Are Neurons Actually Collapsed? On the Fine-Grained Structure in Neural Representations

Yongyi Yang [1] **Jacob Steinhardt** [2] **Wei Hu** [1]

Abstract

Recent work has observed an intriguing "Neural Collapse" phenomenon in well-trained neural networks, where the last-layer representations of training samples with the same label collapse into each other. This appears to suggest that the last-layer representations are completely determined by the labels, and do not depend on the intrinsic structure of input distribution. We provide evidence that this is not a complete description, and that the apparent collapse hides important fine-grained structure in the representations. Specifically, even when representations apparently collapse, the small amount of remaining variation can still faithfully and accurately captures the intrinsic structure of input distribution. As an example, if we train on CIFAR-10 using only 5 coarse-grained labels (by combining two classes into one superclass) until convergence, we can reconstruct the original 10-class labels from the learned representations via unsupervised clustering. The reconstructed labels achieve 93% accuracy on the CIFAR-10 test set, nearly matching the normal CIFAR-10 accuracy for the same architecture. We also provide an initial theoretical result showing the fine-grained representation structure in a simplified synthetic setting. Our results show concretely how the structure of input data can play a significant role in determining the fine-grained structure of neural representations, going beyond what Neural Collapse predicts.

1. Introduction

Much of the success of deep neural networks has, arguably, been attributed to their ability to learn useful *representations*, or *features*, of the data (Rumelhart et al., 1985). Although neural networks are often trained to optimize a single objective function with no explicit requirements on the inner representations, there is ample evidence suggesting that these learned representations contain rich information about the input data (Levy & Goldberg, 2014; Olah et al., 2017). As a result, formally characterizing and understanding the structural properties of neural representations is of great theoretical and practical interest, and can provide insights on how deep learning works and how to make better use of these representations.

One intriguing phenomenon recently discovered by Papyan et al. (2020) is *Neural Collapse*, which identifies structural properties of last-layer representations during the terminal phase of training (i.e. after zero training error is reached). The simplest of these properties is that the last-layer representations for training samples with the same label collapse into a single point, which is referred to as "variability collapse (NC1)." This is surprising, since the collapsed structure is not necessary to achieve small training or test error, yet it arises consistently in standard architectures trained on standard classification datasets.

A series of recent papers were able to theoretically explain Neural Collapse under a simplified model called the *unconstrained feature model* or *layer-peeled model* (see Section 2 for a list of references). In this model, the last-layer representation of each training sample is treated as a free optimization variable and therefore the training loss essentially has the form of a matrix factorization. Under a variety of different setups, it was proved that the solution to this simplified problem should satisfy Neural Collapse. Although Neural Collapse is relatively well understood in this simplified model, this model completely ignores the role of the input data because the loss function is independent of the input data. Conceptually, this suggests that **Neural Collapse is only determined by the labels** and may happen regardless of the input data distribution. Zhu et al. (2021) provided further empirical support of this claim via a random labeling experiment.

On the other hand, it is conceivable that the **intrinsic structure of the input distribution *should* play a role** in determining the structure of neural net representations. For

[1]University of Michigan, Ann Arbor, Michigan, USA [2]UC Berkeley, Berkeley, California, USA. Correspondence to: Wei Hu <vvh@umich.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

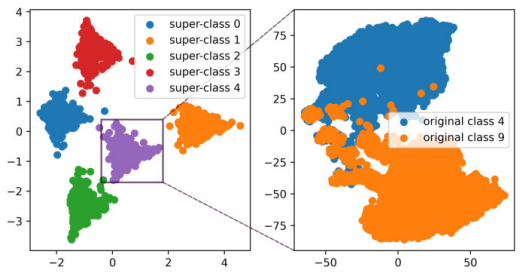


Figure 1. Fine-grained clustering structure of the last-layer representations of ResNet-18 trained on Coarse CIFAR-10 (5 superclasses). Left figure: PCA visualization for all training samples. Right figure: t-SNE visualization for all training samples in superclass 4 (which consists of original classes 4 and 9).

example, if a class contains a heterogeneous set of input data (such as different subclasses), it is possible that their heterogeneity is also respected in their feature representations (Sohoni et al., 2020). However, this appears to contradict Neural Collapse, because Neural Collapse would predict that all the representations collapse into each other as long as they have the same class label. This dilemma motivates us to study the following main question in this paper:

*How can we reconcile the roles of the **intrinsic structure of input distribution** vs. the **explicit structure of the labels** in determining the last-layer representations in neural networks?*

Our methodology and findings. To study the above question, we design experiments to manually create a *mismatch* between the intrinsic structure of the input distribution and the explicit labels provided for training in standard classification datasets and measure how the last-layer representations behave in response to our interventions. This allows us to isolate the effect of the input distribution from the effect of labels. As an illustrative example, for the CIFAR-10 dataset (a 10-class classification task), we alter its labels in two different ways, resulting in a coarsely-labeled and a finely-labeled version:

- Coarse CIFAR-10: combine every two class labels into one and obtain a 5-class task (see Figure 2 for an illustration);
- Fine CIFAR-10: split every class label randomly into two labels and obtain a 20-class task.

We train standard network architectures (e.g. ResNet, DenseNet) using SGD on these altered datasets. Our main findings are summarized below.

First, both the intrinsic structure of the input distribution and the explicit labels provided in training clearly affect the structure of the last-layer representations. The effect of input distribution emerges earlier in training, while the effect of labels appears at a later stage. For example, for both Coarse CIFAR-10 and Fine CIFAR-10, at some point the representations naturally form 10 clusters according to the original CIFAR-10 labels (which comes from the intrinsic input structure), even though 5 or 20 different labels are provided for training. Later in training (after 100% training accuracy is reached), the representations collapse into 5 or 20 clusters driven by the explicit labels provided, as predicted by Neural Collapse.

Second, even after Neural Collapse has occurred according to the explicit label information, the seemingly collapsed representations corresponding to each label can still exhibit *fine-grained structures* determined by the input distribution. As an illustration, Figure 1 visualizes the representations from the last epoch of training a ResNet-18 on Coarse CIFAR-10. While globally there are 5 separated clusters as predicted by Neural Collapse, if we zoom in on each cluster, it clearly consists of two subclusters which correspond to the original CIFAR-10 classes. We also find that this phenomenon persists even after a very long training period (e.g. 1,000 epochs), indicating that the effect of input distribution is not destroyed by that of the labels, at least not within a normal training budget.

To further validate our finding that significant input information is present in the last-layer representations despite Neural Collapse, we perform a simple *Cluster-and-Linear-Probe (CLP)* procedure on the representations from ResNet-18 trained on Coarse CIFAR-10, in which we use an unsupervised clustering method to reconstruct the original labels, and then train a linear classifier on top of these representations using the reconstructed labels. We find that CLP can achieve $> 93\%$ accuracy on the original CIFAR-10 test set, matching the standard accuracy of ResNet-18, even though only 5 coarse labels are provided the entire time.

Theoretical result in a synthetic setting. To complement our findings, we provide a theoretical explanation of the fine-grained representation structure in a simplified synthetic setting — a one-hidden-layer neural network trained on coarsely labeled Gaussian mixture data. We prove that such a network trained by gradient descent produces separable hidden-layer representations for different clusters even if they are given the same label for training.

Takeaway. While Neural Collapse is an intriguing phenomenon that consistently happens, we provide concrete evidence showing that it is not the most comprehensive description of the behavior of last-layer representations in practice, as it fails to capture the possible fine-grained prop-

erties determined by the intrinsic structure of the input distribution.

2. Related Work

The Neural Collapse phenomenon was originally discovered by Papyan et al. (2020), and has led to a series of further investigations.

A number of papers Fang et al. (2021); Lu & Steinerberger (2020); Wojtowytsch et al. (2020); Mixon et al. (2022); Zhu et al. (2021); Ji et al. (2021); Han et al. (2021); Zhou et al. (2022); Tirer & Bruna (2022); Yaras et al. (2022) studied a simplified *unconstrained feature model*, also known as *layer-peeled model*, and showed that Neural Collapse provably happens under a variety of settings. This model treats the last-layer representations of all training samples as free optimization variables. By doing this, the loss function no longer depends on the input data, and therefore this line of work is unable to capture any effect of the input distribution on the structure of the representations. Ergen & Pilanci (2021); Tirer & Bruna (2022); Weinan & Wojtowytsch (2022) considered more complicated models but still did not incorporate the role of the input distribution.

Hui et al. (2022) studied the connection of Neural Collapse to generalization and concluded that Neural Collapse occurs only on the training set, not on the test set. Galanti et al. (2021) found that Neural Collapse does generalize to test samples as well as new classes, and used this observation to study transfer learning and few-shot learning.

Sohoni et al. (2020) observed that the last-layer representations of different subclasses within the same class are often separated into different clusters, and used this observation to design an algorithm for improving group robustness. The fine-grained representation phenomenon we observe is in a qualitatively different regime from that of Sohoni et al. (2020). First of all, we focus on the Neural Collapse regime and find that fine-grained representation structure can co-exist with Neural Collapse. Furthermore, Sohoni et al. (2020) looked at settings in which different subclasses have very different accuracies and explicitly attributed the representation separability phenomenon to this performance difference. On the other hand, we find that representation separability can happen even when there is no performance gap between different subclasses.

3. Preliminaries and Setup

Consider a classification dataset $\mathcal{D} = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^n$, where $(\boldsymbol{x}_k, y_k) \in \mathbb{R}^{d'} \times [C]$ is a pair of input features and label, n is the number of samples, d' is the input dimension, and C is the number of classes. Here $[C] = \{1, \ldots, C\}$.

For a given neural network, we denote its last-layer representations corresponding to the dataset $\mathcal{D}$ by $H \in \mathbb{R}^{n \times d}$, i.e. the hidden representation before the final linear transformation, where d is the last-layer dimensionality. For an original class $c \in [C]$, we denote the number of samples in class c by n_c, and the last-layer representation of k-th sample in class c by $\boldsymbol{h}_k^{(c)}$.

3.1. Preliminaries of Neural Collapse

Neural Collapse (Papyan et al., 2020) characterizes 4 phenomena, named NC1-NC4. Here we introduce NC1 and NC2 which concern the structure of the last-layer representations.

NC1, or variability collapse, asserts that the variance of last-layer representations of samples within the same class vanishes as training proceeds. Formally, it can be measured by $\mathsf{NC}_1 = \frac{1}{C}\mathrm{Tr}\left(\Sigma_W \Sigma_B^{\dagger}\right)$ (Papyan et al., 2020; Zhu et al., 2021), which is observed to go to 0. Here Σ_W and Σ_B are defined as

$$\Sigma_W = \frac{1}{C}\sum_{c\in[C]} \frac{1}{n_c}\sum_{i=1}^{n_c}\left(\boldsymbol{h}_i^{(c)} - \boldsymbol{\mu}_c\right)\left(\boldsymbol{h}_i^{(c)} - \boldsymbol{\mu}_c\right)^{\top} \quad (1)$$

and

$$\Sigma_B = \frac{1}{C}\sum_{c\in[C]} \left(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\right)\left(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\right)^{\top}, \quad (2)$$

where $\boldsymbol{\mu}_c = \frac{1}{n_c}\sum_{k=1}^{n_c} \boldsymbol{h}_k^{(c)}$ are the class means and $\boldsymbol{\mu}_G = \frac{1}{n}\sum_{c=1}^{C}\sum_{k=1}^{n_c} \boldsymbol{h}_k^{(c)}$ is the global mean.

NC2 predicts that the class means form a special structure, namely, their normalized covariance converges to the Simplex Equiangular Tight Frame (ETF). This can be characterize by

$$\mathsf{NC}_2 \stackrel{\text{def}}{=} \left\| \frac{MM^{\top}}{\|MM^{\top}\|_{\mathcal{F}}} - \frac{1}{\sqrt{C-1}}\left(I - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^{\top}\right)\right\|_{\mathcal{F}} \to 0 \quad (3)$$

during training, where $M \in \mathbb{R}^{C\times d}$ is the stack of centralized class-means, whose c-th row is $\boldsymbol{\mu}_c - \boldsymbol{\mu}_G$, and $\mathbb{1}_C \in \mathbb{R}^C$ is the all-one vector, and I is the identity matrix.

3.2. Experiment Setup

In our experiment, we explore the role of input distribution and labels through assigning coarser or finer labels to each sample, and then explore the structure of last-layer representation of a model trained on the dataset with coarse or fine labels and see to what extent the information of original labels are preserved.

The coarse labels are created in the following way. Choose a number $\tilde{C}$ divides C, and create coarse labels by

$$\tilde{y}_k = y_k \bmod \tilde{C}, \quad (4)$$

Figure 2. Illustration of Coarse CIFAR-10.

which merges the classes whose indices have the same remainder w.r.t. $\tilde{C}$ and thus creates $\tilde{C}$ super-classes. Since the original indices of classes generally have no special meanings, this process should act similarly to randomly merging classes.[1] We say the samples with the same coarse label belong to the same super-class, and call the dataset $\tilde{\mathcal{D}} = \{(x_k, \tilde{y}_k)\}_{k=1}^n$ the *coarse dataset*, which we use to train the model. Figure 2 provides an illustration of the coarse labels on CIFAR-10 with $\tilde{C} = 5$, which we call Coarse CIFAR-10.

To create fine labels, we randomly split each class into two sub-classes. Specifically, the fine labels are created by

$$\hat{y}_k = y_k + \beta C, \tag{5}$$

where $\hat{y}_k$ is the fine label of sample k and C is the number of original classes and β is a Bernoulli Variable. This process result in a dataset $\hat{\mathcal{D}} = \{(x_k, \hat{y}_k)\}_{k=1}^n$ with $2C$ classes. Same as before, we call $\hat{\mathcal{D}}$ the fine dataset.

4. Exploring the Fine-Grained Representation Structure with Coarse CIFAR-10

In this section, we experiment with coarsely labeled datasets, using Coarse CIFAR-10 as an illustrative example. Specifically, the model is trained on the training set of Coarse CIFAR-10 for a certain number of steps that is sufficient for the model to converge. We then take the last-layer representations of the model throughout training and explore their structure.

In order to make an exhaustive observation, the experiments are performed using different learning rates and weight-decay rates. Specifically, we choose the initial learning rate in $\{10^{-1}, 10^{-2}, 10^{-3}\}$ (we apply a standard learning rate decay schedule) and the weight-decay rate in $\{5\times10^{-3}, 5\times10^{-4}, 5\times10^{-5}\}$ and run all 9 possible combinations of them. The experiments are also conducted on multiple datasets and network architectures. Due to space limit, we defer complete results to Appendix (see Appendices C, E and F). In this section we focus on ResNet-18 on Coarse CIFAR-10, where the original number of classes is $C = 10$ and the number of coarse labels is $\tilde{C} = 5$. In this section, we only report results for one group of representative hyper-parameter combinations: learning rate is 0.1 and weight-decay rate is 5×10^{-4}. Results for other hyper-parameters are presented in Appendices C to H.

First, we verify that Neural Collapse does happen, i.e. the representations converge to 5 clusters, and the class means form a Simplex ETF structure. Specifically, we measure NC_1 and NC_2 defined in Section 3.1, with C replaced by $\tilde{C}$ since we are calculating it on the coarsely labeled dataset. The results are shown in Figure 3, which matches previous results in Papyan et al. (2020); Zhu et al. (2021), which verify Neural Collapse happens.

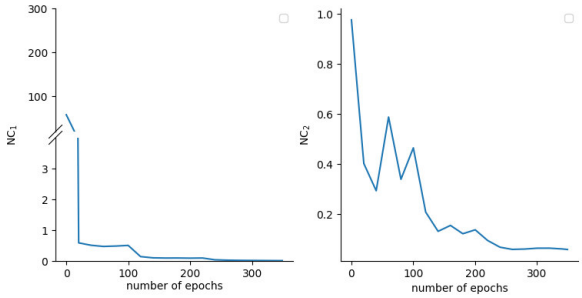


Figure 3. The value of NC_1 and NC_2 w.r.t. number of training epochs.

4.1. Class Distance

Now, we look at the average square Euclidian distance of last-layer representations between each two *original* classes. Formally, we calculate a class distance matrix $D \in \mathbb{R}^{C\times C}$, whose entries are

$$D_{i,j} = \frac{1}{n_i n_j} \sum_{u=1}^{n_i} \sum_{v=1}^{n_j} \left\| \boldsymbol{h}_u^{(i)} - \boldsymbol{h}_v^{(j)} \right\|_2^2, \tag{6}$$

for all $i, j \in [C]$, where $\boldsymbol{h}_k^{(c)}$ represents the last-layer representation of the k-th sample of super-class u.

Since the model is trained on the coarse dataset, Neural Collapse asserts that for every original class pair i, j in the same super-class (including the case of $i = j$), the class distance $D_{i,j}$ should be very small. In Coarse CIFAR-10,

[1] We adopt this deterministic process for simplicity and reproducibility. However, we do provide additional results with random merging in Appendix F.

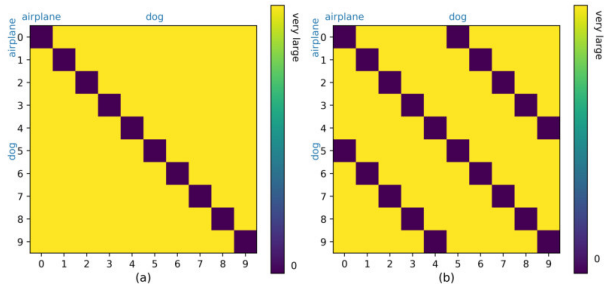


Figure 4. An illustration of predicted class distance matrix heatmaps of Coarse CIFAR-10, each row and column represents an original class. (a): If input distribution dominates the last-layer representations. (b): If Neural Collapse dominates the last-layer representations.

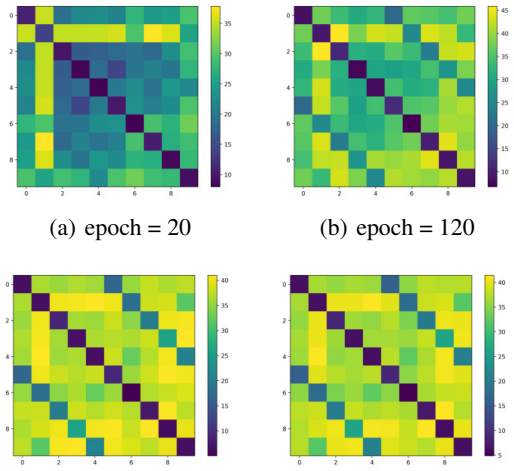


Figure 5. The heatmap of class distance matrix.

this will result in three darks lines (darker color represents lower value) in the heatmap of D since each super class contains two original classes, as illustrated in Figure 4 (b). For example, in Coarse CIFAR-10 the original class "airplane" and "dog" both belong to the super class "airplane & dog", therefore per Neural Collapse's prediction, their last-layer representations would collapse to each other, making the average square distance extremely small compared to other entries. In contrast, if the last-layer representations perfectly reflect the distribution of input, i.e. original classes, the class distance matrix should be a diagonal matrix as shown in Figure 4 (a), because the last-layer representation of samples in each original class only collapse to the class-mean of this original class.

Figure 5 displays the heatmaps of class distance matrix D at different stages in training, which shows that: (i) There are indeed three dark lines that show up, but they do not show up simultaneously. In particular, the central diagonal line – representing the samples in the same original classes – emerges earlier in training. (ii) Even in the final stage of training when the training error is zero, the three lines are not of the same degree of darkness. The central line is clearly darker, indicating a smaller distance within the original classes.

Those observations suggest that the actual behaviour of the last-layer representations is between the cases predicted in Figure 4 (a) and (b): the input distribution and training labels both have an impact on the distribution of the last-layer representations, both of them can be present even after reaching zero training error for a long time, and the impact of input distribution emerges earlier in training. These observations suggest both the existence of Neural Collapse and the inadequacy of Neural Collapse to completely describe the behaviour of last-layer representations.

4.2. Visualization

In this subsection, we take a closer look at the last-layer representations of the model at the end of training by reducing the dimensionality of the last-layer representations to 2 through t-SNE (Van der Maaten & Hinton, 2008) and visualize them. Specifically, we visualize each super-class separately, but color the samples whose original labels are different with different colors.

The visualization results are displayed in Figure 6, from which we observe a distinguishable difference between different original classes — their representations form well-separated clusters in the 2-dimensional space. This suggests that the input distribution information, i.e. the original label information, is well preserved in the last-layer representations.

Training extremely long. In order to explore if the fine-grained structures are still preserved even after a extremely long time of training, we further train the model with 1,000 epochs. The heatmap of the distance matrix is presented in Figure 7. We also produce the t-SNE visualizations, but only include the result of the first super-class in Figure 8 due to space limitation.

4.3. Learning CIFAR-10 from 5 Coarse Labels

As the results in Section 4.2 suggest, even after the training accuracy has reached 100% for a long time, the samples within each super-class still exhibit a clear structure per their original class, and those structures act as clusters after reducing dimensionality. Inspired by this observation, we perform a Cluster-and-Linear-Probe (CLP) test to quantify to what extent the original class information is preserved

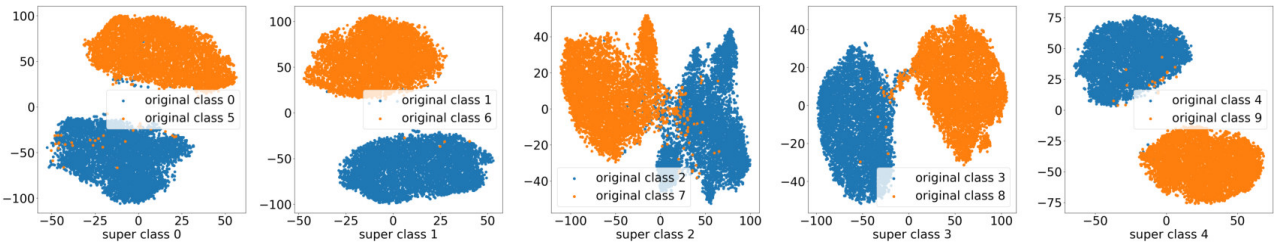


Figure 6. The t-SNE visualization of the last-layer representations of ResNet-18 trained on Coarse CIFAR-10. Each plot corresponds to a super-class.

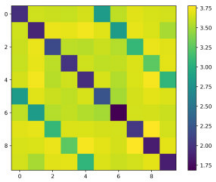


Figure 7. The heatmap of distance matrix of ResNet-18 trained on Coarse CIFAR-10 for 1,000 epochs.

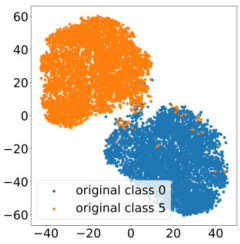


Figure 8. The t-SNE visualization of the last-layer representations of the first super-class of ResNet-18 trained on Coarse CIFAR-10 for 1,000 epochs.

in the last-layer representations. In CLP, we use the representations learned on Coarse CIFAR-10 to reconstruct 10 labels and run a linear probe on these representations using the reconstructed labels. Specifically, we first use t-SNE to reduce the dimensionality to 2 and then use k-means to find 2 clusters in the dimensionality-reduced representations within each super-class. We use the clusters as reconstructed labels to do a linear probe. In linear probe, we train a linear classifier on top of the previously learned representations on the training set with reconstructed labels and evaluate the learned linear classifier on the original test set. Notice that because we do not know the mapping of reconstructed classes to true original classes, we permute each possible mapping and report the highest performance. We also train a linear probe with original training labels as a comparison. The result is shown in Figure 9.

The performance of CLP on the original test set is comparable to linear probe trained on true original labels or even to models originally trained on CIFAR-10. Notice that the representation H is obtained from the model trained with coarse labels, and the label reconstruction only uses information of H and the number of original classes. This means we can achieve very high performance on the original test set even if we only have access to coarse labels. This result further confirms that the input distribution plays an important role in the last-layer representations.

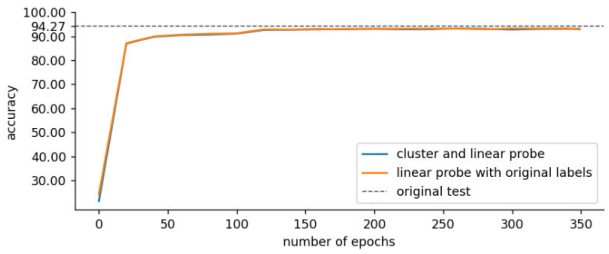


Figure 9. The CLP result. "original test" is the highest test set accuracy achieved by ResNet18 trained on original CIFAR-10 with the same training hyper-parameters.

5. How Does Semantic Similarity Affect the Fine-Grained Structure?

In our experiments with Coarse CIFAR-10, each coarse label is obtained by combining two classes regardless of the semantics. The fact that the neural network can separate the two classes in its representation space implies that the network recognizes these two classes as semantically different (even though they are given the same coarse label). In this section, we explore the following question: If the sub-classes in a super-class have semantic similarity, will the representations still exhibit a fine-grained structure to distinguish them? Intuitively, if the coarse label provided is "natural" and consists of semantically similar sub-classes, it is possible that the neural network will not distinguish between them and just produce truly collapsed representations.

We take a step towards answering this question by looking at ResNet-18 trained on CIFAR-100 using the official 20 super-classes (each super-class contains 5 sub-classes) as labels. Unlike randomly merging classes as we did in Section 4, the official super-classes of CIFAR-100 are natural, merging

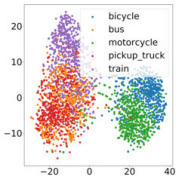


Figure 10. The t-SNE visualization of the last-layer representations of super-class "vehicles 1" of ResNet-18 trained on CIFAR-100 with original super-classes.

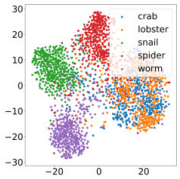


Figure 11. The t-SNE visualization of the last-layer representations of super-class "non-insect invertebrates" of ResNet-18 trained on CIFAR-100 with original super-classes.

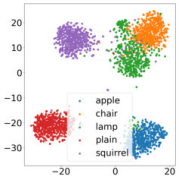


Figure 12. The t-SNE visualization of the last-layer representations of super-class 1 of ResNet-18 trained on CIFAR-100 with random super-classes.

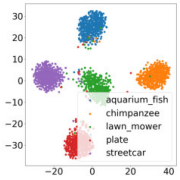


Figure 13. The t-SNE visualization of the last-layer representations of super-class 2 of ResNet-18 trained on CIFAR-100 with random super-classes.

classes with similar semantics (for example, "beaver" and "dolphin" both belong to "aquatic mammals"). This offers a perfect testbed for our question.

We find that ResNet-18 indeed is not able to distinguish all sub-classes in its representation space, but can still produce separable representations if some sub-classes within a super-class are sufficiently different. Interestingly, the notion of semantic similarity of ResNet-18 turns out to agree well with that of humans. Figures 10 and 11 show the t-SNE visualizations of representations from two super-classes. From the visualizations, although there are not as clear clusters as for Coarse CIFAR-10, the representations do exhibit visible separations between certain sub-classes. In Figure 10, "bicycles" and "motorcycles" are entangled together, while they are separated from "bus", "pickup truck", and "train", which is human-interpretable. In Figure 11, "crab" and "lobster" are mixed together, which are both aquatic and belong to malacostraca, while the other three are not and have more differentiative representations.

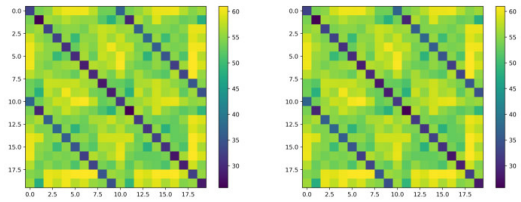
In comparison, when the CIFAR-100 classes are randomly merged into 20 super-classes, we find that the sub-class representations are much better separated (Figures 12 and 13). This is because randomly merged super-classes no longer have semantic similarity in their sub-classes.

These results confirm the intuition that the fine-grained structure in last-layer representations is affected by, or even based on, the semantic similarity between the inputs.

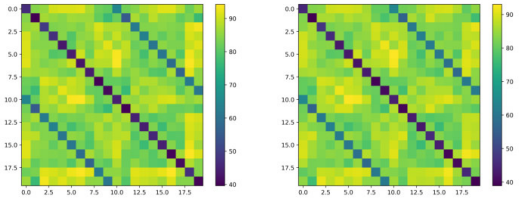
6. Fine-Grained Representation Structure on Fine CIFAR-10

In this section, we consider a finely-labeled dataset. We construct a fine version of CIFAR-10 with the process described in Section 3.2, and call it Fine CIFAR-10. Figure 14 presents the class distance matrices, arranged by the number of training epochs. As before, we only provide the results for a specific training hyper-parameter setting here and defer the full results to Appendix D.

It can be observed that at the early stage of training, there are three dark lines, which indicates the last-layer representations are converging towards 10 clusters instead of 20. At the end of training, this 10-class relationship is still preserved, although with a lighter color. Therefore, both the input distribution and the label information still have a strong influence on the representation structure when fine labels are used for training.



(a) epoch = 20 (b) epoch = 120



(c) epoch = 240 (d) epoch = 350

Figure 14. The heatmap of class distance matrix on Fine CIFAR-10.

7. Theoretical Result in a Synthetic Setting

In this section, we provide a theoretical result to show the fine-grained representation structure for a coarsely labeled dataset, supporting our empirical observations in previous sections. In particular, we consider a one-hidden-layer neural network trained by gradient descent on Gaussian mixture data. We describe our setting below.

Data generation. We consider input data generated from a mixture of 4 separated Gaussian distributions in $\mathbb{R}^d$. We denote the four clusters as $\mathscr{C}_1$, $\mathscr{C}_2$, $\mathscr{C}_3$, and $\mathscr{C}_4$. We give coarse label $+1$ to inputs from $\mathscr{C}_1$ and $\mathscr{C}_2$, and give coarse label -1 to inputs from $\mathscr{C}_3$ and $\mathscr{C}_4$. We denote the n samples as $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. We assume a sample $\boldsymbol{x}_i$ where $i \in \mathscr{C}_p$ ($p \in \{1,2,3,4\}$) is generated according to

$$\boldsymbol{x}_i = \boldsymbol{\mu}^{(p)} + \boldsymbol{\xi}_i, \tag{7}$$

where $\boldsymbol{\xi}_i \sim \mathcal{N}(0, \kappa^2 \boldsymbol{I})$. We assume that the means $\boldsymbol{\mu}^{(p)}$'s are pairwise orthogonal and $\|\boldsymbol{\mu}^{(p)}\|_2 = \tau$. For convenience, assume each cluster $\mathscr{C}_p$ has the same number of samples $|\mathscr{C}_p| = n/4$.

Neural network. We consider training a one-hidden-layer network with m hidden neurons. The first-layer weight matrix is $W = (\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_m) \in \mathbb{R}^{d\times m}$ and is trained by gradient descent. The second-layer weights are fixed to be all ones (which satisfies the ETF structure predicted by Neural Collapse in this setting). The output of the network is

$$f(\boldsymbol{x}, W) = \mathbf{1}^\top h(\boldsymbol{x}, W) = \sum_{r=1}^m \sigma\left(\boldsymbol{w}_r^\top \boldsymbol{x}\right),$$

where $h(\boldsymbol{x}, W) = \left(\sigma(\boldsymbol{w}_r^\top \boldsymbol{x})\right)_{r=1}^m \in \mathbb{R}^m$ is the hidden-layer representation, and

$$\sigma(z) = \begin{cases} \frac{1}{3}z^3 & \text{if } |z| \le 1 \\ z - \frac{2}{3} & \text{if } z \ge 1 \\ z + \frac{2}{3} & \text{if } z \le -1 \end{cases}$$

is the activation function. This activation is a smoothed version of symmetrized ReLU; it and its variants have been adopted in a line of theoretical work (e.g. Allen-Zhu & Li (2020); Zou et al. (2021); Shen et al. (2022)).

Loss function and training algorithm. We train the network by gradient descent on the logistic loss

$$L(W) = \frac{1}{n}\sum_{i=1}^n \ell(f(\boldsymbol{x}_i, W), y_i)$$

where $\ell(\hat{y}, y) = \frac{-y\hat{y}}{2}$ is the unhinged loss (van Rooyen et al., 2015). Notice that when $f(\boldsymbol{x}_i, W)$ is small, the unhinged loss can be viewed as an approximation of logistic loss (Shen et al., 2022). We initialize the first-layer weights i.i.d. from $\mathcal{N}(0, \omega^2)$ and update them using gradient descent with learning rate η.

Our main theorem shows that after training, the hidden-layer representations for $\mathscr{C}_1$ and $\mathscr{C}_2$ will form two separate clusters, even though they are given the same label. (Similar result holds for $\mathscr{C}_3$ and $\mathscr{C}_4$ by symmetry.) In particular, for any three samples $i_1, i_2 \in \mathscr{C}_1$ and $i_3 \in \mathscr{C}_2$, we show that

$$\|h(\boldsymbol{x}_{i_1}) - h(\boldsymbol{x}_{i_2})\|_2 \ll \|h(\boldsymbol{x}_{i_1}) - h(\boldsymbol{x}_{i_3})\|_2.$$

Theorem 7.1. *Consider the synthetic setting describe above and let* $\mathfrak{c} = \frac{8\kappa\sqrt{d}}{\tau}$. *Suppose that the following conditions hold regarding the Gaussian mean length* τ, *Gaussian variance* κ^2, *weight initialization variance* ω^2, *input dimension* d, *number of samples* n, *and number of neurons* m:

1. $n^{\frac{1}{2}} d^{-\frac{1}{4}} \ll \mathfrak{c} \ll n^{\frac{1}{2}} d^{-\frac{1}{6}}$;
2. $d^{1/3} \gg n$;
3. $d^{-\frac{1}{4}} n^{\frac{1}{2}} \mathfrak{c}^{-1} \ll \tau\omega \ll \log^{-\frac{1}{2}}(m)$.

For learning rate $\eta = O\left(\min\left\{\mathfrak{c}^3\tau^{-4}, \mathfrak{c}^2\omega\tau^{-3}, \mathfrak{c}^2\omega\tau\right\}\right)$ *and number of iterations* $T = \Theta(\frac{1}{\eta\omega\tau^3})$, *with high probability, the hidden-layer representation map* $h(\boldsymbol{x}) = h(\boldsymbol{x}, W(T))$ *satisfies that for all* $i_1, i_2 \in \mathscr{C}_1$ *and* $i_3 \in \mathscr{C}_2$, *we have*

$$\|h(\boldsymbol{x}_{i_1}) - h(\boldsymbol{x}_{i_2})\|_2 \ll \|h(\boldsymbol{x}_{i_1}) - h(\boldsymbol{x}_{i_3})\|_2.$$

An example set of parameters that satisfy the above conditions is:

$$\kappa = 1, \tau = d^{0.52}, \omega = d^{-0.53}, m = \log d, n = d^{0.32}.$$

The proof of Theorem 7.1 is given in Appendix A. The main step is to prove that after training, the neurons $\boldsymbol{w}_r$ will be better correlated with the class means $\boldsymbol{\mu}^{(p)}$ than with the individual sample noise $\boldsymbol{\xi}_i$. Therefore, the network will produce more similar representations for samples from the same cluster than for samples from different clusters.

Empirical verification. To verify our theoretical analysis and gain more understanding of the fine-grained structure of last-layer representation we provide some synthetic experiment results under the setting of classifying mixture of Gaussian using a 2-layer MLP in Appendix B, which is analogous (although not exactly the same) to the setting analyzed in our theory. Notice that this synthetic experiment is not only helpful to verify the theory, but also able to let us perform controlled experiments by varying different characteristics of the data distribution, architecture, and algorithmic components, to let us better understand how those hyper-parameters play a role in the final-layer representation, and serve as a starting point for a more thorough understanding of the last-layer representation behavior of neural networks.

8. Discussion

In this paper, we initiated the study of the role of the intrinsic structure of the input data distribution on the last-layer representations of neural networks, and in particular, how to reconcile it with the Neural Collapse phenomenon, which

is only driven by the explicit labels provided in the training procedure. Through a series of experiments, we provide concrete evidence that the representations can exhibit clear fine-grained structure despite their apparent collapse. While Neural Collapse is an intriguing phenomenon and deserves further studies to understand its cause and consequences, our work calls for more scientific investigations of the structure of neural representations that go beyond Neural Collapse.

We note that the fine-grained representation structure we observed depends on the inductive biases of the network architecture and the training algorithm, and may not appear universally. In our experiments on Coarse CIFAR-10, we observe the fine-grained structure for ResNet and DenseNet, but not for VGG (see Appendices G and H for extended results). We also note that certain choices of learning rate and weight-decay rate lead to stronger fine-grained structure than others. We leave a thorough investigation of such subtlety for future work.

Acknowledgements

WH would like to thank Ruoqi Shen for helpful discussions and contributions in the early stage of this project. JS acknowledges support from NSF DMS-2031899.

References

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Chung, F. and Lu, L. Concentration inequalities and martingale inequalities: a survey. *Internet mathematics*, 3(1): 79–127, 2006.

Ergen, T. and Pilanci, M. Revealing the structure of deep neural networks via convex duality. In *International Conference on Machine Learning*, pp. 3004–3014. PMLR, 2021.

Fang, C., He, H., Long, Q., and Su, W. J. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

Galanti, T., György, A., and Hutter, M. On the role of neural collapse in transfer learning. *arXiv preprint arXiv:2112.15121*, 2021.

Han, X., Papyan, V., and Donoho, D. L. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

Hui, L., Belkin, M., and Nakkiran, P. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022.

Ji, W., Lu, Y., Zhang, Y., Deng, Z., and Su, W. J. An unconstrained layer-peeled perspective on neural collapse. *arXiv preprint arXiv:2110.02796*, 2021.

Levy, O. and Goldberg, Y. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pp. 171–180, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-1618. URL https://aclanthology.org/W14-1618.

Lu, J. and Steinerberger, S. Neural collapse with cross-entropy loss. *arXiv preprint arXiv:2012.08465*, 2020.

McKenna, R. Tail bounds on the sum of half normal random variables. http://www.ryanhmckenna.com/2021/12/tail-bounds-on-sum-of-half-normal.html.

Mixon, D. G., Parshall, H., and Pi, J. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20(2):1–13, 2022.

Olah, C., Mordvintsev, A., and Schubert, L. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Shen, R., Bubeck, S., and Gunasekar, S. Data augmentation as feature manipulation. In *International Conference on Machine Learning*, pp. 19773–19808. PMLR, 2022.

Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Tirer, T. and Bruna, J. Extended unconstrained features model for exploring deep neural collapse. *arXiv preprint arXiv:2202.08087*, 2022.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

van Rooyen, B., Menon, A. K., and Williamson, R. C. Learning with symmetric label noise: The importance of being unhinged. In Cortes, C., Lawrence, N. D., Lee, D. D.,

Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 10–18, 2015.

Weinan, E. and Wojtowytsch, S. On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers. In *Mathematical and Scientific Machine Learning*, pp. 270–290. PMLR, 2022.

Wojtowytsch, S. et al. On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers. *arXiv preprint arXiv:2012.05420*, 2020.

Yaras, C., Wang, P., Zhu, Z., Balzano, L., and Qu, Q. Neural collapse with normalized features: A geometric analysis over the riemannian manifold. *arXiv preprint arXiv:2209.09211*, 2022.

Zhou, J., Li, X., Ding, T., You, C., Qu, Q., and Zhu, Z. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. *arXiv preprint arXiv:2203.01238*, 2022.

Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., and Qu, Q. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

Zou, D., Cao, Y., Li, Y., and Gu, Q. Understanding the generalization of adam in learning neural networks with proper regularization. *arXiv preprint arXiv:2108.11371*, 2021.

A. Proof of Theorem 7.1

Recall that we use unhinged loss $\ell(f(\boldsymbol{x}), y) = -\frac{yf(\boldsymbol{x})}{2}$ whose gradient is

$$\frac{\partial \ell\left(f(\boldsymbol{x}), y\right)}{\partial f(\boldsymbol{x})} = \frac{-y}{2}, \tag{8}$$

and our purpose is to prove

$$\|h(\boldsymbol{x}_{i_1}) - h(\boldsymbol{x}_{i_2})\| \ll \|h(\boldsymbol{x}_{i_1}) - h(\boldsymbol{x}_{i_3})\|. \tag{9}$$

A.1. Concentration Lemmas Used

We first introduce some concentration lemmas that we will use.

Lemma A.1 (McKenna). *Suppose $\{\psi_k\}_{k=1}^n$ are a set of independent Gaussian variables that $\psi_k \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \boldsymbol{I})$, then for any $t > 0$,*

$$\mathbb{P}\left\{\sum_{k=1}^n |\psi_k| \geq t\right\} \leq \exp\left(-\frac{t^2}{2\sum_{k=1}^n \sigma_k^2} + n\log 2\right). \tag{10}$$

In other words, with probability at least $\exp(-n)$, the following inequality holds:

$$\sum\nolimits_{k=1}^n |\psi_k| \geq 3\sqrt{n\sum_{i=1}^n \sigma_k^2}. \tag{11}$$

Lemma A.2 (Lemma 4 in (Shen et al., 2022)). *If $\boldsymbol{z}_1 \sim \mathcal{N}(0, \sigma_1^2 \boldsymbol{I})$ and $\boldsymbol{z}_2 \sim \mathcal{N}\left(0, \sigma_2^2 \boldsymbol{I}\right)$ are d-dimensional independent Gaussian vectors, then , we have*

$$\mathbb{P}\left\{\left|\boldsymbol{z}_1^\top \boldsymbol{z}_2\right| \geq 4\sigma_1\sigma_2\sqrt{d\log(2/\delta)}\right\} \leq \delta \tag{12}$$

Lemma A.3 (Corollary 3 in (Shen et al., 2022)). *If $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ is a d-dimensional Gaussian vector, then for large enough d, and $\delta > 2e^{-d/64}$ we have*

$$\frac{1}{2}\sigma^2 d \leq \|\boldsymbol{z}\|^2 \leq 2\sigma^2 d \tag{13}$$

Lemma A.4 (Union Bound). *If there are n variables $\{z_k\}_{k=1}^n$ (not necessarily independent), and each z_k satisfies*

$$\mathbb{P}\left\{z_k \geq \epsilon_k(\delta)\right\} \leq \delta, \tag{14}$$

then

$$\mathbb{P}\left\{\sum_{k=1}^n z_k \geq \sum_{k=1}^n \epsilon_k\left(\delta/n\right)\right\} \leq \delta \tag{15}$$

Proof. We have that

$$\begin{aligned}
\mathbb{P}\left\{\sum_{k=1}^n z_k \leq \sum_{k=1}^n \epsilon_k\left(\delta/n\right)\right\} &\geq \mathbb{P}\bigcap_{k=1}^n \left\{z_k \leq \epsilon_k(\delta/n)\right\} && (16)\\
&\geq \sum_{k=1}^n \mathbb{P}\left\{z_k \leq \epsilon_k(\delta/n)\right\} - n + 1 && (17)\\
&\geq n - n \times \frac{\delta}{n} - n + 1 && (18)\\
&= 1 - \delta. && (19)
\end{aligned}$$

□

Lemma A.5 (Lemma 5 in (Shen et al., 2022)). *If there are i.i.d samples $\{z_i\}_{i=1}^N$, where $z_i \sim \mathcal{N}(0, \sigma^2)$, then with probability at least $1-\delta$ we have*

$$\max_{i=1}^{N} |z_i| \leq \sigma \sqrt{2 \log \frac{2N}{\delta}}. \tag{20}$$

Lemma A.6 (Anti-Concentration). *If a random variable $z \sim \mathcal{N}(0, \sigma^2)$, then for any $\delta > 0$,*

$$\mathbb{P}\{-\delta < z < \delta\} < \frac{\delta}{\sigma}, \tag{21}$$

in other words, with probability higher than $1-\delta$, we have $|z| \geq \sigma\delta$.

Proof.

$$\begin{aligned}
\mathbb{P}\{-\delta < z < \delta\} &= \int_{-\delta}^{\delta} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x && (22)\\
&\leq \int_{-\delta}^{\delta} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{d}x && (23)\\
&= \sigma\delta \frac{2}{\sqrt{2\pi}} && (24)\\
&< \sigma\delta. && (25)
\end{aligned}$$

□

A.2. The Dynamics of Model Parameter

In this section, we will consider the dynamics of the projection of the model parameters on the "cluster mean" direction and the "noise" direction. Specifically, let $\boldsymbol{\mu}_k = \boldsymbol{\mu}^{(p)}$ for $k \in \mathscr{C}_p$, we will investigate these two quantities: $\zeta_k(t) = \boldsymbol{w}_r(t)^\top \boldsymbol{\mu}_k$ and $\phi_k(t) = \boldsymbol{w}_r(t)^\top \boldsymbol{\xi}_i$. Notice that ζ_k is only dependent on the cluster mean of the cluster which $\boldsymbol{x}_k$ belongs, so if $\boldsymbol{x}_k \in \mathscr{C}_s$, we denote ζ_k by $\zeta^{(s)}$. Notice that ζ_k and ϕ_k are actually depend on neuron index r. Throughout this section we fix a neuron index r to perform the analysis.

Notice that since we have assumed $f(\boldsymbol{x})$ is initialized very small, the activation function σ will fall in the interval $[-1, 1]$, and hence be a cubic function. Hereinafter unless explicitly mentioned, we will simply use $\sigma(z) = \frac{1}{3}z^3$ and therefore $\sigma'(x) = x^2$. This simplification will be rigorously justified in the Theorem 7.1.

Since we optimize the objective function by gradient descent with step size η. The update of $\boldsymbol{w}_r(t)$ in each step is $\Delta\boldsymbol{w}(t) = -\eta \frac{\partial L(W)}{\partial \boldsymbol{w}_r(t)}$. For other quantities (i.e. ζ, ϕ), we will use Δ to denote the update. For example $\Delta\zeta^{(s)}(t) = \Delta\boldsymbol{w}_r(t)^\top \boldsymbol{\mu}^{(s)}$.

The gradient of target function w.r.t. $\boldsymbol{w}_r(t)$ is:

$$\begin{aligned}
\frac{1}{\eta}\Delta\boldsymbol{w}_r(t) &= -\frac{\partial L(W)}{\partial \boldsymbol{W}_r(t)} && (26)\\
&= -\frac{1}{n}\sum_{i=1}^{n} \ell'(f(\boldsymbol{x}_i; \boldsymbol{w}_r(t)), y_i) \frac{\partial f(\boldsymbol{x}_i, \boldsymbol{w}_r(t))}{\partial \boldsymbol{w}_r(t)} && (27)\\
&= -\frac{1}{n}\sum_{i=1}^{n} \ell'(f(\boldsymbol{x}_i; \boldsymbol{w}_r(t)), y_i) \sigma'(\boldsymbol{x}_i^\top \boldsymbol{w}_r(t)) \boldsymbol{x}_i && (28)\\
&= \frac{1}{2n}\sum_{i=1}^{n} y_i \sigma'(\boldsymbol{x}_i^\top \boldsymbol{w}_r(t)) \boldsymbol{x}_i && (29)\\
&= \frac{1}{2n}\sum_{i=1}^{n} y_i \sigma'\left(\boldsymbol{\mu}_i^\top \boldsymbol{w}_r(t) + \boldsymbol{\xi}_i^\top \boldsymbol{w}_r(t)\right)(\boldsymbol{\mu}_i + \boldsymbol{\xi}_i) && (30)\\
&= \frac{1}{2n}\sum_{p=1}^{4} \sum_{\boldsymbol{x}_i \in \mathscr{C}_p} y_i \sigma'\left(\boldsymbol{w}_r(t)^\top \boldsymbol{\mu}^{(p)} + \boldsymbol{w}_r(t)^\top \boldsymbol{\xi}_i\right)(\boldsymbol{\mu}^{(p)} + \boldsymbol{\xi}_i). && (31)
\end{aligned}$$

Next, we consider the projection of $\boldsymbol{w}_r(t)$ onto the direction of $\boldsymbol{\mu}^{(s)}$ and $\boldsymbol{\xi}_k$ separately.

Lemma A.7. *For $s \in \{1,2\}$, we have the following inequality holds with probability at least $1-\exp(-n)$:*

$$\left|\frac{1}{\eta}\Delta\zeta^{(s)}(t) - \frac{\tau^2}{2n}\sum_{\boldsymbol{x}_i\in\mathscr{C}_s}\sigma'\left[\zeta^{(s)}(t)+\phi_i(t)\right]\right| \le \frac{3}{2}\mathcal{L}\tau\kappa. \tag{32}$$

Proof. Notice that since $s\in\{1,2\}$, we have $y_i = 1$ for $x_i \in \mathscr{C}_s$. We have that

$$\begin{aligned}
2n\times\frac{1}{\eta}\Delta\boldsymbol{w}_r(t)^\top\boldsymbol{\mu}^{(s)} =& \sum_{p=1}^{4}\sum_{\boldsymbol{x}_i\in\mathscr{C}_p} y_i\sigma'\left(\boldsymbol{w}_r(t)^\top\boldsymbol{\mu}^{(p)} + \boldsymbol{w}_r(t)^\top\boldsymbol{\xi}_i\right)\left(\boldsymbol{\mu}^{(p)\top}\boldsymbol{\mu}^{(s)} + \boldsymbol{\xi}_i^\top\boldsymbol{\mu}^{(s)}\right) && (33)\\
=& \sum_{\boldsymbol{x}_i\in\mathscr{C}_s}\sigma'\left(\boldsymbol{w}_r(t)^\top\boldsymbol{\mu}^{(s)} + \boldsymbol{\xi}_i^\top\boldsymbol{w}_r(t)\right)\tau^2 + \sum_{i=1}^{n} y_i\sigma'\left(\boldsymbol{w}_r(t)^\top\boldsymbol{\mu}^{(s)} + \boldsymbol{w}_r(t)^\top\boldsymbol{\xi}_i\right)\boldsymbol{\xi}_i^\top\boldsymbol{\mu}^{(s)} && (34)\\
\in& \sum_{\boldsymbol{x}_i\in\mathscr{C}_s}\sigma'\left(\boldsymbol{w}_r(t)^\top\boldsymbol{\mu}^{(s)} + \boldsymbol{\xi}_i^\top\boldsymbol{w}_r(t)\right)\tau^2 \pm \left(\mathcal{L}\sum_{i=1}^{n}\left|\boldsymbol{\xi}_i^\top\boldsymbol{\mu}^{(s)}\right|\right) && (35)\\
=& \sum_{\boldsymbol{x}_i\in\mathscr{C}_s}\sigma'\left(\boldsymbol{w}_r(t)^\top\boldsymbol{\mu}^{(s)} + \boldsymbol{\xi}_i^\top\boldsymbol{w}_r(t)\right)\tau^2 \pm \mathcal{L}\beta, && (36)
\end{aligned}$$

where $\beta = \sum_{i=1}^{n}|\boldsymbol{\xi}_i^\top\boldsymbol{\mu}^{(s)}|$. Since $\boldsymbol{\xi}_i^\top\boldsymbol{\mu}^{(s)} \sim \mathcal{N}(0,\kappa^2\tau^2)$, by Lemma A.1, we have that

$$\mathbb{P}\left\{\beta \ge 3n\kappa\tau\right\} \le \exp(-n). \tag{37}$$

Combining Equations (36) and (37), we have that with probability at least $1-\exp(-n)$,

$$2n\times\frac{1}{\eta}\Delta\boldsymbol{w}_r(t)^\top\boldsymbol{\mu}^{(s)} \in \tau^2\sum_{\boldsymbol{x}_i\in\mathscr{C}_s}\sigma'\left(\boldsymbol{w}_r(t)^\top\boldsymbol{\mu}^{(s)} + \boldsymbol{\xi}_i^\top\boldsymbol{w}_r(t)\right) \pm 3n\mathcal{L}\tau\kappa, \tag{38}$$

which proves the proposition. □

Lemma A.8. *For $k\in\mathscr{C}_1\cup\mathscr{C}_2$, if there exists a constant $c^{(n)}>0$ such that $n \le c^{(n)}\frac{\kappa d}{\tau}$, then we have*

$$\mathbb{P}\left\{\frac{\kappa^2 d}{4n}\sigma'(\phi_k(t)+\zeta_k(t)) - \frac{5}{2}\kappa\tau\mathcal{L} \le \frac{1}{\eta}\Delta\phi_k(t) \le \frac{3\kappa^2 d}{4n}\sigma'(\phi_k(t)+\zeta_k(t)) + \frac{5}{2}\kappa\tau\mathcal{L}\right\} \ge 1-\delta, \tag{39}$$

where $\delta = \exp(-n) + 2n\exp\left(-\frac{1}{\mathfrak{c}^2}\right) + 2\exp(-d/64)$.

Proof. For the noise term $\boldsymbol{\xi}_k$, notice that since $k\in\mathscr{C}_1\cup\mathscr{C}_2$, we have $y_k = 1$, and

$$\begin{aligned}
2n\times\frac{1}{\eta}\Delta\boldsymbol{w}_r(t)^\top\boldsymbol{\xi}_k =& \sigma'\left[\boldsymbol{\mu}_k^\top\boldsymbol{w}_r(t) + \boldsymbol{\xi}_k^\top\boldsymbol{w}_r(t)\right]\|\boldsymbol{\xi}_k\|^2 + \sigma'\left[\boldsymbol{\mu}_k^\top\boldsymbol{w}_r(t) + \boldsymbol{\xi}_k^\top\boldsymbol{w}_r(t)\right]\boldsymbol{\xi}_k^\top\boldsymbol{\mu}_k && (40)\\
&+ \sum_{p=1}^{4}\sum_{\substack{\boldsymbol{x}_i\in\mathscr{C}_p\\ i\neq k}} y_i\sigma'(\boldsymbol{w}_r(t)^\top\boldsymbol{\mu}^{(p)} + \boldsymbol{w}_r(t)^\top\boldsymbol{\xi}_i)\left(\boldsymbol{\xi}_k^\top\boldsymbol{\mu}^{(p)} + \boldsymbol{\xi}_k^\top\boldsymbol{\xi}_i\right) && (41)\\
\in& \sigma'\left[\boldsymbol{\mu}_k^\top\boldsymbol{w}_r(t) + \boldsymbol{\xi}_k^\top\boldsymbol{w}_r(t)\right]\|\boldsymbol{\xi}_k\|^2 \pm \left[\mathcal{L}\left|\boldsymbol{\xi}_k^\top\boldsymbol{\mu}_k\right| + \sum_{p=1}^{4}\sum_{\substack{\boldsymbol{x}_i\in\mathscr{C}_p\\ i\neq k}}\mathcal{L}\left|\boldsymbol{\xi}_k^\top\boldsymbol{\mu}^{(p)} + \boldsymbol{\xi}_k^\top\boldsymbol{\xi}_i\right|\right] && (42)\\
\subseteq& \sigma'\left[\boldsymbol{\mu}_k^\top\boldsymbol{w}_r(t) + \boldsymbol{\xi}_k^\top\boldsymbol{w}_r(t)\right]\|\boldsymbol{\xi}_k\|^2 \pm \left[\sum_{i=1}^{n}\mathcal{L}\left|\boldsymbol{\xi}_i^\top\boldsymbol{\mu}_i\right| + \sum_{\substack{1\le i\le n\\ i\neq k}}\mathcal{L}\left|\boldsymbol{\xi}_k^\top\boldsymbol{\xi}_i\right|\right] && (43)\\
=& \sigma'\left[\boldsymbol{\mu}_k^\top\boldsymbol{w}_r(t) + \boldsymbol{\xi}_k^\top\boldsymbol{w}_r(t)\right]\|\boldsymbol{\xi}_k\|^2 \pm \mathcal{L}\beta && (44)
\end{aligned}$$

where

$$\beta = \sum_{i=1}^{n} \left| \boldsymbol{\xi}_k^\top \boldsymbol{\mu}_i \right| + \sum_{\substack{1 \le i \le n \\ i \neq k}} \left| \boldsymbol{\xi}_k^\top \boldsymbol{\xi}_i \right|. \tag{45}$$

Notice that $n \le \frac{c^{(n)} \kappa d}{\tau}$. Let $\delta_1 = \exp(-n)$. From Lemma A.1, with probability at least $1 - \delta_1$, we have

$$\sum\nolimits_{i=1}^{n} \left| \boldsymbol{\xi}_k^\top \boldsymbol{\mu}_i \right| \le 3\tau\kappa n. \tag{46}$$

From Lemma A.2 and Lemma A.1, we have that with probability at least $1 - \delta_2$, we have

$$\sum_{i \neq k} \left| \boldsymbol{\xi}_k^\top \boldsymbol{\xi}_i \right| \le 2n\kappa^2 \sqrt{d \log(2n/\delta_2)}. \tag{47}$$

Take $\delta_2 = 2n \exp\left(-\frac{1}{c^2}\right)$, we get

$$\sum_{i \neq k} \left| \boldsymbol{\xi}_k^\top \boldsymbol{\xi}_i \right| \le 2n\kappa^2 \sqrt{d \log(2n/\delta_2)}. \tag{48}$$

$$\le 2n\kappa^2 \sqrt{d \times \frac{\tau^2}{\kappa^2 d}} \tag{49}$$

$$= 2n\kappa\tau \tag{50}$$

To summery, with probability at least $1 - \delta_1 - \delta_2$, we have

$$\frac{\mathcal{L}}{2n} \beta \le \frac{5}{2} \kappa\tau\mathcal{L}. \tag{51}$$

.

From Lemma A.3, with $\delta_3 > 2\exp(-d/64)$, we have

$$\frac{1}{2}\kappa^2 d \le \|\boldsymbol{\xi}_k\|^2 \le \frac{3}{2}\kappa^2 d. \tag{52}$$

Combining Equations (51) and (52), the proposition is proved.

□

To summarize, Lemma A.7 shows that with probability at least $1 - \delta_1$, where $\delta_1 = -\exp(-n)$, we have

$$\zeta^{(s)}(t+1) - \zeta^{(s)}(t) \in \frac{\tau^2\eta}{2n} \sum_{\boldsymbol{x}_i \in \mathscr{C}_s} \sigma'\left[\zeta^{(s)}(t) + \phi_i(t)\right] \pm \frac{\tau^2\eta}{2} \times \frac{3\mathcal{L}\kappa}{\tau} \tag{53}$$

and with probability at least $1 - \delta_2$, where $\delta_2 = \exp(-n) + 2n\exp\left(-c^{-2}\right) + 2\exp(-d/64)$, we have

$$\phi_k(t+1) - \phi_k(t) = \eta \frac{\mathrm{d}}{\mathrm{d}t} \phi_k(t) \tag{54}$$

$$\in \eta \left[\frac{\kappa^2 d}{4n} \sigma'(\phi_k(t) + \zeta_k(t)) - \frac{5}{2}\kappa\tau\mathcal{L}, \frac{3\kappa^2 d}{4n} \sigma'(\phi_k(t) + \zeta_k(t)) + \frac{5c^{(n)}\kappa^2 d\mathcal{L}}{n} \right] \tag{55}$$

$$\subseteq \eta \left[\frac{\kappa^2 d}{4n} \sigma'(|\phi_k(t)| + |\zeta_k(t)|) - \frac{5c^{(n)}\kappa^2 d\mathcal{L}}{n}, \frac{3\kappa^2 d}{4n} \sigma'(|\phi_k(t)| + |\zeta_k(t)|) + \frac{5}{2}\kappa\tau\mathcal{L} \right]. \tag{56}$$

Lemma A.9. *Let $s \in \{1, 2\}$. Suppose $\zeta^{(s)}$ is initialized by $\zeta^{(s)}(0)$, and there exists constants $c^{(t)} \in \left(0, \frac{8}{1+8\mathfrak{c}}\right)$ and $C^{(\phi)} > 0$ such that the following conditions hold for $t_0 < c^{(t)} \left(\tau^2\eta \left|\zeta^{(s)}(0)\right|\right)^{-1}$:*

1. $\forall t \le t_0$, *if* $\zeta^{(s)}(0) > 0$, *we have* $\left[\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8} - \mathfrak{c}\right)\eta\tau^2 t\right]^{-1} \le \zeta^{(s)}(t) \le \left[\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8} + \mathfrak{c}\right)\eta\tau^2 t\right]^{-1}$, *while if* $\zeta^{(s)}(0) < 0$, *we have* $\zeta^{(s)}(0) \le \zeta^{(s)}(t) \le \zeta^{(s)}(0) + \left(\frac{1}{8} + \mathfrak{c}\right)\eta\tau^2 t|\zeta^{(s)}(0)|^2$,
2. $\forall t \le t_0, \forall i \le n$, $|\phi_i(t)| \le \frac{C^{(\phi)}\mathfrak{c}^2}{n}\left|\zeta^{(s)}(t)\right| + \frac{\mathfrak{c}}{8}|\zeta^{(s)}(0)|$;
3. $\sqrt{\frac{32\mathcal{L}n}{\mathfrak{c}^2\sqrt{d}}} \le |\zeta^{(s)}(0)| \le \frac{\mathfrak{c}\left(8 - c^{(t)}\right)}{16\eta\tau^2}$;
4. $\mathfrak{c} \le \min\left\{\frac{nC^{(\phi)}}{4}, \left(\frac{n}{48}\right)^{\frac{1}{3}}, \frac{1}{8}\right\}$,

and $\zeta^{(s)}$ *is updated as described in Equation* (53)*, then we have:*

a) $\forall t \le t_0, \left|\zeta^{(s)}(t)\right| > \frac{1}{2}\left|\zeta^{(s)}(0)\right|$;

b) $\left(\frac{1}{8} - \frac{\mathfrak{c}}{2}\right)\eta\tau^2\zeta^{(s)}(t_0)^2 \le \zeta^{(s)}(t_0+1) - \zeta^{(s)}(t_0) \le \left(\frac{1}{8} + \frac{\mathfrak{c}}{2}\right)\eta\tau^2\zeta^{(s)}(t_0)^2$;

c) If $\zeta^{(s)}(0) > 0$, *we have*

$$\left[\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8} - \mathfrak{c}\right)\eta\tau^2(t_0+1)\right]^{-1} \le \zeta^{(s)}(t_0+1) \le \left[\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8} + \mathfrak{c}\right)\eta\tau^2(t_0+1)\right]^{-1}, \tag{57}$$

while if $\zeta^{(s)}(0) < 0$ *we have* $\zeta^{(s)}(0) \le \zeta^{(s)}(t_0+1) \le \zeta^{(s)}(0) + \eta\tau^2(t_0+1)|\zeta^{(s)}(0)|^2$.

Proof.

- First, notice that since $t < t_0 < c^{(t)}\left(\eta\tau^2|\zeta^{(s)}(0)|\right)^{-1}$, we have $\eta\tau^2 t|\zeta^{(s)}(0)|^2 \le c^{(t)}|\zeta^{(s)}(0)|$, and

$$\zeta^{(s)}(t) \le \zeta^{(s)}(0) + \left(\frac{1}{8} + \mathfrak{c}\right)\eta\tau^2 t|\zeta^{(s)}(0)|^2 \le \zeta^{(s)}(0) + c^{(t)}\left(\frac{1}{8} + \mathfrak{c}\right)|\zeta^{(s)}(0)| \le 0. \tag{58}$$

  If $\zeta^{(s)}(0) < 0$, then

$$|\zeta^{(s)}(t)| \ge |\zeta^{(s)}(0)| - \eta\tau^2 t|\zeta^{(s)}(0)|^2 \ge \left(1 - c^{(t)}\right)|\zeta^{(s)}(0)| \ge \frac{1}{2}|\zeta^{(s)}(0)|. \tag{59}$$

  While if $\zeta^{(s)}(0) > 0$, then

$$\left|\zeta^{(s)}(t)\right| \ge \left[\left|\zeta^{(s)}(0)\right|^{-1} - \left(\frac{1}{8} - \mathfrak{c}\right)\tau^2\eta t\right]^{-1} \ge |\zeta^{(s)}(0)|. \tag{60}$$

  Equation (59) and Equation (60) together proves Result a).

- Notice that from Condition 4 we have $\frac{\mathfrak{c}^2 C^{(\phi)}}{n} \le \frac{\mathfrak{c}}{4}$. From Condition 2, we have

$$\zeta^{(s)}(t) + \phi_i(t) \ge \left|\zeta^{(s)}(t)\right| - |\phi_i(t)| \ge \left(1 - \frac{\mathfrak{c}}{4}\right)\left|\zeta^{(s)}(t)\right| - \frac{\mathfrak{c}}{8}|\zeta^{(s)}(0)| \ge \left(1 - \frac{\mathfrak{c}}{2}\right)|\zeta^{(s)}(t)| \tag{61}$$

  and

$$\zeta^{(s)}(t) + \phi_i(t) \le \left|\zeta^{(s)}(t)\right| + |\phi_i(t)| \le \left(1 + \frac{\mathfrak{c}}{4}\right)\left|\zeta^{(s)}(t)\right| + \frac{\mathfrak{c}}{8}|\zeta^{(s)}(0)| \le \left(1 + \frac{\mathfrak{c}}{2}\right)|\zeta^{(s)}(t)|. \tag{62}$$

  Therefore we have

$$\frac{1}{n}\sum_{\boldsymbol{x}_i \in \mathscr{C}_s} \sigma'\left[\zeta^{(s)}(t) + \phi_i(t)\right] = \frac{1}{n}\sum_{\boldsymbol{x}_i \in \mathscr{C}_s}\left[\zeta^{(s)}(t) + \phi_i(t)\right]^2 \tag{63}$$

$$\in \frac{1}{n}\sum_{\boldsymbol{x}_i \in \mathscr{C}_s}\left(1 \pm \frac{\mathfrak{c}}{2}\right)^2\zeta^{(s)}(t)^2 \tag{64}$$

$$= \frac{1}{4}\left(1 \pm \frac{\mathfrak{c}}{2}\right)^2\zeta^{(s)}(t)^2. \tag{65}$$

$$\subseteq \frac{1}{4}(1 \pm 2\mathfrak{c})\zeta^{(s)}(t)^2. \tag{66}$$

From Condition 3, Condition 4 and Result a), we have

$$\frac{3\mathcal{L}\kappa}{\tau} \leq \frac{\mathfrak{c}}{8}\zeta^{(s)}(0)^2 \leq \frac{\mathfrak{c}}{2}\zeta^{(s)}(t)^2. \tag{67}$$

Subtracting Equation (66) and Equation (67) into the update rule Equation (53), we have

$$\zeta^{(s)}(t+1) - \zeta^{(s)}(t) \in \frac{\tau^2\eta}{2}\left(\frac{1}{4} \pm \frac{\mathfrak{c}}{2}\right)\zeta^{(s)}(t)^2 \pm \frac{\tau^2\eta\mathfrak{c}}{4}\zeta^{(s)}(t)^2 \tag{68}$$

$$= \tau^2\eta\zeta^{(s)}(t)^2 \times \left(\frac{1}{8} \pm \frac{\mathfrak{c}}{2}\right), \tag{69}$$

which proves Result b).

- For the greater-or-equal part of Result c), if $\zeta^{(s)}(0) < 0$, then we have $\zeta^{(s)}(t_0+1) \geq \zeta^{(s)}(t_0) \geq \zeta^{(s)}(0)$. In the following we assume $\zeta^{(s)}(0) > 0$. From Result b) we have

$$\zeta^{(s)}(t_0+1) \geq \zeta^{(s)}(t_0) + \left(\frac{1}{8} - \frac{\mathfrak{c}}{2}\right)\eta\tau^2\zeta^{(s)}(t_0)^2 \tag{70}$$

$$\geq \frac{1}{\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8} - \mathfrak{c}\right)\tau^2\eta t_0} + \frac{\left(\frac{1}{8} - \frac{\mathfrak{c}}{2}\right)\tau^2\eta}{\left(\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8} - \mathfrak{c}\right)\tau^2\eta t_0\right)^2} \tag{71}$$

$$= \frac{\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8} - \mathfrak{c}\right)\tau^2\eta t_0 + \left(\frac{1}{8} - \frac{\mathfrak{c}}{2}\right)\tau^2\eta}{\left(\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8} - \mathfrak{c}\right)\tau^2\eta t_0\right)^2} \tag{72}$$

$$\overset{\text{(i)}}{\geq} \frac{1}{\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8} - \mathfrak{c}\right)\tau^2\eta(t_0+1)}, \tag{73}$$

which proves the greater-or-equal part of Result c). To see (i), let $A = \zeta^{(s)}(0)^{-1} - \left(\frac{1}{8} - \mathfrak{c}\right)\tau^2\eta t_0$. Since $t_0 < c^{(t)}\left[\tau^2\eta|\zeta^{(s)}(0)|\right]^{-1}$ and $\zeta^{(s)}(0) \leq \frac{\mathfrak{c}\left(8-c^{(t)}\right)}{16\eta\tau^2}$, we have

$$A \geq \left[1 - c^{(t)}\left(\frac{1}{8} - \mathfrak{c}\right)\right]\zeta^{(s)}(0)^{-1} \geq \left(1 - \frac{c^{(t)}}{8}\right)\zeta^{(s)}(0)^{-1} \geq \frac{2}{\mathfrak{c}}\eta\tau^2, \tag{74}$$

and

$$\left[A + \left(\frac{1}{8} - \mathfrak{c}\right)\tau^2\eta + \frac{\mathfrak{c}}{2}\tau^2\eta\right]\left[A - \left(\frac{1}{8} - \mathfrak{c}\right)\tau^2\eta\right] - A^2 = \frac{\mathfrak{c}}{2}A\tau^2\eta - \left(\frac{1}{8} - \mathfrak{c}\right)^2\tau^4\eta^2 - \frac{\mathfrak{c}}{2}\tau^4\eta^2 \tag{75}$$

$$\geq \tau^4\eta^2\left[1 - \left(\left[\frac{1}{8} - \mathfrak{c}\right]^2 + \frac{\mathfrak{c}}{2}\right)\right] \tag{76}$$

$$\geq 0, \tag{77}$$

which proves the greater-or-equal part of Result c).

- For the less-or-equal part of Result c), if $\zeta^{(s)}(0) < 0$, simply notice that $|\zeta^{(s)}(t_0)| \leq |\zeta^{(s)}(0)|$, so

$$\zeta^{(s)}(t_0+1) \leq \zeta^{(s)}(0) + \left(\frac{1}{8} + \mathfrak{c}\right)\eta\tau^2 t_0\zeta^{(s)}(0)^2 + \left(\frac{1}{8} + \mathfrak{c}\right)\eta\tau^2\zeta^{(s)}(t_0)^2 \leq \zeta^{(s)}(0) + \left(\frac{1}{8} + \mathfrak{c}\right)\eta\tau^2(t_0+1)\zeta^{(s)}(0)^2, \tag{78}$$

which proves the result. If $\zeta^{(s)}(0) > 0$, then we have

$$\zeta^{(s)}(t_0+1) \le \zeta^{(s)}(t_0) + \left(\frac{1}{8}+\frac{\mathfrak{c}}{2}\right)\eta\tau^2\zeta^{(s)}(t_0)^2 \tag{79}$$

$$\le \frac{1}{\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8}+\mathfrak{c}\right)\tau^2\eta t_0} + \frac{\left(\frac{1}{8}+\frac{\mathfrak{c}}{2}\right)\tau^2\eta}{\left(\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8}+\mathfrak{c}\right)\tau^2\eta t_0\right)^2} \tag{80}$$

$$= \frac{\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8}+\mathfrak{c}\right)\tau^2\eta t_0 + \left(\frac{1}{8}+\frac{\mathfrak{c}}{2}\right)\tau^2\eta}{\left(\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8}+\mathfrak{c}\right)\tau^2\eta t_0\right)^2} \tag{81}$$

$$\le \frac{\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8}+\mathfrak{c}\right)\tau^2\eta(t_0-1)}{\left(\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8}+\mathfrak{c}\right)\tau^2\eta t_0\right)^2} \tag{82}$$

$$\overset{\text{(ii)}}{<} \frac{1}{\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8}+\mathfrak{c}\right)\tau^2\eta(t_0+1)}, \tag{83}$$

which proves the less-or-equal part of Result c). To see (ii), notice that

$$\left[\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8}+\mathfrak{c}\right)\tau^2\eta(t_0-1)\right]\left[\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8}+\mathfrak{c}\right)\tau^2\eta(t_0+1)\right] \tag{84}$$

$$= \left[\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8}+\mathfrak{c}\right)\tau^2\eta t_0\right]^2 - \left(\frac{1}{8}+\mathfrak{c}\right)^2\tau^4\eta^2 \tag{85}$$

$$< \left[\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8}+\mathfrak{c}\right)\tau^2\eta t_0\right]^2. \tag{86}$$

□

Next, we will consider the dynamics of $\boldsymbol{w}_r$ projected to the direction of $\boldsymbol{\xi}_i$.

Lemma A.10. *Let $s \in \{1,2\}$ and $k \in \mathscr{C}_s$. Suppose ζ_k is initialized as $\zeta_k(0)$ and $\phi_k(k)$ is initialized as $\phi_k(0)$, and there exists constants $c^{(t)} \in \left(0, \frac{8}{1+8\mathfrak{c}}\right)$ and $8 \le C^{(\phi)} \le n$ such that the following conditions hold for $t_0 \le c^{(t)}\left(\tau^2\eta|\zeta_k(0)|\right)^{-1}$:*

1. *$\forall t \le t_0$, if $\zeta_k(0) > 0$, we have $\left[\zeta_k(0)^{-1} - \left(\frac{1}{8}-\mathfrak{c}\right)\eta\tau^2 t\right]^{-1} \le \zeta_k(t) \le \left[\zeta_k(0)^{-1} - \left(\frac{1}{8}+\mathfrak{c}\right)\eta\tau^2 t\right]^{-1}$, while if $\zeta_k(0) < 0$, we have $\zeta_k(0) \le \zeta_k(t) \le \zeta_k(0) + \left(\frac{1}{8}+\mathfrak{c}\right)\eta\tau^2 t|\zeta_k(0)|^2$,*
2. *$t \le t_0$, $|\phi_k(t)| \le \frac{C^{(\phi)}\mathfrak{c}^2}{n}|\zeta_k(t)| + \frac{\mathfrak{c}}{8}|\zeta_k(0)|$, and $|\phi_k(0)| \le \frac{\mathfrak{c}}{8}|\zeta_k(0)|$;*
3. *$\sqrt{\frac{32\mathcal{L}n}{\mathfrak{c}^2\sqrt{d}}} \le |\zeta_k(0)| \le \frac{\mathfrak{c}\left(8-c^{(t)}\right)}{16\eta\tau^2}$;*
4. *$\mathfrak{c} \le \min\left\{\frac{1}{4} - \frac{2}{C^{(\phi)}}, \frac{nC^{(\phi)}}{4}, \left(\frac{n}{48}\right)^{\frac{1}{3}}, \frac{1}{8}\right\}$,*

and ϕ_k is updated through Equation (56), ζ_k is updated through Equation (53), then

$$|\phi_k(t_0+1)| \le \frac{C^{(\phi)}\mathfrak{c}^2}{n}|\zeta_k(t_0+1)| + \frac{\mathfrak{c}}{8}|\zeta_k(0)|. \tag{87}$$

Proof.

- From Lemma A.9, we have:

$$\forall t \le t_0, |\zeta_k(t)| > \frac{1}{2}|\zeta_k(0)| \tag{88}$$

and

$$\left(\frac{1}{8}-\frac{\mathfrak{c}}{2}\right)\eta\tau^2\zeta_k(t_0)^2 \le \zeta_k(t_0+1) - \zeta_k(t_0) \le \left(\frac{1}{8}+\frac{\mathfrak{c}}{2}\right)\eta\tau^2\zeta_k(t_0)^2. \tag{89}$$

- Let $s = \frac{5}{2}\kappa\tau\mathcal{L}$, from Condition 4 and the update rule Equation (56) we have

$$\phi_k(t+1) - \phi_k(t) \in \eta\left[\frac{\kappa^2 d}{4n}\sigma'(|\phi_k(t)| + |\zeta_k(t)|) - s, \frac{3\kappa^2 d}{4n}\sigma'(|\phi_k(t)| + |\zeta_k(t)|) + s\right]. \tag{90}$$

Given Condition 2 and Condition 4, we have

$$|\phi_k(t)| + |\zeta_k(t)| \leq \left(1 + \frac{C^{(\phi)}\mathfrak{c}^2}{n} + \frac{\mathfrak{c}}{4}\right)|\zeta_k(t)| \leq 2|\zeta_k(t)|. \tag{91}$$

Since $\sqrt{\frac{32\mathcal{L}n}{\mathfrak{c}^2\sqrt{d}}} \leq |\zeta_k(0)| \leq 2|\zeta_k(t)|$, we have

$$s = \frac{5}{2}\kappa\tau\mathcal{L} \leq \frac{5\kappa^2 d}{n}\zeta_k(0)^2 \leq \frac{20\kappa^2 d}{n}\zeta_k(t)^2. \tag{92}$$

Combining Equations (91) and (92)we have

$$\frac{3\kappa^2 d}{4n}\sigma'(\phi_k(t) + \zeta_k(t)) + s \leq \frac{3\kappa^2 d}{2n}\zeta(t)^2 + \frac{20\kappa^2 d}{n}\zeta(t)^2 \leq \frac{30\kappa^2 d}{n}\zeta(t)^2. \tag{93}$$

On the other hand, we have

$$\frac{\kappa^2 d}{4n}\sigma'(\phi_k(t) - \zeta_k(t)) - s \geq \frac{\kappa^2 d}{2n}\zeta(t)^2 - \frac{20\kappa^2 d}{n}\zeta(t)^2 \geq -\frac{30\kappa^2 d}{2n}\zeta(t)^2. \tag{94}$$

In summary we have

$$|\phi_k(t+1) - \phi_k(t)| \leq \frac{30\eta\kappa^2 d}{n}\zeta(t)^2. \tag{95}$$

- Notice that $C^{(\phi)} \geq \frac{1}{\frac{1}{8} - \frac{\mathfrak{c}}{2}}$. From Equation (89), we know that $\zeta_k(t_0)^2 \leq \frac{C^{(\phi)}}{\eta\tau^2}\left[\zeta(t_0+1) - \zeta(t_0)\right]$. If $\zeta_k(0) > 0$, we have

$$\begin{aligned}
|\phi_k(t+1)| &\leq |\phi_k(t)| + |\phi_k(t+1) - \phi_k(t)| && (96)\\
&\leq \frac{\mathfrak{c}}{8}\zeta_k(0) + \frac{C^{(\phi)}\kappa^2 d}{n\tau^2}\zeta_k(t) + \frac{30\eta\kappa^2 d}{n}\zeta_k(t)^2 && (97)\\
&\leq \frac{\mathfrak{c}}{8}\zeta_k(0) + \frac{C^{(\phi)}\kappa^2 d}{n\tau^2}\zeta_k(t) + \frac{30\eta\kappa^2 d}{n} \times \frac{C^{(\phi)}}{\eta\tau^2}\left[\zeta(t_0+1) - \zeta(t_0)\right] && (98)\\
&= \frac{\mathfrak{c}}{8}\zeta_k(0) + \frac{C^{(\phi)}\kappa^2 d}{n\tau^2}\zeta_k(t+1). && (99)\\
&\leq \frac{\mathfrak{c}}{8}\zeta_k(0) + \frac{C^{(\phi)}\mathfrak{c}^2}{n}\zeta_k(t+1). && (100)
\end{aligned}$$

While if $\zeta_k(0) < 0$, then we have $|\zeta_k(t)| \leq |\zeta_k(0)|$, and since $c^{(t)} \leq C^{(\phi)}$, we have

$$\begin{aligned}
|\phi_k(t+1)| &\leq |\phi_k(0)| + \sum_{i=1}^{t_0}\frac{\eta\kappa^2 d}{n}\zeta_k(t)^2 && (101)\\
&\leq \frac{\mathfrak{c}}{8}|\zeta_k(0)| + t_0 \times \frac{\eta\kappa^2 d}{n}\zeta_k(0)^2 && (102)\\
&\leq \frac{\mathfrak{c}}{8}|\zeta_k(0)| + \frac{c^{(t)}\kappa^2 d}{n\tau^2}|\zeta_k(0)| && (103)\\
&\leq \frac{\mathfrak{c}}{8}|\zeta_k(0)| + \frac{2c^{(t)}\kappa^2 d}{n\tau^2}|\zeta_k(t+1)| && (104)\\
&\leq \frac{\mathfrak{c}}{8}|\zeta_k(0)| + \frac{C^{(\phi)}\mathfrak{c}^2}{n}|\zeta_k(t+1)|. && (105)
\end{aligned}$$

□

Combining Lemma A.9 and Lemma A.10, we have the following conclusion:

Corollary A.11. *Let $s \in \{1,2\}$. Suppose $\zeta^{(s)}$ is initialized as $\zeta^{(s)}(0)$ and $\phi_k(k)$ is initialized as $\phi_k(0)$ for any $k \in \mathscr{C}_s$, and there exists constants $c^{(t)} \in \left(0, \frac{8}{1+8\mathfrak{c}}\right)$ and $8 \le C^{(\phi)} \le n$ such that the following conditions hold for $t_0 \le c^{(t)} \left(\tau^2\eta|\zeta^{(s)}(0)|\right)^{-1}$ and any $\boldsymbol{x}_k \in \mathscr{C}_s$:*

1. $\sqrt{\frac{32\mathcal{L}n}{\mathfrak{c}^2\sqrt{d}}} \le |\zeta^{(s)}(0)| \le \frac{\mathfrak{c}\left(8-c^{(t)}\right)}{16\eta\tau^2}$;
2. $|\phi_k(0)| \le \frac{\mathfrak{c}}{8}|\zeta^{(s)}(0)|$;
3. $\mathfrak{c} \le \min\left\{\frac{1}{4} - \frac{2}{C^{(\phi)}}, \frac{nC^{(\phi)}}{4}, \left(\frac{n}{48}\right)^{\frac{1}{3}}, \frac{1}{8}\right\}$;

and ϕ_k is updated through Equation (56), $\zeta^{(s)}$ is updated through Equation (53), then we have for all $t \le t_0$ and $\boldsymbol{x}_k \in \mathscr{C}_s$, we have the following conclusions:

a) if $\zeta^{(s)}(0) > 0$, then $\left[\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8} - \mathfrak{c}\right)\eta\tau^2 t\right]^{-1} \le \zeta^{(s)}(t) \le \left[\zeta^{(s)}(0)^{-1} - \left(\frac{1}{8} + \mathfrak{c}\right)\eta\tau^2 t\right]^{-1}$, while if $\zeta^{(s)}(0) < 0$, then $\zeta^{(s)}(0) \le \zeta^{(s)}(t) \le 0$;

b) $|\phi_k(t)| \le \frac{C^{(\phi)}\mathfrak{c}^2}{n}\left|\zeta^{(s)}(t)\right| + \frac{\mathfrak{c}}{8}|\zeta^{(s)}(0)|$.

A.3. The Analysis of Initialization

Next, we consider the initialization conditions in Corollary A.11. In this section, we will show that, either the initialization conditions in Corollary A.11 will be satisfied at some point, or ζ and ϕ will stay small all the time.

Notice that $\zeta_k(t)$ is initialized as $\zeta_k(0) = \boldsymbol{w}(0)^\top\boldsymbol{\mu}_k \sim \mathcal{N}(\mathbf{0}, \tau^2\omega^2)$ and $\phi_k(t)$ is initialized as $\phi_k(0) = \boldsymbol{w}(0)^\top\boldsymbol{\xi}_k$ where $\boldsymbol{w}(0) \sim \mathcal{N}(\mathbf{0}, \omega^2\boldsymbol{I})$ and $\boldsymbol{\xi}_k \sim \mathcal{N}\left(\mathbf{0}, \kappa^2\boldsymbol{I}\right)$. We have

Lemma A.12. *For $s \in \{1,2\}$, and any constant P we have $\forall k \in \mathscr{C}_s, \phi_k(0) \le P\mathfrak{c}\zeta^{(s)}(0)$ with probability higher than $1-\delta$, where $\delta = \frac{n}{2}\exp(-d/64) + \frac{n}{2}\exp(-2P\mathfrak{c}^{-1}) + \mathfrak{c}^{1/2}$.*

Proof. Let event $\mathcal{A} = \{\frac{1}{2}\kappa^2 d \le \|\boldsymbol{\xi}_k\|^2 \le 2\kappa^2 d\}$. From Lemma A.3, we have $\mathbb{P}(\mathcal{A}) \ge 1 - 2\exp(-d/64)$. Conditioned on $\mathcal{A}$, for any constant S we have $\mathbb{P}\{|\phi_k(0)| \le 2S\omega\kappa\sqrt{d}\} \ge 1 - 2\exp\left(-2S^2\right)$. From union bound we have $|\phi_k(0)| \le 2S\omega\kappa\sqrt{d}$ holds for all $k \in \mathscr{C}_s$ with probability at least $1 - \frac{n}{2}\exp(-d/64) - \frac{n}{2}\exp\left(-2S^2\right)$.

From Lemma A.6, for any constant T we have $\mathbb{P}\{|\zeta_k(0)| \ge T\omega\tau\} \le 1 - T$.

Let $S = \frac{P}{2}\mathfrak{c}^{-1/2}$ and $T = \mathfrak{c}^{1/2}$ we have with probability at least $1 - \frac{n}{2}\exp(-d/64) - \frac{n}{2}\exp(-2P\mathfrak{c}^{-1}) - \mathfrak{c}^{1/2}$,

$$\forall k \in \mathscr{C}_s, \frac{|\phi_k(0)|}{|\zeta^{(s)}(0)|} \le \frac{2S\omega\kappa\sqrt{d}}{T\omega\tau} = P\mathfrak{c}, \tag{106}$$

which proves the proposition.

□

Lemma A.13. *For any $s \in \{1,2\}$, if $|\zeta^{(s)}(0)| \ge \sqrt{\frac{18\mathcal{L}n}{\mathfrak{c}^2\sqrt{d}}}$, then the following conclusions hold with probability at least $1-\delta$, where $\delta = 2\exp\left(-\frac{\mathfrak{c}^2\left(1-\frac{c^{(t)}}{8}\right)^2}{8\eta^2\omega^2\tau^6}\right) + \frac{n}{2}\exp(-d/64) + \frac{n}{2}\exp(-\frac{1}{4\mathfrak{c}}) + \mathfrak{c}^{1/2}$:*

a) $|\zeta^{(s)}(0)| < \frac{\mathfrak{c}\left(8-c^{(t)}\right)}{16\eta\tau^2}$;

b) $\forall k \in \mathscr{C}_s, |\phi_k(0)| \le \frac{\mathfrak{c}}{8}|\zeta_k(0)|$.

Proof. From Chernoff inequality, we have $|\zeta^{(s)}(0)| \leq \frac{\mathfrak{c}\left(8-c^{(t)}\right)}{16\eta\tau^2}$ with probability at least $1-\delta_1$, where $\delta_1 = 2\exp\left(-\frac{\mathfrak{c}^2\left(1-\frac{c^{(t)}}{8}\right)^2}{8\eta^2\omega^2\tau^6}\right)$. From Lemma A.12, we have $\forall k \in \mathscr{C}_s, |\phi_k(0)| \leq \frac{\mathfrak{c}}{8} \times \sqrt{\frac{18\mathcal{L}n}{\mathfrak{c}^2\sqrt{d}}} \leq \frac{\mathfrak{c}}{8}|\zeta_k(0)|$ holds with probability at least $1-\delta_2$, where $\delta_2 = \frac{n}{2}\exp(-d/64) + \frac{n}{2}\exp(-\frac{1}{4\mathfrak{c}}) + \mathfrak{c}^{1/2}$. The proposition is proved through union bounding the probabilities of these two quantities. □

Lemma A.14. *Let $q = \frac{32\mathcal{L}n}{\mathfrak{c}^2\sqrt{d}}$. For $s \in \{1,2\}$, if $|\zeta^{(s)}(0)| \leq \sqrt{q}$, and there exists a constant $c^{(t)} \in \left(0, \frac{8}{1+8\mathfrak{c}}\right)$ such that $t_0 \leq c^{(t)}\left(\eta\tau^2\sqrt{q}\right)^{-1}$ we have $|\zeta^{(s)}(t_0-1)| \leq \sqrt{q}$, $|\zeta^{(s)}(t_0)| \geq \sqrt{q}$ and :*

1. $\mathfrak{c} \leq \frac{1}{8}$;
2. $\eta \leq \min\left\{1, \frac{\mathfrak{c}\left(8-c^{(t)}\right)}{16\tau^2\left(\sqrt{q}+\tau^2 q\right)}\right\}$,

then the following conclusions hold with probability at least $1-\delta$, where $\delta = \frac{n}{2}\exp(-d/64) + \frac{n}{2}\exp(-\frac{1}{4\mathfrak{c}}) + \mathfrak{c}^{1/2}$:

a) $|\zeta^{(s)}(t_0)| < \frac{\mathfrak{c}\left(8-c^{(t)}\right)}{16\eta\tau^2}$;

b) $\forall k \in \mathscr{C}_s, |\phi_i(t_0)| \leq \frac{\mathfrak{c}}{8}\sqrt{q}$.

Moreover, if $|\zeta^{(s)}(t)| < \sqrt{q}$ holds for all $t < c^{(t)}\left(\eta\tau^2\sqrt{q}\right)^{-1}$, then we also have $\forall i \in \mathscr{C}_s, \phi_i(t) \leq \frac{\mathfrak{c}}{8}\sqrt{q}$.

Proof. If there does not exists a t_0 such that $\zeta^{(s)}(t_0+1)| \geq \sqrt{q}$, then we set $t_0 = c^{(t)}\left(\eta\tau^2\sqrt{q}\right)^{-1}$ in the following. With this notation we have $|\zeta^{(s)}(t)| \leq \sqrt{q}$ for all $t < t_0$.

- Consider deduction on ϕ_k. For a specific $t < t_0$ if for all $t' < t$ we have $|\phi_k(t')| \leq \frac{\mathfrak{c}}{8} \times \sqrt{q}$, then we have

$$|\phi_k(t)| \leq |\phi_k(0)| + \sum_{t'=1}^{t-1}\left[\frac{3\eta\kappa^2 d}{4n}\sigma'\left(|\phi_k(t')| + |\zeta^{(s)}(t')|\right) + \frac{5}{2}\eta\kappa\tau\mathcal{L}\right] \tag{107}$$

$$\leq |\phi_k(0)| + \sum_{t'=1}^{t-1}\left[\frac{3\eta\kappa^2 d}{4n}\left[\left(1+\frac{\mathfrak{c}}{8}\right)\sqrt{q}\right]^2 + \frac{5\eta\kappa^2 dq}{n}\right] \tag{108}$$

$$\leq |\phi_k(0)| + t_0\eta\left[\frac{3\kappa^2 dq}{2n} + \frac{5\kappa^2 dq}{n}\right] \tag{109}$$

$$\leq |\phi_k(0)| + 7c^{(t)}\mathfrak{c}^4\sqrt{q} \tag{110}$$

$$\leq |\phi_k(0)| + \frac{7\mathfrak{c}}{64}\sqrt{q}. \tag{111}$$

  From Lemma A.12, we have $\forall k \in \mathscr{C}_s, \phi_k(0) \leq \frac{\mathfrak{c}}{64}\sqrt{q}$ holds with probability at least $1-\delta_1$, where $\delta_1 = \frac{n}{2}\exp(-d/64) + \frac{n}{2}\exp(-\frac{1}{32\mathfrak{c}}) + \mathfrak{c}^{1/2}$. Through union bound, we have all $\frac{n}{4}$ satisfies this condition with probability at least $1-\frac{n}{4}\delta_1$.

- Next, using Equation (53) and Condition 2, we have

$$|\zeta^{(s)}(t_0)| \leq |\zeta^{(s)}(t_0)-1| + \times\frac{\tau^2\eta}{2n} \times \frac{n}{4} \times \left(\sqrt{q}+\frac{\mathfrak{c}}{8}\sqrt{q}\right)^2 + \frac{3\kappa\tau q\eta}{2} \tag{112}$$

$$\leq \sqrt{q} + \frac{\eta\tau^2 q}{2} + \frac{3\mathfrak{c}\tau^2 q\eta}{2\sqrt{d}} \tag{113}$$

$$\leq \frac{\mathfrak{c}\left(8-c^{(t)}\right)}{16\eta\tau^2}. \tag{114}$$

Consider the union bound of the two inequalities, the proposition is proved. □

A.4. The Bound of the Representation Distance

In this section, we put all the results together and prove our main theorem. In this section we will consider all neurons. For $\zeta^{(s)}$ w.r.t. the r-th neuron, we denote it as $\zeta^{(s,r)}$, and for ϕ_k we denote it as $\phi_k^{(r)}$. Similar to before if $k \in \mathscr{C}_s$ we also write $\zeta^{(s,r)}$ as $\zeta_k^{(r)}$.

Theorem A.15. *Suppose the model and training process is described as before, and following conditions hold*

1. $n \leq d^{1/3}$;
2. $\sqrt{\frac{32\mathcal{L}n}{d^{\frac{1}{2}}}} \leq \mathfrak{c} \leq \min\left\{\sqrt{\frac{n}{d^{\frac{1}{3}}}}, \frac{1}{8}\right\}$
3. $\eta \leq \min\left\{1, \frac{\mathfrak{c}^2 \times \left(8\mathfrak{c}+4\omega\tau\sqrt{\log(4m)}\right)}{4\tau^2(\tau^2+1)}\right\}$;
4. $\frac{2}{\tau} \times \sqrt{\frac{32\mathcal{L}n}{\mathfrak{c}^2\sqrt{d}}} \leq \omega \leq \frac{1}{4\tau\sqrt{\log 4m}}$,

then for all $t < t_0 = \frac{1-4\omega\tau\sqrt{\log 4m}}{\frac{1}{8}+\mathfrak{c}} \times \left(2\eta\tau^3\omega\sqrt{\log(4m)}\right)^{-1}$, *we have*

$$\frac{\|h(\boldsymbol{x}_{i_1}) - h(\boldsymbol{x}_{i_3})\|}{\|h(\boldsymbol{x}_{i_1}) - h(\boldsymbol{x}_{i_2})\|} \geq \sqrt{\frac{1}{24\mathcal{L}'}}\left[\frac{1}{\frac{\mathfrak{c}}{2} + \frac{10}{\sqrt{8\mathcal{L}}} \times \frac{\mathfrak{c}^3\sqrt{d}}{n^{1.5}}}\right]^3 = O\left(\min\left\{\mathfrak{c}^{-3}, \mathfrak{c}^{-9}n^{\frac{9}{2}}d^{-\frac{3}{2}}\right\}\right) \tag{115}$$

With probability at least $1-\delta$, *where*

$$\delta = 4mt_0\exp\left(-\mathfrak{c}^{-2}\right) + mn\exp\left(-\mathfrak{c}^{-1}\right) + m\mathfrak{c}^{\frac{1}{2}} + 2m\exp\left[-\frac{\mathfrak{c}^2 \times \left(8\mathfrak{c}+4\omega\tau\sqrt{\log(4m)}\right)}{16\eta^2\omega^2\tau^6}\right] + 2m^{-1}. \tag{116}$$

Proof.

Let $q = \frac{32\mathcal{L}n}{\mathfrak{c}^2\sqrt{d}} \leq 1$ and $c^{(t)} = \frac{1-4\omega\tau\sqrt{\log 4m}}{\frac{1}{8}+\mathfrak{c}} \leq 8$, we have $\eta \leq \frac{\mathfrak{c}\left(8-c^{(t)}\right)}{16\tau^2\left(\sqrt{q}+\tau^2 q\right)}$ and $\omega\tau \geq 2\sqrt{q}$.

- First, notice that both Lemma A.7 and Lemma A.8 holds with high probability. Specifically, Lemma A.7 holds with probability at least $1-\exp(-n)$, and Lemma A.8 holds with probability at least $1-\delta_1'$, where $\delta_1' = \exp(-n) + 2n\exp\left(-\frac{1}{\mathfrak{c}^2}\right) + 2\exp(-d/64) \leq 3n\exp\left(-\frac{1}{\mathfrak{c}^2}\right)$. We have that Corollary A.11 holds with probability at least $1-mt_0\delta_1$, where $\delta_1 = 4n\exp(-n)$.

- From Lemma A.5, we have $\forall r \leq m, \left|\zeta^{(1,r)}(0)\right| \leq 2\omega\tau\sqrt{\log 4m}$ with probability at least $1-\delta_2$, where $\delta_2 = \frac{1}{m}$. Similarly the probability of $\forall r \leq m, \left|\zeta^{(2,r)}(0)\right| \leq 2\omega\tau\sqrt{\log 4m}$ is also at least $1-\delta_2$. In this case for any s, r we have $t_0 \leq c^{(t)}\left(\eta\tau^2\left|\zeta^{(s,r)}(0)\right|\right)^{-1}$.

  If in addition Corollary A.11 holds, then from Lemmas A.13 and A.14 and Corollary A.11, for any $s \in \{1,2\}$ and $i \in \mathscr{C}_s$ we have $|\phi_i(t)| \leq \frac{C^{(\phi)}\mathfrak{c}^2}{n}\left|\zeta^{(s)}(t)\right| + \frac{\mathfrak{c}}{8}\sqrt{q}$ and $\zeta^{(s,r)}(t) \leq \left[1-\left(\frac{1}{8}+\mathfrak{c}\right)c^{(t)}\right]^{-1}\left|\zeta^{(1,r)}(0)\right|$ with probability at

least $1-\delta_3$, where $\delta_3 = 2\exp\left(-\frac{\mathfrak{c}^2\left(1-\frac{c^{(t)}}{8}\right)^2}{8\eta^2\omega^2\tau^6}\right) + n\exp(-\frac{1}{4\mathfrak{c}}) + \mathfrak{c}^{1/2}$. In this case we have

$$\forall r \le m, \forall s \in \{1,2\}, \forall t \le t_0, \left|\zeta^{(s,r)}(t_0)\right| \le \left(\left|\zeta^{(s,r)}(0)\right|^{-1} - \left(\frac{1}{8}+\mathfrak{c}\right)^{-1}\eta\tau^2 t\right)^{-1} \tag{117}$$

$$\le \left[1 - c^{(t)}\left(\frac{1}{8}+\mathfrak{c}\right)\right]^{-1}\left|\zeta^{(s,r)}(0)\right| \tag{118}$$

$$\le \left[1 - c^{(t)}\left(\frac{1}{8}+\mathfrak{c}\right)\right]^{-1} \times 2\omega\tau\sqrt{\log 4m} \tag{119}$$

$$\le \frac{1}{2} \tag{120}$$

with probability at least $1-2\delta_2 - m\delta_3$.

With $|\zeta^{(s,r)}(t)| \le \frac{1}{2}$ and $|\phi_i(t)| \le |\zeta^{(s,r)}(t)| \le \frac{1}{2}$ where $i \in \mathscr{C}_s$, we have $|\boldsymbol{w}(t)^\top \boldsymbol{x}_i| \le 1$, which falls in the range where $\sigma(\boldsymbol{w}(t)^\top \boldsymbol{x}_i) = \frac{1}{3}\left[\boldsymbol{w}(t)^\top \boldsymbol{x}_i\right]^3$, which fulfills our assumption in Section A.2. In the following we will assume $\sigma(z) = \frac{1}{3}z^3$.

- If the conclusion of Lemmas A.13 and A.14 and Corollary A.11 holds, for any $t \le t_0$ we have

$$|h(\boldsymbol{x}_{i_1}) - h(\boldsymbol{x}_{i_2})|_r = \left|\sigma\left(\boldsymbol{w}_r(t)^\top\boldsymbol{\mu}^{(1)} + \boldsymbol{w}_r(t)^\top\boldsymbol{\xi}_{i_1}\right) - \sigma\left(\boldsymbol{w}_r(t)^\top\boldsymbol{\mu}^{(1)} + \boldsymbol{w}_r(t)^\top\boldsymbol{\xi}_{i_2}\right)\right| \tag{121}$$

$$\le \mathcal{L}'\sigma\left|\boldsymbol{w}_r(t)^\top\boldsymbol{\xi}_{i_1} - \boldsymbol{w}_r(t)^\top\boldsymbol{\xi}_{i_2}\right| \tag{122}$$

$$= \mathcal{L}'\sigma\left(\left|\phi_{i_1}^{(r)}(t)\right| + \left|\phi_{i_2}^{(r)}(t)\right|\right) \tag{123}$$

$$\le \mathcal{L}'\sigma\left(\frac{\mathfrak{c}}{4}\sqrt{q} + \frac{2C^{(\phi)}\mathfrak{c}^2}{n}\left|\zeta^{(1,r)}(t)\right|\right) \tag{124}$$

$$\le \mathcal{L}'\sigma\left(\frac{\mathfrak{c}}{4}\sqrt{q} + \frac{C^{(\phi)}\mathfrak{c}^2}{n}\right). \tag{125}$$

- Notice that if Corollary A.11 holds, then the sign of $\zeta^{(s,r)}(t)$, as well as $\zeta^{(s,r)}(t) - \phi_i^{(s)}(t)$, is determined by its initialization $\zeta^{(s,r)}(t)$. Since $\zeta^{(s,r)}(0) \sim \mathcal{N}(0, \omega^2\tau^2)$, we with probability at least 0.5 that $\zeta^{(2,r)}(0) \le 0$ and since $\omega \ge \frac{2\sqrt{q}}{\tau}$, from Lemma A.6 we have $\zeta^{(1,r)}(0) \ge \sqrt{q}$ with probability at least $\frac{1}{2}$. We denote $\mathcal{A}$ to be the neuron indices r who satisfies $\zeta^{(2,r)}(0) < 0$ and $\zeta^{(1,r)}(0) \ge \sqrt{q}$. For each $r \le m$, we have $\mathbb{P}\{r \in \mathcal{A}\} \ge \frac{1}{4}$. By calculating the concentration of Binomial distribution (Chung & Lu, 2006), we have $|\mathcal{A}| \ge \frac{m}{8}$ with probability at least $1-\delta_4$, where $\delta_4 = \exp(-m/32)$. If $r \in \mathcal{A}$, then we have

$$|h(\boldsymbol{x}_{i_1}) - h(\boldsymbol{x}_{i_3})|_r = \left|\sigma\left(\boldsymbol{w}_r(t)^\top\boldsymbol{\mu}^{(1)} + \boldsymbol{w}_r(t)^\top\boldsymbol{\xi}_{i_1}\right) - \sigma\left(\boldsymbol{w}_r(t)^\top\boldsymbol{\mu}^{(2)} + \boldsymbol{w}_r(t)^\top\boldsymbol{\xi}_{i_2}\right)\right| \tag{126}$$

$$\ge \sigma\left(\left|\zeta^{(1,r)}(t) + \phi_{i_1}^{(r)}(t)\right|\right). \tag{127}$$

$$\ge \sigma\left(\left|\zeta^{(1,r)}(t)\right| - \left|\phi_{i_1}^{(r)}(t)\right|\right). \tag{128}$$

$$\ge \sigma\left[\left(1 - \frac{C^{(\phi)}\mathfrak{c}^2}{n}\right)\left|\zeta^{(1,r)}(t)\right| - \frac{\mathfrak{c}}{8}|\zeta_k(0)|\right] \tag{129}$$

$$\ge \sigma\left(\frac{1}{2}\zeta_k(t)\right) \tag{130}$$

$$\ge \sigma\left(\frac{1}{2}\left[\zeta^{(1,r)}(0)^{-1} - \left(\frac{1}{8}-\mathfrak{c}\right)\eta\tau^2 t\right]^{-1}\right) \tag{131}$$

$$= \sigma\left(\frac{1}{2}\left[1 - c^{(t)}\left(\frac{1}{8}-\mathfrak{c}\right)\right]^{-1}\left|\zeta^{(1,r)}(0)\right|\right). \tag{132}$$

Notice that $\forall r \in \mathcal{A}, \left|\zeta^{(1,r)}(0)\right| \geq \sqrt{q}$, we have

$$\left\|h\left(\boldsymbol{x}_{i_1}\right)-h\left(\boldsymbol{x}_{i_3}\right)\right\|^2=\sum_{r=1}^m\left|h(\boldsymbol{x}_{i_1})-h\left(\boldsymbol{x}_{i_3}\right)\right|_r^2 \tag{133}$$

$$\geq \sum_{r\in\mathcal{A}}\left|h(\boldsymbol{x}_{i_1})-h\left(\boldsymbol{x}_{i_3}\right)\right|_r^2 \tag{134}$$

$$\geq \frac{m}{8}\times\sigma\left(\frac{1}{2}\left[1-c^{(t)}\left(\frac{1}{8}-\mathfrak{c}\right)\right]^{-1}\sqrt{q}\right)^2. \tag{135}$$

$$\geq \frac{m}{8}\times\sigma\left(\frac{1}{2}\sqrt{q}\right)^2. \tag{136}$$

- To put things together, notice that since $\sigma(z)=\frac{1}{3}z^3$, we have $\sigma(a)/\sigma(b)=3\sigma(a/b)$. Let $\delta=mt_0\delta_1+2\delta_2+m\delta_3+\delta_4$, with probability at least $1-\delta$ we have

$$\frac{\left\|h\left(\boldsymbol{x}_{i_1}\right)-h\left(\boldsymbol{x}_{i_3}\right)\right\|^2}{\left\|h\left(\boldsymbol{x}_{i_1}\right)-h\left(\boldsymbol{x}_{i_2}\right)\right\|^2}\geq\frac{3\times\frac{m}{8}}{m\mathcal{L}'}\times\sigma\left(\frac{\frac{1}{2}\sqrt{q}}{\frac{\mathfrak{c}}{4}\sqrt{q}+\frac{C^{(\phi)}\mathfrak{c}^2}{n}}\right)^2 \tag{137}$$

$$=\frac{3}{8\mathcal{L}'}\sigma\left(\frac{1}{\frac{\mathfrak{c}}{2}+\frac{2C^{(\phi)}\mathfrak{c}^2}{n\sqrt{q}}}\right)^2 \tag{138}$$

$$\geq\frac{3}{8\mathcal{L}'}\sigma\left(\frac{1}{\frac{\mathfrak{c}}{2}+\frac{2C^{(\phi)}\mathfrak{c}^2}{n\sqrt{\frac{32\mathcal{L}n}{\mathfrak{c}^2d^{1/2}}}}}\right)^2 \tag{139}$$

$$=\frac{3}{8\mathcal{L}'}\sigma\left(\frac{1}{\frac{\mathfrak{c}}{2}+\frac{C^{(\phi)}}{\sqrt{8\mathcal{L}}}\times\frac{\mathfrak{c}^3\sqrt{d}}{n^{1.5}}}\right)^2. \tag{140}$$

Notice that $\sigma(z)=\frac{1}{3}z^3$, we have

$$\frac{\left\|h\left(\boldsymbol{x}_{i_1}\right)-h\left(\boldsymbol{x}_{i_3}\right)\right\|}{\left\|h\left(\boldsymbol{x}_{i_1}\right)-h\left(\boldsymbol{x}_{i_2}\right)\right\|}\geq\sqrt{\frac{1}{24\mathcal{L}'}}\left[\frac{1}{\frac{\mathfrak{c}}{2}+\frac{C^{(\phi)}}{\sqrt{8\mathcal{L}}}\times\frac{\mathfrak{c}^3\sqrt{d}}{n^{1.5}}}\right]^3 \tag{141}$$

$$=O\left(\min\left\{\mathfrak{c}^{-3},\mathfrak{c}^{-9}n^{\frac{9}{2}}d^{-\frac{3}{2}}\right\}\right), \tag{142}$$

and simply taking $C^{(\phi)}=10$ proves the proposition.

□

B. Experiment on Mixture of Gaussian with Coarse Labels

In this section, we reproduce the experiments under the setting of classifying mixture of Gaussian using a 2-layer MLP.

Formally, we create a dataset from a mixture of Gaussian, where the input from the c-th cluster is generated from $\mathcal{N}(\boldsymbol{\mu}^{(c)},\boldsymbol{I})$, and each cluster mean $\boldsymbol{\mu}^{(c)}$ is drawn i.i.d. from $\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I})$. The class label of each datapoint is the index of the cluster it belongs to. The larger σ^2 is, the larger the separation between each two clusters is, and the more likely it is to observe a fine-grained representation structure when given coarse labels.

We perform the same coarsening process described in Section 3.2 (by combining two classes into one super-class) and train a 2-layer MLP on the coarsely labeled dataset. We measure the significance of the fine-grained structure using the ratio between {the average squared distance between representations in the same super-class but different sub-classes} and {the

average squared distance between representations in the same subclass}, which we call the Mean Squared Distance Ratio (MSDR). Mathematically, it is defined as

$$\text{MSDR} = \frac{\text{average}_{i \neq j \text{ in same super-class}}\{D_{i,j}\}}{\text{average}_i\{D_{i,i}\}},$$

where $D_{i,j}$ is defined in (6). A larger MSDR means that the fine-grained structure is more pronounced, while $\text{MSDR} \approx 1$ indicates no fine-grained structure.

By varying training and data-generating parameters, we investigate factors that impact the significance of the fine-grained structure. Specifically, we vary the input dimension, hidden dimension in the network, and weight decay rate, and plot how MSDR scales with σ^2. The results are shown in Figures 15 to 17. Note that in each figure we use two different scales in the x-axes to differentiate the cases of $\sigma^2 < 2$ and $\sigma^2 > 2$.

From the figures, we observe that both input and hidden dimensions exhibit a clear positive correlation with MSDR. On the other hand, the weight decay rate does not have an impact on MSDR in this setting.

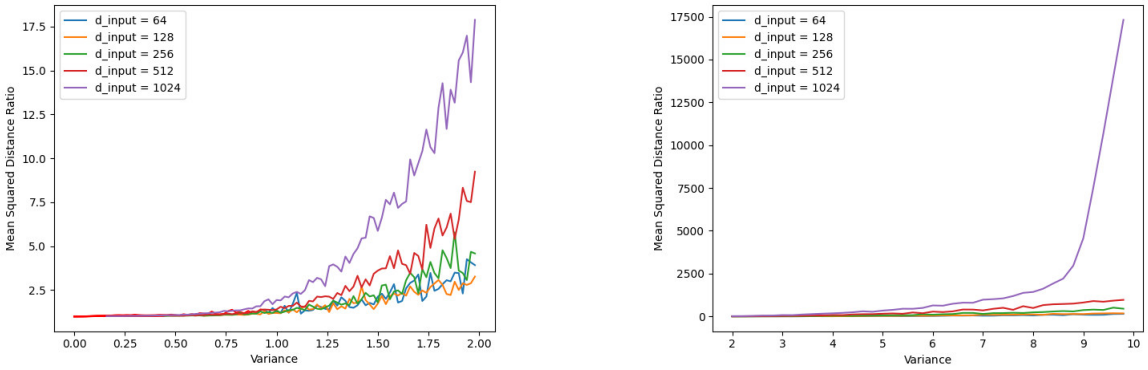


Figure 15. Mean Squared Distance Ratio vs. variance σ^2 for different input dimensions. Red lines on the left end are cases where the training accuracy does not reach 100%.

Training details. We generate data from 8 clusters, each having 500 samples. We train the model with gradient descent for 1,000 steps. The results are averaged over 10 runs. When varying one hyper-parameter, other hyper-parameters are set to their default values: $d_{\text{input}} = 512, d_{\text{hidden}} = 512$, weight decay $= 0$.

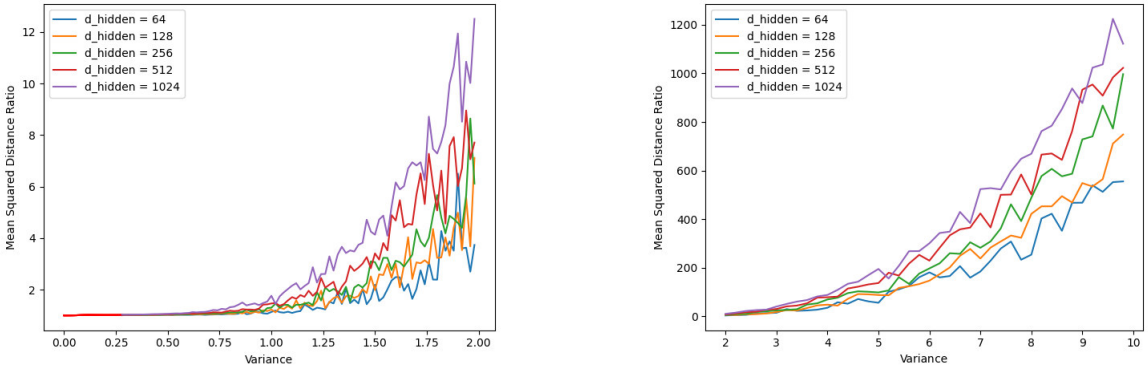


Figure 16. Mean Squared Distance Ratio vs. variance σ^2 for different hidden dimensions. Red lines on the left end are cases where the training accuracy does not reach 100%.

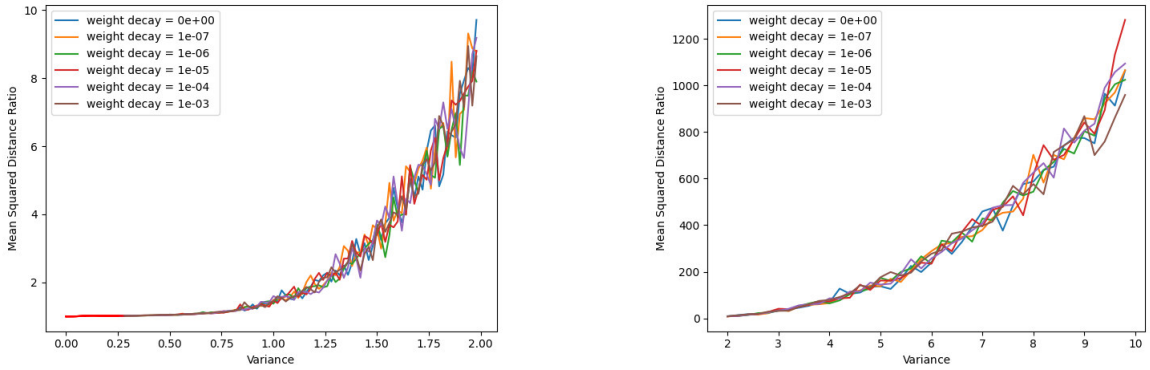


Figure 17. Mean Squared Distance Ratio vs. variance σ^2 for different weight decay rates. Red lines on the left end are cases where the training accuracy does not reach 100%.

B.1. Modeling Semantic Similarity

In Section 5, we showed that the emergence of fine-grained representations depends on the semantic similarity between sub-classes. Now we take a step in investigating this question by creating "similar" and "dissimilar" sub-classes in the Gaussian mixture model considered in this section.

In particular, we use the same data-generating process described above, except that half of the super-classes will be altered so that they consist of "similar" sub-classes, and we say that the other super-classes consist of "dissimilar" sub-classes. The way to generate similar sub-classes it to first sample $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ and then generate two means $\boldsymbol{\mu}^{(c)}, \boldsymbol{\mu}^{(c')} \sim \mathcal{N}(\boldsymbol{\mu}, \tau^2 \boldsymbol{I})$. Therefore, we can vary τ^2 to control the level of similarity between similar sub-classes.

We use default hyper-parameters described above, fix $\sigma^2 = 4$, and vary τ^2. Figure 18 shows the Mean Squared Distance Ratio for similar and dissimilar subclasses, respectively. We see that fine-grained structure within a super-class does require sufficient dissimilarity between its sub-classes, which agrees with our observation from Section 5 on CIFAR-100.

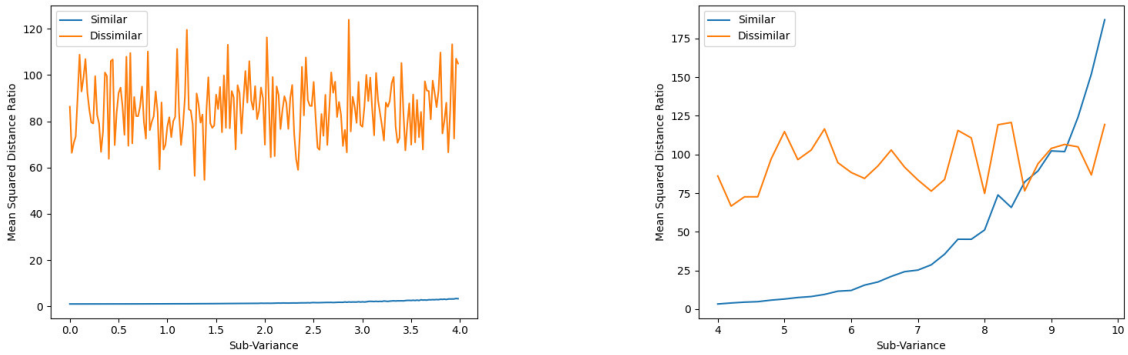


Figure 18. Mean Squared Distance Ratio vs. sub-variance τ^2. The setting is described in Appendix B.1.

C. Complete Coarse CIFAR-10 Experiment Results

In the following sections, we provide extended experiment results. As mentioned in Section 3.2, we permute learning rate in $\{10^{-1}, 10^{-2}, 10^{-3}\}$ and weight decay rate in $\{5 \times 10^{-3}, 5 \times 10^{-4}, 5 \times 10^{-5}\}$. Generally, the results will be shown in a 3×3 table, of which each grid represents the result of one hyper-parameter combination, with each row has the same learning rate and each column has the same weight decay rate.

In this section, we repeat the experiments in Section 4 with all learning rate and weight-decay rate combinations. Firstly, we present the training statistics (accuracy, loss) of all hyper-parameters in Figures 19 and 20 as an reference. It can be observed that all hyper-parameter groups achieved very low training error except the first one (weight decay = 5×10^{-3}, learning rate

$= 10^{-1}$). In fact, the last two hyper-parameter combinations (learning rate = 10^{-3}, weight decay $\in \{5 \times 10^{-4}, 5 \times 10^{-5}\}$) didn't achieve exactly 0 training error (their training error are $< 0.5\%$ but not exactly 0), and the other 6 hyper-parameter combinations all achieved exactly 0 training error.

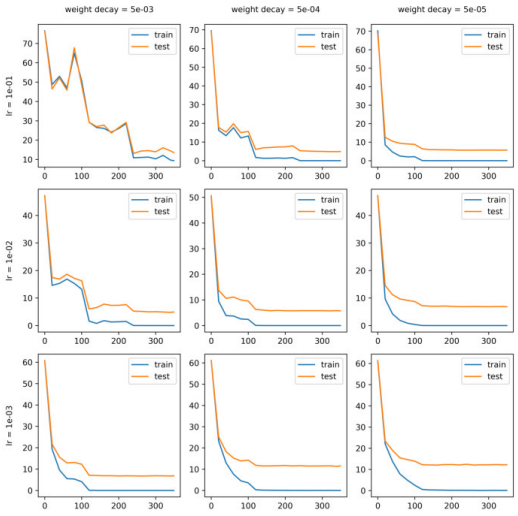


Figure 19. Training and test error during training.

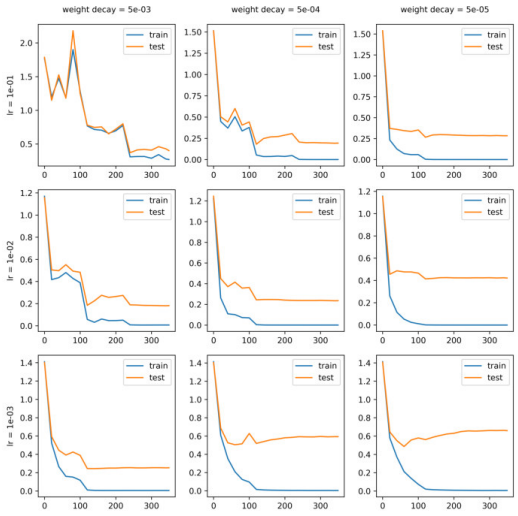


Figure 20. Training and test loss during training.

C.1. Class Distance

Here we present the visualization of the heatmap of class distance matrix D which is defined in Section 4.1. We choose 4 epochs to show the trend during training. The results are presented in Figures 21 to 24, whose epoch numbers are $20, 120, 240$ and 349 respectively.

C.2. Visualization

In this section, we present the t-SNE visualization result of ResNet-18 on Coarse CIFAR-10 in Figures 25 to 27. The results are divided into three groups, each of which has the same learning rate and the format of each group is the same as Figure 6.

C.3. Cluster-and-Linear-Probe

The Cluster-and-Linear-Probe test results of ResNet-18 trained on Coarse CIFAR-10 with all hyper-parameter combinations are presented in Figure 28.

D. Complete Class Distance Result of Fine CIFAR-10

In this section, we provide the visualization of the class distance matrix of Fine CIFAR-10 with all hyper-parameter combinations, which has been partially displayed in Section 6. As before, multiple epochs during training are selected to display a evolutionary trend of the class distance matrices. The results are presented in Figures 29 to 31, whose epoch numbers are $20, 200$ and 350 respectively.

E. Experiment of ResNet-18 on Coarse CIFAR-100

In this section, we report extended experiment results on Coarse CIFAR-100. The same with the case of Coarse CIFAR-10, we construct CIFAR-100 through the label coarsening process described in Section 3.2 and choose $\tilde{C} = 20$, so that every 5 original classes are merged into one super-class. We repeat most of experiments in Section 4.

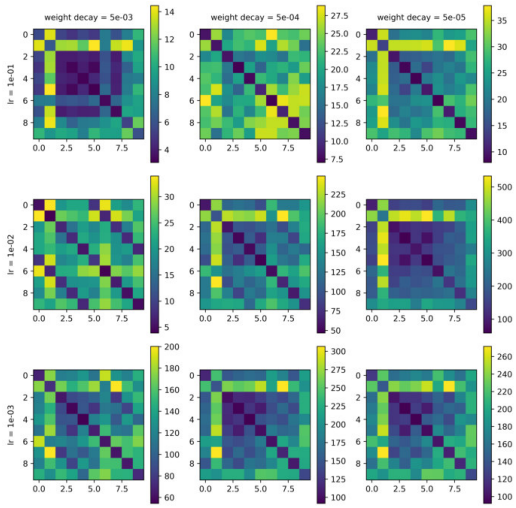


Figure 21. The heatmaps of class distance matrices of different hyper-parameter combinations at epoch 20.

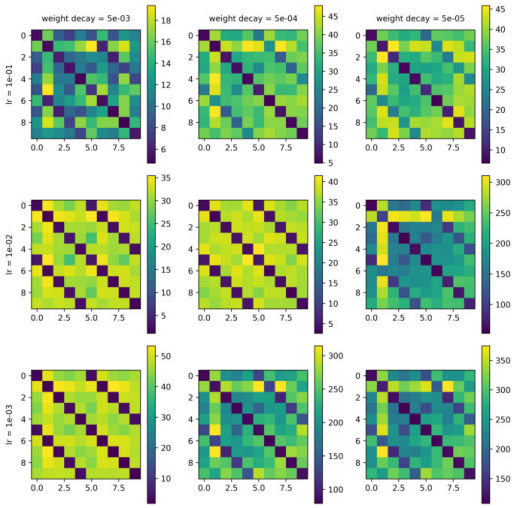


Figure 22. The heatmaps of class distance matrices of different hyper-parameter combinations at epoch 120.

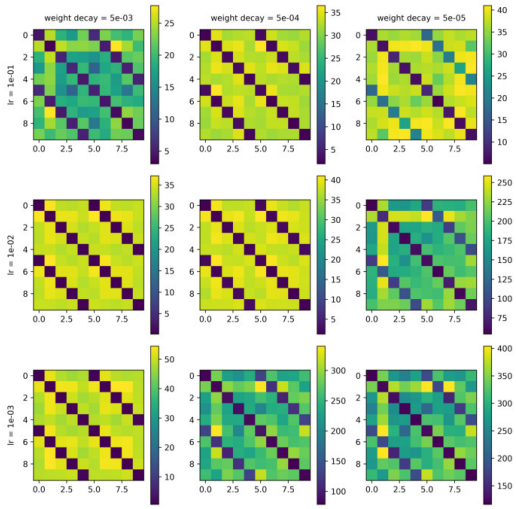


Figure 23. The heatmaps of class distance matrices of different hyper-parameter combinations at epoch 240.

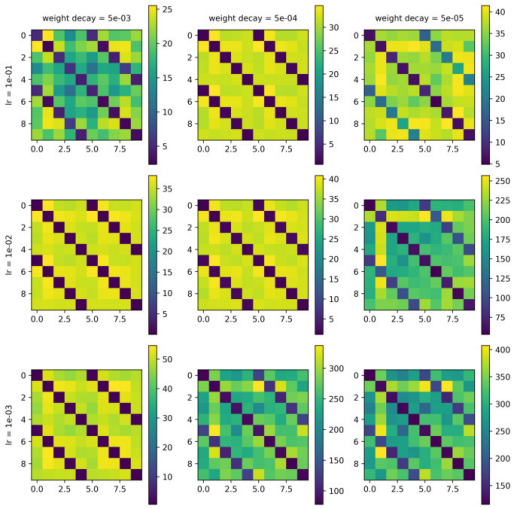


Figure 24. The heatmaps of class distance matrices of different hyper-parameter combinations at epoch 349.

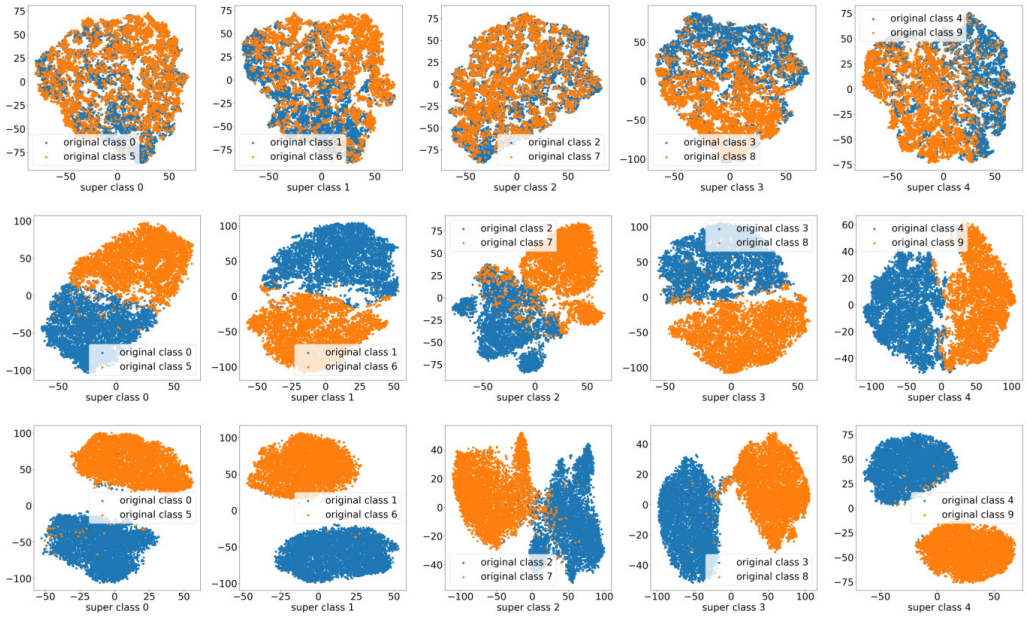


Figure 25. Visualization of last layer representations of ResNet-18 trained on Coarse CIFAR-10 with learning rate = 0.1. Each row represents a hyper-parameter combination and each column represents a super-class. The weight decay rates from top to bottom are $5 \times 10^{-3}, 5 \times 10^{-4}, 5 \times 10^{-5}$.

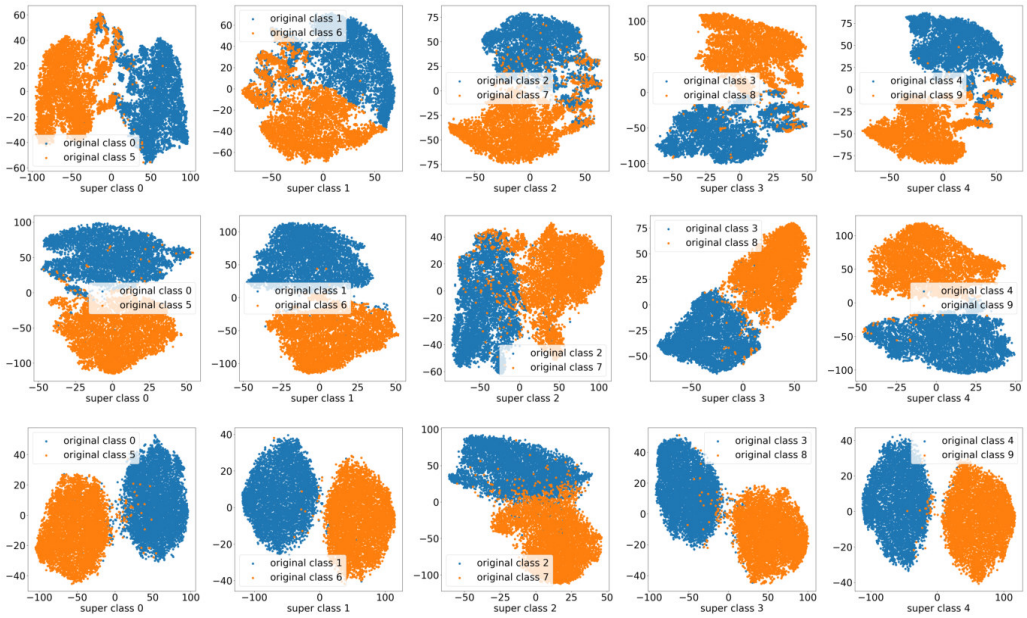


Figure 26. Visualization of last layer representations of ResNet-18 trained on Coarse CIFAR-10 with learning rate = 0.01. Each row represents a hyper-parameter combination and each column represents a super-class. The weight decay rates from top to bottom are $5 \times 10^{-3}, 5 \times 10^{-4}, 5 \times 10^{-5}$.

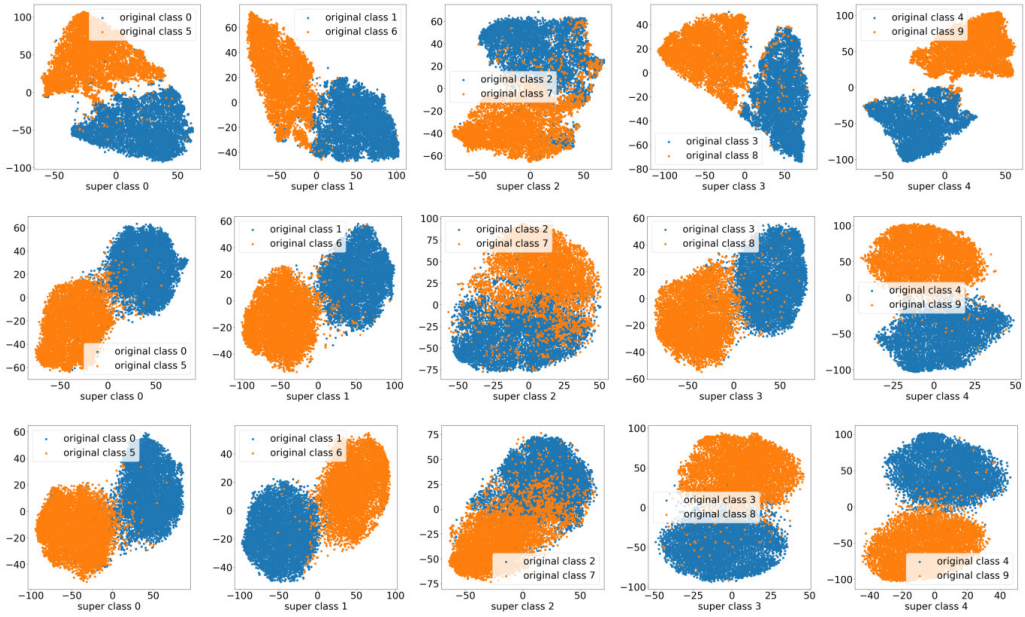


Figure 27. Visualization of last layer representations of ResNet-18 trained on Coarse CIFAR-10 with learning rate = 0.001. Each row represents a hyper-parameter combination and each column represents a super-class. The weight decay rates from top to bottom are $5 \times 10^{-3}, 5 \times 10^{-4}, 5 \times 10^{-5}$.

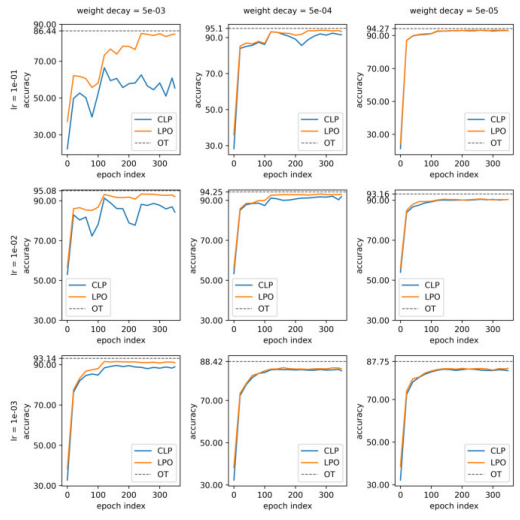


Figure 28. The result of Cluster-and-Linear-Probe test. In the figure"CLP" refers to Cluster-and-Linear-Probe, "LPO" refers to linear probe with original labels and "OT" refers to the test set accuracy of model trained on original CIFAR-10.

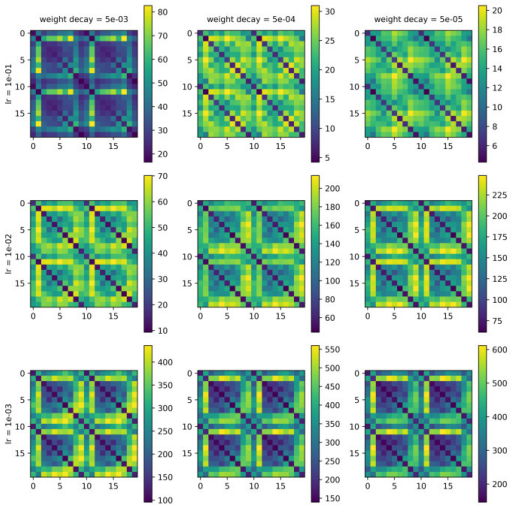


Figure 29. The heatmaps of class distance matrices of different hyper-parameter combinations on Fine CIFAR-10 at epoch 20.

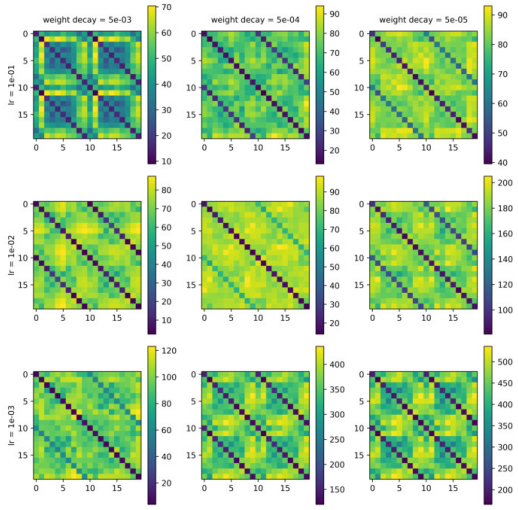


Figure 30. The heatmaps of class distance matrices of different hyper-parameter combinations on Fine CIFAR-10 at epoch 200.

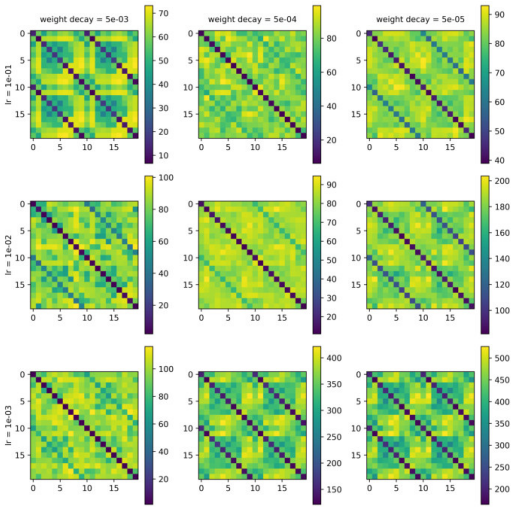


Figure 31. The heatmaps of class distance matrices of different hyper-parameter combinations on Fine CIFAR-10 at epoch 350.

E.1. Class Distance

The heatmaps of distance matrices are presented in Figures 32, 33 and 35, whose epoch number are $20, 200$ and 350 respectively.

E.2. Visualization

In this section, we present the t-SNE visualization result of ResNet-18 on Coarse CIFAR-100. We only put the result for learning rate = 0.01 and weight decay = 1e-4 here as a demonstration in Figure 34.

E.3. Cluster-and-Linear-Probe

Notice that we omit the visualization result of Coarse CIFAR-100 since there are too many figures. We present the Cluster-and-Linear-Probe results to reflect the clustering property of last-layer representations learned on Coarse CIFAR-100. The CLP results are presented in Figure 36.

F. Random Coarse CIFAR-10

In this section, as mentioned in Section 3.2, we make our experiment more complete by performing a random combination of labels on CIFAR-10 rather than using a determined coarsening process as in the main paper. The dataset construction is almost the same as the process of assigning coarse labels described in Section 3.2, except here we randomly shuffle the class indices before coarsening them.

The class distance matrices of three difference epochs are shown in Figures 37 to 39. From the results we can see, although there are no longer three dark lines, for each row there are generally two dark blocks, represents the original classes belongs to the same super-class, and the same observations in Section 4 can still be made here.

G. Experiment with DenseNet

We also perform our experiments with different neural network structures for completeness. In this section, we show the result with DenseNet-121 on Coarse CIFAR-10. The experiments with DenseNet is supportive to our observations in the main paper.

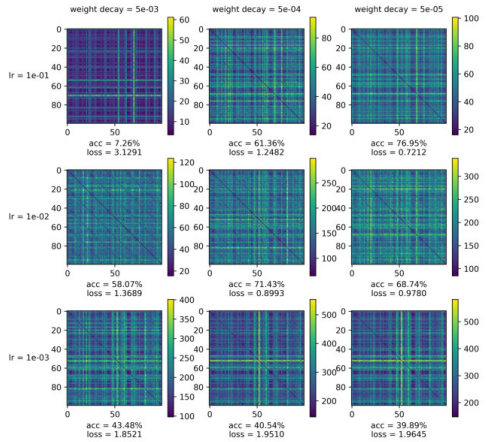


Figure 32. The heatmaps of class distance matrices of different hyper-parameter combinations on Coarse CIFAR-100 at epoch 20.

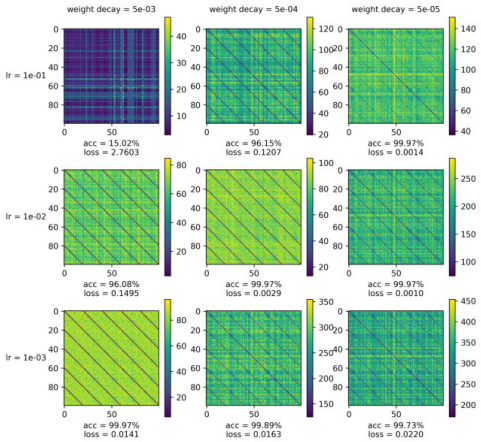


Figure 33. The heatmaps of class distance matrices of different hyper-parameter combinations on Coarse CIFAR-100 at epoch 200.

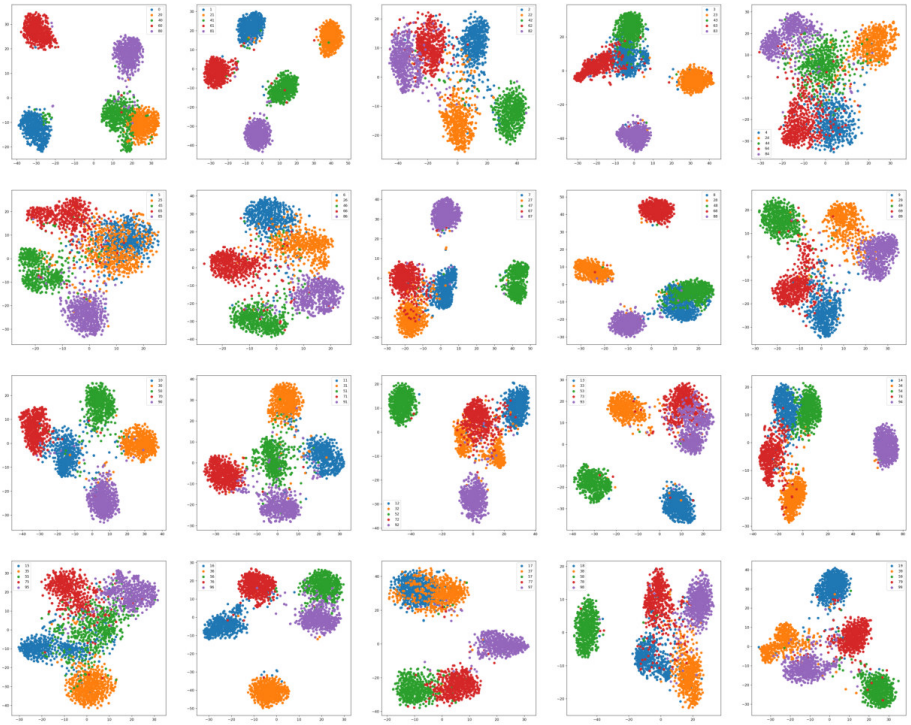


Figure 34. Visualization of last layer representations of ResNet-18 trained on Coarse CIFAR-100. Each grid represents a super-class and each color in a grid represents a sub-class.

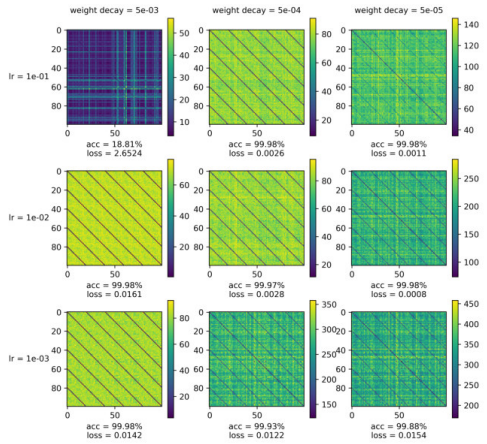


Figure 35. The heatmaps of class distance matrices of different hyper-parameter combinations on Coarse CIFAR-100 at epoch 350.

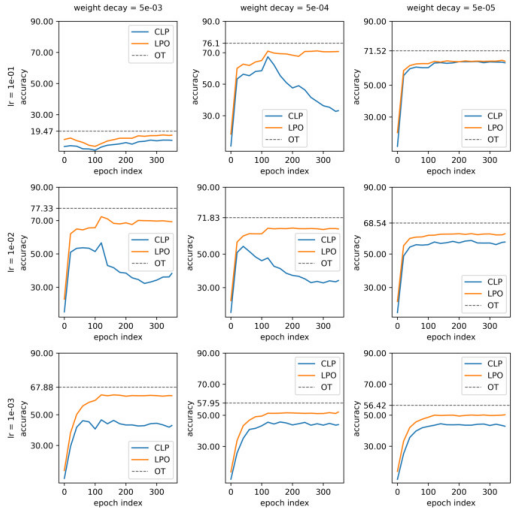


Figure 36. The result of Cluster-and-Linear-Probe test. In the figure"CLP" refers to Cluster-and-Linear-Probe, "LPO" refers to linear probe with original labels and "OT" refers to the test set accuracy of model trained on original CIFAR-100.

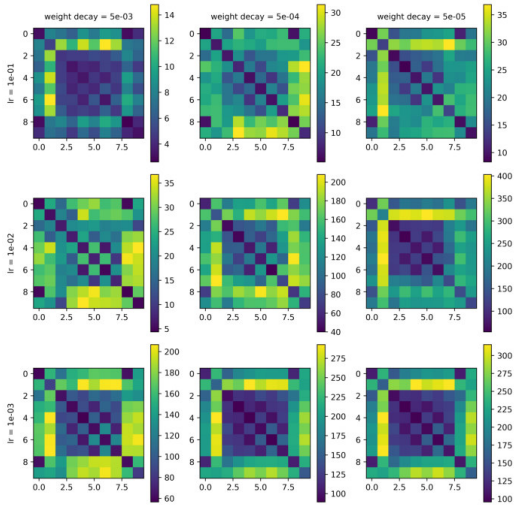


Figure 37. The heatmaps of class distance matrices of different hyper-parameter combinations on Random Coarse CIFAR-10 at epoch 20.

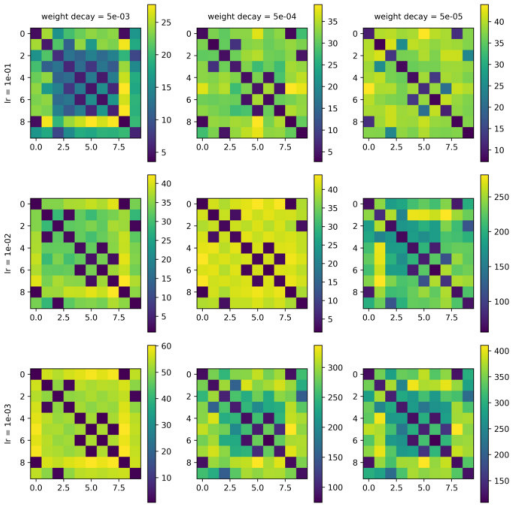


Figure 38. The heatmaps of class distance matrices of different hyper-parameter combinations on Random Coarse CIFAR-10 at epoch 200.

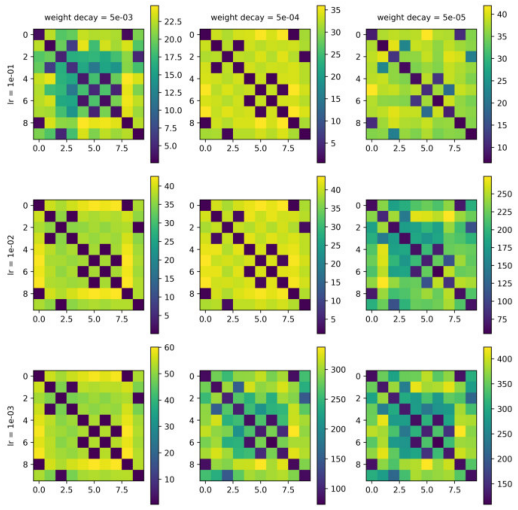


Figure 39. The heatmaps of class distance matrices of different hyper-parameter combinations on Random Coarse CIFAR-10 at epoch 350.

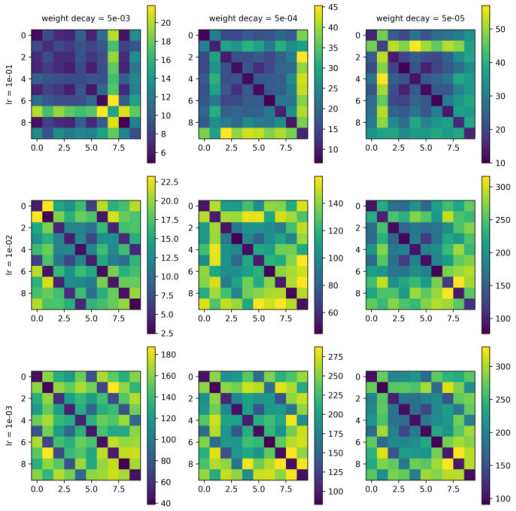


Figure 40. The heatmaps of class distance matrices of different hyper-parameter combinations with DenseNet-121 on Coarse CIFAR-10 at epoch 20.

G.1. Class Distance

The class distance matrices of three epochs during training are presented in Figures 40 to 42.

G.2. Cluster-and-Linear-Probe

The Cluster-and-Linear-Probe test results are presented in Figure 43.

H. Experiment with VGG

We also extend our experiments to VGG-18. Interestingly, VGG to some extent is a counter example of the observations made in the main paper: it only displays Neural Collapse, and can not distinguish different original classes within one super-class, even in an early stage of training. The reason why VGG is abnormal requires further exploration.

The class distance matrices of three epochs with VGG during training are shown in Figures 44 to 46. It can be observed that the three dark lines appears almost at the same time and always be of nearly the same darkness. This represents the trend predicted by Neural Collapse (Figure 4 (a)), but rejects the prediction made by (Figure 4 (b)).

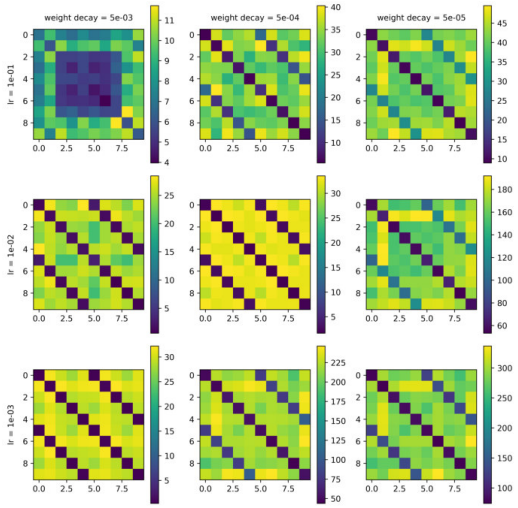


Figure 41. The heatmaps of class distance matrices of different hyper-parameter combinations with DenseNet-121 on Coarse CIFAR-10 at epoch 20.

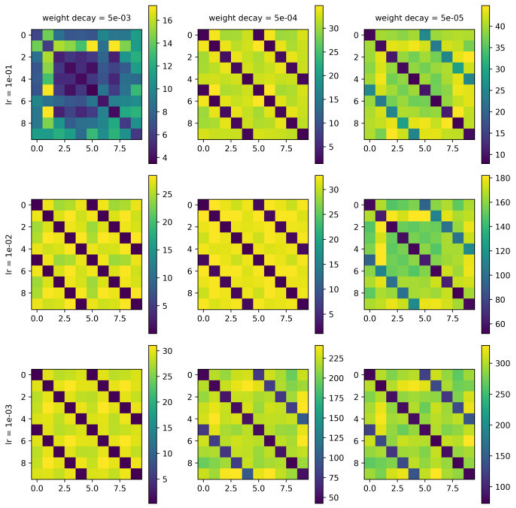


Figure 42. The heatmaps of class distance matrices of different hyper-parameter combinations with DenseNet-121 on Coarse CIFAR-10 at epoch 20.

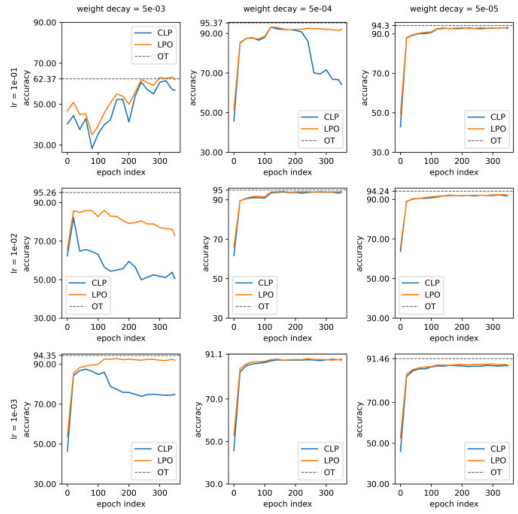


Figure 43. The result of Cluster-and-Linear-Probe test with DenseNet-121. In the figure"CLP" refers to Cluster-and-Linear-Probe, "LPO" refers to linear probe with original labels and "OT" refers to the test set accuracy of model trained on original CIFAR-10.

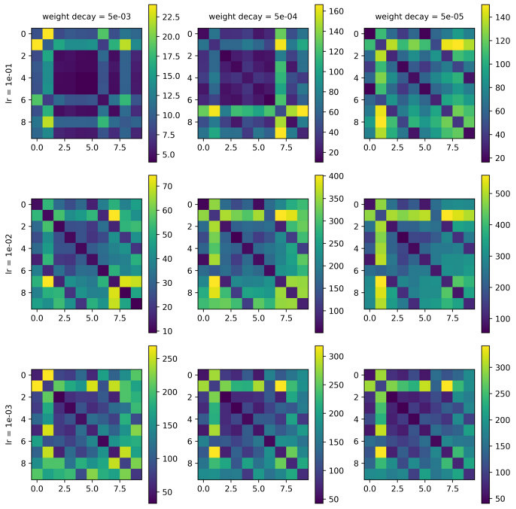


Figure 44. The heatmaps of class distance matrices of different hyper-parameter combinations with VGG-18 on Coarse CIFAR-10 at epoch 20.

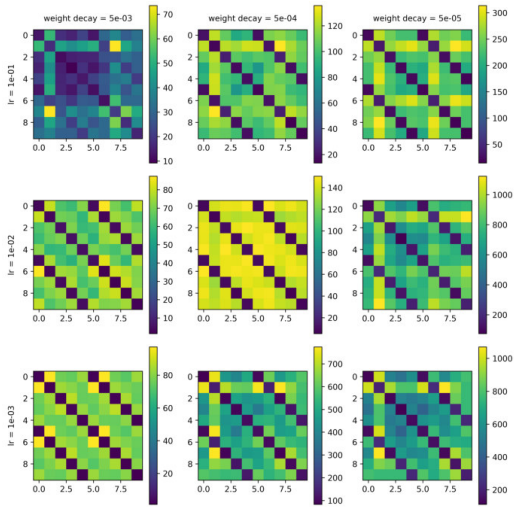


Figure 45. The heatmaps of class distance matrices of different hyper-parameter combinations with VGG-18 on Coarse CIFAR-10 at epoch 200.

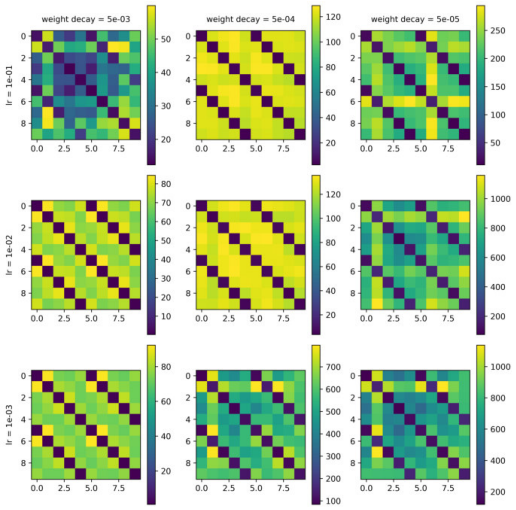


Figure 46. The heatmaps of class distance matrices of different hyper-parameter combinations with VGG-18 on Coarse CIFAR-10 at epoch 350.