Random Matrix Theory in a nutshell

Part II: Random Matrices

Manuela Girotti

based on M. Girotti's PhD thesis,
A. Kuijlaars' and M. Bertola's lectures from Les Houches Winter School 2012,
and B. Eynard's notes from IPhT Saclay 2015

Contents

1 Introduction

A random matrix is a matrix whose elements are randomly distributed. A random matrix model is characterized by a matrix ensemble $\mathcal{E}$ and a probability measure $\mathrm{d}\mu(M)$ for $M \in \mathcal{E}$ (the *random matrix law*), thus the matrix itself is a random variable.

Let $\mathfrak{M}$ be a space of matrices of given size: e.g.

- Hermitian matrices ($M = M^\dagger$) of size $n \times n$: $\mathfrak{M} = \{ M \in \mathrm{Mat}_n(\mathbb{C}) \mid M_{ij} = M_{ji}^* \}$ (**Unitary ensemble**)
- Symmetric matrices ($M = M^T$) of size $n \times n$: $\mathfrak{M} = \{ M \in \mathrm{Mat}_n(\mathbb{R}) \mid M_{ij} = M_{ji} \}$ (**Orthogonal ensemble**)
- Symplectic matrices: $M^T J = J M^T$, with $J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \otimes \mathbf{1}_n$ of size $2n \times 2n$ (**Symplectic ensemble**)
- Rectangular matrices of size $n \times K$
- $\mathfrak{M} = \mathrm{Mat}_n(\mathbb{C})$
- etc.

Remark 1. *The first three examples are extensively studied and their names refer to the compact group that leaves the measure invariant.*

A simple way to define a probability measure on these ensembles relies on the remark that each of these spaces is a vector space and thus carries a natural flat Lebesgue measure (invariant by translations) which we shall denote by $\mathrm{d}M$.

Then, starting from $\mathrm{d}M$, we can equip each of these spaces with a probability measure of the form

$$\mathrm{d}\mu(M) = F(M)\mathrm{d}M$$

where $F : \mathfrak{M} \to \mathbb{R}_+$ is some suitable ($L^1(\mathrm{d}M)$) function of total integral 1 (this is a measure which is *absolutely continuous* with respect to Lebesgue measure).

The main objective in Random Matrix Theory typically is to study the statistical properties of the spectra (for square matrices ensembles) or singular values (for rectangular ensembles). In order to do so, we need to develop an understanding of the joint probability distribution functions (jpdf) of the eigen-/singular-values.

Additional interest is the study of the properties of the said statistics when the size of the matrix ensemble tends to infinity (under suitable assumption on the probability measure).

Random Matrices are one of those transversal theories who may appear in different fields of Mathematics and Physics, providing unexpected links between for example Probability, Number Theory and Integrable Systems.

Among the vast literature on Random Matrix Theory, we can mention the book by Mehta [5] and the book by Anderson, Guionnet and Zeitouni [3].

2 A little bit of history

The first appearance of the concept of a random matrix dates back to the Fifties and it is due to the physicist E.P. Wigner. In the field of Nuclear Physics, Wigner wished to describe the general properties of the energy levels of highly excited states of heavy nuclei, as measured in nuclear reactions. In particular, he wanted to study the spacings between those energy levels.

Such a complex nuclear system is usually represented by a Hermitian operator $\mathcal{H}$, called the Hamiltonian, defined on an infinite-dimensional Hilbert space and governed by physical laws. However, except for very specific and simple cases, $\mathcal{H}$ is unknown or very hard to compute.

On the other hand, the real quantities of interest are the eigenvalues of $\mathcal{H}$, which represent the energy levels:

$$\mathcal{H}v = \lambda v \tag{1}$$

where v is the eigenfunction associated to the eigenvalue λ.

Wigner argued that one should regard a specific Hamiltonian $\mathcal{H}$ as behaving like a large-dimension matrix with random entries. Such a matrix is thought as a member of a large class of Hamiltonians, all of which would have similar general properties as the specific Hamiltonian $\mathcal{H}$ in question ([12]). As a consequence, the eigenvalues of $\mathcal{H}$ could then be approximated by the eigenvalues of a large random matrix and the spacings between energy levels of heavy nuclei could be modelled by the spacings between successive eigenvalues of a random $n \times n$-matrix as $n \to +\infty$.

It turns out that the ensemble of the random eigenvalues is a determinantal point process. Therefore, studying the spacings or gaps between eigenvalues means studying the gap probabilities of the determinantal system. Furthermore, the distribution of the largest eigenvalue obeys a different law on its own and is governed by the so called "Tracy-Widom" distribution ([11]), which can still be considered as a gap probability on an interval of the type $[s, +\infty)$, $s \in \mathbb{R}$ (the eigenvalues, or in general the points of a DPP, are always confined in finite positions on the real line).

3 Unitary matrices

We will focus from now on on the case of Hermitian matrices (Unitary ensemble). This is a vector space with the real diagonal entries $\{M_{ii}\}_{i=1}^n$ and the real and imaginary part of the upper diagonal elements $\{\Re M_{ij},\ \Im M_{ij}\}_{i<j}$ as independent coordinates:

$$M_{ij} = \Re M_{ij} + i\Im M_{ij}, \qquad \text{with} \qquad \Re M_{ij} = \Re M_{ji},\ \Im M_{ij} = -\Im M_{ij},\ n = 1, \dots, n.$$

Its dimension is equal to

$$\dim \mathfrak{M} = \frac{n(n+1)}{2} + \frac{n(n-1)}{2} = n^2$$

and the corresponding Lebesgue measure reads

$$\mathrm{d}M = \prod_{i=1}^{n} \mathrm{d}M_{ii} \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \mathrm{d}\Re M_{ij} \mathrm{d}\Im M_{ij}. \tag{2}$$

We also recall the following properties of Hermitian matrices:

Theorem 2 (**Spectral Theorem**). *Any Hermitian matrix can be diagonalized by a Unitary matrix*

$$U \in \mathcal{U}(n) = \{\, U \in \mathrm{GL}_n(\mathbb{C}) \mid U^\dagger U = U U^\dagger = \mathbf{1}_n \,\}$$

and its eigenvalues are real:

$$M = U^\dagger X U, \qquad X = \mathrm{diag}\,\{x_1, \dots, x_n\}\,, \qquad x_j \in \mathbb{R}. \tag{3}$$

Remark 3. *The diagonalization is not unique even if X has distinct eigenvalues, because of the ordering of the eigenvalues, so in general there are $n!$ distinct diagonalizations.*

Additionally, the Lebesgue measure (2) is invariant under conjugation with a unitary matrix

$$\mathrm{d}M = \mathrm{d}(UMU^\dagger), \quad U \in \mathcal{U}(n) \tag{4}$$

(more generally, for all ensembles of square matrices the Lebesgue measure is invariant under conjugation: $\mathrm{d}M = \mathrm{d}(CMC^{-1})$).

In view of these properties, we can perform a strategic change of variables

$$\begin{aligned} M &\mapsto (X, U) \\ \{M_{ii},\ i = 1, \dots, n;\ \Re M_{ij}, \Im M_{ij},\ i < j\} &\mapsto \{x_1, \dots, x_n;\ u_{ij}\}\,, \end{aligned} \tag{5}$$

where u_{ij} are the parameters that parametrize the unitary group. Under such transformation, the Lebesgue measure reads (thanks to the Weyl integration formula)

$$\mathrm{d}M = c_n\, \Delta(X)^2\, \mathrm{d}X\, \mathrm{d}U \tag{6}$$

where $c_n = \frac{\pi^{n(n-1/2)}}{\prod_{j=1}^n j!}$,

$$\Delta(X) = \prod_{1 \le i < j \le n} (x_i - x_j) = \det\left[x_a^{b-1}\right]_{1 \le a,b \le n} \tag{7}$$

is the *Vandermonde determinant* and $\mathrm{d}U$ is the Haar measure on $\mathcal{U}(n)$.

Remark 4. *Similarly, for the other two main cases*

$$\begin{array}{ll} \textit{Orthogonal} & \mathrm{d}M \sim |\Delta(X)| \mathrm{d}X \mathrm{d}U \\ \textit{Symplectic} & \mathrm{d}M \sim \Delta(X)^4 \mathrm{d}X \mathrm{d}U \end{array}$$

where dU is the Haar measure in the respective compact group ($\mathcal{O}(n)$ or $\mathcal{S}p(2n)$). Since the exponent of the Vandermonde determinant $\Delta(X)$ is $\beta = 1, 2, 4$ (Orthogonal, Unitary, Symplectic ensembles), they are also universally known as the $\beta = 1, 2, 4$ ensembles.

3.1 Eigenvalues distribution

We now want to equip the space $\mathfrak{M}$ with a probability measure. Consider again a measure $\mathrm{d}\mu(M)$ that is absolutely continuous with respect to Lebesgue:

$$\mathrm{d}\mu(M) = F(M) \mathrm{d}M \tag{8}$$

with $F \in L^1(\mathfrak{M}, \mathrm{d}M)$ and $\int F(M)\mathrm{d}M = 1$. Thus, under the "change of variable" performed before,

$$\mathrm{d}\mu(\vec{x}, U) = c_n F(U^\dagger X U)\Delta(X)^2 \,\mathrm{d}x_1 \dots \mathrm{d}x_n \,\mathrm{d}U$$

If we are interested only on the eigenvalues one can study the reduced measure

$$\begin{aligned}\mathrm{d}\mu(\vec{x}) &= \Delta(X)^2 \,\mathrm{d}x_1 \dots \mathrm{d}x_n \times \left(\int_{\mathcal{U}(n)} c_n F(U^\dagger X U) \,\mathrm{d}U \right) \\ &= \Delta(X)^2 \, \widetilde{F}(x_1, \dots, x_n) \,\mathrm{d}x_1 \dots \mathrm{d}x_n \end{aligned} \tag{9}$$

where $\widetilde{F}$ needs to be a symmetric function of the n arguments.

The connection to Orthogonal Polynomials becomes possible when $\widetilde{F}$ is the product of a single function of the individual eigenvalues:

$$\widetilde{F}(x_1, \dots, x_n) \sim \prod_{j=1}^{n} e^{-V(x_j)} \tag{10}$$

($V(x)$ is called the *potential*).

Remark 5. *A sufficient condition for the probability (9) to be well-defined is that*

$$\lim_{|x| \to +\infty} \frac{V(x)}{\ln(1 + x^2)} = +\infty. \tag{11}$$

A standard example is when $V(x)$ is a polynomial of even degree, with positive leading coefficient (e.g. $V(x) = x^2$).

In conclusion, the probability measure on the space of matrices (8) induces a joint probability density on the eigenvalues given by

$$\mathrm{d}\mu(x_1, \dots, x_n) = \frac{1}{Z_n} \Delta(x_1, \dots x_n)^2 \prod_{j=1}^{n} e^{-V(x_j)} \mathrm{d}x_1 \dots \mathrm{d}x_n \tag{12}$$

with $Z_n = \int_{\mathbb{R}^n} \mathrm{d}\mu(x_1, \dots, x_n)$ a suitable normalization constant (*partition function*).

Paradigma. The **Gaussian Unitary Ensemble** (GUE) is the ensemble on Hermitian matrices equipped with the probability measure

$$\mathrm{d}\mu(M) = \frac{1}{Z_n} e^{-\frac{1}{2}\operatorname{Tr} M^2} \mathrm{d}M. \tag{13}$$

Since

$$\begin{aligned}\operatorname{Tr} M^2 = \sum_{i=1}^{n} (M^2)_{ii} = \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} M_{ji} = \sum_{i=1}^{n} M_{ii}^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |M_{ij}|^2 = \\ = \sum_{i=1}^{n} M_{ii}^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left[(\Re M_{ij})^2 + (\Im M_{ij})^2 \right], \end{aligned} \tag{14}$$

the probability measure (13) factorizes as a product of Gaussians

$$\mathrm{d}\mu(M) = \frac{1}{Z_n}\prod_{i=1}^{n} e^{-\frac{1}{2}M_{ii}^2}\mathrm{d}M_{ii}\prod_{i=1}^{n-1}\prod_{j=i+1}^{n}\left(e^{-(\Re M_{ij})^2}\mathrm{d}\Re M_{ij}\right)\left(e^{-(\Im M_{ij})^2}\mathrm{d}\Im M_{ij}\right). \tag{15}$$

Therefore, in GUE all the entries $\{\Re M_{ij},\ \Im M_{ij}\}_{i<j}$ and $\{M_{ii}\}$ are mutually independent normal random variable with zero mean and different variances for the diagonal and off-diagonal entries:

$$\Re M_{ij},\ \Im M_{ij} \sim \mathcal{N}\left(0, \frac{1}{2}\right) \qquad M_{ii} \sim \mathcal{N}(0,1)\,. \tag{16}$$

Furthermore, the induced joint probability density of the eigenvalues is

$$\mathrm{d}\mu(x_1,\ldots,x_n) = \frac{1}{Z_n}\prod_{1\le i<j\le n}(x_i-x_j)^2\, e^{-\frac{1}{2}\sum_{j=1}^{n}x_j^2}\,\mathrm{d}x_1\ldots\mathrm{d}x_n. \tag{17}$$

3.2 Dyson's Theorem

We start from the eigenvalue distribution. Calling $W(X)$ the Vandermonde matrix

$$W(X) = \begin{bmatrix} 1 & x_1 & \ldots & x_1^{n-1} \\ 1 & x_2 & \ldots & x_2^{n-1} \\ \vdots & & & \\ 1 & x_n & \ldots & x_n^{n-1} \end{bmatrix} \tag{18}$$

then, the Vandermonde determinant is $\Delta(X) = \det W(X)$:

$$\begin{aligned}
\mathrm{d}\mu(x_1,\ldots,x_n) &= \frac{1}{Z_n}\Delta(x_1,\ldots x_n)^2\prod_{i=1}^{n}e^{-V(x_i)}\mathrm{d}x_1\ldots\mathrm{d}x_n \\
&= \frac{1}{Z_n}\Big(\det\left[x_i^{j-1}\right]_{1\le i,j\le n}\Big)^2\prod_{i=1}^{n}e^{-V(x_i)}\mathrm{d}x_1\ldots\mathrm{d}x_n \\
&= \frac{1}{Z_n}\prod_{i=1}^{n}e^{-\frac{V(x_i)}{2}}\det\left[x_i^{j-1}\right]_{1\le i,j\le n}\cdot\det\left[x_i^{j-1}\right]_{1\le i,j\le n}\prod_{i=1}^{n}e^{-\frac{V(x_i)}{2}}\mathrm{d}x_1\ldots\mathrm{d}x_n \\
&= \frac{1}{Z_n}\det\left[e^{-\frac{V(X)}{2}}\right]\det\left[x_i^{j-1}\right]_{1\le i,j\le n}\cdot\det\left[x_i^{j-1}\right]_{1\le i,j\le n}\det\left[e^{-\frac{V(X)}{2}}\right]\mathrm{d}x_1\ldots\mathrm{d}x_n \\
&= \frac{1}{Z_n}\det\left[x_i^{j-1}e^{-\frac{V(x_i)}{2}}\right]_{1\le i,j\le n}\cdot\det\left[x_i^{j-1}e^{-\frac{V(x_i)}{2}}\right]_{1\le i,j\le n}\mathrm{d}x_1\ldots\mathrm{d}x_n
\end{aligned} \tag{19}$$

where we used $V(X) = \mathrm{diag}\{V(x_1),\ldots,V(x_n)\}$.

Proposition 6. *The partition function is*

$$Z_n = n!\det\mathbb{M}, \tag{20}$$

where the matrix $\mathbb{M}$ *has entries*

$$\mathbb{M}_{ab} = \int_{\mathbb{R}^n} x^{a+b}e^{-V(x)}\mathrm{d}x \qquad 0\le a,b\le n-1. \tag{21}$$

Proof. From the definition of the constant Z_n and (19), use the definition of the determinant of a matrix to obtain the right-hand-side. □

Even better,

Proposition 7.

$$\frac{1}{Z_n}\prod_{1\le i<j\le n}(x_1,\dots x_n)^2\prod_{i=1}^{n}e^{-V(x_i)}=\frac{1}{n!}\det\Big[K_n(x_i,x_j)\Big]_{1\le i,j\le n} \tag{22}$$

where

$$K(x,y)=e^{-\frac{V(x)+V(y)}{2}}\sum_{j,k=0}^{n-1}x^j[\mathbb{M}]^{-1}_{jk}y^k \tag{23}$$

Proof.

$$\frac{1}{n!}\det\Big[K_n(x_a,x_b)\Big]_{1\le i,j\le n}=\frac{1}{n!}\det\left[\sum_{j,k}e^{-\frac{V(x_a)}{2}}x_a^j[\mathbb{M}]^{-1}_{jk}x_b^k e^{-\frac{V(x_b)}{2}}\right]=\frac{1}{n!}\det\left[e^{-\frac{V(X)}{2}}W\mathbb{M}^{-1}W^Te^{-\frac{V(X)}{2}}\right]$$
$$=\frac{1}{n!\det\mathbb{M}}\left[\det W(X)\right]^2e^{-\operatorname{Tr}V(X)}=\frac{1}{Z_n}\Delta^2(X)e^{-\operatorname{Tr}V(X)} \tag{24}$$

□

Finally,

Proposition 8. *The kernel $K(x,y)$ has the following properties:*

1. *(***reproducibility***)* $\int_{\mathbb{R}}K(x,z)K(z,y)\mathrm{d}z=K(x,y)$
2. *(***normalization***)* $\int_{\mathbb{R}}K(x,x)\mathrm{d}x=n$
3. *(***marginals***)* $\int_{\mathbb{R}}\det\left[K(x_i,x_j)\right]_{i,j\le r}\mathrm{d}x_r=(n-r-1)\det\left[K(x_i,x_j\right]_{i,j\le r-1}$
4. *(***marginals***)* $\int_{\mathbb{R}^{n-r}}\det\left[K(x_i,x_j)\right]_{i,j\le n}\mathrm{d}x_{r+1}\dots\mathrm{d}x_n=(n-r)!\det\left[K(x_i,x_j\right]_{i,j\le r-1}$

Conclusion 1: Dyson's theorem says that the joint probability density function of the eigenvalues and all its marginals are in a determinantal form. Therefore, the set of random eigenvalues of a (unitary) matrix ensemble is a determinantal point process!

Conclusion 2: As already seen in the DPP part of the notes, the whole statistical information is contained in the kernel $K(x,y)$.

4 Asymptotics and universality

As it was said in the introduction, a typical question one asks when dealing with random matrices is what happens to the statistical properties of the eigenvalues when the size of the matrix ensemble tends to infinity (or equivalently, when the number of eigenvalues grows).

The goal is to find **asymptotic behaviours** or asymptotic properties of the probability distribution and its related quantities.

A property is considered **universal** if it only depends on the matrix ensemble, and not – or almost not – on the probability measure (in particular, we have independency with respect to the choice of the potential $V(x)$).

This is a quite vague description of the picture and in fact there are different ways to study the asymptotics of a matrix ensemble.

4.1 Macroscopic behaviour

We are interested in the distribution of the eigenvalues as a whole, when the dimension of the matrix grows.

Consider a matrix ensemble and denote the (ordered) eigenvalues by $x_1 \leq x_2 \leq \ldots \leq x_n$. The empirical spectral distribution of the eigenvalues is defined by

$$\mathrm{d}\mu_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x)\mathrm{d}x \tag{25}$$

where δ_x is the Dirac delta function centered at x.

Numerical evaluations shows that as we increase the size of the matrix and at the same time we scale down the phase space by a suitable power of n, we can notice a limiting shape appearing from the hystograms of the eigenvalues: see Figure 1.

This means that the (rescaled) eigenvalues do not escape at infinity. In many cases of interest the eigenvalue density has a finite limit as $n \to +\infty$, called the **equilibrium density** $\rho(x)$.

We saw that the ensemble of eigenvalues is a determinantal point process and in particular the 1-point correlation function (or density function) can be given in terms of the correlation kernel as

$$\rho_1(x) = K_n(x, x). \tag{26}$$

Proposition 9. *The equilibrium density (if it exists) can be computed as*

$$\lim_{n\to+\infty} \frac{1}{n} K_n(x, x) = \rho(x). \tag{27}$$

4.1.1 Wigner's semicircle law

One notable example is the case where we consider the GUE ensemble. It is possible to show that the second moment of the eigenvalue distribution measure of the matrix M behaves like n ($\mathbb{E}\left[M^2\right] \sim n$) and therefore it is divergent. On the other hand, if we "smartly" rescale the matrices

$$\widetilde{M} = \frac{1}{\sqrt{n}} M, \tag{28}$$

then the corresponding eigenvalue distribution density has finite moments. Its limit distribution has a very peculiar shape as a semicircle.

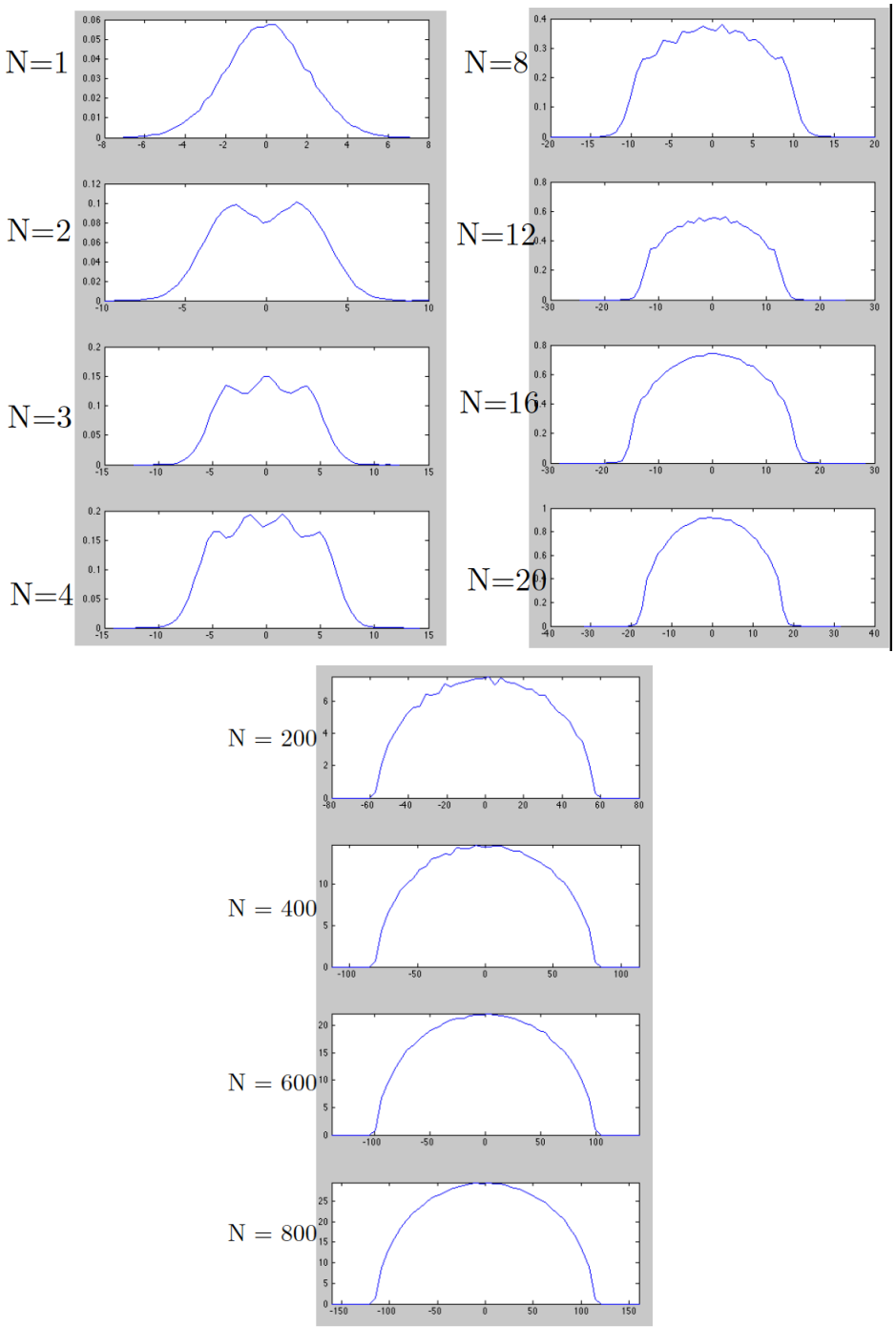


Figure 1: Histograms of the eigenvalues of GUE matrices as the size of the matrix increases (such histograms are produced with rescaling the spectrum by a factor $1/\sqrt{n}$). Numerical simulation with MATLAB (courtesy of prof. Ken McLaughlin).

Theorem 10 (**Wigner's semicircle law**). *Consider the GUE ensemble of size n with matrices $\frac{1}{\sqrt{n}}M$, then the spectrum distribution converges weakly as $n \to +\infty$ to the following deterministic probability density*

$$\rho(x) = \begin{cases} \frac{1}{2\pi}\sqrt{4-x^2} & \text{if } |x| \leq 2 \\ 0 & \text{otherwise.} \end{cases} \tag{29}$$

More precisely, $\forall\ f \in C^0(\mathbb{R})$ bounded, $\forall\ \epsilon > 0$

$$\lim_{n\to+\infty} \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int f(t)\rho(t)\ \mathrm{d}t\right| \geq \epsilon\right] = 0. \tag{30}$$

Observation on the proof. There are several ways to prove the theorem. Historically this was proven using the so-called "moments method", but it can also be proved using the Stieltjes transform (see Appendix A) and other more recent methods. □

The semicircle law is characteristic for a large class of random matrices. The minimum requirements are that the matrices are hermitian (or symmetric, if we are considering matrices with real entries), with mean zero and finite variance independent entries: $\mathbb{E}[M_{ij}] = 0$, $\mathbb{E}[M_{ij}^2] < +\infty$, $i = 1, \dots n$, $j = i, \dots, n$. Such type of matrices are generally called **Wigner matrices**.

This is already one example of the **universality** feature mentioned at the beginning of this section.

Remark 11. *If we consider random square matrices with independent entries but without symmetry (i.e. the entries M_{ij} are independent for all i, j) another universal pattern emerges, the so-called circular law.*

In particular, if the entries M_{ij} have zero mean and finite variance, the empirical density of eigenvalues converges to the uniform measure on the unit disk in the complex plane.

If independence is dropped, one can get many different density profiles.

4.2 Microscopic behaviour

Another aspect of interest about the distribution of eigenvalues is the local (infinitesimal) behaviour of the eigenvalue distribution in specific points of the spectrum in the limit as $n \to +\infty$.

Focusing on the GUE ensemble, the two settings that we can consider are points that lie in the interior of the spectrum (the **bulk**) or that lie on the boundary of the spectrum (the **edge**, meaning the largest or the smallest eigenvalue). In order to study their statistical behaviour we will again make use of the results seen in the DPP theory.

We recall the following result about transformations of DPPs.

Proposition 12. *Let $\mathbb{P}$ and $\mathbb{P}_n$ be determinantal point processes with kernels K and K_n respectively. Let K_n converge to K*

$$\lim_{n\to\infty} K_n(x, y) = K(x, y) \tag{31}$$

uniformly over compact subsets of $\mathbb{R} \times \mathbb{R}$. Then, the point processes $\mathbb{P}_n$ converge to $\mathbb{P}$ weakly.

Given a fixed reference point x^* of the spectrum, center and scale the DPP of eigenvalues, i.e. apply the change of variables

$$x \mapsto Cn^{\gamma}(x - x^*) \tag{32}$$

to the correlation kernel, with suitable values of $C,\ \gamma > 0$, depending on the RM model and on where we are focusing on (the edges behaviour or the bulk behaviour). We can now perform the limit:

$$\lim_{n\to\infty} \frac{1}{Cn^\gamma} K_n \left(x^* + \frac{x}{Cn^\gamma}, x^* + \frac{y}{Cn^\gamma} \right) = K(x,y) \tag{33}$$

with x, y the new local coordinates of the limit-DPP.

4.2.1 Bulk universality

A point x^* lies in the bulk of the spectrum if the equilibrium density doesn't vanish $\rho(x^*) \neq 0$ (we are actually requiring that the density doesn't vanish in a whole neighbourhood of x^*).

Pick a point x^* in the bulk of the spectrum and introduce the following change of variables:

$$x = x^* + \frac{\xi}{n\rho(x^*)} \qquad y = x^* + \frac{\eta}{n\rho(x^*)} \tag{34}$$

Theorem 13 (**Bulk universality at the origin).** *For the Unitary Ensemble, the local behaviour in the bulk of the spectrum is described by a DPP with correlation kernel given by*

$$\lim_{n\to+\infty} \frac{1}{n\rho(x^*)} K_n \left(x^* + \frac{\xi}{n\rho(x^*)}, x^* + \frac{\eta}{n\rho(x^*)} \right) = K_{\text{sine}}(\xi,\eta) \tag{35}$$

with

$$K_{\text{sine}}(\xi,\eta) = \frac{\sin\left(\pi(\xi-\eta)\right)}{\pi(\xi-\eta)}. \tag{36}$$

This results holds regardless on the choice of the potential $V(x)$ (despite some reasonable properties that we require $V(x)$ to have). Here is another universality results.

4.2.2 Soft-edge universality

In the case of points x^* close to a spectral edge a, the definition of an edge microscopic limit will depend on the behaviour of the equilibrium density $\bar{\rho}(x)$ near a.

For a generic potential V, we have **regular edges** or **soft edges** if the density vanishes as a square root: $\rho(x) \sim \sqrt{x-a}$. On the other hand for special choices of the potential, we can have that the vanishing of the density has a different regime $\rho(x) \sim (x-a)^{\frac{p}{q}}$ for some positive integers p, q; in this case, the edges are called **critical**.

In the case of UE with regular one-cut potential (e.g. $V(x) = x^2$), both edges (on $a = \pm 2\sqrt{n}$) are regular and the microscopic limit is described as

Theorem 14 (**Soft-edge universality).**

$$\lim_{n\to+\infty} \frac{1}{2n^{\frac{1}{6}}} K_n \left(\pm 2\sqrt{n} + \frac{\xi}{2n^{\frac{1}{6}}}, \pm 2\sqrt{n} + \frac{\eta}{2n^{\frac{1}{6}}} \right) = K_{\text{Airy}}(\xi,\eta) \tag{37}$$

with

$$K_{\text{Airy}}(\xi,\eta) = \frac{\text{Ai}(\xi)\text{Ai}'(\eta) - \text{Ai}'(\xi)\text{Ai}(\eta)}{\xi-\eta}. \tag{38}$$

Note 15. *The function* $\mathrm{Ai}(z)$ *is the Airy function. It satisfies the second-order ODE*

$$y'' = zy, \quad \text{such that } \lim_{z\to+\infty} y(z) = 0. \tag{39}$$

It can be represented as a contour integral

$$\mathrm{Ai}(z) = \int_\gamma e^{\frac{\zeta^3}{3} - z\zeta} \frac{\mathrm{d}\zeta}{2\pi i} \tag{40}$$

where the curve $\gamma \subseteq \mathbb{C}$ *is an oriented contour starting at* ∞ *with argument* $-\frac{\pi}{3}$ *and ending at* ∞ *with argument* $\frac{\pi}{3}$*.*

The corresponding gap probabilities of this (limit) DPP describe the local behaviour of the largest eigenvalue in the spectrum and it infinitesimal random oscillations are described by the celebrated Tracy–Widom distribution:

Theorem 16 (**Tracy–Widom distribution).** *Consider the semi-infinite interval* $[s, +\infty)$*, then the distribution of the largest eigenvalue of the GUE ensemble obeys the following law*

$$\det\left(\mathbf{1}_{L^2(\mathbb{R})} - K_{\mathrm{Ai}}\Big|_{[s,+\infty)}\right) = \exp\left\{-\int_s^\infty (x-s)q^2(x)\mathrm{d}x\right\} \tag{41}$$

where $q(x)$ *is the Hasting-Mc Leod solution to the Painlevé II equation:*

$$\begin{aligned} &q''(x) = 2q^3(x) + sq(x) \\ &q(x) \sim \mathrm{Ai}(x) \qquad x \to +\infty. \end{aligned} \tag{42}$$

Remark 17 (**food for thought).** *The Tracy–Widom distribution can be seen as the* τ*-function of the Painlevé II integrable system.*

5 A zoo of random matrix models

In these notes we only focused on unitary ensembles, i.e. squared Hermitian matrices with a given (smooth) potential $V(x)$. On the other hand, there is a vast variety of possible matrix models and the sky is the limit.

Here we'll mention just a few.

5.1 Wishart ensemble

The Wishart ensemble is the ensemble of matrices of the form $M = XX^T$, where X is a rectangular matrix $X \in \mathrm{Mat}_{n\times m}(\mathbb{R})$ with i.i.d entries, $\mathbb{E}[X_{ij}] = 0$, $\mathbb{E}[X_{ij}^2] = 1$.

The corresponding eigenvalue distribution has the following expression

$$\mathrm{d}\mu(x) = \prod_{i<j} |x_i - x_j| \prod_i x_i^\alpha e^{\frac{1}{2}\sum_i x_i} \mathrm{d}x_1 \dots \mathrm{d}x_n. \tag{43}$$

If we properly rescale the ensemble by

$$\widetilde{M} = \frac{1}{n} M \tag{44}$$

and take the limit as $n \to +\infty$, while assuming that $\frac{n}{m} \to \kappa$ ($\kappa \in (0,1]$), then the limit distribution of eigenvalues is defined on a bounded interval $[a_-, a_+]$ depending on κ

$$a_- = \left(1 - \sqrt{\kappa}\right)^2, \ a_+ = \left(1 + \sqrt{\kappa}\right)^2 \tag{45}$$

and it is equal to

$$\rho_{\mathrm{MP}}(x) = \frac{1}{2\pi\kappa} \frac{\sqrt{(a_+ - x)(x - a_-)}}{x}. \tag{46}$$

This distribution is called Marchenko-Pastur law (see Figure 2).

More formally,

Theorem 18 (**Marchenko-Pastur law**). *Let $X \in \mathrm{Mat}_{p,n}(\mathbb{R})$ with i.i.d. zero mean, unit variance, entries.*

In the limit as $p, n \to +\infty$, with $\frac{p}{n} \to \kappa \in \mathbb{R}$, the empirical spectral distribution $\mathrm{d}\mu_p$ of $\frac{1}{p} X X^T$ converges weakly, in probability, as $p \to \infty$ to the distribution $\mathrm{d}\mu_{\mathrm{MP}}$ with density function

$$\rho_{\mathrm{MP}}(x) = \begin{cases} \frac{1}{2\pi\kappa} \frac{\sqrt{(a_+ - x)(x - a_-)}}{x} & x \in [a_-, a_+] \setminus \{0\} \\ \max\{0, 1 - \kappa^{-1}\} & x = 0 \end{cases} \tag{47}$$

One way to prove this theorem is via Stieltjes transform (see Appendix A).

In particular, if $\kappa = 1$, then the distribution has a square-root singularity at $x = 0$:

$$\rho_{\mathrm{MP},1}(x) = \frac{1}{2\pi} \sqrt{\frac{4 - x}{x}} \qquad x \in (0, 4]. \tag{48}$$

In this configuration, the point $x = 0$ is called **hard-edge** and the local infinitesimal behaviour in a neighbourhood of $x = 0$ (in the limit as $n \to +\infty$) is described by a universal kernel called Bessel kernel

$$K_{\mathrm{Bessel}}(x, y) = \frac{J_\alpha\left(\sqrt{x}\right) \sqrt{y} J'_\alpha\left(\sqrt{y}\right) - J'_\alpha\left(\sqrt{x}\right) \sqrt{x} J_\alpha\left(\sqrt{y}\right)}{2(x - y)} \qquad x, y \in \mathbb{R}_+, \ \alpha > -1. \tag{49}$$

Note 19. *The function $J_\alpha(z)$ is the Bessel function (of first kind). It satisfies the second-order ODE*

$$z^2 y'' + z y' + (z^2 - \alpha^2) y = 0, \tag{50}$$

such that $\lim_{z \to 0} y(z) < \infty$ for integer or positive α. It admits a series expansion at $x = 0$ of the form

$$J_\alpha(z) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n! \Gamma(n + \alpha + 1)} \left(\frac{z}{2}\right)^{2n + \alpha} \tag{51}$$

(alternatively, it admits a representation in terms of a contour integral).

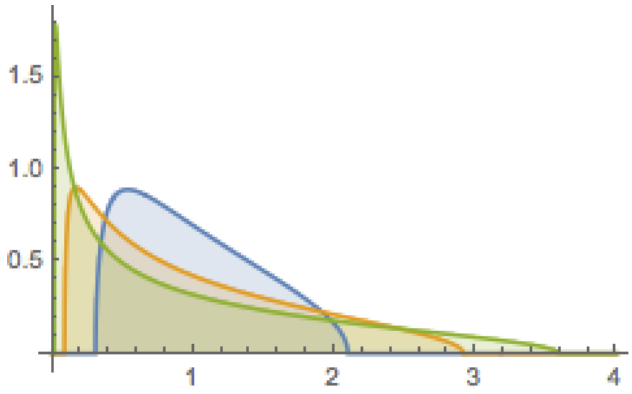


Figure 2: Marchenko-Pastur distribution for different values of κ (picture taken from WolframAlpha website).

5.2 (Gaussian) β-ensembles.

$$\mathrm{d}\mu(x) = \prod_{i<j} |x_i - x_j|^\beta \, e^{-\frac{\beta}{4}\sum_i x_i^2} \mathrm{d}x_i \dots \mathrm{d}x_n \tag{52}$$

As seen before, for $\beta = 1$, the matrix ensemble is called Gaussian Orthogonal Ensemble and it is the ensemble of real symmetric matrices (its distribution is invariant under orthogonal conjugation). For $\beta = 4$, the ensemble is given by quaternionic Hermitian matrices (its distribution is invariant under conjugation by the symplectic group) and it is called Gaussian Symplectic Ensembles.

For general $\beta > 1$, it is possible to realize this distribution as the distribution of eigenvalues of certain random tri-diagonal matrices with independent entries (Dumitriu, Edelman [4]).

The adjective "Gaussian" refers to the fact that we're still considering a quadratic potential $V(x) = x^2$ in the definition of the probability measure.

5.3 Multi-matrix models and external field.

The matrix models which we have considered so far could be called one-matrix models, as the corresponding integrals involved only one matrix. A natural generalization is to consider integrals over multiple matrices, and the corresponding multi-matrix models. For example, a two-matrix model can be defined from the ensemble

$$\mathcal{E} = \mathfrak{M} \times \mathfrak{M},$$

where $\mathfrak{M}$ is the set of all $n \times n$ Hermitian matrices, with measure (for example)

$$\mathrm{d}\mu(M_1, M_2) = e^{-\operatorname{Tr}\left[V_1(M_1) + V_2(M_2) - M_1 M_2\right]} \mathrm{d}M_1 \mathrm{d}M_2 \tag{53}$$

where V_1 and V_2 are two potentials. This can be generalized to the matrix chain on $\otimes^k \mathfrak{M}$ with or without the so-called "external field", a deterministic fixed matrix which breaks the invariance under conjugation of the original model: for example,

$$\mathrm{d}\mu(M) = e^{\operatorname{Tr}(M^2) - AM} \mathrm{d}M. \tag{54}$$

6 Bonus: applications

We conclude these notes with a few applications of random matrices to Statistics and Machine Learning.

6.1 Principal Component Analysis (PCA) [6]

Consider $X \in \mathbb{R}^p$ a random vector with covariance matrix

$$\Sigma = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T\right].$$

The goal of Principal Component Analysis is to learn a low dimensional representation of the vector X such that the residual variance is minimized. This is solved by finding a sequence of orthonormal eigenvectors $\{v_k\}_{k=1,\dots,p}$ of Σ.

In practice we observe n i.i.d. realizations $X^{(i)}$, $i = 1, \dots, n$ and Σ is unknown. We estimate the v_k's by their empirical counterparts $\hat{v}_k$ defined as as a sequence of orthonormal eigenvectors of the empirical covariance matrix

$$S = \frac{1}{n}\sum_{i=1}^{n}\left(X^{(i)} - \bar{X}\right)\left(X^{(i)} - \bar{X}\right)^T,$$

with eigenvalues $\{\hat{\lambda}_k\}$; here $\bar{X} \in \mathbb{R}^p$ is the mean of the $X^{(i)}$'s over each entry/feature. Then, S can be viewed as a random matrix.

In general, one can test the existence of a one-dimensional signal (i.e. correlation between the features/entries of X) over a Gaussian white noise. In the case were there is no correlation, we can reasonably argue that the n i.i.d. realizations of the random vector $X \in \mathbb{R}^p$ are sampled from a multivariate distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and the corresponding empirical covariance matrix nS follows a Wishart distribution $W_p(n-1, \Sigma)$.

We test the null hypothesis $H_0 : \{\Sigma = I_p\}$ (no correlations) against the alternative $H_1 : \{\Sigma = \lambda\theta\theta^T + I_p\}$, where $\theta \in \mathbb{R}^p$. Denote by $\hat{\lambda}$ the largest eigenvalue of S. Under H_0 the asymptotic distribution of $\hat{\lambda}$ is given by the Tracy-Widom law TW_1. An asymptotic test of level $1-\epsilon$, $0 < \epsilon < 1$, is to accept H_0 if $\hat{\lambda}$ is smaller or equal to the $(1-\epsilon)$-quantile of TW_1, and to accept it otherwise.

It is worth mentioning the following fact: consider i.i.d. data $X_1, \dots, X_n \in \mathbb{R}^p$ with $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_1 X_1^T] = \Sigma$. Then, the (strong) law of large numbers tells us that the sample covariance matrix S converges almost surely in the limit as $n \to \infty$ to the true covariance matrix Σ:

$$S \overset{\text{a.s.}}{\to} \Sigma, \qquad \text{equivalently } \|S - \Sigma\| \overset{\text{a.s.}}{\to} 0.$$

However, the law is not true anymore when considering the limit as both $n, p \to \infty$, while keeping their ratio constant ($p/n \to c \in \mathbb{R}$): in this case $\|S - \Sigma\| \nrightarrow 0$.

For example, consider $X_1, \dots, X_n \in \mathbb{R}^p$ i.i.d., $X_1 \sim \mathcal{N}(0, I_p)$, and $p = p(n)$ such that $p/n \to c > 1$. Then, we have joint pointwise convergence

$$\max_{i,j=1,\dots,p} |[S - I_p]_{ij}| = \max_{i,j=1,\dots,p} \left|\frac{1}{n} X_{j,\cdot} X_{i,\cdot}^T - \delta_{ij}\right| \to 0$$

but the eigenvalues do not match:

$$0 = \lambda_1(S) = \ldots = \lambda_{p-n}(S) \leq \lambda_{p-n+1}(S) \leq \ldots \leq \lambda_p(S)$$

(indeed, the data matrix $[X_1|X_2|\ldots|X_n] \in \mathrm{Mat}_{p,n}(\mathbb{R})$ a rectangular matrix with more rows than columns, thus implying that its Gram matrix $S = XX^T$ is noninvertible) while

$$\lambda_1(I_p) = \lambda_2(I_p) = \ldots = \lambda_p(I_p) \equiv 1.$$

6.2 Geometry of NN Loss Surfaces [7]

Consider a NN with one hidden layer, without biases, and with ReLU activation function:

$$\hat{y}_{i\mu} = \sum_{k=1}^{n_1} W_{ik}^{(2)} \left[z_{k\mu}^{(1)} \right]_+, \qquad z_{k\mu}^{(1)} = \sum_{\ell=1}^{n_0} W_{k\ell}^{(1)} x_{\ell\mu} \tag{55}$$

and consider the least-square-error loss function

$$\mathcal{L} = \frac{1}{2m} \sum_{i,\mu=1}^{n_2,m} (\hat{y}_{i\mu} - y_{i\mu})^2 \tag{56}$$

(with m the number of input samples). We are interested in the regime $n = n_1 = n_2 = n_0$ and $n, m \gg 1$, but constant ratio of parameter to to data points $\phi = \frac{2n^2}{nm} = \frac{2n}{m} \in \mathbb{R}$.

The Hessian of $\mathcal{L}$ $H[\mathcal{L}]_{\alpha,\beta} = \frac{\partial^2 \mathcal{L}}{\partial\alpha\partial\beta}$ can be decomposed into the sum of two matrices

$$H[\mathcal{L}] = H_0 + H_1 \tag{57}$$

where

$$[H_0]_{\alpha,\beta} = \frac{1}{m} \left[JJ^T \right]_{\alpha,\beta}$$
$$[H_1]_{\alpha,\beta} = \frac{1}{m} \sum_{i,\mu=1}^{n_2,m} (\hat{y}_{i\mu} - y_{i\mu}) \cdot \frac{\partial \hat{y}_{i\mu}}{\partial\alpha\partial\beta}$$

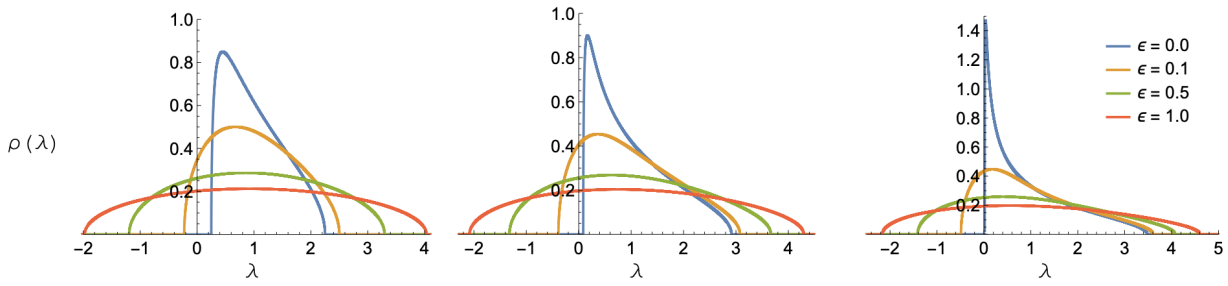
Under some mild (justifiable) assumption, we can assume that H_0 and H_1 are respectively a real Wishart matrix and a real Wigner matrix (at least locally near the critical points). Additionally we assume that H_0 and H_1 are **freely independent**.

Therefore, the spectrum $\rho(\lambda; \mathcal{L}, \phi)$ of the Hessian can be easily computed with the help of the Stieltjes (and $\mathcal{R}$) transform.

Of particular interest is the *index* α of the Hessian (i.e. the fraction of negative eigenvalues),

$$\alpha(\mathcal{L}; \phi) = \int_{-\infty}^{0} \rho(\lambda; \mathcal{L}, \phi) \ \mathrm{d}\lambda$$

because it measures the number of "descent directions", which is crucial in optimization: previous work showed that critical points with many descent directions have large loss value.



(a) $\phi = 1/4$ (b) $\phi = 1/2$ (c) $\phi = 3/4$

Figure 3: Spectral distribution of the Wishart+Wigner approximation of the Hessian for different values of ϕ. As the value of the loss function increases, the spectrum becomes more semi-circular and negative eigenvalues emerge. (From [7])

It is then possible to derive a prediction (also supported by numerical simulations) that for critical points of small index (i.e. possible good candidates for the minimizer of the cost function) we have

$$\alpha\left(\mathcal{L};\phi\right) \approx \alpha_0\left(\phi\right)\left|\frac{\mathcal{L}-\mathcal{L}_c}{\mathcal{L}_c}\right|^{\frac{3}{2}} \tag{58}$$

where $\mathcal{L}_c$ is the value of the loss function below which all critical points are minimizers: in particular,

$$\mathcal{L}_c = \mathcal{L}_c\left(\phi\right) = \frac{1}{16}\left(1 - 20\phi - 8\phi^2 + (1+8\phi)^{\frac{3}{2}}\right)$$

(note that $\mathcal{L}_c \to 0$ as $\frac{n}{m} \to 1$).

Note 20 (Food for thought). *Note that the $\frac{3}{2}$-exponent appearing in (58) appears also in a refined version of the model and, more remarkably, in the context of field theory of Gaussian random functions. Could there be a KP-universality hidden here?*

6.3 Nonlinear RMT for Deep Learning [8]

Consider a NN of the form

$$Y = f(WX) \tag{59}$$

where f is the nonlinear activation function, W is a Gaussian random weight matrix, X is a Gaussian random data matrix (both matrices have Gaussian entries with mean zero and prescribed variance). We are assuming that the dimension of both the number of parameters n and the number of samples m grows to infinity at the same rate ($n_0, n_1, m \to +\infty$, $\frac{n_0}{m} = \phi \in \mathbb{R}$, $\frac{n_1}{n_0} = \psi \in \mathbb{R}$). The focus is on the spectral composition ρ_M of the Gram matrix

$$M := \frac{1}{m} Y^T Y.$$

As before, Stieltjes transform (Appendix A) comes into help:

$$G(z) := \frac{1}{n_1}\mathbb{E}\left[\mathrm{Tr}\left[(M - z\mathbf{1}_{n_1})^{-1}\right]\right],$$

where the expectation is taken w.r.t. W and X. After we take the large size limit ($n_0, n_1, m \to +\infty$), we'll get back to the (limiting) spectral density via the Inverse Stieltjes Transform. As extra ingredients, before computing the IS transform, we will first perform an asymptotic expansion of $G(z)$ as a power series

$$G(z) = \sum_{k=0}^{\infty} \frac{m_k}{z^{k+1}}, \qquad z \to \infty$$

where the coefficients m_k are the moments of the distribution

$$m_k = \int t^k \rho_M(t)\mathrm{d}t = \frac{1}{n_1}\mathbb{E}\left[\mathrm{Tr}[M^k]\right]$$

(this is known as the "moment methods").

The main result is the following: in the large size limit, the Stieltjes transform of the density ρ_M Gram matrix satisfies a simple quartic polynomial expression depending exclusively on ϕ, ψ and two quantities η, ζ that only depends on the nonlinearity f (see [8, Formula (10)]):

$$G(z) = \frac{\psi}{z} P\left((z\psi)^{-1}\right) + \frac{1-\psi}{z} \tag{60}$$

where P is the solution of a given (implicit) equation.

Consider two special cases:

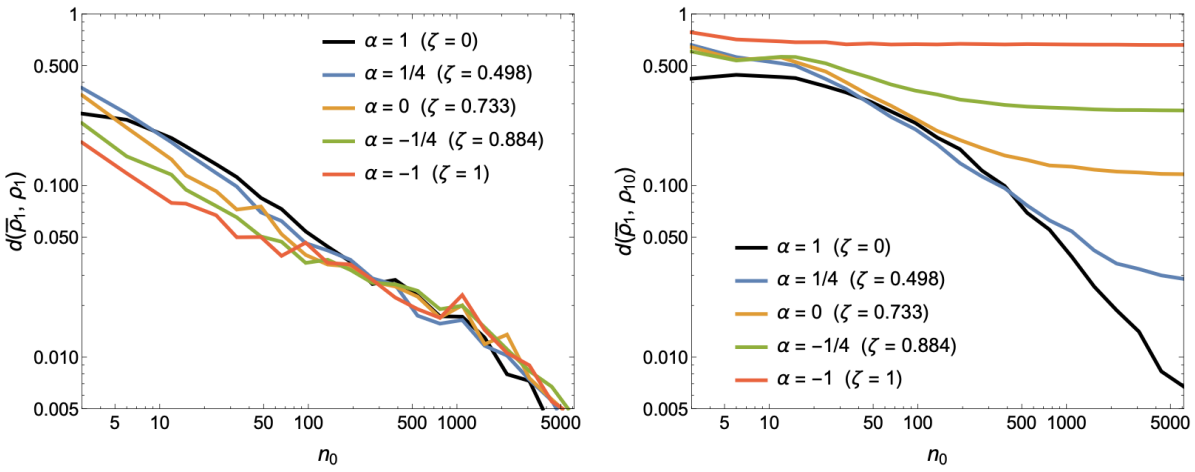
1. $\zeta = \eta$ (theoretically interesting, but possibly with limited span of application): $\eta = \zeta$ if and only if f is linear. In this case, $M = (WX)(WX)^T$ product of Gaussian random matrices. The distribution of its singular values has been widely studied and derived in [1] and [2], although these results were not mentioned in this paper.

2. $\zeta = 0$: in this case the equation describing $G(z)$ coincides with the Stieltjes transform of the Marchenko-Pastur distribution with parameter $\kappa = \phi/\psi$ and in particular if $\psi = 1$, it additionally coincides with the limiting distribution of XX^T, implying that YY^T and XX^T have the same limiting spectral distribution.

   The interesting implication is that nonlinear functions f for which $\zeta = 0$ are isospectral transformations.

   If we consider now a deep forward NN with ℓ-th layer $Y^\ell = f(W^\ell Y^{\ell-1})$, $Y^0 = X$, we may wonder whether (by smartly choosing the nonlinearity f so that $\zeta = 0$) the distribution of the eigenvalues of the ℓ-th data covariance matrix $Y^\ell (Y^\ell)^T$ is (approximately) the same as the distribution of the eigenvalues of the input Gram matrix XX^T.

   Indeed, having similar distribution, indicates that the input signals are not distorted or stretched as they propagate through the network (highly skewed distribution shows poor conditioning to the point that learning may not happen). Batch normalization techniques arose to address this same issue. See Figure 4 for some numerical results.



(a) $L = 1$ (b) $L = 10$

Figure 4: Statistical distance (i.e. the L^1 norm of the difference) between the first-layer (a) and the tenth-layer (b) empirical eigenvalue distributions of the data covariance matrices and the theoretical prediction given by equation (60) for the first-layer limiting distribution $\bar{\rho}_1$, as the network width increases. The different curves correspond to different piecewise linear activation functions ($\alpha = -1$ is linear, $\alpha = 0$ is ReLU, $\alpha = 1$ is absolute value). We can see that in (a) for all α we have good convergence to $\bar{\rho}_1$, while in (b) the convergence only happens when $\alpha = 1$ (i.e., $\zeta = 0$). (From [8])

An additional result presented in [8] is the following: consider the ridge-regularized least-square loss function for a single-layer network

$$L(W_2) := \frac{1}{2n_2 m} \left\| \mathcal{Y} - W_2^T Y \right\|_F + \gamma \left\| W_2 \right\|_F^2, \quad Y = f(W_1 X)$$

where W_1 is a matrix of random weights and W_2 is the matrix of parameters to be learned (a setting similar to the Random Kitchen Sinks [9]).

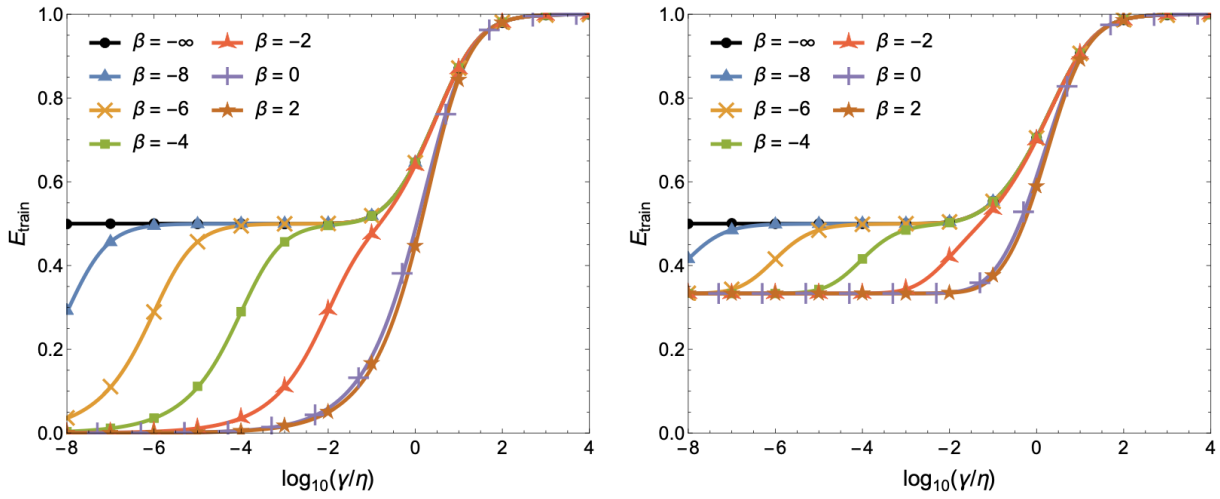
It can be shown that the training loss of this problem is related to $-\gamma^2 G'(-\gamma)$: for fixed value of γ, the training loss is lower (i.e. the memorization capacity is higher) if $\frac{\eta}{\zeta}$ is higher (ideally going to infinity if $\zeta \to 0$?), a condition satisfied by a large class of functions, for example if f is "close" to be an even function. See Figure 5.

A A few facts about the Stieltjes transform

For the sake of simplicity, consider a measure μ that is absolutely continuous with respect to the Lebesgue measure (i.e. it admits a density function $\rho(t) \in L^1_{\text{loc}}(\mathbb{R})$ such that $\mathrm{d}\mu(t) = \rho(t)\mathrm{d}t$)

Definition 21. For a (real) probability measure μ with support $\operatorname{supp}\mu$, its **Stieltjes transform** m_μ is defined for $z \in \mathbb{C} \setminus \operatorname{supp}\mu$ as

$$m_\mu(z) = \int \frac{1}{t - z} \mathrm{d}\mu(t) \tag{61}$$



(a) $\phi = \frac{1}{2}, \psi = \frac{1}{2}$ (b) $\phi = \frac{1}{2}, \psi = \frac{3}{4}$

Figure 5: Memorization performance of random feature networks versus γ. Theoretical curves are solid lines and numerical results are the points. $\beta = \log_{10}(\eta/\zeta - 1)$ distinguishes the class of nonlinearities. In (a) there are more random features than data points, allowing for perfect memorization, unless the function is linear ($\beta = -\infty$). In (b) there are fewer features than data points. For fixed γ, curves with larger values of β (i.e. smaller value of ζ) have lower training loss. (From [8])

Focussing on the special case of (symmetric) matrices, we have the following result:

Proposition 22. *Given a symmetric matrix $M \in \mathrm{Mat}_{p,p}(\mathbb{R})$ with empirical spectral distribution $\mu = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i(M)}$, its Stieltjes transform is*

$$m_\mu(z) = \frac{1}{p} \operatorname{Tr}\left[(M - z\mathbf{1}_p)^{-1}\right]. \tag{62}$$

Proof.

$$\begin{aligned} m_\mu(z) = \int \frac{1}{t-z} \mathrm{d}\mu(t) = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{\lambda_i(M) - z} &= \frac{1}{p} \operatorname{Tr}\left[(\operatorname{diag}\{\lambda_i(M)\} - z\mathbf{1}_p)^{-1}\right] \\ &= \frac{1}{p} \operatorname{Tr}\left[(M - z\mathbf{1}_p)^{-1}\right] \end{aligned} \tag{63}$$

□

The measure itself can be recovered via the **Inverse Stieltjes Transform**:

Proposition 23. *Given a measure μ with density function $\rho(t)$ and its Stieltjes transform m_μ, then the following holds:*

$$\rho(t) = \lim_{\epsilon \searrow 0} \frac{1}{\pi} \Im\left[m_\mu(t + i\epsilon)\right] \tag{64}$$

The idea of the proof of Theorems 10 and 18 is the following: compute the Stieltjes transform of the empirical spectral distribution, calculate its limit as $n \to \infty$ (with $\frac{p}{n} \to \kappa$) and finally "get back" to the measure on the real line by calculating the Inverse Stieltjes transform.

References

[1] G. Akemann, J. R. Ipsen, and M. Kieburg. Products of rectangular random matrices: Singular values and progressive scattering. *Phys. Rev. E*, 88(5):52–118, 2013.

[2] G. Akemann, M. Kieburg, and L. Wei. Singular value correlation functions for products of Wishart random matrices. *J. Phys. A: Math. Theor.*, 46(27), 2013.

[3] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2009.

[4] I. Dumitriu and A. Edelman. Matrix models for β-ensembles. *J. Math. Phys.*, 43:5830–5847, 2008.

[5] M. L. Mehta. *Random Matrices*. Elsevier/Academic Press, Amsterdam, third edition, 2004.

[6] D. Paul and A. Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 2013.

[7] J. Pennington and Y. Bahri. Geometry of neural network loss surfaces via Random Matrix Theory. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, International Convention Centre, Sydney, Australia, 2017. PMLR.

[8] J. Pennington and P. Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems 30*, pages 2637–2646. Curran Associates, Inc., 2017.

[9] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *NIPS*, 2008.

[10] M.E.A. Seddik, C. Louart, M. Tamaazousti, and R. Couillet. Random Matrix Theory proves that deep learning representations of GAN-data behave as Gaussian mixtures. *arXiv:2001.08370*, 2020.

[11] C. Tracy and H. Widom. Level spacing distributions and the Airy kernel. *Comm. Math. Phys.*, 159(1):151–174, 1994.

[12] E. P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Ann. Math.*, 62:548–564, 1955.