Electron. Commun. Probab. **24** (2019), no. 66, 1–7.
https://doi.org/10.1214/19-ECP262
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

A note on the Pennington-Worah distribution*

S. Péché†

Abstract

This paper is concerned with a new expression of the so-called Pennington-Worah distribution, characterizing the asymptotic empirical eigenvalue distribution of some non linear random matrix ensembles. More precisely consider $M = \frac{1}{m}YY^*$ with $Y = f(WX)$ where W and X are random rectangular matrices with i.i.d. centered entries. The function f is applied pointwise and can be seen as an activation function in (random) neural networks. The asymptotic empirical distribution of this ensemble has been computed in [16] and [3]. Here it is related to the Marcenko-Pastur distribution and information plus noise matrices.

Keywords: random matrices; machine learning.
AMS MSC 2010: 60E05.
Submitted to ECP on May 21, 2019, final version accepted on August 21, 2019.

1 Introduction

The scope of this article is to describe the limiting empirical eigenvalue distribution (e.e.d.) of some non linear random matrix ensembles considered in [16]. Such ensembles have been introduced as new approaches to understand deep learning using the theory of random matrices: we refer the reader to the above cited article as well as [9], [8], [14], and [13] for a more complete introduction to the subject. These non linear random matrix ensembles can be defined as follows:

Consider a real random matrix $X \in \mathbb{R}^{n_0 \times m}$ with $i.i.d$ elements with distribution ν_1. Let also $W \in \mathbb{R}^{n_1 \times n_0}$ be a real random matrix with $i.i.d$ entries with distribution ν_2. The entries are normalized so that

$$\int x d\nu_i(x) = 0, \quad \int x^2 d\nu_i(x) = 1 \text{ for } i = 1, 2.$$

There are also some technical assumptions on the tail of these distributions: assume that there exist constants ϑ_w, $\vartheta_x > 0$ and $\alpha > 1$ such that for any $t > 0$

$$\mathbb{P}\left(|W_{11}| > t\right) \leqslant e^{-\vartheta_w t^\alpha} \quad \text{and} \quad \mathbb{P}\left(|X_{11}| > t\right) \leqslant e^{-\vartheta_x t^\alpha}. \tag{1.1}$$

Regarding the (activation) function f, one assumes that there exist positive constants C_f and c_f and $A_0 > 0$ such that for any $A \geqslant A_0$ and any $n \in \mathbb{N}$

$$\sup_{x \in [-A, A]} |f^{(n)}(x)| \leqslant C_f A^{c_f n}. \tag{1.2}$$

*S.P. is supported by the Institut Universitaire de France.
†LPSM, Université Paris Diderot. E-mail: peche@lpsm.paris

In particular this implies that f is real analytic.

Define then

$$M = \frac{1}{m} YY^* \in \mathbb{R}^{n_1 \times n_1} \quad \text{with} \quad Y = f\left(\frac{WX}{\sqrt{n_0}}\right) \tag{1.3}$$

where f is applied entrywise. We suppose that the dimensions of both the columns and the rows of each matrix grow together in the following sense: there exist positive constants ϕ and ψ such that

$$\frac{n_0}{m} \xrightarrow[m\to\infty]{} \phi, \quad \frac{n_0}{n_1} \xrightarrow[m\to\infty]{} \psi$$

Denote by $(\lambda_1, \ldots, \lambda_{n_1})$ the eigenvalues of M given by (1.3) and define its e.e.d. by

$$\mu_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{\lambda_1}. \tag{1.4}$$

In order that the entries of Y are roughly centered (using the central limit theorem), we assume that

$$\int f(x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx = 0. \tag{1.5}$$

Last we set:

$$\theta_1(f) = \int f^2(x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \quad \text{and} \quad \theta_2(f) = \left(\int f'(x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \right)^2. \tag{1.6}$$

[3] and [16] have shown the following result.

Theorem 1.1. *There exists a deterministic compactly supported measure μ_f such that $\mu_{n_1}^{(f)} \xrightarrow[n_1\to\infty]{} \mu_f$ weakly in probability. The measure μ_f is characterized through a self-consistent equation for its Stieljes transform G: for $z \in \mathbb{C} \setminus \mathbb{R}$, we set*

$$G(z) := \int \frac{d\mu_f(x)}{x - z}, \qquad H(z) := \frac{\psi - 1}{\psi} + \frac{z}{\psi} G(z),$$
$$H_\phi(z) := 1 - \phi + \phi H(z) \quad \textit{and} \quad H_\psi(z) := 1 - \psi + \psi H(z)$$

We then have the following fourth-order self-consistent equation:

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

where $\theta_1(f)$ and $\theta_2(f)$ are defined in (1.6).

Theorem 1.1 states that the Stieltjes transform G of the distribution μ_f is the solution of a quartic equation depending on $\theta_1(f)$ and $\theta_2(f)$ only. The limit is thus universal.

Remark 1.2. The quartic equation is not exactly the one stated in [16]. The correct statement is given in [3].

When $\theta_2(f) = 0$, the probability distribution μ_f can then be shown to be the Marcenko-Pastur distribution with parameter $c = \frac{\phi}{\psi}$. Indeed the quartic fixed point equation for G then reduces in this case to Marcenko-Pastur [15] fixed point equation:

$$zm(z)^2 + (z - (1 - 1/c))m(z) + 1/c = 0.$$

When $\theta_1(f) = \theta_2(f) = 1$, the probability distribution μ_f can then be shown to coincide with the limiting e.e.d. of the *linear* random matrix ensemble $\frac{1}{n_1} WX(WX)^*$ which has first been computed in [1] (see also [11]). We call such a distribution the product Wishart distribution. In the general case, the limiting e.e.d. μ_f is an interpolation of these two limiting distributions.

Remark 1.3. The result of [3] can be adapted to the case of complex sample covariance matrices $W = W_1 + iW_2$, $X = X_1 + iX_2$ for some independent matrices W_1, W_2 whose entries have distribution ν_2 (and similarly for X_1 and X_2). In that case, the limiting distribution is either the Marcenko-Pastur distribution or the product Wishart one.

The aim of this article is to describe the limiting distribution μ_f, in terms of the two extremal distributions which are the Marcenko-Pastur one and the product Wishart one. First, one can observe, from their definition (1.6), that $\theta_1(f) \geq \theta_2(f)$. To state our main result, we need some more definitions. In addition to the matrices W and X, we consider an additionnal matrix Z:

$$Z \text{ is a } n_1 \times m \text{ Gaussian random matrix with i.i.d. entries } Z_{ij} \sim \mathcal{N}(0,1). \quad (A_1).$$

The three matrices W, X and Z are assumed to be independent.

Theorem 1.4. *The probability distribution μ_f is also the limiting e.e.d. of the information plus noise sample covariance matrix*

$$M = \frac{1}{m}\left(\sqrt{\theta_2(f)}\frac{WX}{\sqrt{n_0}} + \sqrt{\theta_1(f)-\theta_2(f)}Z\right)\left(\sqrt{\theta_2(f)}\frac{WX}{\sqrt{n_0}} + \sqrt{\theta_1(f)-\theta_2(f)}Z\right)^*, \quad (1.7)$$

where W, X (resp. Z) are independent Gaussian random matrices as in (1.1) (resp. as in (A_1)).

Remark 1.5. Theorem 1.4 has a similar flavor to that of [12], where kernel matrices are considered. Indeed, in both cases, the limiting empirical eigenvalue distribution can be related to that of a linear model of random matrices. This is also in the same vein as [14] where the same phenomenon arises.

Remark 1.6. Theorem 1.4 states indeed that μ_f is related to the rectangular free convolution of the pushforward for both the Marchenko-Pastur distribution and the product Wishart distribution (see [6] Chapter 3 e.g.). This can be related to the results of [2].

Remark 1.7. Possible outliers for information plus noise random matrices as in (1.7) have been studied in [5]. We refer the reader to Theorem 4.2 therein: in particular this may suggest when possible outliers may arise for a deformation of a non linear random matrix ensemble of the form

$$\tilde{M} = \frac{1}{m}\left(f\left(\frac{WX}{\sqrt{n_0}}\right) + B\right)\left(f\left(\frac{WX}{\sqrt{n_0}}\right) + B\right)^*,$$

for some (deterministic) matrix B.

The rest of the article is dedicated to the proof of this theorem. The intuition comes from the combinatorial argument we give in subsection 2.2.

2 Proof of Theorem 1.4

In this section, we assume that $\theta_1 - \theta_2 = 1$ (which can be achieved by scaling).

2.1 Stieltjes transforms

Information plus noise matrices have been studied e.g. in [10], [4] and [5]. We refer the reader for more references therein. Consider a $n_1 \times m$ random matrix Z with i.i.d. centered entries of variance 1. Let now A be a (possibly deterministic) matrix such that the e.e.d. of $\frac{1}{m}AA^*$ converges weakly to a probability distribution ν. The information plus noise matrix is then defined as the sample covariance matrix

$$M = \frac{1}{m}(A+Z)(A+Z)^*.$$

We denote by $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{n_1}$ the ordered eigenvalues of M and the associated e.e.d.

$$\mu_{n_1} := \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{\mu_i}.$$

Then [10] show the following.

Proposition 2.1. *Assume that $\frac{m}{n_1} \to c_1 < 1$ as $m \to \infty$. There exists a probability distribution μ such that $\mu_{n_1} \to \mu$ weakly in probability when $n_1 \to \infty$. The Stieltjes transform $m(z) = \int \frac{1}{z-x} d\mu(x), z \in \mathbb{C} \setminus \mathbb{R}$ satisfies the fixed point equation*

$$m(z) = \int \frac{1 - cm(z)}{(1 - cm(z))^2 z - (1-c)(1 - cm(z)) - t} d\nu(t).$$

In other words Proposition 2.1 states that

$$\frac{m(z)}{1 - cm(z)} = G_A(\mathbf{z}), \text{ where } \mathbf{z} = (1 - cm(z))^2 z - (1-c)(1 - cm(z)), \tag{2.1}$$

where

$$G_A(z) := \int \frac{1}{x - z} d\nu(x).$$

We are now choosing A to be the random matrix $\frac{\sqrt{\alpha_2}\, WX}{\sqrt{n_0}}$ for some α_2 to be defined. For ease, let W and X be Gaussian random matrices (with respective size $n_1 \times n_0$ and $n_0 \times m$). Denote by G the asymptotic Stieltjes transform of the e.e.d. of $\alpha_2 WX(WX)^*/m$. From Theorem 1.1, we first observe that $H_\psi(z) = zG(z)$, and $H_\phi(z) = 1 + \frac{\phi}{\psi}(zG(z) - 1)$. Thus the Stieltjes transform of the asymptotic e.e.d. of $\frac{1}{m} WX(WX)^*$ satisfies the following equation:

$$\psi z^2 (zG(z) - 1) = z^2 G(z) \left(1 + \frac{\phi}{\psi}(zG(z) - 1)\right) \alpha_2 (\psi + zG(z) - 1),$$

which can be rewritten setting $c := \frac{\phi}{\psi}$,

$$\alpha_2 c z^2 G^3(z) + \alpha_2 z G^2 \left(1 - c + c(\psi - 1)\right) + G\left(\alpha_2 (1-c)(\psi - 1) - z\psi\right) + \psi = 0. \tag{2.2}$$

Replacing z with $\mathbf{z}$ in (2.2), we then use the fact $G(\mathbf{z}) = \dfrac{m(z)}{1 - cm(z)}$. One can check that the resulting equation is indeed a quartic equation (and not of degree 5), which holds true only because the change of variables $\mathbf{z} = (1 - cm(z))^2 z - (1-c)(1 - cm(z))$ has the same parameter c as in (2.2). After some heavy computations, we obtain when $\alpha_2 = \theta_2$

$$\begin{aligned} &\psi + m(-\psi z - \theta_2(1-c)(1-\psi) + (1-c)\psi) + m^2(\theta_2(1-2c)z + \psi c z - \theta_2(1-c)^2) \\ &+ m^3(\theta_2(1-c)^2 z^2 - 2c(1-c)z\theta_2) - m^4 \theta_2 c^2 z^2 = 0. \end{aligned} \tag{2.3}$$

This is indeed the quartic equation from Theorem 1.1 when $\theta_1 - \theta_2 = 1$. This finishes the proof of Theorem 1.4.

2.2 Moments

Theorem 1.4 has now been proved. We explain the intuition yielding this result. To that aim, we turn back to the moments of the distribution μ_f. Let $q \in \mathbb{N}$ be the order of a moment. Let C_q be the cycle of length $2q$ with vertices labeled $i_1, j_1, i_2, j_2, \ldots, i_q, j_q$ in order. To state the result we need a few definitions.

Definition 2.2. *An admissible graph is a connected graph built up from simple even cycles obtained from C_q by identifying i-vertices and j-vertices. The cycles are joined to another by at most a common vertex and each* red *edge belongs to a unique cycle.*

An example of an admissible graph is given in Figure 1. We recall from [3] the following result. One has that

$$\int x^q d\mu_f(x) = \sum_{I_i,I_j=0}^{q} \sum_{b=0}^{I_i+I_j+1} \mathcal{A}(q, I_i, I_j, b)\theta_1(f)^b \theta_2(f)^{q-b} \psi^{I_i+1-q} \phi^{I_j}, \tag{2.4}$$

where $\mathcal{A}(q, I_i, I_j, b)$ denotes the number of admissible graphs with $2q$ edges, I_i i-identifications, I_j j-identifications and b cycles of size 2.

We now consider the same moment evaluation for the random matrix

$$M = \frac{1}{m}\left(\sqrt{\theta_2(f)}\frac{WX}{\sqrt{n_0}} + Z\right)\left(\sqrt{\theta_2(f)}\frac{WX}{\sqrt{n_0}} + Z\right)^*.$$

The associated spectral moment of order q is then

$$\begin{aligned}&\frac{1}{n_1 m^q}\mathbb{E}\mathrm{Tr}\Big[\left(\sqrt{\theta_2(f)}\frac{WX}{\sqrt{n_0}} + Z\right)\left(\sqrt{\theta_2(f)}\frac{WX}{\sqrt{n_0}} + Z\right)^*\Big]^q \\ &= \frac{1}{n_1 m^q}\sum_{i_1,\dots i_q}\sum_{j_1,\dots j_q}\mathbb{E}\prod_{t=1}^{q} A_{i_t j_t} A_{i_{t+1} j_t},\end{aligned} \tag{2.5}$$

where in the last line $i_{q+1} = i_1$ and $A = \sqrt{\theta_2(f)}\frac{WX}{\sqrt{n_0}} + Z$.

We now use the independence of the three matrices W, X, and Z and the fact that the entries are centered. We observe that expanding any A_{ij} in terms of entries W_{il}, X_{lj}, $l = 1, \dots, n_0$ and Z_{ij}, each such entry has to arise at least twice in the whole summand (so that the contribution to the expectation is not zero). Consider the following encoding: to a possible set of indices $\{i_1, \dots i_q\}$ $\{j_1, \dots j_q\}$ (identifications inside each set are allowed), we draw the graph obtained by connecting i_t to j_t and j_t to i_{t+1}. For ease, edges of this graph are colored red. In all cases, this yields a connected graph which is cyclic. We are then going to consider the contribution of each graph to the expectation: thus one has to assign to each red edge either a "WX" label or a Z label: this comes from the fact that A is the sum of the two corresponding matrices. Finally for each graph, we count the number of possible labellings of the vertices and the combined sum of all these contributions will eventually give the expected trace.

Consider the simple cycle C_q corresponding to the case where all i-indices and j-indices are pairwise distinct: there are no identifications. It is not difficult to check that in this case, the only contributing term comes from assigning a "WX" label to each edge corresponding to A. Indeed it is not difficult to check that if there is at least one Z label the associated expectation is null. Thus one is left with the same contribution as that of the non-linear matrix model in [3]. In particular one has that

$$\mathbb{E}\prod_{t=1}^{q} A_{i_t j_t} A_{i_{t+1} j_t} = \frac{\theta_2^q n_0}{n_0^q},$$

for such a cycle: the l index is necessarily the same along the whole cycle so that each W or X entry arises twice.

Consider now an admissible graph: this graph is then a tree of cycles. Some of these cycles have length 2 (denote by b the number of such cycles) and let c be the number of the remaining cycles (which have an even length ≥ 4 in all cases). Similarly for a cycle of length $2l_1 \geq 4$, the contribution to the expectation is not zero iff all red edges

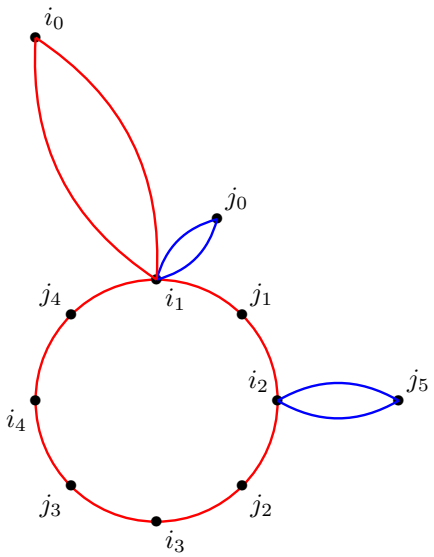

Figure 1: An admissible graph: edges from cycles of length 2 are assigned a Z (blue edge) or WX label (in red) and those of length greater a WX label. Here $b = c = 2$.

are assigned a "WX" label and the l-index is necessarily constant along the cycle. On the other side, for a cycle of length 2 red edges can be assigned any label. However so that the edges all arise with a multiplicity at least 2, the two edges of the cycle bear the same label necessarily.

As a consequence, when the red graph associated to the i-indices and j-indices is admissible, the contribution to the expectation in (2.5) can be easily proved to be given by

$$\theta_2^{q-b}(1+\theta_2)^b \frac{n_0^c}{n_0^{q-b}}.$$

Indeed, in view of [3], the l-index is constant along any cycle of length greater than 3. Counting now the possible labelling of the i and j-vertices: note that i-indices can describe $\{1, 2, \dots, n_1\}$ while j indices run from 1 to m. The final contribution of admissible graphs is then

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

where the error term comes from the number of possible indices i and j only. We note that this is the same as the moment (2.4).

The proof is finished provided we can show that the contribution of non-admissible graphs obtained from C_q by i-identifications and j-identifications is negligible. The arguments can then be copied from [3]. Non admissible graphs can be obtained from admissible ones by making some further identifications in such a way that one does not obtain a tree of cycles. There are then more than one way to run through the graph and some edges may arise more than twice in the summand. This is however compensated by the fact that ones loses for each additional identification a power of m. We skip the detail of the proof. The contribution of non admissible graphs can be proved to be negligible.

References

[1] Akemann, G. Ipsen, J. R. and Kieburg, M. Products of rectangular random matrices: Singular values and progressive scattering. *Phys. Rev. E*, **88**, (2013) 052–118.

[2] Benaych-Georges, F., On a surprising relation between the Marchenko-Pastur law, rectangular and square free convolutions. *Ann. Inst. Henri Poincaré Probab. Stat.*, **46**, no. 3, (2010), 644–652.

[3] Benigni, L. and Peche, S. Eigenvalue distribution of non linear models of matrix ensembles. *arXiv preprint*, (2019).

[4] Capitaine, M. Limiting eigenvectors of outliers for spiked information-plus-noise type matrices. *Lecture Notes in Math. Springer, Séminaire de Probabilités*, **XLIX**, (2018), 119–164.

[5] Capitaine, M. Exact separation phenomenon for the eigenvalues of large information-plus-noise type matrices, and an application to spiked models. *Indiana Univ. Math. J.*, **63**, (2014), 1875–1910.

[6] Capitaine, M. Deformed ensembles, polynomials in random matrices and free probability theory. *Habilitation thesis, HAL Id: tel-01978065, version 1*, (2017).

[7] Cébron, G., Dahlqvist, A. and Male C. Universal constructions for spaces of traffics. *arXiv preprint*, (2016).

[8] Choromanska, A., Henaff, M., Mathieu, M., Ben Arous, G. and LeCun, Y. The loss surfaces of multilayer networks, *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, AISTATS 2015*, (2015).

[9] Cirac, C., Cranmer,K., Daudet, L., Schuld, M., Vogt-Maranto, L. and Zdeborová, L. Machine learning and the physical sciences. *arXiv preprint arXiv:1903.10563*, (2019).

[10] Dozier, B. and Silverstein, J. On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices. *Journal of Multivariate Analysis*, **98**, (2007), 678–694.

[11] Dupic, T. and Castillo, I. P. Spectral density of products of Wishart dilute random matrices. Part I: the dense case. *arXiv preprint*, (2014)

[12] El Karoui, N. The spectrum of kernel random matrices. *Ann. Statist.*, **38**, no. 1 (2010), 1–50.

[13] Hanin, B. and Nica, M. Products of many large random matrices and gradients in deep neural networks. *arXiv preprint*, (2018).

[14] Louart, C., Liao, Z. and Couillet, R. A random matrix approach to neural networks. *Ann. Appl. Probab.*, **28**, no. 2 (2018), 1190–1248.

[15] Marčenko, V. A. and Pastur, L. A. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, **72**, no. 114 (1967), 507–536.

[16] Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. *Advances in Neural Information Processing Systems*, (2017), 2637–2646.

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)
- Secure publication (LOCKSS[1])
- Easy interface (EJMS[2])

Economical model of EJP-ECP

- Non profit, sponsored by IMS[3], BS[4] , ProjectEuclid[5]
- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

[1]LOCKSS: Lots of Copies Keep Stuff Safe `http://www.lockss.org/`
[2]EJMS: Electronic Journal Management System `http://www.vtex.lt/en/ejms.html`
[3]IMS: Institute of Mathematical Statistics `http://www.imstat.org/`
[4]BS: Bernoulli Society `http://www.bernoulli-society.org/`
[5]Project Euclid: `https://projecteuclid.org/`
[6]IMS Open Access Fund: `http://www.imstat.org/publications/open.htm`