

New Evidence of the Two-Phase Learning Dynamics of Neural Networks

Zhanpeng Zhou^{1*}, Yongyi Yang², Mahito Sugiyama^{3,4}, Junchi Yan^{1*}

¹Shanghai Jiao Tong University, ²University of Michigan, ³National Institute of Informatics,

⁴The Graduate University for Advanced Studies, SOKENDAI

{zzp1012, yanjunchi}@sjtu.edu.cn

Abstract

Understanding how deep neural networks learn remains a fundamental challenge in modern machine learning. A growing body of evidence suggests that training dynamics undergo a distinct phase transition, yet our understanding of this transition is still incomplete. In this paper, we introduce an interval-wise perspective that compares network states across a time window, revealing two new phenomena that illuminate the two-phase nature of deep learning. i) **The Chaos Effect**. By injecting an imperceptibly small parameter perturbation at various stages, we show that the response of the network to the perturbation exhibits a transition from chaotic to stable, suggesting there is an early critical period where the network is highly sensitive to initial conditions; ii) **The Cone Effect**. Tracking the evolution of the empirical Neural Tangent Kernel (eNTK), we find that after this transition point the model’s functional trajectory is confined to a narrow cone-shaped subset: while the kernel continues to change, it gets trapped into a tight angular region. Together, these effects provide a structural, dynamical view of how deep networks transition from sensitive exploration to stable refinement during training.

1 Introduction

Modern neural networks have become widely used in a broad range of fields [5, 28, 24]. However, a thorough understanding of how their capabilities and behaviors are developed throughout the training process remains incomplete, especially for deep models [27, 34]. Many recent studies have suggested, either implicitly or explicitly, that there is a *phase transition* point during the NN training, where the model’s properties and behaviors undergo substantial shifts before and after this time point. For example, Cohen et al. [7], Damian et al. [8], Wang et al. [29] showed that during training, the network first enters a progressive sharpening phase, and after which, the sharpness stabilizes and remains roughly constant for the rest of the training. Achille et al. [1] identified a *critical learning period* early in training, during which exposure to low-quality data can cause irreversible damage, while similar exposure later in training can be reversed.

Despite abundant evidence for the *two-phase phenomenon* [34, 7, 29, 12, 35, 1], a complete characterization and understanding of this phenomenon still lags behind. Moreover, most existing studies adopt a “point-wise” perspective: they primarily focus on examining specific properties of the network at isolated time points. This perspective, while informative, offers only a static snapshot of the model’s behavior, and does not capture the temporal dynamics of learning: how a property emerges, evolves, or vanishes as training progresses.

*Corresponding authors. This work extends our workshop paper: *On the Cone Effect in the Learning Dynamics* (accepted by ICLR 2025 Workshop DeLTa).

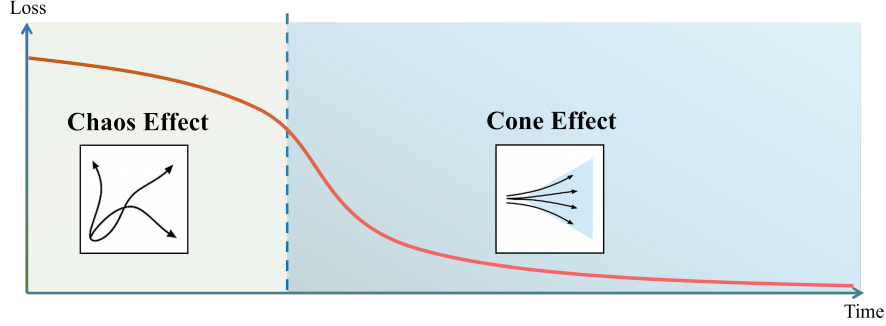


Figure 1: An illustration of the two-phase training dynamics. The optimization trajectory initially passes through a chaotic training phase, termed as the *chaos effect*; then undergoes a more stable, non-chaotic regime, namely the *cone effect*.

In this paper, to gain a deeper understanding of the two-phase phenomenon, we introduce two new empirical observations that exhibit characteristics of the phase transition. Crucially, these are what we call “interval-wise” phenomena: rather than analyzing a property at specific time points, we compare the model’s behavior across two different time points of training. We show that this novel approach reveals patterns that are otherwise invisible to point-wise analysis and offers new insights into the learning dynamics of neural networks. Specifically, we identify and investigate two distinct behaviors: *the Chaos Effect* and *the Cone Effect*. See Figure 1 for an illustration.

• **The Chaos Effect.** First, we observe that the learning dynamics of neural networks transition from a chaotic to a stable, non-chaotic regime during training. Specifically, we train two networks that are initialized identically and trained with the same stochastic gradient noise. At a specific time t_0 , we apply a small perturbation to the parameters of one of them, and then we compare the resulting parameters at a later time t_1 . Particularly, we observe an *inflection point* during the training process. We find that when t_0 is in the early stage of training, specifically before the inflection point, even a tiny perturbation leads to a significant divergence from the original training trajectory. This phenomenon indicates a high sensitivity of learned parameters to initial conditions, which is a hallmark of chaotic systems in physics. However, if t_0 is later in the training (after the inflection point), the divergence is minimal, suggesting that the system becomes increasingly stable as training progresses.

• **The Cone Effect.** Second, we discover that after the early training phase, the learning dynamics of neural networks keep constrained in a narrow cone in the function space. Specifically, we train a network, and starting from a chosen time point τ , we track the empirical Neural Tangent Kernel (eNTK) at later time steps and measure their deviation from the eNTK at τ . We observe that when τ is sufficiently large, the subsequent eNTKs remain confined within a narrow cone around the eNTKs at time τ . In contrast, if the τ is chosen in the early stages of training, the eNTKs experience chaotic and large unstructured changes over time, and no such confinement is observed.

Our contributions. In summary, we present an empirical study of neural network learning dynamics, with the aim of guiding future theoretical and experimental investigations in deep learning. Our main contributions are:

- *Interval-Wise Analysis Framework.* We propose a novel interval-wise framework for analyzing learning dynamics in neural networks;
- *Chaos and Cone Effects.* We identify and characterize two distinct phenomena: the Chaos Effect and the Cone Effect. Both exhibit clear two-phase behavior;
- *Structural Insights into Training Evolution.* We demonstrate how these effects uncover new structural properties of the neural network’s temporal evolution.

Roadmap. This rest of the paper is organized as follows. In Section 2, we recall the existing related works. In Section 3, we introduce our notation and the quantitative measures used in our empirical study. In Section 4, we present and analyze the Chaos Effect. In Section 5, we describe the Cone Effect. Finally, in Section 6, we summarize our findings and discuss potential limitations.

2 Related Work

As discussed in Section 1, many studies have suggested neural network training undergoes a phase transition in practice. In our view, this body of work can be roughly grouped into the following four categories:

- **Transition of the First-order Quantities.** A recent line of research investigated phase transitions in neural network training using gradient-based metrics. For example, Fort et al. [11] identified a clear branching point in the evolution velocity of the empirical Neural Tangent Kernel (eNTK), transitioning from a fast to a slow regime. Jastrzebski et al. [18] observed that the largest eigenvalue of the gradient covariance matrix increases monotonically in the early phase and then enters a regime of sustained oscillation.
- **Transition of the Second-order Quantities** Extensive works analyzed the Hessian spectrum of neural networks to characterize training phase transitions. For example, Cohen et al. [7], Wang et al. [29], Damian et al. [8], Frankle et al. [13] reported a two-phase pattern in the largest Hessian eigenvalue (the “sharpness”) during gradient descent: an initial phase of steady growth in sharpness accompanied by smooth loss reduction, followed by a plateau where sharpness fluctuates around a critical threshold. Ghorbani et al. [14] studied the full Hessian spectrum, clearly revealing a structural phase transition. Gur-Ari et al. [15] showed that the subspace spanned by the top Hessian eigenvectors stabilizes shortly after initialization, indicating a transition in the dominant eigenspace.
- **Transition of the Training Trajectories.** Another line of research focused on the phase transition behavior from the perspective of optimization trajectory and loss landscape geometry. For instance, Frankle et al. [12, 13], Fort et al. [11], Zhou et al. [35] introduced a *spawning method*: a network is initialized and trained for a few epochs, then spawned into two copies that continue training independently under different sources of SGD randomness (e.g., mini-batch order and data augmentation). They showed that when spawning occurs late in training, the two models, despite far in Euclidean distance, remain in the same basin of the loss landscape; whereas early spawning yields models in distinct basins. Singh et al. [26] analyzed the directionality of the optimization path and identified a corresponding phase transition from this directional viewpoint.
- **Transition in Behaviors/Ability.** From a more functional perspective, [34] demonstrate that neural networks learn concepts sequentially based on their signal strength. There exists a critical moment when one concept has been fully learned while another remains unlearned, clearly marking a phase boundary. [1] analyze how models respond to temporary corruptions in training data. They identify a “critical period” early in training during which such temporary corruption can cause permanent damage, while similar interventions later in training are reversible.

3 Preliminary and Methodology

Basic Notations. Throughout this paper, we focus on a classification task. Denote $[k] = \{1, 2, \dots, k\}$. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be the training set of size n , where $\mathbf{x}_i \in \mathbb{R}^{d_0}$ represents the i -th input and $y_i \in [c]$ represents the corresponding target. Here, c is the number of classes. Let $f : \mathcal{D} \times \mathbb{R}^p \rightarrow \mathbb{R}$ be the NN model, and thus $f(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}$ denotes the output of model f on the input \mathbf{x} with parameter $\boldsymbol{\theta} \in \mathbb{R}^p$. Let $\ell(f(\mathbf{x}_i, \boldsymbol{\theta}), y_i)$ be the loss at the i -th data point, simplified to $\ell_i(\boldsymbol{\theta})$. The total loss over the dataset \mathcal{D} is then denoted as $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \ell_i(\boldsymbol{\theta})$. We also use $\text{Err}_{\mathcal{D}}(\boldsymbol{\theta})/\text{Acc}_{\mathcal{D}}(\boldsymbol{\theta})$ to denote the classification error/accuracy of the network $f(\boldsymbol{\theta}; \cdot)$ on the training set \mathcal{D} .

Additionally, we use bold lowercase letters (e.g., \mathbf{x}) to denote vectors, and bold uppercase letters (e.g., \mathbf{A}) to represent matrices. For a matrix \mathbf{A} , let $\|\mathbf{A}\|_2$, $\|\mathbf{A}\|_F$, and $\text{Tr}(\mathbf{A})$ denote its operator norm, Frobenius norm, and trace respectively.

Parameter Dissimilarity. Following Singh et al. [26], we first introduce the *parameter dissimilarity* to measure the directionality of the optimization process. Specifically, given the training trajectory consisting of a sequence of checkpoints $\{\boldsymbol{\theta}_t\}_{t=0}^T$, we use the pairwise cosine dissimilarity to capture

the directional aspect of the trajectory. For any two time points $i, j \in [T]^2$, we define

$$(\mathbf{C})_{i,j} := 1 - \cos\langle \text{vec}(\boldsymbol{\theta}_i), \text{vec}(\boldsymbol{\theta}_j) \rangle = 1 - \langle \text{vec}(\boldsymbol{\theta}_i), \text{vec}(\boldsymbol{\theta}_j) \rangle / (\|\text{vec}(\boldsymbol{\theta}_i)\|_2 \|\text{vec}(\boldsymbol{\theta}_j)\|_2), \quad (1)$$

where $\text{vec}(\boldsymbol{\theta})$ denotes the flattened parameters of the network. We note that the sequence of checkpoints $\{\boldsymbol{\theta}_t\}_{t=1}^T$ represents only a subset of the entire training trajectory that encountered in practice, sampled at intervals of k points. The pairwise parameter dissimilarity matrix \mathbf{C} will serve as a qualitative measure for analyzing the directionality of the training dynamics.

Kernel Distance. Similar to Fort et al. [11], we use the *kernel distance* to quantify the evolution of neural networks in the function space. Consider minimizing the total loss $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$ using gradient flow; the evolution of the network function $f_t(\mathbf{x}_i) = f(\mathbf{x}_i, \boldsymbol{\theta}_t)$ can be written as:

$$\frac{df_t(\mathbf{x}_i)}{dt} = - \sum_{j=1}^n \left\langle \frac{\partial f_t(\mathbf{x}_i)}{\partial \boldsymbol{\theta}}, \frac{\partial f_t(\mathbf{x}_j)}{\partial \boldsymbol{\theta}} \right\rangle \frac{\partial \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_t)}{\partial f_t(\mathbf{x}_i)}. \quad (2)$$

Denoting $\mathbf{u}_t = \{f_t(\mathbf{x}_i)\}_{i=1}^n \in \mathbb{R}^n$ as the network outputs for all inputs, then a more compact form of Equation (2) is given by:

$$\frac{d\mathbf{u}_t}{dt} = -\mathbf{H}(\boldsymbol{\theta}_t) \nabla_{\mathbf{u}_t} \mathcal{L}(\boldsymbol{\theta}_t), \quad \text{where } (\mathbf{H}(\boldsymbol{\theta}_t))_{i,j} = \left\langle \frac{\partial f_t(\mathbf{x}_i)}{\partial \boldsymbol{\theta}}, \frac{\partial f_t(\mathbf{x}_j)}{\partial \boldsymbol{\theta}} \right\rangle, \quad (3)$$

Here, the kernel matrix $\mathbf{H}(\boldsymbol{\theta}_t) \in \mathbb{R}^{n \times n}$ is often termed as the *empirical neural tangent kernel* (eNTK). To study the evolution of neural network in function space, we measure the pairwise distance between the eNTK matrices at two different time points, namely *kernel distance*. For two time points $i, j \in [T]^2$, we define:

$$(\mathbf{S})_{i,j} := 1 - \frac{\langle \mathbf{H}(\boldsymbol{\theta}_i), \mathbf{H}(\boldsymbol{\theta}_j) \rangle}{\|\mathbf{H}(\boldsymbol{\theta}_i)\|_F \|\mathbf{H}(\boldsymbol{\theta}_j)\|_F}. \quad (4)$$

We will use the kernel distance matrix \mathbf{S} to analyze the training dynamics in function space.

Loss Barriers. In addition to the directional and functional aspect, we also investigate the geometry of the neural network’s loss landscape, focusing specifically on the regions encountered during the training process. In particular, we examine the *loss barriers* [12, 2] between any two points along the training trajectory. For any two points $i, j \in [T]^2$, we define:

$$(\mathbf{B})_{i,j} := \max_{\alpha} \mathcal{L}_{\mathcal{D}'}(\alpha \boldsymbol{\theta}_i + (1 - \alpha) \boldsymbol{\theta}_j) - \frac{1}{2} (\mathcal{L}_{\mathcal{D}'}(\boldsymbol{\theta}_i) + \mathcal{L}_{\mathcal{D}'}(\boldsymbol{\theta}_j)), \quad (5)$$

where \mathcal{D}' denotes the unseen test set. Indeed, the loss barrier is typically high for two independently trained neural networks, indicating that the two models reside in different and isolated “valleys”. Once the loss barrier approaches zero, i.e., $(\mathbf{B})_{i,j} \approx 0$, we say the two models $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ are linearly connected in the loss landscape [12, 35].

Disagreement Rate. Lastly, we concern about the similarity of the outputs at two points along the optimization trajectory. Specifically, we introduce the *disagreement rate* on the test data for any two points $i, j \in [T]^2$:

$$(\mathbf{D})_{i,j} := \mathbb{E}_{\mathbf{x} \in \mathcal{D}'} [\mathbf{1}(f(\mathbf{x}, \boldsymbol{\theta}_i) \neq f(\mathbf{x}, \boldsymbol{\theta}_j))], \quad (6)$$

where $\mathbf{1}(\cdot)$ is the indicator function and \mathcal{D}' is the test set. Notably, Jiang et al. [19] demonstrated that the test error of deep models can be approximated by the disagreement rate between two independently trained models on the same dataset, i.e., $\text{Err}_{\mathcal{D}'}(\boldsymbol{\theta}) \approx \text{Err}_{\mathcal{D}'}(\boldsymbol{\theta}') \approx \mathbb{E}_{\mathbf{x} \in \mathcal{D}'} [\mathbf{1}(f(\mathbf{x}, \boldsymbol{\theta}) \neq f(\mathbf{x}, \boldsymbol{\theta}'))]$.

Main Experimental Setup. We train the VGG-16 architecture [25] and the ResNet-20 architecture [16] on the CIFAR-10 dataset. Data augmentation techniques include random horizontal flips and random 32×32 pixel crops. Optimization is done using SGD with momentum (momentum set to 0.9). A weight decay of 1×10^{-4} is applied. The learning rate is initialized at 0.1 and is dropped by 10 times at 80 and 120 epochs. The total number of epochs is 160.

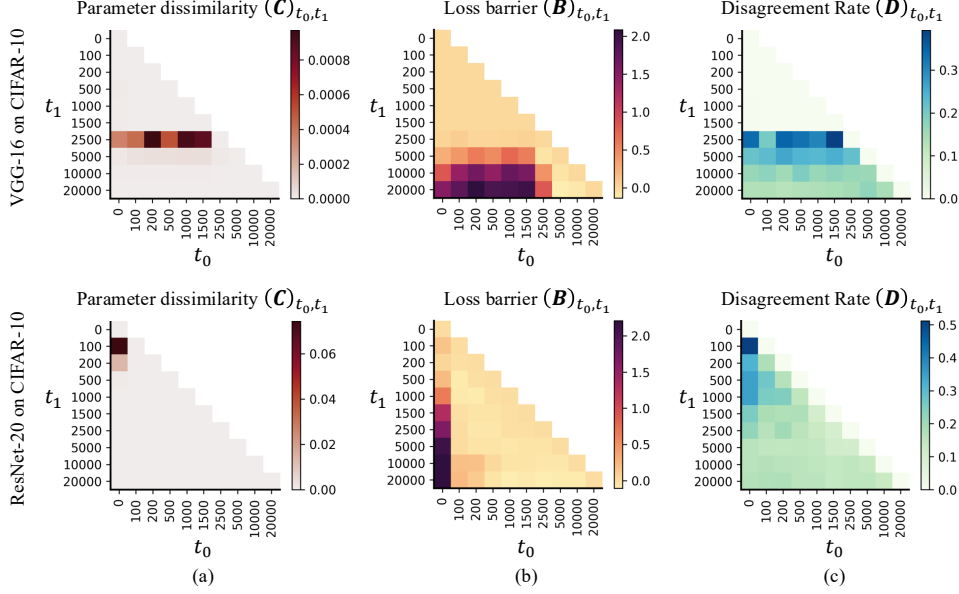


Figure 3: **The sensitivity of learning dynamics to tiny perturbations.** We apply the perturbation ϵ at the time point t_0 and compare resulting models at t_1 with different metrics. We set $\|\epsilon\|_0 = 10^{-7}$. Our results are reported for VGG-16 and ResNet-20 on CIFAR-10. **(a)** The parameter dissimilarity $(C)_{t_0, t_1}$. **(b)** The loss barrier $(B)_{t_0, t_1}$. **(c)** The disagreement rate $(D)_{t_0, t_1}$. Note that the t_0 and t_1 are presented in iterations, not epochs.

4 The Chaos Effect: Sensitivity of Learning Dynamics to Small Perturbations

Our investigation of the two-phase learning dynamics is motivated by a simple question:

How do injected noise influence the learning dynamics of neural networks? Is the training trajectory robust to small perturbation encountered during optimization.

To answer this question, in the section, we empirically investigate the sensitivity of neural network learning dynamics to small perturbations. Surprisingly, we find that an inflection point emerges during the training process. We observe that even a tiny perturbation applied before the inflection point can cause a significant divergence from the original optimization trajectory. In contrast, applying perturbations after the inflection point have far less impact. Thus we conjecture that the inflection point serves as a hallmark of the transition from a chaotic to a non-chaotic regime. We refer to this phenomenon as the *chaos effect*.

Experimental Design. We train two networks with identical initializations and the same stochastic gradient noise. However, at a specific time t_0 , we introduce a small perturbation ϵ to the parameters of one network, such that $\theta'_{t_0} = \theta_{t_0} + \epsilon$. We then compare the resulting models at a later time t_1 ($t_1 > t_0$), with the parameters θ'_{t_1} and θ_{t_1} respectively. The experimental design is illustrated in Figure 2. We consider a tiny perturbation here, where $\|\epsilon\|_0 = 10^{-7}$. We vary the time t_0, t_1 and compare the two resulting models using different metrics. The results are presented below.

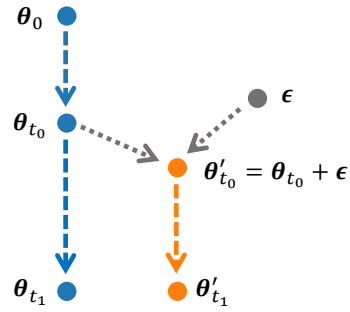


Figure 2: **The illustration of the injected perturbation.** θ_0 denotes the initialization. For both θ and θ' , the same stochastic gradient noise are applied during training.

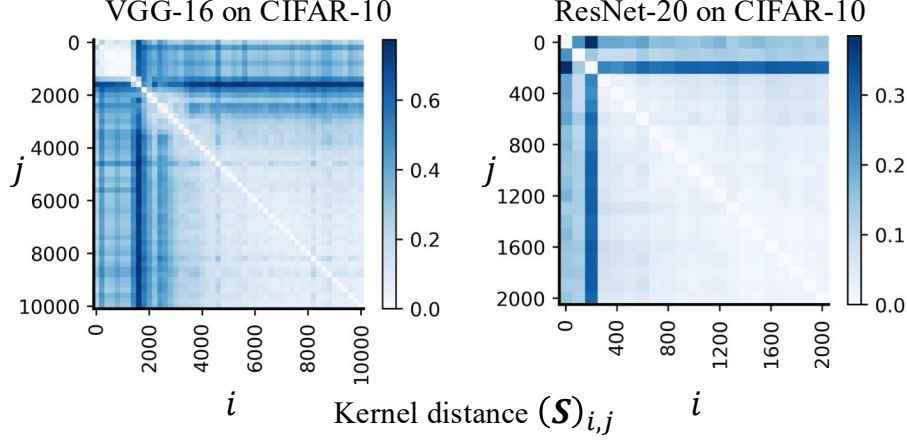


Figure 4: **The kernel distance between every pair of two points at the optimization trajectory $\{\theta_t\}_{t=1}^T$.** Our results are reported for both VGG-16 and ResNet-20 on CIFAR-10. Note that the i and j are presented in iterations, not epochs.

Finding I. Optimization trajectory changes its direction at an inflection point. As shown in Figure 3 (a), we first present the parameter dissimilarity for any pair of t_0 and t_1 (with $t_1 \geq t_0$). Notably, across different training settings, there exists a specific time t_1 at which the value of $(C)_{t_0, t_1}$ remains relatively high for all choices of t_0 . For example, in the case of VGG-16 on CIFAR-10, the dissimilarity $(C)_{t_0, t_1}$ consistently reaches its maximum when $t_1 = 2500$ iteration, regardless of the value of t_0 . Typically, a high value of $(C)_{t_0, t_1}$ indicates the directional change along the optimization trajectory. Therefore, it is evident that the optimization trajectory changes its direction at a fixed point, namely the *inflection point*.

Finding II. Tiny perturbations applied before the inflection point leads to significant loss barriers and disagreement rate. In Figure 3 (b), we also report the loss barrier between each pair of time points t_0, t_1 . We observe that even a tiny perturbation ($\|\epsilon\|_0 = 10^{-7}$) applied at an early time point t_0 could result in a substantial loss barrier between the resulting parameters θ_{t_1} and θ'_{t_1} in the later stage of training. This indicates that the two solutions likely reside in different, isolated “valleys” of the loss landscape. In Figure 3 (c), we further evaluate the disagreement rate between θ_{t_1} and θ'_{t_1} . The observed high disagreement rate firmly validates the functional dissimilarity between θ_{t_1} and θ'_{t_1} . Together, these results indicate that the learning dynamics pass through a chaotic regime during the early phase of training, where small perturbations might lead to pronounced loss barriers and functional divergence later on, namely the *chaos effect*.

Conjecture I. The inflection points marks the transition from a chaotic to a non-chaotic regime. Interestingly, the results observed for loss barriers and disagreement rate exhibit patterns similar to those for parameter dissimilarity. Taking VGG-16 on CIFAR-10 as an example, we observe that a significant loss barrier $(B)_{t_0, t_1}$ emerges only when $t_0 \leq 2500$ iterations and $t_1 > 2500$ iterations. Similar observations are also noted for the disagreement rate. Recall that the 2500 iteration marks the inflection point for VGG-16 on CIFAR-10. Therefore, we conjecture that the inflection point serves as a hallmark of the transition from a chaotic to a non-chaotic training regime.

To further verify this conjecture, we compute the kernel distance between every pair of points along the optimization trajectory within a single training run. Specifically, we train the neural network and obtain a sequence of checkpoints $\{\theta_t\}_{t=1}^T$. Then we measure the kernel distance for any two points $i, j \in [T]^2$, i.e., $(S)_{i, j}$. As shown in Figure 4, we observe that the eNTKs evolve significantly during the early phase of training, indicating a chaotic regime. Subsequently, the evolution of the eNTKs stabilizes, signifying a transition to a non-chaotic training phase. Notably, the transition point in the evolution of the eNTKs aligns with the inflection points identified in earlier experiments. For example, in the case of ResNet-20 on CIFAR-10, both the eNTK transition and the inflection point

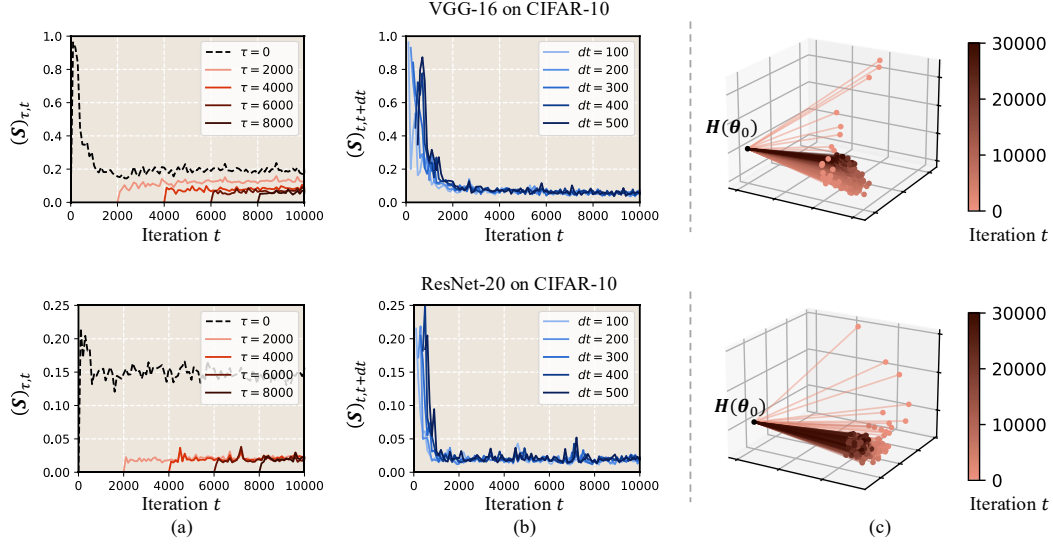


Figure 5: **Constrained learning dynamics in the second phase.** (a) The kernel distance between the current iterate θ_t and a reference point θ_τ v.s. training iteration t , where τ is varied. (b) The kernel distance between two adjacent iterates θ_t and θ_{t+dt} vs. training iteration t , where dt is varied. (c) The visualization of the changes of the eNTK matrices $H(\theta_t)$. The black dot represents position of the eNTK matrix at initialization, i.e., $H(\theta_0)$. The other dot represents the relative position of $H(\theta_t)$ at $t > 0$, with darker color indicating larger iteration.

occur around $100 \sim 500$ iteration. These results provide strong support for our conjecture that the inflection point marks the boundary between chaotic and non-chaotic training phases.

Section 4 Key Takeaways:

- Inflection points emerge during training, marking significant changes in the direction of the optimization trajectory.
- Tiny perturbations applied before the inflection point lead to substantial divergence later in training, indicating a chaotic regime in the early phase.
- The inflection point signifies a transition from a chaotic to a stable, non-chaotic training phase.

5 The Cone Effect: Constrained Learning Dynamics in the Second Phase

We have seen that the learning dynamics of neural networks undergo a transition from a highly chaotic to a more stable, non-chaotic phase, marked by a consistent inflection point. During the chaotic phase, the network evolves rapidly in function space, as evidenced by the significant changes in the eNTKs (see Figure 4). However, a natural question arises:

What characterizes the training dynamics during the subsequent stable, non-chaotic phase?

In this section, we dig deeper into the learning dynamics in the “second” phase. Surprisingly, we note that, contrary to the typical assumption of the *lazy training regime*, which we will elaborate on shortly, the neural network continues to evolve. However, this evolution is no longer unconstrained; instead, it is confined within a narrow, “cone”-like region in function space, a phenomenon we refer to as the *cone effect*.

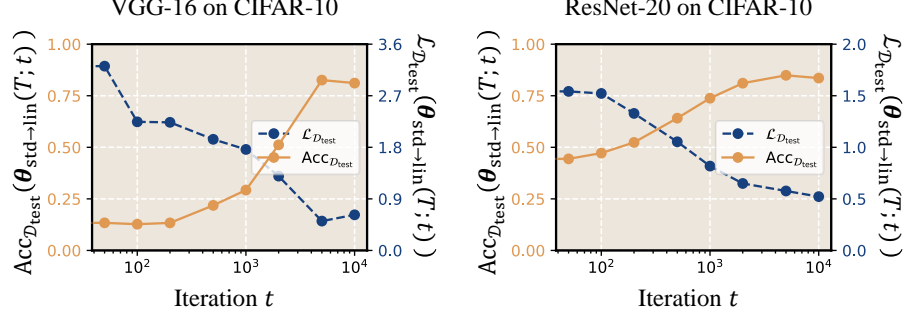


Figure 6: **The non-linear advantage of the cone effect.** Test accuracy $\text{Acc}_{\mathcal{D}_{\text{test}}}(\theta_{\text{std} \rightarrow \text{lin}}(T; t))$ (left) and Test loss $\mathcal{L}_{\mathcal{D}_{\text{test}}}(\theta_{\text{std} \rightarrow \text{lin}}(T; t))$ vs. the switching iteration t . $\theta_{\text{std} \rightarrow \text{lin}}(T; t)$ represents the model initially trained with standard method up to iteration t , followed by linearized training up to iteration T , where T is set to 10^4 .

The Lazy Regime. Numerous theoretical studies [10, 21, 9, 3, 37] proved that over-parameterized models can achieve zero training loss with minimal parameter variation. Moreover, Jacot et al. [17], Yang [33], Arora et al. [4], Lee et al. [20] showed that the learning dynamics of infinitely wide neural networks can be captured by a frozen kernel at initialization and are considered linear in theory. This behavior, often termed as *lazy regime*, typically occurs in over-parameterized models with large initialization and is considered undesirable in practice [6].

Previous studies often hypothesized that the learning dynamics during the later stages of training [32, 11, 22] or during fine-tuning [30, 31] are approximately linear, suggesting the presence of a *lazy training regime*. However, recent works [23, 36, 7, 29] have challenged this view, showing that linearized dynamics alone are insufficient to capture the behavior of neural networks in the later phases of training. Our observations are consistent with these findings: rather than exhibiting lazy dynamics, we identify the *cone effect* during the second training phase. This constrained yet non-trivial evolution plays a crucial role in shaping the network’s generalization ability.

Beyond the Lazy Regime: The Cone Effect. Through the kernel distance, we delve into the second phase of the learning dynamics. First, we compute the distance between the kernel matrices at two adjacent points θ_t and θ_{t+dt} . In Figure 5 (b), we observe that, across different values of dt , the kernel distance $(S)_{t,t+dt}$ between adjacent iterates θ_t and θ_{t+dt} is significant in the early training phase and then drops quickly to a low but non-negligible value. This result aligns with our previous results in Figure 3, where in the early training phase, the model evolves significantly in the function space. However, surprisingly, we note that in the later training phase, the values of $(S)_{t,t+dt}$ are upper-bounded by the same value for different dt . One possible explanation of this phenomenon is that during the second training phase, the eNTK matrix evolves in a constrained space.

To validate this, we further measure how the distance between the kernel matrices at the current iterate θ_t and a referent point θ_τ changes during training. As shown in Figure 5 (a), for different referent points $\tau \in \{2000, 4000, 6000, 8000\}$, the kernel distance between the current iterate and the reference point, i.e., $S(\theta_t, \theta_\tau)$, first increases and then keeps nearly constant in training. This result supports our claim and suggests that during later training phase, beyond the lazy regime, the model operates in a constrained function space. The visualization in Figure 5 (c) further confirms the existence of the cone effect, where a clear “cone” pattern is observed during the evolution of eNTK matrices.

The Non-linear Advantages of the Cone Effect. Despite the model evolving in a constrained function space in the second phase, it still provides significant advantages over completely lazy regime. To verify this, we consider a “switching” training method: we first train a neural network with standardized training method, and then switch to the linearized training (corresponding to the completely lazy regime) until T iterations. We vary the switching point t and obtain different solutions $\theta_{\text{std} \rightarrow \text{lin}}(T; t)$. In Figure 6, we observe that the test performance of $\theta_{\text{std} \rightarrow \text{lin}}(T; t)$ generally

increases with t , especially when $t > 2000$. This result implies that the cone effect in the later training phase still offers significant advantages over the entirely lazy regime.

Section 5 Key Takeaways:

- Despite the stable, non-chaotic dynamics in the later training phase, the neural networks the neural network continues to evolve within a narrow, constrained “cone”-like region of the function space, namely the cone effect.
- Compared to purely linearized training, the cone effect offers significant advantages for the performance of the final solution.

6 Conclusion and Limitations

In this paper, we introduced an interval-wise perspective on neural network training dynamics, moving beyond traditional point-wise analyses to reveal new insights into the learning process. Through this lens, we identified two novel empirical phenomena that characterize a two-phase transition in deep learning: the Chaos Effect and the Cone Effect. The Chaos Effect highlights a critical early period during which the training trajectory is highly sensitive to small perturbations, indicating chaotic behavior. In contrast, the Cone Effect demonstrates that, after this transition point, the model’s functional evolution becomes increasingly constrained within a narrow region of the function space, even though learning continues.

Together, these findings suggest a transition from an exploratory, unstable phase to a more stable, refinement-oriented phase during training. Our interval-wise analysis framework not only captures dynamic behaviors missed by point-wise approaches but also opens new directions for understanding and improving the training of deep neural networks. We hope this work inspires further research into the temporal structure of learning dynamics and the underlying mechanisms driving phase transitions in modern deep learning.

Limitations. We note that our current work primarily focuses on empirical findings, despite strongly related to optimization theory and NTK theory, we defer a thorough theoretical analysis to future work. We also note that our current experiments mainly focus on image classification tasks, though aligning with existing empirical studies on training dynamics [11]. We leave the exploration of empirical evidence beyond image classification as future direction.

References

- [1] Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep neural networks. *arXiv preprint arXiv:1711.08856*, 2017.
- [2] Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha S. Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *ICLR*. OpenReview.net, 2023.
- [3] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 09–15 Jun 2019.
- [4] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [5] Pedro Celard, Eva Lorenzo Iglesias, José Manuel Sorribes-Fdez, Rubén Romero, A Seara Vieira, and Lourdes Borrajo. A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. *Neural Computing and Applications*, 35(3):2291–2323, 2023.
- [6] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

- [7] Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.
- [8] Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=nhKHA59gXz>.
- [9] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 09–15 Jun 2019.
- [10] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.
- [11] Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:5850–5861, 2020.
- [12] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269. PMLR, 2020.
- [13] Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The early phase of neural network training. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=Hk11iRNFwS>.
- [14] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019.
- [15] Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.
- [17] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [18] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. *arXiv preprint arXiv:2002.09572*, 2020.
- [19] Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=Wv0GCEAQhxl>.
- [20] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [21] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [22] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=siCt4xZn5Ve>.
- [23] Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=0A9f2jZDGW>.

- [24] Xiaoli Ren, Xiaoyong Li, Kaijun Ren, Junqiang Song, Zichen Xu, Kefeng Deng, and Xiang Wang. Deep learning-based weather prediction: a survey. *Big Data Research*, 23:100178, 2021.
- [25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1409.1556>.
- [26] Sidak Pal Singh, Bobby He, Thomas Hofmann, and Bernhard Schölkopf. The directionality of optimization trajectories in neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=JY6P45sFDS>.
- [27] Namjoon Suh and Guang Cheng. A survey on statistical theory of deep learning: Approximation, training dynamics, and generative models. *Annual Review of Statistics and Its Application*, 12, 2024.
- [28] Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28694–28698, 2025.
- [29] Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability. *Advances in Neural Information Processing Systems*, 35:9983–9994, 2022.
- [30] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998. PMLR, 2022.
- [31] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7959–7971, June 2022.
- [32] Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/6651526b6fb8f29a00507de6a49ce30f-Paper.pdf.
- [33] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.
- [34] Yongyi Yang, Core Francisco Park, Ekdeep Singh Lubana, Maya Okawa, Wei Hu, and Hidenori Tanaka. Swing-by dynamics in concept learning and compositional generalization. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [35] Zhanpeng Zhou, Yongyi Yang, Xiaojiang Yang, Junchi Yan, and Wei Hu. Going beyond linear mode connectivity: The layerwise linear feature connectivity. *Advances in neural information processing systems*, 36:60853–60877, 2023.
- [36] Zhanpeng Zhou, Zijun Chen, Yilan Chen, Bo Zhang, and Junchi Yan. On the emergence of cross-task linearity in pretraining-finetuning paradigm. In *ICML*, 2024. URL <https://openreview.net/forum?id=qg6A1npEQH>.
- [37] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109:467–492, 2020.