walks on this graph. Crucially, because the matrix entries are independent, centred random variables, $\mathbb{E}[\mu_{i_1i_2}\mu_{i_2i_3}\ldots\mu_{i_ki_1}] = 0$ for all paths except those for which every edge is traversed at least twice, possibly in reverse. There can therefore be at most $k/2$ unique edges and hence at most $k/2+1$ distinct vertices in paths giving a non-zero contribution to the expectation value.

For a sequence $i_1i_2\ldots i_ki_1$ we define its *weight* $w$ to be the number of distinct indices. The terms making a non-zero contribution to (3.14) have $w \leq k/2+1$. Two sequences $i_1i_2\ldots i_ki_1$ and $i'_1i'_2\ldots i'_ki'_1$ are termed *equivalent* if there exists a bijection on the set $\{1, 2, \ldots, n\}$ mapping each $i_j$ to $i'_j$. Sequences that are equivalent have essentially the same weight and their contributions to (3.14) are asymptotically the same. Note that the total number of distinct equivalent classes depends on $k$ but is independent of $n$, because each class has a representative where all $i_1i_2\ldots i_k$ are in $\{1, 2, \ldots, k\}$.

The next step is to show that those paths with weight $w < k/2+1$ make a contribution to (3.14) that vanishes in the limit as $n \to \infty$. As noted above, $\mathbb{E}[\mu_{i_1i_2}\mu_{i_2i_3}\ldots\mu_{i_ki_1}]$ is uniformly bounded by $B_k$. For a given index sequence $i_1i_2\ldots i_ki_1$ of weight $w$, the number of sequences equivalent to it is $n(n-1)(n-2)\ldots(n-w+1) \leq n^w$. The contribution of each term in this equivalence class to (3.14) is therefore $\leq B_k n^{w-k/2-1}$ and if $w < k/2+1$ this tends to zero as $n \to \infty$. Therefore, because the number of distinct equivalent classes is independent of $n$, the total contribution from paths with $w < k/2+1$ tends to zero in the limit, and when $k$ is even it does so at least as fast as $1/n$.

It follows that the only sequences making a non-zero contribution to (3.14) in the limit $n \to \infty$ are those for which $w = k/2+1$. When $k$ is odd this equation has no solution, because $w$ is an integer. Hence when $k$ is odd the left hand side of (3.14) 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.

To see this note that for any finite connected graph $G = (V, E)$ consisting of a set of vertices $V$ and a set of edges $E$,

$$|V| \leq |E| + 1, \tag{3.15}$$

with equality if and only if $G$ is a tree. If $G$ is a tree, it follows easily by induction that $|E| = |V| - 1$. If $G$ is not a tree, let $T = (V, E')$ denote a spanning tree of $G$; i.e. a tree which has the same vertices as $G$. Then $|E| \geq |E'| = |V| - 1$. If $|E| = |V| - 1$ then we would have $|E| = |E'|$, and so $G = T$.

When $k$ is even and $w = k/2+1$, the sequence $i_1i_2\ldots i_ki_1$ 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 (3.14) only off-diagonal matrix entries contribute, and these are precisely paired, so each appears squared. Therefore, using the fact that[13] $\mathbb{E}|\mu_{ij}|^2 \sim 1$ for $i \neq j$, and that the off-diagonal matrix entries are independent, we have

$$\mathbb{E}[\mu_{i_1i_2}\mu_{i_2i_3}\ldots\mu_{i_ki_1}] \sim 1. \tag{3.16}$$

Recall that for a given index sequence $i_1i_2\ldots i_ki_1$ of weight $w$, the number of sequences equivalent to it is $n(n-1)(n-2)\ldots(n-w+1)$, and that we now have $w = k/2+1$, so that this number grows like $n^{1+k/2}$, as $n \to \infty$ which exactly cancels the factor multiplying the expectation in (3.14).

The problem has therefore reduced to that of counting the number of sequences $i_1i_2\ldots i_ki_1$ corresponding to paths of length $k$ on the tree with $k/2+1$ vertices and $k/2$ distinct edges, each traversed

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

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 $j$th 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 $\pm 1$. 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

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\mathrm{Tr}M^k &= \#\{\text{path sequences of length } k\}\\ &= \#\{\text{Dyck paths of length } k\}\\ &= C_{k/2},\end{aligned} \tag{3.17}$$

and when $k$ is odd, the limit is zero. This proves Theorem 2 for convergence in expectation.

We next indicate how to prove almost sure convergence to the same limit. The strategy here is, having computed the expectation value of the moments, now 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. This is proved by appealing to the Borel-Cantelli lemma. It is an example of a phenomenon known as *concentration of measure* that plays an important role in RMT.

The variance is

$$\mathrm{var}\left[\frac{1}{n}\mathrm{Tr}M^k\right] = \mathbb{E}\left[\frac{1}{n}\mathrm{Tr}M^k\right]^2 - \left[\mathbb{E}\frac{1}{n}\mathrm{Tr}M^k\right]^2 \tag{3.18}$$

and so

$$\begin{aligned}\mathrm{var}\left[\frac{1}{n}\mathrm{Tr}M^k\right] = \frac{1}{n^{2+k}}\sum_{i_1,i_2,\dots,i_k=1}^{n}\sum_{j_1,j_2,\dots,j_k=1}^{n} &\mathbb{E}[\mu_{i_1 i_2}\mu_{i_2 i_3}\dots\mu_{i_k i_1}\mu_{j_1 j_2}\mu_{j_2 j_3}\dots\mu_{j_k j_1}]\\ &- \mathbb{E}[\mu_{i_1 i_2}\mu_{i_2 i_3}\dots\mu_{i_k i_1}]\mathbb{E}[\mu_{j_1 j_2}\mu_{j_2 j_3}\dots\mu_{j_k j_1}].\end{aligned} \tag{3.19}$$

As in the calculation of the expectation value of the moments above, one can analyse this sum in terms of paths on a graph. In the case of the first term on the right hand side of (3.19), the vertices of the graph in question are labelled by $\{i_1, i_2, \dots, i_k, j_1, j_2, \dots, j_k\}$. For this graph to contribute to the sum, it must be connected[14], otherwise the first term factorises 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 arguments above, the terms making a non-zero contribution to (3.19) have $w \le k+1$. The fact that the summand is uniformly bounded implies that the variance tends to zero at least as fast as $1/n$. However, we shall need a better bound, and we can find one as follows. For the reasons explained in the previous calculation, the largest contribution is expected to come from paths with $w = k+1$, which is when the graph is a tree and each edge appears exactly twice, traversed once in each direction. However, in this case no such terms exist. To see this note that for a pair $\{i_1, i_2, \dots, i_k\}$ and $\{j_1, j_2, \dots, j_k\}$ to make a non-zero contribution, they must have at least one edge in common; otherwise the first term in (3.19) factorizes and is cancelled by the second term. However, if each edge is to appear exactly twice in the $i$-path or the $j$-path, this is impossible unless the graph has a closed cycle; and this is a contradiction, because the graph is a tree.

It follows that there are no contributing paths with $w = k+1$, and so in fact the variance tends to zero as $n \to \infty$ at least as fast as $1/n^2$.

[14] and so $i_1 = j_1$

We now apply *Chebychev's inequality*, which asserts that for a random variable $X$ with zero mean and variance $\sigma^2$

$$\mathbb{P}(|X| \geq \delta) \leq \frac{\sigma^2}{\delta^2} \tag{3.20}$$

for any $\delta > 0$[15]. In our case, the fact that the variance tends to zero at least as fast as $1/n^2$ then means that for any $\delta > 0$

$$\mathbb{P}\left( \left| \frac{1}{n} \mathrm{Tr} M^k - \mathbb{E} \frac{1}{n} \mathrm{Tr} M^k \right| > \delta \right) \leq \frac{\alpha_k}{\delta^2 n^2} \tag{3.21}$$

where $\alpha_k$ is a uniform bound on the summand in (3.19).

The *Borel-Cantelli lemma* asserts that if $E_1, E_2, \ldots$ is a sequence of events in some probability space, and if the sum of their probabilities is finite, i.e. if

$$\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty, \tag{3.22}$$

then

$$\mathbb{P}(\limsup_{n \to \infty} E_n) = 0 \tag{3.23}$$

i.e. the probability that infinitely many of them occur is 0.

We can apply the Borel-Cantelli lemma to (3.21) precisely because the sum of the right-hand side over $n$ converges[16], giving that the convergence in Theorem 2 holds almost surely; that is,

$$\lim_{n \to \infty} \left| \frac{1}{n} \mathrm{Tr} M^k - \mathbb{E} \frac{1}{n} \mathrm{Tr} M^k \right| = 0 \tag{3.24}$$

almost surely.

We have therefore established Theorem 2 for polynomials $f(x)$. It remains to show that the result extends to all bounded continuous functions $f(x)$. One can do this using the Weierstrass approximation theorem, which asserts that in a bounded domain $f(x)$ can be approximated by a polynomial. Specifically, for any $B$ and $\delta > 0$, one can find a polynomial $p_\delta(x)$ such that

$$\sup_{|x| \leq B} |f(x) - p_\delta(x)| < \frac{\delta}{6}. \tag{3.25}$$

We now introduce some additional notation. Let

$$\Theta(z) = \begin{cases} 1 & \text{if } z \leq 0 \\ 0 & \text{if } z > 0. \end{cases} \tag{3.26}$$

We define the *empirical spectral distribution function* to be

$$\mu_M(x) := \frac{1}{n} \sum_{j=1}^{n} \Theta(\lambda_j - x). \tag{3.27}$$

So this is the fraction of eigenvalues of $M$ less than or equal to $x$. Note that

$$\frac{1}{n} \sum_{j=1}^{n} \phi(\lambda_j) = \int_{-\infty}^{\infty} \phi(x) \mathrm{d}\mu_M(x). \tag{3.28}$$

We also define

$$\bar{\mu}(x) := \mathbb{E} \frac{1}{n} \sum_{j=1}^{n} \Theta(\lambda_j - x). \tag{3.29}$$

We start with the following result which we will need.

[15] Chebychev's inequality follows from Markov's inequality: $\mathbb{P}(|X|^2 > (\delta\sigma)^2) \leq \mathbb{E}[|X|^2]/(\delta\sigma)^2$.

[16] This explains why the straightforward bound, $1/n$, was not good enough for our purposes and why we needed the stronger bound (3.21) .

**Lemma 3.** *Let $p \in \mathbb{N}$ and $\epsilon > 0$. For $B > 4$*

$$\limsup_{n\to\infty} \mathbb{P}\left(\int_{|x|>B} |x|^p \mathrm{d}\mu_M(x) > \epsilon\right) = 0 \tag{3.30}$$

*Proof.* By Markov's inequality

$$\mathbb{P}\left(\int_{|x|>B} |x|^p \mathrm{d}\mu_M(x) > \epsilon\right) \leq \frac{1}{\epsilon}\mathbb{E}\left(\int_{|x|>B} |x|^p \mathrm{d}\mu_M(x)\right) \tag{3.31}$$

and

$$\int_{|x|>B} |x|^p \mathrm{d}\mu_M(x) \leq \frac{1}{B^p}\int x^{2p}\mathrm{d}\mu_M(x), \tag{3.32}$$

which follows by multiplying the integrand by $x^p/B^p$.

Hence

$$\mathbb{P}\left(\int_{|x|>B} |x|^p \mathrm{d}\mu_M(x) > \epsilon\right) \leq \frac{1}{\epsilon B^p}\mathbb{E}\left(\int x^{2p}\mathrm{d}\mu_M(x)\right) = \frac{1}{\epsilon B^p}\frac{\mathbb{E}\mathrm{Tr}M^{2p}}{n}. \tag{3.33}$$

We have shown above that the right hand side converges to $C_p/\epsilon B^p$, and so, for $n$ sufficiently large, using the fact that $C_p \leq 4^p$,

$$\mathbb{P}\left(\int_{|x|>B} |x|^p \mathrm{d}\mu_M(x) > \epsilon\right) \leq \frac{1}{\epsilon}\left(\frac{4}{B}\right)^p. \tag{3.34}$$

In the range of integration, $|x|^p$ is strictly increasing, and so the left hand side increases as $p$ increases. However, the right hand side decreases as $p$ increases. This is only possible if the limsup of the sequence on the left is actually 0, as was to be shown.

□

We return now to proving that Theorem 2 extends to all bounded continuous functions $f(x)$, making use of this Lemma. Let $f(x)$ be a (fixed) bounded continuous function, fix $B > 4$ and $\delta > 0$. We have by the triangle inequality that

$$\begin{aligned}\left|\int f(x)\mathrm{d}\mu_M(x) - \int f(x)\mathrm{d}\sigma(x)\right| \leq &\left|\int f(x)\mathrm{d}\mu_M(x) - \int p_\delta(x)\mathrm{d}\mu_M(x)\right| \\ &+ \left|\int p_\delta(x)\mathrm{d}\mu_M(x) - \int p_\delta(x)\mathrm{d}\sigma(x)\right| \\ &+ \left|\int p_\delta(x)\mathrm{d}\sigma(x) - \int f(x)\mathrm{d}\sigma(x)\right|,\end{aligned} \tag{3.35}$$

where $\sigma(x)$ is the semicircle law. Therefore,

$$\begin{aligned}\mathbb{P}\left(\left|\int f(x)\mathrm{d}\mu_M(x) - \int f(x)\mathrm{d}\sigma(x)\right| > \delta\right) \leq &\mathbb{P}\left(\left|\int f(x)\mathrm{d}\mu_M(x) - \int p_\delta(x)\mathrm{d}\mu_M(x)\right| > \delta/3\right) \\ &+ \mathbb{P}\left(\left|\int p_\delta(x)\mathrm{d}\mu_M(x) - \int p_\delta(x)\mathrm{d}\sigma(x)\right| > \delta/3\right) \\ &+ \mathbb{P}\left(\left|\int p_\delta(x)\mathrm{d}\sigma(x) - \int f(x)\mathrm{d}\sigma(x)\right| > \delta/3\right).\end{aligned} \tag{3.36}$$

By design of (3.25), the last term is identically zero. In the first term we can use the triangle inequality

$$\left|\int f(x)\mathrm{d}\mu_M(x) - \int p_\delta(x)\mathrm{d}\mu_M(x)\right| \leq \int_{|x|\leq B} |f(x) - p_\delta(x)|\mathrm{d}\mu_M(x) + \int_{|x|>B} |f(x) - p_\delta(x)|\mathrm{d}\mu_M(x). \tag{3.37}$$

Hence

$$\mathbb{P}\left(\left|\int f(x)\mathrm{d}\mu_M(x) - \int p_\delta(x)\mathrm{d}\mu_M(x)\right| > \delta/3\right) \leq \mathbb{P}\left(\int_{|x|\leq B} |f(x) - p_\delta(x)|\mathrm{d}\mu_M(x) > \delta/6\right) + \mathbb{P}\left(\int_{|x|>B} |f(x) - p_\delta(x)|\mathrm{d}\mu_M(x) > \delta/6\right). \quad (3.38)$$

Again, it follows from (3.25) that the first term is identically zero.

We are therefore left with

$$\mathbb{P}\left(\left|\int f(x)\mathrm{d}\mu_M(x) - \int f(x)\mathrm{d}\sigma(x)\right| > \delta\right) \leq \mathbb{P}\left(\left|\int p_\delta(x)\mathrm{d}\mu_M(x) - \int p_\delta(x)\mathrm{d}\sigma(x)\right| > \delta/3\right) + \mathbb{P}\left(\int_{|x|>B} |f(x) - p_\delta(x)|\mathrm{d}\mu_M(x) > \delta/6\right). \quad (3.39)$$

We have already shown that the first term tends to zero as $n \to \infty$. In the second term, because $f(x)$ is bounded and $p_\delta(x)$ is a polynomial, we have $|f(x) - p_\delta(x)| \leq c|x|^k$ for some positive constant $c$ and some integer $k$. Therefore

$$\mathbb{P}\left(\int_{|x|>B} |f(x) - p_\delta(x)|\mathrm{d}\mu_M(x) > \delta/6\right) \leq \mathbb{P}\left(\int_{|x|>B} c|x|^k\mathrm{d}\mu_M(x) > \delta/6\right). \quad (3.40)$$

It follows from Lemma 3 that as $n \to \infty$, the lim sup of the righthand side is 0, concluding the proof of Theorem 2.

We note finally that establishing Theorem 2 implies that the measure $\mathrm{d}\mu_M(x)$ almost surely converges weakly to the measure $\mathrm{d}\sigma(x)$ when $n \to \infty$.

## 3.3 The 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.

Let $X$ be a centred complex random variable with mean 0 and variance 1, and let $M$ be the random $n \times n$ matrix with entries being i.i.d. copies of $X$. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of $\frac{1}{\sqrt{n}}M$. Define the empirical spectral distribution function $\mu_M$ by

$$\mu_M(s,t) = \frac{1}{n}\#\{k \leq n | \mathrm{Re}(\lambda_k) \leq s; \mathrm{Im}(\lambda_k) \leq t\}. \quad (3.41)$$

We shall compare this to the uniform distribution on the unit disk in the complex plane

$$\nu(s,t) = \frac{1}{\pi}\mathrm{meas}\{z \in \mathbb{C} : |z| \leq 1; \mathrm{Re}(z) \leq s; \mathrm{Im}(z) \leq t\}. \quad (3.42)$$

The *weak circular law* is said to hold for $X$ if for any fixed $s, t$ $\mu_M$ converges to $\nu$ in probability. The *strong circular law* is said to hold for $X$ if, with probability 1, $\mu_M$ converges uniformly to $\nu$ over the unit disk as $n \to \infty$. Both laws hold under the general assumptions on $X$ stated.

## 3.4 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.

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

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\lambda I x \tag{3.43}$$

where $\lambda$ 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 (3.43) 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

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\lambda I x + Ax \tag{3.44}$$

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

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 $A$ represents interactions between the different species.

Now, if $A$ 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 $A$ 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 $A$ is not symmetric, so the matrix entries are independent random variables, the circular law 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 i.i.d. 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.

# 4 Gaussian random matrices

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 2. 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 (3.14).

## 4.1 Wick's theorem

Let $X$ be a real standard normal random variable, i.e. one with mean 0 and variance 1, so that

$$\mathbb{P}(a \leq X \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-y^2/2} \mathrm{d}y. \tag{4.1}$$

Hence

$$\mathbb{E}[X^n] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^n e^{-y^2/2} \mathrm{d}y = \begin{cases} 0 & \text{if } n \text{ is odd} \\ (n-1)!! & \text{if } n \text{ is even} \end{cases} \tag{4.2}$$

where, when $m$ is odd, $m!! = m(m-2)(m-4)\dots 1$. Therefore $\mathbb{E}[X^n]$ is equal to the number of ways of splitting $n$ objects into disjoint pairs. For example $\mathbb{E}[X^4] = 3$, corresponding to the 3 ways to split the numbers $1, 2, 3, 4$ into disjoint pairs (i.e in this case [(1, 2)(3, 4)], [(1, 3)(2, 4)], [(1, 4)(2, 3)]).

We label the different ways of splitting $n$ objects into disjoint pairs by $\pi$ and the set of all pairings by $\mathcal{P}_2(n)$, so $\#\mathcal{P}_2(n) = (n-1)!!$ when $n$ is even and zero otherwise.

The following is then a straightforward application of (4.2) due to Isserlis in 1918 and Wick in 1950 (and many others).

**Theorem 4.** *Let $Y_1, \dots, Y_p$ be independent standard Gaussian random variables and consider $x_1, \dots, x_n \in \{Y_1, \dots, Y_p\}$. Then*

$$\mathbb{E}[x_1 x_2 \dots x_n] = \sum_{\pi \in \mathcal{P}_2(n)} \prod_{(i,j) \in \pi} \mathbb{E}[x_i x_j] \tag{4.3}$$

Note that because it is linear in the variables $x_1, \dots, x_n$, 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.

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.

Importantly, the covariances in the Wick formula can be computed straightforwardly from (4.2):

$$\mathbb{E}[x_i x_j] = \begin{cases} 1 & \text{if } x_i = x_j \\ 0 & \text{if } x_i \neq x_j. \end{cases} \tag{4.4}$$

## 4.2 The genus expansion for Gaussian random matrices

Recall that the spectral moments are given by

$$\mathbb{E}\left[\frac{1}{n}\mathrm{Tr}M^k\right] = \sum_{i_1, i_2, \dots, i_k = 1}^{n} \frac{1}{n^{1+k/2}} \mathbb{E}[\mu_{i_1 i_2} \mu_{i_2 i_3} \dots \mu_{i_k i_1}]. \tag{4.5}$$

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 complex Hermitian Gausssian random matrices, for example, we can apply Wick's theorem to analyse this expression much more explicitly.

We have from Wick's theorem that

$$\mathbb{E}[\mu_{i_1 i_2}\mu_{i_2 i_3}\dots\mu_{i_k i_1}] = \sum_{\pi\in\mathcal{P}_2(k)}\prod_{(a,b)\in\pi}\mathbb{E}[\mu_{i_a,i_{a+1}}\mu_{i_b,i_{b+1}}] \tag{4.6}$$

where $i_{k+1} = i_1$. Hence, using (4.4),

$$\mathbb{E}\left[\frac{1}{n}\mathrm{Tr}M^k\right] = \sum_{\pi\in\mathcal{P}_2(k)}\sum_{i_1,i_2,\dots,i_k=1}^{n}\frac{1}{n^{1+k/2}}\prod_{(a,b)\in\pi}\delta_{i_a,i_{b+1}}\delta_{i_{a+1},i_b}, \tag{4.7}$$

where $\delta_{i,j}$ denotes the Kronecker $\delta$-symbol.

We can think of $\pi$ as a particular kind of permutation, 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 $\pi(b) = a$. Therefore

$$\prod_{(a,b)\in\pi}\delta_{i_a,i_{b+1}}\delta_{i_{a+1},i_b} = \prod_{a=1}^{k}\delta_{i_a,i_{\pi(a)+1}}. \tag{4.8}$$

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

$$\mathbb{E}\left[\frac{1}{n}\mathrm{Tr}M^k\right] = \sum_{\pi\in\mathcal{P}_2(k)}\sum_{i_1,i_2,\dots,i_k=1}^{n}\frac{1}{n^{1+k/2}}\prod_{a=1}^{k}\delta_{i_a,i_{\gamma\pi(a)}}. \tag{4.9}$$

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

$$\prod_{a=1}^{k}\delta_{i_a,i_{\gamma\pi(a)}} = \begin{cases}1 & \text{if } \mathbf{i} \text{ is constant on the cycles of } \gamma\pi\\ 0 & \text{otherwise.}\end{cases} \tag{4.10}$$

We therefore have that

$$\mathbb{E}\left[\frac{1}{n}\mathrm{Tr}M^k\right] = \frac{1}{n^{1+k/2}}\sum_{\pi\in\mathcal{P}_2(k)}\#\{\mathbf{i}:\{1,\dots,k\}\to\{1,\dots,n\} : \mathbf{i} \text{ is constant on the cycles of } \gamma\pi\}. \tag{4.11}$$

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 $\sigma$. Then we have

$$\mathbb{E}\left[\frac{1}{n}\mathrm{Tr}M^k\right] = \sum_{\pi\in\mathcal{P}_2(k)} n^{\#(\gamma\pi)-k/2-1}. \tag{4.12}$$

Equation (4.12) is an exact formula for the spectral moments of $M$. 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

$$\mathbb{E}\left[\frac{1}{n}\mathrm{Tr}M^{2m}\right] = \sum_{\pi\in\mathcal{P}_2(2m)} n^{\#(\gamma\pi)-m-1}. \tag{4.13}$$

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

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 $\tilde{S}_\pi$ obtained by this gluing procedure. The number of distinct vertices in the image $G_\pi$ of the $2m$-gon in $\tilde{S}_\pi$ 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 $\tilde{S}_\pi$, $\chi(\tilde{S}_\pi)$, is an even integer which may be defined as follows: if $G$ is *any* embedded polygonal complex in $\tilde{S}_\pi$, then

$$\chi(\tilde{S}_\pi) = (\text{number of vertices in } G) - (\text{number of edges in } G) + (\text{number of faces of } G). \tag{4.14}$$

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

$$\chi(\tilde{S}_\pi) = 2 - 2g(\tilde{S}_\pi). \tag{4.15}$$

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

$$2 - 2g(\tilde{S}_\pi) = \chi(\tilde{S}_\pi) = \#(\gamma\pi) - m + 1. \tag{4.16}$$

We therefore have, finally, that

$$\mathbb{E}\left[\frac{1}{n}\mathrm{Tr}M^{2m}\right] = \sum_{\pi \in \mathcal{P}_2(2m)} n^{-2g(\tilde{S}_\pi)} = \sum_{g \geq 0} \tau_g(m) n^{-2g} \tag{4.17}$$

where

$$\tau_g(m) = \#\{\text{genus-}g \text{ surfaces obtained by gluing together pairs of edges of a } 2m\text{-gon}\}. \tag{4.18}$$

This remarkable formula therefore implies that averaging $\mathrm{Tr}M^{2m}$ 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!

We see immediately that since $g \geq 0$,

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\mathrm{Tr}M^{2m}\right] = \tau_0(m) = \#\{\text{spheres obtained by gluing together pairs of edges of a } 2m\text{-gon}\} \tag{4.19}$$

and so we see that by Wigner's theorem

$$\#\{\text{spheres obtained by gluing together pairs of edges of a } 2m\text{-gon}\} = C_m \tag{4.20}$$

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.

# 5 The Marchenko-Pastur law for Wishart random matrices

The Marchenko-Pastur[17] law is the analogue for Wishart random matrices of the Wigner semicircle law for Hermitian matrices.

## 5.1 The Marchenko-Pastur Theorem

Let $X$ be a $p \times n$ matrix with entries $X_{ij}$ that are i.i.d. real random variables with $\mathbb{E}X_{ij} = 0$ and $\mathbb{E}X_{ij}^2 = 1$. Denote by $\sigma_n$ the $p \times p$ matrix

$$\sigma_n = \frac{1}{n} X X^{\mathrm{T}} \in \mathbb{R}^{p \times p} \tag{5.1}$$

and denote by $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_p$ the eigenvalues of $\sigma_n$.

Defining the random spectral distribution by

$$\mu_\sigma(x) := \frac{1}{p} \sum_{j=1}^{p} \Theta(\lambda_j - x), \tag{5.2}$$

The Marchenko-Pastur law asserts that

**Theorem 5.** *For $\sigma_n$ and $\mu_\sigma$ as defined above, when $p \to \infty$ and $n \to \infty$ such that $p/n \to \gamma \in (0, 1]$, $\mu_\sigma \to \mu$ in expectation and almost surely, where $\mu$ is the deterministic measure satisfying*

$$\frac{\mathrm{d}\mu}{\mathrm{d}x} = \begin{cases} \frac{1}{2\pi\gamma x}\sqrt{(a_+ - x)(x - a_-)} & \text{if } a_- \leq x \leq a_+ \\ 0 & \text{otherwise,} \end{cases} \tag{5.3}$$

*where $a_- = (1 - \sqrt{\gamma})^2$ and $a_+ = (1 + \sqrt{\gamma})^2$. When $\gamma > 1$, one needs to add $(1 - \gamma^{-1})\delta(x)$ to the right-hand side of (5.3), where $\delta(x)$ denotes a Dirac delta-function/mass at the origin.*

The additional factor $(1 - \gamma^{-1})\delta(x)$ which needs to be added to the right-hand side of (5.3) when $\gamma > 1$ is explained by the fact that, since the rank of $\sigma_n$ is the larger 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 (5.3) coincides with the semicircle law under the mapping $x \to x^2$.

We illustrate the Marchenko-Pastur law by showing the results of numerical experiments. These involved generating random matrices $\sigma_n$, plotting histograms of the eigenvalues, and comparing with (5.3). Figure 5 shows the result when $p = 10^3$ and $n = 10^4$, compared to (5.3) with $\gamma = 0.1$. Similarly Figure 6 shows the result when $p = 1000$ and $n = 2000$, compared to (5.3) with $\gamma = 0.5$.

One can prove the Marchenko-Pastur formula in a similar way to the proof of the semicircle law already described. The main steps are:

- proving that the spectral moments converge in expectation to the moments of the limiting deterministic Marchenko-Pastur distribution when $n \to \infty$;
- showing that the variance vanishes in the limit when $n \to \infty$, and then using the Borel-Cantelli lemma to conclude that the moments converge almost surely;
- demonstrating that the distribution is determined by its moments.

[17] Sometimes written Marčenko-Pastur.

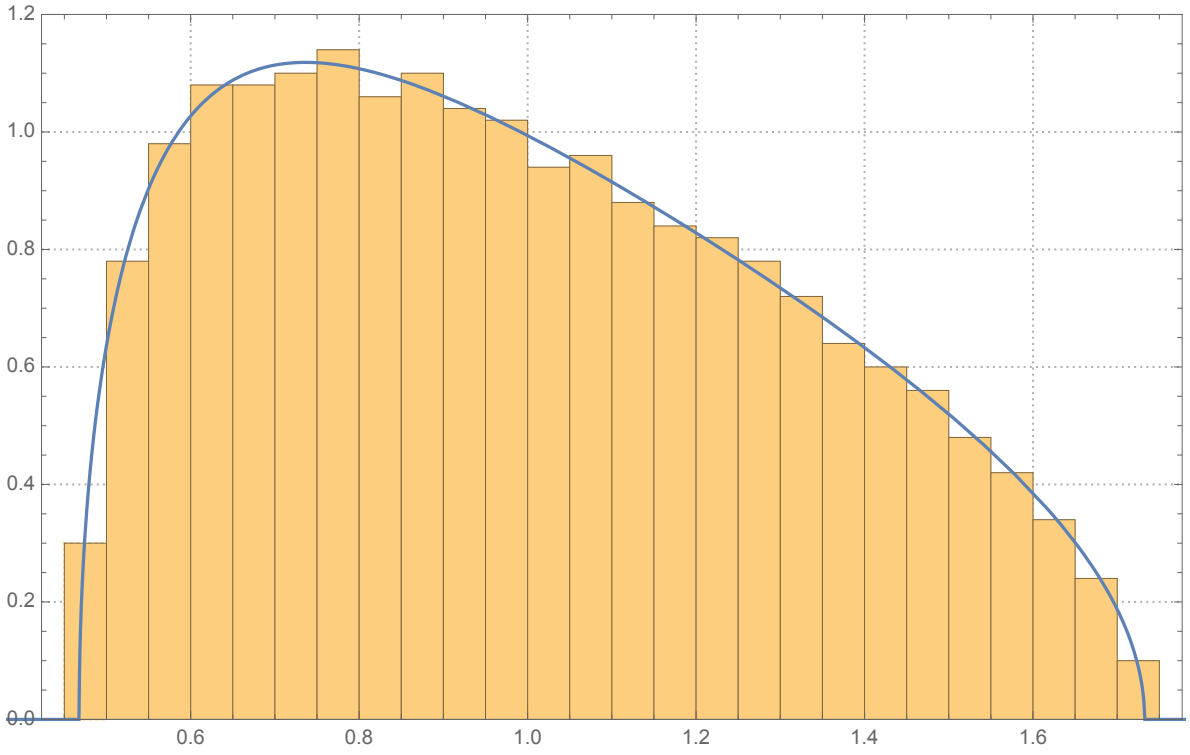


Figure 5: Eigenvalue density when $p = 10^3$ and $n = 10^4$, compared to (5.3) with $\gamma = 0.1$.

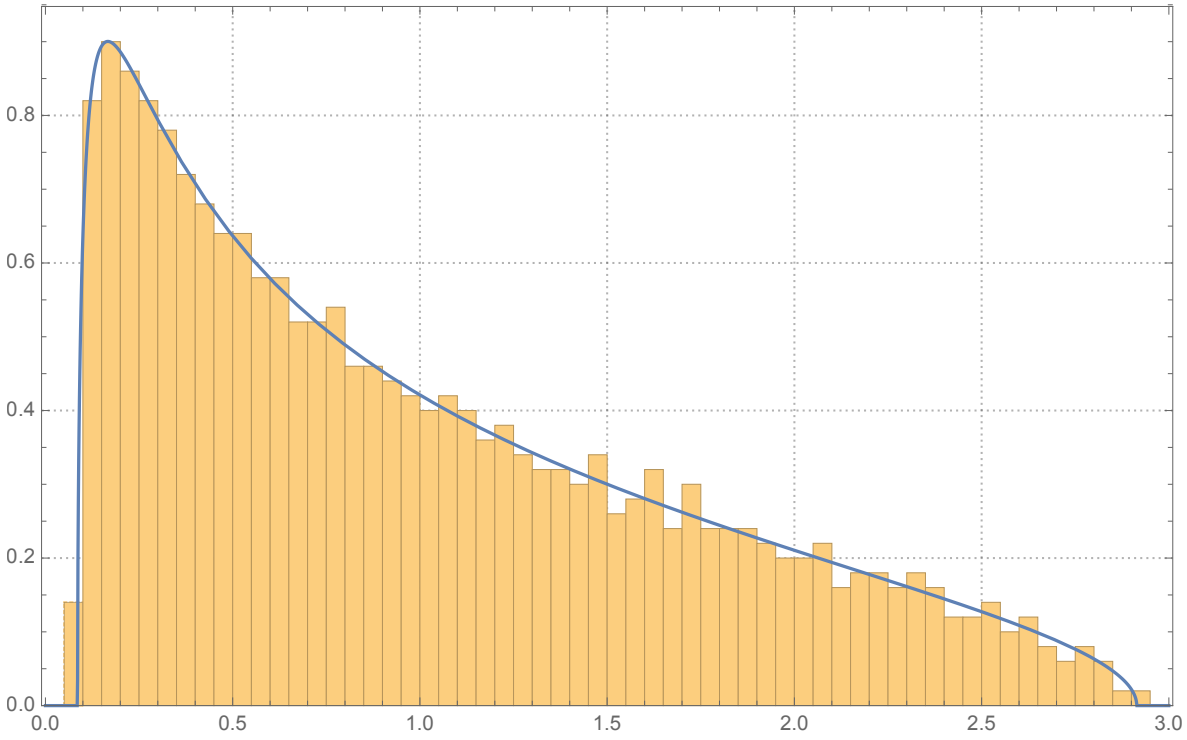


Figure 6: Eigenvalue density when $p = 1000$ and $n = 2000$, compared to (5.3) with $\gamma = 0.5$.

We will describe the proof of the first step: the convergence of the spectral moments to those of the limiting deterministic Marchenko-Pastur distribution in expectation. The second and third steps follow very similar lines to the corresponding ones in the proof of the semicircle law and so we shall not repeat them here. We focus on the case when $\gamma \leq 1$.

The moments of the Marchenko-Pastur distribution are, recalling that $a_- = (1 - \sqrt{\gamma})^2$ and $a_+ = (1 + \sqrt{\gamma})^2$, given by

$$\begin{aligned} \tilde{\alpha}_k &= \int_{a_-}^{a_+} x^k \frac{1}{2\pi\gamma x} \sqrt{(a_+ - x)(x - a_-)} \mathrm{d}x \\ &= \sum_{r=0}^{k-1} \frac{\gamma^r}{r+1} \binom{k}{r} \binom{k-1}{r}. \end{aligned} \tag{5.4}$$

To see this note that $a_- + a_+ = 2(1 + \gamma)$ and that $a_- a_+ = (1 - \gamma)^2$, and hence

$$\tilde{\alpha}_k = \frac{1}{2\pi} \int_{-2}^{2} (\sqrt{\gamma} y + 1 + \gamma)^{k-1} \sqrt{4 - y^2} \mathrm{d}y. \tag{5.5}$$

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

$$\binom{m+n}{k} = \sum_{r=0}^{k} \binom{m}{r} \binom{n}{k-r} \tag{5.6}$$

then leads immediately to the result.

The goal is to show that

$$\mathbb{E}\frac{1}{p}\sum_{i=1}^{p}\lambda_i^k = \mathbb{E}\frac{1}{p}\mathrm{Tr}\left(\frac{1}{n}XX^{\mathrm{T}}\right)^k \tag{5.7}$$

converges to (5.4) 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 and using the independence of these to identify the products that make the largest contributions. So,

$$\mathbb{E}\frac{1}{p}\mathrm{Tr}\left(\frac{1}{n}XX^{\mathrm{T}}\right)^k = \frac{1}{pn^k}\sum_{\mathcal{I}\mathcal{J}}\mathbb{E}\left[X_{i_1j_1}X_{i_2j_1}X_{i_2j_2}X_{i_3j_2}\dots X_{i_kj_k}X_{i_1j_k}\right] \tag{5.8}$$

where $\mathcal{I}\in[p]^k$ and $\mathcal{J}\in[n]^k$.

It is helpful to picture this in terms of cycles of length $2k$ on a directed bipartite graph with one set of vertices labelled by $i_1, i_2, \dots, i_k$ and the second set by $j_1, j_2, \dots, j_k$, with directed edges 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 $i_1$.

Since $\mathbb{E}X_{ij} = 0$ and the $X_{ij}$ are independent, the summand in (5.8) vanishes unless each edge appears at least twice in a cycle. Hence there can be at most $k$ edges and therefore $k+1$ vertices in the *skeleton* of the graph; i.e. in the subgraph consisting of the vertices and edges visited, with multiplicities removed. The number of vertices corresponds to the weight of the contributing cycle.

We consider first the case when the weight $w < k+1$. Let the cycle visit $l_{\mathcal{I}}$ $i$-vertices and $l_{\mathcal{J}}$ $j$-vertices (so $w = l_{\mathcal{I}} + l_{\mathcal{J}}$). The total number of ways of choosing these sites is bounded from above by a constant (i.e. a factor independent of $n$) times $p^{l_{\mathcal{I}}}n^{l_{\mathcal{J}}}$. Hence the contribution of all such terms in (5.8) is bounded from above by a constant (i.e. a factor independent of $n$) times $p^{l_{\mathcal{I}}}n^{l_{\mathcal{J}}}/pn^k$, 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 $k+1$ edges and $k$ vertices. The graph then is a double tree, namely a tree where vertices can be labelled either by $i$ or $j$ and these are visited alternately on the cycle, which passes through each edge exactly twice. So, if we start at an $i$-vertex, other $i$-vertices can be reached if and only if the number of steps taken is even, and $j$-vertices reached if and only if the number of steps taken is odd.

We call the vertices numbered in order of appearance, up to an overall shift in the numbers assigned, the *shape* of the double tree. The question then is: how many double trees are there with a given shape? The answer is that we need to choose $r+1$ $i$-vertices from $[p]$ and $k-r$ $j$-vertices from $[n]$, and there are

$$p(p-1)\dots(p-r)n(n-1)\dots(n-k+r+1) \tag{5.9}$$

ways of doing this. Note that if we set $p\sim\gamma n$, then as $n\to\infty$

$$p(p-1)\dots(p-r)n(n-1)\dots(n-k+r+1) = pn^k\gamma^r(1+O(1/n)), \tag{5.10}$$

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 $\mathcal{G}(r,k)$ denotes the number of different shapes of double trees with $r+1$ $i$-edges and $k-r$ $j$-edges, we have

$$\mathbb{E}\frac{1}{p}\sum_{i=1}^{p}\lambda_i^k = \sum_{r=0}^{k-1}\gamma^r\mathcal{G}(r,k)(1+O(1/n)). \tag{5.11}$$

We therefore need to show that

$$\mathcal{G}(r,k) = \frac{1}{r+1}\binom{k}{r}\binom{k-1}{r} \tag{5.12}$$

if we are to demonstrate that the spectral moments converge to (5.4).

In order to do this we establish a bijection with a set of paths of $2k$ steps, or equivalently a set of sequences, much as we did in the case of the semicircle law. In this case the paths (or sequences) are generated as follows. Take

$$s_m \in \begin{cases} \{-1, 0\} & \text{if } m \text{ is odd} \\ \{0, 1\} & \text{if } m \text{ is even,} \end{cases} \tag{5.13}$$

and take $s_{2k} = 0$. We then consider all sequences of steps $s_m$ constrained such that: (i) for any $t = 1, 2, \ldots, 2k$

$$\sum_{m=0}^{t} s_m \geq 0, \tag{5.14}$$

that is, the path never crosses 0; (ii) $\#\{m : s_m = 1\} = \#\{m : s_m = -1\} = r$, that is, there are exactly $r$ positive steps and $r$ negative steps; and (iii)

$$\sum_{m=0}^{2k} s_m = 0. \tag{5.15}$$

To see that this set of sequences is in one-to-one correspondence with the set of double trees that we wish to count, we start by constructing a mapping from sequences of this kind to double trees in the following way. If $m$ is odd and $s_m = 0$ go from the $i$-vertex to a new $j$-vertex, and if $s_m = -1$ go back to an adjacent, previously-constructed $j$-vertex. If $m$ is even and $s_m = 1$, go from a $j$-vertex to a new $i$-vertex, and if $s_m = 0$ go back to an adjacent, previously-constructed $i$-vertex.

In the reverse direction, consider a path on the double tree as described above, starting at an $i$-vertex. After an even number $2m$ of edge crossings one sits at an $i$-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 $j$-vertex, having just left an $i$-vertex. If this edge crossing is the last exit from the $i$-vertex in question, set $s_{2m-1} = -1$, and otherwise set $s_{2m-1} = 0$. It is easy to see that this construction gives sequences satisfying the constraints (ii) and (iii) set out above. We need, therefore, to check (i). 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 $\sum_{m=1}^{2t} s_m = 0$ and $s_{2t-1} = -1$. Now, we could use the sequence up to $2t$ 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.

In order to evaluate $\mathcal{G}(r, k)$, we therefore need to count the number of sequences $\{s_m\}_{m=1}^{2k}$. We begin by counting sequences that do not necessarily satisfy constraint (i). Recall that $s_{2k} = 0$. Since $s_{2k} \neq +1$ , we have to count the number of ways of allocating $r$ numbers $+1$ to $k - 1$ positions, and $r$ numbers $-1$ to $k$ positions. The total number of sequences not necessarily satisfying constraint (i) is therefore $\binom{k-1}{r}\binom{k}{r}$.

We now count the number of sequences that fail to satisfy constraint (i). For each of these there must exist a first $t$ such that $\sum_{m=1}^{2t-1} s_m = -1$. Given this, construct a new sequence $\{s'_m\}_{m=1}^{2k}$ in the following way. Set $s'_m = s_m$ for $m = 1, 2, \ldots, 2t - 1$, and $s'_{2k} = 0$. For $t \leq j \leq k - 1$ put $(s'_{2j}, s'_{2j+1}) = (1, -1)$ if $(s_{2j}, s_{2j+1}) = (1, -1)$, $(s'_{2j}, s'_{2j+1}) = (0, 0)$ if $(s_{2j}, s_{2j+1}) = (0, 0)$, $(s'_{2j}, s'_{2j+1}) = (1, 0)$ if $(s_{2j}, s_{2j+1}) = (0, -1)$, and $(s'_{2j}, s'_{2j+1}) = (0, -1)$ if $(s_{2j}, s_{2j+1}) = (1, 0)$. The set of sequences $\{s_m\}_{m=1}^{2k}$ that fail to satisfy constraint (i) is in bijection with the set of sequences $\{s'_m\}_{m=1}^{2k}$, but the latter can be counted straightforwardly to be the number of ways of choosing $r - 1$ out of $k - 1$ $+1$s, and $r + 1$ out of $k$ $-1$s. The number of ways of doing this is $\binom{k-1}{r-1}\binom{k}{r+1}$. Hence the number of sequences satisfying all of the conditions, and hence $\mathcal{G}(r, k)$, is

$$\binom{k-1}{r}\binom{k}{r} - \binom{k-1}{r-1}\binom{k}{r+1} = \frac{1}{r+1}\binom{k}{r}\binom{k-1}{r}. \tag{5.16}$$

□

## 5.2 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 correlation matrix and so 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 [8], and for specific applications to mathematical finance (where this is as an important technique), see [3, 9].

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 (i.e. at the edge of the support).

# 6 Stieltjes and $\mathcal{R}$-transforms

## 6.1 The Stieltjes transform and the semicircle law

Let $\mu$ be a non-negative finite measure on $\mathbb{R}$.

**Definition.** The Stieltjes transform of $\mu$ is, for $z \in \mathbb{C}\backslash\mathbb{R}$

$$g_\mu(z) = \int_{\mathbb{R}} \frac{1}{x-z} \mathrm{d}\mu(x). \tag{6.1}$$

If all of the moments of $\mu$, $m_k = \int_{\mathbb{R}} x^k \mathrm{d}\mu(x)$, are bounded[18], then for sufficiently large $z$

$$g_\mu(z) = -\int_{\mathbb{R}} \frac{1}{z(1-x/z)} \mathrm{d}\mu(x) = -\frac{1}{z}\sum_{k=0}^{\infty} \frac{m_k}{z^k} \tag{6.2}$$

and so $g_\mu(z)$ can be thought of as the moment generating function of $\mu$.

Note that when $\mathrm{Im} z \neq 0$, it follows immediately from the fact that $|(x-z)|^{-1} \leq |\mathrm{Im}(z)|^{-1}$ that $|g_{\mu_M}(z)| \leq |\mathrm{Im}(z)|^{-1}$ when $\mu$ is a probability measure. Moreover, when $\mathrm{Im} z > 0$, $g_\mu(z)$ is continuous and analytic.

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

When the measure is the empirical spectral distribution function $\mu_M(x)$ defined in (3.27),

$$g_{\mu_M}(z) = \frac{1}{n}\sum_j \frac{1}{\lambda_j - z} = \frac{1}{n}\mathrm{Tr}\frac{1}{M - zI}. \tag{6.3}$$

Note for future reference that it follows from this formula that if $\mathrm{Im} z > 0$, then $\mathrm{Im} g_{\mu_M}(z) > 0$. Also, if $\mu_M$ is compactly supported, then as $z \to \infty$

$$g_{\mu_M}(z) \sim -\frac{1}{z}. \tag{6.4}$$

When $\mu$ is the measure corresponding to the semicircle law (3.3),

$$g_\sigma(z) = \frac{1}{2\pi}\int_{-2}^{2} \frac{1}{x-z}\sqrt{4-x^2}\mathrm{d}x = \frac{1}{\pi}\int_0^\pi \frac{2}{2\cos y - z}\sin^2 y \mathrm{d}y. \tag{6.5}$$

Setting $w = e^{iy}$ then gives

$$g_\sigma(z) = -\frac{1}{4\pi i}\oint_{|w|=1} \frac{(w^2-1)^2}{w^2(w^2+1-zw)}\mathrm{d}w. \tag{6.6}$$

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 gives

$$g_\sigma(z) = \frac{-z + \sqrt{z^2-4}}{2} \tag{6.7}$$

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

$$g_\sigma(z) = -\frac{1}{z} - \frac{1}{z}g_\sigma(z)^2 \tag{6.8}$$

with (6.7) being the solution satisfying $\mathrm{Im} g_\sigma(z) > 0$ when $\mathrm{Im} z > 0$.

[18] or, more generally, if $m_k = O(C^k)$ for some constant $C > 0$ when $k \to \infty$

(We remark in passing that for the Marchenko-Pastur distribution, the formula analogous to (6.8) is

$$zg(z)^2 + (z - 1 + \gamma)g(z) + \gamma = 0, \tag{6.9}$$

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

Importantly, the Stieltjes transform can be inverted in the following way. Let us define a sub-probability measure to be a measure $\mu$ on $\mathbb{R}$ such that $\mu(\mathbb{R}) \leq 1$. For such a measure consider an interval $[a, b]$ such that $\mu\{a\} = \mu\{b\} = 0$. Then

$$\mu[a, b] = \lim_{\eta \to 0} \int_a^b \frac{1}{\pi} \mathrm{Im} g_\mu(x + i\eta) dx. \tag{6.10}$$

The proof of this inversion formula follows from interchanging the order of integration in

$$\int_a^b \frac{1}{\pi} \mathrm{Im} g_\mu(x + i\eta) dx = \int_a^b \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\eta}{(\lambda - x)^2 + \eta^2} \mathrm{d}\mu(\lambda) \mathrm{d}x. \tag{6.11}$$

This gives

$$\int_a^b \frac{1}{\pi} \mathrm{Im} g_\mu(x + i\eta) dx = \int_{-\infty}^{\infty} V(\lambda) \mathrm{d}\mu(\lambda) \tag{6.12}$$

where

$$V(\lambda) = \frac{1}{\pi} \left[ \arctan\left( \frac{b - \lambda}{\eta} \right) - \arctan\left( \frac{a - \lambda}{\eta} \right) \right] = \frac{1}{\pi} \arctan\left( \frac{\eta(b - a)}{(b - \lambda)(a - \lambda) + \eta^2} \right). \tag{6.13}$$

As $\eta \to 0$, $V(\lambda) \to 1$ for $\lambda \in (a, b)$, $V(a) = V(b) \to 1/2$, and $V(\lambda) \to 0$ otherwise. Since $V(\lambda)$ is uniformly bounded and, for any $\eta$, $V(\lambda) \sim \eta/\lambda^2$ as $\lambda \to \infty$, it can be majorized uniformly in $\eta$ by a positive integrable function. Hence it follows from the dominated convergence theorem that

$$\int_{-\infty}^{\infty} V(\lambda) \mathrm{d}\mu(\lambda) \to \mu[a, b]. \tag{6.14}$$

Note that this means that if two sub-probability measures $\mu$ and $\nu$ have $g_\mu = g_\nu$, then $\mu = \nu$.

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; instead the Stieltjes transform approach is more rooted in analysis and probability theory. We sketch here the proof of the semicircle law for a specific class of Wigner random matrices – those with Gaussian entries. 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 with

$$g_{\mu_M}(z) = \frac{1}{n} \mathrm{Tr} \frac{1}{M - zI} := \frac{1}{n} \mathrm{Tr} G_M(z) \tag{6.15}$$

where $G_M$ is termed the *resolvent* of $M$. For this calculation we will specialise to the case where $M$ is a real symmetric matrix with entries that are i.i.d. Gaussian random variables with $\mathbb{E}M_{ij} = 0$ and $\mathbb{E}M_{ij}^2 = 1/n$. (Note the scaling of the variance, which is the one that leads to the semicircle law when $n \to \infty$ and which corresponds to scaling the eigenvalues with $1/\sqrt{n}$ with respect to when the variance is 1.) Clearly

$$\mathbb{E} g_{\mu_M}(z) = \frac{1}{n} \mathbb{E} \mathrm{Tr} G_M(z). \tag{6.16}$$

We now use the fact, which follows straightforwardly from the definition of the resolvent, that

$$G_{M+A}(z) - G_M(z) = -G_{M+A}(z) A G_M(z) \tag{6.17}$$

and so, setting $A = -M$

$$G_M(z) = -\frac{1}{z}I + \frac{1}{z}MG_M(z). \tag{6.18}$$

Another consequence of (6.17) , which follows from setting, for a fixed pair $i$ and $j$, $A_{ij} = A_{ji} = \epsilon$ and $A_{uv} = 0$ otherwise and then taking the limit as $\epsilon \to 0$, is that

$$\frac{\partial (G_M)_{uv}}{\partial M_{ij}} = -(G_M)_{ui}(G_M)_{jv} - (G_M)_{uj}G_M)_{iv}. \tag{6.19}$$

We note now that if $x$ is a normal random variable with mean zero and variance $\sigma^2$, then for any function $f$ that is differentiable and grows no faster than a polynomial, it follows from integrating by parts that $\mathbb{E}xf(x) = \sigma^2 \mathbb{E}f'(x)$.

Combining the last results together, we then have that

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\mathrm{Tr}G_M(z) &= -\frac{1}{z} + \frac{1}{nz}\mathbb{E}\mathrm{Tr}(MG_M) \\
&= -\frac{1}{z} + \frac{1}{nz}\sum_{ij}\mathbb{E}\left(M_{ij}(G_M)_{ji}\right) \\
&= -\frac{1}{z} + \frac{1}{n^2 z}\sum_{ij}\mathbb{E}\left(\frac{\partial (G_M)_{ji}}{\partial M_{ij}}\right) \\
&= -\frac{1}{z} + \frac{1}{n^2 z}\sum_{ij}\mathbb{E}\left[-(G_M)_{ji}(G_M)_{ij} - (G_M)_{jj}(G_M)_{ii}\right] \\
&= -\frac{1}{z} - \frac{1}{n^2 z}\mathbb{E}\left[\mathrm{Tr}G_M^2\right] - \frac{1}{z}\mathbb{E}\left[\left(\frac{1}{n}\mathrm{Tr}G_M\right)^2\right].
\end{aligned} \tag{6.20}$$

Now note that

$$\begin{aligned}
\left|\frac{1}{n^2 z}\left[\mathrm{Tr}G_M^2\right]\right| &= \frac{1}{n|z|}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|\lambda_i - z|^2}\right| \\
&\leq \frac{1}{n|z||\mathrm{Im}z|^2} \\
&\to 0
\end{aligned} \tag{6.21}$$

when $n \to \infty$.

Hence we have that

$$\mathbb{E}g_{\mu_M}(z) = -\frac{1}{z} - \frac{1}{z}\mathbb{E}\left(g_{\mu_M}(z)^2\right) + E(z,n) \tag{6.22}$$

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

We now want to establish that we can replace $\mathbb{E}\left(g_{\mu_M}(z)^2\right)$ with $\left(\mathbb{E}g_{\mu_M}(z)\right)^2$, with an error that tends to zero as $n \to \infty$. Note that $g_{\mu_M}(z)$ depends on $n$, because $M$ is an $n \times n$ matrix. In the following calculation the $n$-dependence will be important, so we record that dependence explicitly, augmenting the notation to read $g_{\mu_{M_n}}(z)$. The strategy will be to show that replacing $M_n$ by an $(n-1) \times (n-1)$ minor does not change the Stieltjes transform appreciably when $n$ is large – i.e. that $g_{\mu_{M_{n-1}}}(z)$ is close to $g_{\mu_{M_n}}(z)$. This will allow us to use an estimate to establish that $g_{\mu_{M_n}}(z)$ lies sufficiently close to $\mathbb{E}\left(g_{\mu_{M_n}}(z)\right)$ for us to prove what we need.

Let $z = a + ib$, with $b > 0$. We denote the eigenvalues of $M_n$ by $\{\lambda_j^{(n)}\}_{j=1}^n$ and the eigenvalues of $M_{n-1}$ by $\{\lambda_j^{(n-1)}\}_{j=1}^{n-1}$. Now, Cauchy's Interlace Theorem implies that the eigenvalues of $M_n$ and those of $M_{n-1}$ interlace, that is $\lambda_j^{(n)} \leq \lambda_j^{(n-1)} \leq \lambda_{j+1}^{(n)}$ for $j = 1, \ldots, n-1$. Hence the difference

$$\sum_{j=1}^{n-1}\frac{b}{(\lambda_j^{(n-1)} - a)^2 + b^2} - \sum_{j=1}^{n}\frac{b}{(\lambda_j^{(n)} - a)^2 + b^2} \tag{6.23}$$

is bounded in $n$ , 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. 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

$$g_{\mu_{M_n}}(z) = g_{\mu_{M_{n-1}}}(z) + O\left(\frac{1}{n}\right). \tag{6.24}$$

$g_{\mu_{M_n}}(z)$ is therefore said to be *stable* in $n$. Note as well that the right-hand side of this equation depends only on the top left $(n-1) \times (n-1)$ minor of $M$ and is independent of its $n$th row and $n$th column. We can therefore apply the following inequality, due to McDiarmid.

**Lemma 6** (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 $i$,*

$$|F(X_1, \ldots, X_n) - F(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)| \leq c_i \tag{6.25}$$

*for all $X_i, X_i' \in R_i$. Then for any $\kappa > 0$*

$$\mathbb{P}\left(|F(X_1, \ldots, X_n) - \mathbb{E}F(X_1, \ldots, X_n)| \geq \kappa\sigma\right) \leq C\exp(-c\kappa^2) \tag{6.26}$$

*for some $C > 0$ and $c > 0$, where $\sigma^2 = \sum_{i=1}^n c_i^2$.*

We can apply this inequality in the following way. First note that (6.24) still applies if we resample the $n$th row and $n$th column. Denoting the Stieltjes transform of the resampled matrix by $g_{\mu_{M'}}(z)$, it follows from (6.24) that $g_{\mu_M}(z) = g_{\mu_{M'}}(z) + O(1/n)$. Moreover, this applies to resampling any row and column with the same label. Therefore, applying McDiarmid's inequality with $X_j = (M_{j,j}, M_{j,j+1}, \ldots, M_{j,n})$ gives

$$\mathbb{P}(|g_{\mu_{M_n}}(z) - \mathbb{E}(g_{\mu_{M_n}}(z)| \geq \kappa/\sqrt{n}) \leq C\exp(-c\kappa^2). \tag{6.27}$$

This means that as $n \to \infty$ $g_{\mu_{M_n}}(z)$ lies sufficiently close to $\mathbb{E}\left(g_{\mu_{M_n}}(z)\right)$ that we can replace $\mathbb{E}\left(g_{\mu_M}(z)^2\right)$ with $\left(\mathbb{E}g_{\mu_M}(z)\right)^2$ in (6.22). Hence we have that

$$\mathbb{E}g_{\mu_M}(z) = -\frac{1}{z} - \frac{1}{z}\left(\mathbb{E}g_{\mu_M}(z)\right)^2 + E'(z, n) \tag{6.28}$$

where $E'(z, n) \to 0$ as $n \to \infty$. Hence $\mathbb{E}g_{\mu_M}(z)$ converges to the solution of

$$s(z) = -\frac{1}{z} - \frac{1}{z}s(z)^2 \tag{6.29}$$

which we have already seen is the Stieltjes transform of the semicircle distribution, once one specifies that $\mathrm{Im}s(z) > 0$ when $\mathrm{Im}z > 0$. Therefore, because the Stieltjes transform is invertible, we see that the empirical spectral measure converges weakly to the semicicle distribution when $n \to \infty$ in expectation.

Finally, taking $\kappa = n^{1/3}$ in (6.27) and applying the Borel-Cantelli lemma, we see that almost surely

$$g_{\mu_{M_n}}(z) - \mathbb{E}[g_{\mu_{M_n}}(z)] \leq O(n^{-1/6}) \tag{6.30}$$

for all large $n$. Therefore $g_{\mu_{M_n}}(z) - \mathbb{E}[g_{\mu_{M_n}}(z)] \to 0$ almost surely for all $z$ with $\mathrm{Im}z > 0$.

Note that one can view (6.27) and (6.30) as examples of concentration of measure in the context of the Stieltjes transform.

## 6.2 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

$$B_\mu(-g_\mu(z)) = z, \tag{6.31}$$

is sometimes referred to as the *Blue function*.

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

$$R_\mu(z) = B_\mu(z) - \frac{1}{z} \tag{6.32}$$

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

$$R_{sc}(z) = z \tag{6.33}$$

and similarly for the Marchenko-Pastur distribution

$$R_{MP}(z) = \frac{1}{1-\gamma z}. \tag{6.34}$$

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.

# 7 Eigenvalue statistics

In this section we will focus on matrices from the GOE, the GUE, and the CUE. We start with the GOE and the GUE and turn to the CUE at the end of the section.

We have seen that GOE and GUE matrices satisfy the semicircle law, so that if we take the eigenvalues of a GOE or a GUE matrix (in which the matrix elements have a variance independent of the matrix size $n$) 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 on the scale of their mean spacing? This is a much smaller scale than that for which the semicircle law holds. It might be assumed that if the matrix elements are i.i.d. random variables, then the eigenvalues will also behave like i.i.d. random variables on this scale; that is, that they behave like a *Poisson point process*. We shall see, however, that this is not the case, and that instead the eigenvalues are strongly correlated in an interesting and significant way. This is an important feature of random matrices.

## 7.1 Poisson point processes

A Poisson point process on $\mathbb{R}$ with parameter $\lambda > 0$ corresponds to a set of random points such that the probability that the number of points in an interval $(a, b]$, $N(a, b)$, is exactly equal to $k$ is

$$\mathbb{P}(N(a,b) = k) = \frac{[\lambda(b-a)]^k}{k!} \exp[-\lambda(b-a)]. \tag{7.1}$$

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

The probability density $P_{\text{Poiss}}(s)$ that the spacing between two nearest-neighbouring points is $s$ may be computed as follows. Let us fix an eigenvalue at some point. $P_{\text{Poiss}}(s)\mathrm{d}s$ is then the probability that the interval of length $s$ is empty, and that there is an eigenvalue in the next infinitesimal interval $\mathrm{d}s$. This is

$$P_{\text{Poiss}}(s)\mathrm{d}s = \exp[-\lambda s]\lambda \mathrm{d}s. \tag{7.2}$$

Clearly

$$\int_0^\infty P_{\text{Poiss}}(s)\mathrm{d}s = 1. \tag{7.3}$$

If now we choose to set the mean spacing to be 1, i.e.

$$\int_0^\infty s P_{\text{Poiss}}(s)\mathrm{d}s = 1, \tag{7.4}$$

this corresponds to setting $\lambda = 1$, in which case $P_{\text{Poiss}}(s) = e^{-s}$. Hence the probability to find nearest-neighbouring random points a distance $s$ apart is maximal at $s = 0$ and decreases as $s$ increases.

We shall later want to compare this to how the eigenvalues of GOE and GUE matrices are distributed.

## 7.2 Joint eigenvalue density

We have seen earlier that the GOE and GUE measures can be expressed, up to a normalization constant, in the form

$$e^{-\frac{\beta}{4}\mathrm{Tr}(M^2)}\mathrm{d}M \tag{7.5}$$

where for the GOE $\beta = 1$, for the GUE $\beta = 2$, and $\mathrm{d}M$ is the Lebesgue measure on the appropriate space of matrix entries. Our aim now is to re-express this in terms of the Lebesgue measure on the space of the eigenvalues. Specifically, denoting the eigenvalues of a matrix $M$ by $\lambda_1, \lambda_2, \ldots, \lambda_n$, define a *class function* of $M$ to be a function, $f(M) = f(\lambda_1, \ldots, \lambda_n)$, that depends only on the eigenvalues and which is symmetric with respect to permutations of its arguments. The following theorem gives a formula for averaging class functions with respect to the GUE and the GOE.

**Theorem 7.** *The average of $f(\lambda_1, \ldots, \lambda_n)$ over the GUE and GOE can be expressed as the multiple integral*

$$\mathbb{E}[f(\lambda_1, \lambda_2, \ldots, \lambda_n)] = c_n^{(\beta)} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\lambda_1, \ldots, \lambda_n) \prod_{i<j} |\lambda_i - \lambda_j|^\beta e^{-\frac{\beta}{4}\sum_i \lambda_i^2} \mathrm{d}\lambda_1 \ldots \mathrm{d}\lambda_n \tag{7.6}$$

*where $\beta = 1$ 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:

$$\frac{1}{c_n^{(\beta)}} = (2\pi)^{n/2} \beta^{-n/2-\beta n(n-1)/4} [\Gamma(1+\beta/2)]^{-n} \prod_{j=1}^{n} \Gamma(1+\beta j/2) \tag{7.7}$$

although we shall not make explicit use of this fact and do not include the proof here.

In the case of the GUE, the idea behind the proof of Theorem 7 is based on the observation that 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 $\lambda_1, \ldots, \lambda_n$.

We first note that when the eigenvalues of $M$ are distinct, the decomposition $M = UDU^\dagger$ is unique up to (i) permuting the eigenvalues, and (ii) multiplying $U$ by a diagonal matrix with entries $e^{i\theta_1}, e^{i\theta_2}, \ldots, e^{i\theta_n}$, and that this holds almost surely in the GUE (and the GOE). To see this, observe first that the joint distribution of the entries of $M$ has a smooth density. Since the eigenvalues are continuous functions of the matrix entries, and because the event of two eigenvalues coinciding has codimension $\geq 1$ in the $n$-dimensional variety of eigenvalues, it follows that the eigenvalues are almost surely distinct [19]. So we can ignore this situation.

Let $y_1, \ldots, y_n, w_1, \ldots, w_{n(n-1)}$ be the local parameters on $(\mathbb{R}^n/S_n) \times (U(n)/\mathbb{T}^n)$. Our goal now is to compute the Jacobian $\det\left(\frac{\partial M_{ij}}{\partial y_\alpha}, \frac{\partial M_{ij}}{\partial w_\beta}\right)$. To this end we write $M$ as an element of $\mathbb{R}^{n^2}$, setting

$$\phi(M) = \left(\frac{M_{11}}{\sqrt{2}}, \frac{M_{22}}{\sqrt{2}}, \ldots, \frac{M_{nn}}{\sqrt{2}}, \mathrm{Re}M_{12}, \mathrm{Im}M_{12}, \mathrm{Re}M_{13}, \mathrm{Im}M_{13}, \ldots, \mathrm{Re}M_{n-1,n}, \mathrm{Im}M_{n-1,n}\right). \tag{7.8}$$

It is easily seen that

$$|\phi(M)|^2 = \frac{1}{2}\mathrm{Tr}[M^2] = \frac{1}{2}\mathrm{Tr}[(UMU^\dagger)^2] = |\phi(UMU^\dagger)|^2 \tag{7.9}$$

for any $n \times n$ unitary matrix $U$, because the trace is invariant under unitary conjugation. Consequently, the linear transformation $\tau_U : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$, $\tau_U(y) = \phi(U^\dagger \phi^{-1}(y) U)$ is isometric; that is $\det \tau_U = 1$. Hence

$$\det\left(\frac{\partial M_{ij}}{\partial y_\alpha}, \frac{\partial M_{ij}}{\partial w_\beta}\right) = \det\left(\tau_U\left(\frac{\partial M_{ij}}{\partial y_\alpha}, \frac{\partial M_{ij}}{\partial w_\beta}\right)\right). \tag{7.10}$$

Clearly we can also think of $\tau_U$ as acting on matrices $X$ according to $\tau_U(X) = \tau_U(\phi(X))$. Now,

$$\tau_U\left(\frac{\partial M}{\partial y_i}\right) = \tau_U\left(U\frac{\partial D}{\partial y_i}U^\dagger\right) = \frac{\partial D}{\partial y_i}. \tag{7.11}$$

[19] 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 $M$, $\det(M - xI)$, 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 $M$.

Note that acting with $\phi$ on the right-hand side gives a vector with all $n^2$ coordinates except the $i$th equal to zero. We also have that

$$\tau_U\left(\frac{\partial M}{\partial w_\beta}\right) = \tau_U\left(\frac{\partial U}{\partial w_\beta}DU^\dagger + UD\frac{\partial U^\dagger}{\partial w_\beta}\right) = U^\dagger\frac{\partial U}{\partial w_\beta}D + D\frac{\partial U^\dagger}{\partial w_\beta}U. \tag{7.12}$$

Since $U^\dagger U = I$, we have, differentiating with respect to $w_\beta$,

$$U^\dagger\frac{\partial U}{\partial w_\beta} + \frac{\partial U^\dagger}{\partial w_\beta}U = 0. \tag{7.13}$$

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

$$\tau_U\left(\frac{\partial M}{\partial w_\beta}\right) = T_\beta D - DT_\beta. \tag{7.14}$$

Bearing in mind that $D$ is diagonal,

$$\left(\tau_U\left(\frac{\partial M}{\partial w_\beta}\right)\right)_{ij} = (T_\beta)_{ij}(y_j - y_i). \tag{7.15}$$

Therefore, the matrix $\tau_U\left(\frac{\partial M_{ij}}{\partial y_\alpha}, \frac{\partial M_{ij}}{\partial w_\beta}\right)$ has the following form

$$\begin{pmatrix} I_n & 0_n & 0_n & 0_n & \dots \\ 0_n & \mathrm{Re}(T_1)_{12}(y_2-y_1) & \mathrm{Im}(T_1)_{12}(y_2-y_1) & \mathrm{Re}(T_1)_{13}(y_3-y_1) & \dots \\ 0_n & \mathrm{Re}(T_2)_{12}(y_2-y_1) & \mathrm{Im}(T_2)_{12}(y_2-y_1) & \mathrm{Re}(T_2)_{13}(y_3-y_1) & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \tag{7.16}$$

where $0_n$ is the $n \times n$ zero matrix.

Hence, computing the determinant, we have that

$$\det\left(\frac{\partial M_{ij}}{\partial y_\alpha}, \frac{\partial M_{ij}}{\partial w_\beta}\right) = \prod_{i<j}(y_j - y_i)^2 \det\begin{pmatrix} \mathrm{Re}(T_1)_{12} & \mathrm{Im}(T_1)_{12} & \dots & \\ \mathrm{Re}(T_2)_{12} & \mathrm{Im}(T_2)_{12} & \dots & \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}. \tag{7.17}$$

The point is that the problem has now factorized and we are left with a product of $\prod_{i<j}(y_j - y_i)^2$ and the determinant, which is a function only of the $w$-variables. Hence in computing $\mathbb{E}[f(\lambda_1, \lambda_2, \dots, \lambda_n)]$, 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 $y_1, \dots, y_n$, picking up a factor $\prod_{i<j}(y_j - y_i)^2$. 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), $y_1, \dots, y_n$ 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 $\det\left(\frac{\partial M_{ij}}{\partial y_\alpha}, \frac{\partial M_{ij}}{\partial w_\beta}\right)$. 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 (7.16) but missing the columns involving $\mathrm{Im}(T_\beta)$ because the orthogonal matrices can be taken to be real valued. Therefore each factor $(y_j - y_i)$ appears only once and the Jacobian in this case factorizes as the product of $\prod_{i<j}|y_j - y_i|$ and a function only of the $w$-variables. Hence again in computing $\mathbb{E}[f(\lambda_1, \lambda_2, \dots, \lambda_n)]$, 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 $y_1, \dots, y_n$, picking up a factor $\prod_{i<j}|y_j - y_i|$. This proves the theorem for the GOE.

We turn next to the Circular Unitary Ensemble of random unitary matrices $M$. In this case too the matrices in question can be diagonalized by a unitary transformation: $M = UDU^\dagger$, where $D$ is diagonal with entries $e^{i\theta_1}, e^{i\theta_2}, \dots, e^{i\theta_n}$. The ensemble itself is, by construction, invariant under unitary transformations. A calculation that is essentially the same as for the GUE then gives, for a class function $f(\theta_1, \dots, \theta_n)$, that

**Theorem 8.** *The average of $f(\theta_1, \dots, \theta_n)$ over the CUE can be expressed as the multiple integral*

$$\mathbb{E}[f(\theta_1, \dots, \theta_n)] = \frac{1}{(2\pi)^n n!} \int_0^{2\pi} \cdots \int_0^{2\pi} f(\theta_1, \dots, \theta_n) \prod_{j<k} |e^{i\theta_j} - e^{i\theta_k}|^2 \mathrm{d}\theta_1 \dots \mathrm{d}\theta_n. \tag{7.18}$$

This is known as the *Weyl integration formula.*

Finally, for the Wishart Ensemble, the corresponding result is[20]

**Theorem 9.** *The average of $f(\lambda_1, \dots, \lambda_p)$ over the Wishart Ensemble can be expressed as the multiple integral*

$$\mathbb{E}[f(\lambda_1, \lambda_2, \dots, \lambda_p)] = c_{n,p}^{(\mathrm{Wishart})} \int_0^\infty \cdots \int_0^\infty f(\lambda_1, \dots, \lambda_p) \prod_{i<j} |\lambda_i - \lambda_j| \prod_k \lambda_k^{(n-p-1)/2} e^{-\frac{1}{2}\lambda_k} \mathrm{d}\lambda_1 \dots \mathrm{d}\lambda_p \tag{7.19}$$

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

One of the most important consequences of Theorem 7, Theorem 8 and Theorem 9 is that the eigenvalues of random matrices repel each other. It is clear from Theorem 7 that the probability that $\lambda_i - \lambda_j \in (s, s + \mathrm{d}s)$ vanishes like $s^\beta$ as $s \to 0$; so it vanishes quadratically for the GUE and linearly for the GOE. The same quadratic repulsion may be seen in Theorem 8 and Theorem 9. Note that this is unlike the Poisson point process, where the probability to find nearest-neighbouring random points a distance $s$ apart is maximal at $s = 0$.

## 7.3 The method of orthogonal polynomials for the GUE

For the GUE, the integrand in Theorem 7 involves $\prod_{i<j} |\lambda_i - \lambda_j|^2$. It is a key observation that the latter factor can be written as a determinant. Specifically, it is the square of the *Vandermonde determinant*:

$$\prod_{i<j} |\lambda_i - \lambda_j|^2 = \begin{vmatrix} 1 & 1 & \dots & 1 \\ \lambda_1 & \lambda_2 & \dots & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & \dots & \lambda_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \dots & \lambda_n^{n-1} \end{vmatrix}^2. \tag{7.20}$$

(To see this, note that the determinant is a homogeneous polynomial in the variables $\lambda_1, \lambda_2, \dots, \lambda_n$ 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 $\prod_{i<j}(\lambda_j - \lambda_i)$.)

It is a second key observation that by combining rows in the determinant we also have

$$\prod_{i<j} |\lambda_i - \lambda_j|^2 = \begin{vmatrix} p_0(\lambda_1) & p_0(\lambda_2) & \dots & p_0(\lambda_n) \\ p_1(\lambda_1) & p_1(\lambda_2) & \dots & p_1(\lambda_n) \\ p_2(\lambda_1) & p_2(\lambda_2) & \dots & p_2(\lambda_n) \\ \vdots & \vdots & \ddots & \vdots \\ p_{n-1}(\lambda_1) & p_{n-1}(\lambda_2) & \dots & p_{n-1}(\lambda_n) \end{vmatrix}^2 \tag{7.21}$$

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

where $p_k$ is a monic polynomial of degree $k$. Hence the integrand in Theorem 7 is the function we wish to average, $f(\lambda_1, \lambda_2, \ldots, \lambda_n)$, multiplied by

$$\begin{vmatrix} p_0(\lambda_1)e^{-\frac{1}{4}\lambda_1^2} & p_0(\lambda_2)e^{-\frac{1}{4}\lambda_2^2} & \ldots & p_0(\lambda_n)e^{-\frac{1}{4}\lambda_n^2} \\ p_1(\lambda_1)e^{-\frac{1}{4}\lambda_1^2} & p_1(\lambda_2)e^{-\frac{1}{4}\lambda_2^2} & \ldots & p_1(\lambda_n)e^{-\frac{1}{4}\lambda_n^2} \\ p_2(\lambda_1)e^{-\frac{1}{4}\lambda_1^2} & p_2(\lambda_2)e^{-\frac{1}{4}\lambda_2^2} & \ldots & p_2(\lambda_n)e^{-\frac{1}{4}\lambda_n^2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n-1}(\lambda_1)e^{-\frac{1}{4}\lambda_1^2} & p_{n-1}(\lambda_2)e^{-\frac{1}{4}\lambda_2^2} & \ldots & p_{n-1}(\lambda_n)e^{-\frac{1}{4}\lambda_n^2} \end{vmatrix}^2 . \tag{7.22}$$

This is true for any monic polynomials $p_k$. There is, however, one choice that considerably simplifies the calculations we shall want to perform: we henceforth take $\{p_k(x)\}$ to be orthogonal with respect to the measure $e^{-x^2/2}\mathrm{d}x$. Specifically, setting

$$\rho(\lambda_1, \ldots, \lambda_n) = c_n^{(2)} \begin{vmatrix} p_0(\lambda_1)e^{-\frac{1}{4}\lambda_1^2} & p_0(\lambda_2)e^{-\frac{1}{4}\lambda_2^2} & \ldots & p_0(\lambda_n)e^{-\frac{1}{4}\lambda_n^2} \\ p_1(\lambda_1)e^{-\frac{1}{4}\lambda_1^2} & p_1(\lambda_2)e^{-\frac{1}{4}\lambda_2^2} & \ldots & p_1(\lambda_n)e^{-\frac{1}{4}\lambda_n^2} \\ p_2(\lambda_1)e^{-\frac{1}{4}\lambda_1^2} & p_2(\lambda_2)e^{-\frac{1}{4}\lambda_2^2} & \ldots & p_2(\lambda_n)e^{-\frac{1}{4}\lambda_n^2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n-1}(\lambda_1)e^{-\frac{1}{4}\lambda_1^2} & p_{n-1}(\lambda_2)e^{-\frac{1}{4}\lambda_2^2} & \ldots & p_{n-1}(\lambda_n)e^{-\frac{1}{4}\lambda_n^2} \end{vmatrix}^2 \tag{7.23}$$

we have

$$\rho(\lambda_1, \ldots, \lambda_n) = c_n^{(2)} \left( \prod_{i=1}^{n-1} \kappa_i^2 \right) \left( \det A^{\mathrm{T}} A \right) \tag{7.24}$$

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

$$A_{ij} = \frac{1}{\kappa_{j-1}} p_{j-1}(\lambda_i) e^{-\frac{1}{4}\lambda_i^2} \tag{7.25}$$

and the $\kappa_i$ are constants satisfying

$$\int_{-\infty}^{\infty} p_i(x) p_j(x) e^{-x^2/2} \mathrm{d}x = \kappa_i^2 \delta_{ij}. \tag{7.26}$$

If $j \neq i$, this clearly holds if

$$\int_{-\infty}^{\infty} x^j p_i(x) e^{-x^2/2} \mathrm{d}x = 0 \tag{7.27}$$

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

$$p_i(x) = (-1)^i e^{x^2/2} \frac{\mathrm{d}^i}{\mathrm{d}x^i} e^{-x^2/2}. \tag{7.28}$$

These are known as the *Hermite polynomials* and are usually denoted $H_i(x)$[21]. We then also have

$$\kappa_i^2 = \int_{-\infty}^{\infty} p_i^2(x) e^{-x^2/2} \mathrm{d}x = \int_{-\infty}^{\infty} x^i p_i(x) e^{-x^2/2} \mathrm{d}x = i! \sqrt{2\pi}. \tag{7.29}$$

We shall use $p_k(x)$ and $H_k(x)$ interchangeably, since they are equal.

[21] 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.

The normalized Hermite polynomials $\kappa_k^{-1}H_k(x)$ form an orthonormal set of functions with respect to the weight $e^{-x^2/2}$:

$$\int_{-\infty}^{\infty} \kappa_j^{-1}H_j(x)\kappa_k^{-1}H_k(x)e^{-x^2/2}\mathrm{d}x = \delta_{jk}. \tag{7.30}$$

They form an orthonormal basis of the (complex) Hilbert space of functions satisfying

$$\int_{-\infty}^{\infty} |f(x)|^2 e^{-x^2/2}\mathrm{d}x < \infty \tag{7.31}$$

in which the inner product is taken to be

$$\langle f, g\rangle = \int_{-\infty}^{\infty} f(x)\overline{g(x)}e^{-x^2/2}\mathrm{d}x. \tag{7.32}$$

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

$$-\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} + \frac{1}{4}x^2\Psi = (k+\frac{1}{2})\Psi \tag{7.33}$$

which is the Schrödinger equation of the simple harmonic oscillator.

We note now that the entries of the matrix $A^{\mathrm{T}}A$ appearing in (7.24) can be written

$$\left(A^{\mathrm{T}}A\right)_{ij} = \sum_{k=1}^{n} A_{ki}A_{kj} = \sum_{k=0}^{n-1} \phi_k(\lambda_i)\phi_k(\lambda_j). \tag{7.34}$$

The function

$$K_n(x,y) = \sum_{k=0}^{n-1} \phi_k(x)\phi_k(y) = \frac{1}{\sqrt{2\pi}}\sum_{k=0}^{n-1}\frac{1}{k!}p_k(x)p_k(y)e^{-(x^2+y^2)/4} \tag{7.35}$$

is known in the *Christoffel-Darboux kernel* in the theory of orthogonal polynomials. It is the kernel of the projection onto the span of $\phi_1, \ldots, \phi_{n-1}$.

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

**Lemma 10.** *The Hermite polynomials satisfy*

$$p_{k+1}(x) = xp_k(x) - p_k'(x), \tag{7.36}$$

$$xp_k(x) = p_{k+1}(x) + kp_{k-1}(x), \tag{7.37}$$

$$p_k''(x) - xp_k'(x) = -kp_k(x), \tag{7.38}$$

*and for $x \neq y$*

$$\sum_{k=0}^{n-1}\frac{p_k(x)p_k(y)}{k!} = \frac{p_n(x)p_{n-1}(y) - p_{n-1}(x)p_n(y)}{(n-1)!(x-y)}. \tag{7.39}$$

*Proof.* (7.36) follows from differentiating (7.28). (7.37) follows from expanding $xp_k(x)$ in terms of the basis formed by all of the Hermite polynomials and using (7.27). (7.38) follows by differentiating (7.36) and then using (7.37). (7.39) follows from multiplying $K_n$ by $x-y$ and then using (7.37); 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, [14]).

Note that, applying L'Hôpital's rule, it follows from (7.39) that

$$\sum_{k=0}^{n-1}\frac{p_k^2(x)}{k!} = \frac{p_n'(x)p_{n-1}(x) - p_{n-1}'(x)p_n(x)}{(n-1)!}. \tag{7.40}$$

We have from (7.39) that for $x \neq y$

$$K_n(x,y) = \sqrt{n}\frac{\phi_n(x)\phi_{n-1}(y) - \phi_{n-1}(x)\phi_n(y)}{x-y} \tag{7.41}$$

and that

$$K_n(x,x) = \sqrt{n}\left(\phi_n'(x)\phi_{n-1}(x) - \phi_{n-1}'(x)\phi_n(x)\right). \tag{7.42}$$

In terms of the Christoffel-Darboux kernel, we have that the right-hand side of (7.24) can be written as

$$\text{const.}\det\left(K_n(\lambda_i,\lambda_j)\right). \tag{7.43}$$

It follows from the orthonormality of the Hermite polynomials that

$$\int_{-\infty}^{\infty} K_n(x,y)K_n(y,z)\mathrm{d}y = K_n(x,z). \tag{7.44}$$

This allows us to apply the following general lemma

**Lemma 11** (Gaudin's lemma)**.** *Let*

$$J_n = (J_{ij})_{1\leq i,j\leq n} = (k(x_i,x_j))_{1\leq i,j\leq n} \tag{7.45}$$

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

$$\int_{-\infty}^{\infty} k(x,y)k(y,z)\mathrm{d}y = k(x,z). \tag{7.46}$$

*Then*

$$\int_{-\infty}^{\infty} \det(J_n)\mathrm{d}x_n = (r-n+1)\det(J_{n-1}) \tag{7.47}$$

*where $r = \int_{-\infty}^{\infty} k(x,x)\mathrm{d}x$.*

*Proof.* We note first that expanding the determinant gives

$$\begin{aligned}\int_{-\infty}^{\infty} \det(J_n)\mathrm{d}x_n &= \int_{-\infty}^{\infty} \sum_{\sigma\in S_n} \mathrm{sgn}(\sigma)k(x_1,x_{\sigma(1)})\dots k(x_n,x_{\sigma(n)})\mathrm{d}x_n \\ &= \int_{-\infty}^{\infty} \sum_{m=1}^{n} \sum_{\sigma:\sigma(n)=m} \mathrm{sgn}(\sigma)k(x_1,x_{\sigma(1)})\dots k(x_n,x_m)\mathrm{d}x_n. \end{aligned} \tag{7.48}$$

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

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

$$\hat{\sigma}(i) = \begin{cases}\sigma(i), & \text{if } i\neq j \\ m, & \text{if } i=j.\end{cases} \tag{7.49}$$

The map $\{\sigma\in S_n : \sigma(n)=m\} \longrightarrow S_{n-1}$, $\sigma\mapsto\hat{\sigma}$, 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

$$\begin{aligned}&\int_{-\infty}^{\infty} \sum_{\sigma:\sigma(n)=m} \mathrm{sgn}(\sigma)k(x_1,x_{\sigma(1)})\dots k(x_n,x_m)\mathrm{d}x_n \\ &\quad= \int_{-\infty}^{\infty} \sum_{\sigma:\sigma(n)=m} \mathrm{sgn}(\sigma)k(x_1,x_{\sigma(1)})\dots k(x_{n-1},x_{\sigma(n-1)})k(x_j,x_n)k(x_n,x_m)\mathrm{d}x_n \\ &\quad\quad= \sum_{\hat{\sigma}\in S_{n-1}} -\mathrm{sgn}(\hat{\sigma})k(x_1,x_{\hat{\sigma}(1)})\dots k(x_{n-1},x_{\hat{\sigma}(n-1)}) \\ &\quad\quad\quad= -\det J_{n-1}. \end{aligned} \tag{7.50}$$

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 Theorem 7. Note that in the notation introduced above, we have for the GUE that

$$\mathbb{E}[f(\lambda_1, \lambda_2, \ldots, \lambda_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\lambda_1, \ldots, \lambda_n) \rho(\lambda_1, \ldots, \lambda_n) \mathrm{d}\lambda_1 \ldots \mathrm{d}\lambda_n \tag{7.51}$$

with

$$\rho(\lambda_1, \ldots, \lambda_n) = \tilde{c}_n \det\left(K_n(\lambda_i, \lambda_j)_{1 \le i,j \le n}\right) \tag{7.52}$$

where $\tilde{c}_n$ denotes the constant.

In many situations we shall see that we will want to average functions $f$ that depend only on $k < n$ eigenvalues. In this case we need to evaluate the integral of $\rho$ over the remaining $n - k$ variables. This can be achieved using (Gaudin's) lemma 11.

**Lemma 12.** *For the GUE we have*

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \rho(\lambda_1, \ldots, \lambda_n) \mathrm{d}\lambda_{k+1} \ldots \mathrm{d}\lambda_n = \frac{(n-k)!}{n!} \det\left(K_n(\lambda_i, \lambda_j)_{1 \le i,j \le k}\right). \tag{7.53}$$

*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

$$\int_{-\infty}^{\infty} \rho(\lambda_1, \ldots, \lambda_n) \mathrm{d}\lambda_n = \tilde{c}_n \det\left(K_n(\lambda_i, \lambda_j)_{1 \le i,j \le n-1}\right). \tag{7.54}$$

Applying Gaudin's lemma a second time now gives

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \rho(\lambda_1, \ldots, \lambda_n) \mathrm{d}\lambda_{n-1} \mathrm{d}\lambda_n = 2\tilde{c}_n \det\left(K_n(\lambda_i, \lambda_j)_{1 \le i,j \le n-2}\right). \tag{7.55}$$

Repeated applications of Gaudin's lemma therefore give

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \rho(\lambda_1, \ldots, \lambda_n) \mathrm{d}\lambda_{k+1} \ldots \mathrm{d}\lambda_n = (n-k)! \tilde{c}_n \det\left(K_n(\lambda_i, \lambda_j)_{1 \le i,j \le k}\right). \tag{7.56}$$

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

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \rho(\lambda_1, \ldots, \lambda_n) \mathrm{d}\lambda_1 \ldots \mathrm{d}\lambda_n = n! \tilde{c}_n \tag{7.57}$$

and so we see that $\tilde{c}_n = 1/n!$.

□

It is worth pointing out that this is a rather remarkable result: 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*.

## 7.4 Correlation functions

*Definition.* The $k$-point correlation function is defined to be

$$R_k(x_1, \ldots, x_k) := \frac{n!}{(n-k)!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \rho(x_1, \ldots, x_n) \mathrm{d}x_{k+1} \ldots \mathrm{d}x_n \tag{7.58}$$

$$= \det\left(K_n(x_i, x_j)_{1 \le i,j \le k}\right). \tag{7.59}$$

These correlation functions have the following interpretation. For any Borel set $\mathcal{B}$,

$$\int_{\mathcal{B}} R_1(x)\mathrm{d}x = \mathbb{E}[\#\{\text{eigenvalues in } \mathcal{B}\}] \tag{7.60}$$

and

$$\int_{\mathcal{B}}\int_{\mathcal{B}} R_2(x,y)\mathrm{d}x\mathrm{d}y = \mathbb{E}[\#\{\text{ordered pairs of eigenvalues in } \mathcal{B}\}] \tag{7.61}$$

etc.

Clearly if we rescale the eigenvalues by $1/\sqrt{n}$ and normalise the 1-point correlation function by $1/n$, we expect to recover the semicircle law from the 1-point correlation function; that is, we expect

$$\lim_{n\to\infty}\frac{1}{\sqrt{n}}R_1(\sqrt{n}x) = \lim_{n\to\infty}\frac{1}{\sqrt{n}}K_n(\sqrt{n}x,\sqrt{n}x) = \frac{1}{2\pi}\sqrt{4-x^2} \tag{7.62}$$

for $x\in[-2,2]$, 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

$$\Xi_{n+p}(x) = n^{\frac{1}{4}}\phi_{n+p}(x\sqrt{n}) \tag{7.63}$$

for $p=-2,-1,0$. It follows from the properties of the Hermite polynomials listed above that

$$\frac{1}{\sqrt{n}}K_n(\sqrt{n}x,\sqrt{n}x) = \Xi_{n-1}^2(x) - \sqrt{\frac{n-1}{n}}\Xi_{n-2}(x)\Xi_n(x). \tag{7.64}$$

The asymptotic formulae we need[22] are as follows. First, setting $x=2\cos\phi$ with $0<\phi<\pi$, when $n\to\infty$

$$\Xi_{n+p}(x) \sim \frac{1}{\sqrt{\pi\sin\phi}}\cos\left[n\left(\phi-\frac{1}{2}\sin(2\phi)\right)+\left(p+\frac{1}{2}\right)\phi-\frac{\pi}{4}\right] \tag{7.65}$$

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

$$\Xi_{n+p}(x) \sim \frac{e^{(p+1/2)\phi}}{\sqrt{2\pi\sinh\phi}}e^{-\frac{n}{2}(e^{2\phi}+1-2\phi)} \tag{7.66}$$

uniformly for $\phi$ in a compact subset of $(0,\infty)$[23]. Substituting (7.65) into (7.64) gives the semicircle law in $|x<2$, and substituting in (7.65) gives the result when $|x|>2$.

Note that (7.42) gives that

$$\frac{1}{\sqrt{n}}R_1(\sqrt{n}x) = \phi_n'(\sqrt{n}x)\phi_{n-1}(\sqrt{n}x) - \phi_{n-1}'(\sqrt{n}x)\phi_n(\sqrt{n}x) \tag{7.67}$$

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 7 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 (7.67). 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 $|x|\le 1$.

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

[22] These go under the general name *Plancherel-Rotach asymptotic formule* and can be established via a WKB analysis of (7.33).

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

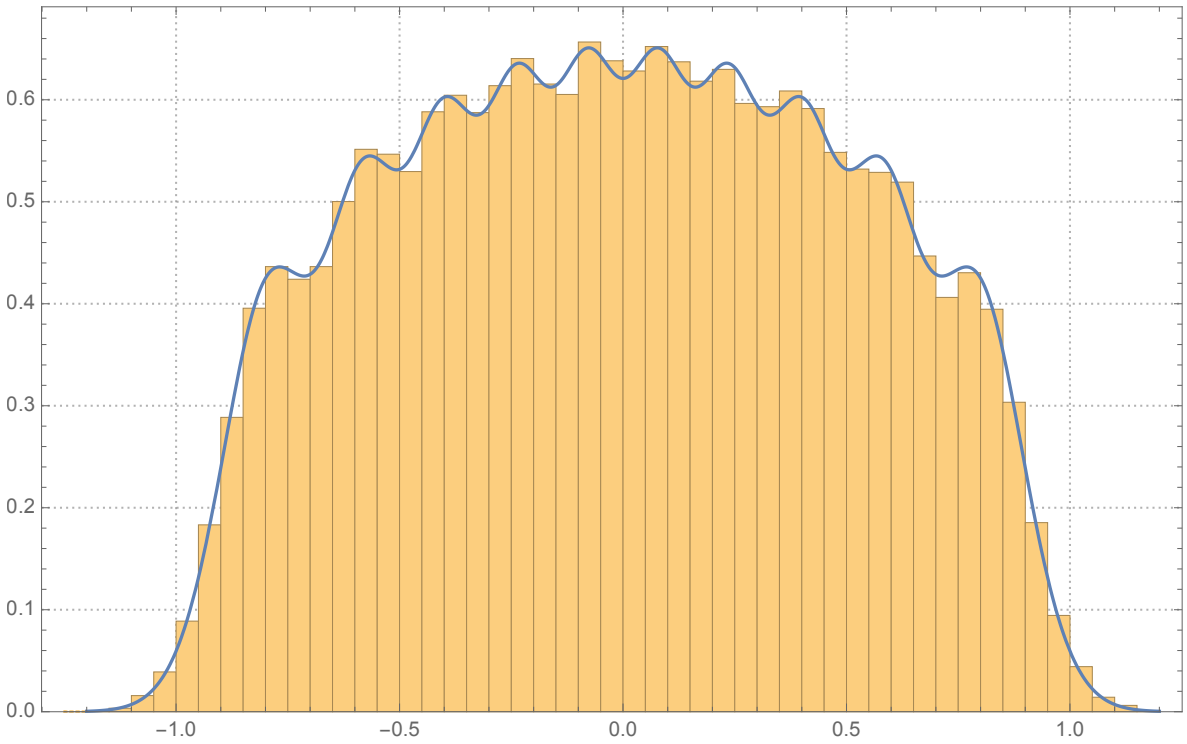


Figure 7: The spectral density of 10,000 GUE matrices of dimension 10 compared to (7.67).

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. Setting $y_i = \sqrt{n}\lambda_i$, we need to consider the $n \to \infty$ asymptotics of

$$\det\left[n^{-1/2}K_n(y_i/\sqrt{n}, y_j/\sqrt{n})\right]_{i,j=1}^{r} := \det\left[\hat{K}_n(y_i, y_j)\right]_{i,j=1}^{r} \tag{7.68}$$

where, when $y_i \neq y_j$,

$$\hat{K}_n(y_i, y_j) = \frac{\phi_n(y_i/\sqrt{n})\phi_{n-1}(y_j/\sqrt{n}) - \phi_{n-1}(y_i/\sqrt{n})\phi_n(y_j/\sqrt{n})}{y_i - y_j}, \tag{7.69}$$

and

$$\hat{K}_n(y, y) = \phi_n'(y/\sqrt{n})\phi_{n-1}(y/\sqrt{n}) - \phi_{n-1}'(y/\sqrt{n})\phi_n(y/\sqrt{n}). \tag{7.70}$$

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

**Theorem 13.** *For any fixed $x, y$*

$$\lim_{n\to\infty} \hat{K}_n(x, y) = \frac{\sin(\pi(x-y))}{\pi(x-y)}. \tag{7.71}$$

This is an immediate consequence of the following lemma.

**Lemma 14.** *For any fixed $x$ and any fixed $l$*

$$\lim_{n\to\infty} \left| n^{1/4}\phi_{n-l}(x/\sqrt{n}) - \pi^{-1/2}\cos\left(x - (n-l)\pi/2\right)\right| = 0. \tag{7.72}$$

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

$$(-1)^m e^{x^2/2} \frac{\mathrm{d}^m}{\mathrm{d}x^m} e^{-x^2/2} = (-1)^m e^{x^2/2} \frac{\mathrm{d}^m}{\mathrm{d}x^m} \int_{-\infty}^{\infty} e^{-ixz - z^2/2} \frac{\mathrm{d}z}{\sqrt{2\pi}}$$
$$= \frac{1}{\sqrt{2\pi}} e^{x^2/2} \int_{-\infty}^{\infty} (iz)^m e^{-ixz - z^2/2} \mathrm{d}z \tag{7.73}$$

and so

$$n^{1/4} \phi_m(x/\sqrt{n}) = (2\pi)^{-3/4} (m!)^{-1/2} e^{x^2/4n} n^{1/4} \int_{-\infty}^{\infty} (iz)^m e^{-ixz/\sqrt{n} - z^2/2} \mathrm{d}z. \tag{7.74}$$

Using Stirling's formula to evaluate the prefactor and Laplace's method[24] to evaluate the integral when $n \to \infty$ proves the lemma. (See, for example, Chapter VIII in [14].)

□

Substituting the asymptotic estimate in the lemma into the formula for $\hat{K}_n$, and using (7.36) and (7.37) to evaluate the asymptotics of the derivative of $\phi$, then leads to the formula in the theorem.

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

$$\frac{\sin(\pi(x-y))}{\pi(x-y)}. \tag{7.75}$$

This is known as the sine kernel. For example, the pair correlation of the scaled eigenvalues is simply

$$1 - \left( \frac{\sin(\pi(x-y))}{\pi(x-y)} \right)^2. \tag{7.76}$$

So, for example, for $f(x, y)$ such that the sum and integral converge,

$$\lim_{n \to \infty} \sum_{ij} f(\sqrt{n}\lambda_i, \sqrt{n}\lambda_j) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \left( 1 - \left( \frac{\sin(\pi(x-y))}{\pi(x-y)} \right)^2 \right) \mathrm{d}x \mathrm{d}y. \tag{7.77}$$

Note that the quadratic repulsion manifest in Theorem 7 is clear here in (7.76), which, if we set $x - y = w$, 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 (7.76) 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.

## 7.5 Scaling at the edge of the spectrum

Having seen how the sine kernel describes correlations in the *bulk* of the spectrum, we now consider how the eigenvalues are distributed near to the edges. By the semicircle law, we expect to have, when $\epsilon$ is small, that

$$\#\{\lambda_i : \lambda_i/\sqrt{n} \geq 2 - \epsilon\} \approx \frac{n}{2\pi} \int_{2-\epsilon}^{2} \sqrt{4 - x^2} \mathrm{d}x \approx \frac{2}{3\pi} n \epsilon^{3/2}. \tag{7.78}$$

We might therefore expect to get a nice limit if we set $\epsilon \propto n^{-3/2}$ or 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 $\lambda_i = 2\sqrt{n} + \alpha_i n^{-1/6}$ and

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

looking at correlations between the values taken by $\alpha_1, \ldots, \alpha_n$. This involves studying the asymptotics of

$$\det\left[n^{-1/6}K_n(2\sqrt{n}+\alpha_i n^{-1/6}, 2\sqrt{n}+\alpha_j n^{-1/6})\right]_{ij=1}^{r}. \tag{7.79}$$

Defining

$$\tilde{K}_n(x,y) = n^{-1/6}K_n(2\sqrt{n}+xn^{-1/6}, 2\sqrt{n}+yn^{-1/6}) \tag{7.80}$$

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

**Theorem 15.** *For any fixed $x, y$*

$$\lim_{n\to\infty}\tilde{K}_n(x,y) = \frac{\mathrm{Ai}(x)\mathrm{Ai}'(y)-\mathrm{Ai}'(x)\mathrm{Ai}(y)}{(x-y)} \tag{7.81}$$

*where*

$$\mathrm{Ai}(x) = \frac{1}{\pi}\int_0^\infty \cos(t^3/3+xt)\mathrm{d}t \tag{7.82}$$

*is the Airy function.*

This is an immediate consequence of the following lemma

**Lemma 16.** *For any fixed $x$*

$$\lim_{n\to\infty}\left|n^{1/12}\phi_n(2\sqrt{n}+x/n^{1/6}) - \mathrm{Ai}(x)\right| = 0. \tag{7.83}$$

We only sketch the proof of this lemma. This starts with (7.74). 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 8.

Substituting the asymptotic estimate in the lemma into the formula for $\tilde{K}_n$ in terms of $\phi$, and using (7.36) and (7.37) to evaluate the asymptotics of the derivative of $\phi$, 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.

## 7.6 Counting statistics

Having understood the structure of spectral correlation functions in the bulk of the spectrum and at the edge, we now turn our attention to counting statistics. Let us start by looking at an interval $I$ in the bulk of the spectrum. Denote by $\nu(M,I)$ the number of eigenvalues of a GUE matrix $M$ lying in $I$. We shall focus on

$$D_m(I) = \mathbb{P}(\nu(M,I)=m). \tag{7.84}$$

Let $\mathbb{I}_I(x)$ denote the characteristic function on $I$, so

$$\mathbb{I}_I(x) = \begin{cases} 1 \text{ if } x\in I \\ 0 \text{ if } x\notin I. \end{cases} \tag{7.85}$$

Then, for example,

$$D_0(I) = \mathbb{E}\prod_{i=1}^{n}(1-\mathbb{I}_I(\lambda_i)). \tag{7.86}$$

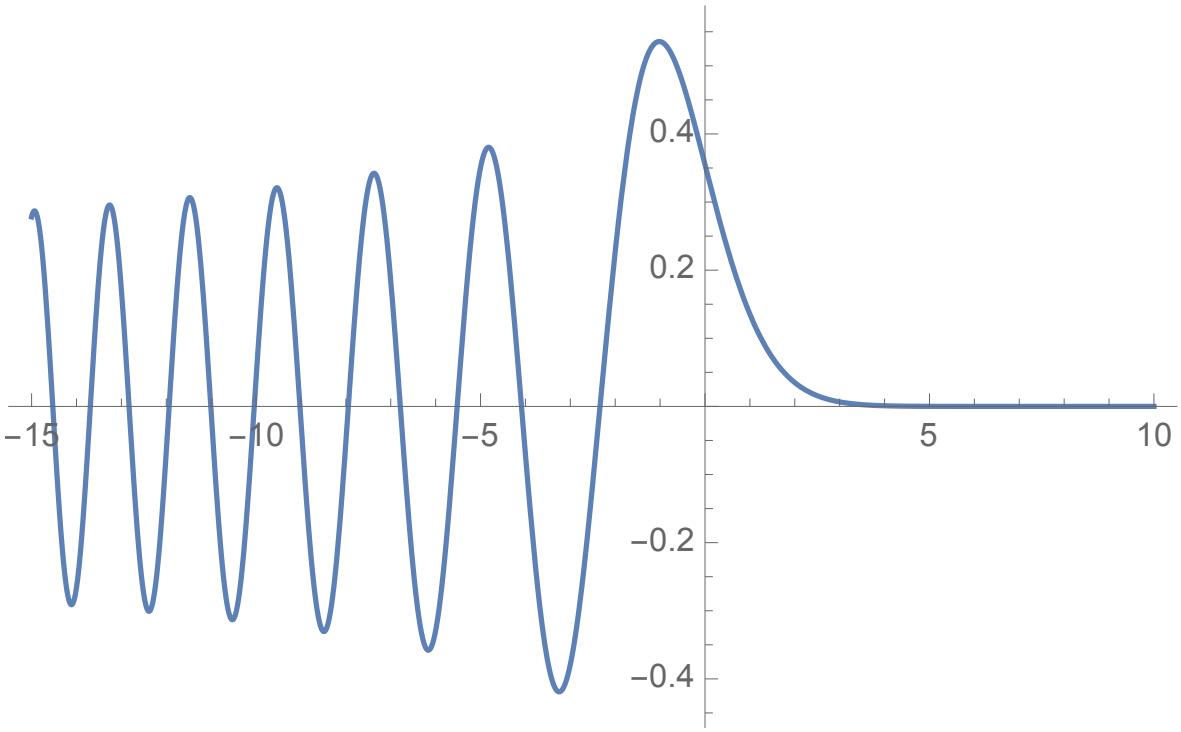


Figure 8: The Airy function $\mathrm{Ai}(x)$.

We define

$$G_I(t) = \mathbb{E}\prod_{i=1}^{n}(1 - t\mathbb{I}_I(\lambda_i)), \tag{7.87}$$

so that $D_0(I) = G_I(1)$. It is easy to see that $-G_I'(1) = D_1(I)$, and that more generally

$$D_m(I) = \frac{(-1)^m}{m!}G_I^{(m)}(1). \tag{7.88}$$

$G_I(t)$ is therefore the generating function for $D_m(I)$.

To calculate $G_I(t)$, we expand the product over the eigenvalues to get[25]

$$\begin{aligned} G_I(t) &= 1 - t\mathbb{E}\sum_i \mathbb{I}_I(\lambda_i) + t^2\mathbb{E}\sum_{i\le j}\mathbb{I}_I(\lambda_i)\mathbb{I}_I(\lambda_j) - \dots \\ &= 1 - t\int_I R_1(x_1)\mathrm{d}x_1 + \frac{t^2}{2}\int_I\int_I R_2(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2 - \frac{t^3}{3!}\int_I\int_I\int_I R_3(x_1,x_2,x_3)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3 + \dots \\ &= 1 + \sum_{k=1}^{\infty}\frac{(-t)^k}{k!}\int_I\cdots\int_I \det\left(K_n(x_i,x_j)\right)_{1\le i,j\le k}\mathrm{d}x_1\dots\mathrm{d}x_k. \end{aligned} \tag{7.89}$$

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

$$\det\left(I - tK_n(x,y)\right) := 1 + \sum_{k=1}^{\infty}\frac{(-t)^k}{k!}\int_I\cdots\int_I \det\left(K_n(x_i,x_j)\right)_{1\le i,j\le k}\mathrm{d}x_1\dots\mathrm{d}x_k. \tag{7.90}$$

[25]Note the the expansion is valid because the rank of $K_n$ is at most $n$.

In the theory of Fredholm determinants, 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 $\mathrm{Tr}(K) = \int K(x,x)\mathrm{d}x$ exists, then $K$ is called *trace class* and one can calculate the Fredholm determinant in a number of (equivalent) ways:

$$\det\left(I - tK_n(x,y)\right) = \exp\left(-\sum_{m=1}^{\infty}\frac{t^m}{m}\mathrm{Tr}K^m\right) \tag{7.91}$$

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

$$\det\left(I - tK_n(x,y)\right) = \prod(1 - t\beta_j). \tag{7.92}$$

The probability $P_n(s)\mathrm{d}s$ that the spacing between adjacent (i.e. nearest neighbouring) eigenvalues lies between $s$ and $s+\mathrm{d}s$ can be calculated in terms of the Fredholm determinant as follows. Consider the interval $I$ to have length $s$. The probability that this interval contains no eigenvalues is $F(s) = D_0(I)$. 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. Now extending the interval at the other end, one sees that

$$P_n(s) = \frac{\mathrm{d}^2F}{\mathrm{d}s^2} = \left.\frac{\mathrm{d}^2G_{I(s)}(t)}{\mathrm{d}s^2}\right|_{t=1}. \tag{7.93}$$

Hence one can calculate $P_n(s)$ from the Fredholm determinant. When the eigenvalues are scaled by $\sqrt{n}$, then when $n \to \infty$ the limiting spacing distribution is the second derivative of the Fredholm determinant of the sine kernel[26]. This distribution is often called the *Gaudin-Mehta* distribution

We illustrate this by showing in Figure 9 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.

The analysis of $D_m(I)$ extends immediately to the edge of the spectrum, with, in this case, the kernel in (7.89) and (7.90) 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

$$F_n^{(\max)}(t) := \mathbb{P}(\lambda_{\max} \le t). \tag{7.94}$$

Then from the scaling at the edge we have that

$$F^{(\max)}(s) = \lim_{n\to\infty} F_n^{(\max)}(2\sqrt{n} + s/n^{1/6}) \tag{7.95}$$

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

$$F^{(\max)}(s) = \det(I - K_{\mathrm{Airy}}) \tag{7.96}$$

where

$$K_{\mathrm{Airy}} = \frac{\mathrm{Ai}(x)\mathrm{Ai}'(y) - \mathrm{Ai}'(x)\mathrm{Ai}(y)}{(x-y)} \tag{7.97}$$

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

In this case too $F^{(\max)}(s)$ can be expressed in terms of a solution of a second order nonlinear o.d.e. which is another of the Painlevé equations. Specifically,

$$F^{(\max)}(s) = \exp\left(-\int_s^{\infty}(x-s)q(x)^2\mathrm{d}x\right) \tag{7.98}$$

[26]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 $\sigma$ 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.

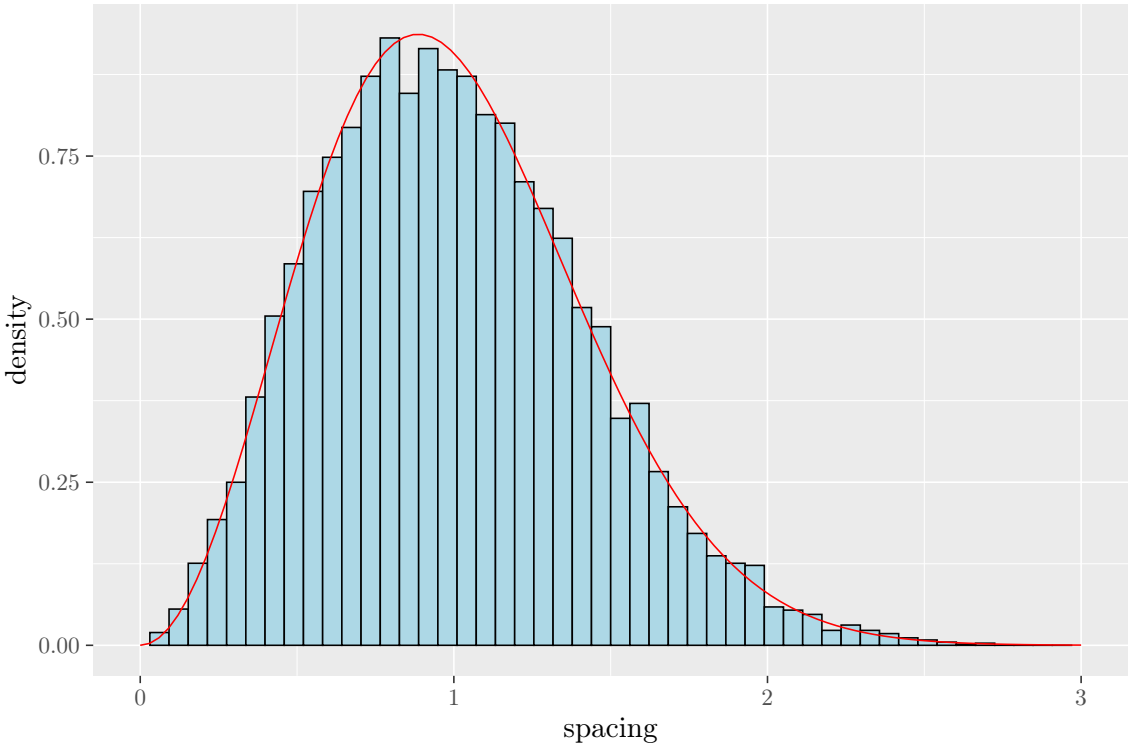


Figure 9: The spacing distribution between neighbouring eigenvalues of 10,000 GUE matrices of dimension $n = 10$ compared to the the Fredholm determinant of the sine kernel. (Plot kindly provided by Johannes Forkel.)

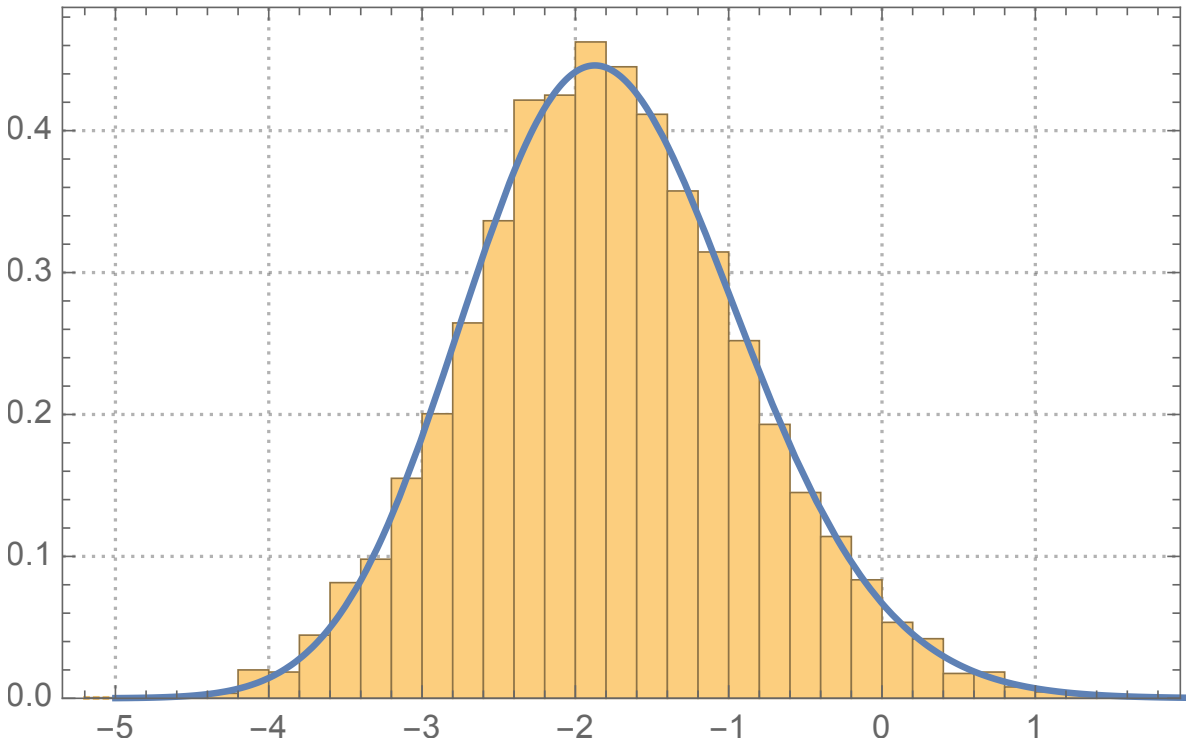


Figure 10: $(\lambda_{\max} - 2\sqrt{n})n^{1/6}$ for 10,000 GUE matrices of size $n = 1000$.

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 10 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.

## 7.7 Other Ensembles

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 $e^{ip\theta}$ – all that is involved is the manipulation of Fourier series and the asymptotic analysis is elementary. So, for the Vandermonde factor in the Weyl integration formula

$$\begin{aligned}\prod_{j<k} |e^{i\theta_j} - e^{i\theta_k}|^2 &= \prod_{j<k} \left(e^{i\theta_j} - e^{i\theta_k}\right) \prod_{j<k} \left(e^{-i\theta_j} - e^{-i\theta_k}\right) \\ &= \det[e^{ip\theta_j}] \det[e^{-ip\theta_j}]. \end{aligned} \tag{7.99}$$

In these determinants $j = 1, 2, \ldots, n$ and $p = 0, 1, \ldots, n-1$, or, equivalently, $p = -\frac{1}{2}(n-1), -\frac{1}{2}(n-1)+1, \ldots, \frac{1}{2}(n-1)$. Hence

$$\prod_{j<k} |e^{i\theta_j} - e^{i\theta_k}|^2 = (2\pi)^n \det\left[\tilde{S}_n(\theta_j - \theta_k)\right]_{j,k=1,\ldots,n} \tag{7.100}$$

with

$$\tilde{S}_n(\theta) = \frac{1}{2\pi} \sum_{p=-\frac{1}{2}(n-1)}^{\frac{1}{2}(n-1)} e^{ip\theta} = \frac{1}{2\pi} \frac{\sin(n\theta/2)}{\sin(\theta/2)}. \tag{7.101}$$

Using its representation as a Fourier series

$$\int_0^{2\pi} \tilde{S}_n(\theta_j - \theta_k)\tilde{S}_n(\theta_k - \theta_l)\mathrm{d}\theta_k = \tilde{S}_n(\theta_j - \theta_l). \tag{7.102}$$

We can therefore apply Gaudin's lemma, but now with integrals over $[0, 2\pi)$. Trivially,

$$\int_0^{2\pi} \tilde{S}_n(0)\mathrm{d}\theta = n \tag{7.103}$$

and so Gaudin's lemma yields, exactly as for the GUE, that the $k$-point correlation function is

$$R_k(\theta_1, \ldots, \theta_k) = \det\left[\tilde{S}_n(\theta_i - \theta_j)\right]_{i,j=1,\ldots,k}. \tag{7.104}$$

This is again a $k \times k$ determinant, but now with a simpler kernel. It is easy to see that

$$R_1(\theta) = \tilde{S}_n(0) = \frac{n}{2\pi} \tag{7.105}$$

which is the analogue of the semicircle law in this setting – the eigenvalues have a constant average density around the unit circle – and that

$$R_2(\theta_1, \theta_2) = \left(\frac{n}{2\pi}\right)^2 - \left(\tilde{S}_n(\theta_1 - \theta_2)\right)^2. \tag{7.106}$$

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

$$\lim_{n\to\infty} \frac{1}{n}\tilde{S}_n(2\pi(\phi_j - \phi_k)/n) = \lim_{n\to\infty} \frac{\sin(\pi(\phi_j - \phi_k)}{n\sin(\pi(\phi_j - \phi_k)/n)} = \frac{\sin(\pi(\phi_j - \phi_k)}{\pi(\phi_j - \phi_k))} \tag{7.107}$$

i.e. 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 11 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 $2 \times 2$ determinant involving the sine kernel.

The method of orthogonal polynomials works as well for the GOE, but is a little more complicated in that case. It works for complex Wishart matrices, when the orthogonal polynomials are the generalized Laguerre polynomials. In the case of real Wishart matrices, as for the GOE, the method also applies, but it is a little more complicated.

## 7.8 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

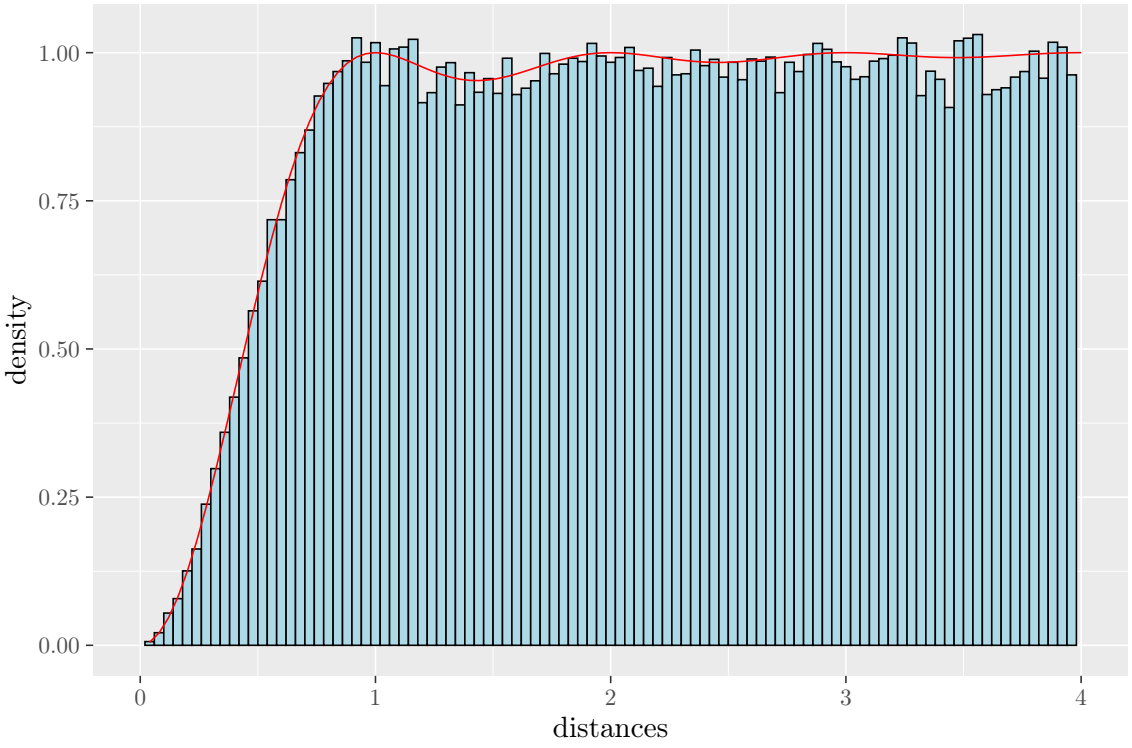


Figure 11: The two-point correlation function of the eigenvalues of 200 CUE matrices of dimension 200 compared to the $2 \times 2$ determinant involving the sine kernel. (Plot kindly provided by Johannes Forkel.)

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.

## 7.9 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.

For more on these applications, see the respective chapters in The Oxford Handbook of Random Matrix Theory [1].

# 8 Dyson Brownian motion

We can express the GOE and GUE measures in terms of the eigenvalues, as embodied in Theorem 7, in the form

$$c_n^{(\beta)} e^{-\beta W(\lambda_1,\ldots,\lambda_n)} \tag{8.1}$$

where

$$W(\lambda_1,\ldots,\lambda_n) = \frac{1}{4}\sum_{j=1}^{n}\lambda_j^2 - \sum_{i<j}\ln|\lambda_i-\lambda_j|. \tag{8.2}$$

One can think of this as a Boltzmann weight where $W(\lambda_1,\ldots,\lambda_n)$ represents the potential energy associated with particles located at positions $\lambda_1,\ldots,\lambda_n$ and with $\beta$ 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 explore the dynamics of the eigenvalues if we allow the matrix elements of a real-symmetric or complex-Hermitian matrix to vary stochastically with time $t$. Specifically, let us consider the matrix $M$ to be a function of $t$ satisfying, for $t \geq 0$, the matrix-valued stochastic differential equation

$$\mathrm{d}M(t) = \frac{1}{\sqrt{n}}\mathrm{d}B(t) - \frac{1}{2}M(t)\mathrm{d}t \tag{8.3}$$

with initial data $M_0$. Here in the real-symmetric case, when $\beta = 1$, $B(t)$ is an $n \times n$ symmetric matrix such that $B_{ij}(t)$, for $1 \leq i < j \leq n$, and $B_{ii}(t)/\sqrt{2}$, for $1 \leq i \leq n$, are independent standard Brownian motions; and in the complex-Hermitian case, when $\beta = 2$, $B(t)$ is an $n \times n$ complex Hermitian matrix such that $\sqrt{2}\mathrm{Re}B_{ij}(t)$ and $\sqrt{2}\mathrm{Im}B_{ij}(t)$, for $1 \leq i < j \leq n$, and $B_{ii}(t)/\sqrt{2}$, for $1 \leq i \leq n$, are independent standard Brownian motions. This defines what is known as a *matrix-valued Ornstein-Uhlenbeck process*. In terms of the matrix elements of $M$, this differential equation reads

$$\mathrm{d}M_{ij}(t) = \frac{1}{\sqrt{n}}\mathrm{d}B_{ij}(t) - \frac{1}{2}M_{ij}(t)\mathrm{d}t \tag{8.4}$$

where $B_{ij}(t)$ has variance $t$ in the complex-Hermitian case, while in the real-symmetric case $B_{ij}(t)$ has variance $t$ when $i \neq j$ and variance $2t$ when $i = j$.

The question we seek to address is: if the matrix $M$ evolves in time according to (8.3), what is the equation satisfied by 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 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. This system of SDEs is the following one.

$$\mathrm{d}\lambda_i = \frac{\sqrt{2}}{\sqrt{\beta n}}\mathrm{d}\tilde{B}_i + \left(-\frac{\lambda_i}{2} + \frac{1}{n}\sum_{j\neq i}\frac{1}{\lambda_i-\lambda_j}\right)\mathrm{d}t \tag{8.5}$$

where $\tilde{B}_i$ is a set of real-valued independent standard Brownian motions. The solution of this equation is known as *Dyson Brownian motion* with parameter $\beta$, after Freeman Dyson, who introduced it in a seminal paper in 1962.

As noted above, it can be proved (see, for example, section 4.3 in [2]) that there exists a unique (strong) solution of this equation in the space of continuous functions. Moreover, if the initial conditions

are such that all of the eigenavalues are simple at $t = 0$, then this remains true for all $t > 0$ – i.e. none of the eigenvalue processes intersect.

Dyson's theorem can be stated as follows.

**Theorem 17.** *Let $M(t)$ solve (8.3) in a strong sense. Then its eigenvalue process satisfies (8.5).*

*Proof..* We denote the eigenvalues of an $n \times n$ real-symmetric or complex-Hermitian matrix $M$ by $(\lambda_m)_{1\leq m\leq n}$ and the corresponding eigenvectors by $(\mathbf{v}_m)_{1\leq m\leq n}$, so that

$$M\mathbf{v}_m = \lambda_m \mathbf{v}_m. \tag{8.6}$$

Note that the eigenvectors are orthonormal, i.e.

$$\mathbf{v}_m^\dagger \mathbf{v}_p = \delta_{m,p}. \tag{8.7}$$

Differentiating these equations with respect to $M_{ij}$ yields

$$\frac{\partial M}{\partial M_{ij}}\mathbf{v}_m + M\frac{\partial \mathbf{v}_m}{\partial M_{ij}} = \frac{\partial \lambda_m}{\partial M_{ij}}\mathbf{v}_m + \lambda_m \frac{\partial \mathbf{v}_m}{\partial M_{ij}}, \tag{8.8}$$

$$\frac{\partial \mathbf{v}_m^\dagger}{\partial M_{ij}}\mathbf{v}_p + \mathbf{v}_m^\dagger \frac{\partial \mathbf{v}_p}{\partial M_{ij}} = 0 \tag{8.9}$$

for $p \neq m$, and

$$\frac{\partial \mathbf{v}_m^\dagger}{\partial M_{ij}}\mathbf{v}_m = 0. \tag{8.10}$$

Hence

$$\frac{\partial \lambda_m}{\partial M_{ij}} = \mathbf{v}_m^\dagger \frac{\partial M}{\partial M_{ij}}\mathbf{v}_m \tag{8.11}$$

and for $p \neq m$

$$\mathbf{v}_p^\dagger \frac{\partial M}{\partial M_{ij}}\mathbf{v}_m + \mathbf{v}_p^\dagger M\frac{\partial \mathbf{v}_m}{\partial M_{ij}} = \lambda_m \mathbf{v}_p^\dagger \frac{\partial \mathbf{v}_m}{\partial M_{ij}}, \tag{8.12}$$

from which we have that

$$\mathbf{v}_p^\dagger \frac{\partial M}{\partial M_{ij}}\mathbf{v}_m + \lambda_p \mathbf{v}_p^\dagger \frac{\partial \mathbf{v}_m}{\partial M_{ij}} = \lambda_m \mathbf{v}_p^\dagger \frac{\partial \mathbf{v}_m}{\partial M_{ij}}. \tag{8.13}$$

Hence

$$\frac{\partial \mathbf{v}_m}{\partial M_{ij}} = \sum_{p\neq m}\left(\mathbf{v}_p^\dagger \frac{\partial \mathbf{v}_m}{\partial M_{ij}}\right)\mathbf{v}_p = \sum_{p\neq m}\frac{1}{\lambda_m - \lambda_p}\left(\mathbf{v}_p^\dagger \frac{\partial M}{\partial M_{ij}}\mathbf{v}_m\right)\mathbf{v}_p. \tag{8.14}$$

In order to keep the notation as simple as possible, we now focus on the real-symmetric case, when $\beta = 1$. The complex-Hermitian calculation follows exactly the same lines. In the real-symmetric case, denoting the $i$th component of $\mathbf{v}_m$ by $v_m^{(i)}$, (8.11) becomes

$$\frac{\partial \lambda_m}{\partial M_{ij}} = v_m^{(i)} v_m^{(j)}(2 - \delta_{ij}), \tag{8.15}$$

and (8.14) becomes

$$\frac{\partial v_m^{(k)}}{\partial M_{ij}} = \sum_{p\neq m}\frac{v_p^{(i)} v_m^{(j)} + v_p^{(j)} v_m^{(i)}(1 - \delta_{ij})}{\lambda_m - \lambda_p} v_p^{(k)}. \tag{8.16}$$

Differentiating the first of these equations and then using the second gives

$$\begin{aligned}\frac{\partial^2 \lambda_m}{\partial M_{lj}\partial M_{ik}} &= (2 - \delta_{ik})\left[\frac{\partial v_m^{(i)}}{\partial M_{lj}} v_m^{(k)} + v_m^{(i)}\frac{\partial v_m^{(k)}}{\partial M_{lj}}\right] \\ &= (2 - \delta_{ik})\sum_{p\neq m}\frac{1}{\lambda_m - \lambda_p}\left[(v_p^{(j)} v_m^{(l)} + v_p^{(l)} v_m^{(j)}(1 - \delta_{jl}))(v_p^{(i)} v_m^{(k)} + v_p^{(k)} v_m^{(i)})\right].\end{aligned} \tag{8.17}$$

We are now in a position to use the assumption that the matrix elements evolve as an Ornstein-Uhlenbeck process (8.4), which means that in a time interval of length $\mathrm{d}t$ their square scales like $\mathrm{d}t$. We therefore expand $\mathrm{d}\lambda_m$ to quadratic order in the $\mathrm{d}M_{ij}$[27]. We then get[28]

$$\mathrm{d}\lambda_m = \sum_{i\leq k} \frac{\partial \lambda_m}{\partial M_{ik}} \mathrm{d}M_{ik} + \frac{1}{2} \sum_{i\leq k} \sum_{j\leq l} \frac{\partial^2 \lambda_m}{\partial M_{lj} \partial M_{ik}} (\mathrm{d}M_{ik})(\mathrm{d}M_{lj}). \tag{8.18}$$

Now, using (8.4), (8.15), (8.17), and the fact that

$$(\mathrm{d}M_{ik})(\mathrm{d}M_{lj}) = \frac{1}{n} \delta_{il}\delta_{kj}(1+\delta_{ik})\mathrm{d}t \tag{8.19}$$

we have, using the fact that $B$ is symmetric, that

$$\begin{aligned} \mathrm{d}\lambda_m =& \sum_{i,k} v_m^{(i)} v_m^{(k)} \left( \frac{1}{\sqrt{n}} \mathrm{d}B_{ik} - \frac{M_{ik}}{2} \mathrm{d}t \right) \\ &+ \frac{1}{2n} \sum_{i,k} \sum_{p\neq m} \frac{1}{\lambda_m - \lambda_p} \left[ |v_p^{(i)}|^2 |v_m^{(k)}|^2 + |v_m^{(i)}|^2 |v_p^{(k)}|^2 \right] \mathrm{d}t. \end{aligned} \tag{8.20}$$

Hence, using the fact that $\mathbf{v}_m$ is the eigenvector of $M$ with eigenvalue $\lambda_m$, and so

$$\sum_{i,k} v_m^{(i)} v_m^{(k)} M_{ik} = \sum_i v_m^{(i)} (M\mathbf{v}_m)^{(i)} = \lambda_m \sum_i |v_m^{(i)}|^2 = \lambda_m, \tag{8.21}$$

we have finally that

$$\mathrm{d}\lambda_m = \frac{1}{\sqrt{n}} \sum_{i,k} v_m^{(i)} v_m^{(k)} \mathrm{d}B_{ik} - \frac{1}{2}\lambda_m \mathrm{d}t + \frac{1}{n} \sum_{p\neq m} \frac{1}{\lambda_m - \lambda_p} \mathrm{d}t. \tag{8.22}$$

The key point now is that, looking at the first term in (8.22), we can define

$$\tilde{B}_m := \sum_{i,k} v_m^{(i)} v_m^{(k)} B_{ik}. \tag{8.23}$$

This is itself a real Gaussian process satisfying $\mathbb{E}\mathrm{d}\tilde{B}_m = 0$ and

$$\begin{aligned} \mathbb{E}\mathrm{d}\tilde{B}_m \mathrm{d}\tilde{B}_p &= \mathbb{E} \sum_{i,k} \sum_{i',k'} v_m^{(i)} v_m^{(k)} \mathrm{d}B_{ik} v_p^{(i')} v_p^{(k')} \mathrm{d}B_{i'k'} \\ &= 2 \sum_{i,k} v_m^{(i)} v_p^{(i)} v_m^{(k)} v_p^{(k)} \mathrm{d}t \\ &= 2\delta_{mp} \mathrm{d}t \end{aligned} \tag{8.24}$$

We see therefore that $\tilde{B}_m = \sqrt{2} B_m$ in distributional sense, where $(B_m)_{m=1}^n$ is the standard Brownian motion in $\mathbb{R}^n$. This completes the proof in the real-symmetric case. A similar calculation along the same lines gives the corresponding result when $\beta = 2$.

□

It is clear from (8.3) that the Ornstein-Uhlenbeck process starting at $t = 0$ from some fixed matrix $M_0$ is equal in distribution at any later fixed time $t$ to the sum of $e^{-t/2} M_0$, and $\sqrt{1-e^{-t}} G$, where $G$ is

[27] This corresponds to applying Ito's formula from stochastic calculus.

[28] There is a technical subtlety that we are ignoring in this step: one needs to prove that the singularities coming from the factors $(\lambda_i - \lambda_j)^{-1}$ do not overwhelm the other terms. To do this one needs to show that under the dynamics the eigenvalues do not get too close. That this is true is a consequence of the eigenvalue repulsion in the GOE and GUE. We do not give the argument here; see, for example, Section 4.3.1 of [2] for the proof that this subtlety can indeed be ignored.

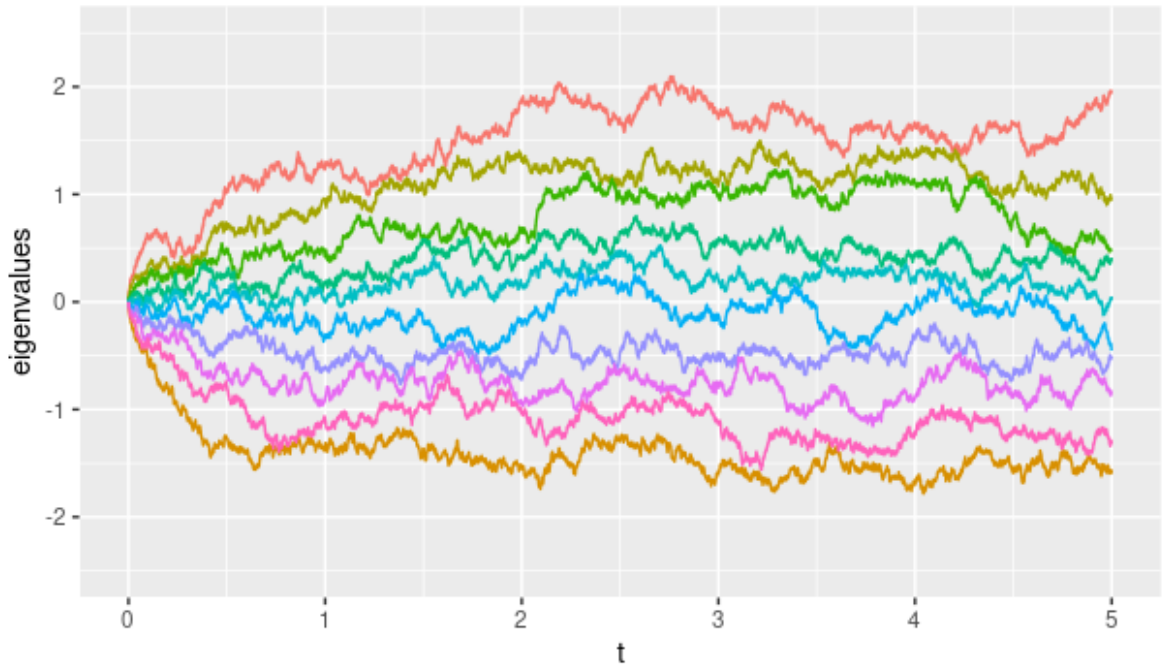


Figure 12: A simulation of GOE Dyson Brownian motion when $n = 10$. (Plot kindly provided by Johannes Forkel.)

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 12, 13 and 14 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 15, 16 and 17, where $M_0$ is taken to be a GOE matrix in each case, again with, respectively, $n = 10$, $n = 20$ and $n = 50$.

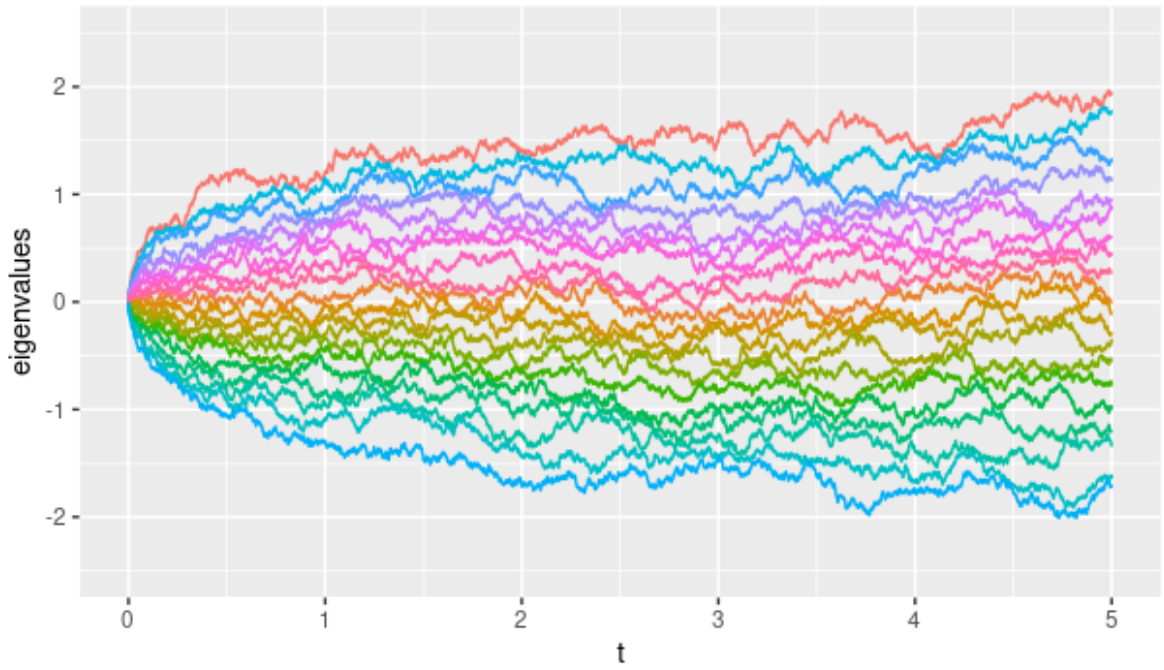


Figure 13: A simulation of GOE Dyson Brownian motion when $n = 20$. (Plot kindly provided by Johannes Forkel.)

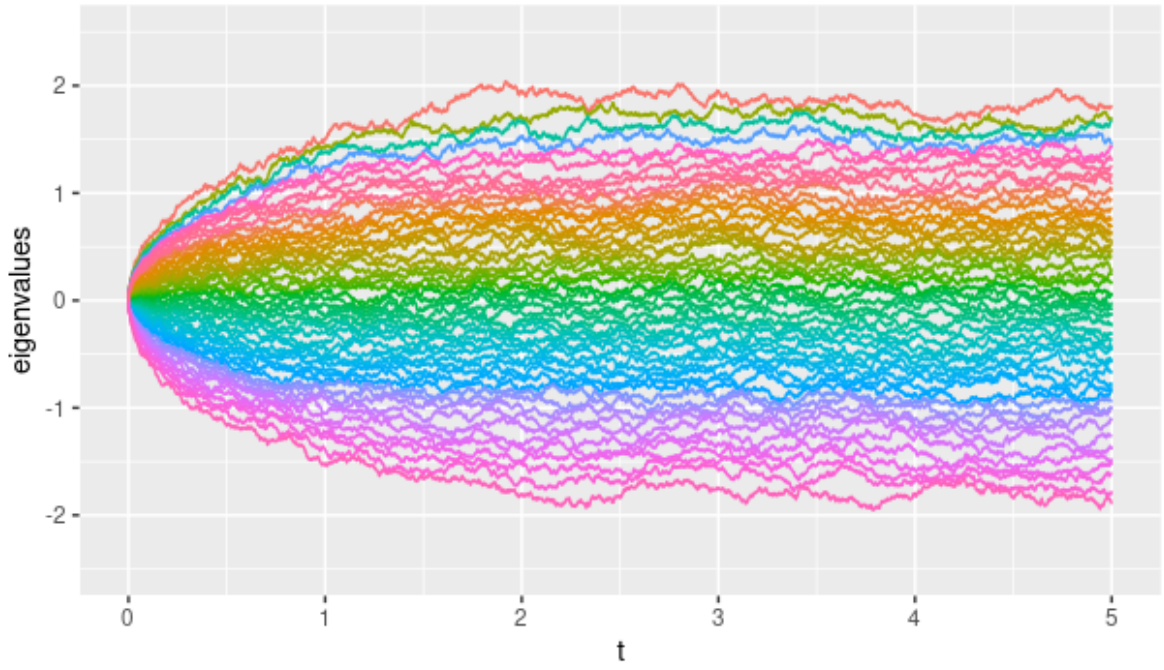


Figure 14: A simulation of GOE Dyson Brownian motion when $n = 50$. (Plot kindly provided by Johannes Forkel.)

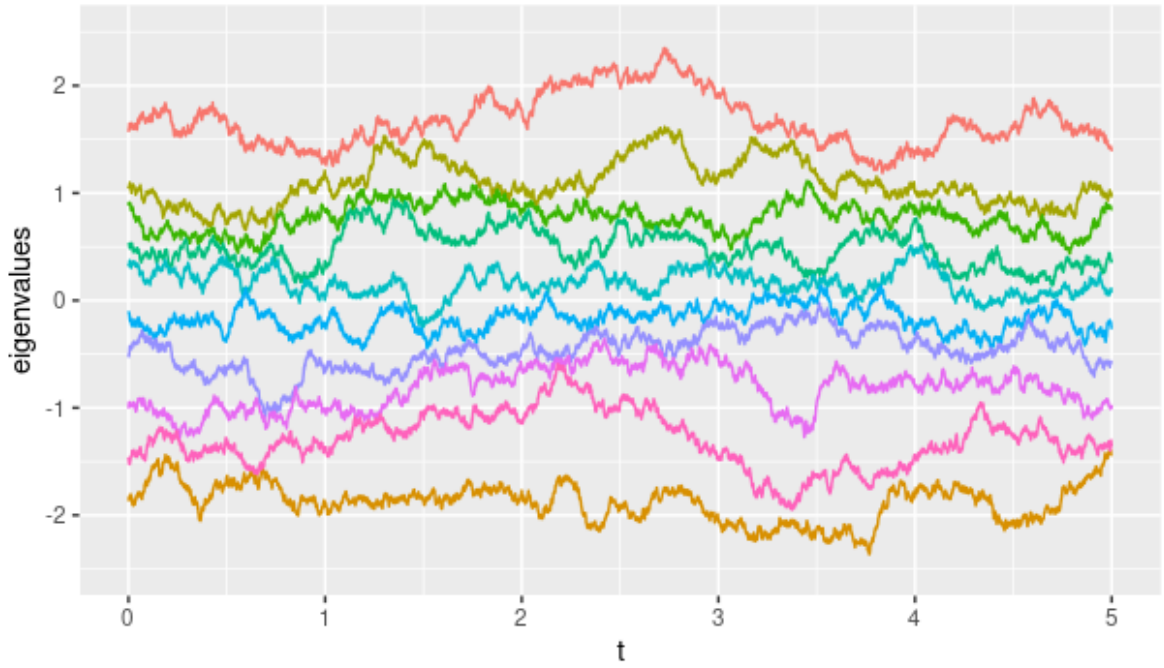


Figure 15: A simulation of GOE Dyson Brownian motion when $n = 10$, starting from a GOE matrix. (Plot kindly provided by Johannes Forkel.)

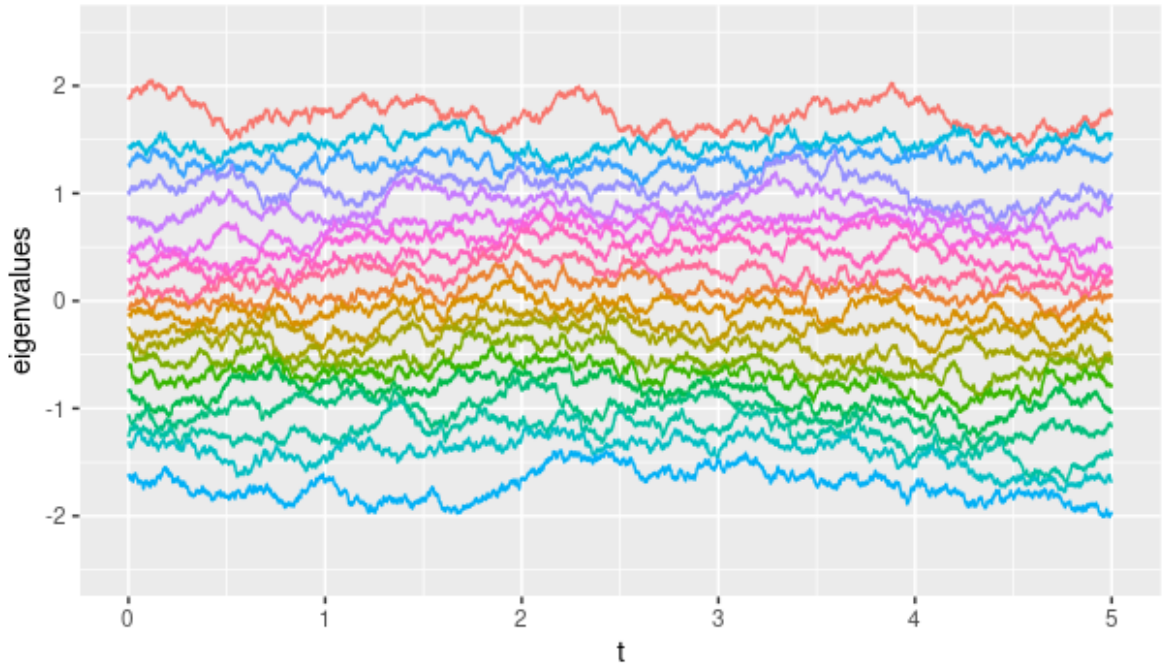


Figure 16: A simulation of GOE Dyson Brownian motion when $n = 20$, starting from a GOE matrix. (Plot kindly provided by Johannes Forkel.)

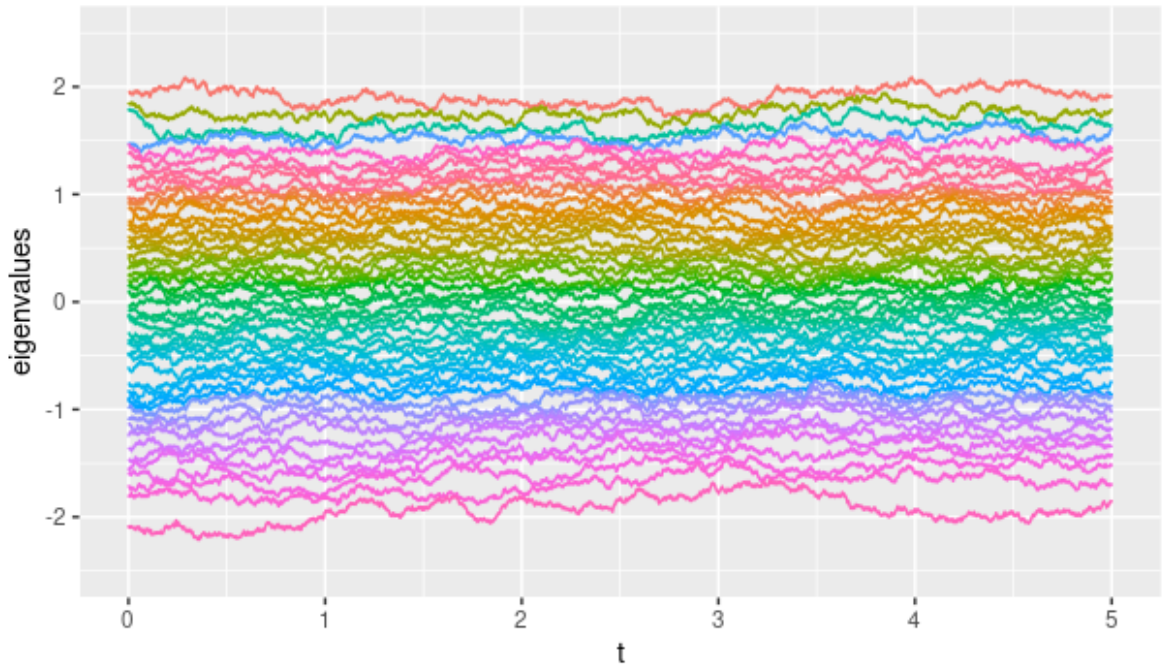


Figure 17: A simulation of GOE Dyson Brownian motion when $n = 50$, starting from a GOE matrix. (Plot kindly provided by Johannes Forkel.)

# 9 Some 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.

## 9.1 Longest increasing subsequences

As discussed in the introduction, let $S_n$ be the group of permutations of $1, 2, \ldots, n$. If $\pi \in S_n$, $\pi(i_1), \ldots, \pi(i_k)$ is an increasing subsequence in $\pi$ 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 $\pi$ 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 $S_n$ with uniform distribution,

$$\mathbb{P}(l_n \leq m) = \frac{\#\{\pi \in S_n : l_n(\pi) \leq m\}}{n!}. \tag{9.1}$$

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

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{l_n - 2\sqrt{n}}{n^{1/6}} \leq s\right) = F^{(\max)}(s) = \det(I - K_{\text{Airy}}) \tag{9.2}$$

where

$$K_{\text{Airy}} = \frac{\text{Ai}(x)\text{Ai}'(y) - \text{Ai}'(x)\text{Ai}(y)}{(x-y)} \tag{9.3}$$

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

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{l_n - 2\sqrt{n}}{n^{1/6}} \le s\right) = \exp\left(-\int_s^\infty (x-s)q(x)^2 \mathrm{d}x\right) \tag{9.4}$$

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.

## 9.2 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 [16]. For more about random growth models and random tilings, see the article by P. Ferrari and H. Spohn in [1].

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 $\alpha$. 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 $s \in \mathbb{R}$,

$$\lim_{\alpha\to\infty} \mathbb{P}\left(\frac{L(\alpha) - 2\sqrt{\alpha}}{\alpha^{1/6}} \le s\right) = F^{(\max)}(s) = \det(I - K_{\mathrm{Airy}}). \tag{9.5}$$

## 9.3 Zeros of the Riemann zeta-function

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

$$\zeta(s) = \sum_{n=1}^\infty \frac{1}{n^s} = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1} \tag{9.6}$$

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 $\mathrm{Re}s = 1/2$; that is, they are all of the form $1/2 + it_n$ 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

$$\#\{n : 0 < t_n \le T\} = \frac{T}{2\pi} \log\left(\frac{T}{2\pi}\right) + O(\log T) \tag{9.7}$$

Therefore, setting

$$w_n = \frac{t_n}{2\pi} \log\left(\frac{|t_n|}{2\pi}\right) \tag{9.8}$$

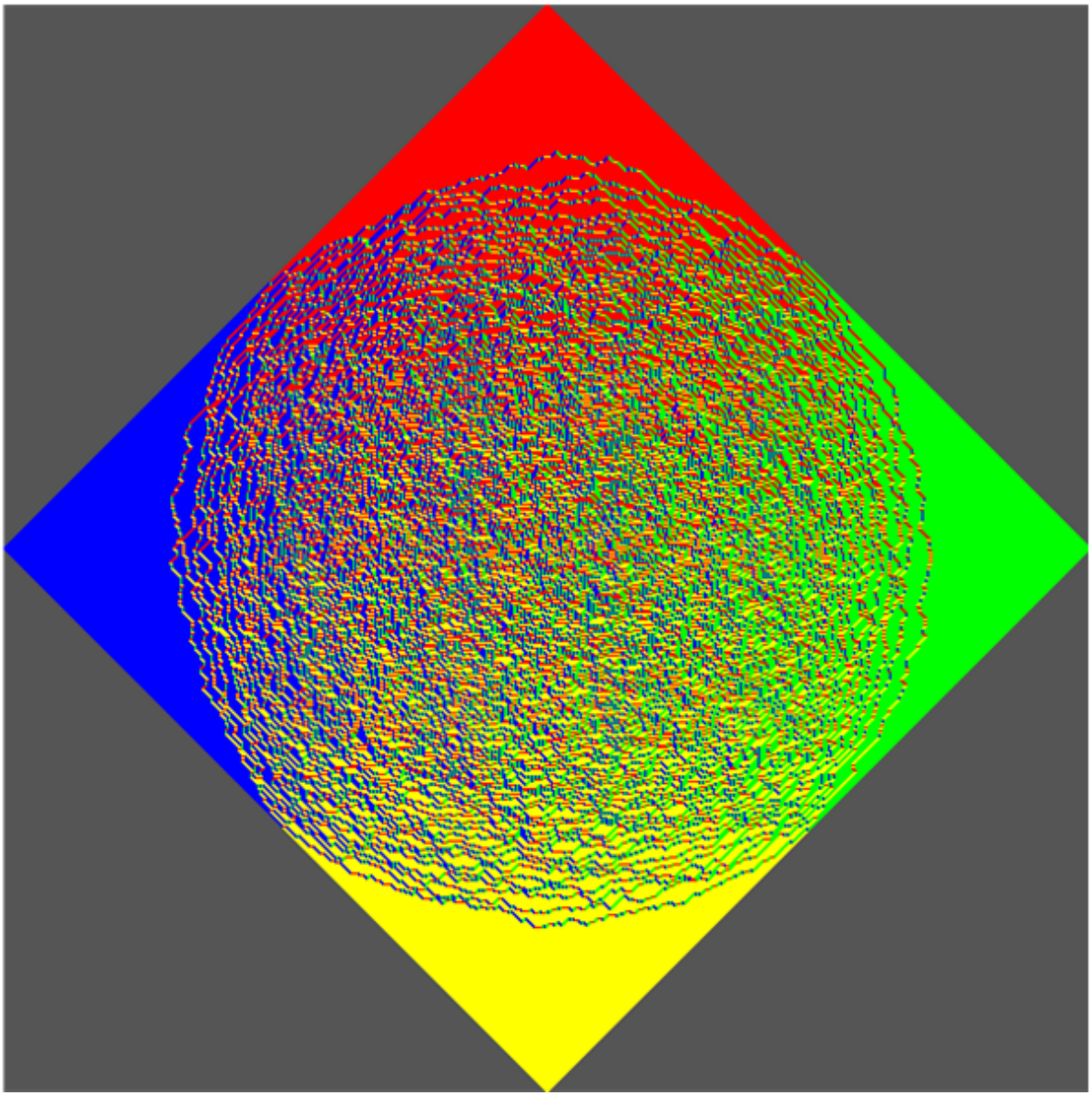

Figure 18: A random tiling, generated by P. Ferrari.

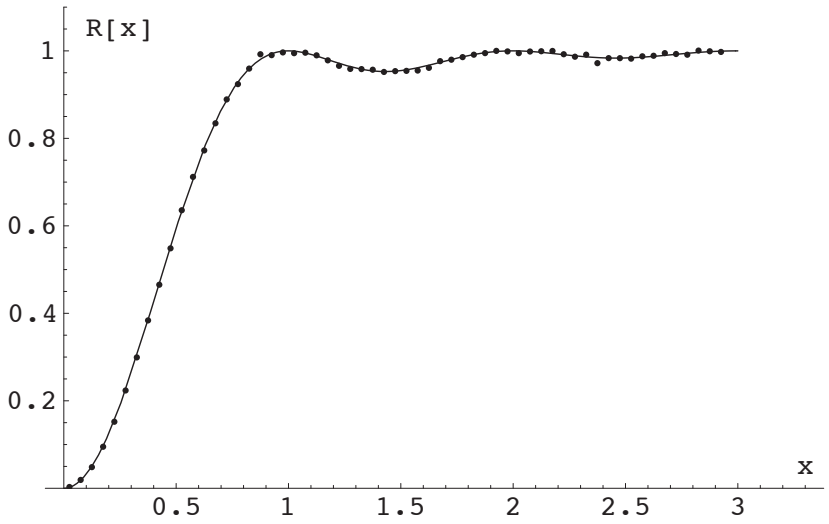


Figure 19: The 2-point correlation function for the Riemann zeros, computed by A. Odlyzko for zeros near to the $10^{20}$th, compared to the GUE 2-point correlation function (9.11).

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

$$\lim_{W\to\infty} \frac{1}{W} \#\{n : 0 < w_n \leq W\} = 1 \tag{9.9}$$

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

$$\lim_{N\to\infty} \frac{1}{N} \sum_{1\leq n<m\leq N} f(w_n - w_m) = \int_{-\infty}^{\infty} f(x) R_2(x) \mathrm{d}x \tag{9.10}$$

It is a theorem due to Montgomery in 1973 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

$$R_2(x) = 1 - \left(\frac{\sin \pi x}{\pi x}\right)^2 \tag{9.11}$$

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 9.11. Assuming this is true, many interesting properties of the zeta function and the primes can be calculated using random matrix theory. See the chapter on connections between number theory and random matrix theory in [1]. This remains a highly active area of research.

# References

[1] G. Akemann, J. Baik & P. Di Francesco, The Oxford Handbook of Random Matrix Theory (Oxford University Press).

[2] G.W. Anderson, A. Guionnet & O. Zeitouni, An Introduction to Random Matrices (Cambridge Studies in Advanced Mathematics).

[3] J.P. Bouchaud & M. Potters, Financial Applications of Random Matrix Theory: a short review, in The Oxford Handbook of Random Matrix Theory, edited by Gernot Akemann, Jinho Baik, & Philippe Di Francesco (OUP), ;also available from arXiv:0910.1205

[4] Z. Bai & J.W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices (Springer).

[5] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous & Y. LeCun, The loss surfaces of multilayer networks, Artificial intelligence and statistics, 192–204 (2015).

[6] P. Diaconis & P.J. Forrester, Hurwitz and the origins of random matrix theory in mathematics, Random Matrices: Theory and Applications, 06, 1730001 (2017); arXiv:1512.09229.

[7] A. Edelman & N. Raj Rao, Random matrix theory, Acta Numerica 14, 233–297 (2005).

[8] I.M. Johnstone, High dimensional statistical inference and random matrices, Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006; arXiv:math/0611589.

[9] L. Laloux, P. Cizeau, M. Potters & J.P. Bouchaud, Random matrix theory and financial correlations, available from https://www.math.nyu.edu/faculty/avellane/LalouxPCA.pdf.

[10] T. Koshy, Catalan Numbers with Applications, Oxford University Press (2008).

[11] R.M. May, Will a large complex system be stable?, Nature 238, 413-414 (1972).

[12] E.S. Meckes, The Random Matrix Theory of the Classical Compact Groups, Cambridge Tracts in Mathematics 218 (CUP, 2019); available online from https://case.edu/artsci/math/esmeckes/books.html

[13] J. von Neumann & H. H. Goldstine, Numerical inverting of matrices of high order, Bull. Amer. Math. Soc. 53, 1021–1099 (1947).

[14] G. Szegö, Orthogonal Polynomials, 4th ed. Providence, RI: Amer. Math. Soc., 42–44, (1975).

[15] R.P. Stanley, Catalan numbers, Cambridge University Press (2015).

[16] K.A. Takeuchi, M. Sano, T. Sasamoto, & H. Spohn, H., Growing interfaces uncover universal fluctuations behind scale invariance, Scientific Reports, 1: 34 (2011); arXiv:1108.2118

[17] E.P. Wigner, On the distribution of the roots of certain symmetric matrices, Annals of Mathematics 67, 325–327 (1958).

[18] J. Wishart, The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population, Biometrika 20A, 32–52 (1928).