Reproducibility Project for Traditional and Heavy Tailed Self Regularization in Neural Network Models.

Kevin Pratama
School of Computer Science
University of Waterloo
Waterloo, ON, N2L 3G1
kpratama@uwaterloo.ca

Abstract

The reproducibility project aims to replicate the experiments described in one of the ICLR reviewed papers, by comparing the weights spectral density in several deep networks trained on CIFAR-10 dataset. The paper under review in which the project is based on aims to describe the underlying process of how deep learning works through heavy-tailed self-regularization. It is built upon previous works, which have also attempted to explain the generalization of deep network through several regularizations techniques. In the reproducibility project, the author aims to evaluate and validate the results from the paper under review. By reproducing the work and examining further the implications of weight patterns through traning, the author hopes to extend the work of the original authors and provide more foundations for future work.

1 Introduction

Generalization has been one of the subjects of intense study in the field of deep learning networks. However, the vast number of adjustable hyperparameters in the network present a challenge to the existing solutions and approaches that attempt to generalize through these hyperparameters. In order to understand how all of the hyperparameters are correlated to the model accuracy, the paper under review proposed a novel approach through self-regularization. That is, by eliminating all explicit regularizations and observing the regularization patterns from deep networks through spectral density of weights.

The proposed approach from the said paper is deemed significant by the author, as it directly maps the spectral density of Random Matrix Theory (RMT) to weight matrices and tries to observe the underlying circumstances in which deep networks generalizes better. Therefore, the author is interested in reproducing the heavy-tailed self-regularization, comparing the observations against the empirical results from the paper, and extending the work through observing the result on other kind of layers.

The reproducibility project relates to the course in regularization and deep learning. This is done through understanding the underlying factors of deep networks, which is the self organization pattern of weight matrices regardless of explicit regularizations.

2 Related Works

There have been several efforts attempting to understand the generalization and overtraining in deep networks. One of the approaches tries to explain the need of understanding kernel learning, which states that generalization is correlated to the properties of kernel function, rather than the optimiza-

Submitted to Introduction to Machine Learning (CS680, 2018 Fall). Do not distribute.

tion process. [2] Other approach describe the generalization using Lipschitz regularization [7] or from an invariance point of view using the Fisher-Rao norm. [6] One of the weaknesses found in these approaches are that they provide explicit regularization techniques. While the techniques avoid overfitting, it still does not explain fully the behaviour of spectral density on weight matrices which lead to overfitting.

However, the proposed approach from the paper under review attempts to interpret the generalization better and how deep learning works by going through the underlying factor beneath regularizations mentioned above. The authors of this paper are motivated by the analogy proposed by Choromanska et al.[3] which suggest that Energy Landscape of a zero-temperature Gaussian Spin Glass which explains the Energy/optimization Landscape of modern DNNs. The authors stated that the Spin Glass may be the key to understanding how deep learning works at its fundamentals.

In particular, the authors of the forementioned paper are concerned with Random Matrix Theory (RMT) and its relations to spectral density of deep networks.[4] Using Machenko-Pastur (MP) Theory, the RMT describes the density patterns of large rectangular matrix W emerging from deep networks when all explicit regularization are removed. These patterns are referred to as the 5+1 training phase: Random-Like, Bleeding-Out, Bulk+Spikes, Bulk-Decay, Heavy-Tailed, and Rank-Collapse. According to the empirical results in the paper, deep networks achieve their best generalizations when the spectral density of weights form the heavy-tailed distribution. [1] This taxonomy corresponds to the theory of self-regularization in deep networks, which is the main topic of the paper. These observations provide the groundwork and motivation for this project.

The Marchenko Pastur Distribution is defined as follows [9]:

$$\begin{aligned} \lambda_{\pm} &= \sigma^2(1 \pm \sqrt{\lambda})^2 \\ \sigma^2 &= \lambda_{max}(1 + \frac{1}{\sqrt{Q}})^{-2} \\ \text{where } Q &= \frac{1}{\lambda} \end{aligned} \tag{1}$$

While the Powerlaw Distribution is defined as follows [8]: Powerlaw Distribution

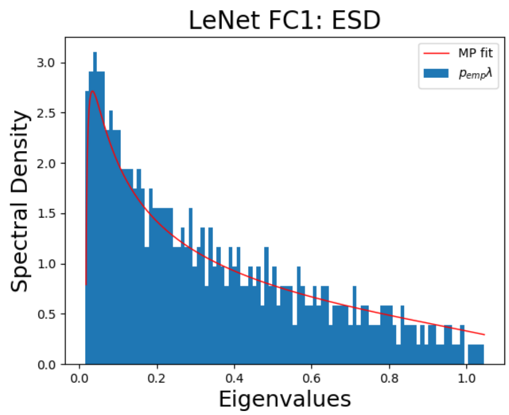
$$p(x) \sim x^{-\alpha} \tag{2}$$

The author of this proposal is interested in providing an implementation of the heavy-tailed self-regularization observations on deep networks. The observations found from the reproducibility project will be validated against the empirical results in the paper, and provide further experiments and observations on other kinds of networks, as discussed in proposed work.

3 Results

The tools used in reproducing the results are PyTorch, Powerlaw [5]

3.1 LeNet5



LeNet Spectral Density on FC1 Layer

The LeNet model was trained in PyTorch, using the same configurations described in the reviewed paper. The LeNet model was trained for 20 epochs using AdaDelta optimizer, achieving 100% accuracy on the training set and 98% accuracy on the test set. After plotting the spectral density of eigenvalues into the histogram, the above graph shows a significantly different pattern from the one described in the paper. The eigenvalue mass exhibits a form of heavy-tailed distribution, rather than the perfect Marchenko-Pastur fit. This may be due to the explicit regularizations used from the original paper, whereas the LeNet reproduced does not use any form of explicit regularization (no dropout or batch normalization). The powerlaw fit gives a value of $\alpha \approx 7.882$.

3.2 Pretrained AlexNet

FC1: Heavy-Tailed

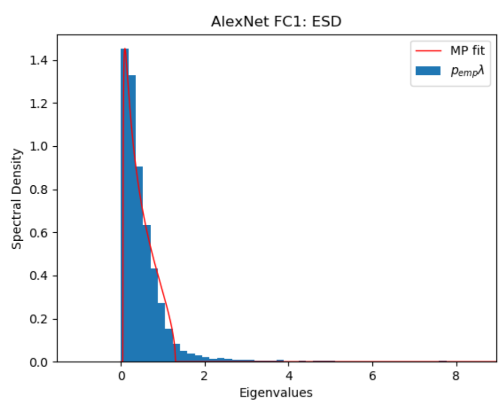
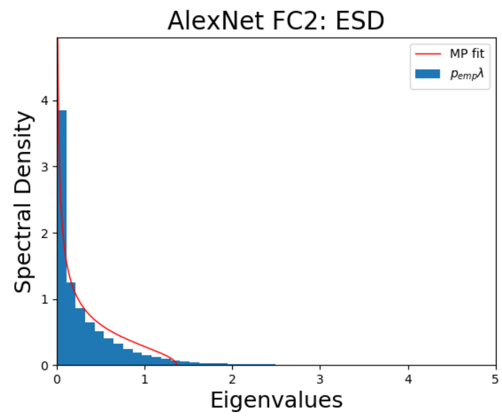


Figure 2: AlexNet Spectral Density on FC1 Layer.

The above figure show the empirical spectral density of eigenvalues of the pretrained AlexNet FC1 Layer on ImageNet dataset. The empirical spectral density of the visible eigenvalues in the range from 0 to 8 exhibit a well-defined heavy-tailed distribution. The best MP fit (in red curve) captures a good part of the eigenvalue mass, but the peak is not filled in. A part of the eigenvalue mass is bleeding out from the bulk as shown in the figure, and the shape of the ESD is convex in the region near and above the best fit of λ_+. Finally, the powerlaw fit for this distribution gives an alpha value of $\alpha \approx 2.288$.

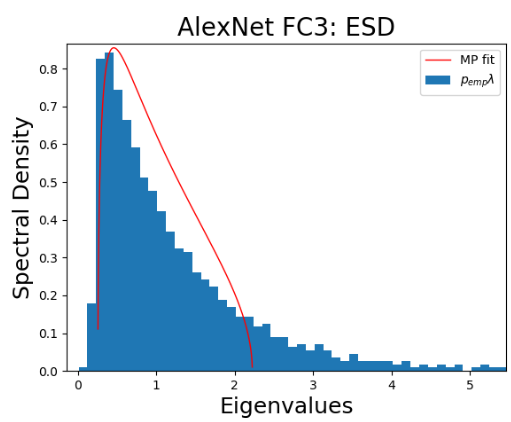
FC2: Heavy-Tailed



AlexNet Spectral Density on FC2 Layer

In the FC2 layer of the pretrained AlexNet model, as shown in the above figure differs significantly from the standard MP theory. After several adjustments in the MP fit through Q and σ, the best MP fit (denoted in red curve) still does not fit the bulk part of the eigenvalue mass. The overall inspection of the FC2 layer agrees similarly according to the reviewed paper. The entire ESD is concave in nearly everywhere, compared to the MP curves which are convex near the bulk edge. The powerlaw fit gives an alpha value of $\alpha \approx 2.245$, smaller than FC1.

FC3: Heavy-Tailed



AlexNet Spectral Density on FC3 Layer

In the final FC3 layer, the empirical spectral density is, again, deviating strongly from the predictions of the MP theory, both for the bulk properties and local edges. The powerlaw fit gives an alpha value of $\alpha \approx 3.02$, which is larger than FC1 and FC2 layers.

Summary

The following is the table of the AlexNet FC layers with phases and powerlaw distributions.

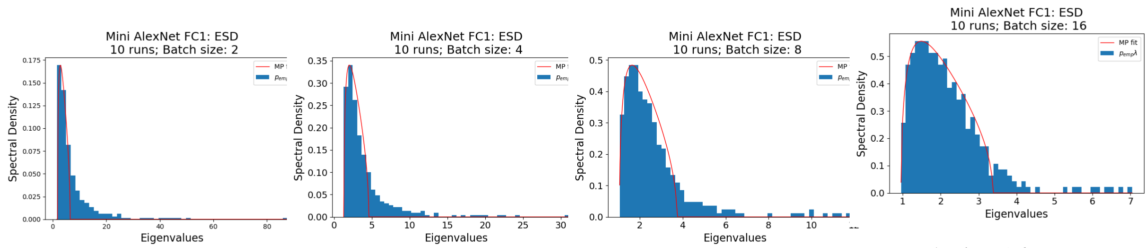
AlexNet FC Phase and Powerlaw Distribution

Layer	**Phase**	**Powerlaw α**	**Soft Rank**
FC1	Heavy-Tailed	2.288	0.033
FC2	Heavy-Tailed	2.245	0.015
FC3	Heavy-Tailed	3.020	0.050

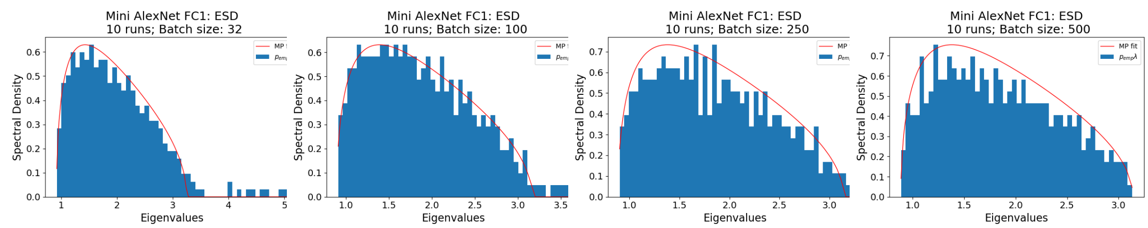
As expected, all of the FC1, FC2, and FC3 layers exhibit a form of heavy-tailed distribution. However, the Marchenko-Pastur distribution fit fails to describe the spectral density of eigenvalues within these layers. This problem can be alleviated with using the powerlaw distribution, with α values ranging from above 2 and slightly above 3.

3.3 Performance with respect to Batch Size

MiniAlexNet - FC1 Layer



Batch size 2 | Batch size 4 | Batch size 8 | Batch size 16



Batch size 32 | Batch size 100 | Batch size 250 | Batch size 500

The above figures showed the empirical spectral density of MiniAlexNet models trained in several batch sizes (2, 4, 8, 16, 32, 100, 250, 500). The majority of the bulk decreases as the batch size decreases.

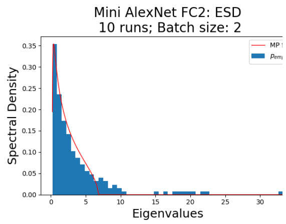
- At batch size b = 250 and larger, the spectral density resembles a pure Marchenko Distribution and exhibit RANDOM-LIKE distribution, as noted in the paper.
- As b decreases, some of the eigenvalues tend to contain more information in the outlier region (spikes). As b = 100, the ESDs resemble BLEEDING-OUT phase.
- When b = 32, the eigenvalues are now separated in two parts: the bulk and the spike. The mass distribution resembles BULK+SPIKES.
- At b = 16 and 8, the spikes in the outlier region become more visible. The spectral density exhibits BULK-DECAY.
- Finally, for b = 4 and 2, the eigenvalues are localized within one region near value 0, where the histogram now resembles HEAVY-TAILED distribution.

Mini AlexNet FC1 Phase and Powerlaw Distribution

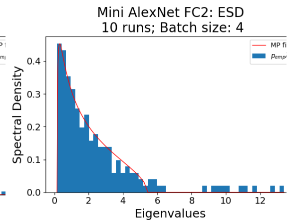
Batch Size	**Phase**	**Powerlaw α**	**Soft Rank**
2	Heavy-Tailed	2.438	0.074
4	Heavy-Tailed	2.803	0.147
8	Bulk Decay	3.565	0.315
16	Bulk Decay	4.944	0.479
32	Bulk+Spikes	6.755	0.642
100	Bleeding-Out	13.92	0.880
250	Random-like	16.628	0.978
500	Random-like	17.317	1.002

In the table, the models with batch sizes 2 and 4 have α values in the range between 2 and below 3.5, which is consistent with the fidnings from AlexNet layers. As α increases, the spectral density exhibits more bulk forms, or earlier phases in the 5+1 training phase.

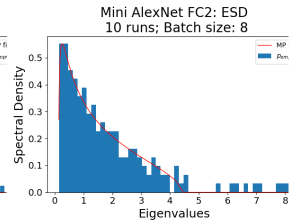
MiniAlexNet - FC2 Layer



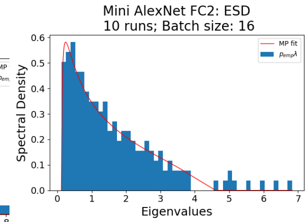
Batch size 2



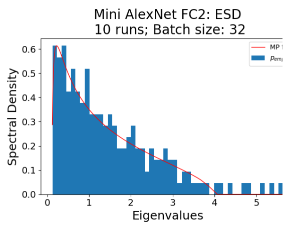
Batch size 4



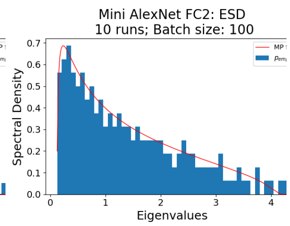
Batch size 8



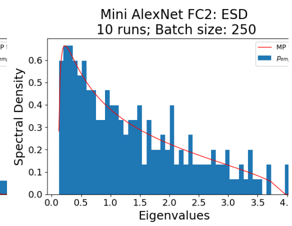
Batch size 16



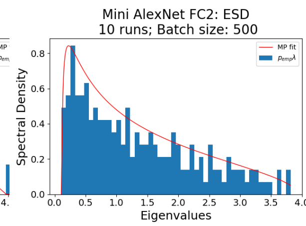
Batch size 32



Batch size 100



Batch size 250



Batch size 500

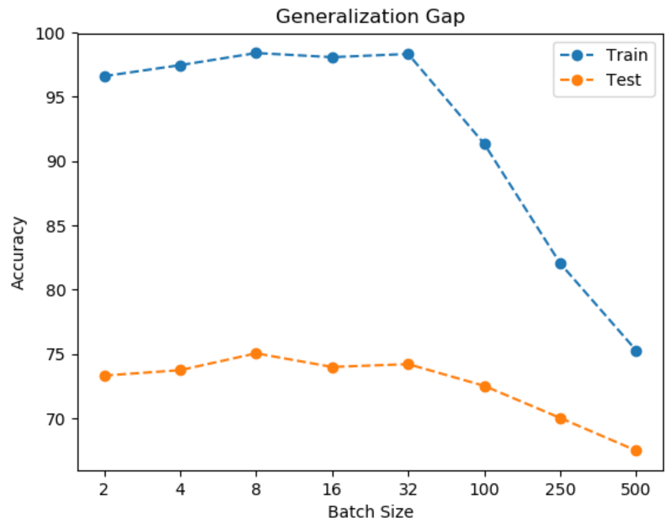
The above figures show the eigenvalues spectral density for FC2 Layer of the Mini AlexNet model. The reproduced results also agree with the findings from the original paper. The overall shape of the spectral density for FC2 is different, as the aspect ratio of the matrix is less. However, similar properties of the 5+1 training phases can still be observed in the figures. Moreover, the spikes or eigenvalues that lie in the outlier region tend to be more localized as batch size decreases.

Mini AlexNet FC2 Phase and Powerlaw Distribution

Batch Size	Phase	Powerlaw α	Soft Rank
2	Heavy-Tailed	2.631	0.204
4	Heavy-Tailed	2.801	0.405
8	Bulk-Decay	3.355	0.540
16	Bulk-Decay	3.807	0.615
32	Bulk Decay	4.689	0.701
100	Bulk+Spikes	4.664	0.942
250	Bleeding-Out	11.951	0.977
500	Bleeding-Out	11.067	1.029

In the table, we can see that the heavy-tailed forms have a powerlaw α between 2 and below 3.3, which is consistent with the findings from FC1 layer and the pretrained AlexNet layers. As α increases, the spectral density exhibits more bulk forms, or earlier phases in the 5+1 training phase.

3.4 Generalization Gap



Generalization Gap of Mini AlexNet with respect to batch size.

The results of the MiniAlexNet training on different batch sizes (2, 4, 8, 16, 32, 64, 100, 250, 500) are displayed in the above figure. The overall reproduced results agree with the findings in the paper, where the training and test accuracy tend to decrease as the batch size increases. As conjectured in the reviewed paper, the network was unable to extract the intricate details of the dataset in large batches. The peak performance hovers around batch value of 8.

4 Conclusion

After implementing the reproducibility for the heavy-tailed regularization behaviors in neural networks, the findings of the modern, DNN models mostly agree with the observations found in the original paper. Deep Neural Networks such as AlexNet tend to exhibit the 5+1 training phases as training accuracy increases, where a well-generalized neural network model forms a heavy-tailed distribution for its eigenvalue mass. One exception, however, is the LeNet model, which exhibits the heavy-tailed distribution compared to the expected perfect Marchenko-Pastur fit in the original observation. Several interesting future work is to further extend the reproducibility project to analyze the ESD of the convolutional layers within CNNs, and other types of neural networks which is recurrent neural networks (RNNs). The author is also interested in analyzing further simpler models such as LeNet, and other modern DNNs on InceptionNet and VGG11Nets.

References

[1] Anonymous. Traditional and heavy tailed self regularization in neural network models. In *Submitted to International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJeFNoRcFQ. under review.

[2] M. Belkin and S. M. S. Ma. To Understand Deep Learning We Need to Understand Kernel Learning. Technical Report Preprint: arXiv:1802.01396, 2018.

[3] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surfaces of multilayer networks. Technical Report Preprint: arXiv:1412.023, 2014.

[4] A. Edelman and Y. Wang. Random matrix theory and its innovative applications. In R. Melnik and I. Kotsireas, editors, *Advances in Applied Mathematics, Modeling, and Computational Science*. Springer, 2013.

[5] E. B. J. Alstott and D. Plenz. powerlaw: a Python package for analysis of heavy-tailed distributions. Technical Report Preprint: arXiv:1305.0215., 2018.

[6] T. Liang, T. A. Poggio, A. Rakhlin, and J. Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *CoRR*, abs/1711.01530, 2017.

[7] A. M. Oberman and J. Calder. Lipschitz regularized deep neural networks converge and generalize. Technical Report Preprint: arXiv:1808.09540, 2018.

[8] A. D. P. Corral and R. F. i Cancho. A practical recipe to fit discrete power-law distributions. Technical Report Preprint: arXiv:1209.1270., 2018.

[9] P. Yaskov. A short proof of the Marchenko-Pastur theorem. Technical Report Preprint: arXiv:1506.04922., 2018.