Non linear random matrix models.

L. Benigni, S. Péché,
Université de Paris-Paris 7

Integrability and Randomness in Mathematical Physics and Geometry.

04/11/2019

Plan

- I. Motivations and existing results.
- II. Model and results.
- III. Some ideas of the proof.

Some motivations:

Neural network : let $x \in \mathbb{R}^n$ be an input column vector, W_1, W_2 be $p \times n$ matrices.
The neural network output is: $s = W_2 f(W_1 x)$, where $f(x) = \max(0, x)$ is applied pointwise.
Multistage architecture: alternated layers with both linear/non linear such fonctionnals $s = W_3 \max(0, W_2 \max(0, W_1 x))...$

W_1, W_2 are the (synaptic) weights to be learned, with e.g. stochastic gradient descent (or other). There are multiple choices for the "activation function" f

- $f(x) = \max(0, x)$ known as ReLU activation function good choice in general (accelerates learning but the process can die)

- sigmoid function $f(x) = (1 + e^{-x})^{-1}$, $f(x) = \tanh(x)$...

To understand how it works: make it random

Random matrices and neural networks:

Let X be a deterministic $n \times p$ matrix of input data, Y be a $d \times p$ matrix (the target dataset).

Let W be a $m \times n$ random matrix with i.i.d. entries: W is the random weight matrix.

Let B be a $m \times d$ matrix. Set $z = B^* f(WX)$.

Aim: minimize

$$\frac{1}{p}\sum_{i=1}^{p}(z_i - y_i)^2 + \gamma ||B||_F^2,$$

for some regularisation factor γ.

Then the optimal B is:

$$p^{-1}M(p^{-1}M^*M + \gamma I)^{-1}Y^*, \text{ where } M = f(WX).$$

The performance depends on the spectral measure of $G = p^{-1}M^*M$.

Random matrices and neural networks II:

Theorem Louart, Liao, Couilet (18)

Assume that:
-W is sub Gaussian ($W_{ij} = g(\mathcal{N}(0,1)_{ij})$ for a Lipschitz function g)
-f is Lipschitz continuous
-m, n, p grow to infinity in the same way (m/n bounded from above and below).
The empirical eigenvalue distribution of $1/pM^*M$ has the same limit as $\bar{\mu}$ defined through its Stieltjes transform by

$$m_{\bar{\mu}}(z) = \frac{1}{p}\mathrm{Tr}\left(\frac{n}{p}\frac{\overline{G}}{1+s(z)} - zI_m\right)^{-1} \quad \text{with} \quad \overline{G} = \mathbb{E}\left[M^*M/p\right]$$

and $s(z)$ is the solution such that $\mathrm{Im}\, s(z) > 0$ of

$$s(z) = \frac{1}{p}\mathrm{Tr}\left(\overline{G}\left(\frac{n}{p}\frac{\overline{G}}{1+s(z)} - zI_m\right)^{-1}\right)$$

Some comments

- The dependence in f comes from the deterministic matrix $\overline{G}$.

- Let T be a deterministic matrix such that $TT^* = \overline{G}$. The limiting e.e.d. is the same as that of a sample covariance matrix with general population of type TXX^*T^*/p (Silverstein, Bai (95)).

- The limit is non universal. See Louart, Liao, Couillet for the effect of the fourth moments of the distribution for the efficiency of the neural networks, as well as that of spikes.

Fully random case

Pennington Worrah (17) consider the model:

W resp. X, is a $m \times n$ (resp. $n \times p$) random matrix with i.i.d. Gaussian $\mathcal{N}(0,1)$ matrices

Y also random and independent of X

Set

$$M = \left(f\Big(\frac{WX}{\sqrt{n}}\Big) \right) ; \quad G = \frac{1}{p} M^* M.$$

Similar minimization problem. The limiting distribution of G describes the performance of learning.

Theorem Pennington Worrah (17)

- There exists a limiting distribution μ_f provided $m/n \to \psi$ and $m/p \to \phi$.
- The Stieltjes transform of μ_f satisfies a quartic fixed point equation.

The model

Assume that

- the entries W_{ij}, X_{ij} are independent random variables which are centered and of variance 1, and

$$\mathbb{P}\left(|W_{11}| > t\right) \leqslant e^{-\vartheta_w t^\alpha} \quad \text{and} \quad \mathbb{P}\left(|X_{11}| > t\right) \leqslant e^{-\vartheta_x t^\alpha}, \alpha > 1.$$

- f is a real analytic function such that

$$\int f(x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx = 0.$$

- $m/n \to \psi > 0$ and $m/p \to \phi > 0$ as $n \to \infty$.

Then the e.e.d. of G converges to a probability distribution μ_f.

What is μ_f?

Same as Pennington Worrah : Set

$$\theta_1(f) = \int f^2(x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \quad \text{and} \quad \theta_2(f) = \left(\int f'(x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \right)^2 .$$

The measure μ_f is characterized through a self-consistent equation for its Stieljes transform S defined for $z \in \mathbb{C} \setminus \mathbb{R}$ by

$$S(z) := \int \frac{d\mu_f(x)}{x - z}, \quad \text{denote also} \quad H(z) := \frac{\psi - 1}{\psi} + \frac{z}{\psi} S(z),$$

$$H_\phi(z) := 1 - \phi + \phi H(z) \quad \text{and} \quad H_\psi(z) := 1 - \psi + \psi H(z)$$

Then

$$H(z) = 1 + \frac{H_\phi(z) H_\psi(z)(\theta_1(f) - \theta_2(f))}{\psi z} + \frac{H_\phi(z) H_\psi(z) \theta_2(f)}{\psi z - H_\phi(z) H_\psi(z) \theta_2(f)}.$$

Some remarks

- The dependence in f lies in the two parameters θ_1 and θ_2 only.

- If $\theta_2 = 0$, one recovers the Marcenko-Pastur distribution.

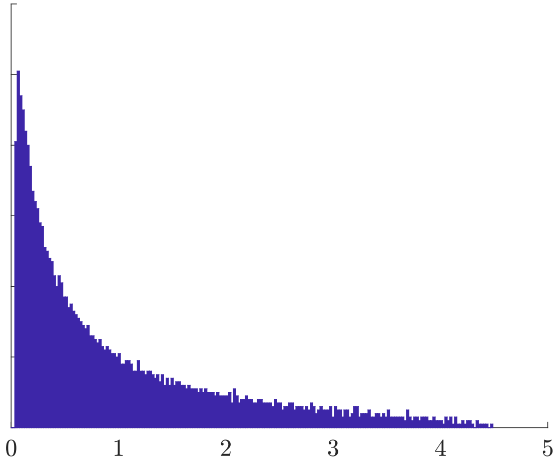
- If $\theta_2 = \theta_1$, μ is the same as the limiting distribution as that of a product Wishart matrix ZZ^*/p with $Z = WX$ (Dupic, Castillo (14))

- In other cases, some kind of interpolation.

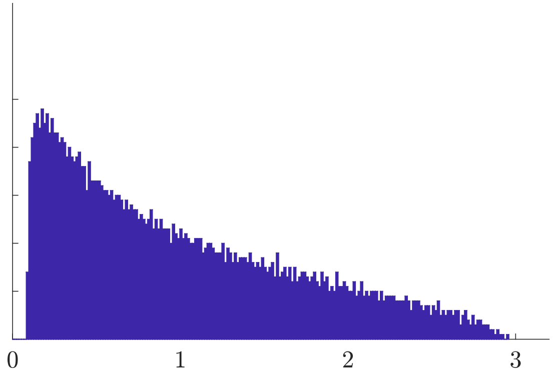
In all cases, we obtain a quartic equation for the Stieltjes transform.

Application: choice for the activation function.

Examples

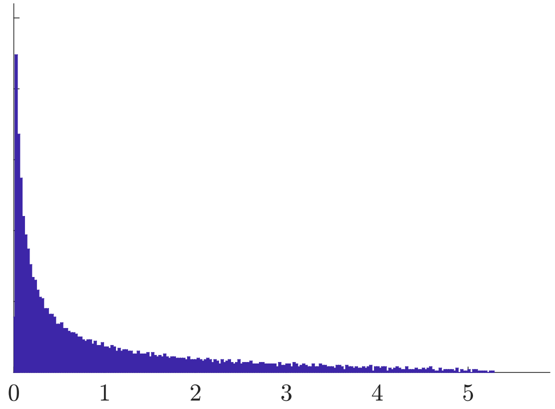


$$f(x) = \frac{x^3}{\sqrt{15}} \quad \psi = 1 \quad \phi = .6$$



$$f(x) = \frac{\cos(x) - \frac{1}{\sqrt{e}}}{\sqrt{\frac{1}{2}\left(1 + \frac{1}{e^2}\right) - \frac{1}{e}}} \quad \psi = 1 \quad \phi = 0.5$$

$$f(x) = \frac{\sin(x)}{\sqrt{\frac{1}{2}\left(1 - \frac{1}{e^2}\right)}} \quad \psi = 1.2 \quad \phi = 0.9$$



In addition the largest eigenvalue sticks to the support of the probability distribution μ_f.

More than one layer

Assume that $X, W^{(i)}, i = 1, \ldots, L$ are independent random matrices ($W^{(i)}$ of size $m_i \times m_{i-1}$) such that the entries satisfy the same decay assumption as before. Set

$$M^{(L)} = \left(f \left(\frac{W^{(L)} M^{(L-1)}}{\sqrt{n \text{Var} M_{12}^{(L-1)}}} \right) \right), \quad M^{(0)} = X; \quad 1 \leq L \leq L_0$$

$$G^{(L)} = \frac{1}{p} M^{(L)} M^{(L)^*}.$$

If $\theta_2(f) = 0$ and f is real analytic and bounded,then the limiting e.e.d. after layer L is the Marcenko-Pastur distribution (whose shape parameter is m_L/n).

Remarks: -all this has been conjectured by Pennington and Worrah.
This result is important for the choice of the activation function (batch normalization).

Some ideas of the proof

The proof is a simple moment method when f is a polynomial. Assume f is an odd monomial: $f(x) = x^k$ and one wants to compute $\mathbb{E}\mathrm{Tr}G^q$ for some given integer q.

Developp the whole trace in terms of W and X entries

$$\frac{1}{m}\mathbb{E}\left[\mathrm{Tr}G^q\right] =$$
$$\frac{1}{mp^q n^{kq}}\mathbb{E}\sum_{i_1,\ldots,i_q}^{m}\sum_{j_1,\ldots,j_q}^{p}\sum_{\substack{\ell_1^1,\ldots\ell_k^1\\ \cdots\\ \ell_1^{2q}\ldots\ell_k^{2q}}}^{n}\prod_{p=1}^{k}W_{i_1\ell_p^1}X_{\ell_p^1 j_1}\prod_{p=1}^{k}W_{i_2\ell_p^2}X_{\ell_p^2 j_1}\cdots\prod_{p=1}^{k}W_{i_1\ell_p^{2q}}X_{\ell_p^{2q} j_q}.$$

Encode each summand into a graph as follows: assume first that the i_k, j_k's are pairwise distinct. We encode the i, j labels into a black graph, decorated by blue edges :

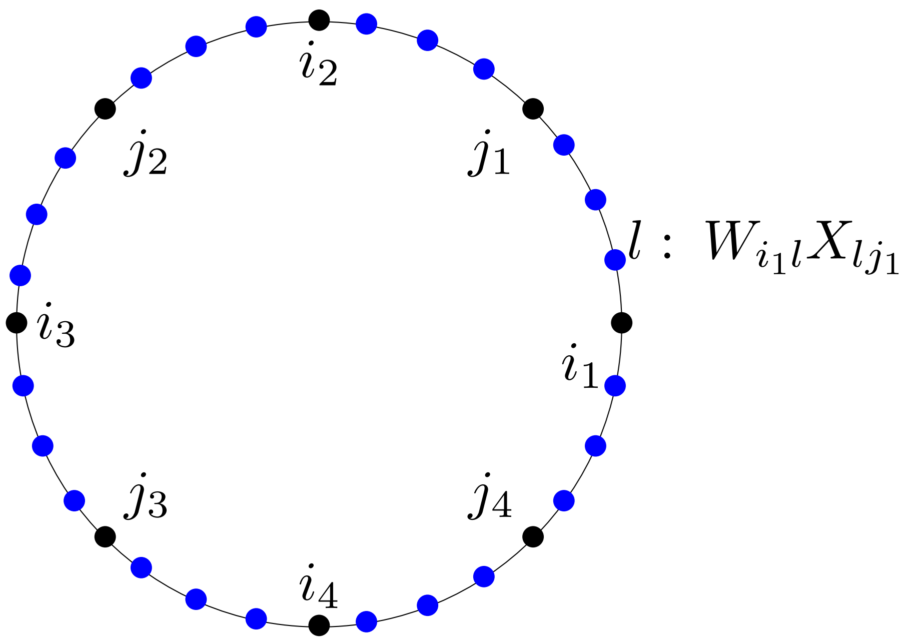


Figure 1: The basic cycle with $2q$ edges. There are k blue vertices between subsequent two black nodes on the main cycle.

A blue point bears an l-label: as each W_{il} and X_{lj} entry has to arise at least twice in the expected trace: the blue points need to be matched.

:: Proof

Typical matchings

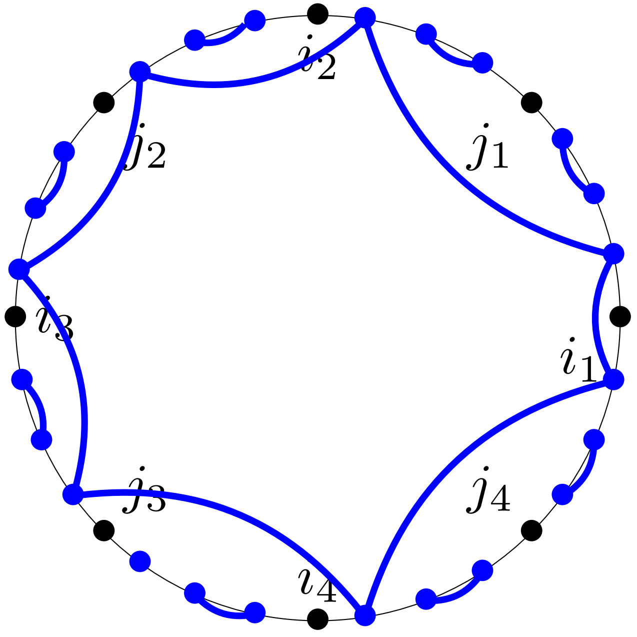
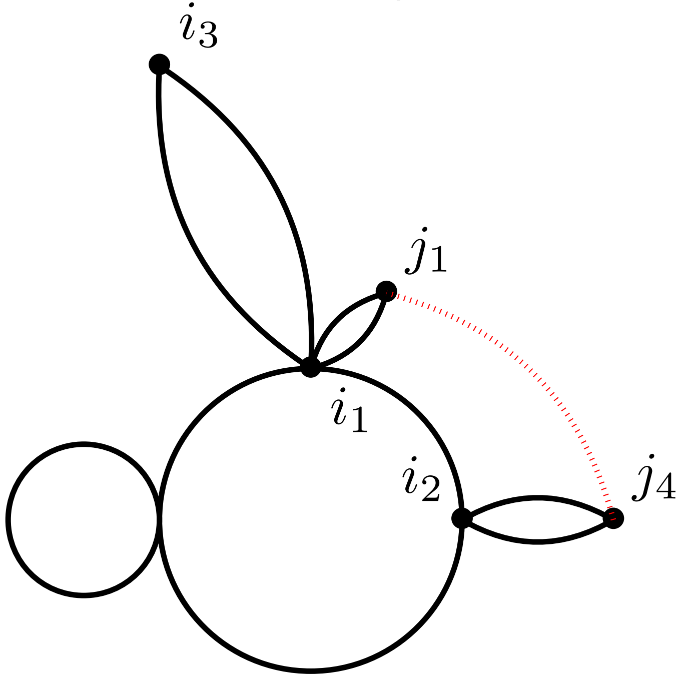


Figure 2: The basic cycle with $2q$ edges: one cycle and perfect matches inside "niches".

Other matchings give a negligible contribution.

Typical black graphs

One can identify some i-indices or j-indices: Any black graph contributing to the expected trace is a "cactus graph" i.e. a tree of cycles (of even length).



A typical matching for such a graph is for each black cycle: a full blue cycle and perfect matching inside niches, except when the black cycle has length 2 (perfect matching of the $2k$ blue edges without a cycle).

Moments

Let $\mathcal{A}(q, I_i, I_j, b)$ denote the number of such cactus graphs which have been obtained from the $2q$ cycle by identifying I_i (resp. I_j) i-vertices (resp. j-vertices) and with b cycles of length 2.

Moments:

$$\int x^q d\mu_f = \sum_{I_i, I_j = 0}^{q} \sum_{b=0}^{I_i + I_j + 1} \mathcal{A}(q, I_i, I_j, b) \theta_1(f)^b \theta_2(f)^{q-b} \psi^{I_i + 1 - q} \phi^{I_j}$$

Remarks: -f is a monomial here but the moment depend on f only through the parameters θ_1 and θ_2.
-If $\theta_2 = 0$ one recovers the number of fat trees (and Narayana numbers) from Chen Yan Yang (08).
-If $\theta_1 = \theta_2$, simple sum over all such graphs as for $(WX)(WX)^*$.

Extension

Case of an even monomial: by centering the monomial w.r.t. the Gaussian distribution, this amounts to ban perfect matchings inside each niche. Thus negligible except those from graphs with cycles of length 2 only.

Arbitrary polynomial: long cycle only odd monomials contribute; cycles of length 2 monomials are both odd or both even.

Extension to real analytic functions f: by Taylor approximation.

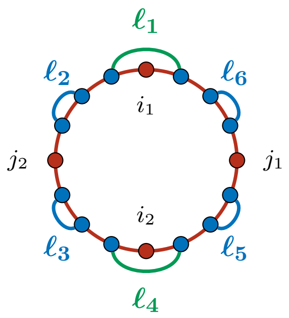
Largest eigenvalue: can push the argument up to q in the order of $\ln n$, which is enough: μ_f has compact support, we call u_+ the top edge. Then one has

$$\mathbb{P}(\lambda_{max} \geq u_+(1+\eta)) \to 0 \quad \forall \eta > 0.$$

Multiple layers: ideas of the proof

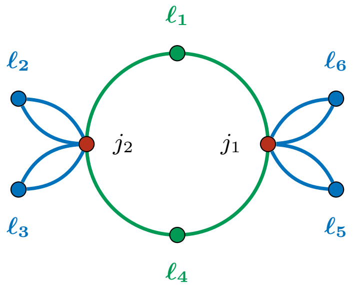
We explain some ideas of the proof for $L = 2$. Again for a monomial develop the trace: then X is first replaced with $M^{(1)}$ whose entries are not independent.

Assume the i and j indices are pairwise distinct. Match the $W^{(2)}$ entries and consider the induced graph on (j, l) indices.



Corresponding moment:

$$M_{\ell_1 j_1} M^2_{\ell_6 j_1} M^2_{\ell_5 j_1} M_{\ell_4 j_1} M_{\ell_4 j_2} M^2_{\ell_3 j_2} M^2_{\ell_2 j_2} M_{\ell_1 j_2}$$



Multiple layers: ideas of the proof II

A typical matching on the $W^{(2)}$ entries produces an admissible induced graph (or is negligible):

-There is a single edge linking two niches adjacent to the same i-labeled vertex which we call a bridge (green on the figure).

-Remaining edges inside a niche are matched according to a perfect matching.

- We can add identifications between bridges only.

Note that the graph can be a forest of such graphs.
Going through more layers results in multiplying the "flowers".

Conclusion

- The method does not work for $f(WX)$, W deterministic or $f(X)$ Wigner (combinatorics is different). In these cases, no universality.

- Main open problem: it does not work for the $\max$ function.

- Complex case and f polynomial: extremal cases only (Marcenko-Pastur or product Wishart) if $\mathbb{E}W_{il}^2 = 0$.

- Possible outliers: models can be determinantal in the complex case, e.g. products of Ginibre matrices (Kuijlaars, Zhang (14); Akeman Burda Kieburg (13). The largest eigenvalue can exhibit a phase transition (Liu, Wang, Wang (18))