A Practical Investigation into the Training Dynamics of Deep Neural Networks

Parisa Mohammadi

Independent Researcher

Course: Advanced Topics in Machine Learning

Abstract—This report investigates the core theoretical principles governing the training dynamics of modern deep neural networks, focusing on the concept of implicit bias in overparameterized models. We explore the mathematical foundations of Information Geometry, the Fisher Information Metric (FIM), and the Neural Tangent Kernel (NTK), contrasting the "lazy" kernel regime with the "rich" feature-learning regime. This theory is made tangible through a three-phase capstone project that visually demonstrates: (1) the effectiveness of geometry-aware Natural Gradient Descent, (2) the evolution of the empirical NTK as a proxy for feature learning, and (3) the multi-stage trajectory of an optimizer revealed through Principal Component Analysis. The results provide a cohesive, practical illustration of how networks find, tune, and refine features to achieve generalization.

Keywords—*Neural Network Dynamics, Information Geometry, Fisher Information Metric, Natural Gradient Descent, Neural Tangent Kernel, Feature Learning, SGD Trajectory, PCA*

1. Introduction: The Paradox of Modern Deep Learning

Modern deep learning is defined by a fascinating paradox. We routinely train massively "overparameterized" models, containing millions of parameters, on datasets with far fewer examples. Classical statistical theory would predict that these models, which can perfectly interpolate the training data, should suffer from catastrophic overfitting and fail to generalize to new, unseen data. Yet, empirically, they often generalize remarkably well.

This observation has launched a quest to understand the source of this success. The prevailing hypothesis is the concept of **implicit bias**. This theory posits that the optimization algorithm itself—most commonly Stochastic Gradient Descent (SGD)—acts as a regularizer. From the infinite landscape of solutions that achieve zero training error, SGD is biased towards a specific subset of solutions with properties that favor generalization. This project aims to explore, implement, and visualize the core theoretical mechanisms that underpin this phenomenon.

2. Theoretical Foundations

Our investigation is grounded in three pillars of modern deep learning theory: the geometric structure of the learning problem, the idealized behavior of infinite-width networks, and the complex dynamics of practical optimization algorithms.

2.1. Information Geometry: The True Map of the Learning Landscape

Information Geometry reframes machine learning by treating a family of probability distributions as a curved differential manifold. The model's parameters (e.g., network weights) serve as coordinates on this "statistical manifold." The key insight is that this space is not flat (Euclidean).

The natural way to measure distance and curvature on this manifold is with the **Fisher Information Metric (FIM)**, which serves as the Riemannian metric tensor. For a model $p(z|\theta)$, the FIM is defined as the variance of the score function:

$$F(\theta) = \mathbb{E}_{z\sim p(z|\theta)}\left[(\nabla_\theta \log p(z|\theta))(\nabla_\theta \log p(z|\theta))^\top\right] \tag{1}$$

The FIM is also the negative expected Hessian of the log-likelihood, directly connecting it to the curvature of the loss surface. A large FIM eigenvalue indicates a "sharp" direction, while a small eigenvalue indicates a "flat" direction. The optimization path of steepest descent on this true, curved manifold is given by the **Natural Gradient**, which preconditions the standard gradient with the inverse of the FIM: $\Delta\theta = -\gamma F(\theta)^{-1}\nabla L(\theta)$.

2.2. The Lazy vs. Rich Regimes of Training

Theoretical analysis reveals two distinct dynamical regimes for network training:

2.2.1. *The Lazy (Kernel) Regime*

In the limit of infinite network width, training dynamics simplify dramatically. The network's parameters move only infinitesimally, and its complex behavior can be linearized. In this "lazy" regime, the network is equivalent to a kernel machine governed by a fixed kernel called the **Neural Tangent Kernel (NTK)**. The NTK is defined as:

$$\Theta(x, x'; w) = \langle \nabla_w f(x; w), \nabla_w f(x'; w)\rangle \tag{2}$$

Because the NTK remains constant, the network does not perform meaningful **feature learning**; it learns a linear classifier on top of fixed, random features.

2.2.2. *The Rich (Feature Learning) Regime*

Practical, finite-width networks operate in the "rich" regime. Here, parameters undergo significant changes, the NTK approximation breaks down, and the network's internal representations dynamically adapt to the data. It is in this regime that deep learning's true power emerges. Recent research provides a compelling narrative for this process, which can be understood in three stages:

1. **Find:** As shown by Glasgow (2023), feature learning provides a quantifiable, exponential improvement in sample complexity. SGD's initial dynamics are a "signal-finding" process that discovers relevant low-dimensional feature subspaces.
2. **Tune:** The work of Xu et al. (2024) on "grokking" provides a clear example of feature tuning. A network first finds a catastrophically overfit, kernel-like solution before its neurons gradually align with the true data features, enabling generalization.
3. **Maximize:** Finally, the analysis by Schechtman & Schreuder (2025) shows that even after achieving zero training error, SGD's implicit bias continues to refine the solution by implicitly maximizing the classification margin.

3. Project Implementation and Results

This project was designed as a capstone implementation to provide tangible, visual evidence of these core theoretical concepts. Each phase was constructed to isolate and demonstrate a specific aspect of training dynamics.

3.1. Phase 1: Visualizing Optimization on a 2D Geometric Landscape

Design: This phase was designed to provide a stark visual contrast between standard Euclidean optimization and geometry-aware optimization. By implementing both Standard and Natural Gradient Descent on a 2D non-convex loss surface with a defined FIM, we aimed to visualize the practical benefit of incorporating the landscape's true geometry.

Results and Analysis: The visualization in Figure 1 perfectly captures the theoretical prediction. The Standard GD path (orange) is inefficient; it is "confused" by the parameterization and takes large steps perpendicular to the valley of the loss function, resulting in slow, oscillating progress. In contrast, the Natural Gradient path (green), by using the inverse of the FIM to rescale its steps, correctly identifies the true path of steepest descent along the curved manifold. It takes a smooth and vastly more efficient route to the minimum.

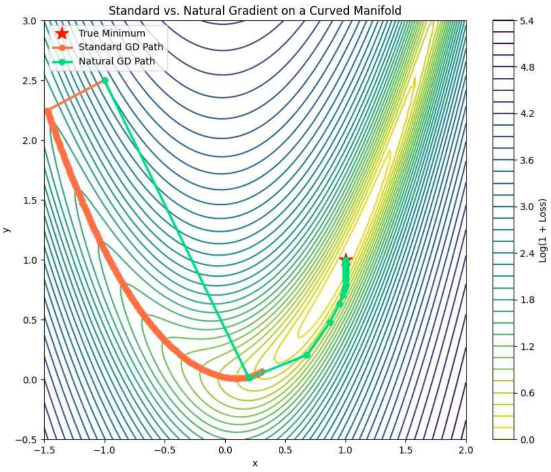

Figure 1. Phase 1 Result: NGD (green) uses the space's geometry to find a direct path to the minimum. Standard GD (orange) is much slower; its inefficient oscillations prevent it from reaching the goal in the allotted time, despite heading in the right general direction.

3.2. Phase 2: Probing the Lazy vs. Rich Regimes with the Empirical NTK

Design: This phase moves from a toy problem to a real neural network. The goal was to compute both the theoretical, static NTK of an infinite-width network and the *empirical* NTK (eNTK) of a finite-width network during training. By tracking the evolution of the eNTK's eigenvalues, we aimed to create a visual proxy for feature learning.

Results and Analysis: Figure 2 shows that the eigenvalues of the theoretical NTK (dashed black lines) are constant, as predicted. However, the eigenvalues of our practical network's eNTK (solid colored lines) are clearly dynamic. They start near their theoretical counterparts at initialization but then decay and diverge as training progresses. This deviation is the visual signature of the network operating in the rich, feature-learning regime. It is tangible proof that the network is not "lazy" but is actively changing its internal representation.

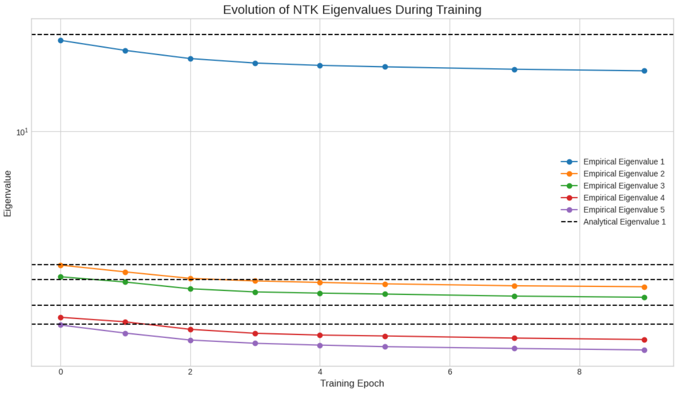

Figure 2. Phase 2 Result: The empirical NTK eigenvalues (solid lines) evolve during training, deviating from the static, theoretical NTK eigenvalues (dashed lines), visually signifying feature learning.

3.3. Phase 3: Visualizing High-Dimensional SGD Trajectories

Design: This final phase was designed to demystify the path of SGD in a high-dimensional parameter space. By recording the entire parameter vector of a CNN at frequent intervals and then applying Principal Component Analysis (PCA), we aimed to project the journey onto a visible 3D space.

Results and Analysis: The 3D trajectory in Figure 3 reveals a striking, non-trivial structure, providing visual evidence for the multi-stage learning narrative. The path clearly shows two distinct phases:

1. **Exploration/Finding:** An initial, long arc where the optimizer travels a large distance. The color change from yellow to purple shows the loss drops rapidly.
2. **Convergence/Tuning:** A subsequent, much tighter spiral where the optimizer has found a low-loss basin and is making small, fine-tuning adjustments.

This plot makes the abstract concept of a high-dimensional optimization path concrete, visually confirming the transition from a broad search for features to a final phase of refinement.

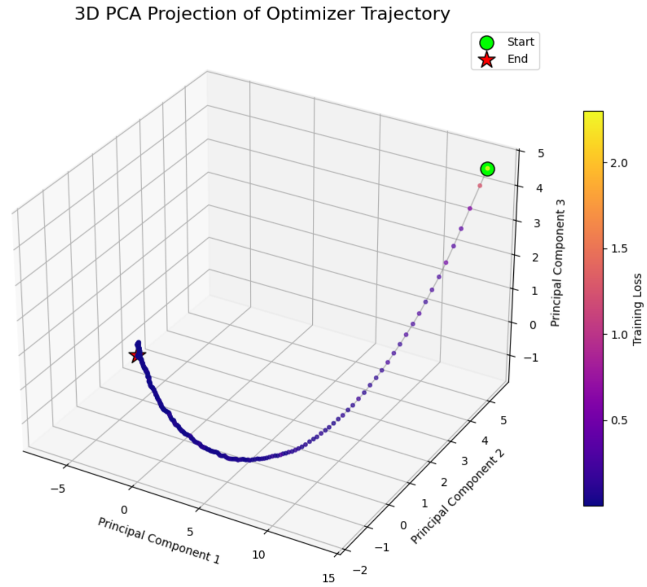

Figure 3. Phase 3 Result: A 3D PCA projection of the optimizer's path, showing an initial phase of rapid exploration (upper arc) followed by a final phase of convergence (lower spiral).

4. Conclusion

This project successfully bridged the gap between the deep theory of neural network training dynamics and practical, empirical visualization. The theoretical review confirmed that the success of over-parameterized models is driven by the implicit bias of SGD, which facilitates a rich, multi-stage feature learning process that is quantifiably superior to static kernel methods.

Our three-phase implementation provided tangible evidence for these theories. We demonstrated the profound efficiency of geometry-aware optimization, visualized feature learning as a concrete deviation from idealized NTK theory, and uncovered the structured, multi-phase path of SGD in its high-dimensional search. Together, these results provide a cohesive and compelling illustration of the core principles governing how modern neural networks learn.