AMATH 590: Final Project Report
Nonlinear Random Matrix Theory for Deep Learning

Nga Yu Lo[1*], Sara M. Ichinaga[1*]

[1] Department of Applied Mathematics, University of Washington, Seattle, WA 98195, United States

Abstract

The application of random matrix theory to deep learning has great potential to improve and expand our understanding of neural networks, as random matrices largely define the initial loss landscape of deep learning models. In a recent work by Pennington and Worah [6], the authors derive an expression for the Stieltjes transform of the Gram matrix $\frac{1}{m}\mathbf{Y}\mathbf{Y}^\top$, where $\mathbf{Y} = f(\mathbf{W}\mathbf{X})$ is the output of a neural network layer with activation f, Gaussian weights $\mathbf{W}$, and Gaussian data $\mathbf{X} \in \mathbb{R}^{n\times m}$. In deriving this result, the authors not only derive the spectral density of the Gram matrix, but they also identify a special family of activation functions for which the authors hypothesize improvements in training speed and training error. In this work, we summarize and reproduce the main results of Pennington and Worah's work [6], in addition to supplementing these results with our own numerical experiments. We generate and visualize the theoretical and empirical spectral density of the Gram matrix for new data parameters, and we test Pennington and Worah's model improvement claims via a noisy MNIST classification task. Overall, we find that we are able to verify Pennington and Worah's results and ultimately build a successful classifier by using the activation functions suggested by their findings. All corresponding code may be found at https://github.com/sichinaga/amath590-rmt-project.

1 Introduction

From applications in image classification [4] and speech recognition [2] to system identification [1] and modeling in latent space [5], there is no doubt that deep learning models are extremely powerful data-driven tools that can be used to model highly complex systems and improve our understanding of the world around us. In general, such models consist of multiple "layers" at which input data $\mathbf{X} \in \mathbb{R}^{n\times m}$ is multiplied by a weight matrix $\mathbf{W}$. A nonlinear activation function f is then applied pointwise to this matrix, yielding the layer output matrix $\mathbf{Y} = f(\mathbf{W}\mathbf{X})$. Typically, the weight matrices $\mathbf{W}$ are initialized randomly and are then gradually updated via a training procedure that utilizes a loss function applied to the actual output and the desired output of the overall model. It is hence natural to consider the application of random matrix theory to neural networks, as random matrices not only describe the initial weights of the model, but also oftentimes the data itself (take for example highly polluted data or noisy signals).

Despite the theoretical complexities introduced by the presence of nonlinear activation functions, Pennington and Worah recently derived an expression for the Stieltjes transform of the Gram matrix $\frac{1}{m}\mathbf{Y}\mathbf{Y}^\top$, assuming Gaussian random weights and Gaussian random data [6]. The authors additionally assumed infinitely-large weight and data matrices by sending the number of samples m and the feature size n to infinity ($m, n \to \infty$) while keeping the ratio n/m fixed to ensure similar rates of growth. In other words, they assume that both the data set and the neural network model are extremely large, which is quite reasonable given the context of most deep learning applications. In obtaining their results, Pennington and Worah found that the Stieltjes transform of the Gram matrix in the large data and model limit depends solely on matrix shape ratios, and the activation function parameters ζ and η, which denote the square of the Gaussian mean of f' and the Gaussian mean of f^2 respectively. This result additionally suggests that for $\zeta = 0$ activation functions, the Stieltjes transform of the Gram matrix becomes Marchenko-Pastur, in which case the spectral density of the Gram matrix is preserved even as data passes through the neural network. This led the authors to hypothesize that $\zeta = 0$ activation functions possess many advantageous training benefits, including training speed ups, reduced training error, and a batch normalization effect that lasts throughout multiple layers of the neural network during the initial stages of training [3].

*Corresponding authors (nylo3@uw.edu and sarami7@uw.edu)

Unfortunately, Pennington and Worah's findings are largely new and the authors were unable to investigate these hypotheses in their work. In this project, we reproduce the work of Pennington and Worah [6] with the ultimate goal of summarizing their main results in addition to validating their findings with supplemental experiments and numerical results.

2 Mathematical Background

2.1 Problem statement

For a random input data matrix $\mathbf{X} \in \mathbb{R}^{n_0 \times m}$ and a random weight matrix $\mathbf{W} \in \mathbb{R}^{n_1 \times n_0}$ whose entries are independent and identically distributed (i.i.d.) Gaussian random variables

$$X_{ij} \overset{\mathscr{L}}{=} \mathcal{N}\left(0, \sigma_x^2\right), \qquad W_{ij} \overset{\mathscr{L}}{=} \mathcal{N}\left(0, \tfrac{\sigma_w^2}{n_0}\right), \tag{1}$$

we define the output of the corresponding neural network layer to be the matrix

$$\mathbf{Y} := f(\mathbf{WX}) \in \mathbb{R}^{n_1 \times m} \tag{2}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a nonlinear activation function that is applied pointwise to the entries of $\mathbf{WX}$. We assume that f is of zero mean and finite moments. That is, we assume

$$\int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} f(\sigma_w \sigma_x z) = 0, \qquad \left| \int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} f(\sigma_w \sigma_x z)^k \right| < \infty, \quad k > 1. \tag{3}$$

We are specifically interested in the spectral density of the Gram matrix

$$\mathbf{M} := \frac{1}{m} \mathbf{Y}\mathbf{Y}^\top \in \mathbb{R}^{n_1 \times n_1} \tag{4}$$

in the regime of infinite data samples m and infinite model parameters n_0, n_1. We additionally assume that these parameters are roughly the same order magnitude and hence grow at the same rate as they approach infinity. As such, we define the dimension ratio parameters

$$\phi := \frac{n_0}{m} \in \mathbb{R}, \qquad \psi := \frac{n_0}{n_1} \in \mathbb{R}, \tag{5}$$

which remain fixed as $n_0, n_1, m \longrightarrow \infty$.

2.2 Computing the spectral density

Where the *empirical spectral density* of $\mathbf{M}$ is defined as

$$\rho_{\mathbf{M}}(t) := \frac{1}{n_1} \sum_{j=1}^{n_1} \delta(t - \lambda_j(\mathbf{M})), \tag{6}$$

for the Dirac delta function δ and eigenvalues λ_j of $\mathbf{M}$, we study the *limiting spectral density* when $n_1 \to \infty$. That is, we are interested in (6) as $n_1 \to \infty$. Given the *Stieltjes transform* G of $\rho_{\mathbf{M}}$

$$G(z) := \int \frac{\rho_{\mathbf{M}}(t)}{z - t} dt = -\frac{1}{n_1} \mathbb{E}\left[\mathrm{Tr}(\mathbf{M} - z\mathbf{I}_{n_1})^{-1}\right], \tag{7}$$

we can recover the limiting spectral density using the following Stieltjes inversion formula:

$$\rho_{\mathbf{M}}(\lambda) = -\frac{1}{\pi} \lim_{\epsilon \to 0^+} G(\lambda + i\epsilon). \tag{8}$$

Hence computing the limiting spectral density comes down to computing $G(z)$. We determine $G(z)$ by expanding (7) as a Laurent series of moments of $\rho_{\mathbf{M}}$, which we note holds for large z.

$$G(z) = \sum_{k=0}^{\infty} \frac{m^k}{z^{k+1}}, \quad m^k := \int dt \rho_{\mathbf{M}} t^k = \frac{1}{n_1} \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{M}^k\right)\right] \tag{9}$$

We then compute $\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{M}^k\right)\right]$ using the *moment method*, which considers the expansion

$$\frac{1}{n_1}\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{M}^k\right)\right] = \frac{1}{n_1}\mathbb{E}\left[\sum_{i_1,\dots i_k\in[1,n_1]\cap\mathbb{Z}} M_{i_1i_2}M_{i_2i_3}\dots M_{i_{k-1}i_k}M_{i_ki_1}\right] \tag{10}$$

$$= \frac{1}{n_1}\frac{1}{m}\mathbb{E}\left[\sum_{\substack{i_1,\dots i_k\in[1,n_1]\cap\mathbb{Z}\\ \mu_1,\dots\mu_k\in[1,m]\cap\mathbb{Z}}} Y_{i_1\mu_1}Y_{i_2\mu_1}Y_{i_2\mu_2}Y_{i_3\mu_2}\dots Y_{i_k\mu_k}Y_{i_1\mu_k}\right]. \tag{11}$$

By taking on this particular perspective of $\mathrm{Tr}\left(\mathbf{M}^k\right)$, and by utilizing the fact that the matrices $\mathbf{W}$ and $\mathbf{X}$ contain i.i.d. Gaussian random variables that relate to $\mathbf{Y}$ and $\mathbf{M}$ via (2) and (4) respectively, we find that we can derive the results of Pennington and Worah [6], which we present in Section 3.

3 Main Results

Theorem 1. *For $\mathbf{M}, \phi, \psi, \sigma_w, \sigma_x$, and constants η, ζ defined as*

$$\eta = \int dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} f(\sigma_w\sigma_x z)^2, \quad \zeta = \left[\sigma_w\sigma_x \int dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} f'(\sigma_w\sigma_x z)\right]^2 \tag{12}$$

the Stieltjes transform of $\mathbf{M}$ is

$$G(z) = \frac{\psi}{z}P\left(\frac{1}{z\psi}\right) + \frac{1-\psi}{z} \tag{13}$$

which satisfies the recurrence relation

$$P(t) := 1 + (\eta-\zeta)P_\phi(t)P_\psi(t)t + \frac{P_\phi(t)P_\psi(t)t\zeta}{1-P_\phi(t)P_\psi(t)t\zeta} \tag{14a}$$

$$P_\phi(t) := 1 + (P(t)-1)\phi, \qquad P_\psi(t) := 1 + (P(t)-1)\psi. \tag{14b}$$

Proof. In lieu of a complete proof, we provide a proof sketch that outlines the major ideas that lead to Theorem 1. For more details and for a more thorough proof, we refer readers to the supplemental material of Pennington and Worah [6].

Our primary goal is to understand the behavior of our expression from the moment method (11). Notice that there is a unique set of indices that $\{i_1, \dots i_k\}$ and $\{\mu_1, \dots \mu_k\}$ can take in each term that contributes to the sum. In addition, since the elements of $Y_{i\mu}$ are i.i.d., the expected value of terms with the same set of indices is the same. Therefore, by considering the frequency of these unique patterns and their corresponding expected value, we can compute the expectation over $\mathbf{W}$ and $\mathbf{X}$ for (11). To formalize this approach, let each pattern be represented by a graph $G_j = (V_j, E_j)$, where the set of vertices V_j accounts for the indices i and μ that appear in the sequence, and the set of edges E_j accounts for values $Y_{i\mu}$ that appear in the sequence. If we let $\mathcal{G}$ denote the set of all possible graph structures, counting repeats, then (11) may alternatively be expressed in terms of the expected value of each graph structure.

$$\frac{1}{n_1}\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{M}^k\right)\right] = \frac{1}{n_1}\frac{1}{m}\sum_{G_j\in\mathcal{G}} \mathrm{E}_{G_j} \tag{15}$$

Now consider a special case in which G_j is a $2k$ cycle. That is, consider a pattern in which all indices i, μ are unique. Using (2) and the known distribution of the entries of $\mathbf{W}$ and $\mathbf{X}$ (1), we can proceed by expressing (11) as a multi-dimensional integral over the elements of $\mathbf{W}$ and $\mathbf{X}$.

$$\mathrm{E}_{2k\mathrm{cycle}} = \mathbb{E}\left[\,(Y_{i_1\mu_1}) \times (Y_{i_2\mu_1}) \times \cdots \times (Y_{i_1\mu_k})\,\right] \tag{16}$$

$$= \int f\left(\sum_\ell W_{i_1\ell}X_{\ell\mu_1}\right) f\left(\sum_\ell W_{i_2\ell}X_{\ell\mu_1}\right)\cdots f\left(\sum_\ell W_{i_1\ell}X_{\ell\mu_k}\right)\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X} \tag{17}$$

$$\mathcal{D}\mathbf{W} = \prod_{i,j=1}^{n_1,n_0} \frac{dW_{ij}}{\sqrt{2\pi\sigma_w^2/n_0}} e^{-\frac{n_0}{2\sigma_w^2}W_{ij}^2} \qquad \mathcal{D}\mathbf{X} = \prod_{i,j=1}^{n_0,m} \frac{dX_{ij}}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{1}{2\sigma_x^2}X_{ij}^2} \tag{18}$$

Although we can be sure that all i, μ are unique in the $2k$ cycle case, we must still account for repeated $Y_{i\mu}$ values. To do so, define the set of all unique values of $Y_{i\mu}$ as $\mathcal{Z}$ and define the matrix $\mathbf{Z} \in \mathbb{R}^{n_1 \times m}$ such that $\mathbf{Z} = \mathbf{W}\mathbf{X}$ and the entries of $\mathbf{Z}$ do not lead to repeat $Y_{i\mu}$ values. To enforce the former, we utilize the Dirac delta function. To enforce the latter, we define the matrix $\mathbf{Z}$ as follows, which then allows us to express our original multi-dimensional integral as the following.

$$\mathrm{E}_{2k\mathrm{cycle}} = \int \prod_{z_{ij} \in \mathcal{Z}} \delta\left(z_{ij} - \sum_{\ell} W_{i\ell} X_{\ell j}\right) F(z)\, \mathcal{D}z \mathcal{D}\mathbf{W} \mathcal{D}\mathbf{X} \tag{19}$$

$$Z_{i\mu} = \begin{cases} z_{i\mu} & Y_{i\mu} \in \mathcal{Z} \\ 0 & \text{else} \end{cases} \qquad \mathcal{D}z = \prod_{z_{ij} \in \mathcal{Z}} dz_{ij} \qquad F(z) := f(z_{i_1\mu_1}) f(z_{i_2\mu_1}) \cdots f(z_{i_1\mu_k}) \tag{20}$$

In lieu of using the delta function explicitly, we instead utilize the Fourier representation of the delta function. This introduces the auxiliary variables λ and their corresponding matrix $\mathbf{\Lambda}$, which we define as follows. We hence incorporate these variables into our multi-dimensional integral.

$$\mathrm{E}_{2k\mathrm{cycle}} = \int e^{-i\mathrm{Tr}(\mathbf{\Lambda}^\top(\mathbf{W}\mathbf{X} - \mathbf{Z}))} F(z)\, \mathcal{D}\lambda \mathcal{D}z \mathcal{D}\mathbf{W} \mathcal{D}\mathbf{X} \tag{21}$$

$$\Lambda_{i\mu} = \begin{cases} \lambda_{i\mu} & Y_{i\mu} \in \mathcal{Z} \\ 0 & \text{else} \end{cases} \qquad \mathcal{D}\lambda = \prod_{\lambda_{ij} \in \mathbf{\Lambda}} \frac{d\lambda_{ij}}{2\pi} \qquad \delta(x) = \frac{1}{2\pi} \int d\lambda e^{i\lambda x} \tag{22}$$

We now perform our integrals with respect to $\mathbf{W}$ and $\mathbf{X}$ to get

$$\mathrm{E}_{2k\mathrm{cycle}} = \int \mathcal{D}\lambda \mathcal{D}z \exp\left[-i\mathrm{Tr}(\mathbf{\Lambda}\mathbf{Z}) - \frac{n_0}{2} \log\det\left| 1 + \frac{\sigma_w^2 \sigma_x^2}{n_0} \mathbf{\Lambda}\mathbf{\Lambda}^\top \right| \right] F(z). \tag{23}$$

Now, as we evaluate our remaining integrals, we must consider the fact that we are particularly interested in the leading-order behavior of our integral as $n_0, n_1, m \to \infty$. Using the method of steepest descent, we note that the dominant contributions to our λ integrals will occur near $\mathbf{\Lambda} = 0$, as this is where n_0 is minimized. We hence narrow in on the regime about $\mathbf{\Lambda} = 0$, which justifies the use of the following summation expression for the log determinant.

$$\mathrm{E}_{2k\mathrm{cycle}} = \int \mathcal{D}\lambda \mathcal{D}z \exp\left[-i\mathrm{Tr}(\mathbf{\Lambda}\mathbf{Z}) - \frac{n_0}{2} \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} \mathrm{Tr}\left(\frac{\sigma_w^2 \sigma_x^2}{n_0} \mathbf{\Lambda}\mathbf{\Lambda}^\top \right)^j \right] F(z). \tag{24}$$

Next we rescale our λ variables to $\tilde{\lambda} = \frac{\sigma_w \sigma_x}{\sqrt{n_0}} \lambda$ to obtain the following. Note that because $n_0 \to \infty$, we are interested in the asymptotic behavior of our integral for very small $\tilde{\lambda}$.

$$\mathrm{E}_{2k\mathrm{cycle}} = \int \mathcal{D}\tilde{\lambda} \mathcal{D}z \exp\left[-i \frac{\sqrt{n_0}}{\sigma_w \sigma_x} \mathrm{Tr}(\tilde{\mathbf{\Lambda}}\mathbf{Z}) - \frac{n_0}{2} \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} \mathrm{Tr}\left(\tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}^\top \right)^j \right] F(z). \tag{25}$$

Finally, we use the fact that the lowest-order $\tilde{\lambda}$ terms of the summation contribute the greatest. We additionally use the fact that we are considering a $2k$-cycle in order to obtain the contributing factors $\tilde{\lambda}_{i\mu}$. From this, we obtain the following leading-order behavior, for which we may evaluate our $\tilde{\lambda}$ integrals to obtain the following expected contribution of the $2k$-cycle when $k > 1$.

$$\mathrm{E}_{2k\mathrm{cycle}} \approx (-1)^k n_0 \int \mathcal{D}\tilde{\lambda} \mathcal{D}z \exp\left[-i \frac{\sqrt{n_0}}{\sigma_w \sigma_x} \mathrm{Tr}(\tilde{\mathbf{\Lambda}}\mathbf{Z}) - \frac{n_0}{2} \mathrm{Tr}\left(\tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}^\top \right) \right] \tilde{\lambda}_{i_1\mu_1} \tilde{\lambda}_{i_1\mu_1} \cdots \tilde{\lambda}_{i_1\mu_k} F(z) \tag{26}$$

$$= n_0^{1-k} \left[\sigma_w \sigma_x \int dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} f'(\sigma_w \sigma_x z) \right]^{2k} = n_0^{1-k} \zeta^k, \quad k > 1 \tag{27}$$

When $k = 1$, we instead obtain the following asymptotics. We refer to this as the "one-block" case, as a 2-cycle refers to a graph that is a consists of 2 vertices separated by one edge length.

$$\mathrm{E}_{1\mathrm{block}} \approx \int \mathcal{D}\tilde{\lambda}\mathcal{D}z \exp\left[-i\frac{\sqrt{n_0}}{\sigma_w\sigma_x}\mathrm{Tr}(\tilde{\mathbf{\Lambda}}\mathbf{Z}) - \frac{n_0}{2}\mathrm{Tr}\left(\tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}^\top\right)\right] F(z) \tag{28}$$

$$= \int dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} f(\sigma_w\sigma_x z)^2 = \eta \tag{29}$$

In summary, we find the the expected contribution of the $2k$-cycle and the 1-block are

$$\mathrm{E}_{2k\mathrm{cycle}} = n_0^{1-k}\zeta^k \qquad\qquad \mathrm{E}_{1\mathrm{block}} = \eta \tag{30}$$

which we may generalize to other graph structures that are combinations of simple cycles and one-blocks. To make this approach more concrete, we present the following definition:

Definition 1 (Admissible graph)**.** *An admissible graph is an outerplanar graph (an undirected graph for which no vertex is completely surrounded by edges) for which all blocks are simple cycles of even length.*

We define admissible graphs because of the following propositions stated by Pennington and Worah:

Proposition 1. *For an admissible graph G with c cyclic blocks, b blocks of exactly size one, and $2k$ edges in total, E_G grows as $n_0^{c-k}\zeta^{c-b}\eta^b$.*

Proposition 2. *For a non-admissible graph G, E_G grows as n_0^{c-k-1}.*

In other words, since admissible graphs may be represented as combinations of cyclic blocks and size-one blocks, we may generalize our results for the $2k$-cycle and the 1-block to admissible graphs with variable amounts of cycles and blocks, yielding Proposition 1. Furthermore, Proposition 2 allows us to ignore the contributions of non-admissible graphs, as their contributions vanish as $n_0 \to \infty$.

When $i_1, \dots i_k, \mu_1, \dots \mu_k$ are unique, the $2k$ indices can be viewed as a graph of a $2k$-sided polygon, and when $k = 1$, we consider this to be a graph of size 1. When these indices are not unique, what we end up with a tree like graph over blocks of even cycles and blocks of sized 1 graphs, which we called admissible graphs. In this way, we count the number of possible graphs and use Proposition 1 to compute G.

$$G(z) = \frac{1}{z} + \sum_{k=1}^{\infty} \frac{1}{z^{k+1}} \sum_{\nu_i,\nu_\mu=0}^{k} \sum_{b=0}^{\nu_i+\nu_\mu+1} p(k,\nu_i,\nu_\mu,b)\eta^b\zeta^{\nu_i+\nu_\mu+1-b}\phi^{\nu_i}\psi^{\nu_\mu} = \frac{1-\psi}{z} + \frac{\psi}{z}\sum_{k=1}^{\infty}\frac{1}{z^k\psi^k}P(k) \tag{31}$$

$$P(k) = \sum_{\nu_i,\nu_\mu=0}^{k} \sum_{b=0}^{\nu_i+\nu_\mu+1} p(k,\nu_i,\nu_\mu,b)\eta^b\zeta^{\nu_i+\nu_\mu+1-b}\phi^{\nu_i}\psi^{\nu_\mu} \tag{32}$$

Writing $P(t) = \sum_k^\infty P(k)t^k$ as a generating function, we sum over all k to yield the recurrence relation and self consistent equations:

$$P(t) := 1 + (\eta - \zeta)P_\phi(t)P_\psi(t)t + \frac{P_\phi(t)P_\psi(t)t\zeta}{1 - P_\phi(t)P_\psi(t)t\zeta} \tag{33a}$$

$$P_\phi(t) := 1 + (P(t) - 1)\phi, \qquad P_\psi(t) := 1 + (P(t) - 1)\psi. \tag{33b}$$

Making the substitution of $P(t)$ back into (31), we get the results of the theorem. □

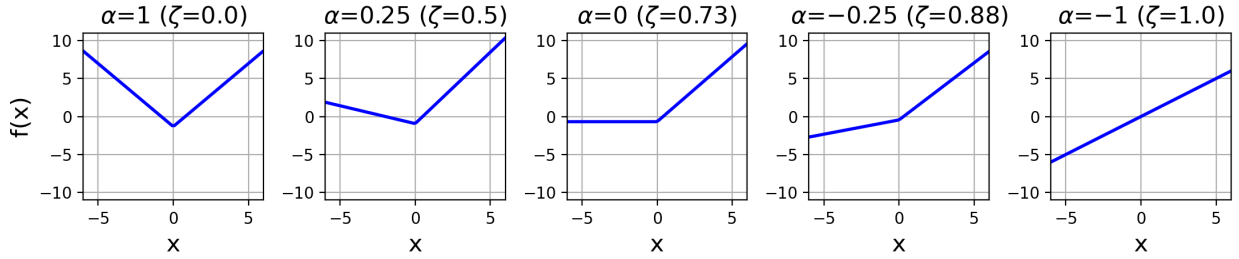

Figure 1: Example ReLU functions (40) for various parameter values α.

3.1 Limiting cases

The constants η and ζ precisely characterize the effect of our activation function f on the limiting spectral density. Consider when $\zeta = 0$, we can choose $\eta = 1$ since η is general scaled by z. In this scenario, (13) is simplified to

$$zG^2 + \left((1-z)\,\frac{\psi}{\phi} - 1\right) G + \frac{\psi}{\phi} = 0. \tag{34}$$

Furthermore, when $\psi = 1$, we get the Marchenko-Pastur distribution that satisfies

$$\phi z G^2 + (1 - z - \phi)\, G + 1 = 0. \tag{35}$$

Then, the limiting spectral density of $\frac{1}{m}\mathbf{X}\mathbf{X}^\top$ is the same as the limiting spectral density of $\frac{1}{m}\mathbf{Y}\mathbf{Y}^\top$, which suggests that the existence of functions f where the singular value distribution of the covariance for $f(\mathbf{W}\mathbf{X})$ is the same as $\mathbf{X}$. This preservation of the covariance has been noted in machine learning literature to improve training performance and speed. For example, batch normalization is commonly used to help preserve the covariance of the input data [3].

4 Numerical Results

4.1 Computing spectral densities

As a proof of concept of the results of Pennington and Worah [6], we begin by examining the spectral density of the gram matrix (4) that we discover empirically via random sampling. Pennington and Worah perform similar experiments and present similar results in their original work [6]. However, details such as spectral density visualizations and methods for computing $G(z)$ from (13) are omitted. Here, we present evidence that the results of Section 3 hold in practice, while additionally providing details and results that supplement those of the original work [6].

First, let $\phi = \psi = 1$. Note that Pennington and Worah examined $\phi = 1$ and $\psi = 3/2$ in their original work [6]. Using (13) and the recurrence relation (14), we find that $G(z)$ must then satisfy

$$z^2(\zeta^2 - \eta\zeta)\big[G(z)\big]^4 + z^2\zeta\big[G(z)\big]^3 + z(\eta - \zeta)\big[G(z)\big]^2 - z\big[G(z)\big] + 1 = 0. \tag{36}$$

In the $\zeta = 0$ case, this expression simplifies drastically to

$$z\eta\big[G(z)\big]^2 - z\big[G(z)\big] + 1 = 0, \tag{37}$$

which then yields the following analytical expression for $G(z)$:

$$G(z) = \frac{z \pm \sqrt{z^2 - 4\eta z}}{2\eta z}, \quad \zeta = 0. \tag{38}$$

Thus when $\zeta = 0$, we may simply use (38) in conjunction with (8) in order to recover the corresponding theoretical spectral density of the gram matrix $\mathbf{G}$. In the $\zeta \neq 0$ case, we unfortunately

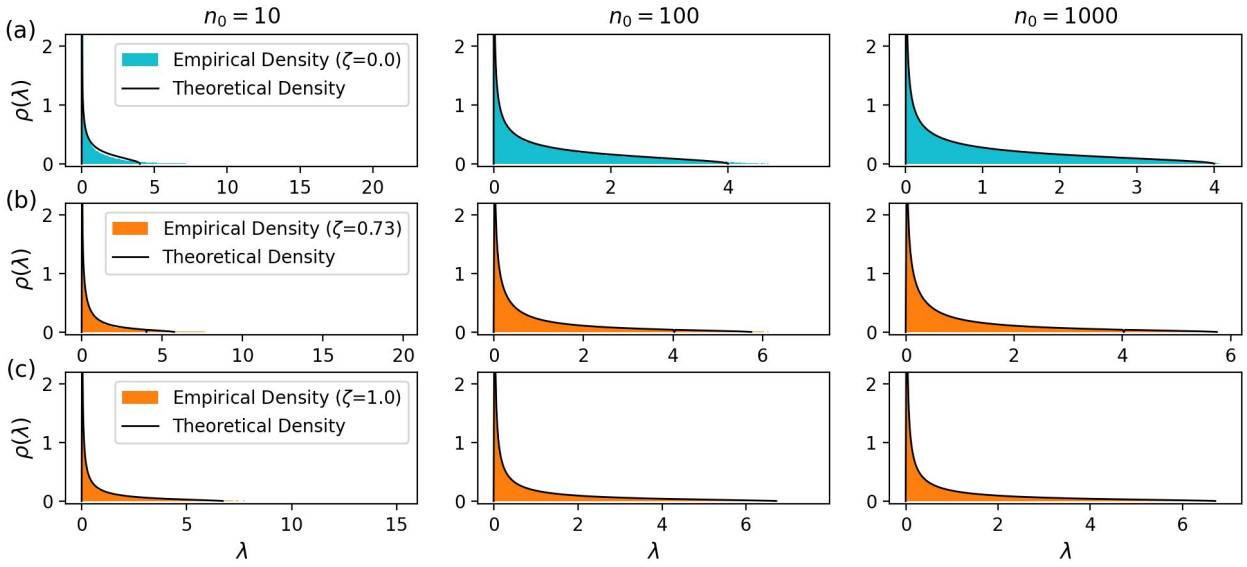

Figure 2: Empirical spectral density of the Gram matrix $\frac{1}{m}\mathbf{Y}\mathbf{Y}^\top$ versus the theoretical density given by Theorem 1 for various matrix sizes n_0 ($\phi = \psi = 1$) and tilted ReLU activation $f = f_\alpha$ (40) with (a) $\alpha = 1$ ($\zeta = 0$), (b) $\alpha = 0$ ($\zeta \approx 0.73$), and (c) $\alpha = -1$ ($\zeta = 1$).

lack a closed form solution for $G(z)$. Hence we instead obtain $G(z)$ from (36) via a root-finding method. More specifically, we deploy Newton's method in order to find the roots of the function

$$h(G) := z^2(\zeta^2 - \eta\zeta)G^4 + z^2\zeta G^3 + z(\eta - \zeta)G^2 - zG + 1 = 0 \tag{39}$$

at various values of $z \in \mathbb{C}$, all while initializing our root-finding routine with (38). This paired with (8) thus allows us to compute the theoretical spectral density for any value of ζ.

In order to investigate activation functions over a variety of ζ values, we use the tilted ReLU function

$$f_\alpha(x) = \frac{[x]_+ + \alpha[-x]_+ - \frac{1}{\sqrt{2\pi}}(1+\alpha)}{\sqrt{\frac{1}{2}(1+\alpha^2) - \frac{1}{2\pi}(1+\alpha)^2}} \quad [x]_+ := \max\{0, x\}, \tag{40}$$

as was done by Pennington and Worah [6]. See Figure 1 for a visualization of the function. The corresponding η, ζ values of this activation may be computed analytically with respect to α as

$$\eta = 1, \qquad \zeta = \frac{(1-\alpha)^2}{2(1+\alpha^2) - \frac{1}{2\pi}(1+\alpha)^2}, \tag{41}$$

in which case we may toggle the α parameter in order to vary ζ without varying η.

For our first experiment, we aim to numerically confirm the legitimacy of Theorem 1 for various values of ζ. To do this, we compute the theoretical density given by Theorem 1 and the Stieltjes inversion formula (8) for $\zeta = 0, 0.73, 1$. We then generate an empirical spectral density for each value of ζ across matrix sizes $n_0 = 10, 100, 1000$ by generating $\frac{10^5}{n_0}$ random Gram matrices and computing their eigenvalues. Comparing the theoretical and empirical spectral densities for each ζ, n_0 pair then yields the results in Figure 2. Overall, we observe that as we increase the size of our data and weight matrices (i.e. as $n_0 \to \infty$), we begin to empirically observe the theoretical density for all ζ values considered. In other words, we observe the results of Theorem 1 for sufficiently-large data and model sizes.

Our next experiment seeks to confirm the special $\zeta = 0$ behavior discussed in Section 3.1. To do this, we re-use the ζ parameters and theoretical densities of the previous experiment, only this time, we compute the spectral density of a "5-pass" Gram matrix for $n_0 = 50, 100, 1000$. That is, we compute the spectral density of $\frac{1}{m}\mathbf{Y}_\ell\mathbf{Y}_\ell^\top$ for $\ell = 5$, where $\mathbf{Y}_\ell = f(\mathbf{W}_\ell\mathbf{Y}_{\ell-1})$, $\mathbf{Y}_0 = \mathbf{X}$. The results of this experiment are summarized in Figure 3. This time, we observe that in the $\zeta \neq 0$

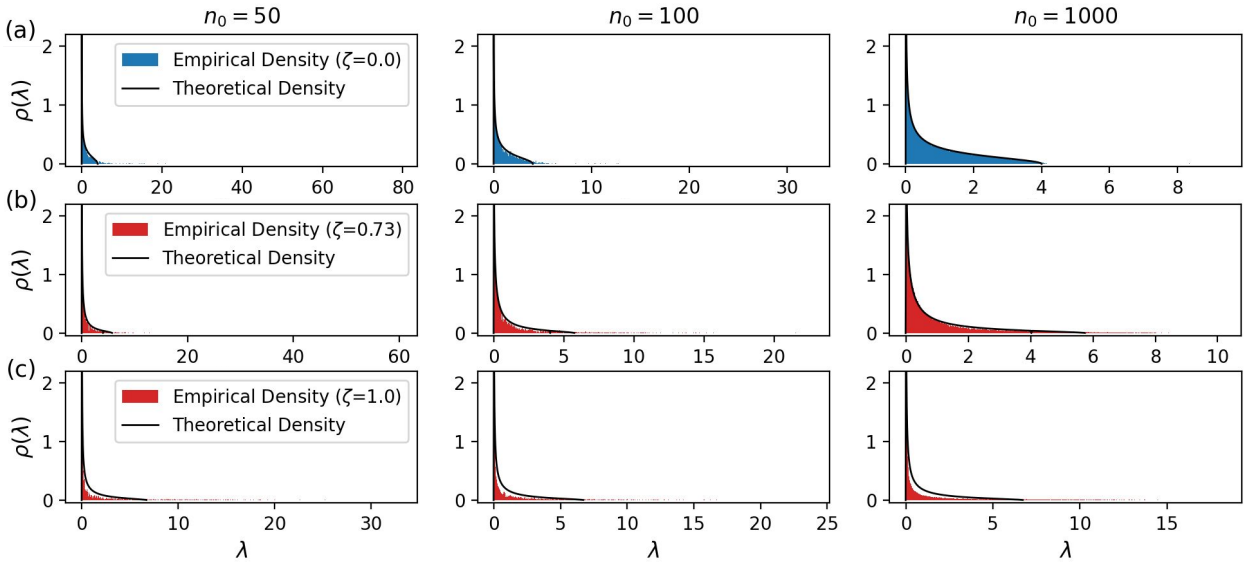

Figure 3: Empirical spectral density of the Gram matrix $\frac{1}{m}\mathbf{Y}\mathbf{Y}^\top$ after five passes through the neural network versus the theoretical density given by Theorem 1 for a single pass through the neural network. We consider various matrix sizes n_0 ($\phi = \psi = 1$) and use the tilted ReLU activation $f = f_\alpha$ (40) with (a) $\alpha = 1$ ($\zeta = 0$), (b) $\alpha = 0$ ($\zeta \approx 0.73$), and (c) $\alpha = -1$ ($\zeta = 1$).

case, we do not empirically observe the theoretical density as $n_0 \to \infty$. However in the $\zeta = 0$ case, we still approximately observe the theoretical density as $n_0 \to \infty$. This suggests and confirms the notion that the spectral density of the Gram matrix is roughly preserved as data passes through the neural network, assuming a $\zeta = 0$ activation function and sufficiently-large matrices.

4.2 MNIST classification

So far, we have been working with Gaussian datasets. In this section, we consider looking at non-Gaussian datasets. In particular, we consider the standard MNIST dataset, which consists of $70,000$ 28×28 pixel images of single digit numbers, each of which come with a corresponding label $y \in \{0, 1, \dots, 9\}$ for classifying the images. While varying the activation functions with values of ζ, we study the effects of ζ on our image classification training performance.

We trained a three-layer neural network with fully connected linear layers and an activation function f. We used an Adam optimizer with a learning rate of 10^{-3} to train our model using cross entropy loss. To fit within the regime of Theorem 1, we add Gaussian noise with variance $\sigma^2 = 1$ to the MNIST images after data normalization (i.e. mean subtraction and variance re-scaling to one). We also subsample 784 data points for our training set and restrict the width of our network to 784, thereby setting $\psi = \phi = 1$. Moreover, we initialized random weights from the standard Gaussian with variance $\frac{1}{n_\ell}$, where n_ℓ is the dimension of the layer input features.

We trained four models, each with a different activation function, for 10 epochs. We show the resulting training loss and accuracy in Figure 4. All of our models converge to about 70% accuracy; however, the model with the activation function of $\zeta = 0$ noticeably achieves the smaller loss and higher accuracy in fewer epochs. Note that we observe similar $\zeta = 0$ improvements, even when using $60,000$ training points as opposed to 784. We additionally observe higher training and test accuracies overall when using tilted ReLU with $\zeta = 0$ as opposed to $\zeta \neq 0$. See the supplementary MNIST notebook in Git for the large training set results.

Figure 5 illustrates the singular value distribution of the output covariance after each layer of the neural network. While we can't expect the distribution to be the same (even in the case of $\zeta = 0$ due to the non-Gaussian nature of MNIST dataset), we do observe that differences between a $\zeta = 0$ activation function and otherwise. In particular, the $\zeta = 0$ function still roughly preserves the shape of the distribution. It is also interesting that after training, the shape of the distribution

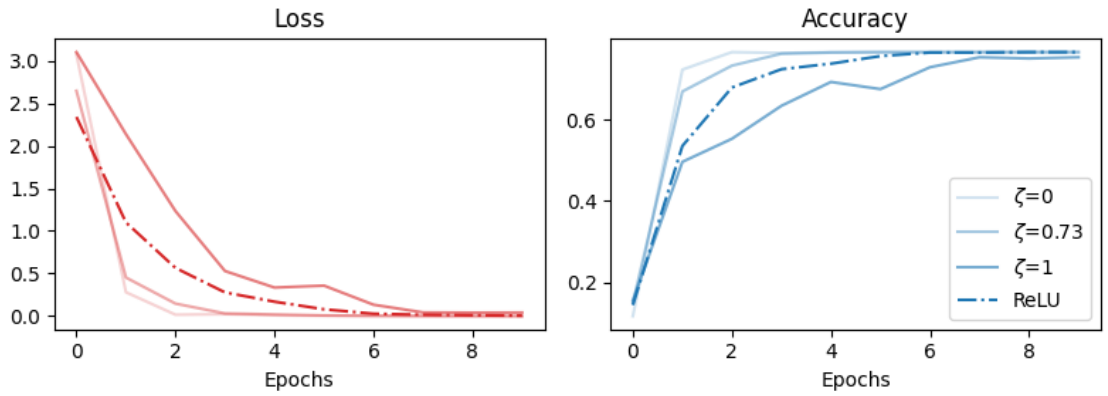

Figure 4: Training performance for ReLU and tilted ReLU activation functions (40) with $\zeta = 0, 0.73, 1$. $\phi = 1, \psi = 1, m = 784$.

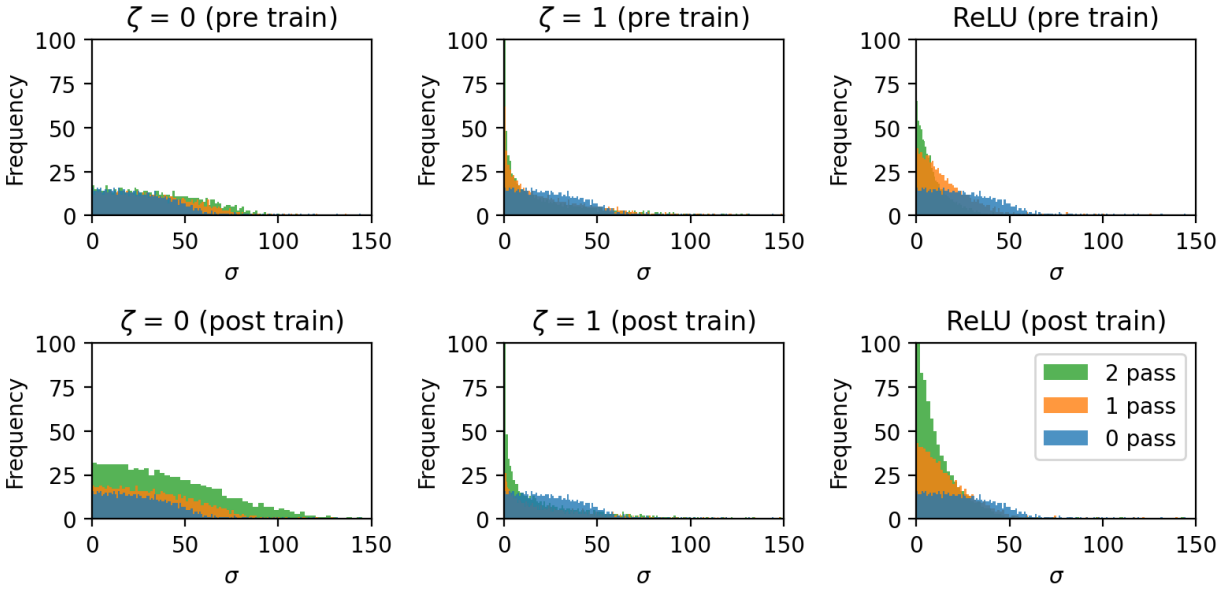

Figure 5: Singular value distribution for a 3-layer network using regular ReLU and tilted ReLU activation functions (40) with $\zeta = 0, 1$.

stays similar to their counterpart prior to training.

5 Conclusion

In their paper, Pennington and Worah studied the spectral density of the Gram matrix $\frac{1}{m}\mathbf{Y}\mathbf{Y}^\top$ where $\mathbf{Y} = f(\mathbf{W}\mathbf{X})$ is the output of a layer in a neural network. Where f is often nonlinear, the main theorem shows that its effect on the spectral density is characterized by constants ζ and η. When $\zeta = 0$, the limiting spectral density of $\frac{1}{m}\mathbf{Y}\mathbf{Y}^\top$ is the same as the spectral density of $\frac{1}{m}\mathbf{X}\mathbf{X}^\top$ (assuming that $\mathbf{X}$ is Gaussian). This suggests the existence of activation functions that do not distort the covariance of the input, a desirable trait to improve training performance in neural networks.

Our numerical results demonstrate the merits of Pennington and Worah's theorem. We also provide a methodology to recover the spectral density using Newton's method. We further examine the effects of activation functions f (whose $\zeta = 0$) on a network trained on a MNIST classification task. Our results suggest these $\zeta = 0$ functions yields faster training time and their preservation properties on Gaussian datasets have possible extensions to non-Gaussian data.

References

[1] K. Champion, B. Lusch, J. N. Kutz, and S. L. Brunton, *Data-driven discovery of coordinates and governing equations*, Proceedings of the National Academy of Sciences, 116 (2019), pp. 22445–22451.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*, IEEE Signal Processing Magazine, 29 (2012), pp. 82–97.

[3] S. Ioffe and C. Szegedy, *Batch normalization: accelerating deep network training by reducing internal covariate shift*, in Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, JMLR.org, 2015, p. 448–456.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, in Advances in Neural Information Processing Systems, vol. 25, Curran Associates, Inc., 2012, pp. 1097–1105.

[5] B. Lusch, J. N. Kutz, and S. L. Brunton, *Deep learning for universal linear embeddings of nonlinear dynamics.*, Nature Communications, 9 (2018), p. 4950.

[6] J. Pennington and P. Worah, *Nonlinear random matrix theory for deep learning*, Journal of Statistical Mechanics: Theory and Experiment, 2019 (2019), p. 124005. https://dx.doi.org/10.1088/1742-5468/ab3bc3.