On Random Matrices
Arising in Deep Neural Networks

L. Pastur

B. Verkin Institute
for Low Temperature Physics and Engineering
Kharkiv, Ukraine

Institute des Hautes Etudes Scientifiques
Bur sur Ivette, France

Random Matrices and Random Landscapes
Conference in honour of Yan Fyodorov's 60th birthday
CSF, Ascona, Switzerland, 25 – 29 July 2022

Outline

- Introduction
- Main Result
- Proof (outline)
- Numerical Results
- Summary

Artificial Neural Networks (NN) are très à la mode. There are various architectures, a typical is *fully connected feed forward NN*. It consists of

(1). **Iteration scheme** (NN dynamics). Given:

- x^0, the *input*, x^L, the *output*,
- $x^l = \{x^l_{j_l}\}^n_{j_l=1}, \ l = 0, \dots L$, the *state* of NN at the lth layer,
- $b^l = \{b^l_{j_l}\}^n_{j_l=1}, \ l = 1, \dots L$, *biases*,
- $W^l = \{W^l_{j_l, j_{l-1}}\}^n_{j_l, j_{l-1}=1}, \ l = 1, \dots L$, *(synaptic) weights*,

consider the recursion of *width n* and of *depth L*

$$y^l = W^l x^{l-1} + b^l, \quad x^l_{j_l} = \varphi(y^l_{j_l}), \; l = 1, \dots, L,$$

where the *nonlinearity* (*activation function*) $\varphi : \mathbb{R} \to \mathbb{R}$, usually a piece-wise differentiable sigmoid, e.g. $\varphi = \mathtt{tanh}, \mathtt{tan}^{-1}$,

$$\text{HardTanh} = 2^{-1}(|x+1| - |x-1|).$$

If the number of layers $L > 1$, NN is called *deep neural network* (DNN).

(2). **Training.** Updates the weigh matrix on the every step of the iteration procedure to reduce the misfit between the output data and the prescribed data by using certain optimization procedures, usually of the least mean square type.

Introduction
Jacobian

An important DNN characteristic is the $n_L \times n_0$ *input-output Jacobian*

$$J^L := \left\{ \frac{\partial x^L_{j_L}}{\partial x^0_{j_0}} \right\}^n_{j_0, j_L = 1} = D^L W^L \cdots D^1 W^1,$$

$$D^l_n = \mathrm{diag}\{\varphi'((W^l x^{l-1})_{j_l} + b^l_{j_l})\}^n_{j_l=1}.$$

Of particular interest are the singular values of J^L, i.e., the square roots of eigenvalues of the positive definite matrix

$$M^L_n = J^L_n (J^L_n)^T$$

in the large width limit $n \to \infty$, in particular, the *Normalized Counting Measure (NCM)* of eigenvalues of M^L_n (a good DNN characteristic)

$$\nu_{M^L_n} = n^{-1} \sum_{t=1}^{n} \delta_{\lambda_{t, M^L_n}}$$

Modern theory operates also with untrained and even random parameters $\{W_n^l, b_n^l\}_{l=1}^L$ of the DNN architecture, hence, RMT (although non-linear if $\varphi \neq x$).

It is usually assumed that $\{W_n^l, b_n^l\}_{l=1}^L$ are i.i.d. in l and either

(i) $W_n^l = n^{-1/2} X_n^l$, the entries of X_n^l are i.i.d. of zero mean and unit variance and the components of b_n^l are i.i.d. of zero mean and variance σ_b^2;

or

(ii) $W_n^l = O_n^l \in SO(n)$ is the Haar distributed, b_n^l are as in (i).

We will deal with the "untrained random" case (i) in the infinite width limit $n \to \infty$.

The case where

$$D_n^l = \text{diag } D_{j_l}^l = \{\varphi'((n^{-1/2}X^l x^{l-1})_{j_l} + b_{j_l}^l)\}_{j_l=1}^n, \; l = 1, ..., L$$

are replaced by $\mathbf{D}_n^l$, non-random or even random but *independent of* W_n^l (*frozen, quenched*), thus

$$J_n^L \text{ is replaced by } \mathbf{J}_n^L := \mathbf{D}_n^L W_n^L \cdots \mathbf{D}_n^1 W_n^1,$$

is known in RMT, see e.g.

G. Akemann et al 2011, F. Goetze et al 2015, R. Mueller 2002. A particular case with $L = 1$ (single layer) dates back to *Marchenko and P. 1967*.

Claim (*J. Pennington et al, arxiv:1802.09979*):

$$\lim_{n\to\infty} \nu_{M_n^L} = \lim_{n\to\infty} \nu_{\mathbf{M}_n^L}$$

where

$$\mathbf{M}_n^L = \mathbf{J}_n^L (\mathbf{J}_n^L)^T, \ \mathbf{J}_n^L = \mathbf{D}_n^L W_n^L \dots \mathbf{D}_n^1 W_n^1,$$

with

$$(\mathbf{D}_n^l)_{j_l} = \varphi'(n^{-1/2} \mathbf{X}_n^l x_n^{l-1})_{j_l} + b_{j_l}^l), \ l = 1, \dots, L$$

and $\mathbf{X}_n^l$ in $\mathbf{D}_n^l$ $l = 1, \dots, L$ are independent of X_n^l (frozen, quenched) but have the same probability distribution, i.e., $\mathbf{X}_n^L$ and X_n^L are stochastically equivalent.

A reason: analogy with the mean field approximation (ideology) of many-body physics.

Write
$$\mathbf{M}_n^l = \mathbf{D}_n^l X_n^l \mathbf{M}_n^{l-1} (X_n^l)^T \mathbf{D}_n^l$$
and observe that $\mathbf{M}_n^{l-1}$ is independent of X_n^l (a matrix Markov chain).
According to RMT, if the NCM's

$$\nu_{\mathbf{K}_n^l}, \quad \nu_{\mathbf{M}_n^{l-1}}$$

of $\mathbf{K}_n^l := (\mathbf{D}_n^l)^2$ and $\mathbf{M}_n^{l-1}$ converge weakly (with probability 1 if random) as $n \to \infty$ to non random measures $\nu_{\mathbf{K}^l}$ and $\nu_{\mathbf{M}^{l-1}}$, then the same holds for $\nu_{\mathbf{M}_n^l}$ and its non random a.s. limit

$$\nu_{\mathbf{M}^l} = \lim_{n\to\infty} \nu_{\mathbf{M}_n^l}$$

is related to $\nu_{\mathbf{K}^l}$ and $\nu_{\mathbf{M}^{l-1}}$ via an analytic procedure (RMT, FP):

$$\nu_{\mathbf{M}^l} = \nu_{\mathbf{K}^l} \diamond \nu_{\mathbf{M}^{l-1}} \Rightarrow \nu_{\mathbf{M}^l} = \nu_{\mathbf{K}^l} \diamond \cdots \diamond \nu_{\mathbf{K}^1}.$$

Main Result

Our work provides a rigorous proof of the claim by updating the conventional RMT techniques (see, e.g. *L.Pastur, M. Shcherbina, Eigenvalue Distribution of Large Random Matrices, AMS, 2011*).

L.Pastur, On random matrices arising in deep neural networks: Gaussian case, Pure and Appl. Funct. Anal. **5** 1395-1424 (2020), arxiv.org:2001.06188

L.Pastur and V. Slavin, On random matrices arising in deep neural networks: general i.i.d. case, RMTA (to appear), arxiv:2011.11439

L.Pastur, On random matrices arising in deep neural networks: orthogonal case, JMP (to appear), arxiv:2201.04543.

We have

$$M_n^l = D_n^l\, W_n^l\, M_n^{l-1}(W_n^l)^T\, D_n^l,\ W_n^l = n^{-1/2}X_n^l$$

Denote

$$M_n^{l-1} =: R_n^l = (S_n^l)^2,\ K_n^l = (D_n^l)^2,$$

write, omitting the superindex l,

$$M_n = (D_n W_n S_n)(D_n W_n S_n)^T$$

and introduce

$$\mathcal{M}_n = (D_n W_n S_n)^T (D_n W_n S_n = S_n W_n^T K_n W_n S_n.$$

It is important that $\mathcal{M}_n$ and M_n have the same NCM's.

Recall that $W_n = n^{-1/2} X_n$ where the entries $\{X_{j\alpha}\}_{j,\alpha=1}^{n}$ of X_n are i.i.d. random variables with zero mean, unit variance and finite fourth moment $m_4 < \infty$ and that

$$K_n = \{K_{jn}\delta_{jk}\}_{j,k=1}^{n}, \; K_{jn} = (\varphi'(n^{-1/2}(X_n x_n)_j + b_j))^2$$

is diagonal and write $\mathcal{M}_n$ as

$$\mathcal{M}_n = \sum_{j=1}^{n} K_{jn} L_j, \;\; L_j = Y_j \otimes Y_j,$$
$$Y_j = n^{-1/2} S_n X_j, \; X_j = \{X_{j\alpha}\}_{\alpha=1}^{n} \in \mathbb{R}^n,$$

i.e., as the sum of rank-one and independent matrices.

Proof (Outline)

Reduction to the Expectation

Use the martingale-difference techniques to get the bounds:

(i) for any n-independent interval $\Delta \in \mathbb{R}$

$$\mathbf{E}\{|\nu_{\mathcal{M}_n}(\Delta) - \mathbf{E}\{\nu_{\mathcal{M}_n}(\Delta)\}|^4\} \leq C_1/n^2,$$

(ii) for the resolvent $G(z) = (\mathcal{M}_n - z)^{-1}$, any $n \times n$ matrix A and $\xi > 0$

$$\mathbf{Var}\{s_n(\xi)\} \leq C_2||A||^2/n\xi^2, \ s_n(\xi) = n^{-1}\mathrm{Tr}AG(-\xi).$$

Bound (i) and the Borel-Cantelli lemma reduce the problem on random $\nu_{\mathcal{M}_n}$ to that on $\overline{\nu}_{\mathcal{M}_n} = \mathbf{E}\{\nu_{\mathcal{M}_n}\}$ and then, by spectral theorem,

$$f_{\mathcal{M}_n}(z) := \int_0^\infty \frac{\overline{\nu}_{\mathcal{M}_n}(d\lambda)}{\lambda - z} = \mathbf{E}\{n^{-1}\mathrm{Tr}G(z)\},$$

showing that it suffices to find $\lim_{n\to\infty} f_{\mathcal{M}_n}(z)$ uniform on a finite interval of $\mathbb{C} \setminus \mathbb{R}_+$.

Proof (Outline)

Rank-one Formula

Use linear agebra for $A = B + KL_Y$, A and B $n \times n$ hermitian with the resolvents $G_A(z)$ and $G_B(z)$, K real and the rank one $L_Y = Y \times Y$ to write the *rank-one perturbation formula*

$$(i) \quad G_A(z) = G_B(z) - \frac{KG_B(z)L_YG_B(z)}{1 + K(G_B(z)Y, Y)}, \ \Im z \neq 0$$

implying for any $n \times n$ matrix C

$$(ii) \quad n^{-1}\mathrm{Tr}G_A(z)C - n^{-1}\mathrm{Tr}G_B(z)C = -\frac{1}{n} \cdot \frac{K(G_B(z)CG_B(z)Y, Y)}{1 + K(G_B(z)Y, Y)}$$

and for positive definite A, B and $K \geq 0$

$$(iii) \quad |n^{-1}\mathrm{Tr}G_A(-\xi)C - n^{-1}\mathrm{Tr}G_B(-\xi)C| \leq ||C||/n\xi, \ \xi > 0.$$

The rank-one formula (i) and the resolvent identity for the pair $(\mathcal{M}_, 0)$

$$G = -z^{-1} + z^{-1} G\mathcal{M}_n = -z^{-1} + z^{-1} \sum_{j=1}^{n} K_{jn} G L_j$$

lead to the basic formula:

$$G(z) = -z^{-1} + z^{-1} \sum_{j=1}^{n} \frac{K_{jn}}{1 + K_{jn} a_{jn}(z)} G_j(z) L_j,$$
$$G_j = G|_{K_{jn}=0}, \ a_{jn} = (G_j Y_j, Y_j).$$

where K_{jn} and Y_j (hence L_j) are independent of G_j (separation of variables!)

To find $f_{\mathcal{M}_n}(z) = \mathbf{E}\{n^{-1}\mathrm{Tr} G(z)\}$ for $n \to \infty$ we can make in the r.h.s. of the formula any change such that the error $\mathcal{E}_n$ satisfies

$$\blacktriangleright \quad \mathcal{E}\{n^{-1}|\mathrm{Tr}\,\mathcal{E}_n|\} = o(1),\ n \to \infty \quad \blacktriangleleft$$

(qf) Denote $\mathbf{E}_j\{\dots\}$ the (conditional) expectation over X_j and observe that for $Y_j = n^{-1/2}SX_j$ and any X_j-independent and bounded A

$$\mathbf{E}_j\{(AY_j, Y_j)\} = n^{-1}\mathrm{Tr} R_n A,\ R_n = S_n^2,$$

$$\mathbf{Var}_j\{(AY_j, Y_j)\} \le (m_4 + 1)||A||^2(n^{-1}\mathrm{Tr} R^2)/n;$$

(ro) $n^{-1}\mathrm{Tr} AG_j = n^{-1}\mathrm{Tr} AG + \varepsilon_n,\ G_j = G|_{K_{jn}=0}$ by formulas (ii)-(iii);

(ml) $\mathbf{Var}_j\{n^{-1}\mathrm{Tr} AG\} \le C||A||^2/n\xi^2$ by the martingale-type bound.

Use the above facts to replace in the basic formula:

$$a_{jn} := (G_j Y_j, Y_j) \overset{\text{qf}}{\Longrightarrow} n^{-1} \mathrm{Tr}\, G_j R \overset{\text{ro}}{\Longrightarrow} n^{-1} \mathrm{Tr}\, GR =: h_n \overset{\text{ml}}{\Longrightarrow} n^{-1} \mathbf{E}\{h_n\} =: \overline{h}_n$$

$$G_j L_j \overset{\text{qf}}{\Longrightarrow} n^{-1} G_j R \overset{\text{ro}}{\Longrightarrow} n^{-1} GR,$$

yielding

$$G = -z^{-1} + z_{-1} k_n GR + \mathcal{E}_n, \quad (*)$$

with

$$k_n = \sum_{j=1}^{n} \frac{K_{jn}}{1 + K_{jn}\overline{h}_n} \quad \text{!! } \textit{DNN to RMT} \text{!!}$$

$$\overset{LLN}{\Longrightarrow} k = \int_0^\infty \frac{\lambda \nu_K(d\lambda)}{\lambda h + 1} \quad (\mathrm{I})$$

Proof (Outline)

Derivation of Main Formulas

Apply then the operation $\mathbf{E}\{n^{-1}\mathrm{Tr}\ ...\}$ to $(*)$ and get

$$f_{\mathcal{M}_n}(z) = -z^{-1} + z^{-1}\overline{k}_n(z)\overline{h}_n(z) + \overline{\varepsilon}_n,$$

hence, after the (sub)limit $n \to \infty$

$$f_{\mathcal{M}}(z) = -z^{-1} + z^{-1}k(z)h(z) \quad (0).$$

Next, we have from $(*)$

$$G(z) = \mathcal{G}(z) + \mathcal{E}_n, \ \mathcal{G}(z) = (\overline{k}_n(z)R - z)^{-1}.$$

Multiply it by R, apply $\mathbf{E}\{n^{-1}\mathrm{Tr} \, ...\}$ and use the independence of R, hence $\mathcal{G}$, on X_n to obtain

$$\overline{h}_n(z) = \int_0^\infty \frac{\lambda \nu_{R_n}(d\lambda)}{\lambda \overline{k}_n(z) - z} + \overline{\varepsilon}_n,$$

and upon the (sub)limit $n \to \infty$

$$h(z) = \int_0^\infty \frac{\lambda \nu_R(d\lambda)}{k(z)\lambda - z}, \quad \text{(II)}.$$

The system

$$k(z) = \int_0^\infty \frac{\lambda \nu_K(d\lambda)}{h(z)\lambda + 1}, \quad (\mathrm{I}),$$
$$h(z) = \int_0^\infty \frac{\lambda \nu_R(d\lambda)}{k(z)\lambda - z}, \quad (\mathrm{II}).$$

is uniquely solvable in an appropriate class of functions analytic in $\mathbb{C} \setminus \mathbb{R}_+$, thereby determines uniquely via

$$(0) \quad f_{\mathcal{M}}(z) = -z^{-1} + z^{-1} k(z) h(z).$$

the Stieltjes transform $f_M = f_{\mathcal{M}}$ of ν_M, hence, ν_M.

The system is equivalent to the result by Pennington et al, expressed in the free probability terms.

Numerical Results

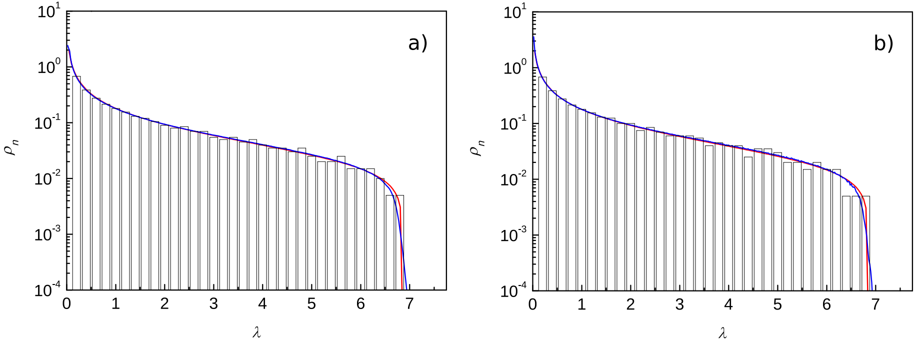

Fig. 1: The eigenvalue density (in the semi-log scale) of the random matrix M_n^L for the Gaussian weights and biases, $L = 2, n = 10^3$. The histograms are the sample densities, the blue lines are arithmetic means ρ_n of $N = 10^3$ samples and the red lines are numerical solutions of the system. a) $\varphi(x) = x$, linear activation function (RMT); b) φ=HardTanh.

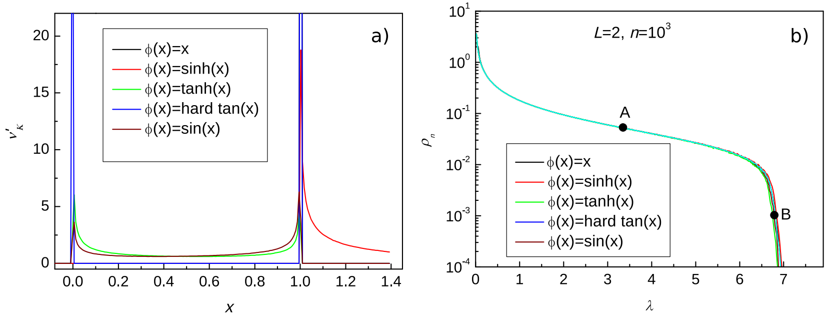

Fig. 2: a) The density ν'_K of the measure ν_K for the indicated activation functions and the Gaussian weights and biases. b) The arithmetic means ρ_n (in the semi-log scale) of the sample eigenvalue densities of $M^2_{10^3}$ over $N = 10^4$ samples for all indicated φ (macroscopic universality).

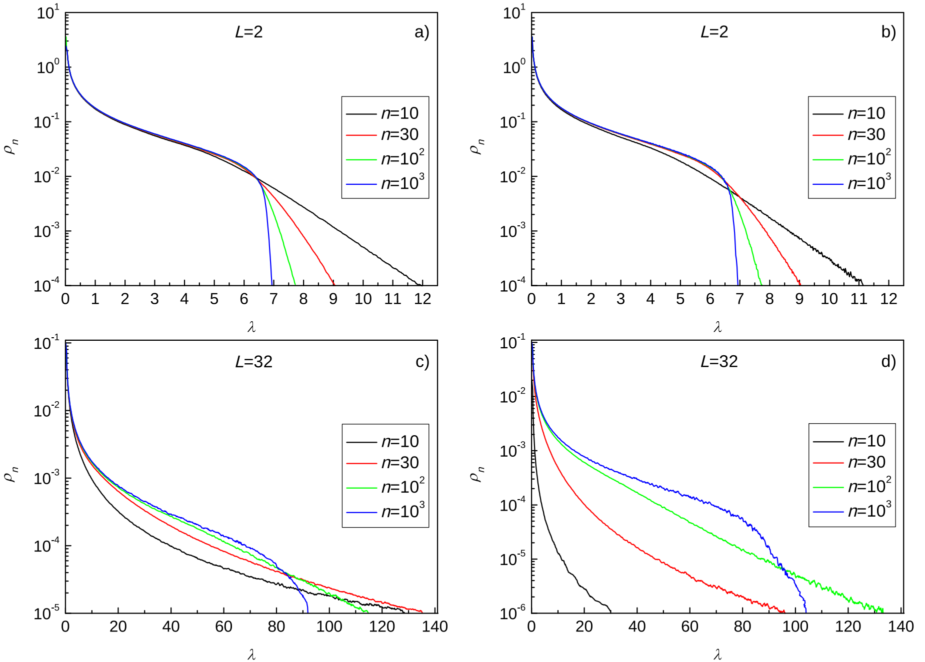
L=2
a)
b)
L=32
c)
d)
n=10
n=30
n=10^2
n=10^3
ρ_n
λ

Figure 3: The arithmetic means ρ_n (in the semi-log scale) of the sample eigenvalue densities of M_n^L with Gaussian weights and biases for various L, n and φ. obtained by averaging over $N = 10^7$ samples for $n = 10,\ 30$, $N = 10^6$ samples for $n = 10^2$ and $N = 10^4$ samples for $n = 10^3$.

The "rows" describe the variation of ρ_n in n and φ for a fixed $L = 2, 32$, while the "columns" describe the variation of ρ_n in n and L for a fixed φ, the linear or the HardTanh. We observe the similarity ("universality") of curves corresponding to different φ', the stronger dependence of curves on n and stronger fluctuations in L, especially near the upper edge of the support and for the (non-smooth) HardTanh φ.

Summary

RMT admits an extension to (untrained) DNN

In particular

- Allowing for the justification of the analog of the mean field approximation
- Extending the macroscopic universality

To YAN

HAPPY BIRTHDAY!

MANY HAPPY RETURNS!