Evolution of Eigenvalue Spectrum of Fully Connected (Dense) Layers in DNNs: The 5+1 Phases of Regularized Training

Vaibhav Dixit

December 13, 2024

1 Abstract

In this project, we conducted an empirical study of the evolution of eigenvalue spectra in deep neural networks (DNN) through different training phases, based on the paper [4] by Martin and Mahoney. The study identifies and validates 5+1 distinct phases of regularized training, characterized by changes in the empirical spectral density (ESD) of weight matrices in fully connected layers. These phases represent increasing levels of self-regularization during the training process, from initial random-like distributions to heavily structured, heavy-tailed distributions characteristic of well-trained networks. Our implementation using Lux.jl [5] examines two distinct architectures: a LeNet5 model trained on MNIST data [3] and a NanoGPT model trained on Shakespeare data [2]. Through this analysis, we demonstrate the universal nature of these training phases across different architectures and domains, while also highlighting architecture-specific variations in how these phases manifest.

2 Introduction

Understanding how neural networks learn and self-regularize during training is crucial to developing better training strategies and evaluating the conclusion of training. While traditional metrics like training loss and validation accuracy provide high-level insights, they offer limited visibility into the internal dynamics of how networks organize their learned representations. The analysis of eigenvalue spectra of weight matrices emerges as a powerful tool for understanding these internal dynamics, offering a window into the progressive structuring of neural networks during training. The eigenvalue spectrum of a neural network's weight matrices contains rich information about how the network processes and transforms data. Changes in this spectrum during training reflect the network's evolving capacity to capture and represent patterns in the training data. By studying these changes through the lens of Random Matrix Theory (RMT), we

can gain insights into the self-regularization processes that occur naturally during training. This study focuses on characterizing and validating the 5+1 phases framework proposed by Martin and Mahoney, which describes distinct stages of training through changes in the empirical spectral density of weight matrices. We implement this analysis using modern deep learning frameworks and apply it to both traditional convolutional architectures and transformer-based models, providing a comprehensive view of how these phases manifest across different neural network paradigms. All the code for the project is available at `https://github.com/Vaibhavdixit02/NN_Eigs_18338`.

3 Methodology

Our analysis methodology integrates theoretical foundations from Random Matrix Theory with empirical observations of neural network training dynamics. The core of our approach revolves around tracking the evolution of eigenvalue spectra in weight matrices throughout the training process. This enables us to understand how the internal structure of neural networks develops and self-organizes during learning. The primary metrics we track are the MP Soft Rank ($\mathcal{R}mp$) and Stable Rank ($\mathcal{R}s$). The MP Soft Rank quantifies the degree to which the empirical spectral density (ESD) of weight matrices adheres to the theoretical Marchenko-Pastur (MP) distribution. Calculated as the ratio between the theoretical MP bulk edge (λ^+) and the largest observed eigenvalue (λ_{max}), this metric provides insight into how far the network has deviated from its initial random state:

$$\mathcal{R}mp(W) = \frac{\lambda^+}{\lambda max}$$

The Stable Rank, computed as the ratio of the Frobenius norm squared to the spectral norm squared, offers a robust measure of the effective dimensionality of the weight matrices:

$$\mathcal{R}_s(W) = \frac{|W|_F^2}{|W|2^2} = \frac{\sum_i \lambda_i}{\lambda max}$$

Throughout the training process, we collect comprehensive data including full eigenvalue spectra of weight matrices, traditional training metrics such as loss and accuracy, and various layer-wise statistics. This multi-faceted approach allows us to correlate the spectral evolution with model performance and learning dynamics.

4 The 5+1 Phases of Training

4.1 Phase 1: Random-like

The initial phase of training represents a state where the network's weight matrices closely resemble random matrices. During this period, the empirical spectral

density exhibits remarkable adherence to the Marchenko-Pastur distribution, a theoretical prediction for random matrices. This alignment is not coincidental but a consequence of choosing the Glorot Normal initialization [1] for the fully-connected layers of our networks. The high MP Soft Rank observed during this phase indicates that the network has yet to develop meaningful structure through learning. The eigenvalues remain tightly bounded within the theoretical limits predicted by Random Matrix Theory, suggesting that the network is essentially operating as a random feature extractor. This phase serves as a crucial baseline against which we can measure the subsequent development of structure in the network.

4.2 Phase 2: Bleeding-out

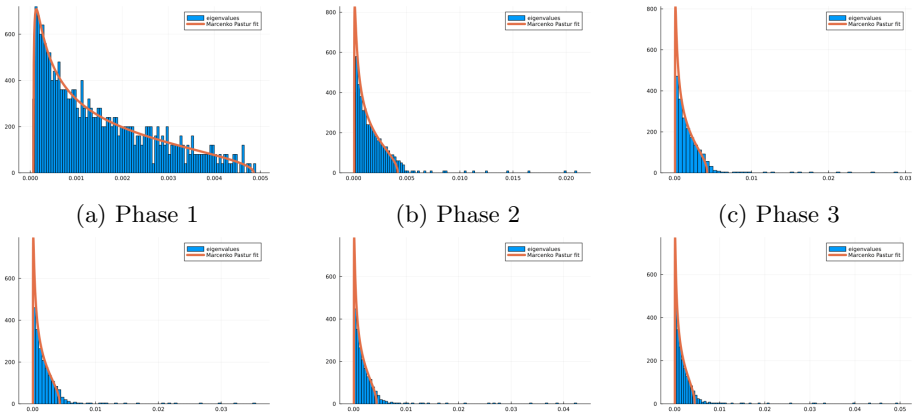
As training progresses, we observe the first signs of departure from pure randomness in what we term the Bleeding-out phase. This transition is characterized by a subtle but significant change in the eigenvalue distribution. The previously well-defined edge of the MP bulk begins to blur as eigenvalues start to extend beyond the theoretical boundary. This eigenvalue "bleeding" represents the network's initial attempts to capture structure in the training data. The formation of a shelf-like distribution of eigenvalues just beyond the MP bulk edge indicates the emergence of weak correlations in the weight matrices. These correlations reflect the network's growing sensitivity to patterns in the input data, marking the beginning of meaningful learning.

4.3 Phase 3: Bulk+Spikes

The Bulk+Spikes phase represents a critical transition in the network's learning process. During this phase, we observe a clear separation between the main bulk of eigenvalues, which still approximately follows the MP distribution, and several distinct eigenvalues that have separated significantly from the bulk. These "spike" eigenvalues correspond to strong, consistent patterns that the network has learned to recognize in the training data. This phase bears striking similarities to the Spiked-Covariance models from random matrix theory, suggesting that the network is implementing a form of Tikhonov-like regularization. The separation between bulk and spike eigenvalues provides a natural "scale" that distinguishes signal from noise in the learned representations.

4.4 Phase 4: Bulk-decay

The Bulk-decay phase marks a significant departure from traditional random matrix behavior. The clear separation between bulk and spike eigenvalues begins to break down, replaced by a continuous decay of eigenvalue density extending into the tail region. This phase represents an intermediate state between the well-structured but relatively simple representations of the Bulk+Spikes phase and the more complex, strongly correlated representations that characterize fully trained networks. The bulk edge becomes increasingly difficult to define



(a) Phase 1 (b) Phase 2 (c) Phase 3

(d) Phase 4 (e) Phase 5 (f) Phase 6

Figure 1: Lenet5 ESD evolution

precisely, and the eigenvalue distribution begins to show early signs of heavy-tailed behavior. This transition suggests that the network is developing more sophisticated, hierarchical representations of the training data.

4.5 Phase 5: Heavy-Tailed

In the Heavy-Tailed phase, we observe a complete breakdown of the MP distribution fit. The eigenvalue distribution now follows a power-law or heavy-tailed distribution, indicating the presence of correlations at all scales in the weight matrices. This phase is characteristic of well-trained modern deep neural networks and suggests the emergence of scale-free properties in the learned representations. The heavy-tailed distribution of eigenvalues implies that the network has developed a rich hierarchy of features, with no single characteristic scale dominating the representation. This property may help explain the remarkable generalization capabilities of modern deep neural networks.

4.6 Phase +1: Rank-collapse

The final phase, Rank-collapse, represents a pathological state of over-regularization. In this phase, we observe a large concentration of eigenvalues near zero, with only a few dominant eigenvalues remaining. This effectively reduces the rank of the weight matrices, severely limiting the network's capacity to represent complex patterns. This phase should generally be avoided in practice, as it indicates that the network has lost much of its representational power. The observation of rank collapse can serve as a useful warning sign during training, indicating that regularization parameters may need to be adjusted.

5 Experimental Results

Our experimental investigation centered on two distinct neural architectures: the classical LeNet5 convolutional network and the more modern NanoGPT transformer model. These architectures, representing different paradigms in deep learning, provided complementary insights into the universality and variation of spectral evolution during training.

5.1 LeNet5 Architecture

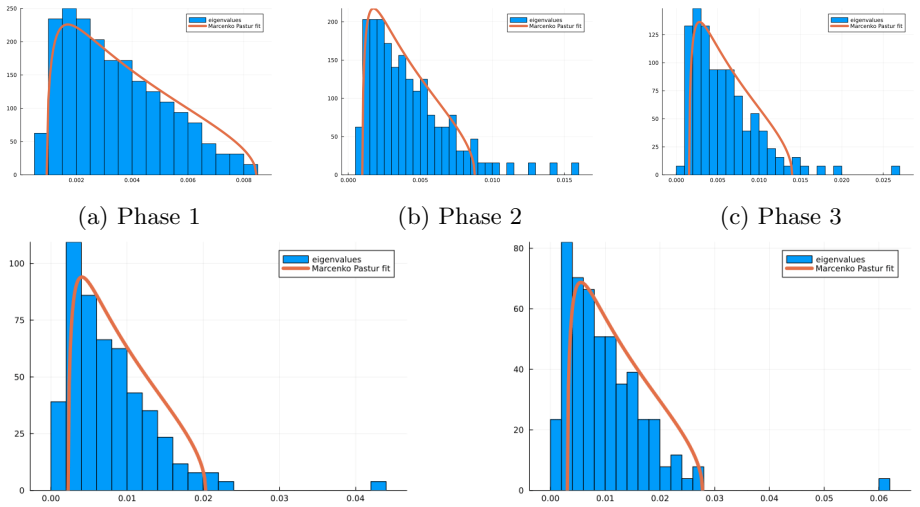
Our implementation of LeNet5 demonstrated a remarkably clear progression through all identified phases. The fully connected layers showed distinct transitions that strongly correlated with the model's learning progress. Beginning from the Random-like phase, we observed systematic evolution of the eigenvalue spectrum through each phase as training progressed. The transition between phases was particularly evident in the first fully connected layer (FC1). During early training, the eigenvalue distribution closely matched the theoretical MP distribution, with an MP Soft Rank near 1.0. As training progressed, we observed clear evidence of the Bleeding-out phase, followed by the emergence of distinct eigenvalue spikes characteristic of the Bulk+Spikes phase. Figure 1 illustrates this progression through all phases. The evolution from Random-like to Heavy-Tailed distributions was particularly pronounced in this architecture, making it an excellent example of the 5+1 phases framework. The clarity of these transitions may be attributed to the relatively simple architecture and well-structured nature of the MNIST dataset.

5.2 NanoGPT Architecture

The transformer-based NanoGPT model exhibited more nuanced behavior in its spectral evolution. While the fundamental phases were still observable, the transitions were less pronounced and the manifestation of heavy-tailed properties was more subtle. This difference suggests that architectural choices significantly influence how self-regularization manifests during training. In the attention layers, we observed a faster transition away from the Random-like phase, possibly due to the self-attention mechanism's ability to quickly capture relationships in the data. The feedforward layers showed behavior more similar to traditional fully connected layers, but with less distinct phase boundaries.

6 Key Findings

Our investigation revealed several significant insights about the nature of neural network training dynamics: First, we confirmed that self-regularization emerges naturally during training, independent of explicit regularization techniques. This process manifests through systematic changes in the eigenvalue spectrum, reflecting the network's progressive organization of learned representations. Second, we found that the progression through phases strongly correlates with



(a) Phase 1 (b) Phase 2 (c) Phase 3

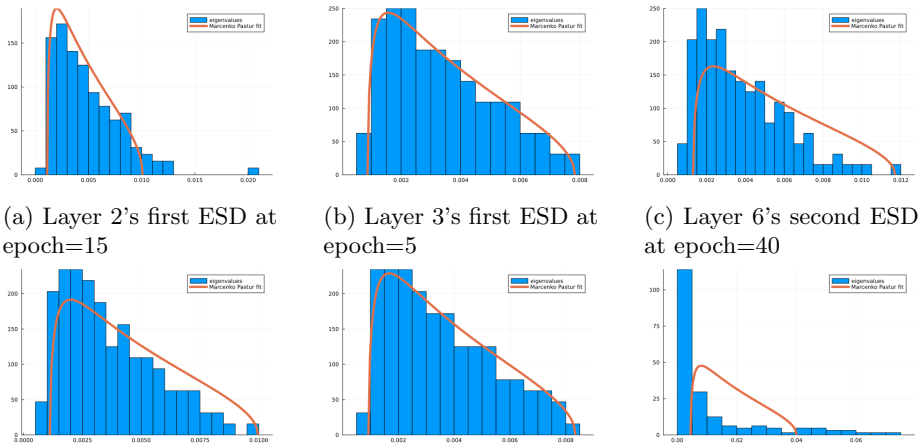
(d) Phase 4 (e) Phase 5

Figure 2: Nanogpt ESD evolution for first Fully Connected layer

model performance and generalization capability. The transition to Heavy-Tailed distributions appears to be a hallmark of successful training in modern architectures, though the clarity of this transition varies with architecture complexity. Third, our results demonstrate that while the 5+1 phases framework applies across different architectures, the manifestation of these phases can vary significantly. Traditional convolutional networks such as LeNet5 show more pronounced transitions, while transformer-based architectures exhibit more subtle spectral evolution patterns. Architecture-dependent variations in phase manifestation suggest that the underlying learning dynamics may be influenced by architectural constraints and inductive biases. This observation has important implications for architecture design and training optimization.

7 Conclusion

The 5+1 phases framework provides valuable insight into the DNN training dynamics through eigenvalue spectrum analysis. Through our comprehensive study of the LeNet5 and NanoGPT architectures, we have demonstrated how this framework can illuminate the internal dynamics of neural network training in different architectural paradigms. Our results confirm that the progression from Random-like to Heavy-Tailed distributions is a fundamental aspect of neural network learning, though its manifestation varies with architectural complexity. The LeNet5 architecture exhibited clear and distinct phase transitions that closely matched the theoretical framework, providing strong validation of the original theory. In contrast, the more complex NanoGPT architecture showed



(a) Layer 2's first ESD at epoch=15

(b) Layer 3's first ESD at epoch=5

(c) Layer 6's second ESD at epoch=40

(d) Layer 5's first ESD at epoch=10

(e) Layer 5's second ESD at epoch=5

(f) Layer 6's second ESD at epoch=50

Figure 3: A sample of Nanogpt ESDs

We observe that for Nanogpt not all the layers show similar evolution, and some do not agree with the proposed hypothesis for the evolution of the ESD.

more subtle transitions, suggesting that architectural sophistication influences the nature of spectral evolution. These architectural differences in spectral evolution patterns point to an important consideration in deep learning: while self-regularization is a universal phenomenon in neural network training, its specific characteristics are shaped by architectural choices and constraints. This understanding could inform future architecture design and training strategies. One particularly noteworthy observation is that the Heavy-Tailed phase, while more pronounced in traditional architectures like LeNet5, appears in some form across both models studied. This suggests that the development of heavy-tailed spectral distributions might be a fundamental characteristic of successful neural network training, regardless of architecture. However, the varying clarity of phase transitions between architectures also raises new questions about how architectural choices influence the self-regularization process. Future work might explore how specific architectural elements, such as attention mechanisms, skip connections, or normalization layers, affect the spectral evolution of neural networks during training. In conclusion, while our study validates the 5+1 phases framework as a valuable tool for understanding neural network training dynamics, it also reveals the framework's nuanced application across different architectural paradigms. These insights not only deepen our theoretical understanding of deep learning but also suggest practical considerations for neural network design and optimization.