
Trap of Feature Diversity in the Learning of MLPs

Dongrui Liu^{a*} Shaobo Wang^{b*} Jie Ren^a Kangrui Wang^a Sheng Yin^a

Huiqi Deng^a Quanshi Zhang^{a†}

^aShanghai Jiao Tong University ^bHarbin Institute of Technology

Abstract

In this paper, we focus on a typical two-phase phenomenon in the learning of multi-layer perceptrons (MLPs), and we aim to explain the reason for the decrease of feature diversity in the first phase. Specifically, people find that, in the training of MLPs, the training loss does not decrease significantly until the second phase. To this end, we further explore the reason why the diversity of features over different samples keeps decreasing in the first phase, which hurts the optimization of MLPs. We explain such a phenomenon in terms of the learning dynamics of MLPs. Furthermore, we theoretically explain why four typical operations can alleviate the decrease of the feature diversity. *The code will be released when the paper is accepted.*

1 Introduction

Deep neural networks (DNNs) have achieved significant success in various tasks. However, the essential reason for the superior performance of DNNs has not been fully investigated. Many studies aim to explain typical phenomena in DNNs, *e.g.* investigating the phenomenon of the lottery ticket hypothesis [16], explaining the double-descent phenomenon [23, 36, 19], understanding the information bottleneck hypothesis [48, 41], exploring the gradient noise and regularization [43, 34], and analyzing the nonlinear learning dynamics [27, 38].

In this paper, we focus on the learning dynamics of multi-layer perceptrons (MLPs). It has been widely discovered that when MLPs have many layers and the optimization is difficult, the learning process of MLPs is more likely to have two phases. As Figure 1(b) shows, the first phase is usually not long, in which the training loss does not decrease. Then, in the second phase, the training loss suddenly begins to decrease. Note that when the task is simple the first phase may be extremely short (*e.g.* a few iterations within an epoch), and thus cannot be observed.

To this end, we discover an interesting phenomenon in the first phase, *i.e.*, as Figure 1(a) shows, *features of different categories become increasingly similar to each other*. The feature diversity keeps decreasing (the cosine similarity between features keeps increasing) until the second phase.

This phenomenon is widely shared by MLPs, convolutional neural networks, and recurrent neural networks (see both Figure 2 and the supplementary material). Neural networks trained with different depths and widths, different activation functions, and different learning rates, on different types of data may all exhibit such a phenomenon, which hurts the optimization.

More crucially, the investigation and alleviation of the two-phase phenomenon are of considerable value, because this is a typical learning-sticking problem with MLPs. As Figure 1(c) shows, when the loss minimization gets stuck, we can consider it as a strong first phase with an infinite length.

*Equal contribution

†This research is done under the supervision of Dr. Quanshi Zhang. He is with the Department of Computer Science and Engineering, the John Hopcroft Center and the MoE Key Lab of Artificial Intelligence, AI Institute, at the Shanghai Jiao Tong University, China. Correspondence to: Quanshi Zhang <zqs1022@sjtu.edu.cn>.

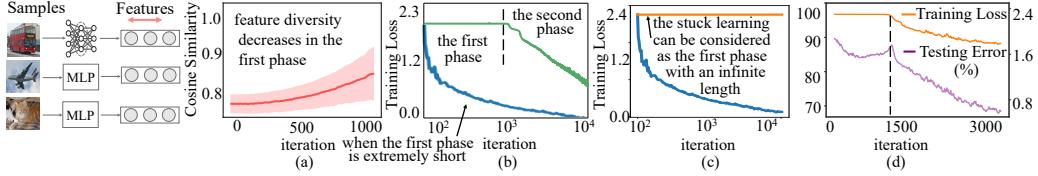


Figure 1: (a, b) We aim to explain the two-phase phenomenon in the training of MLPs, *i.e.*, the reason why the cosine similarity between features of different categories keeps increasing in the first phase in (a). (c) When the loss minimization gets stuck (orange curve), we can consider it as the first phase with an infinite length. The learning-sticking problem can be solved by techniques of shortening the first phase (blue curve). (d) The two-phase phenomenon aligns the double-descent behavior.

Therefore, we aim to investigate the optimization behavior in the first phase, *i.e.*, why and how the feature diversity decreases during the early training process of the MLP. To this end, we find that in intermediate layers of the MLP, both features and parameters are mainly optimized towards a special direction, namely the primary common direction. In this study, we theoretically clarify certain dynamics of the optimization, which increases the likelihood of further enhancing such a primary common direction, just like a self-enhanced system. This may explain the decrease of the feature diversity in early iterations.

Based on our theoretical analysis, we further discover and explain the reason why four typical operations (*i.e.*, batch normalization, momentum, initialization, and L_2 regularization) can effectively alleviate the two-phase phenomenon. The above findings provide theoretical supports for heuristic solutions to the learning-sticking problem.

Although the decrease of feature diversity can be alleviated by traditional operations, the core contribution of this study is to **discover and explain a fundamental yet counter-intuitive two-phase phenomenon with the MLP, which has not been theoretically explained** for a long time. Instead, previous studies simply owed the learning-sticking problem in the first phase to the difficulty of the training task, without insightful analysis or theoretically supported solutions.

Contributions of this study can be summarized as follows. (1) We discover the common phenomenon of feature diversity decreasing in early learning of the MLP, which has been ignored for a long time. (2) We explain this phenomenon from the perspective of learning dynamics. (3) We explain why four types of operations can alleviate the decrease of feature diversity.

2 Related Work

Understanding the optimization and the representation capacity of DNNs is an important direction to explain DNNs. The information bottleneck theory [48, 41] quantitatively explained the information encoded by features in intermediate layers of DNNs. Xu and Raginsky [49], Achille and Soatto [1], and Cheng et al. [8] used the information bottleneck theory to evaluate and improve the DNN’s representation capacity. Arpit et al. [5] analyzed the representation capacity of DNNs with real training data and noises. In addition, several metrics were proposed to measure the generalization capacity or robustness of DNNs, including the stiffness [15], the sensitivity metrics [37], the Fourier analysis [51], and the CLEVER score [47]. Some studies focused on proving the generalization bound of DNNs [32, 31]. In comparison, we explain the MLP from the perspective of the learning dynamics, *i.e.*, we explain the decrease of feature diversity in early iterations of the MLP.

Analyzing the learning dynamics is another perspective to understand DNNs. Many studies analyzed the local minima in the optimization landscape of linear networks [7, 40, 18, 10] and nonlinear networks [9, 25, 39]. Some studies discussed the convergence rate of gradient descent on separable data [45, 50, 35]. Hoffer et al. [20] and Jastrz̄ebski et al. [24] have investigated the effects of the batch size and the learning rate on SGD dynamics. In addition, some studies analyzed the dynamics of gradient descent in the overparameterization regime [4, 22, 30, 14]. Unlike previous studies, we analyze the learning dynamics of features and weights of the MLP, in order to explain the decrease of feature diversity in the first phase.

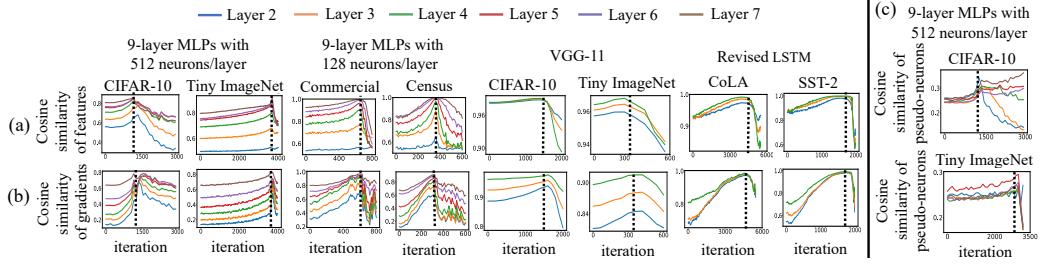


Figure 2: The two-phase phenomenon. (a) Cosine similarity of features between samples in different categories $\mathbb{E}_{x,x' \in X} [\cos(F_t^{(l)}|_x, F_t^{(l)}|_{x'})]$ keeps increasing in the first phase (left to the dotted line), until the second phase. The low cosine similarity indicates the high diversity. (b) Cosine similarity of gradients w.r.t. features between different samples of a category $\mathbb{E}_{x,x' \in X_c} [\cos(\dot{F}_t^{(l)}|_x, \dot{F}_t^{(l)}|_{x'})]$ keeps increasing in the first phase until the second phase, where X_c denotes samples of the category c . (c) Cosine similarity of weight changes between “pseudo-neurons” in a layer $\mathbb{E}_{x \in X} \cos(\Delta w_{t,i}^{(l)}|_x, \Delta w_{t,j}^{(l)}|_x)$ keeps increasing in the first phase. *Please see the supplementary material for results on more DNNs.*

3 Discovering the decrease of feature diversity

The training process of the MLP can usually be divided into the following two phases, according to the training loss. As Figure 1(b) shows, the training loss does not decrease significantly in the first phase, and the training loss suddenly begins to decrease in the second phase. In this paper, we **discover an interesting phenomenon in the first phase that both the diversity of intermediate-layer features over different samples and the diversity of gradients w.r.t. features keep decreasing**. In other words, in the first phase, the cosine similarity between features and the cosine similarity between feature gradients keep increasing.

We consider an MLP f with L concatenated linear layers, each being followed by a ReLU layer. Only the last linear layer is followed by a softmax operation. Let $W_t^{(l)} \in \mathbb{R}^{h \times d}$ denote the weight matrix of the l -th linear layer with h neurons ($1 \leq l \leq L$), and $W_t^{(l)}$ has been learned for t iterations. Given a specific input sample x , the layer-wise forward propagation in the l -th layer is represented as

$$F_t^{(l)} = \text{ReLU}(W_t^{(l)} F_t^{(l-1)}) = D_t^{(l)} W_t^{(l)} F_t^{(l-1)}, \quad (1)$$

where $F_t^{(l)} \in \mathbb{R}^h$ denotes the output feature of the l -th layer after the t -th iteration. $D_t^{(l)}$ denotes a diagonal matrix, which represents gating states in the ReLU layer and $D_{t,(i,i)}^{(l)} \in \{0, 1\}$.

Visualizing that the two-phase phenomenon is widely shared by different DNNs learned for different tasks. We observed such a two-phase phenomenon on MLPs, VGG-11 [42], and the revised long short-term memory (LSTM) on different types of data, including image data (MNIST [29], CIFAR-10 [26], and the Tiny ImageNet dataset [28]), tabular data (two UCI datasets of census income and TV news [6]), and natural language data (CoLA [46], SST-2 [44], and AGNews [13]). We also observed MLPs with Leaky ReLU layers [33], with different learning rates, and with different batch sizes. Figure 2(a,b) shows the two-phase phenomenon on some DNNs, and please see the supplementary material for results on more DNNs. Specifically, given two input samples x_1 and x_2 , the feature similarity between x_1 and x_2 $\cos(F_t^{(l)}|_{x_1}, F_t^{(l)}|_{x_2})$, and the cosine similarity of gradients $\cos(\dot{F}_t^{(l)}|_{x_1}, \dot{F}_t^{(l)}|_{x_2})$ keep increasing, which demonstrates the phenomenon. $\dot{F}_t^{(l)}$ denotes the gradient of the loss w.r.t. the feature $F_t^{(l)}$ in Eq. (1).

Note that such a decrease of feature diversity sometimes appears in very early epochs (or iterations) of the training process. More crucially, as Figure 1(c) shows, when the loss directly decreases in the first epoch, it can be considered as a short first phase in very early iterations. Besides, the learning process is sometimes stuck when the task is difficult. Then we **can consider the learning-sticking problem** as the first phase with an infinite length (please see the supplementary material for more discussions).

Connection to the epoch-wise double descent. Figure 1(d) shows that the epoch-wise double-descent behavior [36, 19] is temporally aligned with the first phase in the aforementioned two-phase phenomenon. Please see the supplementary material for more discussions. Instead of explaining the epoch-wise double descent, in this paper, we mainly explain the decrease of the feature diversity.

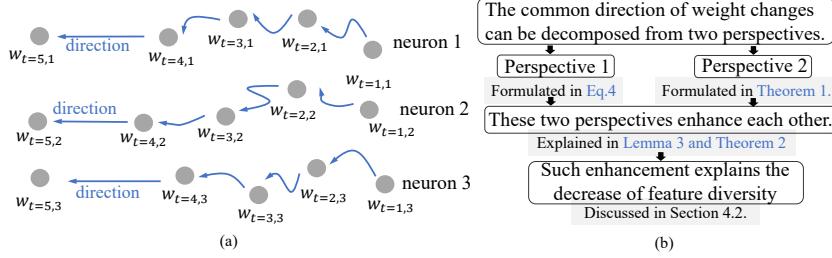


Figure 3: (a) Weights of neurons are changed towards a common direction. (b) The logic of explaining the decrease of feature diversity.

4 Explaining the dynamics of the decrease of feature diversity

In this section, we aim to investigate certain dynamics of network parameters, so as to explain the condition that may boost the likelihood of decreasing the feature diversity in early epochs under some common assumptions. In Section 4.1, we find that the decreasing diversity of feature gradients over different samples is owing to the phenomenon that different neurons in a layer are optimized towards a common direction in the first phase. Then, in Section 4.2, we further clarify learning dynamics that may potentially enhance the significance of the common direction, just like a self-enhanced system. The self-enhanced common direction boosts the likelihood of the feature diversity's decreasing. The overall logic of the explanation is illustrated in Figure 3(b). In Section 4.3, we explain why four types of operations can alleviate the decrease of feature diversity based on our analysis.

4.1 Two perspectives to analyze the common direction of learning effects

The decreasing diversity of feature gradients over different samples is owing to the phenomenon that different neurons in a layer are optimized towards a common direction in the first phase. For example, as Figure 3(a) shows, at the beginning of the learning, different neurons are optimized towards different directions. Along with the learning process, different neurons gradually tend to be optimized along a similar direction. According to the forward propagation in Eq. (1), feature gradients $\dot{F}_t^{(l)}$ at the l -th layer during the back propagation can be rewritten as

$$\dot{F}_t^{(l-1)} = W_t^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)}. \quad (2)$$

In this way, we can consider feature gradients $\dot{F}_t^{(l-1)}$ as the result of a pseudo-forward propagation. In the pseudo-forward propagation, the input is the gradient w.r.t. features of the l -th layer $\dot{F}_t^{(l)}$, and the output is the gradient of the $(l-1)$ -th layer $\dot{F}_t^{(l-1)} \in \mathbb{R}^d$. Accordingly, the equivalent weight matrix is given as $W_t^{(l)\top} = [w_{t,1}^{(l)}, w_{t,2}^{(l)}, \dots, w_{t,d}^{(l)}]^\top \in \mathbb{R}^{d \times h}$, which consists of d “pseudo-neurons.”

Before explaining the common direction phenomenon, let us first clarify that the increasing similarity between feature gradients $\dot{F}_t^{(l-1)}$ of different samples is caused by the experimental observation that different “pseudo-neurons” are approximately optimized to a common direction. Let us consider the following conjecture that different “pseudo-neurons” $[w_{t,1}^{(l)}, w_{t,2}^{(l)}, \dots, w_{t,d}^{(l)}]^\top$ have been roughly optimized towards a common direction $C^{(l)} \in \mathbb{R}^h$ for many epochs. Then, we can roughly represent $w_{t,i}^{(l)} = \beta_i C^{(l)} + \epsilon_i$, where $\beta_i \in \mathbb{R}$, and $\epsilon_i \in \mathbb{R}^h$. Here, $\epsilon_i \in \mathbb{R}^h$ denotes a small residual. According to Eq. (2), we have

$$\dot{F}_t^{(l-1)} = (C^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)}) \cdot \beta + \epsilon D_t^{(l)} \dot{F}_t^{(l)} \quad (3)$$

Thus, if β is large enough (*i.e.*, keeping optimizing $W_t^{(l)\top}$ along the common direction $C^{(l)}$ for a long time), then the feature gradients $\dot{F}_t^{(l-1)}$ of different samples will be roughly parallel to the same vector β . This is because $C^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)}$ is a scalar and the term $\epsilon D_t^{(l)} \dot{F}_t^{(l)}$ is small. In other words, the diversity between feature gradients $\dot{F}_t^{(l-1)}$ of different samples decreases. Here, $\beta = [\beta_1, \beta_2, \dots, \beta_d]$, and $\epsilon = [\epsilon_1, \epsilon_2, \dots, \epsilon_d]^\top$.

Therefore, the first core task of proving the decreasing diversity of feature gradients is to explain the conjecture of the common optimization direction shared by different “pseudo-neurons.” To this end, let us first propose two perspectives to disentangle such a common direction.

Perspective 1. Perspective 1 focuses on the weight change in a certain layer. For clarity, we omit the superscript (l) to simplify the notation in the following paragraphs in Section 4.1, *i.e.*, $\Delta w_{t,i}^{(l)} \xrightarrow{\text{simplify}} \Delta w_{t,i}$, $\Delta W_t^{(l)} \xrightarrow{\text{simplify}} \Delta W_t$, $C^{(l)} \xrightarrow{\text{simplify}} C$. Let $\Delta W_t^\top = [\Delta w_{t,1}, \Delta w_{t,2}, \dots, \Delta w_{t,d}]^\top$ denote

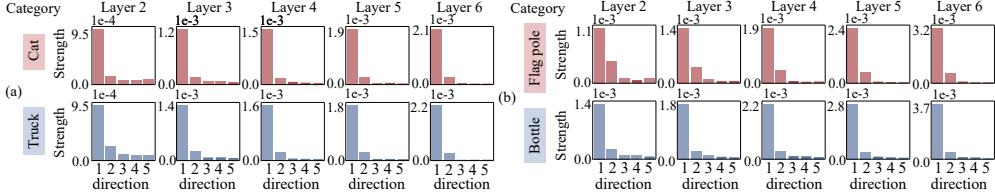


Figure 4: The strength of different common directions in the (a) CIFAR-10 dataset, (b) Tiny ImageNet dataset. We trained 9-layer MLPs, where each layer of the MLP had 512 neurons. We illustrated results on the two categories with the highest training accuracies. $s_i = \|\mathbf{C}_i \Delta \bar{V}_i^\top\|_F$ measures the strength of weight changes along the i -th common direction, where $\Delta \bar{V}_i = \mathbb{E}_t[\Delta \bar{V}_{i,t}]$. The strength of the primary direction was much greater than the strength of other directions. Please see the supplementary material for results on the MNIST.

weight changes of d “pseudo-neurons” in the l -th layer. We decompose ΔW_t^\top into the component along a common direction C and a component along other directions as follows.

$$\Delta W_t^\top = \Delta V_t C^\top + \Delta \varepsilon_t, \quad (4)$$

where $\Delta V_t = [\Delta v_{t,1}, \Delta v_{t,2}, \dots, \Delta v_{t,d}] \in \mathbb{R}^d$ denotes the coefficient vector for weight changes of different “pseudo-neurons” along the common direction C , i.e., using $\Delta v_{t,i} C^\top$ to approximate $\Delta w_{t,i}^\top$. Specifically, $\Delta \varepsilon_t$ is relatively small “noise” term, which is orthogonal to C , i.e., $\Delta \varepsilon_t C = \mathbf{0}$. In this way, the common direction C can be determined by minimizing the “noise” term over different samples across different iterations, as follows.

$$\min_{C, \Delta V_t|x} (\mathbb{E}_{t \in [T_{\text{start}}, T_{\text{end}}]} \mathbb{E}_{x \in X} \|\Delta \varepsilon_t|x\|_F^2), \quad \text{s.t. } \Delta \varepsilon_t|x = \Delta W_t^\top|x - \Delta V_t|x C^\top \quad (5)$$

Lemma 1. (Proof in the supplementary material) For the decomposition $\Delta W_t^\top = \Delta V_t C^\top + \Delta \varepsilon_t$, given weight changes over different samples ΔW_t^\top , we can compute the common direction C in Eq. (5) and obtain $\Delta V_t = \frac{\Delta W_t^\top C}{C^\top C}$ and $\Delta \varepsilon_t = \Delta W_t^\top - \Delta V_t^\top \frac{C C^\top}{C^\top C}$ s.t. $\Delta \varepsilon_t C = \mathbf{0}$. Such settings minimize $\|\Delta \varepsilon_t\|_F$.

Lemma 2. (We can also decompose the weight $W_t^{(l)}$ into the component along the common direction C and the component ε_t in other directions. Proof is in the supplementary material.) Given the weight W_t^\top and the common direction C , the decomposition $W_t^\top = V_t C^\top + \varepsilon_t$ can be conducted as $V_t = \frac{W_t^\top C}{C^\top C}$ and $\varepsilon_t = W_t^\top - W_t^\top \frac{C C^\top}{C^\top C}$ s.t. $\varepsilon_t C = \mathbf{0}$. Such settings minimize $\|\varepsilon_t\|_F$.

• **Experimental verification of the strength of the primary common direction C .** To this end, let us focus on the average weight change over different samples $\Delta \bar{W}_t = \mathbb{E}_{x \in X} \Delta W_t|x$. Then, we decompose $\Delta \bar{W}_t$ into components along five common directions as $\Delta \bar{W}_t = C_1 \Delta \bar{V}_{1,t}^\top + C_2 \Delta \bar{V}_{2,t}^\top + \dots + C_5 \Delta \bar{V}_{5,t}^\top + \Delta \bar{\varepsilon}_{5,t}^\top$, where $C_1 = C$ is termed the *primary common direction*. C_2, C_3, C_4 and C_5 represent the second, third, forth, and fifth common directions, respectively. C_1, C_2, C_3, C_4 , and C_5 are orthogonal to each other. C_i and $\Delta \bar{V}_{i,t}$ are computed based on Eq. (5) when we remove the first $i-1$ components along the direction C, \dots, C_{i-1} from the $\Delta \bar{W}_t$. Figure 4 shows that the strength of the primary common component $C_1 \Delta \bar{V}_1^\top$ was approximately ten times greater than the strength of the secondary common component $C_2 \Delta \bar{V}_2^\top$. Please see the supplementary material for more discussions.

Perspective 2 by using gradients w.r.t. features $\dot{F}_t^{(l+1)}$.

We decompose the weight change by considering the influence of the common direction of the upper layer $C^{(l+1)}$. In order to distinguish variables belonging to different layers, we add the superscript (l) back to $\Delta W_t^{(l)}$, $\Delta V_t^{(l)}$, and $\Delta \varepsilon_t^{(l)}$ to denote the layer in the following paragraphs.

Theorem 1. (Proof in the supplementary material) The weight change made by a sample can be decomposed into $(h+1)$ terms after the t -th iteration as follows.

$$\Delta W_t^{(l)} = \Delta W_{\text{primary},t}^{(l)} + \sum_{k=1}^h \Delta W_{\text{noise},t}^{(l,k)} \xrightarrow{\text{rewritten}} \Gamma_t^{(l)} F_t^{(l-1)\top} + \kappa_t^{(l)\top}, \quad (6)$$

where $\Delta W_{\text{primary},t}^{(l)} = D_t^{(l)} V_t^{(l+1)} C^{(l+1)\top} C^{(l+1)} \Delta V_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2$ denotes the component along the primary common direction, and $\Delta W_{\text{noise},t}^{(l,k)} = D_t^{(l)} \varepsilon_t^{(l+1,k)} \Delta \varepsilon_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2$ denotes the component along the k -th common direction in the noise term. $\varepsilon_t^{(l+1,k)} = \Sigma_{kk} \mathcal{U}_k \mathcal{V}_k^\top$, where the SVD of $\varepsilon_t^{(l+1)} \in \mathbb{R}^{h \times h'}$ is given as $\varepsilon_t^{(l+1)} = \mathcal{U} \Sigma \mathcal{V}^\top$ ($h \leq h'$), and Σ_{kk} denotes

Table 1: Strength of components of weight changes along the common direction and other directions. We trained 9-layer MLPs on the CIFAR-10 dataset and the Tiny ImageNet dataset, respectively. Each layer of the MLP had 512 neurons. The strength of the primary common direction was much greater than those of other directions. The supplementary material provides results on the MNIST dataset and explains the phenomenon that $S_1^{(l)}$, $S_2^{(l)}$, and $S_3^{(l)}$ does not decrease monotonically.

| | Category | Cat | | | | | Truck | | | | |
|---------------|----------------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| | | Layer 2 | Layer 3 | Layer 4 | Layer 5 | Layer 6 | Layer 2 | Layer 3 | Layer 4 | Layer 5 | Layer 6 |
| CIFAR-10 | $S \times 10^{-3}$ | 154.0 \pm 17.1 | 176.5 \pm 16.8 | 201.6 \pm 18.7 | 253.6 \pm 24.6 | 277.4 \pm 25.6 | 169.9 \pm 20.8 | 208.1 \pm 21.5 | 223.6 \pm 20.1 | 248.4 \pm 19.2 | 281.5 \pm 20.4 |
| | $S_{\text{primary}}^{(l)}$ | 11.5 \pm 1.5 | 13.0 \pm 0.9 | 11.6 \pm 1.7 | 16.1 \pm 1.8 | 9.0 \pm 0.8 | 15.6 \pm 2.1 | 14.0 \pm 1.8 | 14.3 \pm 1.1 | 14.3 \pm 1.7 | 10.0 \pm 1.1 |
| | $S_1^{(l)}$ | 12.7 \pm 1.7 | 11.9 \pm 1.3 | 10.9 \pm 1.3 | 11.9 \pm 0.8 | 8.8 \pm 1.1 | 14.4 \pm 1.4 | 15.1 \pm 2.0 | 11.3 \pm 1.4 | 12.3 \pm 0.9 | 12.9 \pm 1.2 |
| | $S_2^{(l)}$ | 11.0 \pm 1.1 | 14.4 \pm 1.7 | 12.5 \pm 2.2 | 13.9 \pm 1.7 | 8.6 \pm 1.1 | 14.3 \pm 2.2 | 12.4 \pm 1.9 | 12.8 \pm 1.6 | 13.1 \pm 1.2 | 9.7 \pm 1.0 |
| | $S_3^{(l)}$ | | | | | | | | | | |
| Tiny ImageNet | Category | Flagpole | | | | | Bottle | | | | |
| | $S \times 10^{-3}$ | Layer 2 | Layer 3 | Layer 4 | Layer 5 | Layer 6 | Layer 2 | Layer 3 | Layer 4 | Layer 5 | Layer 6 |
| | $S_{\text{primary}}^{(l)}$ | 97.8 \pm 3.7 | 143.9 \pm 5.6 | 198.9 \pm 8.1 | 259.8 \pm 10.1 | 322.8 \pm 12.7 | 202.3 \pm 12.2 | 234.4 \pm 13.1 | 276.8 \pm 13.9 | 345.2 \pm 16.6 | 440.2 \pm 22.2 |
| | $S_1^{(l)}$ | 10.6 \pm 0.9 | 9.5 \pm 0.8 | 14.4 \pm 1.4 | 24.9 \pm 1.3 | 8.8 \pm 1.0 | 10.3 \pm 1.4 | 11.2 \pm 1.6 | 12.2 \pm 1.3 | 11.9 \pm 1.1 | 13.2 \pm 1.6 |
| | $S_2^{(l)}$ | 7.5 \pm 0.9 | 7.9 \pm 1.2 | 9.7 \pm 1.2 | 9.2 \pm 1.2 | 8.3 \pm 0.6 | 10.4 \pm 1.1 | 11.6 \pm 1.0 | 13.8 \pm 1.3 | 10.0 \pm 0.8 | 13.6 \pm 1.2 |
| | $S_3^{(l)}$ | 7.1 \pm 0.8 | 9.1 \pm 1.1 | 11.3 \pm 1.0 | 17.9 \pm 2.2 | 16.6 \pm 1.5 | 11.6 \pm 1.4 | 15.7 \pm 1.4 | 10.7 \pm 1.1 | 10.8 \pm 1.2 | 19.8 \pm 1.6 |

the k -th singular value $\in \mathbb{R}$. $\varepsilon_t^{(l+1)} = \sum_k \varepsilon_t^{(l+1,k)}$. \mathcal{U}_k and \mathcal{V}_k denote the k -th column of the matrix \mathcal{U} and \mathcal{V} , respectively. Besides, we have $\forall k \in \{1, 2, \dots, h\}$, $\mathcal{U}_k^\top C^{(l+1)} = 0$. Consequently, we have $\Gamma_t^{(l)} = D_t^{(l)} V_t^{(l+1)\top} C^{(l+1)\top} C^{(l+1)} \Delta V_t^{(l+1)\top} F_t^{(l)} / \|F_t^{(l)}\|_2^2 \in \mathbb{R}^h$, and $\kappa_t^{(l)\top} = D_t^{(l)} \varepsilon_t^{(l+1)} \Delta \varepsilon_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2 \in \mathbb{R}^{h \times d}$.

Given weight changes $\Delta W_t^{(l)}$ made by a sample x , the primary term $\Delta W_{\text{primary},t}^{(l)}$ represents the component of weight changes along the common direction $C^{(l+1)}$. The k -th noise term $\Delta W_{\text{noise},t}^{(l,k)}$ represents the component along the k -th direction \mathcal{U}_k , which is orthogonal to $C^{(l+1)}$.

• *Experimental verification of the significant strength of the component along the common direction $C^{(l+1)}$.* To this end, we computed the average strength of the component along the common direction $C^{(l+1)}$ over all samples in X as $S_{\text{primary}}^{(l)} = \mathbb{E}_{t \in [T_{\text{start}}, T_{\text{end}}]} \mathbb{E}_{x \in X} [\|\Delta W_{\text{primary},t}|_x\|_F]$. Similarly, the strength of the component along the k -th noise direction was computed as $S_k^{(l)} = \mathbb{E}_{t \in [T_{\text{start}}, T_{\text{end}}]} \mathbb{E}_{x \in X} [\|\Delta W_{\text{noise},t}^{(l,k)}|_x\|_F]$. Table 1 illustrates that the strength of the primary component $S_{\text{primary}}^{(l)}$ was more than ten times greater than the strength of components along other noise directions $S_1^{(l)}, S_2^{(l)}$, and $S_3^{(l)}$.

Discussion about comparing with the sum of all other directions' significance. According to Table 1, it seems that the sum of strengths of components along other directions is also large. However, different directions decomposed by the above method are orthogonal to each other. Therefore, weight changes along different directions are independent, and their strengths cannot be summed up. Thus, we can directly compare the strength of the component of weight changes along each direction to verify the significant strength of the primary direction.

4.2 Explaining the enhancement of the significance of the common direction

The previous subsection owes the decreasing diversity of feature gradients to the typical phenomenon that there exists a common optimization direction shared by different “pseudo-neurons.” In the current subsection, we explain that the common optimization direction phenomenon is very likely to be further enhanced, just like a self-enhanced system, when features $F_t^{(l-1)}$ of different samples have been pushed a little bit towards a specific common direction (see the following background assumption). The self-enhancement of the optimization direction will explain the decreasing diversity.

According to Eq. (4) and Eq. (6), weight changes made by the sample x can be given as

$$\text{Perspective 1: } \Delta W_t^{(l)} = C^{(l)} \Delta V_t^{(l)\top} + \Delta \varepsilon_t^{(l)\top} \quad \text{Perspective 2: } \Delta W_t^{(l)} = \Gamma_t^{(l)} F_t^{(l-1)\top} + \kappa_t^{(l)\top} \quad (7)$$

Explaining the guess about the relationship between features and weights. Figure 5 shows a phenomenon that the feature $F_t^{(l-1)}$ is in the similar direction of the vector $\alpha \Delta V_t^{(l)}$, where $\alpha \in \{-1, +1\}$. By comparing the above two perspectives of such a phenomenon, we guess that the common direction $C^{(l)}$ is similar to $\pm \Gamma_t^{(l)}$, and the feature $F_t^{(l-1)}$ is similar to the vector $\pm \Delta V_t^{(l)}$.

Therefore, in this subsection, we aim to explain that the feature $F_t^{(l-1)}$ and the vector $V_t^{(l)}$ become more and more similar to each other in the first phase. This explains the self-enhancement of the significance of the common direction.

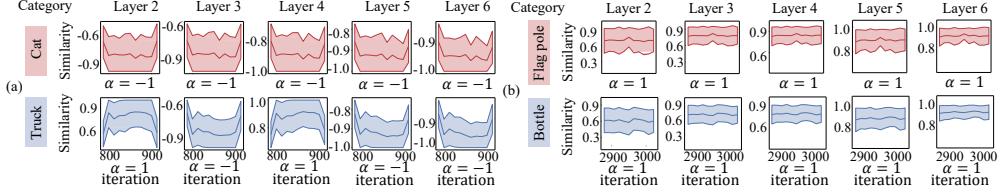


Figure 5: The average cosine similarity between the feature $F_t^{(l-1)}$ and the vector $\Delta V_t^{(l)}$ over different samples in the first phase. We conducted experiments on 9-layer MLPs trained on the (a) CIFAR-10 dataset, and (b) the Tiny ImageNet dataset. The shade in each subfigure represents the standard deviation of the cosine similarity over different samples.

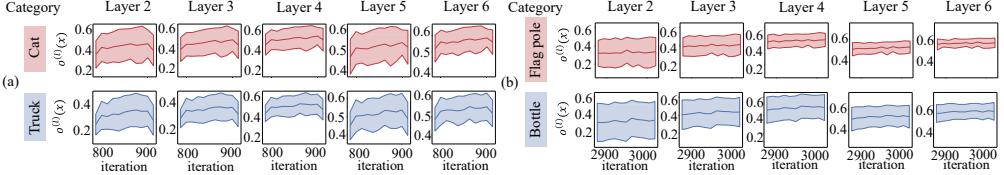


Figure 6: The change of $o^{(l)}$ in the first phase. We trained 9-layer MLPs on the (a) CIFAR-10 and the (b) Tiny ImageNet. Each layer of the MLP had 512 neurons. The supplementary material provides results on the MNIST. The shade represents the standard deviation over different samples.

Discussions on the background assumption. Directly proving how such a “self-enhanced system” emerges from the very beginning of training an initialized MLP is too difficult. Therefore, we hope to explain the significance of the common direction is probably further enhanced under the assumption that features $F_t^{(l-1)}$ of different samples have been pushed a little bit towards a specific common direction. This assumption can derive that there exists at least one learning iteration in the first phase, in which $\Delta F_t^{(l-1)}$ and $F_t^{(l-1)}$ of most samples have similar directions, and $\Delta V_t^{(l)}$ and $V_t^{(l)}$ have similar directions. The derivation is introduced in the supplementary material. The future self-enhancement of $F_t^{(l-1)}$ and $V_t^{(l)}$ is proved under this assumption.

Lemma 3. (Proof in the supplementary material) Given an input sample $x \in X$ and a common direction $C^{(l)}$ after the t -th iteration, if the noise term $\varepsilon_t^{(l)}$ is small enough to satisfy $|\Delta V_t^{(l)\top} F_t^{(l-1)\top} V_t^{(l)\top} C^{(l)\top} \Delta V_t^{(l)\top} C^{(l)\top} F_t^{(l-1)}| \gg |\Delta V_t^{(l)\top} F_t^{(l-1)\top} V_t^{(l)\top} \varepsilon_t^{(l)\top} \Delta \varepsilon_t^{(l)\top} F_t^{(l-1)}|$, we can obtain $\cos(\Delta V_t^{(l)}, F_t^{(l-1)}) \cdot \cos(V_t^{(l)}, \Delta F_t^{(l-1)}) \geq 0$, where $\Delta V_t^{(l)} = \frac{\Delta W_t^{(l)\top} C^{(l)}}{C^{(l)\top} C^{(l)}}$, and $V_t^{(l)} = \frac{W_t^{(l)\top} C^{(l)}}{C^{(l)\top} C^{(l)}}$. $\Delta F_t^{(l-1)}$ denotes the change of features $\Delta F_t^{(l-1)} = F_{t+1}^{(l-1)} - F_t^{(l-1)}$ made by the training sample x after the t -th iteration. To this end, we approximately consider the change of features $\Delta F_t^{(l-1)}$ after the t -th iteration negatively parallel to feature gradients $\dot{F}_t^{(l-1)}$, although strictly speaking, the change of features is not exactly equal to the gradient w.r.t. features.

Theorem 2. (Proof in the supplementary material) Under the background assumption, for any training samples $x, x' \in X_c$ in the category c , if $[C^{(l)\top} D_t^{(l)}|_x \dot{F}_t^{(l)}|_x] \cdot [C^{(l)\top} D_t^{(l)}|_{x'} \dot{F}_t^{(l)}|_{x'}] > 0$ (means that $F_t^{(l)}|_x$ and $F_t^{(l)}|_{x'}$ have kinds of similarity in very early iterations), then $\cos(\alpha_c \Delta V_t^{(l)}|_x, F_t^{(l-1)}|_x) \geq 0$, and $\cos(\alpha_c V_t^{(l)}, \Delta F_t^{(l-1)}|_x) \geq 0$, where $\alpha_c \in \{-1, +1\}$ is a constant shared by all samples in category c .

Explaining the enhancement of the significance of the common direction caused by all training samples in a certain category. Theorem 2 shows that each category has a dominating training direction. Specifically, let us combine Theorem 2 and the background assumption that $\Delta F_t^{(l-1)}$ and $F_t^{(l-1)}$ have similar directions, and $\Delta V_t^{(l)}$ and $V_t^{(l)}$ have similar directions. Thus, we can consider $\cos(\alpha_c V_t^{(l)}, \Delta F_t^{(l-1)}|_x) \geq 0$ in Theorem 2 means that features of training samples in the same category c are all pushed towards a common direction $\alpha_c V_t^{(l)}$, making $\Delta F_t^{(l-1)}|_x$ highly similar to $\alpha_c V_t^{(l)}$. This keeps a high similarity between features within the same category c . On the other hand, $\cos(\alpha_c \Delta V_t^{(l)}|_x, F_t^{(l-1)}|_x) \geq 0$ in Theorem 2 means that training samples in the category c all push $V_t^{(l)}$ towards $\alpha_c \mathbb{E}_{x \in X_c} [F_t^{(l-1)}|_x]$, making $\Delta V_t^{(l)}$ roughly parallel to $\alpha_c \mathbb{E}_{x \in X_c} [F_t^{(l-1)}|_x]$. This phenomenon is verified in Figure 5, where $\cos(\alpha_c \Delta V_t^{(l)}, F_t^{(l-1)})$ is always positive over different samples of the same category. The above analysis also well explains the dynamics behind $\cos(\Delta V_t^{(l)}, F_t^{(l-1)}) \cdot \cos(V_t^{(l)}, \Delta F_t^{(l-1)}) \geq 0$ in Lemma 3.

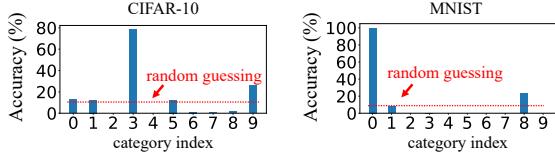


Figure 7: The training accuracy of MLPs on the CIFAR-10 dataset and the MNIST dataset. The accuracy was evaluated at the end of the first phase. The MLP only learned features of a single or two categories in the first phase.

In sum, if we only consider training samples in the sample category c , then features of samples in this category would become increasingly similar to each other. On the other hand, such training samples have similar training effects, i.e., pushing weights of different neurons all towards the average feature.

- Experimental verification of the relationship between the feature $F_t^{(l-1)}$ and the vector $V_t^{(l)}$. To this end, we measured the change of the value $o^{(l)} = \cos(\Delta V_t^{(l)}, F_t^{(l-1)}) \cdot \cos(V_t^{(l)}, \Delta F_t^{(l-1)})$. Figure 6 reports the average $o^{(l)}$ value over different samples at each iteration. For each sample x , $o^{(l)}$ was always positive and usually increased over iterations, which verified Lemma 3. Besides, the assumption for a tiny $\varepsilon_t^{(l)}$ in Lemma 3 was verified by experimental results in the supplementary material.

Assumption 1. We assume that the MLP encodes features of very few (a single or two) categories in the first phase, while other categories have not been learned in this phase.

Figure 7 verifies this assumption. It shows that only a single or two categories exhibit much higher accuracies than the random guessing at the end of the first phase. This indicates that the learning of the MLP is dominated by training samples of a single or two categories in very early iterations.

Enhancement of the significance of the common direction caused by all training samples. The overall learning dynamics in the first phase can be roughly described as follows, by combining Theorem 2 and Assumption 1. The overall learning effects of all training samples are dominated by very few categories \hat{c} . There are two effects. First, features $F_t^{(l-1)}$ of different samples are all pushed towards the vector $\alpha_{\hat{c}} V_t^{(l)}$, where $\alpha_{\hat{c}}$ is determined by the dominating category/categories \hat{c} . Second, $V_t^{(l)}$ is pushed towards $\alpha_{\hat{c}} \mathbb{E}_{x \in X_{\hat{c}}} [F_t^{(l-1)}|_x]$. Therefore, features $F_t^{(l-1)}$ of different samples and $\alpha_{\hat{c}} V_t^{(l)}$ enhance each other, just like a self-enhanced system, in the scenario that $\Delta F_t^{(l-1)}$ and $F_t^{(l-1)}$ of most samples have similar directions, and $\Delta V_t^{(l)}$ and $V_t^{(l)}$ have similar directions (see the background assumption). In other words, the component along the common direction $C^{(l)} \Delta V_t^{(l)\top}$ in $\Delta W_t^{(l)} = C^{(l)} \Delta V_t^{(l)\top} + \Delta \varepsilon_t^{(l)\top}$ will be further enhanced.

Explaining the increasing feature similarity and the increasing gradient similarity. As aforementioned, features $F_t^{(l-1)}$ of different samples are consistently pushed towards the same vector $\alpha_{\hat{c}} V_t^{(l)}$. Therefore, the similarity between features of different samples $\mathbb{E}_{x, x' \in X} [\cos(F_t^{(l-1)}|_x, F_t^{(l-1)}|_{x'})]$ increases in the first phase. On the other hand, the increasing similarity between feature gradients can be explained from two views. (1) The increasing feature similarity over different samples makes different training samples generate similar gating states $D_t^{(l)}$ in each ReLU layer. The increasing similarity between gating states of each ReLU layer over different samples leads to the increasing similarity between feature gradients over different samples $\mathbb{E}_{x, x' \in X_c} [\cos(\dot{F}_t^{(l-1)}|_x, \dot{F}_t^{(l-1)}|_{x'})]$ of the same category. (2) Another view is that the component along the common direction $C^{(l)} V_t^{(l)\top}$ in $W_t^{(l)}$ is enhanced in the first phase. Because $C^{(l)}$ denotes the principle weight direction of each the i -th column $w_{t,i}^{(l)}$ of $W_t^{(l)}$, i.e., each “pseudo-neuron” $w_{t,i}^{(l)}$ is optimized towards the common direction $C^{(l)}$. Eq. (3) has proved that the increasing cosine similarity between $w_{t,i}^{(l)}$ and $C^{(l)}$ for all “pseudo-neurons” will lead to the increasing similarity between feature gradients over different samples.

How to escape from the first phase? In the first phase, the MLP only discovers a single direction to optimize a single or two categories. However, the optimization of a single or two categories will soon saturate, and the gradient mainly comes from training samples of other categories, which disturbs the dominating roles of a single or two categories in the learning of the MLP. Therefore, the learning effects of training samples from different categories may conflict with each other. Thus, the self-enhanced system is destroyed, and the learning of the MLP enters the second phase.

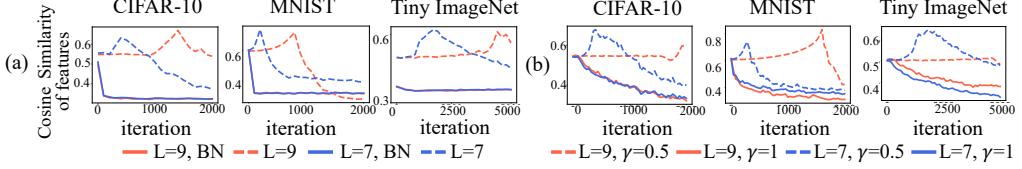


Figure 8: Effects of (a) normalization and (b) initialization. We trained L -layer MLPs, where each layer had 512 neurons. A shorter first phase indicates that the decrease of feature diversity is more alleviated. Effects of normalization and L_2 regularization are shown in the supplementary material.

4.3 Theoretically alleviating the decrease of feature diversity

In previous sections, we have discovered and explained a fundamental yet counter-intuitive two-phase phenomenon with the MLP. This is the distinctive contribution of this study, which has not been theoretically explained for a long time. Besides, we find that we can use the above findings to explain that four typical operations can usually alleviate the decrease of feature diversity, *i.e.*, normalization, momentum, initialization, and L_2 regularization. Although these operations have been widely used, previous studies failed to theoretically explain their effectiveness. To this end, our analysis can explain a high likelihood for such operations to alleviate the decrease of feature diversity, but it is not a proof of a strict sufficient condition or a necessary condition for the feature diversity.

Normalization. Based on theoretical analysis, we explain that normalization operations (*e.g.* batch normalization (BN)) can alleviate the decrease of feature diversity in the first phase. Specifically, according to Theorem 2, the self-enhanced system of decreasing feature diversity requires features $F_t^{(l)}$ of any two training samples x and x' in the same category to be similar to each other. However, the BN layer prevents features $F_t^{(l)}$ of different samples from being similar to each other, because the mean feature $\bar{F}_t^{(l)} = \mathbb{E}_{x \in X}[F_t^{(l)}|_x]$ is subtracted from features of all samples, *i.e.*, $F_t^{(l)}|_x = F_t^{(l)}|_x - \bar{F}_t^{(l)}$. Therefore, features $F_t^{(l)}$ of different samples are no longer similar to each other, thereby being more likely to break the self-enhanced system. Please see the supplementary material for more discussions.

We conducted experiments to verify the above analysis. We compared MLPs trained with and without BN layers. Specifically, we added a BN layer after each linear layer to construct MLPs. Figure 8(a) shows that the feature similarity in MLPs with BN layers kept decreasing. This verified that BN layers alleviated the decrease of feature diversity. Besides, based on our analysis, we further simplified the BN layer, which could also alleviate the decrease of feature diversity. Please see the supplementary material for more details.

Momentum. Based on the theoretical explanation for the decrease of feature diversity, we can explain that momentum in gradient descent can alleviate this phenomenon. Based on Lemma 3, the self-enhanced system of the decreasing of feature diversity requires weights along other directions $\varepsilon_t^{(l)}$ to be small enough. However, because the momentum operation strengthens influences of the initialized noisy weights $W_{t=0}^{(l)}$, it strengthens singular values of $\varepsilon_t^{(l)}$, to some extent, thereby alleviating the decrease of feature diversity. Specifically, a larger coefficient of momentum m usually more alleviate the decrease of feature diversity.

Please see the supplementary material for the detailed discussion. To verify the above analysis, we trained MLPs with $m = 0, 0.5, 0.9$, respectively. The verification of our claim that a larger value of m usually more alleviates the decrease of feature diversity is introduced in the supplementary material.

Initialization. We explain that the initialization of MLPs also affects the decrease of feature diversity. According to Lemma 3, such a self-enhanced system requires weights along other directions $\varepsilon_t^{(l)}$ to be small enough. However, because increasing the variance of the initialized weights $W_{t=0}^{(l)}$ will increase singular values of $\varepsilon_t^{(l)}$, alleviating the decrease of feature diversity. Please see the supplementary material for the detailed discussion.

We conducted experiments to verify the above claim by comparing MLPs trained using different initializations with different variances. We used γ to control the variance of the initialization, *i.e.*, $W_{t=0}^{(l)} \sim \mathcal{N}(0, \gamma\sigma_{var}^2)$, where σ_{var} is a constant. Figure 8(b) verifies that the initialization with a large variance alleviated the decrease of feature diversity.

L_2 regularization. We also explain that a small coefficient of the L_2 regularization can alleviate the decrease of feature diversity. *I.e.*, The total loss $\mathcal{L}(W_t) = \mathcal{L}^{CE}(W_t) + \lambda\|W_t\|_2^2$, where $\mathcal{L}^{CE}(W_t)$

represents the cross entropy loss and $\lambda\|W_t\|_2^2$ denotes the L_2 regularization loss. As aforementioned, the decrease of feature diversity requires singular values of $\varepsilon_t^{(l)}$ to be small enough. However, because a smaller coefficient λ of the L_2 regularization yields larger singular values of $\varepsilon_t^{(l)}$, it alleviates the decrease of feature diversity. Please see the supplementary material for the detailed discussion. The verification that a smaller coefficient λ more alleviated the decrease of feature diversity is introduced in the supplementary material.

5 Conclusion

In this paper, we find that in the early stage of the training process, the MLP exhibits a fundamental yet counter-intuitive two-phase phenomenon, *i.e.*, the feature diversity keeps decreasing in the first phase. We explain this phenomenon by analyzing the learning dynamics of the MLP. Furthermore, we explain the reason why four typical operations can alleviate the decrease of feature diversity.

References

- [1] Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018.
- [2] Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR, 2020.
- [3] Madhu S Advani and Andrew M Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.
- [4] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pages 244–253. PMLR, 2018.
- [5] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242. PMLR, 2017.
- [6] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.
- [7] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.
- [8] Hao Cheng, Dongze Lian, Shenghua Gao, and Yanlin Geng. Evaluating capability of deep neural networks for image classification via information plane. In *European Conference on Computer Vision*, pages 181–195. Springer, 2018.
- [9] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204. PMLR, 2015.
- [10] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances In Neural Information Processing Systems*, 29:2253–2261, 2016.
- [11] Stéphane d’Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR, 2020.
- [12] Stéphane d’Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: where & why do they appear? In *NeurIPS*, 2020.
- [13] Gianna M Del Corso, Antonio Gulli, and Francesco Romani. Ranking a stream of news. In *Proceedings of the 14th international conference on World Wide Web*, pages 97–106, 2005.

- [14] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.
- [15] Stanislav Fort, Paweł Krzysztof Nowak, Stanislaw Jastrzebski, and Srinivas Narayanan. Stiffness: A new perspective on generalization in neural networks. *arXiv preprint arXiv:1901.09491*, 2019.
- [16] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2018.
- [17] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.
- [18] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.
- [19] Reinhard Heckel and Fatih Furkan Yilmaz. Early stopping in deep networks: Double descent and how to eliminate it. In *International Conference on Learning Representations*, 2020.
- [20] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1729–1739, 2017.
- [21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.
- [22] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.
- [23] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pages 4631–4640. PMLR, 2020.
- [24] Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.
- [25] Kenji Kawaguchi. Deep learning without poor local minima. *Advances in Neural Information Processing Systems*, 29:586–594, 2016.
- [26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [27] Andrew K Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. In *International Conference on Learning Representations*, 2018.
- [28] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7:7, 2015.
- [29] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [30] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.
- [31] Xingguo Li, Junwei Lu, Zhaoran Wang, Jarvis Haupt, and Tuo Zhao. On tighter generalization bound for deep neural networks: Cnns, resnets, and beyond. *arXiv preprint arXiv:1806.05159*, 2018.
- [32] Philip M Long and Hanie Sedghi. Generalization bounds for deep convolutional neural networks. In *International Conference on Learning Representations*, 2019.

- [33] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, volume 30, page 3. Citeseer, 2013.
- [34] Michael Mahoney and Charles Martin. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pages 4284–4293. PMLR, 2019.
- [35] Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.
- [36] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2019.
- [37] Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018.
- [38] Jeffrey Pennington, Samuel S Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4788–4798, 2017.
- [39] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pages 4433–4441. PMLR, 2018.
- [40] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.
- [41] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.
- [42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [43] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.
- [44] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.
- [45] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.
- [46] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.
- [47] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. In *International Conference on Learning Representations*, 2018.
- [48] Wolchover. New theory cracks open the black box of deep learning. *Quanta Magazine*, 3, 2017.
- [49] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 2017:2525–2534, 2017.
- [50] Tengyu Xu, Yi Zhou, Kaiyi Ji, and Yingbin Liang. Convergence of sgd in learning relu models with separable data. *arXiv preprint arXiv:1806.04339*, 2018.

- [51] Zhiqin John Xu. Understanding training and generalization in deep learning by fourier analysis. *arXiv preprint arXiv:1808.04295*, 2018.
- [52] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pages 10767–10777. PMLR, 2020.

A Common phenomenon shared by different DNNs for different tasks.

In this section, we aim to demonstrate an interesting two-phase phenomenon when we train an MLP in early iterations. Specifically, the training process of the MLP can usually be divided into the following two phases according to the training loss. In the first phase, the training loss does not decrease significantly, and the training loss suddenly begins to decrease in the second phase. More crucially, the feature diversity decreases in the first phase. This phenomenon is widely shared by different DNNs with different architectures for different tasks.

Let us take the 9-layer MLP trained on the CIFAR-10 dataset for an example, where each layer of the MLP had 512 neurons. As Figure 1(e)(f) shows, before the 1300-th iteration (the first phase), both the feature diversity and the gradient diversity kept decreasing, *i.e.*, both the cosine similarity between features over different samples and the cosine similarity between gradients kept increasing. After the 1300-th iteration (the second phase), the feature diversity and the gradient diversity suddenly began to increase, *i.e.* their similarities began to decrease. Therefore, the MLP had the lowest feature diversity and the lowest gradient diversity at around the 1300-th iteration. Specifically, the training loss was evaluated on the whole training set.

A.1 On the CIFAR-10 dataset

In this subsection, we demonstrated that the two-phase phenomenon was shared by different MLPs on the CIFAR-10 dataset [26]. For different MLPs, we adopted the learning rate $\eta = 0.1$, the batch size $bs = 100$, the SGD optimizer, and the ReLU activation function. Besides, we used two data augmentation methods, including random cropping and random horizontal flipping. The training loss, the testing loss, the training accuracy, the testing accuracy, the cosine similarity of features, and the cosine similarity of feature gradients of MLPs trained on the CIFAR-10 dataset are shown in Figure 1.

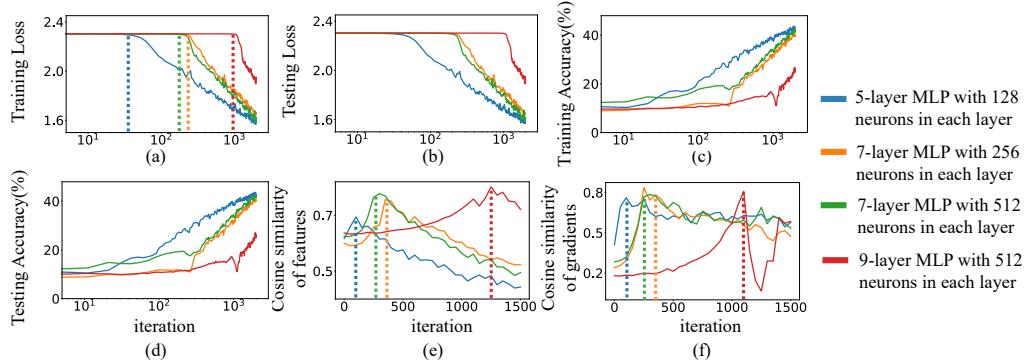


Figure 1: (a) The training loss of four MLPs trained on the CIFAR-10 dataset. (b) The testing loss of four MLPs. (c) Training accuracies of four MLPs. (d) Testing accuracies of four MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the third linear layer of MLPs.

A.2 On the MNIST dataset

In this subsection, we demonstrated that the two-phase phenomenon was shared by different MLPs on the MNIST dataset [29]. For different MLPs, we adopted the learning rate $\eta = 0.01$, the batch size $bs = 100$, the SGD optimizer, and the ReLU activation function. The training loss, the testing loss, the training accuracy, the testing accuracy, the cosine similarity of features, and the cosine similarity of feature gradients of MLPs trained on the MNIST are shown in Figure 2.

A.3 On the Tiny ImageNet dataset

In this subsection, we demonstrated that the two-phase phenomenon was shared by different MLPs on the Tiny ImageNet dataset [28]. Specifically, we randomly selected the following 50 categories,

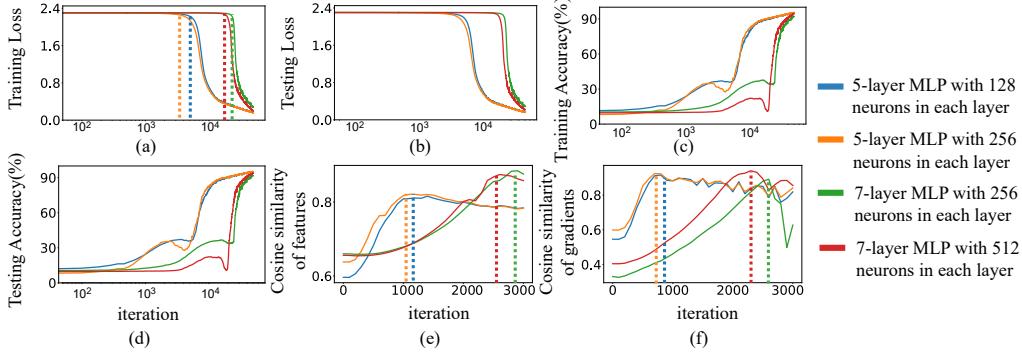


Figure 2: (a) The training loss of four MLPs trained on the MNIST dataset. (b) The testing loss of four MLPs. (c) Training accuracies of four MLPs. (d) Testing accuracies of four MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the third linear layer of MLPs.

orangutan, parking meter, snorkel, American alligator, oboe, basketball, rocking chair, hopper, neck brace, candy store, broom, seashore, sewing machine, sunglasses, panda, pretzel, pig, volleyball, puma, alp, barbershop, ox, flagpole, lifeboat, teapot, walking stick, brain coral, slug, abacus, comic book, CD player, school bus, banister, bathtub, German shepherd, black stork, computer keyboard, tarantula, sock, Arabian camel, bee, cockroach, cannon, tractor, cardigan, suspension bridge, beer bottle, viaduct, guacamole, and iPod for training. For different MLPs, we adopted the learning rate $\eta = 0.1$, the batch size $bs = 100$, the SGD optimizer, and the ReLU activation function. Besides, we used two data augmentation methods, including random cropping and random horizontal flipping. Note that we took a random cropping with 32×32 sizes. The training loss, the testing loss, the training accuracy, the testing accuracy, the cosine similarity of features, and the cosine similarity of feature gradients of MLPs trained on the Tiny ImageNet are shown in Figure 3.

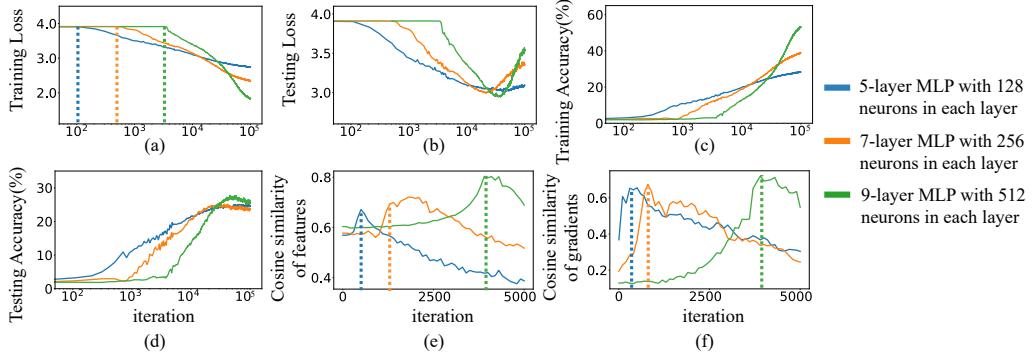


Figure 3: (a) The training loss of three MLPs trained on the Tiny ImageNet dataset. (b) The testing loss of three MLPs. (c) Training accuracies of three MLPs. (d) Testing accuracies of three MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The features and the feature gradient were used in the second linear layer of MLPs.

A.4 On the Census dataset

In this subsection, we demonstrated that the two-phase phenomenon was shared by different MLPs on the UCI census income tabular dataset (Census) [6]. For different MLPs, we adopted the learning rate $\eta = 0.1$, the batch size $bs = 1000$, the SGD optimizer, and the ReLU activation function. The training loss, the testing loss, the training accuracy, the testing accuracy, the cosine similarity of features, and the cosine similarity of feature gradients of MLPs trained on the census are shown in Figure 4.

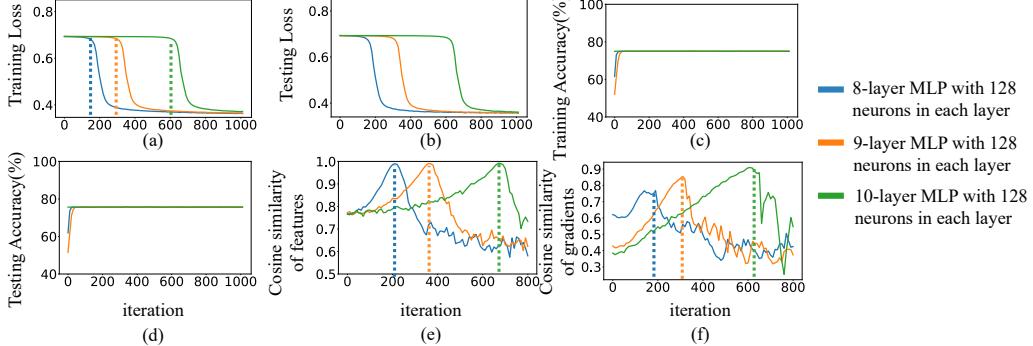


Figure 4: (a) The training loss of three MLPs trained on the Census dataset. (b) The testing loss of three MLPs. (c) Training accuracies of three MLPs. (d) Testing accuracies of three MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the fifth linear layer of MLPs.

A.5 On the Commercial dataset

In this subsection, we demonstrated that the two-phase phenomenon was shared by different MLPs on the UCI TV news channel commercial detection dataset (Commercial) [6]. For different MLPs, we adopted the learning rate $\eta = 0.1$, the batch size $bs = 1000$, the SGD optimizer, and the ReLU activation function. The training loss, the testing loss, the training accuracy, the testing accuracy, the cosine similarity of features, and the cosine similarity of feature gradients of MLPs trained on the census are shown in Figure 5.

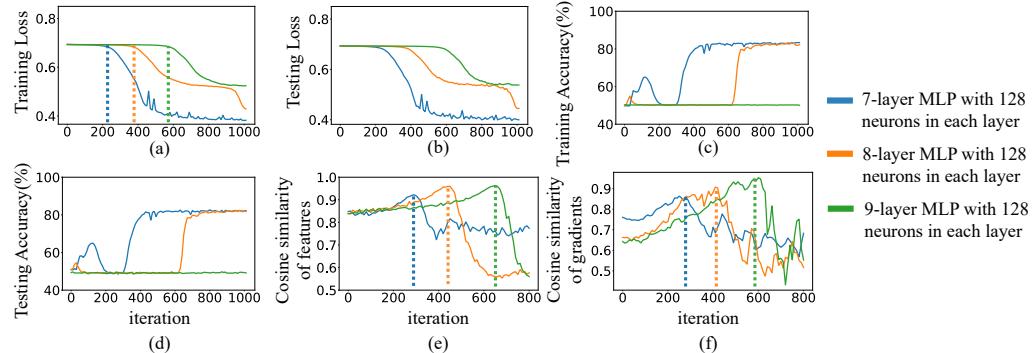


Figure 5: (a) The training loss of three MLPs trained on the Commercial dataset. (b) The testing loss of three MLPs. (c) Training accuracies of three MLPs. (d) Testing accuracies of three MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the fifth linear layer of MLPs.

A.6 On the CoLA dataset

In this subsection, we demonstrated that the two-phase phenomenon was shared by the revised LSTMs on the CoLA dataset [46]. We used two-layer unidirectional LSTMs concatenated with MLPs. Specifically, we trained two LSTMs with 5-layer MLPs, where each layer of the MLP has 256 and 512 neurons. We adopted the learning rate $\eta = 0.1$, the batch size $bs = 1000$, the SGD optimizer, and the ReLU activation function. The training loss, the testing loss, the training accuracy, the testing accuracy, the cosine similarity of features, and the cosine similarity of feature gradients of LSTMs trained on the CoLA are shown in Figure 6. Since training samples in the CoLA dataset were imbalanced, we constructed a new training set by randomly sampling 2000 training samples from two categories, respectively. DNNs were trained on this new training set.

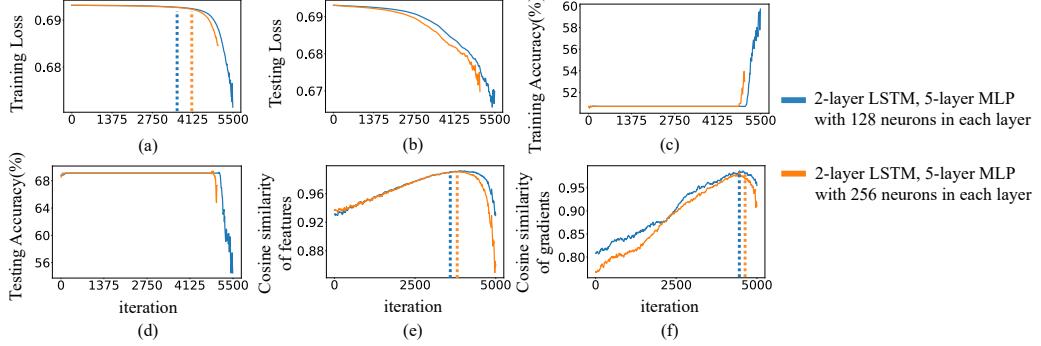


Figure 6: (a) The training loss of two LSTMs trained on the CoLA dataset. (b) The testing loss of two LSTMs. (c) Training accuracies of two LSTMs. (d) Testing accuracies of two LSTMs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the third linear layer of MLPs.

A.7 On the SST-2 dataset

In this subsection, we demonstrated that the two-phase phenomenon was shared by the revised LSTMs on the SST-2 dataset [44]. We used unidirectional LSTMs concatenated with MLPs. Specifically, we trained three LSTMs with 4-layer MLPs, 4-layer MLPs, and 5-layer MLPs, respectively, where each layer of the MLP has 32, 64, 128 neurons. We adopted the learning rate $\eta = 0.1$, the batch size $bs = 500$, the SGD optimizer, and the ReLU activation function. Since the training of LSTMs on the SST-2 with the SGD optimizer is unstable, we randomly selected 15000 training samples from the training set. We trained LSTMs on these 15000 training samples. The training loss, the testing loss, the training accuracy, the testing accuracy, the cosine similarity of features, and the cosine similarity of feature gradients of LSTMs trained on the SST-2 are shown in Figure 7.

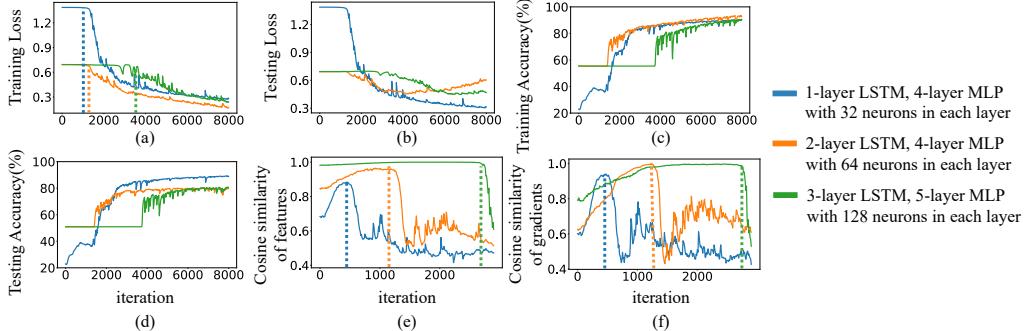


Figure 7: (a) The training loss of three LSTMs trained on the SST-2 dataset. (b) The testing loss of three LSTMs. (c) Training accuracies of three LSTMs. (d) Testing accuracies of three LSTMs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the second linear layer of MLPs.

A.8 On the AGNews dataset

In this subsection, we demonstrated that the two-phase phenomenon was shared by the revised LSTMs on the AGNEWS dataset. We used two-layer unidirectional LSTMs concatenated with MLPs. Specifically, we trained three LSTMs with 4-layer MLPs, 4-layer MLPs, 5-layer MLPs, respectively, where each layer of the MLP had 32, 64, and 128 neurons, respectively. We adopted the learning rate $\eta = 0.1$, the batch size $bs = 500$, the SGD optimizer, and the ReLU activation function. The training loss, the testing loss, the training accuracy, the testing accuracy, the cosine similarity of features, and the cosine similarity of feature gradients of LSTMs trained on the AGNEWS are shown in Figure 8.

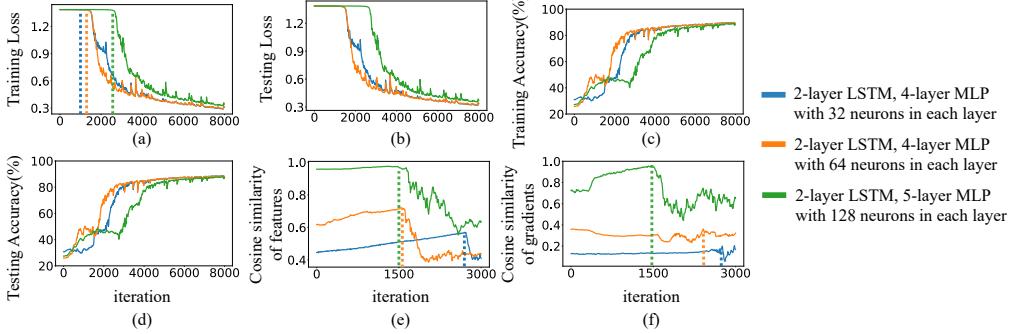


Figure 8: (a) The training loss of three LSTMs trained on the AGNEWS dataset. (b) The testing loss of three LSTMs. (c) Training accuracies of three LSTMs. (d) Testing accuracies of three LSTMs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the second linear layer of MLPs.

A.9 Different training batch sizes

In this subsection, we demonstrated that the two-phase phenomenon was shared by MLPs trained on the CIFAR-10 dataset with different training batch sizes. For different MLPs, we adopted the learning rate $\eta = 0.1$, the SGD optimizer, and the ReLU activation function. Besides, we used two data augmentation methods, including random cropping and random horizontal flipping. We trained three 7-layer MLPs with 256 neurons in each layer, with $bs = 100, 500, 1000$ respectively. The training loss, the testing loss, the training accuracy, the testing accuracy, the cosine similarity of features, and the cosine similarity of feature gradients of MLPs trained with different batch sizes are shown in Figure 9.

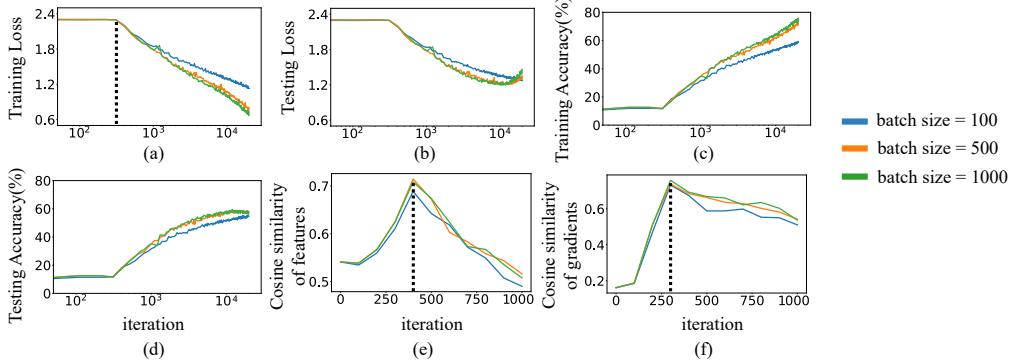


Figure 9: (a) The training loss of three MLPs trained with different batch sizes. (b) The testing loss of three MLPs. (c) Training accuracies of three MLPs. (d) Testing accuracies of three MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the second linear layer of MLPs.

A.10 Different learning rates

In this subsection, we demonstrated that the two-phase phenomenon was shared by MLPs trained on the CIFAR-10 dataset with different learning rates. For different MLPs, we adopted the batch size $bs = 100$, the SGD optimizer, and the ReLU activation function. Besides, we used two data augmentation methods, including random cropping and random horizontal flipping. We trained two 7-layer MLPs with 256 neurons in each layer, with learning rates $\eta = 0.1, 0.01$ respectively. The training loss, the testing loss, the training accuracy, the testing accuracy, the cosine similarity of features, and the cosine similarity of feature gradients of MLPs trained with different learning rates are shown in Figure 10.

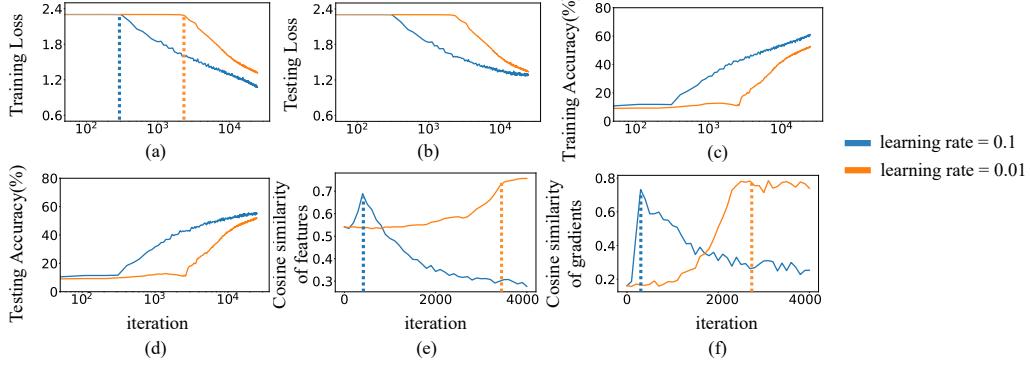


Figure 10: (a) The training loss of two MLPs trained with different learning rates. (b) The testing loss of two MLPs. (c) The training accuracies of two MLPs. (d) The testing accuracies of two MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the second linear layer of MLPs.

A.11 Different activation functions

In this subsection, we demonstrated that the two-phase phenomenon was shared by MLPs with different activation functions. For different MLPs, we adopted the learning rate $\eta = 0.1$, the batch size $bs = 100$, the SGD optimizer. Besides, we used two data augmentation methods, including random cropping and random horizontal flipping. We trained three 9-layer MLPs with 512 neurons in each layer with the ReLU activation function, the Leaky ReLU (slope=0.1) activation function, and the Leaky ReLU (slope=0.01) activation function, respectively. The training loss, the testing loss, the training accuracy, the testing accuracy, the cosine similarity of features, and the cosine similarity of feature gradients of MLPs trained with different activation functions are shown in Figure 11.

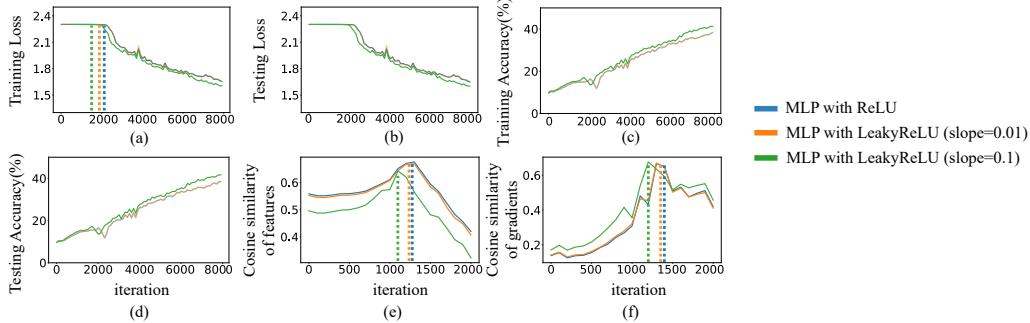
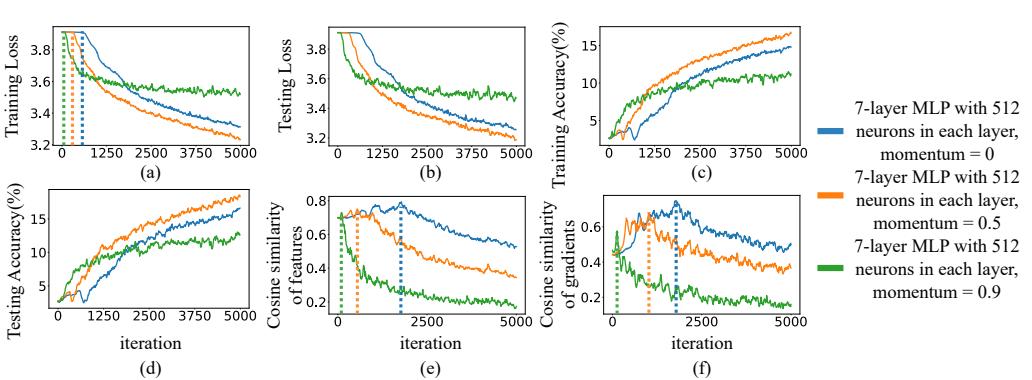
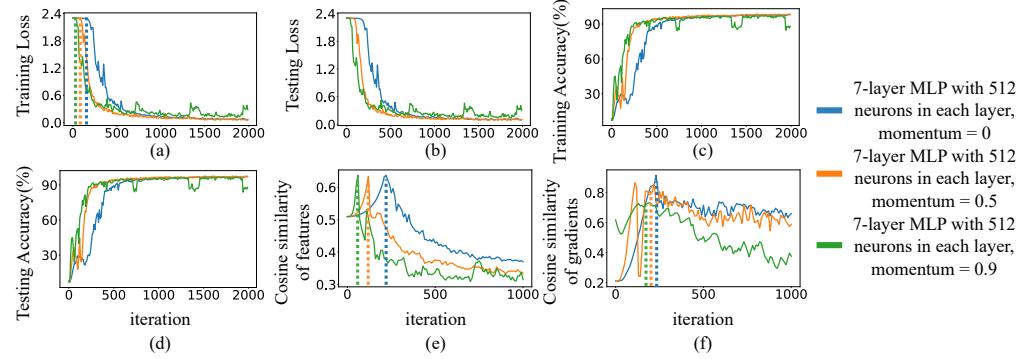
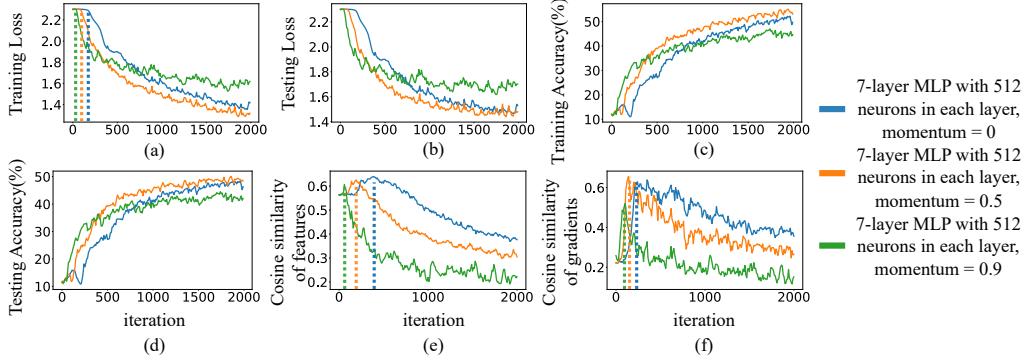


Figure 11: (a) The training loss of three MLPs with different activation functions. (b) The testing loss of three MLPs. (c) Training accuracies of three MLPs. (d) Testing accuracies of three MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the second linear layer of MLPs.

A.12 Different momentums

In this subsection, we demonstrated that the two-phase phenomenon was shared by MLPs trained on the CIFAR-10, MNIST and Tiny Imagenet dataset with different momentums. For different MLPs, we adopted the learning rate $\eta = 0.1$, the batch size $bs = 100$, the SGD optimizer. Besides, we used two data augmentation methods, including random cropping and random horizontal flipping. We trained 7-layer MLPs and 9-layer MLPs with 512 neurons in each layer with the ReLU activation function. The training loss, the testing loss, the training accuracy, the testing accuracy, the cosine similarity of features, and the cosine similarity of feature gradients of MLPs trained with different momentum are shown in Figure 12, Figure 13, Figure 14, Figure 15, Figure 16, and Figure 17.



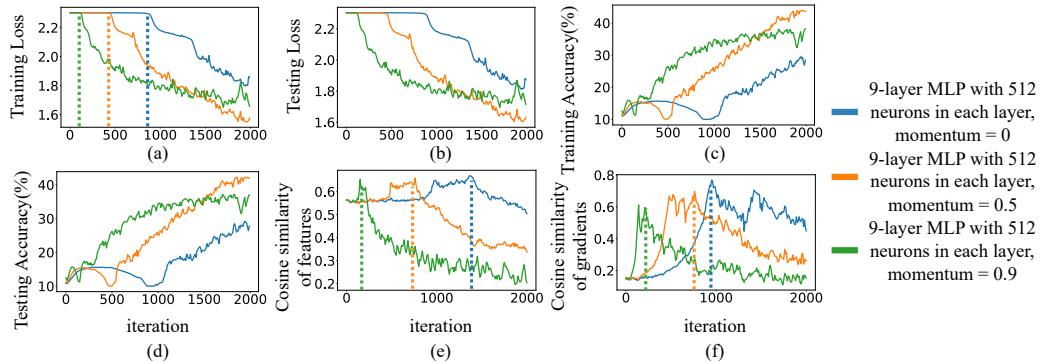


Figure 15: (a) The training loss of three MLPs with different momentums trained on the CIFAR-10 dataset. (b) The testing loss of three MLPs. (c) Training accuracies of three MLPs. (d) Testing accuracies of three MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the second linear layer of MLPs.

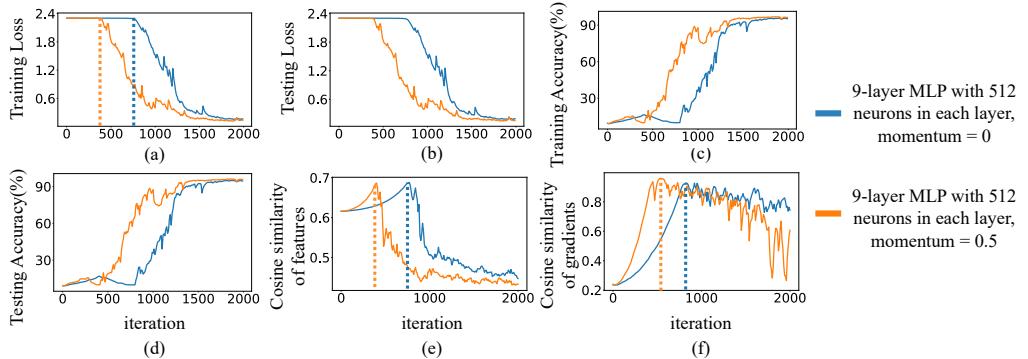


Figure 16: (a) The training loss of two MLPs with different momentums trained on the MNIST dataset. (b) The testing loss of two MLPs. (c) Training accuracies of two MLPs. (d) Testing accuracies of two MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the second linear layer of MLPs.

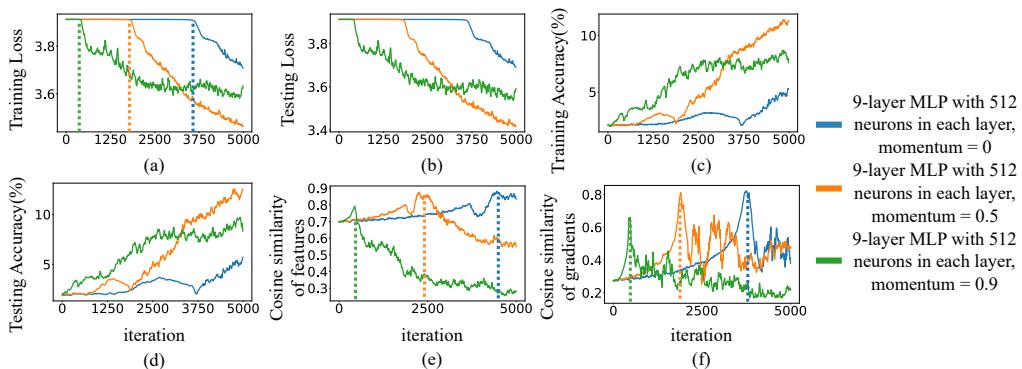


Figure 17: (a) The training loss of three MLPs with different momentums trained on the Tiny ImageNet dataset. (b) The testing loss of three MLPs. (c) Training accuracies of three MLPs. (d) Testing accuracies of three MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the fourth linear layer of MLPs.

A.13 Different weight decays

In this subsection, we demonstrated that the two-phase phenomenon was shared by MLPs trained on the CIFAR-10, MNIST and Tiny Imagenet dataset with different weight decays. For different MLPs, we adopted the learning rate $\eta = 0.1$, the batch size $bs = 100$, the SGD optimizer. Besides, we used two data augmentation methods, including random cropping and random horizontal flipping. We trained 7-layer MLPs and 9-layer MLPs with 512 neurons in each layer with the ReLU activation function. The training loss, the testing loss, the training accuracy, the testing accuracy, the cosine similarity of features, and the cosine similarity of feature gradients of MLPs trained with different weight decays are shown in Figure 18, Figure 19, Figure 20, Figure 21, Figure 22, and Figure 23.

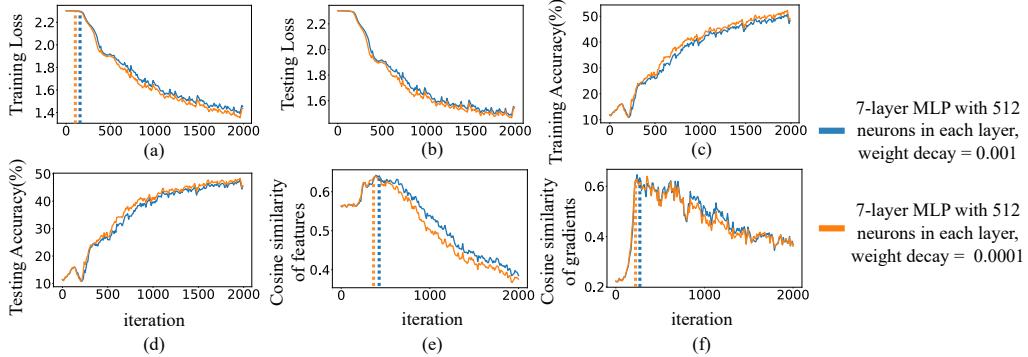


Figure 18: (a) The training loss of two MLPs with different weight decays trained on the CIFAR-10 dataset. (b) The testing loss of two MLPs. (c) Training accuracies of two MLPs. (d) Testing accuracies of two MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the second linear layer of MLPs.

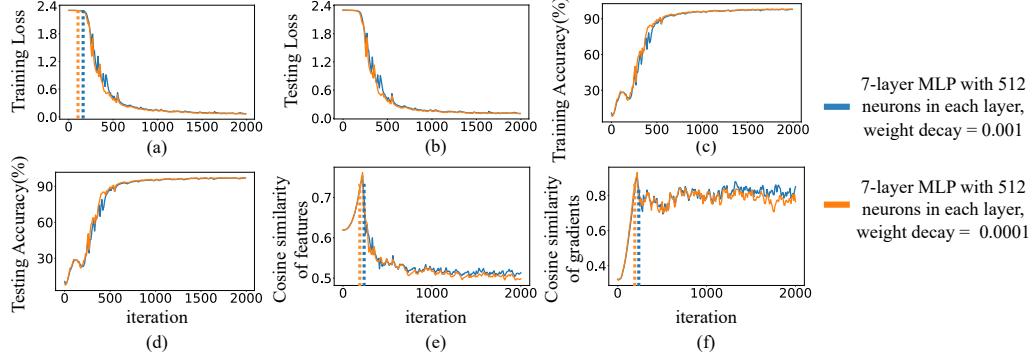


Figure 19: (a) The training loss of two MLPs with different weight decays trained on the MNIST dataset. (b) The testing loss of two MLPs. (c) Training accuracies of two MLPs. (d) Testing accuracies of two MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the second linear layer of MLPs.

B Discussion of the learning-sticking problem

In this section, we aim to discuss the learning-sticking problem in the learning of MLPs. In fact, this problem appears in various DNNs, including MLPs, CNNs, and RNNs, when the task is difficult enough. In this paper, we just take the MLP as an example for discussion without loss of generality. Explaining and solving the occasional sticking of the training of MLPs are of significant values on different tasks. Specifically, previous studies simply owed the learning-sticking problem to the difficulty of the training task and solved this problem by heuristically applying some optimization tricks. They have no insightful analysis or theoretically supported solutions. In comparison, our study explains the learning-sticking problem as the first phase with an infinite length. Moreover, we theoretically explain mechanisms of several heuristic solutions to the learning-sticking problem.

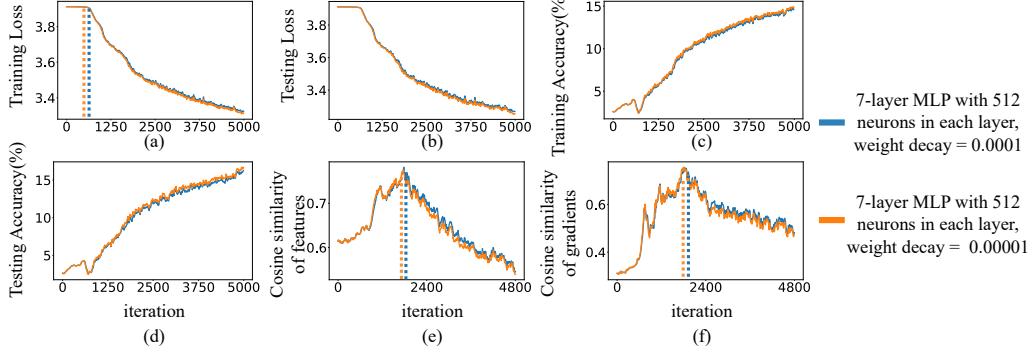


Figure 20: (a) The training loss of two MLPs with different weight decays trained on the Tiny ImageNet dataset. (b) The testing loss of two MLPs. (c) Training accuracies of two MLPs. (d) Testing accuracies of two MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the third linear layer of MLPs.

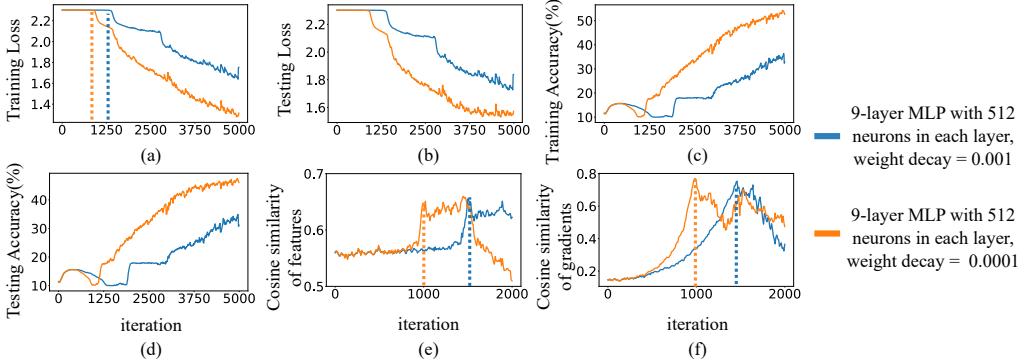


Figure 21: (a) The training loss of two MLPs with different weight decays trained on the CIFAR-10 dataset. (b) The testing loss of two MLPs. (c) Training accuracies of two MLPs. (d) Testing accuracies of two MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the second linear layer of MLPs.

To this end, the learning-sticking problem can be solved based on our study, as shown in Figure 24. Specifically, we trained a 9-layer MLP on the CIFAR-10 dataset, where each layer of the MLP had 512 neurons and its initial weights were sample from $\mathcal{N}(0, \gamma_1 \sigma_{var}^2)$ ($\gamma_1 = 0.1$). We observed that the training of the MLP got stuck (orange curve). According to our study, the technique of increasing the variance of initial weights can shorten the first phase, thereby solving the learning-sticking problem. To this end, we trained another 9-layer MLP, and the only difference from the previous MLP is that the variance of initial weights was increased to $\gamma_2 \sigma_{var}^2$ ($\gamma_2 = 1$). Figure 24(a) shows that the first phase is shortened by increasing the variance of initial weights (another MLP in the blue curve), thereby the learning-sticking problem is solved.

Actually, far beyond solving the learning-sticking problem, the two-phase phenomenon of MLPs is generally considered a counter-intuitive phenomenon. In this paper, our distinctive contribution is to explain the counter-intuitive two-phase phenomenon of MLPs theoretically.

C Double descent

There are usually two types of double-descent phenomena. The model-wise double descent behavior has emerged in many deep learning tasks, which means that as the model size increases, performance first decreases, then increases, and finally decreases [3, 23, 52, 11]. Furthermore, some recent studies discussed the existence of the triple descent curve [12, 2]. Besides, the double descent behavior also occurs with respect to training epochs [36, 19], called epoch-wise double descent, *i.e.* as the epoch increases, the testing error first decreases, then increases, and finally decreases. As Figure 1(d) in

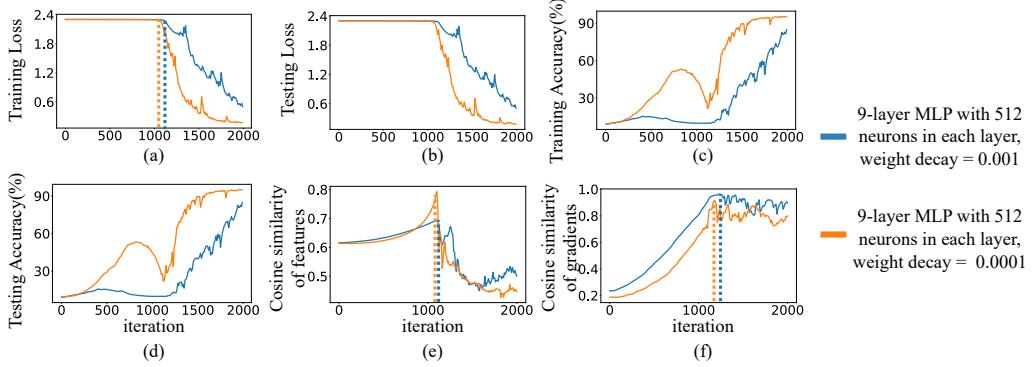


Figure 22: (a) The training loss of two MLPs with different weight decays trained on the MNIST dataset. (b) The testing loss of two MLPs. (c) Training accuracies of two MLPs. (d) Testing accuracies of two MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the second linear layer of MLPs.

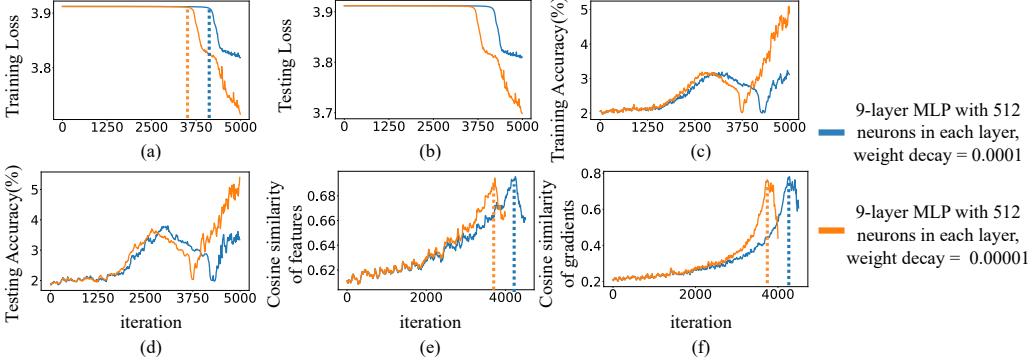


Figure 23: (a) The training loss of two MLPs with different weight decays trained on the Tiny ImageNet dataset. (b) The testing loss of two MLPs. (c) Training accuracies of two MLPs. (d) Testing accuracies of two MLPs. (e) Cosine similarity between features of different categories. (f) Cosine similarity between gradients of different samples in a category. The feature and the feature gradient were used in the third linear layer of MLPs.

the main paper shows, the first and the second stages in the epoch-wise double descent behavior are temporally aligned with the first phase in the aforementioned two-phase phenomenon, where the training loss does not change significantly.

D More results on the MNIST dataset

In this section, we provide more results on the MNIST dataset. Fig 25 and Tables 1 empirically verify the strength of the primary common direction, which are supplementary to Figure 4 and Table 1 in the main paper, respectively. Fig 26 illustrates the change of $o^{(l)} = \cos(\Delta V_t^{(l)}, F_t^{(l-1)}) \cdot \cos(V_t^{(l)}, \Delta F_t^{(l-1)})$ in the first phase, which is supplementary to Figure 6 in the main paper.

Table 1: Strength of components of weight changes along the primary common direction and other directions. We trained a 9-layer MLP on the MNIST dataset. Each layer of the MLP had 512 neurons. It can be observed that the strength of the primary common direction was much greater than those of other directions.

| Category | Eight | | | | | Zero | | | | |
|----------------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| | Layer 2 | Layer 3 | Layer 4 | Layer 5 | Layer 6 | Layer 2 | Layer 3 | Layer 4 | Layer 5 | Layer 6 |
| $S_{\text{primary}}^{(l)}$ | 367.1 \pm 56.8 | 364.5 \pm 52.8 | 381.9 \pm 56.3 | 444.4 \pm 68.7 | 504.0 \pm 81.3 | 441.7 \pm 86.0 | 448.2 \pm 83.5 | 429.0 \pm 78.1 | 493.1 \pm 87.2 | 504.1 \pm 89.0 |
| $S_1^{(l)}$ | 14.9 \pm 0.8 | 15.9 \pm 1.4 | 15.5 \pm 1.1 | 15.6 \pm 1.5 | 13.5 \pm 2.0 | 24.6 \pm 3.1 | 30.0 \pm 4.3 | 18.4 \pm 2.6 | 17.2 \pm 2.2 | 15.6 \pm 1.8 |
| $S_2^{(l)}$ | 16.3 \pm 1.7 | 13.1 \pm 0.9 | 16.4 \pm 0.8 | 18.1 \pm 3.2 | 11.7 \pm 1.6 | 16.6 \pm 1.7 | 23.9 \pm 4.2 | 17.9 \pm 2.4 | 14.3 \pm 1.5 | 12.2 \pm 1.9 |
| $S_3^{(l)}$ | 15.1 \pm 1.5 | 16.3 \pm 1.7 | 13.5 \pm 0.6 | 15.1 \pm 1.4 | 15.0 \pm 1.1 | 29.4 \pm 5.2 | 21.1 \pm 4.2 | 15.5 \pm 1.8 | 21.2 \pm 3.6 | 14.7 \pm 1.6 |

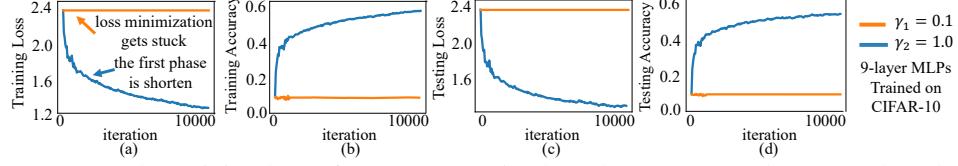


Figure 24: (a) The training loss of two MLPs trained on the CIFAR-10 dataset. When the loss minimization gets stuck (orange curve), we can consider it as the first phase with an infinite length. Therefore, the learning-sticking problem can be solved by techniques of shortening the first phase, such as the technique of increasing the variance of initial weights, which is a theoretically certificated solution in our study (blue curve). (b) The training accuracy of two MLPs. (c) The testing loss of two MLPs. (d) The testing accuracy of two MLPs.

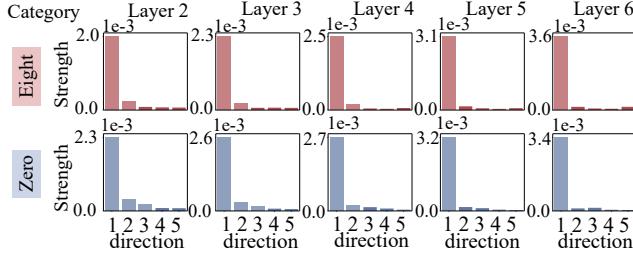


Figure 25: The strength of top-ranked common directions on the MNIST dataset. We trained a 9-layer MLP, where each layer of the MLP had 512 neurons. We computed the strength of common directions on the two categories with the highest training accuracies. $s_i = \|C_i \Delta \bar{V}_i^\top\|_F$ measures the strength of weight changes along the i -th common direction, where $\Delta \bar{V}_i = \mathbb{E}_t[\Delta \bar{V}_{i,t}]$. It can be observed that the strength of the primary direction was much greater than the strength of other directions.

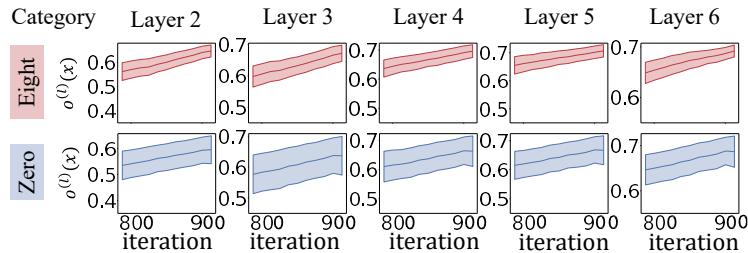


Figure 26: The change of $o^{(l)} = \cos(\Delta V_t^{(l)}, F_t^{(l-1)}) \cdot \cos(V_t^{(l)}, \Delta F_t^{(l-1)})$ in the first phase. We trained a 9-layer MLP on the MNIST dataset. Each layer of the MLP had 512 neurons. The shade represents the standard deviation over different samples.

E Proof for the lemma 1

In this section, we present the detailed proof for Lemma 1.

Lemma 1. *For the decomposition $\Delta W_t^\top = \Delta V_t C^\top + \Delta \varepsilon_t$, given weight changes over different samples ΔW_t^\top , we can compute the common direction C and obtain $\Delta V_t = \frac{\Delta W_t^\top C}{C^\top C}$ and $\Delta \varepsilon_t = \Delta W_t^\top - \Delta W_t^\top \frac{CC^\top}{C^\top C}$ s.t. $\Delta \varepsilon_t C = \mathbf{0}$. Such settings minimize $\|\Delta \varepsilon_t\|_F$.*

proof. Let $\Delta \varepsilon_t^\top[j]$ denote the j -th column of the matrix $\Delta \varepsilon_t^\top \in \mathbb{R}^{h \times d}$. Given a sample x , we can represent $\Delta \varepsilon_t^\top[j]$ by the vector C and a residual term $\Delta \varepsilon_t^\top[j]'$ as follows:

$$\Delta \varepsilon_t^\top[j] = \lambda C + \Delta \varepsilon_t^\top[j]', \quad (1)$$

where $C^\top \Delta \varepsilon_t^\top[j]' = 0$, and λ is a scalar.

Then,

$$\begin{aligned} \|\Delta \varepsilon_t^\top[j]\|_2^2 &= \left\| \lambda C + \Delta \varepsilon_t^\top[j]' \right\|_2^2 \\ &= (\lambda C + \Delta \varepsilon_t^\top[j]')^\top (\lambda C + \Delta \varepsilon_t^\top[j]') \\ &= \lambda^2 C^\top C + (\Delta \varepsilon_t^\top[j]')^\top \Delta \varepsilon_t^\top[j] \\ &= \lambda^2 C^\top C + \left\| \Delta \varepsilon_t^\top[j] \right\|_2^2 \end{aligned} \quad (2)$$

Obviously, $\|\Delta \varepsilon_t^\top[j]\|_2^2$ is the smallest when $\lambda = 0$. In other words, $\Delta \varepsilon_t^\top[j]$ does not contain the component along the direction C and $C^\top \Delta \varepsilon_t^\top[j] = 0$. Therefore, $\|\Delta \varepsilon_t^\top[j]\|_2^2$ reaches its minimum if and only if $\Delta \varepsilon_t C = \mathbf{0}$.

When $\|\Delta \varepsilon_t^\top[j]\|_2^2$ reaches its minimum, $\|\Delta \varepsilon_t\|_F^2$ becomes the smallest. Thus, we have:

$$\begin{aligned} \Delta W_t &= C \Delta V_t^\top + \Delta \varepsilon_t^\top \\ C^\top \Delta W_t &= C^\top C \Delta V_t^\top + C^\top \Delta \varepsilon_t^\top \\ &= C^\top C \Delta V_t^\top + \mathbf{0} \end{aligned} \quad (3)$$

Then, ΔV_t^\top can be represented as follows.

$$\Delta V_t^\top = \frac{C^\top \Delta W_t}{C^\top C} \quad (4)$$

Substituting Eq. 4 into $\Delta W_t = C \Delta V_t^\top + \Delta \varepsilon_t^\top$, we have

$$\Delta \varepsilon_t = \Delta W_t^\top - \Delta W_t^\top \frac{CC^\top}{C^\top C} \quad (5)$$

F Proof for the lemma 2

In this section, we present the detailed proof for Lemma 2.

Lemma 2. *(We can also decompose the weight $W_t^{(l)}$ into the component along the common direction C and the component ε_t in other directions.) Given the weight W_t^\top and the common direction C , the decomposition $W_t^\top = V_t C^\top + \varepsilon_t$ can be conducted as $V_t = \frac{W_t^\top C}{C^\top C}$ and $\varepsilon_t = W_t^\top - W_t^\top \frac{CC^\top}{C^\top C}$ s.t. $\varepsilon_t C = \mathbf{0}$. Such settings minimize $\|\varepsilon_t\|_F$.*

proof. Let $\varepsilon_t^\top[j]$ denote the j -th column of the matrix $\varepsilon_t^\top \in \mathbb{R}^{h \times d}$. We can represent $\varepsilon_t^\top[j]$ by the vector C and a residual term $\varepsilon_t^\top[j]'$ as follows:

$$\varepsilon_t^\top[j] = \lambda C + \varepsilon_t^\top[j]', \quad (6)$$

where $C^\top \varepsilon_t^\top [j]' = 0$ and λ is a scalar.

Then,

$$\begin{aligned}
\|\varepsilon_t^\top [j]\|_2^2 &= \left\| \lambda C + \varepsilon_t^\top (x)[j]' \right\|_2^2 \\
&= (\lambda C + \varepsilon_t^\top [j]')^\top (\lambda C + \varepsilon_t^\top [j]') \\
&= \lambda^2 C^\top C + (\varepsilon_t^\top [j]')^\top \varepsilon_t^\top [j]' \\
&= \lambda^2 C^\top C + \left\| \varepsilon_t^\top [j] \right\|_2^2
\end{aligned} \tag{7}$$

Obviously, $\|\varepsilon_t^\top [j]\|_2^2$ becomes the smallest when $\lambda = 0$. In other words, $\varepsilon_t^\top [j]$ does not contain the component along the direction C and $C^\top \varepsilon_t^\top [j] = 0$. Therefore, $\|\varepsilon_t^\top [j]\|_2^2$ reaches its minimum if and only if $\varepsilon_t C = \mathbf{0}$.

When $\|\varepsilon_t^\top [j]\|_2^2$ reaches its minimum, $\|\varepsilon_t\|_F^2$ becomes the smallest. Thus, we have:

$$\begin{aligned}
W_t &= CV_t^\top + \varepsilon_t^\top \\
C^\top W_t &= C^\top CV_t^\top + C^\top \varepsilon_t^\top \\
&= C^\top CV_t^\top + \mathbf{0}
\end{aligned} \tag{8}$$

Then, V_t^\top can be written as follows.

$$V_t^\top = \frac{C^\top W_t}{C^\top C} \tag{9}$$

Substituting Eq. 9 into $W_t = CV_t^\top + \varepsilon_t^\top$, we have

$$\varepsilon_t = W_t^\top - V_t^\top \frac{CC^\top}{C^\top C} \tag{10}$$

G Decomposition of common directions

Actually, the estimation of the common direction C is similar to the singular value decomposition (SVD), although there are slight differences.

We compute the average weight change $\Delta \bar{W}_t = \mathbb{E}_{x \in X} \Delta W_t|_x$, where $\Delta W_t|_x$ denotes the weight change made by the sample x . Then, we decompose $\Delta \bar{W}_t$ into components along five common directions as $\Delta \bar{W}_t = C_1 \Delta \bar{V}_{1,t}^\top + C_2 \Delta \bar{V}_{2,t}^\top + \dots + C_5 \Delta \bar{V}_{5,t}^\top + \Delta \bar{\varepsilon}_t^\top$, where $C_1 = C$ is termed the *primary common direction*. C_1, C_2, C_3, C_4 , and C_5 are orthogonal to each other. C_2, C_3, C_4 and C_5 represent the second, third, forth, and fifth common directions, respectively. C_i represents the i -th common direction. $\Delta \bar{V}_{i,t}$ denotes the average weight change along the i -th common direction decomposed from $\Delta \bar{W}_t$.

Specifically, we first decompose the average weight change $\Delta \bar{W}_t$ after the t -th iteration as $\Delta \bar{W}_t = C \Delta \bar{V}_t^\top + \Delta \bar{\varepsilon}_t^\top$. We remove all components along the common direction C from $\Delta \bar{W}_t$, and obtain $\Delta \bar{W}_{\text{new},t} = \Delta \bar{W}_t - C \Delta \bar{V}_t^\top = \Delta \bar{\varepsilon}_t^\top$. Then, we further decompose $\Delta \bar{W}_{\text{new},t} = C_2 \Delta V_{2,t}^\top + \Delta \varepsilon_{2,t}^\top$. In this way, we can consider C_2 as the secondary common direction, while $C_1 = C$ is termed as the primary common direction. Thus, we conduct this process recursively and obtain common directions $\{C_1, C_2, \dots, C_5\}$. Accordingly, $\Delta \bar{W}_t$ is decomposed into $\Delta \bar{W}_t = C_1 \Delta \bar{V}_{1,t}^\top + C_2 \Delta V_{2,t}^\top + \dots + C_5 \Delta V_{5,t}^\top + \Delta \varepsilon_{5,t}^\top$.

H Decomposition of the weight change made by a sample x

H.1 Proof for Theorem 1.

In this subsection, we present the detailed proof for Theorem 1.

Theorem 1. *The weight change made by a sample can be decomposed into $(h + 1)$ terms after the t -th iteration as follows.*

$$\Delta W_t^{(l)} = \Delta W_{\text{primary},t}^{(l)} + \sum_{k=1}^h \Delta W_{\text{noise},t}^{(l,k)} \xrightarrow{\text{rewritten}} \Gamma_t^{(l)} F_t^{(l-1)\top} + \kappa_t^{(l)\top}, \quad (11)$$

where $\Delta W_{\text{primary},t}^{(l)} = D_t^{(l)} V_t^{(l+1)\top} C^{(l+1)\top} \Delta V_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2$ denotes the component along the primary common direction, and $\Delta W_{\text{noise},t}^{(l,k)} = D_t^{(l)} \varepsilon_t^{(l+1,k)} \Delta \varepsilon_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2$ denotes the component along the k -th common direction in the noise term. $\varepsilon_t^{(l+1,k)} = \Sigma_{kk} \mathcal{U}_k \mathcal{V}_k^\top$, where the SVD of $\varepsilon_t^{(l+1)} \in \mathbb{R}^{h \times h'}$ is given as $\varepsilon_t^{(l+1)} = \mathcal{U} \Sigma \mathcal{V}^\top$ ($h \leq h'$), and Σ_{kk} denotes the k -th singular value $\in \mathbb{R}$. $\varepsilon_t^{(l+1)} = \sum_k \varepsilon_t^{(l+1,k)}$. \mathcal{U}_k and \mathcal{V}_k denote the k -th column of the matrix \mathcal{U} and \mathcal{V} , respectively. Besides, we have $\forall k \in \{1, 2, \dots, h\}$, $\mathcal{U}_k^\top C^{(l+1)} = 0$. Consequently, we have $\Gamma_t^{(l)} = D_t^{(l)} V_t^{(l+1)\top} C^{(l+1)\top} \Delta V_t^{(l+1)\top} F_t^{(l)} / \|F_t^{(l)}\|_2^2 \in \mathbb{R}^h$, and $\kappa_t^{(l)\top} = D_t^{(l)} \varepsilon_t^{(l+1)\top} \Delta \varepsilon_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2 \in \mathbb{R}^{h \times d}$.

proof. We can represent weight matrix as $W_t^{(l)} = C^{(l)} V_t^{(l)\top} + \varepsilon_t^{(l)\top}$. In addition, according to the back propagation and chain rule, we have $\Delta W_t^{(l)} = -\eta D_t^{(l)} \dot{F}_t^{(l)} F_t^{(l-1)\top}$, where $\dot{F}_t^{(l)} = \frac{\partial \text{Loss}}{\partial F_t^{(l)}}$, and η denotes the learning rate.

According to Lemma 1 and Lemma 2, we have $\Delta \varepsilon_t^{(l+1)} C^{(l+1)} = \mathbf{0}$ and $\varepsilon_t^{(l+1)} C^{(l+1)} = \mathbf{0}$. After the t -th iteration, the weight change made by a training sample x can be computed as follows.

$$\begin{aligned} \Delta W_t^{(l)} &= -\eta D_t^{(l)} \dot{F}_t^{(l)} F_t^{(l-1)\top} \\ &= -\eta D_t^{(l)} W_t^{(l+1)\top} D_t^{(l+1)} \dot{F}_t^{(l+1)} F_t^{(l-1)\top} \\ &= D_t^{(l)} W_t^{(l+1)\top} \Delta W_t^{(l+1)} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2 \\ &= D_t^{(l)} [V_t^{(l+1)\top} C^{(l+1)\top} + \varepsilon_t^{(l+1)}] [C^{(l+1)\top} \Delta V_t^{(l+1)\top} + \Delta \varepsilon_t^{(l+1)\top}] F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2 \\ &= D_t^{(l)} [V_t^{(l+1)\top} C^{(l+1)\top} C^{(l+1)\top} \Delta V_t^{(l+1)\top} + V_t^{(l+1)\top} C^{(l+1)\top} \Delta \varepsilon_t^{(l+1)\top} \\ &\quad + \varepsilon_t^{(l+1)\top} C^{(l+1)\top} \Delta V_t^{(l+1)\top} + \varepsilon_t^{(l+1)\top} \Delta \varepsilon_t^{(l+1)\top}] F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2 \\ &= D_t^{(l)} [V_t^{(l+1)\top} C^{(l+1)\top} C^{(l+1)\top} \Delta V_t^{(l+1)\top} + \varepsilon_t^{(l+1)\top} \Delta \varepsilon_t^{(l+1)\top}] F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2 \\ &= D_t^{(l)} V_t^{(l+1)\top} C^{(l+1)\top} C^{(l+1)\top} \Delta V_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2 \\ &\quad + D_t^{(l)} \varepsilon_t^{(l+1)\top} \Delta \varepsilon_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2 \end{aligned} \quad (12)$$

$\varepsilon_t^{(l+1,k)} = \Sigma_{kk} \mathcal{U}_k \mathcal{V}_k^\top$, where the singular value decomposition of $\varepsilon_t^{(l+1)}$ is given as $\varepsilon_t^{(l+1)} = \mathcal{U} \Sigma \mathcal{V}^\top$, and Σ_{kk} denotes the k -th singular value. \mathcal{U}_k and \mathcal{V}_k denote the k -th column of the matrix \mathcal{U} and \mathcal{V} , respectively. We can derive the following equations.

$$\begin{aligned} \Delta W_t^{(l)} &= D_t^{(l)} V_t^{(l+1)\top} C^{(l+1)\top} C^{(l+1)\top} \Delta V_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2 \\ &\quad + D_t^{(l)} \varepsilon_t^{(l+1)\top} \Delta \varepsilon_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2 \\ &= D_t^{(l)} V_t^{(l+1)\top} C^{(l+1)\top} C^{(l+1)\top} \Delta V_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2 \\ &\quad + \sum_{k=1}^h D_t^{(l)} \varepsilon_t^{(l+1,k)} \Delta \varepsilon_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2. \\ &= \Delta W_{\text{primary},t}^{(l)} + \sum_{k=1}^h \Delta W_{t,\text{noise}}^{(l,k)} \end{aligned} \quad (13)$$

In addition, if we set $\Gamma_t^{(l)} = D_t^{(l)} V_t^{(l+1)} C^{(l+1)\top} C^{(l+1)} \Delta V_t^{(l+1)\top} F_t^{(l)} / \|F_t^{(l)}\|_2^2$, and $\kappa_t^{(l)\top} = D_t^{(l)} \varepsilon_t^{(l+1)} \Delta \varepsilon_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2$. Then we can re-write the Eq. (13) as follows.

$$\Delta W_t^{(l)} \xrightarrow{\text{rewritten}} \Gamma_t^{(l)} F_t^{(l-1)\top} + \kappa_t^{(l)\top} \quad (14)$$

H.2 The explanation for the phenomenon that $S_1^{(l)}$, $S_2^{(l)}$, and $S_3^{(l)}$ does not decrease monotonically.

In this subsection, we explain the phenomenon that $S_1^{(l)}$, $S_2^{(l)}$, and $S_3^{(l)}$ does not decrease monotonically in Table 1 in the supplementary material and Table 1 in the main paper (Page 6). In fact, we first decompose $\varepsilon_t^{(l+1)} = \sum_k \varepsilon_t^{(l+1,k)}$ according to the SVD. Then $\Delta W_{\text{noise},t}^{(l,k)}$ is computed as $\Delta W_{\text{noise},t}^{(l,k)} = D_t^{(l)} \varepsilon_t^{(l+1,k)} \Delta \varepsilon_t^{(l+1)\top} F_t^{(l)} F_t^{(l-1)\top} / \|F_t^{(l)}\|_2^2$. Accordingly, the strength of weight changes along the primary direction is computed as $S_{\text{primary}}^{(l)} = \mathbb{E}_{t \in [T_{\text{start}}, T_{\text{end}}]} \mathbb{E}_{x \in X} [\|\Delta W_{\text{primary},t}^{(l,k)}\|_F]$. The strength of weight changes along the k -th noise direction is computed as $S_k^{(l)} = \mathbb{E}_{t \in [T_{\text{start}}, T_{\text{end}}]} \mathbb{E}_{x \in X} [\|\Delta W_{\text{noise},t}^{(l,k)}\|_F]$. In this way, $S_1^{(l)}$, $S_2^{(l)}$, and $S_3^{(l)}$ do not decrease monotonically, although $\|\varepsilon_t^{(l+1,1)}\|_F$, $\|\varepsilon_t^{(l+1,2)}\|_F$, and $\|\varepsilon_t^{(l+1,3)}\|_F$ are directly decomposed from $\varepsilon_t^{(l+1)}$ based on the SVD and decrease monotonically.

I Analysis based on Eq. (3) in the main paper and explanation for the parallelism.

According the Eq. (3) in the main paper, we have

$$\dot{F}_t^{(l-1)} = (C^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)}) \cdot \beta + \epsilon D_t^{(l)} \dot{F}_t^{(l)} \quad (15)$$

Thus, if $C^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)}$ is large enough (*i.e.*, keeping optimizing $W_t^{(l)\top}$ along the common direction $C^{(l)}$ for a long time), then the feature gradients $\dot{F}_t^{(l-1)}$ of different samples will be roughly parallel to the same vector β . This is because $C^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)}$ is a scalar and the term $\epsilon D_t^{(l)} \dot{F}_t^{(l)}$ is small. In other words, the diversity between feature gradients $\dot{F}_t^{(l-1)}$ of different samples decreases. Here, $\beta = [\beta_1, \beta_2, \dots, \beta_d]$, and $\epsilon = [\epsilon_1, \epsilon_2, \dots, \epsilon_d]^\top$.

J Discussion on the background assumption.

In the above section, we demonstrate that on the ideal state, *i.e.*, $W_t^{(l)\top}$ has been optimized towards the common direction $C^{(l)}$ for a long time, we can consider that the feature gradients $\dot{F}_t^{(l-1)}$ of different samples will be roughly parallel to the same vector β . In this way, we can explain that the diversity between feature gradients $\dot{F}_t^{(l-1)}$ of different samples decreases.

In comparison, in the current section, we mainly discuss the trustworthiness of the background assumption in Line 214 in the main paper. We aim to discuss that on the assumption that features $F_t^{(l-1)}$ of different samples have been pushed a little bit towards a specific common direction, we can find at least one learning iteration in the first phase where $\Delta F_t^{(l-1)}$ and $F_t^{(l-1)}$ of most samples have similar directions, and $V_t^{(l)}$ and $\Delta V_t^{(l)}$ have similar directions. The assumption that features $F_t^{(l-1)}$ of different samples have been pushed a little bit towards a specific common direction is an intermediate state between the chaotic initial state of the MLP and the ideal state introduced in the above section. In this way, we can assume that $C^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)}$ is large.

According to Eq. (2) in the main paper and Lemma 2, we have $\dot{F}_t^{(l-1)} = W_t^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)}$ and $W_t^{(l)\top} = V_t^{(l)} C^{(l)\top} + \varepsilon_t^{(l)\top}$. Thus, we have

$$\begin{aligned} \dot{F}_t^{(l-1)} &= W_t^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)} \\ &= (V_t^{(l)} C^{(l)\top} + \varepsilon_t^{(l)\top}) D_t^{(l)} \dot{F}_t^{(l)} \\ &= V_t^{(l)} C^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)} + \varepsilon_t^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)} \end{aligned} \quad (16)$$

If the scalar $C^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)}$ is large, we can roughly consider

$$\begin{aligned}\dot{F}_t^{(l-1)} &\approx V_t^{(l)} C^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)} \\ &= V_t^{(l)} \cdot (C^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)}) // V_t^{(l)}\end{aligned}\tag{17}$$

It means that the feature gradient $\dot{F}_t^{(l-1)}$ is roughly parallel to the vector $V_t^{(l)}$. Furthermore, the feature gradient $\dot{F}_t^{(l-1)}$ and the change of feature $\Delta F_t^{(l-1)}$ can be considered negatively parallel to each other, we have

$$\Delta F_t^{(l-1)} // \dot{F}_t^{(l-1)} // V_t^{(l)}\tag{18}$$

Similarly, we have $\Delta F_{t+1}^{(l-1)} // V_{t+1}^{(l)}$. Therefore, we can roughly consider that $V_t^{(l)} \approx k_t \Delta F_t^{(l-1)}$, and $V_{t+1}^{(l)} \approx k_{t+1} \Delta F_{t+1}^{(l-1)}$, where $k_t, k_{t+1} \in \mathbb{R}$ are two scalars. Then, we can derive that

$$\Delta V_t^{(l)} = V_{t+1}^{(l)} - V_t^{(l)} \approx k_{t+1} \Delta F_{t+1}^{(l-1)} - k_t \Delta F_t^{(l-1)}\tag{19}$$

If features $F_t^{(l-1)}$ of different samples have been pushed a little bit towards a specific common direction, then it is easy to find at least one learning iteration that $\Delta F_t^{(l-1)}$ and $F_t^{(l-1)}$ of most samples have similar directions, *i.e.* $\Delta F_t^{(l-1)} // F_t^{(l-1)}$. Meanwhile, we can find at least one learning iteration in the first phase where the change of feature in t -th iteration $\Delta F_t^{(l-1)}$ and $(t+1)$ -th iteration $\Delta F_{t+1}^{(l-1)}$ are roughly the same. In other words, $\Delta F_t^{(l-1)} \approx \Delta F_{t+1}^{(l-1)}$. Thus, we have

$$\Delta V_t^{(l)} \approx (k_{t+1} - k_t) \Delta F_t^{(l-1)} // \Delta F_t^{(l-1)} // V_t^{(l)}\tag{20}$$

In this way, we can obtain that $V_t^{(l)}$ and $\Delta V_t^{(l)}$ have similar directions.

K Proof for Lemma 3

In this section, we present the detailed proof for Lemma 3.

Lemma 3. *Given an input sample $x \in X$ and a common direction $C^{(l)}$ after the t -th iteration, if the noise term $\varepsilon_t^{(l)}$ is small enough to satisfy $|\Delta V_t^{(l)\top} F_t^{(l-1)} V_t^{(l)\top} V_t^{(l)\top} C^{(l)\top} C^{(l)} \Delta V_t^{(l)\top} F_t^{(l-1)}| \gg |\Delta V_t^{(l)\top} F_t^{(l-1)} V_t^{(l)\top} \varepsilon_t^{(l)\top} \Delta \varepsilon_t^{(l)\top} F_t^{(l-1)}|$, we can obtain $\cos(\Delta V_t^{(l)}, F_t^{(l-1)}) \cdot \cos(V_t^{(l)}, \Delta F_t^{(l-1)}) \geq 0$, where $\Delta V_t^{(l)} = \frac{\Delta W_t^{(l)\top} C^{(l)}}{C^{(l)\top} C^{(l)}}$, and $V_t^{(l)} = \frac{W_t^{(l)\top} C^{(l)}}{C^{(l)\top} C^{(l)}}$. $\Delta F_t^{(l-1)}$ denotes the change of features $\Delta F_t^{(l-1)} = F_{t+1}^{(l-1)} - F_t^{(l-1)}$ made by the training sample x after the t -th iteration. To this end, we approximately consider the change of features $\Delta F_t^{(l-1)}$ after the t -th iteration negatively parallel to feature gradients $\dot{F}_t^{(l-1)}$, although strictly speaking, the change of features is not exactly equal to the gradient w.r.t. features.*

proof. Given a sample x , we can prove that $\cos(\Delta V_t^{(l)}, F_t^{(l-1)}) \cdot \cos(V_t^{(l)}, \Delta F_t^{(l-1)}) \geq 0$.

According to chain rule, we have

$$\Delta W_t^{(l)} = -\eta D_t^{(l)} \dot{F}_t^{(l)} F_t^{(l-1)\top} \quad (21)$$

According to Lemma 1 and Lemma 2, we have $C^{(l)\top} \Delta \varepsilon_t^{(l)\top} = 0$ and $\varepsilon_t^{(l)\top} C^{(l)} = 0$. Then, we have

$$\cos(\Delta V_t^{(l)}, F_t^{(l-1)}) \cdot \cos(V_t^{(l)}, \dot{F}_t^{(l-1)}) = \left[\frac{\Delta V_t^{(l)\top} F_t^{(l-1)}}{\|\Delta V_t^{(l)}\|_2 \|F_t^{(l-1)}\|_2} \right] \cdot \left[\frac{V_t^{(l)\top} \dot{F}_t^{(l-1)}}{\|V_t^{(l)}\|_2 \|\dot{F}_t^{(l-1)}\|_2} \right] \quad (22)$$

Therefore, we have

$$\begin{aligned} & \text{sign}(\cos(\Delta V_t^{(l)}, F_t^{(l-1)}) \cdot \cos(V_t^{(l)}, \dot{F}_t^{(l-1)})) \\ &= \text{sign}([\Delta V_t^{(l)\top} F_t^{(l-1)}] \cdot [V_t^{(l)\top} \dot{F}_t^{(l-1)}] / (\|\Delta V_t^{(l)}\|_2 \|F_t^{(l-1)}\|_2 \|V_t^{(l)}\|_2 \|\dot{F}_t^{(l-1)}\|_2)) \\ &= \text{sign}([\Delta V_t^{(l)\top} F_t^{(l-1)}] \cdot [V_t^{(l)\top} W_t^{(l)\top} D_t^{(l)} \dot{F}_t^{(l)}] / (\|\Delta V_t^{(l)}\|_2 \|F_t^{(l-1)}\|_2 \|V_t^{(l)}\|_2 \|\dot{F}_t^{(l-1)}\|_2)) \\ &= \text{sign}([\Delta V_t^{(l)\top} F_t^{(l-1)}] \cdot [V_t^{(l)\top} (V_t^{(l)\top} C^{(l)\top} + \varepsilon_t^{(l)\top}) D_t^{(l)} \dot{F}_t^{(l)}] / (\|\Delta V_t^{(l)}\|_2 \|F_t^{(l-1)}\|_2 \|V_t^{(l)}\|_2 \|\dot{F}_t^{(l-1)}\|_2)) \\ &= \text{sign}([\Delta V_t^{(l)\top} F_t^{(l-1)}] \cdot [V_t^{(l)\top} (V_t^{(l)\top} C^{(l)\top} + \varepsilon_t^{(l)\top}) (\Delta W_t^{(l)} F_t^{(l-1)} / (-\eta \|F_t^{(l-1)}\|_2^2))]) \\ &/ (\|\Delta V_t^{(l)}\|_2 \|F_t^{(l-1)}\|_2 \|V_t^{(l)}\|_2 \|\dot{F}_t^{(l-1)}\|_2)) \\ &= \text{sign}([\Delta V_t^{(l)\top} F_t^{(l-1)}] \cdot [(V_t^{(l)\top} V_t^{(l)\top} C^{(l)\top} + V_t^{(l)\top} \varepsilon_t^{(l)\top}) \Delta W_t^{(l)} F_t^{(l-1)}] \\ &/ (-\eta \|F_t^{(l-1)}\|_2^2 \|\Delta V_t^{(l)}\|_2 \|F_t^{(l-1)}\|_2 \|V_t^{(l)}\|_2 \|\dot{F}_t^{(l-1)}\|_2)) \\ &= \text{sign}([\Delta V_t^{(l)\top} F_t^{(l-1)}] \cdot [(V_t^{(l)\top} V_t^{(l)\top} C^{(l)\top} + V_t^{(l)\top} \varepsilon_t^{(l)\top}) (C^{(l)\top} \Delta V_t^{(l)\top} + \Delta \varepsilon_t^{(l)\top} F_t^{(l-1)})] \\ &/ (-\eta \|F_t^{(l-1)}\|_2^2 \|\Delta V_t^{(l)}\|_2 \|F_t^{(l-1)}\|_2 \|V_t^{(l)}\|_2 \|\dot{F}_t^{(l-1)}\|_2)) \\ &= \text{sign}([\Delta V_t^{(l)\top} F_t^{(l-1)}] \cdot [(V_t^{(l)\top} V_t^{(l)\top} C^{(l)\top} C^{(l)\top} \Delta V_t^{(l)\top} + V_t^{(l)\top} \varepsilon_t^{(l)\top} \Delta \varepsilon_t^{(l)\top} \\ &+ V_t^{(l)\top} V_t^{(l)\top} C^{(l)\top} \Delta \varepsilon_t^{(l)\top} + V_t^{(l)\top} \varepsilon_t^{(l)\top} C^{(l)\top} \Delta V_t^{(l)\top}) F_t^{(l-1)}] / (-\eta \|F_t^{(l-1)}\|_2^2 \|\Delta V_t^{(l)}\|_2 \|\dot{F}_t^{(l-1)}\|_2 \|V_t^{(l)}\|_2 \|F_t^{(l-1)}\|_2)) \\ &= \text{sign}([\Delta V_t^{(l)\top} F_t^{(l-1)}] \cdot [(V_t^{(l)\top} V_t^{(l)\top} C^{(l)\top} C^{(l)\top} \Delta V_t^{(l)\top} + V_t^{(l)\top} \varepsilon_t^{(l)\top} \Delta \varepsilon_t^{(l)\top} F_t^{(l-1)}]) \\ &/ (-\eta \|F_t^{(l-1)}\|_2^2 \|\Delta V_t^{(l)}\|_2 \|F_t^{(l-1)}\|_2 \|V_t^{(l)}\|_2 \|\dot{F}_t^{(l-1)}\|_2)) \\ &= \text{sign}([\Delta V_t^{(l)\top} F_t^{(l-1)}] \cdot [V_t^{(l)\top} V_t^{(l)\top} C^{(l)\top} C^{(l)\top} \Delta V_t^{(l)\top} F_t^{(l-1)} + V_t^{(l)\top} \varepsilon_t^{(l)\top} \Delta \varepsilon_t^{(l)\top} F_t^{(l-1)}] \\ &/ (-\eta \|F_t^{(l-1)}\|_2^2 \|\Delta V_t^{(l)}\|_2 \|F_t^{(l-1)}\|_2 \|V_t^{(l)}\|_2 \|\dot{F}_t^{(l-1)}\|_2)) \\ &= \text{sign}([\Delta V_t^{(l)\top} F_t^{(l-1)}] \cdot [V_t^{(l)\top} V_t^{(l)\top} C^{(l)\top} C^{(l)\top} \Delta V_t^{(l)\top} F_t^{(l-1)} + \Delta V_t^{(l)\top} F_t^{(l-1)} V_t^{(l)\top} \varepsilon_t^{(l)\top} \Delta \varepsilon_t^{(l)\top} F_t^{(l-1)}] \\ &/ (-\eta \|F_t^{(l-1)}\|_2^2 \|\Delta V_t^{(l)}\|_2 \|F_t^{(l-1)}\|_2 \|V_t^{(l)}\|_2 \|\dot{F}_t^{(l-1)}\|_2)) \end{aligned} \quad (23)$$

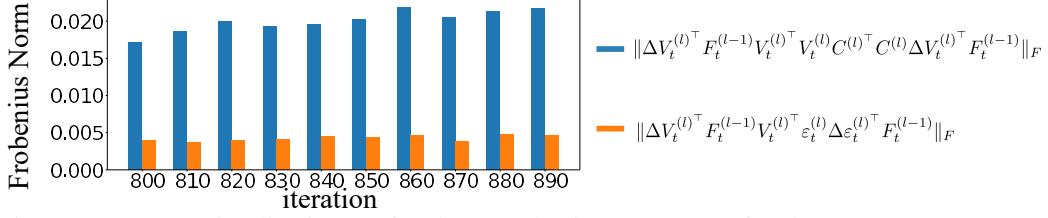


Figure 27: Visualization of the Frobenius norm of the two components $\Delta V_t^{(l)\top} F_t^{(l-1)} V_t^{(l)\top} V_t^{(l)} C^{(l)\top} C^{(l)} \Delta V_t^{(l)\top} F_t^{(l-1)}$ and $\Delta V_t^{(l)\top} F_t^{(l-1)} V_t^{(l)\top} \varepsilon_t^{(l)} \Delta \varepsilon_t^{(l)\top} F_t^{(l-1)}$. We trained a 9-layer MLP on the MNIST dataset, where each layer had 512 neurons. Iterations were chosen at the end of the first phase.

According to our assumption, the noise term $\varepsilon_t^{(l)}$ is small enough to satisfy $|\Delta V_t^{(l)\top} F_t^{(l-1)} V_t^{(l)\top} V_t^{(l)} C^{(l)\top} C^{(l)} \Delta V_t^{(l)\top} F_t^{(l-1)}| \gg |\Delta V_t^{(l)\top} F_t^{(l-1)} V_t^{(l)\top} \varepsilon_t^{(l)} \Delta \varepsilon_t^{(l)\top} F_t^{(l-1)}|$. This assumption is verified in Figure 27. Then we can ignore the last term and obtain

$$\begin{aligned}
& \text{sign}([\Delta V_t^{(l)\top} F_t^{(l-1)} V_t^{(l)\top} V_t^{(l)} C^{(l)\top} C^{(l)} \Delta V_t^{(l)\top} F_t^{(l-1)} + \Delta V_t^{(l)\top} F_t^{(l-1)} V_t^{(l)\top} \varepsilon_t^{(l)} \Delta \varepsilon_t^{(l)\top} F_t^{(l-1)}]) \\
& /(-\eta \|F_t^{(l-1)}\|_2^2 \|\Delta V_t^{(l)}\|_2 \|F_t^{(l-1)}\|_2 \|V_t^{(l)}\|_2 \|\dot{F}_t^{(l-1)}\|_2) \\
& \approx \text{sign}([\Delta V_t^{(l)\top} F_t^{(l-1)} V_t^{(l)\top} V_t^{(l)} C^{(l)\top} C^{(l)} \Delta V_t^{(l)\top} F_t^{(l-1)}] \\
& (-\eta \|F_t^{(l-1)}\|_2^2 \|\Delta V_t^{(l)}\|_2 \|F_t^{(l-1)}\|_2 \|V_t^{(l)}\|_2 \|\dot{F}_t^{(l-1)}\|_2)) \leq 0
\end{aligned} \tag{24}$$

Thus,

$$\text{sign}(\cos(\Delta V_t^{(l)}, F_t^{(l-1)}) \cdot \cos(V_t^{(l)}, \dot{F}_t^{(l-1)})) \leq 0 \tag{25}$$

In this paper, we approximately consider $\Delta F_t^{(l-1)}$ and $\dot{F}_t^{(l-1)}$ are negatively parallel to each other. Thus, we have $\text{sign}(\cos(\Delta V_t^{(l)}, F_t^{(l-1)}) \cdot \cos(V_t^{(l)}, \Delta F_t^{(l-1)})) = \text{sign}(\cos(\Delta V_t^{(l)}, F_t^{(l-1)}) \cdot (-\cos(V_t^{(l)}, \dot{F}_t^{(l-1)}))) \geq 0$.

L Proof for Theorem 2

In this section, we aim to prove that training samples of the same category have the same effect in the first phase.

Theorem 2. Under the background assumption, for any training samples $x, x' \in X_c$ in the category c , if $[C^{(l)\top} D_t^{(l)}|_x \dot{F}_t^{(l)}|_x] \cdot [C^{(l)\top} D_t^{(l)}|_{x'} \dot{F}_t^{(l)}|_{x'}] > 0$ (means that $F_t^{(l)}|_x$ and $F_t^{(l)}|_{x'}$ have kinds of similarity in very early iterations), then $\cos(\alpha_c \Delta V_t^{(l)}|_x, F_t^{(l-1)}|_x) \geq 0$, and $\cos(\alpha_c V_t^{(l)}, \Delta F_t^{(l-1)}|_x) \geq 0$, where $\alpha_c \in \{-1, +1\}$ is a constant shared by all samples in category c .

proof. Given a sample x and a sample x' from the same category, we can prove that $\cos(\Delta V_t^{(l)}|_x, F_t^{(l-1)}|_x) \cdot \cos(\Delta V_t^{(l)}|_{x'}, F_t^{(l-1)}|_{x'}) \geq 0$.

$$\begin{aligned}
& \text{sign}(\cos(\Delta V_t^{(l)}|_x, F_t^{(l-1)}|_x) \cdot \cos(\Delta V_t^{(l)}|_{x'}, F_t^{(l-1)}|_{x'})) \\
& = \text{sign}([\Delta V_t^{(l)\top}|_x F_t^{(l-1)}|_x \cdot [\Delta V_t^{(l)\top}|_{x'} F_t^{(l-1)}|_{x'}]] \\
& = \text{sign}([\frac{C^{(l)\top} \Delta W_t^{(l)}|_x}{C^{(l)\top} C^{(l)}} F_t^{(l-1)}|_x \cdot [\frac{C^{(l)\top} \Delta W_t^{(l)}|_{x'}}{C^{(l)\top} C^{(l)}} F_t^{(l-1)}|_{x'}]]) \\
& = \text{sign}([C^{(l)\top} \Delta W_t^{(l)}|_x F_t^{(l-1)}|_x \cdot [C^{(l)\top} \Delta W_t^{(l)}|_{x'} F_t^{(l-1)}|_{x'}]] \\
& = \text{sign}([C^{(l)\top} (-\eta D_t^{(l)}|_x \dot{F}_t^{(l)}|_x F_t^{(l-1)\top}|_x) F_t^{(l-1)}|_x \cdot [C^{(l)\top} (-\eta D_t^{(l)}|_{x'} \dot{F}_t^{(l)}|_{x'} F_t^{(l-1)\top}|_{x'}) F_t^{(l-1)}|_{x'}]]) \\
& = \text{sign}([C^{(l)\top} D_t^{(l)}|_x \dot{F}_t^{(l)}|_x \cdot [C^{(l)\top} D_t^{(l)}|_{x'} \dot{F}_t^{(l)}|_{x'}])
\end{aligned} \tag{26}$$

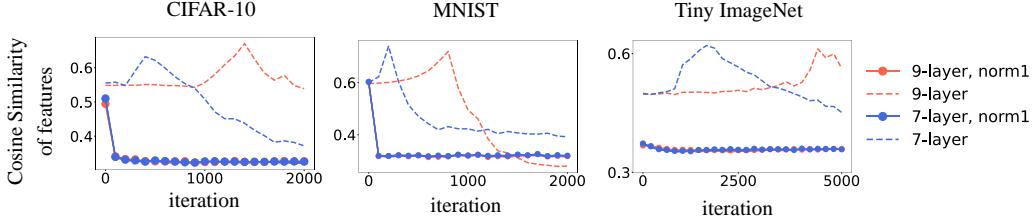


Figure 28: Cosine similarity of features between samples in different categories. We trained 7-layer MLPs and 9-layer MLPs on the CIFAR-10, the MNIST, and the Tiny ImageNet dataset.

According to the assumption that $F_t^{(l)}|_x$ and $F_t^{(l)}|_{x'}$ have kinds of similarity, we can consider $[C^{(l)\top} D_t^{(l)}|_x \dot{F}_t^{(l)}|_x] \cdot [C^{(l)\top} D_t^{(l)}|_{x'} \dot{F}_t^{(l)}|_{x'}] > 0$. In this way, for the category c , there exists a constant α_c , which satisfies $\text{sign}(\cos(\alpha_c \Delta V_t^{(l)}|_x, F_t^{(l-1)}|_x)) \geq 0$, where $\alpha_c \in \{-1, +1\}$ and training sample $x \in X_c$ belongs to the category c .

According to Lemma 3, we have $\cos(\Delta V_t^{(l)}|_x, F_t^{(l-1)}|_x) \cdot \cos(V_t^{(l)}, \Delta F_t^{(l-1)}|_x) \geq 0$. Thus, we have $\text{sign}(\cos(\alpha_c \Delta V_t^{(l)}|_x, F_t^{(l-1)}|_x) \cdot \cos(\alpha_c V_t^{(l)}, \Delta F_t^{(l-1)}|_x)) \geq 0$. In addition, the above proof indicates that $\text{sign}(\cos(\alpha_c \Delta V_t^{(l)}|_x, F_t^{(l-1)}|_x)) \geq 0$. Therefore, we have $\text{sign}(\cos(\alpha_c V_t^{(l)}|_x, \Delta F_t^{(l-1)}|_x)) \geq 0$

M Discussion for four typical operations

M.1 Normalization

The output feature of the l -th linear layer *w.r.t.* the input sample x can be described as $[f_1, f_2, \dots, f_h] = W_t^{(l)} F_t^{(l-1)} \in \mathbb{R}^h$, where f_i denotes the i -th dimension of the feature. In this way, the batch normalization operation can be formulated as $BN(f_i) = \gamma_{scale}[(f_i - \mu_i)/\sigma_i] + \beta_{shift}$, where γ_{scale} and β_{shift} denote the scaling and the shifting parameters, respectively. In this way, the batch normalization operation subtracts the mean feature $\bar{F}_t^{(l)} = \mathbb{E}_{x \in X}[F_t^{(l)}|_x]$ from features of all samples. Therefore, features of different samples in a same category are no longer similar to each other.

We also propose a simplified normalization operation to alleviate the decrease of feature diversity in the first phase. The simplified normalization operation is given as $\text{norm}_1(f_i) = (f_i - \mu_i)/\sigma_i$, where μ_i and σ_i denote the mean value and the standard deviation of f_i over different samples, respectively. This operation is similar to the batch normalization [21], but we do not compute the scaling and shifting parameters in the batch normalization.

In order to verify the simplified normalization operation can alleviate the decrease of feature diversity during the training process of the MLP, we trained 7-layer MLPs and 9-layer MLPs with and without the normalization operation. Specifically, for the normalization operation norm_1 , we added the normalization operation after each linear layer, except the last linear layer. Each linear layer in the MLP had 512 neurons. Figure 28 shows that the feature similarity in MLPs with normalization operations kept decreasing, while the feature similarity of the MLP without normalization operations kept increasing. This indicated that normalization operations alleviate the decreasing of feature diversity.

M.2 Momentum

We can explain that momentum in gradient descent can alleviate this phenomenon. Based on Lemma 3, the self-enhanced system of the decreasing of feature diversity requires singular values of weights along other directions $\varepsilon_t^{(l)}$ to be small enough. However, because the momentum operation strengthens influences of the initialized noisy weights $W_{t=0}^{(l)}$, it strengthens singular values of $\varepsilon_t^{(l)}$, to some extent, thereby alleviating the decrease of feature diversity.

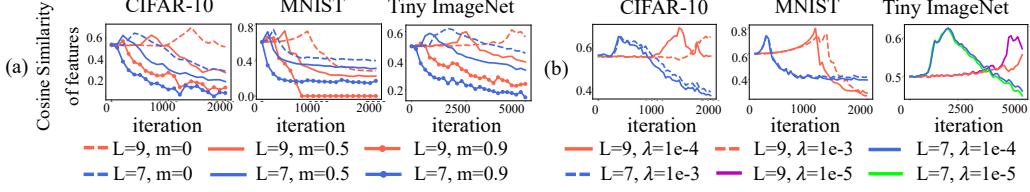


Figure 29: Effects of (a) momentum and (b) L_2 regularization. We trained L -layer MLPs, where each layer had 512 neurons. A shorter first phase indicates that the decrease of feature diversity is more alleviated.

Specifically, consider the momentum with the coefficient m , the dynamics of weights W_{t+1} can be described as,

$$W_{t+1} = W_t - \eta \frac{\partial \text{Loss}}{\partial W_t} - m \frac{\partial \text{Loss}}{\partial W_{t-1}}, \quad (27)$$

where η denotes the learning rate. Because we only focus on weights in a single layer, without causing ambiguity, we omit the superscript (l) to simplify the notation in this subsection. In this way, we can write the gradient descent as

$$W_{T+1} = W_0 + \eta \sum_t^T \frac{1 - m^{T+1-t}}{1 - m} \frac{\partial \text{Loss}}{\partial W_t}. \quad (28)$$

Since $0 < m < 1$, the coefficient $\frac{1-m^{T+1-t}}{1-m}$ decreases when the variable t increases. Thus, a large m represents that influences of W_0 on W_{T+1} are significant. Because ε_{T+1} is decomposed from W_{T+1} and singular values of ε_{T+1} are mainly determined by the noisy W_0 . Accordingly, singular values of ε_{T+1} are relatively large, which disturb the self-enhanced system and alleviate the decrease of feature diversity. Figure 29(a) verifies that a larger value of m usually more alleviates the decrease of feature diversity.

M.3 Initialization

We explain that the initialization of MLPs also affects the decrease of feature diversity. According to Lemma 3, such self-enhanced system requires singular values of weights along other directions $\varepsilon_t^{(l)}$ to be small enough. However, because increasing the variance of the initialized weights $W_0^{(l)}$ will increase singular values of $\varepsilon_t^{(l)}$ based on Lemma 2, alleviating the decrease of feature diversity. Specifically, we initialize weights with Xavier normal distribution [17], i.e. $W_0 \sim \mathcal{N}(0, \gamma \sigma_{var}^2)$, where $\sigma_{var} = \sqrt{\frac{2}{fan_{out} + fan_{in}}}$. fan_{in} and fan_{out} denote the input dimension and the output dimension of the linear layer, respectively. In this way, a large γ yields large singular values of initial weights W_0 . Based on Lemma 2, we also have $\varepsilon_0^{(l)} = W_0^{(l)\top} - W_0^{(l)\top} \frac{C^{(l)} C^{(l)\top}}{C^{(l)\top} C^{(l)}}$. Large singular values of initial weights W_0 lead to large singular values of $\varepsilon_0^{(l)}$. Therefore, a large variance of initialized weights disturbs the self-enhanced system and alleviates the decrease of feature diversity.

M.4 L_2 regularization

L_2 regularization is equivalent to the weight decay in the case of gradient descent. The loss function with L_2 regularization and cross entropy loss can be formulated as $\mathcal{L}_t(W_t) = \mathcal{L}_t^{CE}(W_t) + \lambda \|W_t\|_2^2$, where W_t denotes weights of MLPs. In this way, we have the following iterates by using gradient descent

$$\begin{aligned} W_{t+1} &= W_t - \eta \nabla \mathcal{L}_t(W_t) \\ &= W_t - \eta \nabla \mathcal{L}_t^{CE}(W_t) - 2\eta \lambda W_t \\ &= (1 - 2\eta \lambda) W_t - \eta \nabla \mathcal{L}_t^{CE}(W_t), \end{aligned} \quad (29)$$

According to Lemma 3, such self-enhanced system requires singular values of weights along other directions $\varepsilon_t^{(l)}$ to be small enough. Based on Lemma 2, we also have $\varepsilon_t^{(l)} = W_t^{(l)\top} - W_t^{(l)\top} \frac{C^{(l)} C^{(l)\top}}{C^{(l)\top} C^{(l)}}$. In this way, a smaller λ yields larger singular values of $\varepsilon_t^{(l)}$, which disturbs the self-enhanced system

and alleviates the decrease of feature diversity. Figure 29(b) Figure 9(d) verifies that a smaller coefficient λ more alleviated the decrease of feature diversity.