Analyzing Random Features via Linearization and the Matrix Dyson Equation

Konstantinos Christopher Tsiolis
Department of Mathematics and Statistics
McGill University
kc.tsiolis@mail.mcgill.ca

Abstract

Inspired by recent work using random matrix theory to study neural networks, we use the linearization trick and the matrix Dyson equation to find the limiting spectral distribution for the covariance matrix of Gaussian random features. Compared to prior work, this approach leads to a much shorter derivation of fixed point equations specifying the Stieltjes transform of the target distribution, which can then be solved numerically. Our empirical results suggest that our method achieves a faithful approximation to the limiting distribution that is closely matched by the empirical spectral distribution as dimension grows.

1 Introduction

Deep neural networks have been the driver of tremendous progress in machine learning over the past decade. They are at the core of advances in computer vision [10, 8, 5], natural language processing [4, 3], and reinforcement learning [11], among many other application areas. However, a comprehensive theoretical explanation of the empirically observed success of deep learning remains elusive. Those analyses which are present in the literature tend to focus on constrained cases that involve many assumptions which do not line up with state-of-the-art models used in practice [16].

A key contributor to the success of deep learning is the ability of neural networks to generalize to unseen data. The theory of generalization in machine learning is well-established under the framework of Probably Approximately Correct (PAC) learning [18]. There exist generalization bounds which capture the gap between training and test error of a learning algorithm [17]. However, these bounds depend on the complexity of the hypothesis space, i.e., the class of candidate functions that are considered in an attempt to fit the training data. Neural networks constitute a rich function class whose complexity has only so far been estimated in limited cases with strong assumptions on the size and architecture of the network. For example, [12] prove a lower bound on the Rademacher complexity of two-layer neural networks with a ReLU activation function.

Recently, a novel line of research has approached the question of generalization from the perspective of random matrix theory. [14] analyze Gaussian random feature models, which resemble neural networks at initialization. Their analysis of the limiting spectral distribution of a kernel matrix related to the output of the random feature model yields a closed-form expression of the training error of such a model in the case of ridge regression. [2] find the limiting spectral distribution for random feature models in the non-Gaussian case using the moment method. [15] find the same limiting distribution using a combination of the resolvent method and the cumulant method. Taking this idea a step further, [1] use techniques from random matrix theory and the neural tangent kernel [9] (a kernel formed by

the Jacobian of a neural network's output with respect to the input) to express the generalization error (test error) of two-layer neural networks in the infinite width and infinite data limit.

In this work, we aim to reproduce the results obtained in [14, 2, 15] for random feature models. We make the simplifying assumption that both the features and weights are Gaussian and that the activation function is the identity. Like [15], we approach this with the resolvent method, but forgo the cumulant expansion in favour of the matrix Dyson equation (see [7] for a comprehensive overview). Solving this equation gives an approximation to the Stieltjes transform of the limiting spectral distribution we are interested in. In our simulations, we show how the approximation we obtain compares to the empirical spectral distribution in high dimensions.

2 Setup

We adopt the same setup as in [14, 2, 15], with some simplifying assumptions. Let $W \in \mathbb{R}^{n_1 \times n_0}$ and $X \in \mathbb{R}^{n_0 \times m}$. The entries of W and X are drawn i.i.d. from a normal distribution with mean 0 and variance σ_w^2 and σ_x^2 respectively. We may view this from the perspective of a one-layer linear neural network with input X, where n_0 is the feature dimension and m is the number of samples, and with weight matrix W, where n_1 is the output dimension. We consider the $n_1 \times m$ matrix $Y = f(\frac{1}{\sqrt{n_0}} WX)$, where f is an activation function applied elementwise. For simplicity, we assume in this work that f is the identity map. Then, the random matrix of interest is the symmetric polynomial $p := \frac{1}{m} YY^T \in \mathbb{R}^{n_1 \times n_1}$.

3 Solving the Matrix Dyson Equation

Before beginning our derivations, we briefly outline our strategy. Recall that for any $z \in \mathbb{H}$ (i.e., $z \in \mathbb{C}$ with $\Im(z) > 0$), the *Stieltjes transform* of a probability measure μ on $\mathbb{R}$, $s(z;\mu)$, is $\int_{\mathbb{R}} \frac{1}{z-x} \mu(dx)$. For the empirical spectral measure of p, which we denote by μ_{n_1}, we have $s(z;\mu_{n_1}) = \operatorname{tr}(zI_{n_1} - p)^{-1}$. The matrix $(zI_{n_1} - p)^{-1}$ is called the *resolvent* of p. Showing that for all $z \in \mathbb{H}$, $s(z;\mu_{n_1})$ converges in probability to $s(z;\mu)$ for some probability measure μ on $\mathbb{R}$ is equivalent to showing that μ_{n_1} converges to μ weakly in probability. (To be more specific, here we are interested in the limit as $n_1 \to \infty$ while the ratios $\phi := \frac{n_0}{m}$ and $\psi := \frac{n_0}{n_1}$ are held constant). Furthermore, for every $x \in \mathbb{R}$, we can recover μ via the *Stieltjes inversion* $\lim_{\eta \to 0} \Im(s(x + i\eta; \mu))$.

Expanding p, we have that $p = \frac{1}{mn_0} WXX^TW^T$. Let $d = n_1 + 2n_0 + m$. We define the linearization $\hat{p} \in \mathbb{R}^{d \times d}$ of p to be

$$\hat{p} = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{\sqrt{n_0}} W \\ 0 & 0 & \frac{1}{\sqrt{m}} X & -I_{n_0} \\ 0 & \frac{1}{\sqrt{m}} X^T & -I_m & 0 \\ \frac{1}{\sqrt{n_0}} W^T & -I_{n_0} & 0 & 0 \end{bmatrix}. \tag{1}$$

We define $\Lambda \in \mathbb{R}^{d \times d}$ with the same block structure as $\hat{p}$ and such that the top-left $n_1 \times n_1$ block is I_{n_1} and all other blocks are zero. We remark that the resolvent of p is equal to the top-left block of $(\Lambda z - \hat{p})^{-1}$, and thus the Stieltjes transform of interest is simply the Stieltjes transform of that block.

Since W and X are assumed to be mean zero, we have that

$$H := \hat{p} - \mathbb{E}[\hat{p}] = \begin{bmatrix} 0 & 0 & 0 & \frac{1}{\sqrt{n_0}} W \\ 0 & 0 & \frac{1}{\sqrt{m}} X & 0 \\ 0 & \frac{1}{\sqrt{m}} X^T & 0 & 0 \\ \frac{1}{\sqrt{n_0}} W^T & 0 & 0 & 0 \end{bmatrix}. \tag{2}$$

For $z \in \mathbb{H}$ (i.e., $z \in \mathbb{C}$ with $\Im(z) > 0$), the matrix Dyson equation is

$$-M = (-\Lambda z + \mathcal{E} + \mathcal{S}(M))^{-1}, \tag{3}$$

where $\mathcal{E} := \mathbb{E}[\hat{p}]$ and $\mathcal{S}(M) := \mathbb{E}_H[HMH]$. We aim to solve this equation for M. An important result in random matrix theory shows concentration of the Stieltjes transform of $(\Lambda z - \hat{p})^{-1}$ around $M(z)$, thus motivating our approach [13]. We state the result below for completeness.

Theorem 1. *If S_{ij} has $\max|S_{ij}| \leq 1/n$, there is $\delta > 0$ so for $z \in \mathbb{H}$, $|\Im z| = \eta > 0$, $||\Pi||_{HS}^2 \leq n$, we have*

$$\mathbb{P}\left(\left|\frac{1}{n}\mathrm{tr}\big(\Pi\big((\Lambda z - \hat{p})^{-1} - M(z)\big)\big)\right| > n^{-1/2}\right) \leq \exp(-\delta\eta^4 n).$$

Now, viewing M as a 4×4 block matrix (in the same fashion as $\hat{p}$), we have

$$HMH = \begin{bmatrix} \frac{1}{n_0}WM^{(44)}W^T & \frac{1}{\sqrt{mn_0}}WM^{(43)}X^T & \frac{1}{\sqrt{mn_0}}WM^{(42)}X & \frac{1}{n_0}WM^{(41)}W \\ \frac{1}{\sqrt{mn_0}}XM^{(34)}W^T & \frac{1}{m}XM^{(33)}X^T & \frac{1}{m}XM^{(32)}X & \frac{1}{\sqrt{mn_0}}XM^{(31)}W \\ \frac{1}{\sqrt{mn_0}}X^TM^{(24)}W^T & \frac{1}{m}X^TM^{(23)}X^T & \frac{1}{m}X^TM^{(22)}X & \frac{1}{\sqrt{mn_0}}X^TM^{(21)}W \\ \frac{1}{n_0}W^TM^{(14)}W^T & \frac{1}{\sqrt{mn_0}}W^TM^{(13)}X^T & \frac{1}{\sqrt{mn_0}}W^TM^{(12)}X & \frac{1}{n_0}W^TM^{(11)}W \end{bmatrix}.$$

We observe that only those blocks which contain two copies of W or two copies of X have nonzero expectation, since the entries of W and X are independent.

For example, consider

$$\mathbb{E}\big[(WM^{(44)}W^T)_{ij}\big] = \mathbb{E}\left[\sum_{k=1}^{n_0}\sum_{l=1}^{n_0} W_{ik}M_{kl}^{(44)}W_{lj}^T\right] = \sum_{k=1}^{n_0}\sum_{l=1}^{n_0} M_{kl}^{(44)}\mathbb{E}[W_{ik}W_{jl}].$$

Since the entries of W are independent, we only obtain a non-zero term in the case $i = j$ and $k = l$, and this term has the form $M_{kk}^{(44)}\sigma_w^2$. This implies that $\mathbb{E}[WM^{(44)}W^T] = \sigma_w^2\mathrm{tr}(M^{(44)})I_{n_1}$.

With this reasoning, we obtain

$$\mathcal{S}(M) = \begin{bmatrix} \frac{1}{n_0}\sigma_w^2\mathrm{tr}(M^{(44)})I_{n_1} & 0 & 0 & \frac{1}{n_0}\sigma_w^2(M^{(41)})^T \\ 0 & \frac{1}{m}\sigma_x^2\mathrm{tr}(M^{(33)})I_{n_0} & \frac{1}{m}\sigma_x^2(M^{(32)})^T & 0 \\ 0 & \frac{1}{m}\sigma_x^2(M^{(23)})^T & \frac{1}{m}\sigma_x^2\mathrm{tr}(M^{(22)})I_m & 0 \\ \frac{1}{n_0}\sigma_w^2(M^{(14)})^T & 0 & 0 & \frac{1}{n_0}\sigma_w^2\mathrm{tr}(M^{(11)})I_{n_0} \end{bmatrix}.$$

Since the solution to the matrix Dyson equation is unique, we assume $M^{(14)} = M^{(23)} = M^{(32)} = M^{(41)} = 0$. We will be justified in doing so if we find a solution.

Plugging into Equation 3, we have

$$-M = \begin{bmatrix} (\frac{1}{n_0}\sigma_w^2\mathrm{tr}(M^{(44)}) - z)I_{n_1} & 0 & 0 & 0 \\ 0 & \frac{1}{m}\sigma_x^2\mathrm{tr}(M^{(33)})I_{n_0} & 0 & -I_{n_0} \\ 0 & 0 & (\frac{1}{m}\sigma_x^2\mathrm{tr}(M^{(22)}) - 1)I_m & 0 \\ 0 & -I_{n_0} & 0 & \frac{1}{n_0}\sigma_w^2\mathrm{tr}(M^{(11)})I_{n_0} \end{bmatrix}^{-1}.$$

To simplify the matrix inversion, let $a = \frac{1}{n_0}\sigma_w^2\mathrm{tr}(M^{(44)}) - z$, $b = \frac{1}{m}\sigma_x^2\mathrm{tr}(M^{(33)})$, $c = \frac{1}{m}\sigma_x^2\mathrm{tr}(M^{(22)}) - 1$, and $d = \frac{1}{n_0}\sigma_w^2\mathrm{tr}(M^{(11)})$. Then, we can simply invert the matrix

$$\begin{bmatrix} -a & 0 & 0 & 0 \\ 0 & -b & 0 & 1 \\ 0 & 0 & -c & 0 \\ 0 & 1 & 0 & -d \end{bmatrix}$$

to obtain fixed point equations for the entries of M.

We have

$$\begin{aligned} M_{ii}^{(11)} &= \frac{bcd - c}{ac(1-bd)} = \frac{c(bd-1)}{ac(1-bd)} = -\frac{1}{a} && \Rightarrow \mathrm{tr}(M^{(11)}) = -\frac{n_1}{a} \\ M_{ii}^{(22)} &= \frac{acd}{ac(1-bd)} = \frac{d}{1-bd} && \Rightarrow \mathrm{tr}(M^{(22)}) = n_0\frac{d}{1-bd} \\ M_{ii}^{(33)} &= \frac{abd - a}{ac(1-bd)} = -\frac{1}{c} && \Rightarrow \mathrm{tr}(M^{(33)}) = -\frac{m}{c} \\ M_{ii}^{(44)} &= \frac{abc}{ac(1-bd)} = \frac{b}{1-bd} && \Rightarrow \mathrm{tr}(M^{(44)}) = n_0\frac{b}{1-bd}. \end{aligned}$$

This leads to the system of equations

$$d = \frac{\gamma_1}{a} \qquad c = \gamma_2 \frac{d}{1-bd} - 1 \qquad b = \frac{\gamma_3}{c} \qquad a = \gamma_4 \frac{b}{1-bd} - z,$$

where

$$\gamma_1 = -\frac{n_1}{n_0}\sigma_w^2 \qquad \gamma_2 = \frac{n_0}{m}\sigma_x^2 \qquad \gamma_3 = -\sigma_x^2 \qquad \gamma_4 = \sigma_w^2.$$

(Notice the constants only depend on the dimension through the fixed ratios $\phi = \frac{n_0}{m}$ and $\psi = \frac{n_0}{n_1}$). This can be simplified into a system of two nonlinear equations in two variables,

$$d = \frac{\gamma_1}{\gamma_4(\frac{b}{1-bd}) - z} \qquad b = \frac{\gamma_3}{\gamma_2(\frac{d}{1-bd}) - 1}.$$

We numerically solve this system using a modified form of fixed point iteration, shown in Algorithm 1, which allows us to recover an approximation to the inverse Stieltjes transform directly.

Algorithm 1: Numerical Computation of the Limiting Distribution

Input: n_0, n_1, m, σ_w, σ_x, λ;
$\gamma_1 \leftarrow -\frac{n_1}{n_0}\sigma_w^2$, $\gamma_2 \leftarrow \frac{n_0}{m}\sigma_x^2$, $\gamma_3 \leftarrow -\sigma_x^2$, $\gamma_4 \leftarrow \sigma_w^2$;
$b \leftarrow 0$, $d \leftarrow 0$;
$\mathcal{I} \leftarrow$ logarithmically spaced values from 10^4 to 10^{-8};
for $k \in \mathcal{I}$ **do**
  $d \leftarrow \frac{\gamma_1}{\gamma_4(\frac{b}{1-bd}) - (\lambda + ki)}$;
  $b \leftarrow \frac{\gamma_3}{\gamma_2(\frac{d}{1-bd}) - 1}$;
end
$s \leftarrow \Im(\frac{1}{\pi}d)$;
Output: s

4 Results

We ran Algorithm 1 for various settings of n_0, n_1, m, σ_w, σ_x. The results are shown in Figure 1. The Jupyter notebook used to generate the plots is available in the supplementary material.

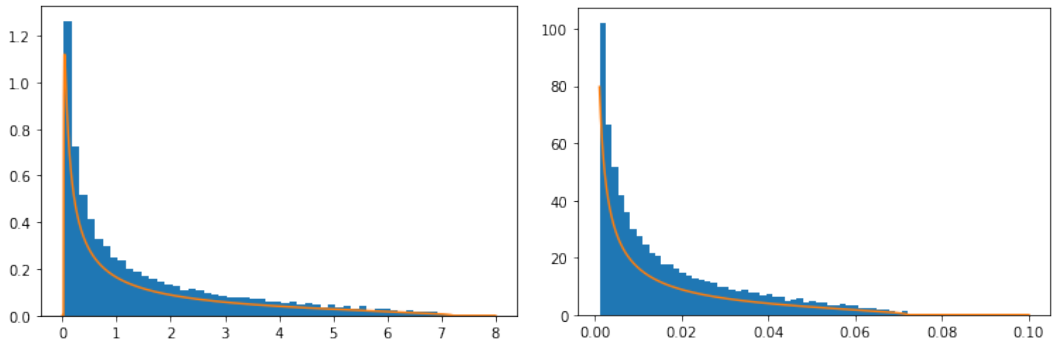
We also aimed to numerically solve the fixed point equation from [14], which was re-written in an equivalent form in [15]. However, due to large numerical errors, we were unable to do so.

5 Discussion

The plots in Figure 1 confirm that the limiting distribution found via the matrix Dyson equation indeed captures the shape of the empirical spectral distribution. As the dimensions n_0, n_1, and m grow, the empirical spectral distribution fits the limiting distribution more and more closely. Thus, our relatively short derivation allows us to capture the spectrum of p without resorting to tedious computation of moments (as in [14] and [2]) or cumulants (as in [15]). As an added bonus, the system of fixed point equations obtained can be numerically solved in straightforward fashion via Algorithm 1, while working with the quartic fixed point equation of [14, 15] leads to overflow.

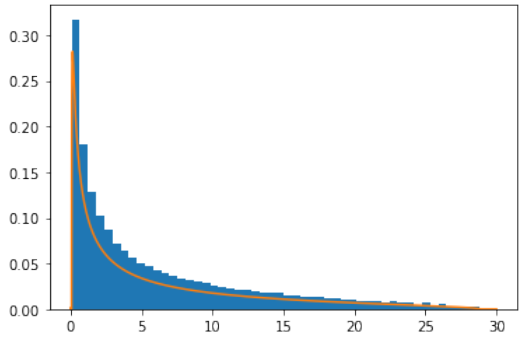
The limiting global law of $(WX)(WX)^T$ has also been studied in [6], and the same self-consistent equation for the Stieltjes transform of μ as in [14, 15] is found. This is then used to derive the distribution in closed form. A key point is that μ has a similar form Marchenko-Pastur law, but is not identical to it, in spite of what the plots in Figure 1 might suggest.

To take these ideas further, one must ask what μ is when f is a non-linear activation function (e.g. ReLU, sigmoid). [14, 2, 15] find that μ depends on f only through two scalar quantities involving

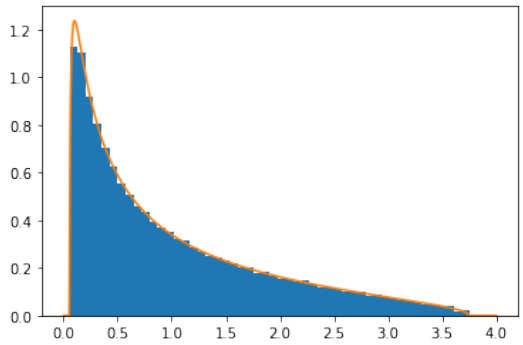


(a) $n_0 = 2000$, $n_1 = 3000$, $m = 4000$, $\sigma_w = 1, \sigma_x = 1$

(b) $n_0 = 2000$, $n_1 = 3000$, $m = 4000$, $\sigma_w = 1, \sigma_x = 0.1$



(c) $n_0 = 2000$, $n_1 = 3000$, $m = 4000$, $\sigma_w = 1, \sigma_x = 2$



(d) $n_0 = 5000$, $n_1 = 2500$, $m = 10\,000$, $\sigma_w = 1, \sigma_x = 1$

Figure 1: Plots of limiting (orange) and empirical (blue) spectral distribution (as generated by Algorithm 1) for different choices of n_0, n_1, m, σ_w, and σ_x. λ values are on the horizontal axis and density is on the vertical axis.

the Gaussian integral. To replicate this result in our framework, one has to reconcile the nonlinearity f with the linearization $\hat{p}$. One proposal is to work with the partial linearization

$$\hat{p} = \begin{bmatrix} 0 & f(WX) \\ \big(f(WX)\big)^T & -I_{n_1} \end{bmatrix},$$

i.e., the linearization trick for a gram matrix AA^T, where we take $A = f(WX)$. We intend to explore this in future work.

6 Conclusion

In this work, we provided an alternative method to determine the limiting spectral distribution of the kernel $\frac{1}{mn_0}(WX)(WX)^T$, where W and X are random matrices each having i.i.d. mean-zero Gaussian entries. We used the linearization trick and expressed the solution to the matrix Dyson equation in terms of two fixed point equations. With these fixed point equations in hand, we used an algorithm to simultaneously compute the Stieltjes transform of the limiting distribution and the inverse transform. Our empirical results show that the distribution obtained indeed closely matched the empirical spectral distribution in large dimension. With this being only a special case of the distribution studied in [14, 2, 15], we intend on extending this analysis to the case where a non-linear activation function f is applied to WX, which simulates a shallow neural network.

7 Acknowledgements

The author wishes to thank Prof. Elliot Paquette for proposing this project and for providing guidance at each step of its development during office hours.

References

[1] B. Adlam and J. Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR, 2020.

[2] L. Benigni and S. Péché. Eigenvalue distribution of nonlinear models of random matrices. *arXiv preprint arXiv:1904.03090*, 2019.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[6] T. Dupic and I. P. Castillo. Spectral density of products of wishart dilute random matrices. part i: the dense case. *arXiv preprint arXiv:1401.7802*, 2014.

[7] L. Erdos. The matrix dyson equation and its applications for random matrices. *arXiv preprint arXiv:1903.10060*, 2019.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[11] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[12] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

[13] E. Paquette. The resolvent method. *MATH 598: Random Matrix Theory*, 2022.

[14] J. Pennington and P. Worah. Nonlinear random matrix theory for deep learning. *Advances in neural information processing systems*, 30, 2017.

[15] V. Piccolo and D. Schröder. Analysis of one-hidden-layer neural networks via the resolvent method. *Advances in Neural Information Processing Systems*, 34, 2021.

[16] D. A. Roberts, S. Yaida, and B. Hanin. The principles of deep learning theory. *arXiv preprint arXiv:2106.10165*, 2021.

[17] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[18] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.