Why do we need more Random Matrix Theory in Deep Learning

Manuela Girotti

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 $\mathrm{d}\mu(\boldsymbol{M})$ for $\boldsymbol{M} \in \mathcal{E}$ (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 ($\boldsymbol{M} = \boldsymbol{M}^\dagger$) of size $n \times n$: $\mathfrak{M} = \{ \boldsymbol{M} \in \mathrm{Mat}_n(\mathbb{C}) \mid M_{ij} = M_{ji}^* \}$ (**Unitary ensemble**)
- Symmetric matrices ($\boldsymbol{M} = \boldsymbol{M}^\top$) of size $n \times n$: $\mathfrak{M} = \{ \boldsymbol{M} \in \mathrm{Mat}_n(\mathbb{R}) \mid M_{ij} = M_{ji} \}$ (**Orthogonal ensemble**)
- Symplectic matrices: $\boldsymbol{M}^\top \boldsymbol{J} = \boldsymbol{J}\boldsymbol{M}^\top$, with $\boldsymbol{J} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \otimes \boldsymbol{I}_n$ 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 $\mathrm{d}\boldsymbol{M}$.

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

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

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

One of the main objectives in Random Matrix Theory 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. In particular, the interest lies in the study of the properties of these 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 appear in different fields of Mathematics and Physics, providing unexpected links between for example Probability, Number Theory and Integrable Systems. Additionally, in recent years there has been an increasing interest from the Machine Learning community into theoretical results in random matrices (Cheng, Singer [6]; Pennington, Worah [19]; Louart *et al.* [13], to name a few) as a powerful tool to rigorously analyze generalization performances of neural networks and other models.

Among the vast literature on Random Matrix Theory, we can mention the following classical books by Mehta [16], Anderson, Guionnet and Zeitouni [2] and Bai and Silverstein [5]. This paper relies on these main sources (and others that will be mentioned along the way) for the theoretical part, while for the applicative part we will refer to some recent papers in the domain of ML.

1.1 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 ([24]). 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:

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

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. 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 point process with a specific structure, called determinantal (see, for example [22]), and its statistical properties can be efficiently studied in this framework.

1.2 A statistical and machine learning perspective

Consider $\boldsymbol{x} \in \mathbb{R}^p$ a random vector according to a prescribed probability distribution with mean $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$ and covariance matrix as

$$\mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^\top\right] = \boldsymbol{\Sigma}.$$

In practical experiments, we usually observe n i.i.d.[1] realizations $\boldsymbol{x}^{(i)} \sim \boldsymbol{x}$, $i = 1, \dots, n$ (a **sample set**), with which we can estimate the unknown $\boldsymbol{\Sigma}$ by calculating the **sample covariance matrix** (also called Gram matrix or design matrix)

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{x}^{(i)} - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}^{(i)} - \bar{\boldsymbol{x}}\right)^\top,$$

with $\bar{\boldsymbol{x}} \in \mathbb{R}^p$ is the sample mean of the $\boldsymbol{x}^{(i)}$'s over each entry/feature. Being the $\boldsymbol{x}^{(i)}$'s random, the matrix $\boldsymbol{S}$ can be rightfully viewed as a random matrix itself.

Let us consider a simple scenario where $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_p)$ and we sample n i.i.d. realizations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^p$ (collected into a $p \times n$ matrix $\boldsymbol{X}$). Then, the strong law of large numbers tells us

[1] The assumption that the observed data is a collection of i.i.d. sample points is not true in general and cases with interdependency and non-identical distribution of the samples are increasingly more common in modern applications of Machine Learning (e.g. transfer learning, domain adaptation/generalization, etc.). Indeed, training and test data may come from different distributions (dataset bias or domain shift) or the data may exhibit temporal or spatial correlations.

that the sample covariance matrix $\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$ converges almost surely in the limit as $n \to \infty$ to the true covariance matrix $\boldsymbol{I}_p$:

$$\boldsymbol{S} \xrightarrow{\text{a.s.}} \boldsymbol{I}_p, \qquad \text{equivalently } \|\boldsymbol{S} - \boldsymbol{I}_p\| \xrightarrow{\text{a.s.}} 0 \quad \text{(in spectral norm)}.$$

However, if we consider the limit as both $n, p \to \infty$, while keeping their ratio constant ($p/n \to \kappa \in \mathbb{R}$), the law of large numbers does not hold true anymore.

As an example, consider the case where $p = \kappa n$ for some $\kappa > 1$. It is easy to verify the joint pointwise convergence

$$\max_{i,j=1,\dots,p} \left| [\boldsymbol{S} - \boldsymbol{I}_p]_{ij} \right| = \max_{i,j=1,\dots,p} \left| \frac{1}{n} x_{j,\cdot} x_{i,\cdot}^\top - \delta_{ij} \right| \to 0, \quad \text{as } n \to \infty,$$

but the eigenvalues do not match: indeed, since $p > n$, the Gram matrix is singular

$$0 = \lambda_1(\boldsymbol{S}) = \dots = \lambda_{p-n}(\boldsymbol{S}) \le \lambda_{p-n+1}(\boldsymbol{S}) \le \dots \le \lambda_p(\boldsymbol{S}),$$

while $\lambda_1(\boldsymbol{I}_p) = \lambda_2(\boldsymbol{I}_p) = \dots = \lambda_p(\boldsymbol{I}_p) \equiv 1$. Therefore, $\boldsymbol{S}$ does not converge in spectral norm to the identity matrix.

It is important to stress that by letting also p going to infinity, the dimension of the matrix itself changes, and consequently the number of eigenvalues grows to infinity. It has been known for a long time that when both dimensions $p, n \to +\infty$, while still keeping their ratio finite $p/n \to \kappa \in \mathbb{R}_+$, the eigenvalues of $\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$ (the normalization $\frac{1}{n}$ is crucial) are distributed according to the following law (Marchenko-Pastur, '67 [14]):

$$\rho_{\text{MP};\kappa}(t) = \max\left\{0, 1 - \kappa^{-1}\right\} \delta_0(t) + \frac{1}{2\pi\kappa} \frac{\sqrt{(a_+ - t)(t - a_-)}}{t} \chi_{[a_-, a_+]}(t)$$

where $\chi_B(t)$ is the characteristic function over the set B that takes values 1 if $t \in B$ and zero otherwise and the endpoints $a_\pm$ are equal to $a_\pm = (1 \pm \sqrt{\kappa})^2$.

The Marchenko-Pastur law will be discussed in more details in Section 2.4.2.

to expand a bit more? add some bla bla to draw some conclusion...

Weight matrices of Deep Neural Network Maybe add a short summary of the transition form Wishart to heavy-tailed for the weight matrices in NN by Martin and Mahoney, "Implicit self-regularization in deep neural networks: evidence of Random Matrix Theory and implications for learning" [15].

to summarize

Gradual shifting form Marchenko Pastur distribution of the weights at initialization (all $\sim \mathcal{N}(0, 1)$) into heavy-tailed distribution at the end of training (5 phases).

2 Random matrix models

bla bla INTRO!!!

2.1 Unitary matrices

Consider the space of $n \times n$ complex Hermitian matrices

$$\mathfrak{H}_n = \{ \boldsymbol{M} \in \mathrm{Mat}_{n,n}(\mathbb{C}) \mid \boldsymbol{M} = \boldsymbol{M}^\dagger \}. \tag{1}$$

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

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

We also recall the following properties of Hermitian matrices:

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

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

and its eigenvalues are real:

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

Remark 3. *The diagonalization is not unique even if $\boldsymbol{X}$ 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

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

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

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

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

where u_{ij} are the parameters that parametrize the unitary group. Under such transformation, the Lebesgue measure reads (thanks to the Weyl integration formula[2])

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

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

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

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

[2]The Weyl formula is a fundamental result in the Theory of Lie Groups. We refer to [1] for all the details.

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

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

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 $\Delta(\boldsymbol{X})$ is $\beta = 1, 2, 4$ (Orthogonal, Unitary, Symplectic ensembles), they are also universally known as the $\beta = 1, 2, 4$ ensembles.

2.1.1 Eigenvalues distribution

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

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

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

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

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

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

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

From now on, we will assume that $\widetilde{F}$ is of the form:

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

for some function $V(x)$ called *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*).

The above probability distribution has a well-known connection to the theory of Orthogonal Polynomials. We report a few classical results in Appendix B.1 and we refer to the book by Deift [9] for a more comprehensive exposition on the topic.

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

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

Since

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

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

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

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

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

Remark 6. *As discussed in Remark 5, we can obtain similar eigenvalues distributions for the Orthogonal ($\beta = 1$) and Symplectic ($\beta = 4$) Ensembles:*

$$\mathrm{d}\mu(x) = \prod_{i<j}|x_i - x_j|^{\beta}\prod_{j=1}^{n} e^{-V_\beta(x_j)}\mathrm{d}x_i\ldots\mathrm{d}x_n. \tag{18}$$

In particular, if the potential V is a quadratic $V(x)\sim x^2$, the corresponding distribution gains the adjective "Gaussian". In this case, it is possible to show that for general $\beta > 1$ the distribution

$$\mathrm{d}\mu(x) = \prod_{i<j}|x_i - x_j|^{\beta}\, e^{-\frac{\beta}{4}\sum_i x_i^2}\mathrm{d}x_i\ldots\mathrm{d}x_n \tag{19}$$

is the eigenvalue distribution of certain random tri-diagonal matrices with independent entries (Dumitriu, Edelman [10]).

2.2 Wishart matrices

taken from Couillet's slides (intro part) + complemented with Guionnet, Zeitouni book [2].

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

The corresponding eigenvalue distribution has the following expression

$$\mathrm{d}\mu(\boldsymbol{x}) = \prod_{i<j} |x_i - x_j| \prod_i x_i^{\alpha} e^{\frac{1}{2}\sum_i x_i} \mathrm{d}x_1 \dots \mathrm{d}x_n. \tag{20}$$

add more stuff, check Couillet's book

2.3 Spiked Models

taken from Couillet's slides (intro) + bibliography therein...

I saw some spiked models applications in generalization/deep learning – to add!

2.4 Asymptotics and universality: macroscopic behaviour

A typical question one may ask 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.

In this Section, we are interested in the global (macroscopic) distribution of the properly rescaled eigenvalues, when the dimension of the matrix grows. We refer to Appendix B.2 for a discussion about the infinitesimal fluctuations of the eigenvalues within their spectrum or at their boundary.

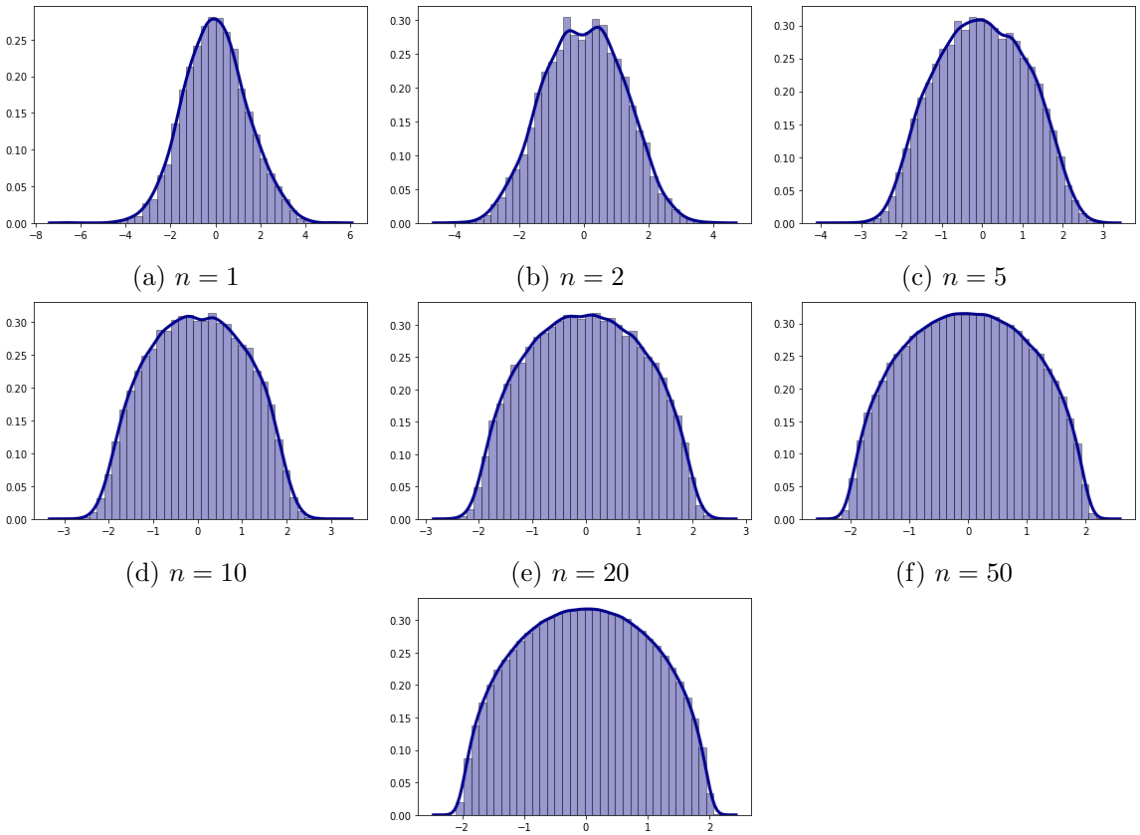
Consider a matrix ensemble and denote the (ordered) eigenvalues by $x_1 \leq x_2 \leq \ldots \leq x_n$. The empirical spectral distribution of the eigenvalues is defined by

$$\mathrm{d}\mu_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{21}$$

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 for the histogram of the eigenvalues of the Gaussian Orthogonal Ensemble.

In many cases of interest the eigenvalue density has a finite limit as $n \to +\infty$, called the **equilibrium density** $\rho(x)$.

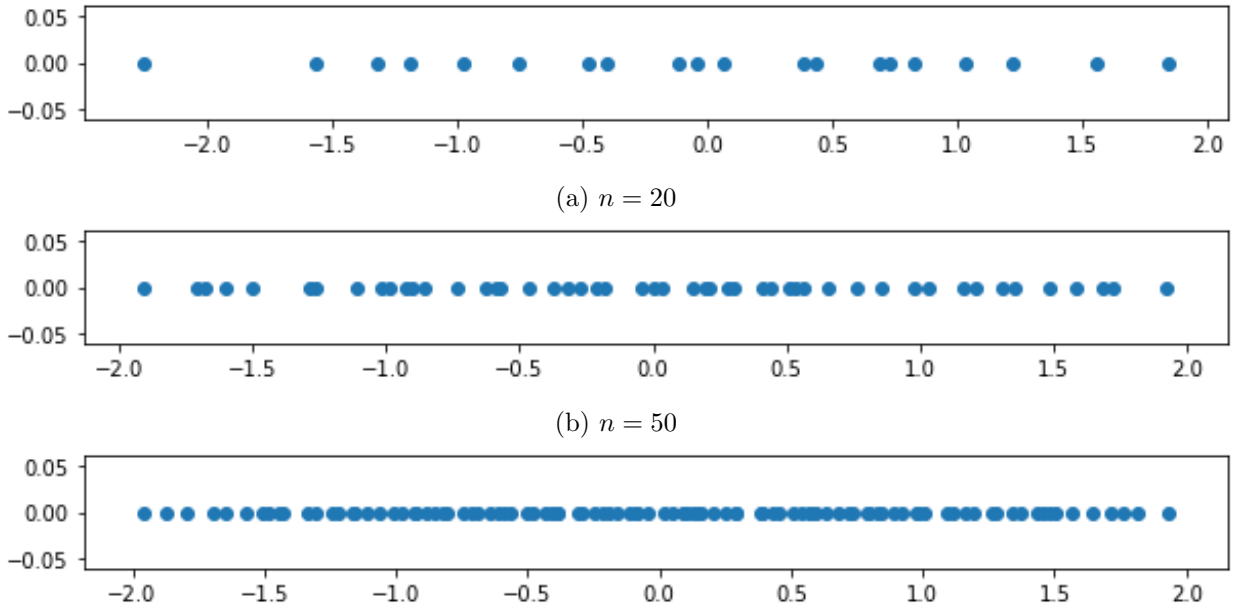


(a) $n = 1$ (b) $n = 2$ (c) $n = 5$

(d) $n = 10$ (e) $n = 20$ (f) $n = 50$

(g) $n = 100$

Figure 1: Histograms and density plots of the eigenvalues of GOE matrices as the size of the matrix n increases. Each graph has been realized with 4000 numerical simulations.



(a) $n = 20$

(b) $n = 50$

(c) $n = 100$

Figure 2: Eigenvalues of one realization of a GOE matrix of size $n \times n$ (properly rescaled)..

2.4.1 Wigner's semicircle law

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

$$\widetilde{\boldsymbol{M}} = \frac{1}{\sqrt{n}} \boldsymbol{M}, \tag{22}$$

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

The limit semicircle distribution appears as a limit distribution for a vast class of random matrices, not only GUE, that satisfies the following minimal requirements: $\boldsymbol{X} \in \mathrm{Mat}_{n,n}(\mathbb{C})$ (or $\boldsymbol{X} \in \mathrm{Mat}_{n,n}(\mathbb{R})$) is a Wigner matrix if it is Hermitian (or symmetric) with entries

$$X_{ij} = X_{ji}^{\dagger} = \begin{cases} \frac{1}{\sqrt{n}} Z_{ij} & i < j \\ \frac{1}{\sqrt{n}} Y_i & i = j \end{cases} \qquad \forall\, i, j = 1, \dots, n, \tag{23}$$

where $\{Z_{ij}\}_{i<j}$ and $\{Y_i\}$ are a collection of i.i.d. complex (or real) random variables with mean zero, finite moments and such that $\mathbb{E}[Z_{ij}^2] = 0$, $\mathbb{E}[|Z_{ij}|^2] = 1$.

Note 7. *The GUE ensemble is a special case of complex Wigner matrices where $\{Z_{ij}\}$ and $\{Y_i\}$ are Gaussians.*

Theorem 8 (Wigner, '55 [24]). *The empirical spectral density of a (real or complex) Wigner matrix converges weakly in probability as $n \to +\infty$ to the following deterministic probability density*

$$\rho(t) = \frac{1}{2\pi}\sqrt{4-t^2}\chi_{[-2,2]}(t) \tag{24}$$

where $\chi_{[[-2,2]]}(t)$ is the characteristic function of the interval $[-2,2]$. More precisely, $\forall\ f \in C^0(\mathbb{R})$ bounded, $\forall\ \epsilon > 0$

$$\lim_{n\to+\infty} \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int f(t)\rho(t)\ \mathrm{d}t\right| \geq \epsilon\right] = 0. \tag{25}$$

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 Section A) and other more recent methods. □

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

2.4.2 Marchenko-Pastur Law

Consider again a sample covariance matrix

$$\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top,$$

with random matrix $\boldsymbol{X} \in \mathrm{Mat}_{p,n}(\mathbb{C})$ (or $\boldsymbol{X} \in \mathrm{Mat}_{p,n}(\mathbb{R})$) having i.i.d., zero mean and unit variance entries. $\boldsymbol{S}$ belongs to a rescaled version of the Wishart ensemble.

As we already discussed in the Introduction (Section 1.2), in the double limit as $p, n \to +\infty$, the Law of Large Numbers does not apply. On the other hand, if we plot a histogram of the eigenvalues of $\boldsymbol{S}$, we can still notice that a limiting shape arises: see Figure 3.

The rigorous claim is the following

Theorem 9 (Marchenko, Pastur [14]). *Let $\boldsymbol{X} \in \mathrm{Mat}_{p,n}(\mathbb{R})$ be a random matrix whose entries are i.i.d. with zero mean and unit variance. As $p, n \to +\infty$ with $p/n \to \kappa \in \mathbb{R}_+$, the empirical spectral distribution of the Gram matrix $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top \in \mathrm{Mat}_{p,p}(\mathbb{R})$ satisfies*

$$\mu \to \mu_{\mathrm{MP};\kappa} \quad \textit{almost surely, weakly,} \tag{26}$$

where $\mu_{\mathrm{MP};\kappa}$ is a compactly supported probability measure with density function

$$\rho_{\mathrm{MP};\kappa}(t) = \max\left\{0, 1-\kappa^{-1}\right\}\delta_0(t) + \frac{1}{2\pi\kappa}\frac{\sqrt{(a_+-t)(t-a_-)}}{t}\chi_{[a_-,a_+]}(t) \tag{27}$$

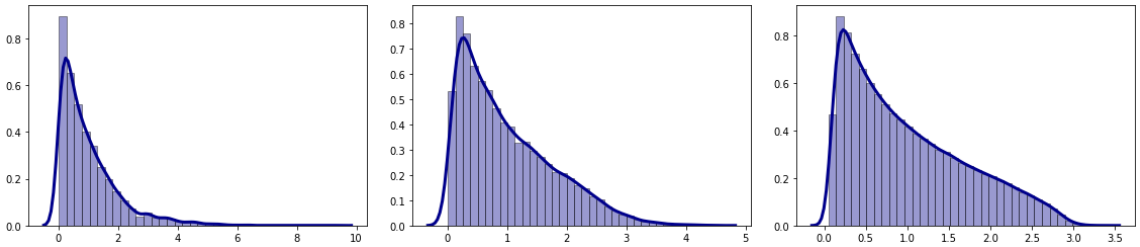
with $\chi_{[a_-,a_+]}(t)$ the characteristic function of the interval $[a_-, a_+]$ and

$$a_- = \left(1-\sqrt{\kappa}\right)^2,\ a_+ = \left(1+\sqrt{\kappa}\right)^2. \tag{28}$$

We will report in Appendix A a simple proof using the method of Stieltjes transform.

Notice that in the special case $\kappa = 1$ the distribution has a square-root singularity at $x = 0$:

$$\rho_{\mathrm{MP},1}(x) = \frac{1}{2\pi}\sqrt{\frac{4-x}{x}} \qquad x \in (0,4]. \tag{29}$$



(a) $p = 1, n = 2$ (b) $p = 5, n = 10$ (c) $p = 50, n = 100$

Figure 3: Histograms and density plots of the eigenvalues of Wishart matrices $\boldsymbol{Y} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$ as the dimension of the matrix $p, n \to +\infty$ and $\frac{p}{n} = \kappa \in \mathbb{R}_+$. In this case $\kappa = \frac{1}{2}$. Each graph has been realized with 4000 numerical simulations.

3 Applications in Machine Learning

We conclude these notes with a few notable applications of random matrices to Machine Learning (Neural Networks in particular). more bla bla INTRO and MOTIVATION!!!! OVERVIEW!!!

3.1 Training error

Consider a two-layer neural network with linear output function and no biases:

$$\hat{\boldsymbol{y}} = \boldsymbol{A}^\top f(\boldsymbol{W}\boldsymbol{x}) \in \mathbb{R}^o, \qquad \boldsymbol{x} \in \mathbb{R}^p \text{ (data point)}, \tag{30}$$

with weight matrices $\boldsymbol{W} \in \mathrm{Mat}_{N,p}(\mathbb{R})$, $\boldsymbol{A} \in \mathrm{Mat}_{N,o}(\mathbb{R})$ and activation function $f : \mathbb{R} \to \mathbb{R}$ applied entry-wise.

Given a data set $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1\dots,n}$ with $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbb{R}^p \times \mathbb{R}^o \ \forall\ i$, we want to perform a regression task using the above function (30) as predictor. We will assume that the weight matrix $\boldsymbol{W}$ is random, but fixed, and we will only train the second layer of weights by minimizing the regularized mean square error

$$\mathcal{L}(\boldsymbol{A}; \boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\|^2 + \lambda \|\boldsymbol{A}\|_F^2 = \frac{1}{n}\left\|\boldsymbol{Y} - \boldsymbol{A}^\top f(\boldsymbol{W}\boldsymbol{X})\right\|_F^2 + \lambda \|\boldsymbol{A}\|_F^2 \tag{31}$$

where $\|\boldsymbol{A}\|_F$ is the Frobenius norm of the matrix $\boldsymbol{A}$. The model was first introduced by Rahimi and Recht [21] and it is known as *random features model*.

We will be interested in the regime where the number n and size p of the samples as well as the number of neurons N grows to infinity, while still keeping finite their respective ratio:

$$\frac{p}{n} \to \gamma_1 \qquad \text{and} \qquad \frac{N}{n} \to \gamma_2. \tag{32}$$

The cost function (31) can be analytically minimized with exact solution given by

$$\boldsymbol{A}^* = \frac{1}{n}\boldsymbol{\Sigma}\left(\frac{1}{n}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma} + \lambda \boldsymbol{I}_n\right)^{-1}\boldsymbol{Y}^\top = \frac{1}{n}\boldsymbol{\Sigma}\,\boldsymbol{Q}_{\boldsymbol{\Sigma}}(-\lambda)\,\boldsymbol{Y}^\top, \tag{33}$$

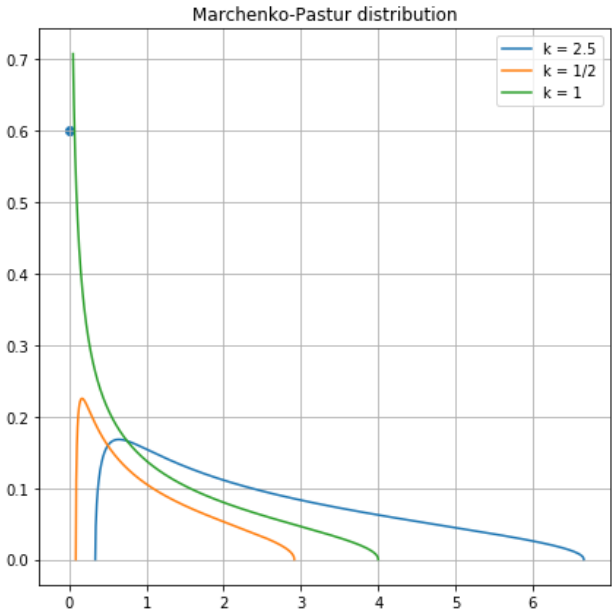


Figure 4: Marchenko-Pastur distribution for different values of κ.

with $\boldsymbol{\Sigma} := f(\boldsymbol{W}\boldsymbol{X})$ the random feature map and $\boldsymbol{Q}_{\boldsymbol{\Sigma}}(\lambda) = \left(\frac{1}{n}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma} - \lambda \boldsymbol{I}_n\right)^{-1}$ the resolvent of $\frac{1}{n}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}$.

We are interested studying the training error

$$E_{\text{train}}(\boldsymbol{X}, \boldsymbol{Y}; \lambda) = \frac{1}{n}\left\|\boldsymbol{Y} - (\boldsymbol{A}^*)^\top f(\boldsymbol{W}\boldsymbol{X})\right\|_F^2 = \frac{\lambda^2}{n}\operatorname{Tr}\left[\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{Q}_{\boldsymbol{\Sigma}}^2\right] \tag{34}$$

as a proxy for evaluating the memorization capacity of the model. Notice that here we are considering the sample data as a fixed set, however E_{train} is still a random object due to the weight matrix $\boldsymbol{W}$.

We will use a concentration of measure approach, described by Louart, Liao, Couillet, '17 [12], that extends some results already seen for the proof of the Marchenko-Pastur law. We will lay down here only a road map that will lead to the final result (Theorem **??**), while we refer to the original paper [12] for all the details and the proofs.

One of the key points will be the study of the limiting distribution of the matrix $\frac{1}{n}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}$ and consequently the study of its resolvent $\boldsymbol{Q}_{\boldsymbol{\Sigma}}$: as in Appendix A, we can rewrite $\boldsymbol{Q}_{\boldsymbol{\Sigma}}(z)$ as

$$\boldsymbol{Q}_{\boldsymbol{\Sigma}}(\lambda) = \left(\frac{1}{n}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma} - \lambda\boldsymbol{I}_n\right)^{-1} = \left(\frac{1}{n}\boldsymbol{\Sigma}_{-i}^\top\boldsymbol{\Sigma}_{-i} + \frac{1}{n}\boldsymbol{v}_i\boldsymbol{v}_i^\top - \lambda\boldsymbol{I}_n\right)^{-1} \tag{35}$$

where we split the matrix $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{v}_i^\top \\ \boldsymbol{\Sigma}_{-i} \end{bmatrix},$$

with $\boldsymbol{v}_i = f(\boldsymbol{X}^\top\boldsymbol{w}_i)$, $\boldsymbol{w}_i$ the i-th row of $\boldsymbol{W}$ $(i = 1, \dots, N)$, and $\boldsymbol{\Sigma}_{-i}$ the remaining submatrix.

Using Sherman-Morrison formula, we obtain

$$\boldsymbol{Q}_{\boldsymbol{\Sigma}}(\lambda) = \boldsymbol{Q}_{-i}(\lambda) - \frac{\boldsymbol{Q}_{-i}(\lambda)\frac{1}{n}\boldsymbol{v}_i\boldsymbol{v}_i^\top \boldsymbol{Q}_{-i}(\lambda)}{1+\frac{1}{n}\boldsymbol{v}_i^\top \boldsymbol{Q}_{-i}(\lambda)\boldsymbol{v}_i} \tag{36}$$

with $\boldsymbol{Q}_{-i}(\lambda) = \left(\frac{1}{n}\boldsymbol{\Sigma}_{-i}^\top \boldsymbol{\Sigma}_{-i} - \lambda \boldsymbol{I}_n\right)^{-1}$.

At this point, we would like to use a concentration of measure argument for the quadratic form $\frac{1}{n}\boldsymbol{v}_i^\top \boldsymbol{Q}_{-i}\boldsymbol{v}_i$: if the entries of $\boldsymbol{v}_i$ were i.i.d., we could use the Trace Lemma 21 in Appendix A in order to conclude that $\frac{1}{n}\boldsymbol{v}_i^\top \boldsymbol{Q}_{-i}\boldsymbol{v}_i - \frac{1}{n}\operatorname{Tr}\boldsymbol{Q}_{-i} \to 0$ almost surely as $n \to \infty$ (notice that $\boldsymbol{Q}_{-i}$ is independent from $\boldsymbol{v}_i$). However, $\boldsymbol{v}_i = f(\boldsymbol{X}\boldsymbol{w}_i)$, therefore its entries are interrelated; therefore we need a "nonlinear" version of the Trace Lemma:

Proposition 10 (Lemma 1, [12]). *Consider a matrix $\boldsymbol{A} \in \mathrm{Mat}_{n,n}(\mathbb{R})$ with bounded norm $\|\boldsymbol{A}\| \leq 1$ and a random vector $\boldsymbol{v} = f(\boldsymbol{X}^\top \boldsymbol{w}) \in \mathbb{R}^n$ with $\boldsymbol{w}$ with i.i.d. normal entries and f Lipschitz continuous, then*

$$\mathbb{P}\left(\left|\frac{1}{n}\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} - \frac{1}{n}\operatorname{Tr}\boldsymbol{\Phi}\boldsymbol{A}\right| \geq t\right) \leq Ce^{-cN\min\{t,t^2\}} \tag{37}$$

for some positive constant $C, c > 0$, with

$$\boldsymbol{\Phi} = \mathbb{E}_{\boldsymbol{w}}\left[\boldsymbol{v}\boldsymbol{v}^\top\right] = \mathbb{E}_{\boldsymbol{w}}\left[f(\boldsymbol{X}^\top \boldsymbol{w})f(\boldsymbol{w}^\top \boldsymbol{X})\right]. \tag{38}$$

Thanks to the above result it is possible to derive the limiting spectral measure of the "Gram matrix" $\frac{1}{n}\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma}$:

Theorem 11 (Theorem 2, [12]). *Let μ_n be the empirical spectral density of $\frac{1}{n}\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma}$. Then, for every bounded continuous function f*

$$\int f \mathrm{d}\mu_n - \int f \mathrm{d}\bar{\mu}_n \to 0, \quad \text{as } n \to \infty \text{ with probability one,} \tag{39}$$

where $\bar{\mu}_n$ is the measure defined by the Stieltjes transform

$$m_{\bar{\mu}_n}(z) = \frac{1}{n}\operatorname{Tr}\left(\frac{N}{n}\frac{\boldsymbol{\Phi}}{1+\delta_z} - z\boldsymbol{I}_n\right)^{-1}, \qquad z \in \mathbb{C}_+ \tag{40}$$

and δ_z is the unique solution in $\mathbb{C}_+$ of the equation

$$\delta_z = \frac{1}{n}\operatorname{Tr}\boldsymbol{\Phi}\left(\frac{N}{n}\frac{\boldsymbol{\Phi}}{1+\delta_z} - z\boldsymbol{I}_n\right)^{-1}. \tag{41}$$

Finally, we can state the final result about the asymptotic behaviour of the training error of our model:

Theorem 12 (Theorem 3, [12]). *Let $\boldsymbol{W} \sim \mathcal{N}(0, \boldsymbol{I}_N)$, f Lipschitz continuous and $\boldsymbol{X}$ of bounded norm. Then, as $n, p, N \to \infty$ with $\frac{p}{n} \to \gamma_1$ and $\frac{N}{n} \to \gamma_2$, for all $\epsilon > 0$*

$$N^{\frac{1}{2}-\epsilon}\left(E_{\text{train}} - \bar{E}_{\text{train}}\right) \to 0, \quad \text{as } N \to \infty, \tag{42}$$

where

$$\bar{E}_{\text{train}} = \frac{\lambda^2}{n} \operatorname{Tr}\left[\boldsymbol{Y}^\top \boldsymbol{Y} \bar{\boldsymbol{Q}} \left(\frac{\frac{1}{N}\operatorname{Tr} \boldsymbol{\Psi}\bar{\boldsymbol{Q}}^2}{1 - \frac{1}{N}\operatorname{Tr}\boldsymbol{\Psi}^2\bar{\boldsymbol{Q}}^2} + \boldsymbol{I}_n \right) \bar{\boldsymbol{Q}} \right] \tag{43}$$

$$\bar{\boldsymbol{Q}} = (\boldsymbol{\Psi} + \lambda \boldsymbol{I}_n)^{-1} \tag{44}$$

$$\boldsymbol{\Psi} = \frac{N}{n} \frac{\boldsymbol{\Phi}}{1+\delta} \tag{45}$$

$$\delta = \frac{1}{n} \operatorname{Tr} \boldsymbol{\Phi} \bar{\boldsymbol{Q}} \tag{46}$$

Very rough sketch of the proof. We recall that $E_{\text{train}} = \frac{\lambda^2}{n} \operatorname{Tr} \boldsymbol{Y}^\top \boldsymbol{Y} \boldsymbol{Q}_{\boldsymbol{\Sigma}}^2$, with $\boldsymbol{Q}_{\boldsymbol{\Sigma}}$ the resolvent of $\frac{1}{n}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}$. Using concentration of measure arguments, we can first show that in the large n, p, N limit (Corollary 3, [12])

$$\mathbb{P}\left(\left| \operatorname{Tr} \boldsymbol{Y}^\top \boldsymbol{Y} \boldsymbol{Q}_{\boldsymbol{\Sigma}}^2 - \operatorname{Tr} \boldsymbol{Y}^\top \boldsymbol{Y} \mathbb{E}[\boldsymbol{Q}_{\boldsymbol{\Sigma}}^2] \right| > t \right) \leq C e^{-cNt^2},$$

for some $C, c > 0$, and subsequently (Proposition 1, [12] with $\boldsymbol{A} = \boldsymbol{I}_n$) $\forall\ \epsilon > 0$, there exists $\tilde{c} > 0$ such that

$$\left\| \mathbb{E}[\boldsymbol{Q}_\Sigma^2] - \left(\bar{\boldsymbol{Q}}\bar{\boldsymbol{Q}} + \frac{\frac{1}{N}\operatorname{Tr}\boldsymbol{\Psi}\bar{\boldsymbol{Q}}^2}{1 - \frac{1}{N}\operatorname{Tr}\boldsymbol{\Psi}^2\bar{\boldsymbol{Q}}^2} \bar{\boldsymbol{Q}}^2 \right) \right\| \leq \tilde{c} N^{-\frac{1}{2}+\epsilon}.$$

The result follows. □

Therefore, the evaluation of the training error can be estimated by the evaluation of the quantity $\bar{E}_{\text{train}}$ which is explicit and it only depends on the matrix

$$\boldsymbol{\Phi} = \mathbb{E}_{\boldsymbol{w}}\left[\boldsymbol{v}\boldsymbol{v}^\top \right] = \mathbb{E}_{\boldsymbol{w}}\left[f(\boldsymbol{X}^\top \boldsymbol{w}) f(\boldsymbol{w}^\top \boldsymbol{X}) \right]. \tag{47}$$

In particular, for $\boldsymbol{W} \sim \mathcal{N}(0, \boldsymbol{I}_N)$ and ReLU activation function $f(t) = \max\{0, t\}$, the matrix $\boldsymbol{\Phi}$ has entries (Table 1, [12])

$$\begin{aligned} \Phi_{\boldsymbol{a},\boldsymbol{b}} &= \frac{1}{(2\pi)^{p/2}} \int_{\mathbb{R}^p} f(\boldsymbol{w}^\top \boldsymbol{a}) f(\boldsymbol{w}^\top \boldsymbol{b}) \mathrm{d}\boldsymbol{w} \\ &= \frac{1}{2\pi} \|\boldsymbol{a}\| \|\boldsymbol{b}\| \left[\frac{\boldsymbol{a}^\top \boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|} \arccos\left(-\frac{\boldsymbol{a}^\top \boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|} \right) + \sqrt{1 - \left(\frac{\boldsymbol{a}^\top \boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|} \right)^2} \right]. \end{aligned} \tag{48}$$

Up to now we assumed the given data matrix $\boldsymbol{X}$ to be fixed (deterministic). We can further inspect the behaviour of the training error by considering $\boldsymbol{X}$ to be a random object: suppose that the entries of $\boldsymbol{X}$ are i.i.d. Gaussians with zero mean and variance $\sigma_{\boldsymbol{X}}^2$ ($X_{ij} \sim \mathcal{N}(0, \sigma_{\boldsymbol{X}}^2)$). We are further assuming the target matrix $\boldsymbol{Y}$ to have Gaussian entries, independent from $\boldsymbol{X}$.

In the same setting as before, the average training error with estimator $\boldsymbol{A}^*$ is

$$\begin{aligned} E_{\text{train}} &= \mathbb{E}_{\boldsymbol{W},(\boldsymbol{X},\boldsymbol{Y})}\left[\frac{1}{n}\left\|\boldsymbol{Y}-(\boldsymbol{A}^*)^\top\boldsymbol{\Sigma}\right\|_F^2\right] = \lambda^2\mathbb{E}_{\boldsymbol{W},(\boldsymbol{X},\boldsymbol{Y})}\left[\frac{1}{n}\operatorname{Tr}\left[\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{Q}_{\boldsymbol{\Sigma}}^2(-\lambda)\right]\right] \\ &= \lambda^2\mathbb{E}_{\boldsymbol{W},\boldsymbol{X}}\left[\frac{1}{n}\operatorname{Tr}\boldsymbol{Q}_{\boldsymbol{\Sigma}}^2(-\lambda)\right] = -\lambda^2\frac{\partial}{\partial\lambda}\mathbb{E}_{\boldsymbol{W},\boldsymbol{X}}\left[\frac{1}{n}\operatorname{Tr}\boldsymbol{Q}_{\boldsymbol{\Sigma}}(-\lambda)\right] \\ &= -\lambda^2\frac{\partial}{\partial\lambda}\mathbb{E}_{\boldsymbol{W},\boldsymbol{X}}\left[\frac{1}{n}\operatorname{Tr}\left(\frac{1}{n}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}+\lambda\boldsymbol{I}_n\right)^{-1}\right] \end{aligned} \tag{49}$$

where in the last line we can discern a Stieltjes transform (see Appendix A for all the details).

Recall the Stieltjes transforms of the (non-linear) Gram matrices $\frac{1}{n}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ and $\frac{1}{n}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}$:

$$g(\lambda) = \frac{1}{N}\operatorname{Tr}\left[\left(\frac{1}{n}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top-\lambda\boldsymbol{I}_N\right)^{-1}\right] \qquad \text{and} \qquad \tilde{g}(\lambda) = \frac{1}{n}\operatorname{Tr}\left[\left(\frac{1}{n}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}-\lambda\boldsymbol{I}_n\right)^{-1}\right] \tag{50}$$

which are related by the following equality (see Proposition 19)

$$g(\lambda) = \frac{n}{N}\tilde{g}(\lambda) - \frac{N-n}{N}\frac{1}{\lambda}. \tag{51}$$

Thus,

$$\begin{aligned} E_{\text{train}} &= -\lambda^2\frac{\partial}{\partial\lambda}\mathbb{E}_{\boldsymbol{W},\boldsymbol{X}}\left[\tilde{g}(-\lambda)\right] = -\lambda^2\frac{\partial}{\partial\lambda}\mathbb{E}_{\boldsymbol{W},\boldsymbol{X}}\left[\frac{N}{n}g(-\lambda)-\frac{N-n}{n}\frac{1}{\lambda}\right] \\ &= -\lambda^2\frac{N}{n}\frac{\partial}{\partial\lambda}\mathbb{E}_{\boldsymbol{W},\boldsymbol{X}}\left[g(-\lambda)\right]-\frac{N-n}{n} \end{aligned} \tag{52}$$

It remains to evaluate the quantity

$$G(z) := \mathbb{E}_{\boldsymbol{W},\boldsymbol{X}}\left[g(z)\right] = \mathbb{E}_{\boldsymbol{W},\boldsymbol{X}}\left[\frac{1}{N}\operatorname{Tr}\left[\left(\frac{1}{n}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top-\lambda\boldsymbol{I}_N\right)^{-1}\right]\right] \tag{53}$$

in the limit as $n, p, N \to +\infty$ with $p/n \to \gamma_1$, $N/n \to \gamma_2$.

The following result will show that the limiting Stieltjes transform, which we still call G with abuse of notation, can be described as the solution to a quartic polynomial expression and it depends exclusively on γ_1, γ_2 and two parameters related to the nonlinearity f:

$$\eta := \int f(\sigma_{\boldsymbol{X}}\sigma_{\boldsymbol{W}}t)^2 e^{-\frac{t^2}{2}}\frac{\mathrm{d}t}{\sqrt{2\pi}} \qquad \text{Gaussian mean of } f^2 \tag{54}$$

$$\zeta := \left(\sigma_{\boldsymbol{X}}\sigma_{\boldsymbol{W}}\int f'(\sigma_{\boldsymbol{X}}\sigma_{\boldsymbol{W}}t)e^{-\frac{t^2}{2}}\frac{\mathrm{d}t}{\sqrt{2\pi}}\right)^2 \qquad \text{square of the Gaussian mean of } f'. \tag{55}$$

It therefore exhibits universal properties.

Theorem 13 (Pennington, Worah, '17 [19])**.** *Consider two random matrices $\boldsymbol{X} \in \mathrm{Mat}_{p,n}(\mathbb{R})$ (data) and $\boldsymbol{W} \in \mathrm{Mat}_{N,p}(\mathbb{R})$ (weights), with $X_{ij} \sim \mathcal{N}(0, \sigma_{\boldsymbol{X}}^2)$ and $W_{ij} \sim \mathcal{N}(0, \frac{\sigma_{\boldsymbol{W}}^2}{p})$, and define the matrix $\boldsymbol{\Sigma} = \sigma(\boldsymbol{W}\boldsymbol{X})$ for some function $\sigma : \mathbb{R} \to \mathbb{R}$ with zero Gaussian mean and finite Gaussian moments.*

In the regime (32), the averaged Stieltjes transform $G(z)$ of the matrix $\frac{1}{n}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ satisfies the following equation

$$G(z) = \frac{\gamma_1}{\gamma_2}\frac{1}{z}P\left(\frac{\gamma_2}{\gamma_1}\frac{1}{z}\right) + \left(1 - \frac{\gamma_1}{\gamma_2}\right)\frac{1}{z} \tag{56}$$

where P satisfies the implicit equation

$$P = 1 + (\eta - \zeta)zP_{\gamma_1}P_{\gamma_1/\gamma_2} + \frac{P_{\gamma_1}P_{\gamma_1/\gamma_2}z\zeta}{1 - P_{\gamma_1}P_{\gamma_1/\gamma_2}z\zeta} \tag{57}$$

$$P_{\gamma_1} = 1 + (P-1)\gamma_1 \qquad P_{\gamma_1/\gamma_2} = 1 + (P-1)\gamma_1/\gamma_2. \tag{58}$$

Observation on the proof. The proof is based on the moment methods: the calculations of the moments $m_k = \frac{1}{N}\mathbb{E}[\operatorname{Tr}\boldsymbol{M}^k]$ of the distribution of the matrix $\boldsymbol{M} := \frac{1}{n}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ via counting of certain connected outer-planar graph and evaluating multidimensional integrals related to those graphs. □

Note that if $\zeta = 0$, the expression for $G(z)$ reduces to the equation for the Marchenko-Pastur distribution (up to rescaling $\eta = 1$), although the function f is not necessarily the identity function.

Corollary 14. Given the random feature model (30), in the regime (32), the training error is equal to

$$E_{\text{train}} = 1 - \gamma_2 - \lambda^2\gamma_2\frac{\partial}{\partial\lambda}G(-\lambda) \tag{59}$$

where G satisfies equation (56)–(58).

3.2 Loss landscape

Consider again a two-layer neural network with linear output function and no biases as in (**??**):

$$\hat{\boldsymbol{y}} = \boldsymbol{A}^\top f(\boldsymbol{W}\boldsymbol{x}) \in \mathbb{R}^o, \qquad \boldsymbol{x} \in \mathbb{R}^p \text{ (data point)}, \tag{60}$$

with weight matrices $\boldsymbol{W} \in \mathrm{Mat}_{N,p}(\mathbb{R})$, $\boldsymbol{A} \in \mathrm{Mat}_{N,o}(\mathbb{R})$, as before, and ReLU activation function $f(x) = \max\{0, x\}$. We are additionally assuming $p = N = o$ (a square network).

For a given set of sample data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, we are again interested in a regression task with (unregularized) mean square error as loss function[3]

$$\mathcal{L} = \frac{1}{2n}\left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\|_F^2 \tag{61}$$

in the regime as $n, N \to +\infty$, while $\frac{N}{n} \to \gamma \in \mathbb{R}_+$.

Understanding of the (high dimensional) landscape described by the loss function $\mathcal{L}$ can be fundamental for optimization and generalization purposes. The analysis of the Hessian of $\mathcal{L}$, and of its spectrum in particular, can shed light on the complexity of the geometry of $\mathcal{L}$ and on the properties of stationary points.

[3]The factor of $\frac{1}{2}$ is simply for cosmetic purposes.

If we rewrite the set of parameters $\{\boldsymbol{W}, \boldsymbol{A}\}$ in form of a single column vector $\boldsymbol{\theta}$, the Hessian can be decomposed into the sum of two matrices

$$\boldsymbol{H}[\mathcal{L}] = \left[\frac{\partial^2 \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta}\right]_{\alpha,\beta=1,\dots,2N^2} = \boldsymbol{H}_0 + \boldsymbol{H}_1 \tag{62}$$

where

$$H_{0;\alpha,\beta} = \frac{1}{n}\left[\boldsymbol{J}\boldsymbol{J}^\top\right]_{\alpha,\beta}$$
$$H_{1;\alpha,\beta} = \frac{1}{n}\sum_{i,j=1}^{n,N}(\hat{y}_{ij} - y_{ij}) \cdot \frac{\partial \hat{y}_{ij}}{\partial\theta_\alpha \partial\theta_\beta}$$

and $\boldsymbol{J} = \boldsymbol{J}[\mathcal{L}]$ is the Jacobian of the loss. By construction, both matrices are real symmetric and $\boldsymbol{H}_0$ is positive semi-definite.

As a first toy model, we will assume some additional strong assumptions (Pennington, Bahri, '17 [18]): the data matrix $\boldsymbol{X}$ and the weights matrices $\boldsymbol{W}$, $\boldsymbol{A}$ have i.i.d. normal entries; the residuals $\epsilon_{ij} := \hat{y}_{ij} - y_{ij}$ are i.i.d. normal with tunable variance: $\epsilon_{ij} \sim \mathcal{N}(0, 2\delta)$ for some $\delta > 0$. Additionally, we assume $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$ to be (asymptotically) freely independent[4], and $\boldsymbol{J}$, $\boldsymbol{H}_1$ to have i.i.d. normal entries: $J_{i,j} \sim \mathcal{N}(0,1)$, $H_{1;ij} \sim \mathcal{N}(0, 2\delta)$. In this setting, we therefore need to study the spectrum of the sum of a "Wishart + Wigner" matrix.

We recall the limiting distributions of the spectra of both ensembles, namely the Marchenko-Pastur law and the Wigner's semicircle Law:

$$\rho_{\mathrm{MP};\gamma}(t) = \max\left\{0, 1-\gamma^{-1}\right\}\delta_0(t) + \frac{1}{2\pi\gamma}\frac{\sqrt{(a_+ - t)(t - a_-)}}{t}\chi_{[a_-,a_+]}(t) \tag{63}$$

$$\rho_{\mathrm{W}}(t) = \frac{1}{4\pi\delta}\sqrt{8\delta - t^2}\,\chi_{[-2\sqrt{2\delta},2\sqrt{2\delta}]} \tag{64}$$

where $a_\pm = \left(1 \pm \sqrt{\gamma}\right)^2$.

We will now resort to the fact that $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$ are freely independent to easily compute the sum distribution via the $\mathcal{R}$ transform (see Appendix B.3): we recall that given two Hermitian random matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathrm{Mat}_{n,n}(\mathbb{C})$ (or $\boldsymbol{A}, \boldsymbol{B} \in \mathrm{Mat}_{n,n}(\mathbb{R})$) with limiting spectral density $\mu_{\boldsymbol{A}}$ and $\mu_{\boldsymbol{B}}$ respectively, then if they are asymptotically free almost everywhere,

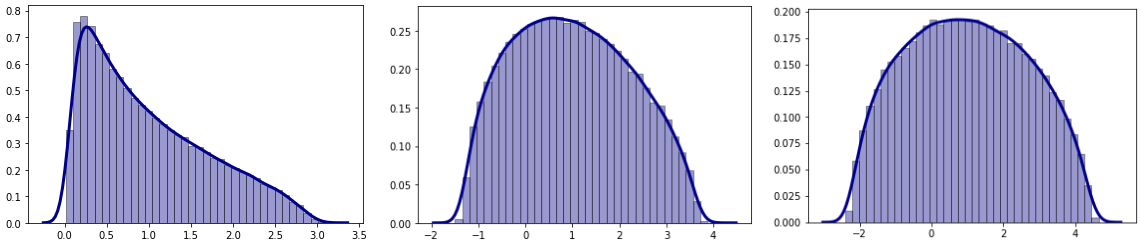
$$\mathcal{R}_{\boldsymbol{A}+\boldsymbol{B}}(z) = \mathcal{R}_{\boldsymbol{A}}(z) + \mathcal{R}_{\boldsymbol{B}}(z) \tag{65}$$

where $\mathcal{R}_{\boldsymbol{A}}, \mathcal{R}_{\boldsymbol{B}}, \mathcal{R}_{\boldsymbol{A}+\boldsymbol{B}}$ are the $\mathcal{R}$ transform of the limiting spectral density of $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{A}+\boldsymbol{B}$ respectively.

In our case, we have that the $\mathcal{R}$ transforms of the Marchenko-Pastur and Wigner laws read

$$\mathcal{R}_{\boldsymbol{H}_0}(z) = \frac{1}{1-\gamma z}, \qquad \mathcal{R}_{\boldsymbol{H}_1}(z) = 2\delta z. \tag{66}$$

[4]Free independence between non-commutative random variables can be though as the analogue of the classical notion of independence. See Appendix B.3 for more details.



(a) $2\delta = 0.1$. (b) $2\delta = 1$. (c) $2\delta = 1.5$

Figure 5: Histogram and density plot of the eigenvalues of 400 realizations of the sum of a Wigner matrix and a Wishart matrix with parameters $p = 110$, $n = 220$, $\gamma = \frac{p}{n} = \frac{1}{2}$, and different values of δ.

therefore

$$\mathcal{R}_{\boldsymbol{H}[\mathcal{L}]}(z) = \frac{1}{1-\gamma z} + 2\delta z. \tag{67}$$

and the Stieltjes transform satisfies the cubic equation

$$2\delta g^3_{\boldsymbol{H}[\mathcal{L}]} - (2\delta + z\gamma) g^2_{\boldsymbol{H}[\mathcal{L}]} + (z + \gamma - 1) g_{\boldsymbol{H}[\mathcal{L}]} - 1 = 0, \tag{68}$$

with asymptotic behaviour $g_{\boldsymbol{H}[\mathcal{L}]}(z) \sim \frac{1}{z}$ as $z \to \infty$. See Figure 5 for an example of the limiting distribution.

It is natural to expect (and it can be inspected numerically, see [18]) that δ plays the role of an interpolation parameter: for small values of δ, the distribution resembles the Marchenko-Pastur distribution, while for big values of δ the distribution resembles a semicircle.

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

$$\alpha(\delta, \gamma) = \int_{-\infty}^{0} \rho(t; \delta, \gamma) \ \mathrm{d}t,$$

as it measures the number of "descent directions", which is crucial in optimization: previous works (e.g. Choromanska et al., '15 [7]) showed that critical points with many descent directions have large loss value; therefore, identifying critical points with small index can be a valid methodology for choosing good candidates for the minimizer of the loss function.

On the other hand, many of the previous assumptions ought to be relaxed in order to have a more realistic setting for a neural network and the research in this direction is still open. As an example, we mention a slightly more advanced analysis in Pennington, Bahri. '17 [18] where they model the distribution of $\boldsymbol{H}_1$ as a product of Wishart matrices, and the paper by Pennington, Worah, '18 [20] where they study more closely the spectrum of $\boldsymbol{H}_0$.

3.3 Generalization

Check Couillet's slides in section "Random Matrix Analysis for Learning Dynamics of Neural Networks – Generalization performance", taken from the paper "The Dynamics of Learning: A Random Matrix Approach" (Liao, Couillet)

honorable mention: double descent curve is explained through singularity of the inverse of random (Gram) matrix

From "Surprises in high-dimensional ridgeless least squares interpolation" (Hastie, Montanari, et al. [11]) and "The generalization error for random features regression: precise asymptotics and double descent curve" (Mei, Montanari, [17]) and others about double descent, maybe

completely to do

4 Some conclusions

Take-away messages:

- asymptotic "concentration effect"(?) for large n, p yields simplification in analyses and models.
- ability to detect, estimate non-trivial phase transition phenomena, like double descent, when $p, n \to \infty$
- access to limiting performances and not only bounds! This will gives us more insight on hyperparameter optimization and algorithm improvement.
- strong coincidence with real datasets, easy link between theory and practice (mention some random papers).

Also, perspectives and open problems: not sure what to make of it in my own words...

- neural nets: loss landscape, gradient descent dynamics and deep learning!
- generalized linear models
- more general problems from convex optimization (often of implicit solution)
- more difficult: problem raised from non-convex optimization problems
- transfer learning, active learning, generative networks (GAN)
- robust statistics in machine learning
- etc.

A The Stieltjes Transform and proof of the Marchenko-Pastur Law

As in the Wigner semicircle law, also for the Marchenko-Pastur law, its proof can be carried out using the moments method . Here we will report a simple proof using the method of the Stietljes Transform. We will first recall a few fundamental properties of the Stieltjes theory and then proceed with the proof.

A few facts about the Stieltjes transform. For the sake of simplicity, we may 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 $\mathrm{d}\mu(t) = \rho(t)\mathrm{d}t$), however the following discussion holds for general measures.

Definition 15. For a (real) probability measure μ with support $\operatorname{supp}\mu$, its **Stieltjes transform** g_μ is defined for $z \in \mathbb{C} \setminus \operatorname{supp}\mu$ as

$$g_\mu(z) = \int_{\mathbb{R}} \frac{1}{t-z}\mathrm{d}\mu(t) \tag{69}$$

We can rewrite the Stieltjes transform as

$$g_\mu(z) = -\frac{1}{z}\int_{\mathbb{R}} \frac{1}{1-\frac{t}{z}}\mathrm{d}\mu(t) = -\sum_{k=0}^{\infty} \frac{m_k}{z^{k+1}}, \tag{70}$$

where $m_k = \int_{\mathbb{R}} t^k \mathrm{d}\mu(t)$ is the k-th moments of the measure μ, provided that they are finite and that the resulting series converges. Also, note that since the measure μ is supported on $\mathbb{R}$, we have $g_\mu(\bar{z}) = \overline{g_\mu(z)}$.

The fundamental result about the Stieltjes transform is that the measure itself can be recovered via the **Inverse Stieltjes Transform**:

Proposition 16. *Given a measure μ with its Stieltjes transform g_μ, then $\forall\ a, b \in \mathbb{R}$, $a < b$, continuity points of μ,*

$$\mu([a,b]) = \lim_{\epsilon \searrow 0} \frac{1}{\pi}\int_a^b \Im\left[g_\mu(t+i\epsilon)\right]\mathrm{d}t. \tag{71}$$

In particular, if the measure μ admits a density function $\rho(t)$, then

$$\rho(t) = \lim_{\epsilon \searrow 0} \frac{1}{\pi}\Im\left[g_\mu(t+i\epsilon)\right] \tag{72}$$

for any continuity point of the measure.

Proof. We will only prove the first identity. We first notice that

$$\begin{aligned}
\frac{1}{\pi}\int_a^b \Im\left[g_\mu(t+i\epsilon)\right]\mathrm{d}t &= \frac{1}{\pi}\int_a^b \Im\left[\int_{\mathbb{R}} \frac{1}{x-(t+i\epsilon)}\mathrm{d}\mu(x)\right]\mathrm{d}t = \frac{1}{\pi}\int_a^b \Im\left[\int_{\mathbb{R}} \frac{1}{x-(t+i\epsilon)}\mathrm{d}\mu(x)\right]\mathrm{d}t \\
&= \frac{1}{\pi}\int_a^b \int_{\mathbb{R}} \frac{\epsilon}{(x-t)^2+\epsilon^2}\mathrm{d}\mu(x)\,\mathrm{d}t = \frac{1}{\pi}\int_{\mathbb{R}}\int_a^b \frac{\epsilon}{(x-t)^2+\epsilon^2}\mathrm{d}t\,\mathrm{d}\mu(x) \\
&= \int_{\mathbb{R}} f_\epsilon(x;a,b)\,\mathrm{d}\mu(x)
\end{aligned} \tag{73}$$

with

$$f_\epsilon(x;a,b) = \frac{1}{\pi}\left[\arctan\left(\frac{b-x}{\epsilon}\right) - \arctan\left(\frac{a-x}{\epsilon}\right)\right] = \frac{1}{\pi}\arctan\left(\frac{\epsilon(b-a)}{\epsilon^2+(b-x)(a-x)}\right).$$

It is easy to see that as $\epsilon \to 0_+$

$$f_\epsilon(x;a,b) \to \begin{cases} 0 & x \in \mathbb{R} \setminus [a,b] \\ 1 & x \in (a,b) \\ \frac{1}{2} & x = a, b. \end{cases} \tag{74}$$

Moreover, since f_ϵ is uniformly bounded ($0 \leq f_\epsilon(x;a,b) \leq 1 \ \forall\, x \in \mathbb{R}, \ \forall\, \epsilon$) and $f_\epsilon(x) \in \mathcal{O}(\frac{1}{x^2})$ as $x \to \pm\infty$, by dominate convergence theorem we can conclude that

$$\int_{\mathbb{R}} f_\epsilon(x;a,b)\, \mathrm{d}\mu(x) \to \int_a^b \mathrm{d}\mu(x) = \mu([a,b]). \tag{75}$$

□

A direct consequence of this proposition is the fact that if two probability measures have the same Stieltjes transform, then they must coincide. Additionally, we recall now a useful theorem about convergence of measures: if we have a sequence of measures $\{\mu_n\}$ whose Stieltjes transform $\{g_{\mu_n}\}$ converges to a function g, then $\{\mu_n\} \to \mu$ where μ is a measure with Stieltjes transform g. More precisely,

Proposition 17. *Let $\{\mu_n\}$ be a sequence of probability measures, with Stieltjes transforms $\{g_{\mu_n}\}$. Suppose there is a probability measure μ with Stieltjes transform g_μ, such that $g_{\mu_n}(z) \to g_\mu(z)$ $\forall\ z \in A \subseteq \mathbb{C}$, where the set A contains at least one accumulation point. Then, $\mu_n \to \mu$ weakly.*

Proof. We refer to the proof in [2, Theorem 2.4.4]. □

We will focus now on the case of (symmetric) random matrices.

Proposition 18. *Given a symmetric matrix $\boldsymbol{M} \in \mathrm{Mat}_{p,p}(\mathbb{R})$ with empirical spectral distribution $\mu(t) = \frac{1}{p}\sum_{i=1}^p \delta_{\lambda_i(\boldsymbol{M})}(t)$, its Stieltjes transform is*

$$g_\mu(z) = \frac{1}{p} \operatorname{Tr}\left[(\boldsymbol{M} - z\boldsymbol{I}_p)^{-1}\right] \tag{76}$$

where the function $(\boldsymbol{M} - z\boldsymbol{I}_p)^{-1}$ is the resolvent of $\boldsymbol{M}$.

Proof.

$$\begin{aligned} g_\mu(z) = \int \frac{1}{t-z}\mathrm{d}\mu(t) = \frac{1}{p}\sum_{i=1}^p \frac{1}{\lambda_i(\boldsymbol{M}) - z} &= \frac{1}{p}\operatorname{Tr}\left[(\operatorname{diag}\{\lambda_i(\boldsymbol{M})\} - z\boldsymbol{I}_p)^{-1}\right] \\ &= \frac{1}{p}\operatorname{Tr}\left[(\boldsymbol{M} - z\boldsymbol{I}_p)^{-1}\right] \end{aligned} \tag{77}$$

□

In particular, if $\boldsymbol{M}$ is a Gram matrix, the following result holds.

Proposition 19. *Given $\boldsymbol{X} \in \mathrm{Mat}_{p,n}(\mathbb{R})$, let μ be the empirical spectral distribution of $\boldsymbol{X}\boldsymbol{X}^\top$ and $\tilde{\mu}$ be the empirical spectral distribution of $\boldsymbol{X}^\top\boldsymbol{X}$. Then,*

$$g_\mu(z) = \frac{n}{p} g_{\tilde{\mu}}(z) - \frac{p-n}{p}\frac{1}{z}. \tag{78}$$

Proof. It easily follows from the expression of the Stieltjes transform:

$$\frac{1}{p}\sum_{i=1}^{p} \frac{1}{\lambda_i(\boldsymbol{X}\boldsymbol{X}^\top) - z} = \frac{1}{p}\sum_{i=1}^{n} \frac{1}{\lambda_i(\boldsymbol{X}^\top\boldsymbol{X}) - z} + \frac{1}{p}(p-n)\frac{1}{-z} \tag{79}$$

□

Proof of the Marchenko-Pastur law We are now ready to prove the Marchenko-Pastur law. The idea of the proof 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.

Proof. We will follow the proof presented in [8, Section 3.2]. The proof is a bit lengthy, but quite educational, as it uses some simple linear algebra identities and some concentration of measure results. Along the proof, we will require the entries of the matrix $\boldsymbol{X}$ to have finite 8th order moment. However, this hypothesis can be loosened.

Step 1.

Let $\mu_p = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i(1/n\boldsymbol{X}\boldsymbol{X}^\top)}$ be the empirical spectral distribution of the matrix $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$. We will start with rewriting the Stieltjes transform of the empirical spectral distribution as

$$g_{\mu_p}(z) = \frac{1}{p}\operatorname{Tr}\left[\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_p\right)^{-1}\right] = \frac{1}{p}\sum_{j=1}^{p}\left[\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_p\right)^{-1}\right]_{jj}. \tag{80}$$

and we want to analyze each element in the summation.

Let us split the matrix $\boldsymbol{X}$ as

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{y}^\top \\ \boldsymbol{Y} \end{bmatrix}$$

for some $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \mathrm{Mat}_{p-1,n}(\mathbb{R})$, so that for $\Im[z] > 0$

$$\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_p\right)^{-1} = \begin{bmatrix} \frac{1}{n}\boldsymbol{y}^\top\boldsymbol{y} - z & \frac{1}{n}\boldsymbol{y}^\top\boldsymbol{Y}^\top \\ \frac{1}{n}\boldsymbol{Y}\boldsymbol{y} & \frac{1}{n}\boldsymbol{Y}\boldsymbol{Y}^\top - z\boldsymbol{I}_{p-1} \end{bmatrix}^{-1}. \tag{81}$$

Lemma 20. *The following identity holds:*

$$\left[\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_p\right)^{-1}\right]_{11} = \frac{1}{-z - z\frac{1}{n}\boldsymbol{y}^\top(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y} - z\boldsymbol{I}_n)^{-1}\boldsymbol{y}} \tag{82}$$

Proof. Recall the block matrix inverse formula:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(A - BD^{-1}C)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix} \tag{83}$$

for any matrix subdivision A, B, C, D of a squared matrix.

Thanks to the representation (81) and by the above formula, we have

$$\left[\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_p\right)^{-1}\right]_{11} = \frac{1}{-z + \frac{1}{n}\boldsymbol{y}^\top\left[\boldsymbol{I}_n - \boldsymbol{Y}^\top(\frac{1}{n}\boldsymbol{Y}\boldsymbol{Y}^\top - z\boldsymbol{I}_{p-1})^{-1}\frac{1}{n}\boldsymbol{Y}\right]\boldsymbol{y}}$$
$$= \frac{1}{-z - z\frac{1}{n}\boldsymbol{y}^\top(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y} - z\boldsymbol{I}_n)^{-1}\boldsymbol{y}} \tag{84}$$

where the last equality follows from the identity $I - A(I + BA)^{-1}B = (I + AB)^{-1}$ (where I is the identity matrix with the right dimensionality). □

Step 2.

Lemma 21 (**Trace Lemma**, [4]). *Consider $\{x_n\}_{n\in\mathbb{N}}$ a sequence of random vectors $\boldsymbol{x}_n \in \mathbb{C}^n$, with i.i.d. entries of zero mean, unit variance and finite 8th order moment, and $\{\boldsymbol{A}_n\}_{n\in\mathbb{N}}$ a sequence of matrices $\boldsymbol{A}_n \in \mathrm{Mat}_{n,n}(\mathbb{C})$ independent from $\boldsymbol{x}_n$ with uniformly bounded spectral norm ($\limsup_n \|\boldsymbol{A}_n\| < \infty$). Then,*

$$\frac{1}{n}\boldsymbol{x}_n^\top\boldsymbol{A}_n\boldsymbol{x}_n - \frac{1}{n}\operatorname{Tr}\boldsymbol{A}_n \to 0 \qquad \text{almost surely, as } n \to +\infty. \tag{85}$$

Proof. The proof relies on the Markov Inequality and the Borel-Cantelli Lemma. We refer to [8, Section 3.2] for the details. □

Applying the Lemma above to the quantity at hand (82), we have

$$\left|\left[\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_p\right)^{-1}\right]_{11} - \frac{1}{-z - z\frac{1}{n}\operatorname{Tr}\left[(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y} - z\boldsymbol{I}_n)^{-1}\right]}\right| \to 0 \tag{86}$$

almost surely, as $n \to \infty$, as long as the denominator of the difference has imaginary part uniformly away from zero.

Step 3. Given the limit

$$\left|\left[\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_p\right)^{-1}\right]_{11} - \frac{1}{-z - z\frac{1}{n}\operatorname{Tr}\left[(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y} - z\boldsymbol{I}_n)^{-1}\right]}\right| \to 0 \tag{87}$$

we would like to estimate the second term using the original matrix $\boldsymbol{X}$ and not the submatrix $\boldsymbol{Y}$. To do so, we recall the following Lemma.

Lemma 22 (Rank-1 perturbation, [3]). *Consider $\boldsymbol{A}, \boldsymbol{B} \in \mathrm{Mat}_{n,n}(\mathbb{C})$ Hermitian, semi-positive definite matrices, with μ the empirical spectral distribution of $\boldsymbol{A}$, and $\boldsymbol{x} \in \mathbb{R}^n$. Then, for $z \in \mathbb{C} \setminus \mathrm{supp}(\mu)$*

$$\left|\frac{1}{n}\operatorname{Tr}\left[\boldsymbol{B}(\boldsymbol{A} + \boldsymbol{x}\boldsymbol{x}^\top - z\boldsymbol{I}_n)^{-1}\right] - \frac{1}{n}\operatorname{Tr}\left[\boldsymbol{B}(\boldsymbol{A} - z\boldsymbol{I}_n)^{-1}\right]\right| \le \frac{1}{n}\frac{\|\boldsymbol{B}\|}{\operatorname{dist}(z, \mathrm{supp}(\mu))} \tag{88}$$

with $\|\boldsymbol{B}\|$ the spectral norm of $\boldsymbol{B}$. In particular, if $\limsup \|\boldsymbol{B}\| < \infty$, then

$$\left|\frac{1}{n}\operatorname{Tr}\left[\boldsymbol{B}(\boldsymbol{A} + \boldsymbol{x}\boldsymbol{x}^\top - z\boldsymbol{I}_n)^{-1}\right] - \frac{1}{n}\operatorname{Tr}\left[\boldsymbol{B}(\boldsymbol{A} - z\boldsymbol{I}_n)^{-1}\right]\right| \to 0 \tag{89}$$

as $n \to \infty$.

In the present case we have $\boldsymbol{B} = \boldsymbol{I}_n$ and $\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} = \frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y} + \frac{1}{n}\boldsymbol{y}\boldsymbol{y}^\top$, therefore

$$\left[\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_p\right)^{-1}\right]_{11} - \frac{1}{-z - z\frac{1}{n}\operatorname{Tr}\left[(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} - z\boldsymbol{I}_n)^{-1}\right]} \to 0 \qquad \text{almost surely.} \tag{90}$$

Step 4. We can now recall the definition of the Stieltjes transform:

$$\frac{1}{-z - z\frac{1}{n}\operatorname{Tr}\left[(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} - z\boldsymbol{I}_n)^{-1}\right]} = \frac{1}{-z - z g_{\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}}(z)} = \frac{1}{-z + 1 - \frac{p}{n} - z\frac{p}{n} g_{\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top}(z)} \tag{91}$$

where the second identity follows from Proposition 19. In conclusion,

$$\left[\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_p\right)^{-1}\right]_{11} - \frac{1}{-z + 1 - \frac{p}{n} - z\frac{p}{n} g_{\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top}(z)} \to 0 \qquad \text{almost surely.} \tag{92}$$

Note that the second term in the difference is independent on choice of the entry of the matrix $\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_p\right)^{-1}$ and we can repeat the same passages for all the diagonal entries $\left[\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_p\right)^{-1}\right]_{jj}$, $j = 2, \dots, p$. Indeed, for the (j, j) entry it suffices to notice that we can rewrite the matrix $\boldsymbol{X}$ as

$$\boldsymbol{X} = \boldsymbol{E}_j \boldsymbol{X}^{(j)} \boldsymbol{E}_j, \qquad \text{with } \boldsymbol{X}^{(j)} = \begin{bmatrix} \boldsymbol{y}_j^\top \\ \boldsymbol{Y}^{(j)} \end{bmatrix} \tag{93}$$

where $\boldsymbol{X}^{(j)} \in \operatorname{Mat}_{p,n}(\mathbb{R})$ is the matrix $\boldsymbol{X}$ with the first and jth row exchanged and $\boldsymbol{E}_j \in \operatorname{Mat}_{n,n}(\mathbb{R})$ is the corresponding elementary matrix.

Therefore, we obtain

$$\left[\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_p\right)^{-1}\right]_{jj} - \frac{1}{-z + 1 - \frac{p}{n} - z\frac{p}{n} g_{\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top}(z)} \to 0 \qquad \text{almost surely,} \tag{94}$$

for all $j = 1, \dots, n$. Summing up all the terms and averaging, we obtain

$$\frac{1}{p}\sum_{j=1}^{p}\left[\left(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_p\right)^{-1}\right]_{jj} - \frac{1}{-z + 1 - \frac{p}{n} - z\frac{p}{n} g_{\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top}(z)} \to 0 \qquad \text{almost surely,} \tag{95}$$

i.e.

$$g_{\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top}(z) - \frac{1}{-z + 1 - \frac{p}{n} - z\frac{p}{n} g_{\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top}(z)} \to 0 \qquad \text{almost surely,} \tag{96}$$

for $\Im[z] > 0$, where we used the fact that the intersection of countably many sets of probability one on which the result holds is itself of probability one.

It can be proven that $g_{\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top}(z) \leq \frac{1}{\Im[z]}$ $\forall n$, therefore the sequence $\{g_{\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top}\}$ is uniformly bounded in a compact set and it converges (possibly via a subsequence) to a limit function $g(z)$:

$$g_{\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top}(z) \to g(z) \qquad \text{almost surely,} \tag{97}$$

with $g(z)$ solution to the implicit equation

$$g(z) = \frac{1}{-z + 1 - \kappa - z\kappa g(z)}. \tag{98}$$

We can explicitly solve the equation, which yields

$$\begin{aligned} g(z) &= \frac{1-\kappa}{2\kappa z} - \frac{1}{2\kappa} + \frac{\sqrt{(1-\kappa-z)^2 - 4\kappa z}}{2\kappa z} \\ &= \frac{1-\kappa}{2\kappa z} - \frac{1}{2\kappa} + \frac{\sqrt{(z-(1+\sqrt{\kappa})^2)\,(z-(1-\sqrt{\kappa})^2)}}{2\kappa z} \end{aligned} \tag{99}$$

where we choose $\mathbb{R}_+$ to be the branch cut of the square root.

We can now calculate the Inverse Stieltjes transform to recover the limit probability measure: for $t \in (1-\sqrt{\kappa}, 1+\sqrt{\kappa})$

$$\begin{aligned} \lim_{\epsilon \searrow 0} \frac{1}{\pi} \Im\left[g(t+i\epsilon)\right] &= \lim_{\epsilon \searrow 0} \frac{1}{\pi} \Im\left[\frac{(\kappa-1)\epsilon}{2\kappa(t^2+\epsilon^2)} + \frac{it}{2\kappa(t^2+\epsilon^2)} \sqrt{\left((1+\sqrt{\kappa})^2 - t - i\epsilon\right)\left(t + i\epsilon - (1-\sqrt{\kappa})^2\right)}\right] \\ &= \frac{1}{2\pi\kappa t} \sqrt{\left((1+\sqrt{\kappa})^2 - t\right)\left(t - (1-\sqrt{\kappa})^2\right)}, \end{aligned} \tag{100}$$

for $t \in (1-\sqrt{\kappa}, 1+\sqrt{\kappa})^c \setminus \{0\}$

$$\begin{aligned} \lim_{\epsilon \searrow 0} \frac{1}{\pi} \Im\left[g(t+i\epsilon)\right] &= \lim_{\epsilon \searrow 0} \frac{1}{\pi} \Im\left[\frac{(\kappa-1)\epsilon}{2\kappa(t^2+\epsilon^2)} - \frac{i\epsilon}{2\kappa(t^2+\epsilon^2)} \sqrt{\left(t + i\epsilon - (1+\sqrt{\kappa})^2\right)\left(t + i\epsilon - (1-\sqrt{\kappa})^2\right)}\right] \\ &= 0, \end{aligned} \tag{101}$$

and for $t = 0$

$$\begin{aligned} \lim_{\epsilon \searrow 0} \epsilon \Im\left[g(i\epsilon)\right] &= \lim_{\epsilon \searrow 0} \epsilon \Im\left[\frac{(\kappa-1)}{2\kappa\epsilon} - \frac{1}{2\kappa\epsilon} \sqrt{\left((1+\sqrt{\kappa})^2 - i\epsilon\right)\left((1-\sqrt{\kappa})^2 - i\epsilon\right)}\right] \\ &= \lim_{\epsilon \searrow 0} \epsilon \Im\left[\frac{(\kappa-1)}{2\kappa\epsilon} - \frac{1}{2\kappa\epsilon}\left[(1-\kappa) - \frac{i\epsilon}{1-\kappa} + \mathcal{O}(\epsilon^2)\right]\right] \\ &= \frac{\kappa-1}{\kappa}, \end{aligned} \tag{102}$$

provided that $\kappa > 1$ (a density function cannot be negative)[5].

Step 5.

The final step requires to prove that almost sure convergence of the Stieltjes transform induces weak convergence of the probability measure almost surely. We refer to [8, Section 3.2] for all the details, which are a bit technical. □

[5] A more rigorous argument to prove the density at $t = 0$ is by noticing that when $p > n$, the empirical spectral density of $\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$ does have zero eigenvalues and, more precisely, the proportion of zero eigenvalues is equal to $\frac{p-n}{p}$ (see the discussion in the introduction, Section 1.2); furthermore, the exact value $1 - \kappa^{-1}$ follows from imposing the measure to have total mass equal to 1.

B More cool stuff on RMT

B.1 Random Matrix Theory and Orthogonal Polynomials

We will focus on the case for Unitary matrices. Recall the eigenvalue distribution:

$$\mathrm{d}\mu(x_1,\dots,x_n) = \frac{1}{Z_n}\Delta(x_1,\dots x_n)^2 \prod_{j=1}^{n} e^{-V(x_j)}\mathrm{d}x_1\dots\mathrm{d}x_n \tag{103}$$

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

Our first goal is to evaluate Z_n. We notice that $\Delta(\boldsymbol{x})$ is rightfully called Vandermonde determinant, as it can be viewed as the determinant of the Vandermonde matrix

$$\Delta(\boldsymbol{x}) = \det W(\boldsymbol{x})$$

with

$$W(\boldsymbol{x}) = \begin{bmatrix} 1 & x_1 & \dots & x_1^{n-1} \\ 1 & x_2 & \dots & x_2^{n-1} \\ \vdots & & & \\ 1 & x_n & \dots & x_n^{n-1} \end{bmatrix}. \tag{104}$$

Therefore,

$$\begin{aligned}
\mathrm{d}\mu(x_1,\dots,x_n) &= \frac{1}{Z_n}\Delta(x_1,\dots x_n)^2 \prod_{i=1}^{n} e^{-V(x_i)}\mathrm{d}x_1\dots\mathrm{d}x_n \\
&= \frac{1}{Z_n}\Big(\det\left[x_i^{j-1}\right]_{1\le i,j\le n}\Big)^2 \prod_{i=1}^{n} e^{-V(x_i)}\mathrm{d}x_1\dots\mathrm{d}x_n \\
&= \frac{1}{Z_n}\prod_{i=1}^{n} e^{-\frac{V(x_i)}{2}} \det\left[x_i^{j-1}\right]_{1\le i,j\le n}\cdot\det\left[x_i^{j-1}\right]_{1\le i,j\le n}\prod_{i=1}^{n} e^{-\frac{V(x_i)}{2}}\mathrm{d}x_1\dots\mathrm{d}x_n \\
&= \frac{1}{Z_n}\det\left[e^{-\frac{V(\boldsymbol{x})}{2}}\right]\det\left[x_i^{j-1}\right]_{1\le i,j\le n}\cdot\det\left[x_i^{j-1}\right]_{1\le i,j\le n}\det\left[e^{-\frac{V(\boldsymbol{x})}{2}}\right]\mathrm{d}x_1\dots\mathrm{d}x_n \\
&= \frac{1}{Z_n}\det\left[x_i^{j-1}e^{-\frac{V(x_i)}{2}}\right]_{1\le i,j\le n}\cdot\det\left[x_i^{j-1}e^{-\frac{V(x_i)}{2}}\right]_{1\le i,j\le n}\mathrm{d}x_1\dots\mathrm{d}x_n
\end{aligned} \tag{105}$$

where we used $V(\boldsymbol{x}) = \mathrm{diag}\{V(x_1),\dots,V(x_n)\}$.

From this expression, we can easily evaluate Z_n:

Proposition 23. *The normalization constant is equal to*

$$Z_n = n!\det \boldsymbol{M}, \tag{106}$$

where $\boldsymbol{M}$ is the moment matrix with entries

$$M_{ij} = \int_{\mathbb{R}^n} x^{i+j}e^{-V(x)}\mathrm{d}x \qquad 0\le i,j\le n-1. \tag{107}$$

Proof. It follows form the general definition of the determinant of a matrix and (105). □

We will rewrite now the probability density function in a compact form that will involve a single determinant of a matrix with entries (i,j) given by a function $K(x_i,x_j)$ called *kernel*.

Proposition 24. *The following identity holds:*

$$\frac{1}{Z_n}\prod_{1\le i<j\le n}(x_1,\dots x_n)^2\prod_{i=1}^{n}e^{-V(x_i)}=\frac{1}{n!}\det\Big[K(x_i,x_j)\Big]_{1\le i,j\le n} \tag{108}$$

where

$$K(x,y)=e^{-\frac{V(x)+V(y)}{2}}\sum_{j,k=0}^{n-1}x^j[\boldsymbol{M}]^{-1}_{jk}y^k. \tag{109}$$

Proof.

$$\begin{aligned}\frac{1}{n!}\det\Big[K_n(x_a,x_b)\Big]_{1\le i,j\le n}&=\frac{1}{n!}\det\left[\sum_{j,k}e^{-\frac{V(x_a)}{2}}x_a^j[\boldsymbol{M}]^{-1}_{jk}x_b^ke^{-\frac{V(x_b)}{2}}\right]\\&=\frac{1}{n!}\det\left[e^{-\frac{V(\boldsymbol{x})}{2}}W(\boldsymbol{x})\boldsymbol{M}^{-1}W(\boldsymbol{x})^Te^{-\frac{V(\boldsymbol{x})}{2}}\right]\\&=\frac{1}{n!\det\boldsymbol{M}}\left[\det W(\boldsymbol{x})\right]^2e^{-\operatorname{Tr}V(\boldsymbol{x})}\\&=\frac{1}{Z_n}\Delta^2(\boldsymbol{x})e^{-\operatorname{Tr}V(\boldsymbol{x})}.\end{aligned} \tag{110}$$

□

Additionally, the above kernel satisfies the following properties.

Proposition 25. *The kernel (109) satisfies:*

1. $\int_{\mathbb{R}}K(x,z)K(z,y)\mathrm{d}z=K(x,y)$ *(reproducibility)*
2. $\int_{\mathbb{R}}K(x,x)\mathrm{d}x=n$ *(normalization)*
3. *(marginals)*

$$\begin{aligned}\int_{\mathbb{R}}\det\left[K(x_i,x_j)\right]_{i,j\le r}\mathrm{d}x_r&=(n-r-1)\det\left[K(x_i,x_j\right]_{i,j\le r-1}\\\int_{\mathbb{R}^{n-r}}\det\left[K(x_i,x_j)\right]_{i,j\le n}\mathrm{d}x_{r+1}\dots\mathrm{d}x_n&=(n-r)!\det\left[K(x_i,x_j\right]_{i,j\le r-1}\end{aligned}$$

The above propositions shows 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 (see [22] for all the details).

For the rest of the section, we shall choose the potential $V(x)$ to be a polynomial of even degree, with positive leading coefficient (e.g. $V(x)=x^{2n}$).

By construction, the moment matrix $\boldsymbol{M}$ is symmetric and it can be proven to be positive definite. Consider the Lower-Diagonal-Upper decomposition (keeping into account the symmetry)

$$\boldsymbol{M} = \boldsymbol{L}\boldsymbol{H}\boldsymbol{L}^\top$$

where $\boldsymbol{L}$ is a lower unipotent matrix (with ones on the diagonal) and $\boldsymbol{H} = \operatorname{diag}\{h_1, \dots, h_n\}$ and $h_j > 0 \ \forall \ j$, then

$$K(x,y) = e^{-\frac{n}{2}(V(x)+V(y))} \begin{bmatrix} 1 & x & \dots & x^{n-1} \end{bmatrix} \boldsymbol{L}^{-\top} \boldsymbol{H}^{-1} \boldsymbol{L}^{-1} \begin{bmatrix} 1 & y & \dots & y^{n-1} \end{bmatrix}^T \tag{111}$$

Definition 26. The polynomials

$$\begin{bmatrix} p_0(x) \\ p_1(x) \\ \vdots \\ p_{n-1}(x) \end{bmatrix} = \boldsymbol{L}^{-1} \begin{bmatrix} 1 \\ x \\ \vdots \\ x^{n-1} \end{bmatrix} \tag{112}$$

are called **orthogonal polynomials (OPs)** for the measure $e^{-V(x)}\mathrm{d}x$.

Therefore, we can rephrase the kernel as

$$K(x,y) = e^{-\frac{n}{2}(V(x)+V(y))} \sum_{j=0}^{n-1} \frac{p_j(x)p_j(y)}{h_j} \tag{113}$$

Proposition 27. *The following properties holds for the OPs $\{p_j(x)\}$ and are equivalent to the above definition:*

- *For each n, $p_j(x)$ is a monic polynomial of degree j:* $\deg p_j(x) = j$ *and* $p_j(x) = x^j + \mathcal{O}(n^{n-1})$.
- *The polynomials $\{p_j(x)\}$ are mutually orthogonal with respect to the measure $e^{-V(x)}\mathrm{d}x$:*

$$\int_{\mathbb{R}} p_i(x)p_j(x)e^{-V(x)}\mathrm{d}x = h_j\delta_{ij}.$$

- *$\{p_j(x)\}$ solve a three terms recurrence relation:*

$$xp_j(x) = p_{j+1} + \alpha_j p_j(x) + \frac{h_j}{h_{j-1}} p_{j-1}(x) \qquad \forall \ j,$$

for a suitable sequence $\{\alpha_j\}$.

We can push the dependency of the kernel on OPs even further and get a simple formula for the kernel.

Proposition 28 (Christoffel–Darboux formula). *For any set of OPs we have*

$$K(x,y) = e^{-\frac{n}{2}(V(x)+V(y))} \frac{1}{h_n} \frac{p_n(x)p_{n-1}(y) - p_n(y)p_{n-1}(x)}{x-y} \tag{114}$$

Proof. The formula follows from the three terms recurrence relation and from cancellations due to the sum becoming telescoping. □

The most notable example is the connection between the Gaussian Unitary Ensemble and the Hermite Polinomials: in this case, the potential V appearing in the density function of the eigenvalues is $V(x) = \frac{x^2}{2}$:

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

and the set of polynomial that is orthogonal with respect to the measure $e^{-\frac{x^2}{2}}\mathrm{d}x$ (i.e. the Gaussian measure) are indeed the Hermite polynomials:

$$H_n(x) = (-1)^n e^{\frac{x^2}{2}} \frac{\mathrm{d}^n}{\mathrm{d}x^n} e^{-\frac{x^2}{2}}, \qquad n \geq 0. \tag{115}$$

B.2 Asymptotics and universality: 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 $n \to +\infty$.

Consider the classical case where the limit spectrum is supported on a bounded interval and it never vanishes within it (GUE or Wishart matrices with $\frac{p}{n} = \kappa \leq 1$, for example). 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 make use of results borrowed from Determinantal Point Process (DPP) theory.

We have seen that the distribution of the eigenvalues of the unitary ensemble (but this applies to other ensembles as well) is a DPP. Indeed, its probability density function can be described in terms of the determinant of a matrix with entries given by a *correlation kernel* $K_n(x, y)$:

$$\frac{1}{Z_n} \prod_{1 \leq i < j \leq n} (x_1, \dots x_n)^2 \prod_{i=1}^n e^{-V(x_i)} = \frac{1}{n!} \det \Big[K_n(x_i, x_j) \Big]_{1 \leq i,j \leq n} \tag{116}$$

We report here the main result about limiting DPPs as the number of points grows to infinity:

Proposition 29. *Let $\mathcal{P}$ and $\mathcal{P}_n$ be determinantal point processes with kernels K and K_n respectively. Let K_n converge to K*

$$\lim_{n \to \infty} K_n(x, y) = K(x, y) \tag{117}$$

uniformly over compact subsets of $\mathbb{R}^2$. Then, $\mathcal{P}_n \to \mathcal{P}$ weakly.

Given a random matrix ensemble, here is the recipe for analyzing the limiting process in different points of its spectrum. Consider a fixed reference point x^* of the spectrum and apply the change of variables

$$x = x^* + \frac{\xi}{Cn^\gamma} \tag{118}$$

to the correlation kernel, with suitable values of $C,\ \gamma > 0$, depending on the random matrix model and on which zone of the spectrum we are focusing on (the edges behaviour or the bulk behaviour).

This is equivalent of centering and scaling the point process of eigenvalues around the reference point x^*. We can now perform the limit:

$$\lim_{n\to\infty} \frac{1}{Cn^\gamma} K_n \left(x^* + \frac{\xi}{Cn^\gamma}, x^* + \frac{\eta}{Cn^\gamma} \right) = K(\xi, \eta) \tag{119}$$

with ξ, η the new local coordinates of the limiting process.

Bulk universality A point x^* lies in the bulk of the spectrum of 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:

$$x = x^* + \frac{\xi}{n\rho(x^*)} \qquad y = x^* + \frac{\eta}{n\rho(x^*)} \tag{120}$$

Theorem 30 ([16])**.** *For the Unitary Ensemble, the local behaviour in the bulk of the spectrum is described by a DPP with correlation kernel given by*

$$\lim_{n\to+\infty} \frac{1}{n\rho(x^*)} K_n \left(x^* + \frac{\xi}{n\rho(x^*)}, x^* + \frac{\eta}{n\rho(x^*)} \right) = K_{\text{sine}}(\xi, \eta) \tag{121}$$

with

$$K_{\text{sine}}(\xi, \eta) = \frac{\sin\left(\pi(\xi - \eta)\right)}{\pi(\xi - \eta)}. \tag{122}$$

Notice that this results holds regardless on the choice of the potential $V(x)$, therefore it is universal.

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 $\rho(x)$ 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**.

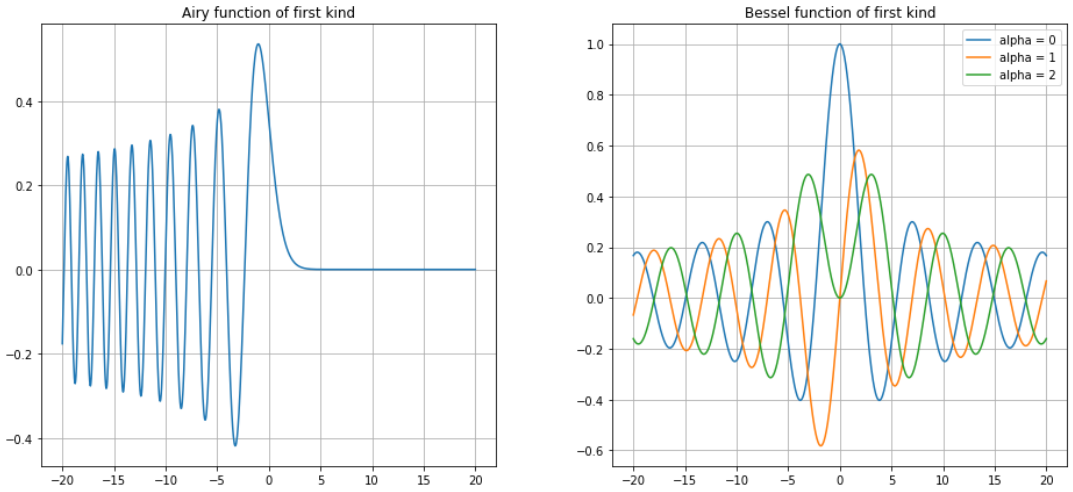
In the case of Unitary Ensemble with regular one-cut potential (e.g. $V(x) = x^2$), both edges (on $a = \pm 2\sqrt{n}$) are regular and the microscopic limit is described in the following theorem.

Theorem 31 ([16])**.**

$$\lim_{n\to+\infty} \frac{1}{2n^{\frac{1}{6}}} K_n \left(\pm 2\sqrt{n} + \frac{\xi}{2n^{\frac{1}{6}}}, \pm 2\sqrt{n} + \frac{\eta}{2n^{\frac{1}{6}}} \right) = K_{\text{Airy}}(\xi, \eta) \tag{123}$$

with

$$K_{\text{Airy}}(\xi, \eta) = \frac{\text{Ai}(\xi)\text{Ai}'(\eta) - \text{Ai}'(\xi)\text{Ai}(\eta)}{\xi - \eta}. \tag{124}$$



(a) Airy. (b) Bessel.

Figure 6: Plots of the Airy and Bessel functions Ai and J_α (for different values of α)

The function $\mathrm{Ai}(z)$ is the Airy function (see Figure 6a). It satisfies the second-order ODE

$$y'' = zy, \quad \text{such that } \lim_{z\to+\infty} y(z) = 0. \tag{125}$$

It can be represented as a contour integral

$$\mathrm{Ai}(z) = \int_\gamma e^{\frac{\zeta^3}{3} - z\zeta} \frac{\mathrm{d}\zeta}{2\pi i} \tag{126}$$

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

We briefly cite here the celebrated **Tracy–Widom distribution** that describes the local behaviour of the largest eigenvalue $\zeta_{\max}$ in the spectrum and its infinitesimal random oscillations:

Theorem 32 (Tracy, Widom, '94 [23])**.** *Consider the semi-infinite interval $[s, +\infty)$, then the distribution of the largest eigenvalue of the GUE ensemble obeys the following law*

$$\mathbb{P}\left(\zeta_{\max} < s\right) = \exp\left\{-\int_s^\infty (x-s)q^2(x)\mathrm{d}x\right\} \tag{127}$$

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

$$\begin{aligned} q''(x) &= 2q^3(x) + sq(x) \\ q(x) &\sim \mathrm{Ai}(x) \qquad x \to +\infty. \end{aligned} \tag{128}$$

Hard edge universality One instance of the so called "critical edges" can be seen in the Marchenko-Pastur distribution for sample covariance matrices when the value of the parameter κ is equal to 1. We recall the general expression for the density function of the MP-distrubtion:

$$\rho_{\mathrm{MP},\kappa}(x) = \begin{cases} \dfrac{1}{2\pi\kappa} \dfrac{\sqrt{(a_+ - x)(x - a_-)}}{x} & x \in [a_-, a_+] \setminus \{0\} \\ \max\{0, 1 - \kappa^{-1}\} & x = 0 \end{cases} \tag{129}$$

with $a_\pm = (1 \pm \sqrt{\kappa})^2$. If we set $\kappa = 1$, then the distribution has a square-root singularity at $x = 0$:

$$\rho_{\mathrm{MP},1}(x) = \frac{1}{2\pi}\sqrt{\frac{4-x}{x}} \qquad x \in (0, 4]. \tag{130}$$

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

$$K_{\mathrm{Bessel}}(\xi, \eta) = \frac{J_\alpha\left(\sqrt{\xi}\right)\sqrt{\eta}J'_\alpha\left(\sqrt{\eta}\right) - J'_\alpha\left(\sqrt{\xi}\right)\sqrt{\xi}J_\alpha\left(\sqrt{\eta}\right)}{2(\xi - \eta)} \qquad \xi, \eta \in \mathbb{R}_+,\ \alpha > -1. \tag{131}$$

The function $J_\alpha(z)$ is the Bessel function of first kind (see Figure 6b). It satisfies the second-order ODE

$$z^2 y'' + zy' + (z^2 - \alpha^2)y = 0, \tag{132}$$

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

$$J_\alpha(z) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\Gamma(n + \alpha + 1)}\left(\frac{z}{2}\right)^{2n+\alpha} \tag{133}$$

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

B.3 A few facts about Free Probability

We refer to Speicher [**?**] for a thorough exposition. Here we will report only the main results specialized to the set of Hermitian matrices (check Couillet-Debbah's book, Section 4.3 and 4.4)

B.4 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.

Need to mention a few more that are relevant in ML (check Couillet's slides...) – for example, the circular law

Circular ensemble. 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*.

add a few more words on it? If useful in ML...

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.

Multi-matrix models and external field. add references

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)

$$\mathrm{d}\mu(\boldsymbol{M}_1, \boldsymbol{M}_2) = e^{-\operatorname{Tr}\left[V_1(\boldsymbol{M}_1)+V_2(\boldsymbol{M}_2)-\boldsymbol{M}_1\boldsymbol{M}_2\right]}\mathrm{d}\boldsymbol{M}_1\mathrm{d}\boldsymbol{M}_2 \tag{134}$$

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,

$$\mathrm{d}\mu(\boldsymbol{M}) = e^{\operatorname{Tr}(\boldsymbol{M}^2)-\boldsymbol{A}\boldsymbol{M}}\mathrm{d}\boldsymbol{M}. \tag{135}$$

References

[1] J. F. Adams. *Lectures on Lie Groups*. Chicago University Press, 1969.

[2] 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.

[3] Z. D. Bai and J. W. Silverstein. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2), 1995.

[4] Z. D. Bai and J. W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large dimensional sample covariance matrices. *Ann. Prob.*, 26(1), 1998.

[5] Z. D. Bai and J. W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer, 2006.

[6] X. Cheng and A. Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(4), 2013.

[7] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun. The loss surfaces of multilayer networks. *JMLR*, 38, 2015.

[8] R. Couillet and M. Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

[9] P. Deift. Orthogonal polynomial and random matrices: a riemann-hilbert approach. In *Courant Lecture Notes in Mathematics*, volume 3. Amer. Math. Soc., Povidence R.I., 1999.

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

[11] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv:1903.08560*, 2019.

[12] C. Louart, Z. Liao, and R. Couillet. A random matrix approach to neural networks. *Annals of Applied Probability*, 28(2), 2017.

[13] C. Louart, Z. Liao, and R. Couillet. A random matrix approach to neural networks. *Ann. Appl. Probab.*, 28(2), 2018.

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

[15] C. H. Martin and M. W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *arXiv:1810.01075*, 2018.

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

[17] S. Mei and A. Montanari. The generalization error for random features regression: precise asymptotics and double descent curve. *arXiv:1908.05355*, 2019.

[18] 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.

[19] 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.

[20] J. Pennington and P. Worah. The spectrum of the Fisher information matrix of a single-hidden-layer neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

[21] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 21, 2008.

[22] A. Soshnikov. Determinantal random point fields. *Russian Mathematical Surveys*, 55:923 – 975, 2000.

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

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