The Benefits of Mixup for Feature Learning

Difan Zou [1] **Yuan Cao** [2] **Yuanzhi Li** [3] **Quanquan Gu** [4]

Abstract

Mixup, a simple data augmentation method that randomly mixes two data points via linear interpolation, has been extensively applied in various deep learning applications to gain better generalization. However, the theoretical underpinnings of its efficacy are not yet fully understood. In this paper, we aim to seek a fundamental understanding of the benefits of Mixup. We first show that Mixup using different linear interpolation parameters for features and labels can still achieve similar performance to the standard Mixup. This indicates that the intuitive linearity explanation in Zhang et al. (2018) may not fully explain the success of Mixup. Then we perform a theoretical study of Mixup from the feature learning perspective. We consider a feature-noise data model and show that Mixup training can effectively learn the rare features (appearing in a small fraction of data) from its mixture with the common features (appearing in a large fraction of data). In contrast, standard training can only learn the common features but fails to learn the rare features, thus suffering from bad generalization performance. Moreover, our theoretical analysis also shows that the benefits of Mixup for feature learning are mostly gained in the early training phase, based on which we propose to apply early stopping in Mixup. Experimental results verify our theoretical findings and demonstrate the effectiveness of the early-stopped Mixup training.

[1]Department of Computer Science & Institute of Data Science, The University of Hong Kong [2]Department of Statistics and Actuarial Science & Department of Mathematics, The University of Hong Kong [3]Machine Learning Department, Carnegie Mellon University [4]Department of Computer Science, University of California, Los Angeles. Correspondence to: Difan Zou <dzou@cs.hku.hk>.

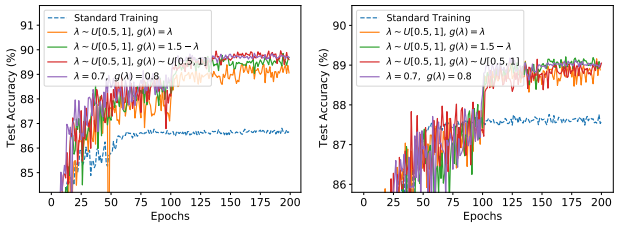
Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

1 Introduction

The Mixup method (Zhang et al., 2018) is a popular data augmentation technique in deep learning, known to yield notable improvements in generalization and robustness across multiple domains, such as image recognition (Berthelot et al., 2019), natural language processing (Guo et al., 2019b; Chen et al., 2020a), and graph learning (Han et al., 2022a). Unlike traditional data augmentation approaches that require domain knowledge of the dataset (e.g., random rotation and cropping for image data, and randomly modifying edges for graph data), Mixup relies on convex combinations of both features and labels from a pair of randomly selected training data points. As a result, this technique does not require any specialized knowledge or expertise to be performed.

Despite the remarkable empirical success of Mixup, there is a considerable gap in the theoretical understanding of this technique. In the original work of Mixup (Zhang et al., 2018), it has been argued that the efficacy of Mixup can be attributed to its inductive bias, which encourages the trained model to behave linearly, leading to (relatively) simple decision boundaries. This inductive bias has been further supported by a series of works (Guo et al., 2019a; Zhang et al., 2020; 2022; Chidambaram et al., 2021), which prove that the Mixup behaves similarly to standard training for linear models. In particular, Mixup applies the same linear interpolation on the features and labels of a pair of training data points $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$: denoted by $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ and labels $\lambda y_1 + (1-\lambda)y_2$, where $\lambda \in [0.5, 1]$ is randomly chosen. Then, the trained neural network (NN) model F is naturally encouraged to conduct the mapping $F(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \to \lambda y_1 + (1-\lambda)y_2$ for all $\lambda \in [0.5, 1]$, $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$, implying that F tends to behave linearly at least within the line segments between all training data pairs.

Although linearity is a nice inductive bias that tends to learn the models with low complexities, we are not clear about whether such an intuition from the algorithm design (i.e., performing the same linear interpolation for features and labels) can indeed explain the improvement in generalization. To examine this, we conduct a proof-of-concept experiment on CIFAR-10 dataset. Instead of using the same linear interpolation in the feature and label space, we implement the interpolations using different λ's for features



(a) ResNet18 (b) VGG16

Figure 1. Test accuracy achieved by Mixup training with different configurations of λ and $g(\lambda)$. The results are evaluated by training ResNet18 and VGG16 on CIFAR-10 dataset without random crop & flip data augmentation and weight decay regularization. We consider 5 different configurations: (1) $\lambda = g(\lambda) = 1$, i.e., standard training; (2) $\lambda = g(\lambda) \sim U[0.5, 1]$, i.e., standard Mixup; (3) $\lambda \sim U[0.5, 1]$ and $g(\lambda) = 1.5 - \lambda$; (4) $\lambda \sim U[0.5, 1]$ and $g(\lambda) \sim U[0.5, 1]$; (5) $\lambda = 0.7$ and $g(\lambda) = 0.8$. It is clear that the performance gain of Mixup does not require setting $g(\lambda) = \lambda$.

and labels, i.e., we implement the Mixup data augmentation on the features and labels as: $\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ and $g(\lambda)y_1 + [1-g(\lambda)]y_2$ for some nonlinear or even random function $g(\cdot) : \mathbb{R}^{[0.5,1]} \to \mathbb{R}^{[0.5,1]}$. Our results, shown in Figure 1, demonstrate that the substantial performance gain of Mixup training over standard training does not require $g(\lambda) = \lambda$. Other choices, such as fixed or independently random λ and $g(\lambda)$, can lead to comparable or even better performance.

Therefore, it demands seeking a more fundamental understanding of Mixup that is beyond the linearization illustration. To address this issue, we draw inspiration from a recent work (Shen et al., 2022), which regards standard image data augmentation as a form of feature manipulation. This perspective offers a general framework to investigate the behavior of various data augmentation techniques, including Mixup in deep learning. In particular, they consider a multi-view data model that consists of multiple feature vectors and noise vectors with different strengths and frequencies. More specifically, the feature vectors are categorized as the common ones (i.e., "easy to learn" features) and the rare ones (i.e., "hard to learn" features): the former refers to the feature appearing in a large fraction of data (thus contribute a lot to the gradient updates), and the latter refers to the features occurring in a small fraction of data (thus have limited contribution to the gradient). They further assume that the common features are the ones with rare orientations compared to the rare features and they can be balanced by applying data augmentations. For example, the common feature of a cow could be the left-facing cow, while the rare feature could be the right-facing cow, which can be generated by applying a horizontal flip to the common feature.

However, in many cases, the common and rare features may not be easily balanced by standard data augmentations. Let's still take the cow image as an example, the common and rare features could be brown cows and black cows, or front-view cows and side-view cows. Then the standard rotation or flip operations clearly cannot convert the common features to rare ones. We conjecture that Mixup may exhibit certain benefits in tackling this type of feature, as it has been shown to improve test accuracy when combined with standard data augmentations (Zhang et al., 2018). This motivates the problem setup considered in this study.

Particularly, we perform the theoretical study of the learning dynamics of Mixup based on a similar multi-view data model (see Definition 3.1 for more details): each data point will either contain a common feature vector with a relatively high probability $1-\rho$, or a rare feature vector with a relatively low probability ρ. The remaining components will be filled with random noise or feature noise. We then consider a two-layer convolutional neural network (CNN) model and study the learning behaviors of both standard training and Mixup training using gradient descent. The main contributions of this paper are highlighted as follows:

- We identify that the linearity illustration may not be able to fully elucidate the exceptional performance of Mixup. In particular, we show that using the same linear interpolations for both features and labels is not necessary, while some other choices, e.g., independently random linear interpolations, can also lead to substantial performance gains compared to standard training.
- We prove a negative result (Theorem 4.1) for standard training, demonstrating its inability to learn the rare features of the multi-view distribution. This failure leads to the domination of the rare feature data by its noise components during the test period, resulting in a $\Theta(\rho)$ test error. The reason for this lies in the tendency of the standard training algorithm to memorize the noise component of rare feature data to attain zero training error, while the rare feature itself, which appears in only a small fraction of the data, is not prominent enough to be effectively discovered by the algorithm.
- More importantly, we establish a positive result (Theorem 4.2) for Mixup training by showcasing its ability to attain near-zero test errors on the multi-view distribution. Specifically, we demonstrate that Mixup can successfully mix the common and rare features so that the gradients along these two features are correlated. As a result, the rare feature learning can be boosted by the fast learning of common features, and ultimately reaches a sufficiently high level to overshadow the effects of noise on test data.
- Our theory also suggests that the feature learning (especially the rare feature) benefits of Mixup are mostly gained in the early training phase. Then we develop the early-stopped Mixup, i.e., turning off the Mixup data augmentation after a certain number of iterations. Experimental results show that the test error achieved by early-stopped

Mixup is comparable to or even better than that achieved by standard Mixup (i.e., using Mixup throughout the entire training), which is consistent with the recent findings that over-training with Mixup may hurt the generalization (Liu et al., 2023). This not only corroborates our theoretical findings but also justifies the necessity to study the entire feature learning dynamics of Mixup rather than only the solution to the (equivalent) empirical risk of Mixup.

Notations. We use $\mathrm{poly}(n)$ and $\mathrm{polylog}(n)$ to denote a polynomial function, with a sufficiently large (constant) degree, of n or $\log(n)$ respectively. We use $o(1/\mathrm{polylog}(n))$ (and $\omega(\mathrm{polylog}(n))$) to denote some quantities that decrease (or grow) faster than $1/\log^c(n)$ (or $\log^c(n)$) for any constant c. We use $\widetilde{O}$, $\widetilde{\Omega}$, and $\widetilde{\Theta}$ to hide some log factors in the standard Big-O, Big-Omega, and Big-Theta notations.

2 Related Work

Theoretical Analysis of Mixup. We would like to comment on some recent works that attempt to explain the benefits of Mixup from different angles. To name a few, Thulasidasan et al. (2019) showed that the models trained by Mixup are substantially better calibrated, i.e., the softmax logits are closer to the actual likelihood than that obtained by standard training. Carratino et al. (2020) studied the regularization effect of Mixup training and connected it to multiple known data-dependent regularization schemes such as label smoothing. Following the same direction, Park et al. (2022) further developed a unified analysis for a class of Mixup methods, including the original one and CutMix (Yun et al., 2019), and proposed a hybrid version of Mixup that achieves better test performance. Chidambaram et al. (2021) studied the Mixup-optimal classifier and characterized its performance on original training data points. However, these works mostly focus on the solution to certain Mixup-version regularized empirical risk, while our experiments on early-stopped Mixup suggest that the entire learning dynamics could be more important.

Very recently, Chidambaram et al. (2022) conducted feature learning-based analyses for Mixup and demonstrated its benefits. However, we would like to clarify some differences in our theoretical analysis. Firstly, in terms of the Mixup method, they considered only the mid-point Mixup, where $\lambda = g(\lambda) = 0.5$, while we allow a more general choice of $\lambda \in (0.5, 1)$. Secondly, for the data model, we followed Shen et al. (2022) by considering a data model with two features of different frequencies (common and rare), feature noise, and random noise, while the random noise component, which plays an important role in memorizing all training data points (Allen-Zhu & Li, 2020b; Shen et al., 2022), was ignored in Chidambaram et al. (2022). Finally, in terms of theoretical analysis, Chidambaram et al. (2021) and our paper also differ due to our distinct data models. Specifically, their study focuses on the competence between learning two symmetric features, whereas our focus is on the competition between rare feature learning and noise memorization. In conclusion, while Chidambaram et al. (2022) and our work share a similar high-level spirit for understanding the benefits of Mixup, we approach this problem from different angles.

Data Augmentation. There are also many works studying the effect of standard data augmentation methods (i.e., performed within the data points) from different perspectives, such as regularization effect (Bishop, 1995; Dao et al., 2019; Wu et al., 2020), algorithm bias (Hanin & Sun, 2021), margins (Rajput et al., 2019), model invariance (Chen et al., 2020b), and feature learning (Shen et al., 2022). We view these works as orthogonal to our work as they mostly concern the data augmentation within the data points (e.g., random perturbation, random rotation, etc), which is different from the cross-data Mixup data augmentation.

Feature Learning in Deep Learning Theory. In the field of deep learning theory, there has emerged a series of works studying feature learning behavior during NN training. They focus on characterizing how different training approaches affect feature learning, such as ensembling & knowledge distillation (Allen-Zhu & Li, 2020b), using adaptive gradients (Zou et al., 2021), mixture of expert (Chen et al., 2022), and contrastive learning (Wen & Li, 2021). We point out that feature learning in Mixup is more complicated as the learning dynamics for different features are heavily coupled.

3 Problem Setting.

As mentioned in the introduction section, we theoretically investigate the behaviors of standard training and Mixup training on a multi-view data model. In this section, we will first deliver a detailed set up of the multi-view data model and then introduce the two-layer CNN model as well as the gradient descent algorithms of standard training and Mixup training.

3.1 Data Model

In this work, we consider a binary classification problem on the data $(\mathbf{x}, y) \in \mathbb{R}^{dP} \times \{1, 2\}$, where $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(P)})$ has P patches and $y \in \{1, 2\}$ denotes the data label. For ease of presentation, we define the data of label $y = 1$ as the *positive data* and the data of label $y = 2$ as the *negative data*. Moreover, the data will be randomly generated according to the following detailed process.

Definition 3.1. Let $\mathcal{D}$ denote the data distribution, from which a data point $(\mathbf{x}, y) \in \mathbb{R}^{dP} \times \{1, 2\}$ is randomly generated as follows:

1. Generate $y \in \{1, 2\}$ uniformly.
2. Generate $\mathbf{x}$ as a vector with P patches $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(2)}) \in (\mathbb{R}^d)^P$, where
   - **Feature Patch.** One patch, among all P patches, will

be randomly selected as the feature patch: with probability $1-\rho$ for some $\rho \in (0,1)$, this patch will contain a *common feature* ($\mathbf{v}$ for positive data, $\mathbf{u}$ for negative data); otherwise, this patch will contain a *rare feature* ($\mathbf{v}'$ for positive data, $\mathbf{u}'$ for negative data).

- **Feature Noise.** For all data, a feature vector from $\alpha \cdot \{\mathbf{u}, \mathbf{v}\}$ is randomly sampled and assigned to up to b patches.
- **Noise patch.** The remaining patches (those haven't been assigned with a feature or feature noise) are random Gaussian noise $\sim N(\mathbf{0}, \sigma_p^2 \cdot \mathbf{H})$, where $\mathbf{H} = \mathbf{I} - \frac{\mathbf{u}\mathbf{u}^\top}{\|\mathbf{u}\|_2^2} - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2} - \frac{\mathbf{v}'\mathbf{v}'^\top}{\|\mathbf{v}'\|_2^2} - \frac{\mathbf{u}'\mathbf{u}'^\top}{\|\mathbf{u}'\|_2^2}$.

We assume all feature vectors are orthonormal, i.e., $\|\mathbf{a}\|_2 = 1$ and $\langle \mathbf{a}, \boldsymbol{b} \rangle = 0$ for all $\mathbf{a}, \boldsymbol{b} \in \{\mathbf{v}, \mathbf{u}, \mathbf{v}', \mathbf{u}'\}$ and $\mathbf{a} \neq \boldsymbol{b}$. Moreover, we set $d = \omega(n^6)$, $P, b = \mathrm{polylog}(n)$, $\rho = \Theta(n^{-3/4})$, $\sigma_p = \Theta(d^{-1/2} n^{1/4})$, and $\alpha = \Theta(1/n)$[1].

Our data model includes three types of critical vectors: common features, rare features, and noise vectors (the feature noise vectors can be categorized into common features since they are only different in terms of strength). All of them can be leveraged to fit the training data points and thus achieve a small training accuracy/loss. However, in order to achieve a nearly perfect test accuracy, one has to learn both common features and rare features as overfitting the random noise vectors of training data points will make no contribution or even be detrimental to the test performance, then the prediction will be heavily affected by the feature noise. Given our data model in Definition 3.1, we aim to show that Mixup is able to learn all informative features while standard training may only learn a part of them.

The feature-noise data model has been widely adopted to study many algorithmic aspects of deep learning, including adversarial training (Allen-Zhu & Li, 2020a), momentum (Jelassi & Li, 2022), ensemble and knowledge distillation (Allen-Zhu & Li, 2020b), benign overfitting (Cao et al., 2022), and data augmentation (Shen et al., 2022). Our data model mostly follows from the one considered in Shen et al. (2022), which also includes the design of common features and rare features for studying the learning behaviors of data augmentation (that is performed within one single data point, e.g., random flip/rotation). However, instead of assuming that the rare features ($\mathbf{v}'$ and $\mathbf{u}'$) can be re-generated by applying data augmentation on the common features ($\mathbf{v}$ and $\mathbf{u}$), we make nearly no assumption on their relationships. Therefore, learning the rare features in our model can be regarded as a harder problem, and our theoretical analyses for Mixup are orthogonal to those in Shen et al. (2022).

[1]The choice of these parameters is not unique, here we only pick a feasible one for the ease of presentation.

3.2 Neural Network Function

Two-layer CNN model. We consider a two-layer CNN model F using quadratic activation function $\sigma(z) = z^2$. Note that we consider binary classification problem with $y \in \{1, 2\}$, then given the input feature $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(p)})$, the k-th output of the network ($k \in \{1, 2\}$) is formulated as

$$F_k(\mathbf{W}; \mathbf{x}) = \sum_{p=1}^{P} \sum_{r=1}^{m} (\langle \mathbf{w}_{k,r}, \mathbf{x}^{(p)} \rangle)^2.$$

where $\mathbf{w}_{k,r} \in \mathbb{R}^d$ denotes the neuron weight corresponding to the k-th output, $\mathbf{W}$ denotes the collection of all model weights, and m denotes the NN width, which is set as $m = \mathrm{polylog}(n)$ throughout this paper[2]. Moreover, given the input $\mathbf{x}$, we denote $\mathrm{Logit}_k(\mathbf{W}; \mathbf{x})$ by the logit of the k-th output of the NN model, which can be calculated via performing a softmax function on the NN outputs:

$$\mathrm{Logit}_k(\mathbf{W}; \mathbf{x}) = e^{F_k(\mathbf{W}; \mathbf{x}_i)} / \textstyle\sum_{s \in \{1,2\}} e^{F_s(\mathbf{W}, \mathbf{x}_i)}.$$

Using a polynomial activation function (or ReLU with polynomial smoothing) is not new in deep learning theory. The purpose is to better illustrate/distinguish the feature and noise learning dynamics during the neural network training (Frei et al., 2022; Cao et al., 2022; Shen et al., 2022; Glasgow et al., 2022). Our analysis can also be extended to other polynomial functions $\sigma(x) = x^q$ for some $q > 1$.

3.3 Training Algorithms

Initialization. We assume that the initial weights of the neural network model are generated i.i.d. from the Gaussian initialization: $\mathbf{w}_{k,r}^{(0)} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I})$, where $\sigma_0 = o(d^{-1/2})$.

Standard training. Given the training data points $\mathcal{S} := \{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$, we train the neural network model via applying standard full-batch gradient descent to optimize the following empirical risk function:

$$L_{\mathcal{S}}(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{W}; \mathbf{x}_i, y_i),$$
$$\text{where} \quad \ell(\mathbf{W}; \mathbf{x}_i, y_i) = -\log \frac{e^{F_{y_i}(\mathbf{W}, \mathbf{x}_i)}}{\sum_{k \in \{1,2\}} e^{F_k(\mathbf{W}; \mathbf{x}_i)}}.$$

Starting from the initialization $\mathbf{W}^{(0)}$, the gradient descent of the standard training takes the following update step

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \frac{\eta}{n} \sum_{i=1}^{n} \nabla_{\mathbf{W}} \ell(\mathbf{W}^{(t)}; \mathbf{x}_i, y_i), \qquad (3.1)$$

[2]This choice of network width is to guarantee some nice properties hold with probability at least $1 - 1/\mathrm{poly}(n)$ at the initialization. We can also resort to setting m as some large constant at the price of deriving a constant probability guarantee, e.g., > 0.9.

where η is the learning rate. Then, the detailed calculation of the partial derivative $\nabla_{\mathbf{w}_{k,r}}\ell(\mathbf{W};\mathbf{x}_i,y_i)$ is given by

$$\nabla_{\mathbf{w}_{k,r}}\ell(\mathbf{W};\mathbf{x}_i,y_i) = -2\ell_{k,i}\cdot\sum_{p=1}^{P}\langle\mathbf{w}_{k,r},\mathbf{x}_i^{(p)}\rangle\cdot\mathbf{x}_i^{(p)}.$$

where $\ell_{k,i} = \mathbb{1}_{k=y_i} - \mathrm{Logit}_k(\mathbf{W}^{(t)};\mathbf{x}_i)$.

Mixup Training. Given two training data points $(\mathbf{x}_1,y_1)$ and $(\mathbf{x}_2,y_2)$, Mixup trains a neural network based on the convex combinations of them: $(\lambda\mathbf{x}_1+(1-\lambda)\mathbf{x}_2, \lambda y_1+(1-\lambda)y_2)$ and $((1-\lambda)\mathbf{x}_1+\lambda\mathbf{x}_2,(1-\lambda)y_1+\lambda y_2)$, where we slightly abuse the notation by viewing the labels y_1 and y_2 as their one-hot encoding. Besides, Figure 1 suggested that λ does not need to be randomly sampled to achieve better performance than standard training, we will focus on a fixed constant $\lambda\in(0.5,1)$[3][4] in our theoretical analysis. Finally, if considering all possible combinations of the training data pairs with a fixed λ, the (equivalent) training dataset of Mixup training is $\mathcal{S}_{\mathrm{Mixup}} := \{\mathbf{x}_{i,j},y_{i,j}\}_{i,j\in[n]}$, where we denote $\mathbf{x}_{i,j}$ and $y_{i,j}$ by $\lambda\mathbf{x}_i+(1-\lambda)\mathbf{x}_j$ and $\lambda y_i+(1-\lambda)y_j$ respectively. Motivated by this, we can claim that the Mixup training actually aims to learn the model parameter by optimizing the following loss function:

$$L_{\mathcal{S}}^{\mathrm{Mixup}}(\mathbf{W}) = \frac{1}{n^2}\sum_{i,j\in[n]}\ell(\mathbf{W};\mathbf{x}_{i,j},y_{i,j}), \tag{3.2}$$

where

$$\ell(\mathbf{W};\mathbf{x}_{i,j},y_{i,j}) = \lambda\ell(\mathbf{W};\mathbf{x}_{i,j},y_i)+(1-\lambda)\ell(\mathbf{W};\mathbf{x}_{i,j},y_j).$$

In this paper, in order to better illustrate the key aspect of Mixup training as well as simplify the theoretical analysis, we resort to the gradient descent on the loss function (3.2), which takes the following update step:

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \frac{\eta}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\nabla_{\mathbf{W}}\ell(\mathbf{W}^{(t)};\mathbf{x}_{i,j},y_{i,j}).$$

Then, the detailed calculations of all partial derivatives are given as follows: for any Mixup data $(\mathbf{x}_{i,j},y_{i,j})$, we have

$$\nabla_{\mathbf{w}_{k,r}}\ell(\mathbf{W};\mathbf{x}_{i,j}) = 2\ell_{k,(i,j)}\cdot\sum_{p=1}^{P}\langle\mathbf{w}_{k,r},\mathbf{x}_{i,j}^{(p)}\rangle\cdot\mathbf{x}_{i,j}^{(p)},$$

where $\ell_{k,i}$ is the loss derivative with respect to the network output $F_k(\mathbf{W};\mathbf{x}_{i,j},y_{i,j})$:

$$\ell_{k,(i,j)} = \lambda\,\mathbb{1}_{k=y_i}+(1-\lambda)\,\mathbb{1}_{k=y_j} - \mathrm{Logit}_k(\mathbf{W};\mathbf{x}_{i,j}).$$

[3]If considering random λ, we will need to further take an expectation of $L_{\mathcal{S}}^{\mathrm{Mixup}}(\mathbf{W})$ defined in (3.2). As a result, our analysis, particularly Proposition 5.4 also needs to consider the additional expectation over λ in the definitions of coefficients $\gamma_k^{(t)}(\boldsymbol{b},\mathbf{a})$.

[4]Besides, we can also allow using different λ's for inputs and outputs, given our theoretical analysis in Section 5.2.

4 Main Theory

In this section, we will theoretically characterize the generalization errors achieved by standard training and Mixup training on the multi-view model. In particular, the following Theorem states the negative result of standard training.

Theorem 4.1. Suppose that the training data are generated according to Definition 3.1, let $\eta = 1/\mathrm{poly}(n)$, $T = \mathrm{polylog}(n)/\eta$, and $\{\mathbf{W}_{\mathrm{standard}}^{(t)}\}_{t=0,\ldots,T}$ be the iterates of standard training, then with probability at least $1-1/\mathrm{poly}(n)$, it holds that for all $t\in[0,T]$, $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\mathrm{argmax}_k F_k(\mathbf{W}_{\mathrm{standard}}^{(t)};\mathbf{x})\neq y\big]\geq\frac{\rho}{2.01}$.

Theorem 4.1 basically states that the two-layer CNN model obtained via standard training will lead to at least $\Theta(\rho)$ test error on the data model defined in Definition 3.1. In fact, as we will clarify in Section 5.1, this is due to the fact that the rare feature data will be fitted via their random noise components, while the rare features $\mathbf{v}'$ and $\mathbf{u}'$ will not be learned. Consequently, nearly a half of test rare feature data will be misled by the feature noise components, resulting in a $\Theta(\rho)$ test error.

In comparison, Mixup training can help learn the rare features and thus achieve a smaller generalization error. We formally state this result in the following theorem.

Theorem 4.2. Suppose the training data are generated according to Definition 3.1, let $\eta = \frac{1}{\mathrm{poly}(n)}$, $T = \frac{\mathrm{polylog}(n)}{\eta}$, and $\{\mathbf{W}_{\mathrm{Mixup}}^{(t)}\}_{t=0,\ldots,T}$ be the iterates of Mixup training, then with probability at least $1-\frac{1}{\mathrm{poly}(n)}$, it holds that for some $t\in[0,T]$, $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\mathrm{argmax}_k F_k(\mathbf{W}_{\mathrm{Mixup}}^{(t)};\mathbf{x})\neq y\big] = o\big(\frac{1}{\mathrm{poly}(n)}\big)$.

Theorem 4.2 shows that the two-layer CNN model obtained via Mixup training can achieve nearly zero test error, which is much better than that of standard training as $\rho = \Theta(n^{-3/4}) \gg o(1/\mathrm{poly}(n))$ (see Definition 3.1). In particular, as we will show in Section 5.2, at the core of Mixup training is that it mixes common features and rare features together, thus the learning of these two types of features will be coupled. Consequently, the learning of rare features will be "boosted" by the learning of common features, reaching a sufficiently large level that dominates the effect of feature noise.

5 Overview of the Analysis

According to the data model in Definition 3.1, the critical step of the generalization analysis for standard training and Mixup training is to sharply characterize the magnitude of the feature learning, including both common features ($\mathbf{v}$ and $\mathbf{u}$) and rare features ($\mathbf{v}'$, $\mathbf{u}'$), as well as the noise learning, including all noise vectors $\boldsymbol{\xi}_i^{(p)}$'s (denoted by $\{\boldsymbol{\xi}\}$). Then, the key step to show the generalization gap between standard

training and Mixup training is to identify their difference in terms of feature and noise learning.

5.1 Feature and Noise Learning of Standard Training

According to Definition 3.1, we define $\mathcal{S}_0^+$ and $\mathcal{S}_0^-$ as the set of training data that have strong positive and negative features respectively and $\mathcal{S}_1^+$ and $\mathcal{S}_1^-$ as the set of data that have weak positive and negative features respectively. In the following, the learning patterns of these vectors will be characterized by studying the inner products $\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{a}\rangle$, where $\mathbf{a} \in \{\mathbf{v}, \mathbf{u}, \mathbf{v}', \mathbf{u}'\} \cup \{\boldsymbol{\xi}\}$. Intuitively, a larger inner product implies that the neural network has a stronger learning ability of $\mathbf{a}$. Given the multi-view data model in Definition 3.1 and the update rule (3.1), we have for any $\mathbf{a} \in \{\mathbf{v}, \mathbf{u}, \mathbf{v}', \mathbf{u}'\} \cup \{\boldsymbol{\xi}\}$,

$$\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{a}\rangle = \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{a}\rangle + \frac{2\eta}{n} \cdot \sum_{i\in[n]} \ell_{k,i}^{(t)} \sum_{p=1}^{P} \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{x}_i^{(p)}\rangle \cdot \langle \mathbf{x}_i^{(p)}, \mathbf{a}\rangle. \tag{5.1}$$

Then by the data model in Definition 3.1, we can see that for common feature vector $\mathbf{a} \in \{\mathbf{v}, \mathbf{u}\}$, there will be $\Theta(n)$ training data points contributing to the learning of $\mathbf{a}$; while for rare feature vector $\mathbf{a} \in \{\mathbf{v}', \mathbf{u}'\}$, only $\Theta(\rho n)$ data points contributing to the learning. Besides, since each noise vector $\mathbf{a} \in \{\boldsymbol{\xi}\}$ in the training data point is randomly generated, its learning will largely rely on one single data, i.e., the data consisting of that noise vector. This difference clearly shows that the common features will be preferably discovered and learned during the standard training.

In the following analysis, we will decompose the entire standard training process into three phases, according to the learning of common features and noises. In particular, the **Phase 1** referred to the initial training iterations such that the neural network output, with respect to all input training data, is in the order of $O(1)$. In this phase, the loss derivatives $\ell_i^{(t)}$ will remain in the constant order and all critical vectors will be learned at a fast rate. Then The **Phase 2** is defined as the training period starting from the end of **Phase 1** to the iteration that the neural network output has reached $\widetilde{\Theta}(1)$ for all training inputs. Finally, we refer to **Phase 3** as the training period starting from the end of **Phase 2** to convergence, i.e., the gradient converges to zero.

Standard Training, Phase 1. The following lemma characterizes the learning of all features and noise in Phase 1.

Lemma 5.1. There exists a iteration number $T_0 = \widetilde{\Theta}(1/\eta)$ such that for any $t \le T_0$, it holds that

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle = \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle \cdot \big(1 + \Theta(\eta)\big), \tag{5.2}$$

$$\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{u}\rangle = \langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u}\rangle \cdot \big(1 + \Theta(\eta)\big). \tag{5.3}$$

Besides, for all remaining inner products, it holds that

$$\max_r |\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{a}\rangle| \le \max_r |\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{a}\rangle| \cdot \big[1 + o(\eta/\mathrm{polylog}(n))\big]$$

where $t \le T_0$, $r \in [m]$, $k \in [2]$, $q \in [P]$, $\mathbf{a} \in \{\mathbf{u}, \mathbf{v}, \mathbf{u}', \mathbf{v}'\} \cup \{\boldsymbol{\xi}\}$ are arbitrarily chosen as long as the inner products are different from those in (5.2).

Lemma 5.1 shows the competence results of learning common features, rare features, and noise vectors in Phase 1. In particular, it can be observed that the learning of common features ($\mathbf{v}$, $\mathbf{u}$) enjoys a much faster rate, while other critical vectors, including rare features and noise vectors, will be staying at their initialization levels.

Standard Training, Phase 2. During this phase, the loss derivative will remain in the constant order for the rare feature data, since either the rare feature learning (e.g, $\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle$) or the noise learning (e.g., $\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle$) are still quite small. Recall that the common features have already been fitted during Phase 1, we will then focus on the competence between learning rare features and learning noise vectors in Phase 2. The following lemma characterizes the dynamics of standard training in Phase 2.

Lemma 5.2. There exists a iteration number $T_1 = \widetilde{O}\big(\frac{n}{d\sigma^2\eta}\big)$ such that for any $t \in [T_0, T_1]$, it holds that

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}'\rangle = \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle \cdot \big[1 + \Theta(\rho\eta)\big],$$
$$\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{u}'\rangle = \langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u}'\rangle \cdot \big[1 + \Theta(\rho\eta)\big].$$

Besides, for any $i \in \mathcal{S}_1^+ \cup \mathcal{S}_1^-$, any $q \in [P]$ and $k = y_s$,

$$\max_r |\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)}\rangle| = \max_r |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)}\rangle| \cdot \Big[1 + \widetilde{\Theta}\Big(\frac{\eta d\sigma_p^2}{n}\Big)\Big]$$

Lemma 5.2 shows that for rare feature data points, standard training admits a faster noise learning speed compared to rare feature learning (note that $d\sigma_p^2 \gg \rho$, according to Definition 3.1). This consequently leads to adequate learning of noise ($|\langle \mathbf{w}_{y_i,r}^{(T_1)}, \boldsymbol{\xi}_i^{(p)}\rangle| = \widetilde{\Theta}(1)$ for some $p \in [P]$) and nearly no learning of rare features ($|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{v}'\rangle|, |\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{u}'\rangle| = \widetilde{O}\big(\sigma_0\big)$).

Standard Training, Final Phase. The final phase is defined as the training period after the end of Phase 2 until convergence. In the following lemma, we will show that (1) the convergence can be guaranteed; and (2) the learning of features and noise vectors at Phase 2 will be maintained.

Lemma 5.3. Let T_1 be the iteration number defined in Lemma 5.2, then for any $t = \mathrm{poly}(n) > T_1$ and $k \in \{1, 2\}$,

$$\frac{1}{n} \sum_{\tau=T_1}^{t} \sum_{i=1}^{n} |\ell_{k,i}^{(\tau)}| = \widetilde{O}(1/\eta).$$

Moreover, we have $\sum_{r=1}^{m}(\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle)^2, \sum_{r=1}^{m}(\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u}\rangle)^2 = \widetilde{\Theta}(1)$ and $|\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{v}'\rangle|, |\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{u}'\rangle| = \widetilde{O}(\sigma_0)$.

It can be clearly seen that the gradient descent can converge to the point with a small gradient (the averaged loss derivative will be roughly in the order of $\widetilde{O}(1/(t\eta))$, which approaches zero when t is large). More importantly, the common feature data and rare feature data will be correctly classified by fitting different components: common feature data will be fitted by learning $\mathbf{v}$ and $\mathbf{u}$, while the rare feature data will be fitted by noise memorization (as standard training nearly makes no progress in learning. Consequently, when it comes to a fresh test rare feature data, the model prediction will be heavily affected by the feature noise component, thus leading to an incorrect prediction with a constant probability (the formal proof is deferred to Section B.3).

5.2 Feature and Noise Learning of Mixup Training

As mentioned in Section 3.3, any data pair sampled from training dataset will be considered, which gives in total n^2 Mixup data. Note that we have two types of data in the origin training dataset: common feature data and rare feature data with two labels, denoted by $\mathcal{S}_0^+, \mathcal{S}_0^-, \mathcal{S}_1^+$, and $\mathcal{S}_1^-$ (see Section 5.1), we can also categorize the Mixup data points into multiple sets accordingly. Particularly, let $\mathcal{S}_{*,**}^{\dagger,\dagger\dagger}$ be the set of mixed data $\mathbf{x}_{i,j} = \lambda\mathbf{x}_i + (1-\lambda)\mathbf{x}_j$ with $\mathbf{x}_i \in \mathcal{S}_*^{\dagger}$ and $\mathbf{x}_j \in \mathcal{S}_{**}^{\dagger\dagger}$, we can accordingly categorize all Mixup data with the following 4 classes:

- Mix between two common feature data points, including $\mathcal{S}_{0,0}^{+,+}, \mathcal{S}_{0,0}^{-,-}, \mathcal{S}_{0,0}^{+,-}, \mathcal{S}_{0,0}^{-,+}$, each of them is of size $\Theta(n^2)$.
- Mix between common feature and rare feature data points with the same label, including $\mathcal{S}_{0,1}^{+,+}, \mathcal{S}_{0,1}^{-,-}, \mathcal{S}_{1,0}^{+,+}$, and $\mathcal{S}_{1,0}^{-,-}$, each of them is of size $\Theta(\rho n^2)$.
- Mix between common feature and rare feature data points with different labels, including $\mathcal{S}_{0,1}^{+,-}, \mathcal{S}_{0,1}^{-,+}, \mathcal{S}_{1,0}^{+,-}$, and $\mathcal{S}_{1,0}^{-,+}$, each of them is of size $\Theta(\rho n^2)$.
- Mix between two rare feature data points, including $\mathcal{S}_{1,1}^{+,+}$,$\mathcal{S}_{1,1}^{-,-}$, $\mathcal{S}_{1,1}^{+,-}$ and $\mathcal{S}_{1,1}^{-,+}$, each of them is of size $\Theta(\rho^2 n^2)$.

In contrast to standard training that nearly admits separate learning dynamics for common and rare features, the second and third classes of Mixup training data points, actively mix the common and rare features together. For instance, some data points in $\mathcal{S}_{0,1}^{+,+}$ will contain a data patch of form $\lambda\mathbf{v} + (1-\lambda)\mathbf{v}'$. Then the learning of $\mathbf{v}$ will benefit the learning of $\mathbf{v}'$, since their gradient updates are positively correlated. In the following, we will provide a precise characterization on the learning dynamics of feature and noise vectors.

In particular, noting that we consider the full-batch gradient descent on the entire Mixup training dataset (see Section 3.3), the update formula of all critical vectors are provided as follows: for any $\mathbf{a} \in \{\mathbf{u}, \mathbf{v}, \mathbf{u}', \mathbf{v}'\} \cup \{\boldsymbol{\xi}\}$, we have

$$\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{a}\rangle = \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{a}\rangle - \eta \cdot \langle \nabla_{\mathbf{w}_{k,r}} L(\mathbf{W}^{(t)}), \mathbf{a}\rangle. \quad (5.4)$$

where we denote $L(\mathbf{W}^{(t)})$ as the short-hand notation of $L_{\mathcal{S}}^{\text{Mixup}}$ (defined in (3.2)) for simplifying the notation. More specifically, we summarize the update of all critical vectors (e.g., common features, rare features, and data noise vectors) in the following Proposition.

Proposition 5.4. For any critical vector $\mathbf{a} \in \{\mathbf{v}, \mathbf{u}, \mathbf{v}', \mathbf{u}'\} \cup \{\boldsymbol{\xi}\}$, we have

$$-\langle \nabla_{\mathbf{w}_{k,r}} L(\mathbf{W}^{(t)}), \mathbf{a}\rangle = \sum_{\boldsymbol{b}\in\{\mathbf{v},\mathbf{u},\mathbf{v}',\mathbf{u}'\}\cup\{\boldsymbol{\xi}\}} \gamma_k^{(t)}(\boldsymbol{b}, \mathbf{a})\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{b}\rangle$$

where $\gamma_k^{(t)}(\boldsymbol{b}, \mathbf{a})$ is a scalar output function that depends on $\boldsymbol{b}, \mathbf{a} \in \{\mathbf{v}, \mathbf{u}, \mathbf{v}', \mathbf{u}'\} \cup \{\boldsymbol{\xi}\}$. More specifically, let

$$\mathbf{x}_{i,j}^{(p)} = \theta_{i,j}^{(p)}(\mathbf{v}) \cdot \mathbf{v} + \theta_{i,j}^{(p)}(\mathbf{u}) \cdot \mathbf{u} + \theta_{i,j}^{(p)}(\mathbf{v}') \cdot \mathbf{v}' + \theta_{i,j}^{(p)}(\mathbf{u}') \cdot \mathbf{u}' + \sum_{s=1}^{n}\sum_{q\in[P]} \theta_{i,j}^{(p)}(\boldsymbol{\xi}_s^{(q)}) \cdot \boldsymbol{\xi}_s^{(q)}$$

be a linear expansion of $\mathbf{x}_{i,j}^{(p)}$ on the space spanned by $\{\mathbf{v}, \mathbf{u}, \mathbf{v}', \mathbf{u}'\} \cup \{\boldsymbol{\xi}\}$, we have

$$\gamma_k^{(t)}(\boldsymbol{b}, \mathbf{a}) = \frac{1}{n^2}\sum_{i,j\in[n]} \ell_{k,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\boldsymbol{b}) \cdot \langle \mathbf{x}_{i,j}^{(p)}, \mathbf{a}\rangle.$$

From Proposition 5.4, it can be seen that the learning of common features, rare features, and noise vectors are heavily coupled. Mathematically, the coefficient $\gamma_k^{(t)}(\mathbf{a}, \boldsymbol{b})$ precisely describes how the learning of $\mathbf{a}$ affects the learning of $\boldsymbol{b}$, where $\mathbf{a}, \boldsymbol{b} \in \{\mathbf{v}, \mathbf{u}, \mathbf{v}', \mathbf{u}'\} \cup \{\boldsymbol{\xi}\}$. This effect can be either positive or negative, depending on the sign of $\gamma_k^{(t)}(\mathbf{a}, \boldsymbol{b})$. Then, the next step is to sharply characterize the coefficients $\gamma_k^{(t)}(\boldsymbol{b}, \mathbf{a})$. We will focus on early phase of Mixup training, where the loss derivatives can be regarded as the constant (i.e., approximately 0.5, -0.5, $\lambda - 0.5$, or $0.5 - \lambda$). Particularly, we will consider the training stage such that $\max_{k\in[2],i,j\in[n]} |F_k(\mathbf{W}^{(t)}; \mathbf{x}_{i,j})| \le \zeta$, where $\zeta = o\big(\frac{1}{\text{polylog}(n)}\big)$ is a user-defined parameter. Then based on ζ, we summarize the results of some critical coefficients in the following lemma, while the results for all coefficients are presented in Lemma C.3-C.11.

Lemma 5.5. Assume $\max_{k\in[2],i,j\in[n]} |F_k(\mathbf{W}^{(t)}; \mathbf{x}_{i,j})| \le \zeta$ for some $\zeta \in [\omega(d\sigma_p^2/(Pn)), o(d^{-1/2}\sigma_p^{-1})]$, then,

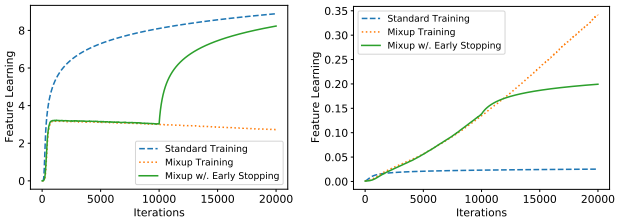
$$\gamma_1^{(t)}(\mathbf{v}, \mathbf{v}), \gamma_2^{(t)}(\mathbf{u}, \mathbf{u}) = \Theta(1),\ \gamma_{y_i}^{(t)}(\boldsymbol{\xi}_i^{(p)}, \boldsymbol{\xi}_i^{(p)}) = \Theta\big(d\sigma_p^2/n\big),$$
$$\gamma_1^{(t)}(\mathbf{v}, \mathbf{v}'), \gamma_2^{(t)}(\mathbf{u}, \mathbf{u}') = \Theta(\rho/P),$$
$$|\gamma_2^{(t)}(\mathbf{u}, \mathbf{v}')|, |\gamma_1^{(t)}(\mathbf{v}, \mathbf{u}')| = O(\zeta\rho/P).$$

The coefficients presented in Lemma 5.5 reveal some key differences between learning common features, rare features, and noise. Let's consider $\mathbf{v}$ without loss of generality. First, similar to the standard training, the learning of common features is much faster than the learning of noises, since the leading terms of common feature learning (i.e., $\gamma_1^{(t)}(\mathbf{v},\mathbf{v})$) and noise learning (i.e., $(\gamma_{y_i}^{(t)}(\boldsymbol{\xi}_i^{(p)},\boldsymbol{\xi}_i^{(p)}))$ satisfy: $\gamma_1^{(t)}(\mathbf{v},\mathbf{v}) \gg \gamma_{y_i}^{(t)}(\boldsymbol{\xi}_i^{(p)},\boldsymbol{\xi}_i^{(p)})$. Second, different from standard training where the rare features are nearly unexplored, Mixup training has the ability to boost the learning of rare features via common feature learning, which is characterized by $\gamma_1^{(t)}(\mathbf{v},\mathbf{v}')\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle$ or $\gamma_2^{(t)}(\mathbf{u},\mathbf{v}')\cdot\langle\mathbf{w}_{2,r}^{(t)},\mathbf{u}\rangle$. Finally, we also show that such a boosting effect is positive: the boosting of $\mathbf{v}'$ to the correct neurons (i.e., $\{\mathbf{w}_{1,r}^{(t)}\}_{r\in[m]}$) is stronger than that to the incorrect neurons (i.e., $\{\mathbf{w}_{2,r}^{(t)}\}_{r\in[m]}$), since $\gamma_1^{(t)}(\mathbf{v},\mathbf{v}') \gg |\gamma_2^{(t)}(\mathbf{u},\mathbf{v}')|$ (recall we pick $\zeta = o\big(\frac{1}{\mathrm{polylog}(n)}\big)$). This implies that the rare features will be effectively discovered by Mixup training, and finally, the neural network will have non-negligible components along the directions of $\mathbf{v}'$ and $\mathbf{u}'$. We formally stated this in the following lemma.

Lemma 5.6. Let ζ be the same as that in Lemma 5.5 and T be the smallest iteration number such that $\max_{k\in[2],i,j\in[n]}|F_k(\mathbf{W}^{(T)};\mathbf{x}_{i,j})| \geq \zeta/2$, then $T = \widetilde{O}(1/\eta)$ and with probability at least $1-1/\mathrm{poly}(n)$,

$$\max_r|\langle\mathbf{w}_{1,r}^{(T)},\mathbf{v}\rangle|,\max_r|\langle\mathbf{w}_{2,r}^{(T)},\mathbf{u}\rangle| = \widetilde{\Omega}(\zeta^{1/2}),$$
$$\max_r|\langle\mathbf{w}_{1,r}^{(T)},\mathbf{v}'\rangle|,\max_r|\langle\mathbf{w}_{2,r}^{(T)},\mathbf{u}'\rangle| = \Omega(\rho\zeta^{1/2})$$
$$\max_r|\langle\mathbf{w}_{2,r}^{(T)},\mathbf{v}\rangle|,\max_r|\langle\mathbf{w}_{1,r}^{(T)},\mathbf{u}\rangle| = \widetilde{O}(\zeta^{3/2}),$$
$$\max_r|\langle\mathbf{w}_{2,r}^{(T)},\mathbf{v}'\rangle|,\max_r|\langle\mathbf{w}_{1,r}^{(T)},\mathbf{u}'\rangle| = \widetilde{O}(\zeta^{3/2}).$$

We can then make a comparison between Lemma 5.3 and Lemma 5.6 to illustrate the similarities and differences between standard training and Mixup training in feature learning. In particular, it is clear that both standard and Mixup training can successfully learn the common features, i.e., the inner products $\langle\mathbf{w}_{1,r}^{(T)},\mathbf{v}\rangle$ and $\langle\mathbf{w}_{2,r}^{(T)},\mathbf{u}\rangle$ are the dominating ones among all critical inner products. While more importantly, the Mixup training can lead to much better rare feature learning compared to standard training: the standard training gives $|\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle|,|\langle\mathbf{w}_{2,r}^{(t)},\mathbf{u}\rangle| = \widetilde{O}(\sigma_0)$ for all iterations; in contrast, the Mixup training gives $|\langle\mathbf{w}_{1,r}^{(T)},\mathbf{v}'\rangle|,|\langle\mathbf{w}_{2,r}^{(T)},\mathbf{u}'\rangle| = \Omega(\rho\zeta^{1/2})$, which are much larger. Consequently, the strength of rare feature learning in Mixup training will dominate the effect of feature noise, thus achieving a nearly zero test error (the formal proof is deferred to Section C.5).



(a) Common Feature Learning (b) Rare Feature Learning

Figure 2. Common feature learning and rare feature learning on synthetic data, all experiments are conducted using full-batch gradient descent. Here we consider three training methods: standard training, Mixup training, and Mixup training with early stopping (at the 10000-th iteration).

5.3 Implications to the Early Stopping of Mixup

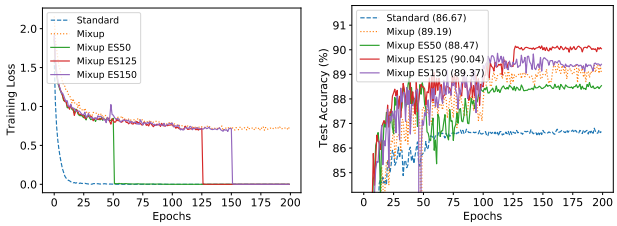
In addition to demonstrating the ability of Mixup in learning rare features, Lemma 5.6 also reveals that the benefits of Mixup training mostly come from its early training phase. Therefore, this motivates us to study the early-stopped Mixup training, i.e., the Mixup data augmentation will be turned off after a number of iterations. Then clearly, after turning off the Mixup data augmentation, the learned features will never be forgotten since the gradient update in this period will be always positively correlated (by (5.1)). This immediately leads to the following fact.

Fact 5.7. Let T be the same as that in Lemma 5.6, then if early stopping Mixup training at the iteration T, we have for any $t > T$, it holds that $\max_r|\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}'\rangle|,\max_r|\langle\mathbf{w}_{2,r}^{(t)},\mathbf{u}'\rangle| = \Omega(\rho\zeta^{1/2})$.

This further implies that applying proper early stopping in Mixup training will not affect the rare feature learning. Besides, turning off Mixup will enhance the learning of common features (since its learning speed will no longer be affected by the mix with rare features and noises), which could potentially lead to even better generalization performance. In the next section, we will empirically justify the effectiveness of applying early stopping in Mixup training.

6 Experiments

Synthetic Data. We first perform numerical experiments on synthetic data to verify our theoretical results. In particular, the synthetic data is generated according to Definition 3.1. In particular, we set dimension $d = 2000$, training sample size $n = 300$, the ratio of rare feature data $\rho = 0.1$, noise strength $\sigma_p = 0.15$, feature noise strength $\alpha = 0.05$, number of total patches $P = 5$, and number feature noise patches $b = 2$. For the two-layer CNN model and the training algorithm, we set network width $m = 10$, and conduct full-batch gradient descent with learning rate $\eta = 0.05$ and total iteration number $T = 20000$. We characterize the learning of common features and rare features via calculating $\sum_{r=1}^m(\langle\mathbf{w}_{1,r},\mathbf{v}\rangle)^2$ and $\sum_{r=1}^m(\langle\mathbf{w}_{1,r},\mathbf{v}'\rangle)^2$ (we only consider $\mathbf{v}$ and $\mathbf{v}'$ as the dynamics for $\mathbf{u}$ and $\mathbf{u}'$ are similar).



(a) Training Loss (b) Test Accuracy

Figure 3. Training loss (the cross-entropy loss on the mixup data/clean data) and test accuracy achieved by Mixup with different early stopping iterations: 0 (standard), 50, 125, 150, 200 (Mixup), numbers in the legend denote the average accuracy of the last 10 iterates. The results are evaluated by training ResNet18 on CIFAR-10 dataset without random crop & flip data augmentation and weight decay regularization.

The results are reported in Figure 2. It is clear that both standard training, Mixup training, and Mixup with early stopping can exhibit sufficiently common feature learning, while the rare feature learning of standard training is much lower than those of Mixup and Mixup with early stopping. This verifies Lemmas 5.3 and 5.6. Besides, we can also see that turning off Mixup after a number of iterations will lead to no decrease in rare feature learning and an increase in common feature learning. This verifies Fact 5.7 and demonstrates the benefits of early stopping.

CIFAR-10 Data. We further perform the Mixup training on CIFAR-10 dataset to evaluate the performance of early stopping, where we use SGD with momentum 0.9 and learning rate 0.1, followed by $\times 0.1$ decaying at the 100-th and 150-th iterations. We first train the ResNet18 model (He et al., 2015) via Mixup without other data augmentations and regularizations. We consider applying early stopping at the 0-th (standard training), 50-th, 125-th, 150-th, and 200-th (Mixup training) iterations and report the training loss and test accuracy in Figure 3. First, it can be observed that the cross-entropy loss on the training data quickly drops to nearly zero after the stopping of Mixup, showing that the neural network has correctly predicted the labels of training data points with high confidence. Besides, the test accuracy results show that such a high-confidence fitting on training data will not affect the test performance, while proper early stopping can even gain further improvements, e.g., Mixup with early stopping at the 125-th iteration achieves substantially higher test accuracy than that of Mixup training. This demonstrates the effectiveness of early-stopped Mixup and backs up our theoretical finding that the benefits of Mixup mainly stem from the early training phase.

7 Conclusion and Future Work

In this work, we attempted to develop a comprehensive understanding of the benefits of Mixup training. We first identified that the benefits cannot be fully explained by the linearity inductive bias of Mixup. Then we theoretically studied the dynamics of Mixup training from a feature learning. We showed that Mixup is more beneficial in learning rare features compared to standard training. Moreover, our analysis revealed that the benefits of Mixup in feature learning mostly stem from early training stages, based on which we developed the early-stopped Mixup. Our experimental results demonstrated that the early-stopped Mixup can achieve a comparable or even better performance than the standard one, which supports our theoretical findings.

One future direction is to theoretically investigate the benefit of a broader class of Mixup methods in various tasks. For instance, Han et al. (2022b); Yao et al. (2022a) proposed different Mixup methods to address the issue of distribution shift; Yao et al. (2022b) proposed C-Mixup methods for regression problems. It is also interesting to explore the theoretical understanding of Mixup for other types of data such as language data and graph data. We believe the theoretical framework developed in this work can be adapted, while a more precise and practical data model needs to be considered.

Acknowledgements

We would like to thank the anonymous reviewers and area chairs for their helpful comments. DZ acknowledges the central funding from HKU. ZL acknowledges National Science Foundation CCF2145703. QG is supported in part by the National Science Foundation IIS-2008981 and the Sloan Research Fellowship. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Allen-Zhu, Z. and Li, Y. Feature purification: How adversarial training performs robust deep learning. *arXiv preprint arXiv:2005.10190*, 2020a.

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

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Bishop, C. M. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

Cao, Y., Chen, Z., Belkin, M., and Gu, Q. Benign overfitting in two-layer convolutional neural networks. *arXiv preprint arXiv:2202.06526*, 2022.

Carratino, L., Cissé, M., Jenatton, R., and Vert, J.-P. On mixup regularization. *arXiv preprint arXiv:2006.06049*, 2020.

Chen, J., Yang, Z., and Yang, D. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2147–2157, 2020a.

Chen, S., Dobriban, E., and Lee, J. H. A group-theoretic framework for data augmentation. *The Journal of Machine Learning Research*, 21(1):9885–9955, 2020b.

Chen, Z., Deng, Y., Wu, Y., Gu, Q., and Li, Y. Towards understanding mixture of experts in deep learning. *arXiv preprint arXiv:2208.02813*, 2022.

Chidambaram, M., Wang, X., Hu, Y., Wu, C., and Ge, R. Towards understanding the data dependency of mixup-style training. In *International Conference on Learning Representations*, 2021.

Chidambaram, M., Wang, X., Wu, C., and Ge, R. Provably learning diverse features in multi-view data with midpoint mixup. *arXiv preprint arXiv:2210.13512*, 2022.

Dao, T., Gu, A., Ratner, A., Smith, V., De Sa, C., and Ré, C. A kernel theory of modern data augmentation. In *International Conference on Machine Learning*, pp. 1528–1537. PMLR, 2019.

Frei, S., Chatterji, N. S., and Bartlett, P. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pp. 2668–2703. PMLR, 2022.

Glasgow, M., Wei, C., Wootters, M., and Ma, T. Max-margin works while large margin fails: Generalization without uniform convergence. *arXiv preprint arXiv:2206.07892*, 2022.

Guo, H., Mao, Y., and Zhang, R. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3714–3722, 2019a.

Guo, H., Mao, Y., and Zhang, R. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*, 2019b.

Han, X., Jiang, Z., Liu, N., and Hu, X. G-mixup: Graph data augmentation for graph classification. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 8230–8248. PMLR, 17–23 Jul 2022a.

Han, Z., Liang, Z., Yang, F., Liu, L., Li, L., Bian, Y., Zhao, P., Wu, B., Zhang, C., and Yao, J. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. In *Advances in Neural Information Processing Systems*, 2022b.

Hanin, B. and Sun, Y. How data augmentation affects optimization for linear regression. *Advances in Neural Information Processing Systems*, 34:8095–8105, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Jelassi, S. and Li, Y. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, pp. 9965–10040. PMLR, 2022.

Liu, Z., Wang, Z., Guo, H., and Mao, Y. Over-training with mixup may hurt generalization. *arXiv preprint arXiv:2303.01475*, 2023.

Park, C., Yun, S., and Chun, S. A unified analysis of mixed sample data augmentation: A loss function perspective. In *Advances in Neural Information Processing Systems*, 2022.

Rajput, S., Feng, Z., Charles, Z., Loh, P.-L., and Papailiopoulos, D. Does data augmentation lead to positive margin? In *International Conference on Machine Learning*, pp. 5321–5330. PMLR, 2019.

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

Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Wen, Z. and Li, Y. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pp. 11112–11122. PMLR, 2021.

Wu, S., Zhang, H., Valiant, G., and Re, C. On the generalization effects of linear transformations in data augmentation. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10410–10420. PMLR, 13–18 Jul 2020.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022a.

Yao, H., Wang, Y., Zhang, L., Zou, J. Y., and Finn, C. C-mixup: Improving generalization in regression. *Advances in Neural Information Processing Systems*, 35: 3361–3376, 2022b.

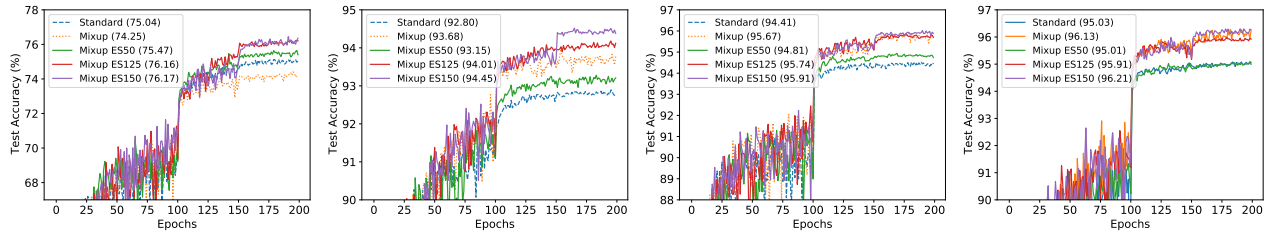
Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., and Zou, J. How does mixup help with robustness and generalization? In *International Conference on Learning Representations*, 2020.

Zhang, L., Deng, Z., Kawaguchi, K., and Zou, J. When and how mixup improves calibration. In *International Conference on Machine Learning*, pp. 26135–26160. PMLR, 2022.

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) LeNet (b) VGG16 (c) ResNet18 (d) ResNet34

Figure 4. Test errors achieved by Mixup training with different early stopping iterations: 0 (standard), 50, 125, 150, 200 (Mixup), numbers in the legend denote the average accuracy of the last 10 iterates. The results are evaluated by training LeNet, VGG16, ResNet18, and ResNet34 on CIFAR-10 dataset with random crop & flip data augmentation and weight decay regularization. Experimental results suggest that applying proper early stopping in Mixup will not downgrade the test performance but can even lead to higher test accuracy, especially for simpler models such as LeNet and VGG16.

A Additional Experiments

We further perform Mixup training for different neural network models and add the random crop/flip data augmentation and weight decay regularization (set as 10^{-4}). In particular, we consider two (relatively) high-capacity models: ResNet18 and ResNet34; and two low-capacity models: LeNet and VGG16. For ResNet18 and ResNet34, we set the learning rate as 0.1; for LeNet and VGG16, we set the learning rate as 0.02 and 0.1 respectively. Then we can clearly see that applying proper early stopping in Mixup will not downgrade the test performance but can even lead to higher test accuracy. In particular, Mixup with early stopping at the 50-th, 125-th, and 150-th iterations can still achieve a substantial performance improvement compared to standard training for LeNet, VGG16, and ResNet18. Moreover, we can also observe that Mixup with early stopping at the 150-th iteration performs better than the standard Mixup for all 4 models, especially for LeNet and VGG16, two relatively simpler models. This justifies our theoretical findings and demonstrates the benefit of early stopping in Mixup.

B Detailed Proof for Standard Training

B.1 Critical Quantities at the Initialization

Before moving on to the detailed characterization of the dynamics of standard training and Mixup training, we first characterize a set of critical quantities at the initialization. Recall (1) the data model in Definition 3.1 that the feature vectors have unit norm and the noise vectors are randomly generated from $N(\mathbf{0}, \sigma_p^2\mathbf{I})$; and (2) the initial model parameter $\mathbf{w}_{k,r}^{(0)}$ is randomly generated from $N(\mathbf{0}, \sigma_0^2\mathbf{I})$, we first give the following lemma that characterizes some critical quantities that will be repeatedly used in the later analysis.

Lemma B.1. With probability at least $1 - 1/\mathrm{poly}(n)$, it holds that for all $i \in [n]$, $k \in [2]$, $r \in [m]$, $\mathbf{a} \in \{\mathbf{v}, \mathbf{u}, \mathbf{v}', \mathbf{u}'\}$,

$$|\langle \mathbf{w}_{k,r}^{(0)}, \mathbf{a}\rangle| = \widetilde{O}(\sigma_0), \quad \sum_{r\in[m]} \big(\langle \mathbf{w}_{k,r}^{(0)}, \mathbf{a}\rangle\big)^2 = \widetilde{\Theta}(\sigma_0^2).$$

Additionally, for any noise patch $\boldsymbol{\xi} \in \{\boldsymbol{\xi}\}$,

$$|\langle \mathbf{w}_{k,r}^{(0)}, \boldsymbol{\xi}\rangle| = \widetilde{O}(d^{1/2}\sigma_p\sigma_0), \quad \sum_{r\in[m]} \big(\langle \mathbf{w}_{k,r}^{(0)}, \boldsymbol{\xi}\rangle\big)^2 = \widetilde{\Theta}(d\sigma_p^2\sigma_0^2).$$

Proof. Note that $\mathbf{w}_{k,r}^{(0)}$ is randomly generated from $N(\mathbf{0}, \sigma_0^2\mathbf{I})$. Then using the fact that $m = \mathrm{polylog}(n)$, $\|\mathbf{a}\|_2^2 = 1$, and $\|\boldsymbol{\xi}\|_2^2 = \Theta(d\sigma_p^2)$ with probability at least $1 - 1/\mathrm{poly}(n)$, applying standard concentration arguments can lead to the desired results.

□

B.2 Feature and Noise Learning of Standard Training

We first restate the feature and noise learning of standard training as follows: for features, we have

$$\begin{aligned}
\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{v} \rangle &= \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{u} \rangle + \frac{2\eta}{n} \cdot \sum_{i\in[n]} \ell_{k,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{v})} \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{v} \rangle \cdot \alpha_{i,p}^2 \|\mathbf{v}\|_2^2 \\
\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{u} \rangle &= \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{u} \rangle + \frac{2\eta}{n} \cdot \sum_{i\in[n]} \ell_{k,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{u})} \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{u} \rangle \cdot \alpha_{i,p}^2 \|\mathbf{u}\|_2^2, \\
\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{v}' \rangle &= \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{v} \rangle + \frac{2\eta}{n} \cdot \sum_{i\in\mathcal{S}_1^+} \ell_{k,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{v}')} \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{v}' \rangle \cdot \|\mathbf{v}'\|_2^2, \\
\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{u}' \rangle &= \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{u}' \rangle + \frac{2\eta}{n} \cdot \sum_{i\in\mathcal{S}_1^-} \ell_{k,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{u}')} \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{u}' \rangle \cdot \|\mathbf{u}'\|_2^2,
\end{aligned} \tag{B.1}$$

where $\mathcal{P}_i(\mathbf{a})$ denotes the set of patches in $\mathbf{x}_i$ containing the feature $\mathbf{a}$ and $\alpha_{i,p}^2 = 1$ if $\mathbf{x}_i^{(p)}$ is a feature patch and $\alpha_{i,p}^2 = \alpha^2$ if $\mathbf{x}_i^{(p)}$ is the feature noise. Additionally, note that the update of rare features only depends on the data in $\mathcal{S}_1^+$ and $\mathcal{S}_1^-$ since the data $(\mathbf{x}_i, y_i)$ in $\mathcal{S}_0^+$ and $\mathcal{S}_0^-$ satisfies $\mathcal{P}_i(\mathbf{v}') = \emptyset$ and $\mathcal{P}_i(\mathbf{u}') = \emptyset$. Similarly, we can also obtain the following result regarding noise learning

$$\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle = \langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle + \frac{2\eta}{n} \cdot \sum_{i=1}^{n} \ell_{k,i}^{(t)} \sum_{p=1}^{P} \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{x}_i^{(p)} \rangle \cdot \langle \mathbf{x}_i^{(p)}, \boldsymbol{\xi}_s^{(q)} \rangle.$$

Moreover, note that if $\mathbf{x}_i^{(p)} \neq \boldsymbol{\xi}_s^{(q)}$ (i.e., $i \neq s$ or $p \neq q$), then $|\langle \mathbf{x}_i^{(p)}, \boldsymbol{\xi}_s^{(p)} \rangle|$ is in the order of $\widetilde{O}(d^{1/2}\sigma_p^2)$. Therefore, we further have

$$\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle = \langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle \cdot \left[1 + \frac{2\eta}{n} \cdot \ell_{k,s}^{(t)} \cdot \|\boldsymbol{\xi}_s^{(q)}\|_2^2 \right] \pm \frac{2\eta}{n} \cdot \sum_{i\neq s || p\neq q} |\ell_{k,i}^{(t)}| \cdot |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_i^{(q)} \rangle| \cdot \widetilde{O}\big(d^{1/2}\sigma_p^2\big). \tag{B.2}$$

Phase 1, Fitting Common Feature Data. The following lemma characterizes the learning of all feature and noise vectors in Phase 1.

Lemma B.2 (Phase 1, Standard Training)**.** Let T_0 be the iteration number such that the neural network output satisfies $|F_k(\mathbf{W}^{(t)}; \mathbf{x}_i)| \leq O(1)$ for all $t \leq T_0$ and $i \in [n]$, then for any $t \leq T_0$, it holds that

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v} \rangle = \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle \cdot \big(1 + \Theta(\eta)\big), \quad \langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{u} \rangle = \langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u} \rangle \cdot \big(1 + \Theta(\eta)\big).$$

Besides, we also have for any $t \leq T_0$, $r \in [m]$, $k \in [2]$, $q \in [P]$, and $s \in [n]$,

$$\begin{aligned}
&|\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v} \rangle| = \widetilde{O}(\sigma_0), \quad |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u} \rangle| = \widetilde{O}(\sigma_0), \\
&|\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{v}' \rangle| = \widetilde{O}(\sigma_0), \quad |\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{u}' \rangle| = \widetilde{O}(\sigma_0), \quad |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle| = \widetilde{O}\big(d^{1/2}\sigma_p\sigma_0\big).
\end{aligned}$$

Proof. First, note that in the first stage, the neural network outputs are in the order of $O(1)$, implying that the loss derivatives satisfy $|\ell_{k,i}^{(t)}| = \Theta(1)$. More specifically, we can get that $\ell_{k,i}^{(t)} = \Theta(1)$ if $k = y_i$ and $\ell_{k,i}^{(t)} = -\Theta(1)$ otherwise. Then by (B.1), we have

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v} \rangle = \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle \cdot \left[1 + \frac{2\eta}{n} \cdot \sum_{i\in\mathcal{S}_0^+} \ell_{1,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{v})} \alpha_{i,p}^2 \|\mathbf{v}\|_2^2 + \frac{2\eta}{n} \cdot \sum_{i\in[n]\backslash\mathcal{S}_0^+} \ell_{1,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{v})} \alpha_{i,p}^2 \|\mathbf{v}\|_2^2 \right].$$

Note that by Definition 3.1, for any data $i \in [n]$ let $\mathcal{P}_i'(\mathbf{v})$ and $\mathcal{P}_i'(\mathbf{u})$ be the set of patches corresponding to the feature noise vectors $\mathbf{v}$ and $\mathbf{u}$ respectively, we have $|\mathcal{P}_i'(\mathbf{v})| \leq b$ and $\sum_{p\in\mathcal{P}_i'(\mathbf{v})} \alpha_{i,p}^2 \leq b\alpha^2 = o\big(1/\mathrm{polylog}(n)\big)$. Additionally, note that $\ell_{1,i}^{(t)} = \Theta(1)$ for $i \in \mathcal{S}_0^+$ and $\mathcal{P}_i(\mathbf{v}) = \mathcal{P}_i'(\mathbf{v})$ for all $i \in [n]\backslash\mathcal{S}_0^+$, we have

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v} \rangle = \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle \cdot \left[1 + \frac{2\eta}{n} \cdot |\mathcal{S}_0^+| \cdot C_v^{(t)} \pm o\big(\eta/\mathrm{polylog}(n)\big) \right] = \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle \cdot \big[1 + \Theta(\eta)\big], \tag{B.3}$$

where $C_v^{(t)} = |\mathcal{S}_0^+|^{-1} \cdot \sum_{i\in\mathcal{S}_0^+} \ell_{1,i}^{(t)}$ remains in the constant level for all $t \le T_0$. Similarly, we can also get that

$$\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{u}\rangle = \langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u}\rangle \cdot \bigg[1 + \frac{2\eta}{n} \cdot |\mathcal{S}_0^-| \cdot C_u^{(t)} \pm o\big(\eta/\mathrm{polylog}(n)\big)\bigg] = \langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u}\rangle \cdot \big[1 + \Theta(\eta)\big], \tag{B.4}$$

where $C_u^{(t)} = |\mathcal{S}_0^-|^{-1} \cdot \sum_{i\in\mathcal{S}_0^-} \ell_{2,i}^{(t)}$ remains in the constant level for all $t \le T_0$. Moreover, in terms of the learning of wrong features, we have

$$\begin{aligned}
\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{v}\rangle &= \langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}\rangle \cdot \bigg[1 + \frac{2\eta}{n} \cdot \sum_{i\in\mathcal{S}_0^+} \ell_{2,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{v})} \alpha_{i,p}^2 \|\mathbf{v}\|_2^2 + \frac{2\eta}{n} \cdot \sum_{i\in[n]\setminus\mathcal{S}_0^+} \ell_{2,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{v})} \alpha_{i,p}^2 \|\mathbf{v}\|_2^2\bigg] \\
&= \langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}\rangle \cdot \bigg[1 - \frac{2\eta}{n} \cdot |\mathcal{S}_0^+| \cdot \Theta(1) \pm o(\eta)\bigg] \\
&= \langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}\rangle \cdot \big[1 - \Theta(\eta)\big].
\end{aligned} \tag{B.5}$$

Then by Lemma B.1, this further implies that for all t in the first stage, we have

$$|\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}\rangle| \le |\langle \mathbf{w}_{2,r}^{(t-1)}, \mathbf{v}\rangle| \le \cdots \le |\langle \mathbf{w}_{2,r}^{(0)}, \mathbf{v}\rangle| = \widetilde{O}(\sigma_0). \tag{B.6}$$

Now we can move on to the learning of rare features and noise vectors. Particularly, for rare features, we have

$$\begin{aligned}
\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}'\rangle &= \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle \cdot \bigg[1 + \frac{2\eta}{n} \cdot \sum_{i\in\mathcal{S}_1^+} \ell_{k,i}^{(t)} \cdot \sum_{p\in\mathcal{P}_i(\mathbf{v}')} \|\mathbf{v}'\|_2^2\bigg] \\
&= \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle \cdot \bigg[1 + \Theta\bigg(\frac{\eta|\mathcal{S}_1^+|}{n}\bigg)\bigg] \\
&= \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle \cdot \big[1 + \Theta(\rho\eta)\big],
\end{aligned}$$

where the second equality is due to $|\mathcal{P}_i(\mathbf{v}')| = \Theta(1)$ and the last equality is due to $|\mathcal{S}_1^+| = \Theta(\rho n)$ with probability at least $1 - 1/\mathrm{poly}(n)$. Therefore, by Lemma B.1, we can then obtain

$$|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle| \le \big[1 + \Theta(\rho\eta)\big]^t \cdot |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle| \le \widetilde{O}(\sigma_0) \cdot e^{\Theta(T_0\eta)} = \widetilde{O}(\sigma_0),$$

where we use the fact that $T_0 = \widetilde{O}(1/\eta)$. Similarly, it also follows that

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}'\rangle = \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle \cdot [1 + \Theta(\rho\eta)] = \widetilde{O}(\sigma_0).$$

Moreover, using the fact that $\ell_{2,i}^{(t)} = -\Theta(1)$ for $i \in \mathcal{S}_1^+$ and $\ell_{1,i}^{(t)} = -\Theta(1)$ for $i \in \mathcal{S}_2^+$, we can follow the same proof in (B.5) and (B.6) and get

$$|\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}'\rangle| \le |\langle \mathbf{w}_{2,r}^{(0)}, \mathbf{v}'\rangle| = \widetilde{O}(\sigma_0),\ |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle| \le |\langle \mathbf{w}_{1,r}^{(0)}, \mathbf{u}'\rangle| = \widetilde{O}(\sigma_0).$$

where the results for $|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle|$ and $|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle|$ are by Lemma B.1.

Finally, regarding the learning of the noise vector $\boldsymbol{\xi}_s^{(q)}$, if $j = y_s$, we have the following by (B.2),

$$\max_{s,r} |\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)}\rangle| \le \max_{s,r} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)}\rangle| \cdot \bigg[1 + \frac{\eta}{n} \cdot \widetilde{\Theta}(d\sigma_p^2) + \frac{\eta}{n} \cdot \widetilde{O}\big(nPd^{1/2}\sigma_p^2\big)\bigg].$$

Note that we have $nP = o(d^{1/2})$, then the above equation further leads to

$$\max_{s,r} |\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)}\rangle| = \max_{s,r} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)}\rangle| \cdot \bigg[1 + \frac{\eta}{n} \cdot \widetilde{\Theta}(d\sigma_p^2)\bigg].$$

Besides, we can also get if $k \neq y_s$,

$$\max_{s,r} |\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle| \leq \max_{s,r} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle| \cdot \bigg[1 - \frac{\eta}{n} \cdot \widetilde{\Theta}(d\sigma_p^2)\bigg].$$

Then for any $t \leq T_0 = \widetilde{O}(1/\eta)$ and any k, we have

$$\begin{aligned}
\max_{s,r} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} | &\leq \max_{s,r} |\langle \mathbf{w}_{k,r}^{(0)}, \boldsymbol{\xi}_s^{(q)} \rangle| \cdot \bigg[1 + \frac{\eta}{n} \cdot \widetilde{\Theta}(d\sigma_p^2)\bigg]^t \\
&\leq \max_{s,r} |\langle \mathbf{w}_{k,r}^{(0)}, \boldsymbol{\xi}_s^{(q)} \rangle| \cdot \bigg[1 + \frac{\eta}{n} \cdot \widetilde{\Theta}(d\sigma_p^2)\bigg]^{T_0} \\
&\leq \max_{s,r} |\langle \mathbf{w}_{k,r}^{(0)}, \boldsymbol{\xi}_s^{(q)} \rangle| \cdot \exp\big\{\widetilde{\Theta}(\eta T_0 d\sigma_p^2/n)\big\} \\
&\leq \max_{s,r} |\langle \mathbf{w}_{k,r}^{(0)}, \boldsymbol{\xi}_s^{(q)} \rangle| \cdot \Theta(1) \\
&= \widetilde{\Theta}(d^{1/2}\sigma_0\sigma_p).
\end{aligned}$$

This completes the proof. □

Lemma B.3. At the end of Phase 1 with maximum iteration number $T_0 = \widetilde{O}(1/\eta)$, we have

$$\sum_{r=1}^m (\langle \mathbf{w}_{1,r}^{(T_0)}, \mathbf{v} \rangle)^2 = \widetilde{\Theta}(1), \ \sum_{r=1}^m (\langle \mathbf{w}_{2,r}^{(T_0)}, \mathbf{u} \rangle)^2 = \widetilde{\Theta}(1);$$

besides, it holds that

$$|\langle \mathbf{w}_{2,r}^{(T_0)}, \mathbf{v} \rangle|, |\langle \mathbf{w}_{1,r}^{(T_0)}, \mathbf{u} \rangle|, |\langle \mathbf{w}_{k,r}^{(T_0)}, \mathbf{u}' \rangle|, |\langle \mathbf{w}_{k,r}^{(T_0)}, \mathbf{v}' \rangle| = \widetilde{O}(\sigma_0); \quad |\langle \mathbf{w}_{k,r}^{(T_0)}, \boldsymbol{\xi} \rangle| = \widetilde{O}(d^{1/2}\sigma_p\sigma_0)$$

for all $k \in [2]$, $r \in [m]$ and $\boldsymbol{\xi} \in \{\boldsymbol{\xi}\}$.

Proof. We first characterize the difference between $C_v^{(t)}$ and $C_u^{(t)}$ in (B.3) and (B.4). Particularly, we consider the iterations that $\max_{i\in[n],k\in[2]} |F_k(\mathbf{W}^{(t)}; \mathbf{x}_i)| \leq \zeta$ for some $\zeta = \Theta\big(1/\log(1/\sigma_0)\big) = \Theta(1/\mathrm{polylog}(n))$, then we can immediately get that it holds that $|\ell_{1,i}^{(t)} - 0.5| \leq O(\zeta)$ for all $i \in \mathcal{S}_0^+$ and $|\ell_{2,i}^{(t)} - 0.5| \leq O(\zeta)$ for all $i \in \mathcal{S}_0^-$. Therefore, we can further get

$$C_v^{(t)} = \frac{1}{|\mathcal{S}_0^+|} \cdot \sum_{i\in\mathcal{S}_0^+} \ell_{1,i}^{(t)} = 0.5 \pm O(\zeta), \quad C_u^{(t)} = \frac{1}{|\mathcal{S}_0^-|} \cdot \sum_{i\in\mathcal{S}_0^-} \ell_{2,i}^{(t)} = 0.5 \pm O(\zeta).$$

Further note that the positive and negative data are independently generated from the data distribution, which implies that with probability at least $1 - 1/\mathrm{poly}(n)$, it holds that $||\mathcal{S}_0^+| - (1-\rho)n/2| \leq \widetilde{O}(n^{1/2})$ and $||\mathcal{S}_0^-| - (1-\rho)n/2| \leq \widetilde{O}(n^{1/2})$. Therefore, applying the fact that $\zeta = \Theta(1/\mathrm{polylog}(n))$, we can obtain the following by (B.3) and (B.4)

$$\begin{aligned}
\sum_{r=1}^m (\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v} \rangle)^2 &= \sum_{r=1}^m (\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle)^2 \cdot \Big[1 + (1-\rho)\eta \pm O(\zeta\eta))\Big] \\
\sum_{r=1}^m (\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{u} \rangle)^2 &= \sum_{r=1}^m (\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u} \rangle)^2 \cdot \Big[1 + (1-\rho)\eta \pm O(\zeta\eta))\Big].
\end{aligned} \tag{B.7}$$

Then let T_0' be the largest iteration number such that $\max_{k,i} |F_k(\mathbf{W}^{(t)}; \mathbf{x}_i)| \leq \zeta$, which clearly satisfies $T_0' < T_0$ (T_0 is defined in Lemma B.2), applying Lemma B.2 and considering the data i with largest neural network output (w.o.l.g assuming it's positive data),

$$\sum_{r=1}^m (\langle \mathbf{w}_{1,r}^{(T_0'+1)}, \mathbf{v} \rangle)^2 \geq c \cdot F_1(\mathbf{W}^{(T_0'+1)}; \mathbf{x}_i) \geq c \cdot \zeta$$

for some absolute constant c. By (B.7), we can immediately obtain that $T_0' = \Theta(\log(\zeta/(m\sigma_0^2))/\eta)$, where we apply the initialization results in Lemma B.1. Besides, we can also obtain that

$$\frac{\sum_{r=1}^m(\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{u}\rangle)^2}{\sum_{r=1}^m(\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle)^2} \geq \frac{\sum_{r=1}^m(\langle \mathbf{w}_{2,r}^{(0)}, \mathbf{u}\rangle)^2}{\sum_{r=1}^m(\langle \mathbf{w}_{1,r}^{(0)}, \mathbf{v}\rangle)^2} \cdot \bigg(\frac{1+(1-\rho)\eta - O(\zeta\eta))}{1+(1-\rho)\eta + O(\zeta\eta)}\bigg)^{T_0'}$$
$$= \widetilde{\Theta}(1) \cdot \big(1 - O(\eta\zeta T_0)\big).$$

Then note that $\zeta = \Theta\big(1/\log(1/\sigma_0)\big)$, we can get $\zeta T_0 \eta = \Theta\big(\zeta\log(\zeta) + \zeta\log(1/(m\sigma_0^2))\big) = o(1)$, which implies that $\sum_{r=1}^m(\langle \mathbf{w}_{2,r}^{(T_0'+1)}, \mathbf{u}\rangle)^2 \geq \Theta(\zeta)$. Finally, by Lemma B.2, we know that $\sum_{r=1}^m(\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle)^2$ and $\sum_{r=1}^m(\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{u}\rangle)^2$ will keep increasing for all $t \leq T_0$. Then based on the definition of T_0 and the fact that $\zeta = \widetilde{\Theta}(1)$, we can conclude that

$$\sum_{r=1}^m(\langle \mathbf{w}_{1,r}^{(T_0)}, \mathbf{v}\rangle)^2 = \widetilde{\Theta}(1),\ \sum_{r=1}^m(\langle \mathbf{w}_{2,r}^{(T_0)}, \mathbf{u}\rangle)^2 = \widetilde{\Theta}(1).$$

The remaining arguments in this lemma directly follow from Lemma B.2, thus we omit their proof here. □

Phase 2. Fitting Rare Feature Data. After **Phase 1**, the neural network output will become larger so that the loss derivatives (i.e, $\ell_{k,i}^{(t)}$) or the output logits may no longer be viewed as a quantity in the constant order. Particularly, as shown in Lemma B.3, when $t > T_0$, the feature learning, i.e., $\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle$ and $\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u}\rangle$ will reach the constant order, implying that $|\ell_{k,i}^{(t)}|$ will be closer to 1 or 0 for all common feature data. Additionally, the loss derivative will remain in the constant order for the rare feature data, since either the rare feature learning (e.g, $\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle$) or the noise learning (e.g., $\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle$) will be in the order of $o\big(1/\mathrm{polylog}(n)\big)$, so that the corresponding neural network outputs are also in the order of $o\big(1/\mathrm{polylog}(n)\big)$. Therefore, we define **Phase 2** by the period that (1) is after **Phase 1** and (2) the neural network outputs for the rare feature data are still in the order of $O\big(1/\mathrm{polylog}(n)\big)$ (or equivalently, the loss derivatives of rare feature data are in the constant order.)

Then, similar to the analysis in Phase 1, we will also characterize the learning of feature and noise separately. Regarding the learning of common feature, by (B.1), we have

$$\begin{aligned}\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{v}\rangle &= \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{v}\rangle + \frac{2\eta}{n} \cdot \sum_{i\in[n]} \ell_{k,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{v})} \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{v}\rangle \cdot \alpha_{i,p}^2 \\ &= \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{v}\rangle \cdot \bigg[1 + \frac{2\eta}{n} \cdot \bigg(\sum_{i\in\mathcal{S}_0^+} \ell_{k,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{v})} \alpha_{i,p}^2 + \sum_{i\in\mathcal{S}_0^-} \ell_{k,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{v})} \alpha_{i,p}^2 + \sum_{\mathcal{S}_1^+\cup\mathcal{S}_1^-} \ell_{k,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{v})} \alpha_{i,p}^2\bigg)\bigg].\end{aligned} \tag{B.8}$$

Similarly, we can also get that

$$\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{u}\rangle = \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{u}\rangle \cdot \bigg[1 + \frac{2\eta}{n} \cdot \bigg(\sum_{i\in\mathcal{S}_0^+} \ell_{k,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{u})} \alpha_{i,p}^2 + \sum_{i\in\mathcal{S}_0^-} \ell_{k,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{u})} \alpha_{i,p}^2 + \sum_{\mathcal{S}_1^+\cup\mathcal{S}_1^-} \ell_{k,i}^{(t)} \sum_{p\in\mathcal{P}_i(\mathbf{u})} \alpha_{i,p}^2\bigg)\bigg]. \tag{B.9}$$

Moreover, according to the data distribution in Definition 3.1, we have

- For any $i \in \mathcal{S}_0^+$, it holds that $\sum_{p\in\mathcal{P}_i(\mathbf{v})} \alpha_{i,p}^2 = \Theta(1)$ and $\sum_{p\in\mathcal{P}_i(\mathbf{u})} \alpha_{i,p}^2 = b\alpha^2 = o\big(1/\mathrm{polylog}(n)\big)$.
- For any $i \in \mathcal{S}_0^-$, it holds that $\sum_{p\in\mathcal{P}_i(\mathbf{u})} \alpha_{i,p}^2 = \Theta(1)$ and $\sum_{p\in\mathcal{P}_i(\mathbf{v})} \alpha_{i,p}^2 = b\alpha^2 = o\big(1/\mathrm{polylog}(n)\big)$.
- For any $i \in \mathcal{S}_1^+ \cup \mathcal{S}_1^-$, it holds that $\sum_{p\in\mathcal{P}_i(\mathbf{u})} \alpha_{i,p}^2 = b\alpha^2 = o\big(1/\mathrm{polylog}(n)\big)$ and $\sum_{p\in\mathcal{P}_i(\mathbf{v})} \alpha_{i,p}^2 = b\alpha^2 = o\big(1/\mathrm{polylog}(n)\big)$

Therefore, we have the following results regarding the relation between $\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{v}\rangle$ and $\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{u}\rangle$.

Lemma B.4. Let $T_1' = O\big(1/(\eta\rho b\alpha^2)\big)$ be a quantity that is greater than T_0, then for any $t \in [T_0, T_1]$, there exists an absolute constant C such that

$$\frac{|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle|}{|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle|} \geq C \cdot \frac{|\langle \mathbf{w}_{1,r}^{(T_0)}, \mathbf{v}\rangle|}{|\langle \mathbf{w}_{1,r}^{(T_0)}, \mathbf{u}\rangle|} = \widetilde{\Omega}\bigg(\frac{1}{\sigma_0}\bigg), \quad \text{and} \quad \frac{|\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u}\rangle|}{|\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}\rangle|} \geq C \cdot \frac{|\langle \mathbf{w}_{2,r}^{(T_0)}, \mathbf{u}\rangle|}{|\langle \mathbf{w}_{2,r}^{(T_0)}, \mathbf{v}\rangle|} = \widetilde{\Omega}\bigg(\frac{1}{\sigma_0}\bigg).$$

Proof. Based on the update rules in (B.8) and (B.9), we have

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle = \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle \cdot \bigg[1 + \frac{\eta}{n} \cdot \bigg(\Theta(1) \cdot \sum_{i\in\mathcal{S}_0^+} \ell_{1,i}^{(t)} + o\big(1/\mathrm{polylog}(n)\big) \cdot \sum_{i\in\mathcal{S}_0^-} \ell_{1,i}^{(t)} \pm O(\rho n b\alpha^2)\bigg)\bigg];$$
$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle = \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle \cdot \bigg[1 + \frac{\eta}{n} \cdot \bigg(\Theta(1) \cdot \sum_{i\in\mathcal{S}_0^-} \ell_{1,i}^{(t)} + o\big(1/\mathrm{polylog}(n)\big) \cdot \sum_{i\in\mathcal{S}_0^+} \ell_{1,i}^{(t)} \pm O(\rho n b\alpha^2)\bigg)\bigg].$$

where we use the fact that $|\ell_{k,i}^{(t)}| \leq 1$. This further implies that

$$\frac{|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle|}{|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle|} = \frac{|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle|}{|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle|} \cdot \underbrace{\frac{1 + \frac{\eta}{n} \cdot \bigg(\Theta(1) \cdot \sum_{i\in\mathcal{S}_0^+} \ell_{1,i}^{(t)} + o\big(1/\mathrm{polylog}(n)\big) \cdot \sum_{i\in\mathcal{S}_0^-} \ell_{1,i}^{(t)} \pm O(\rho n b\alpha^2)\bigg)}{1 + \frac{\eta}{n} \cdot \bigg(\Theta(1) \cdot \sum_{i\in\mathcal{S}_0^-} \ell_{1,i}^{(t)} + o\big(1/\mathrm{polylog}(n)\big) \cdot \sum_{i\in\mathcal{S}_0^+} \ell_{1,i}^{(t)} \pm O(\rho n b\alpha^2)\bigg)}}_{\star}.$$

Note that we have $\ell_{1,i}^{(t)} > 0$ for $i \in \mathcal{S}_0^+$ and $\ell_{1,i}^{(t)} > 0$ for $i \in \mathcal{S}_0^-$. Then it can be readily verified that

$$\Theta(1) \cdot \sum_{i\in\mathcal{S}_0^+} \ell_{1,i}^{(t)} + o\big(1/\mathrm{polylog}(n)\big) \cdot \sum_{i\in\mathcal{S}_0^-} \ell_{1,i}^{(t)} \geq \Theta(1) \cdot \sum_{i\in\mathcal{S}_0^-} \ell_{1,i}^{(t)} + o\big(1/\mathrm{polylog}(n)\big) \cdot \sum_{i\in\mathcal{S}_0^+} \ell_{1,i}^{(t)}.$$

Then we can get that

$$(\star) \geq 1 - \frac{O(\rho\eta b\alpha^2)}{1 + \frac{\eta}{n} \cdot \bigg(\Theta(1) \cdot \sum_{i\in\mathcal{S}_0^-} \ell_{1,i}^{(t)} + o\big(1/\mathrm{polylog}(n)\big) \cdot \sum_{i\in\mathcal{S}_0^+} \ell_{1,i}^{(t)} \pm O(\rho n b\alpha^2)\bigg)} \geq 1 - O(\rho\eta b\alpha^2).$$

Therefore we have for all $t \in [T_0, T_1']$,

$$\frac{|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle|}{|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle|} \geq \frac{|\langle \mathbf{w}_{1,r}^{(T_0)}, \mathbf{v}\rangle|}{|\langle \mathbf{w}_{1,r}^{(T_0)}, \mathbf{u}\rangle|} \cdot \big[1 - O(\rho\eta b\alpha^2)\big]^{T_1' - T_0} \geq \frac{|\langle \mathbf{w}_{1,r}^{(T_0)}, \mathbf{v}\rangle|}{|\langle \mathbf{w}_{1,r}^{(T_0)}, \mathbf{u}\rangle|} \cdot \big[1 - O(\rho\eta b\alpha^2)\big]^{O\big(\frac{1}{\rho\eta b\alpha^2}\big)}.$$

Then applying the fact that $\big[1 - O(\rho\eta b\alpha^2)\big]^{O\big(\frac{1}{\rho\eta b\alpha^2}\big)} \geq C$ holds for some absolute constant C, we are able to complete the proof for bounding $\frac{|\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u}\rangle|}{|\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}\rangle|}$. The results on $\frac{|\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u}\rangle|}{|\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}\rangle|}$ can be obtained similarly. □

In the next step, we will show that the learning of common features $\mathbf{v}$ and $\mathbf{u}$ will not be too large, i.e., exceeding the $\mathrm{polylog}(n)$ order.

Lemma B.5. Let T_1' be the same quantity defined in Lemma B.4, we have for all $t \in [T_0, T_1']$, it holds that

$$|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle|, |\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{u}\rangle| \leq O\big(\mathrm{polylog}(n)\big).$$

Proof of Lemma B.5. Based on the update rules in (B.8) and (B.9), we have

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle = \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle \cdot \bigg[1 + \frac{\eta}{n} \cdot \bigg(\Theta(1) \cdot \sum_{i\in\mathcal{S}_0^+} \ell_{1,i}^{(t)} + o\big(1/\mathrm{polylog}(n)\big) \cdot \sum_{i\in\mathcal{S}_0^-} \ell_{1,i}^{(t)} \pm O(\rho n b\alpha^2)\bigg)\bigg].$$

Using the fact that $\ell_{1,i}^{(t)} < 0$ for all $i \in \mathcal{S}_0^-$, we further have

$$(\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle)^2 \leq (\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle)^2 \cdot \bigg[1 + \frac{\eta}{n} \cdot \bigg(\Theta(1) \cdot \sum_{i \in \mathcal{S}_0^+} \ell_{1,i}^{(t)} + O(\rho n b \alpha^2)\bigg)\bigg]^2$$
$$= (\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle)^2 \cdot \bigg[1 + \frac{\eta}{n} \cdot \bigg(\Theta(1) \cdot \sum_{i \in \mathcal{S}_0^+} \ell_{1,i}^{(t)} + O(\rho n b \alpha^2)\bigg)\bigg],$$

where the second equality holds since $(1 + o(1))^2 = 1 + o(1)$. Further take a summation over $r \in [m]$ leads to

$$\sum_{r=1}^m (\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle)^2 \leq \bigg[\sum_{r=1}^m (\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle)^2\bigg] \cdot \bigg[1 + \frac{\eta}{n} \cdot \bigg(\Theta(1) \cdot \sum_{i \in \mathcal{S}_0^+} \ell_{1,i}^{(t)} + O(\rho n b \alpha^2)\bigg)\bigg]. \tag{B.10}$$

Similarly, we can also get that

$$\sum_{r=1}^m (\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{u}\rangle)^2 \leq \bigg[\sum_{r=1}^m (\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle)^2\bigg] \cdot \bigg[1 + \frac{\eta}{n} \cdot \bigg(\Theta(1) \cdot \sum_{i \in \mathcal{S}_0^-} \ell_{1,i}^{(t)} + O(\rho n b \alpha^2)\bigg)\bigg]. \tag{B.11}$$

Regarding the loss derivative $\ell_{1,i}^{(t)}$, we can get that for any $i \in \mathcal{S}_0^+$,

$$\ell_{1,i}^{(t)} = 1 - \mathrm{Logit}_1(\mathbf{W}^{(t)}; \mathbf{x}_i) = \frac{\exp\big[F_2(\mathbf{W}^{(t)}; \mathbf{x}_i) - F_1(\mathbf{W}^{(t)}; \mathbf{x}_i)\big]}{1 + \exp\big[F_2(\mathbf{W}^{(t)}; \mathbf{x}_i) - F_1(\mathbf{W}^{(t)}; \mathbf{x}_i)\big]} \leq \exp\big[F_2(\mathbf{W}^{(t)}; \mathbf{x}_i) - F_1(\mathbf{W}^{(t)}; \mathbf{x}_i)\big] \tag{B.12}$$

Before moving to the analysis on the feature, we first show that the model weight corresponding to the wrong label will not learn the noise of the data, i.e., $|\langle \mathbf{w}_{2,r}^{(t)}, \boldsymbol{\xi}_s^{(q)}\rangle|$ will be very small for all $q \in [P]$ and $s \in \mathcal{S}_0^+$. Particularly, we have the following by (B.2)

$$\max_{r,s} |\langle \mathbf{w}_{2,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)}\rangle| \leq \max_{r,s} |\langle \mathbf{w}_{2,r}^{(t)}, \boldsymbol{\xi}_s^{(q)}\rangle| \cdot \bigg[1 + \frac{\eta}{n} \cdot \ell_{2,s}^{(t)} \cdot \widetilde{\Theta}(d\sigma_p^2) + \frac{\eta}{n} \cdot \sum_{i \neq s || p \neq q} |\ell_{2,s}^{(t)}| \cdot \widetilde{O}\big(d^{1/2}\sigma_p^2\big)\bigg]$$
$$\leq \max_{r,s} |\langle \mathbf{w}_{2,r}^{(t)}, \boldsymbol{\xi}_s^{(q)}\rangle| \cdot \bigg[1 + \frac{\eta}{n} \cdot \widetilde{O}\big(nPd^{1/2}\sigma_p^2\big)\bigg],$$

where the second inequality is due to $|\ell_{k,i}^{(t)}| \leq 1$ and $\ell_{2,s}^{(t)} < 0$ for $s \in \mathcal{S}_0^+$. Therefore, we can get that for all $t \in \big[T_0, T_1'\big]$, where $T_1' \leq \widetilde{O}\big(1/(\eta P d^{1/2}\sigma_p^2)\big)$, that

$$\max_{r,s} |\langle \mathbf{w}_{2,r}^{(t)}, \boldsymbol{\xi}_s^{(q)}\rangle| \leq \max_{r,s} |\langle \mathbf{w}_{2,r}^{(T_0)}, \boldsymbol{\xi}_s^{(q)}\rangle| \cdot \bigg[1 + \widetilde{O}\big(\eta P d^{1/2}\sigma_p^2\big)\bigg]^{\widetilde{O}\big(\frac{1}{\eta P d^{1/2}\sigma_p^2}\big)}$$
$$\leq C \cdot \max_{r,s} |\langle \mathbf{w}_{2,r}^{(T_0)}, \boldsymbol{\xi}_s^{(q)}\rangle| = \widetilde{O}(d^{1/2}\sigma_p\sigma_0), \tag{B.13}$$

where the last equality is by Lemma B.2. Therefore, we can get the following bound on $F_2(\mathbf{W}^{(t)}; \mathbf{x}_i) - F_1(\mathbf{W}^{(t)}; \mathbf{x}_i)$ for any $i \in \mathcal{S}_0^+$,

$$F_2(\mathbf{W}^{(t)}; \mathbf{x}_i) - F_1(\mathbf{W}^{(t)}; \mathbf{x}_i) = \sum_{r=1}^m \sum_{p=1}^P (\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{x}_i^{(p)}\rangle)^2 - \sum_{r=1}^m \sum_{p=1}^P (\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{x}_i^{(p)}\rangle)^2$$
$$\leq \sum_{r=1}^m \sum_{p \in \mathcal{P}_i(\mathbf{v})} (\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}\rangle)^2 + \alpha^2 \sum_{r=1}^m \sum_{p \in \mathcal{P}_i(\mathbf{u})} (\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u}\rangle)^2$$
$$+ \sum_{r=1}^m \sum_{p \in \mathcal{P}_i(\boldsymbol{\xi})} (\langle \mathbf{w}_{2,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle)^2 - \sum_{r=1}^m (\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle)^2.$$

Then by Lemma B.4 and (B.13), we can further get that

$$F_2(\mathbf{W}^{(t)};\mathbf{x}_i) - F_1(\mathbf{W}^{(t)};\mathbf{x}_i) \leq O(b\alpha^2)\cdot\sum_{r=1}^{m}(\langle\mathbf{w}_{2,r}^{(t)},\mathbf{u}\rangle)^2 - \sum_{r=1}^{m}(\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle) + \widetilde{O}\big(mPd\sigma_p^2\sigma_0^2\big),$$

where we use the fact that $|\mathcal{P}_i(\mathbf{u})| \leq b$ and $(\langle\mathbf{w}_{2,r}^{(t)},\mathbf{v}\rangle/\langle\mathbf{w}_{2,r}^{(t)},\mathbf{u}\rangle)^2 = o(\alpha^2)$ by Lemma B.4. This further implies the following according to (B.12): for all $i\in\mathcal{S}_0^+$,

$$\begin{aligned}\ell_{1,i}^{(t)} &\leq \exp\big[F_2(\mathbf{W}^{(t)};\mathbf{x}_i) - F_1(\mathbf{W}^{(t)};\mathbf{x}_i)\big]\\ &\leq 2\exp\bigg[O(b\alpha^2)\cdot\sum_{r=1}^{m}(\langle\mathbf{w}_{2,r}^{(t)},\mathbf{u}\rangle)^2 - \sum_{r=1}^{m}(\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle)\bigg],\end{aligned}$$

where we use the fact that $mPd\sigma_p^2\sigma_0^2 = o(1)$. Similarly, we can also get that for all $i\in\mathcal{S}_0^-$,

$$\ell_{2,i}^{(t)} \leq 2\exp\bigg[O(b\alpha^2)\cdot\sum_{r=1}^{m}(\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle)^2 - \sum_{r=1}^{m}(\langle\mathbf{w}_{2,r}^{(t)},\mathbf{u}\rangle)\bigg].$$

Consequently, let $a_t := \sum_{r=1}^{m}(\langle\mathbf{w}_{1,r}^{(t+1)},\mathbf{v}\rangle)^2$ and $b_t := \sum_{r=1}^{m}(\langle\mathbf{w}_{2,r}^{(t+1)},\mathbf{u}\rangle)^2$, further applying (B.10) and (B.11) gives

$$\begin{aligned}a_{t+1} &\leq a_t\cdot\Big[1+\Theta(\eta)\cdot\exp\big[O(b\alpha^2)\cdot b_t - a_t\big] + O(\eta\rho b\alpha^2)\Big]\\ b_{t+1} &\leq b_t\cdot\Big[1+\Theta(\eta)\cdot\exp\big[O(b\alpha^2)\cdot a_t - b_t\big] + O(\eta\rho b\alpha^2)\Big].\end{aligned}$$

Then we will first prove a weaker argument on a_t and b_t: for all $t\leq T_1'$ it holds that $a_t, b_t = o\big(1/(b\alpha^2)\big)$. In particular, we will apply standard induction techniques. First, it is easy to verify that this condition holds for $t = T_0$ according to Lemma B.3. Then assuming this condition holds for all $\tau\leq t$, we have $\exp\big[O(b\alpha^2)\cdot b_t\big], \exp\big[O(b\alpha^2)\cdot a_t\big] = \Theta(1)$ and thus

$$\begin{aligned}a_{\tau+1} &\leq a_\tau\cdot\Big[1+\Theta(\eta)\cdot\exp(-a_\tau) + O(\eta\rho b\alpha^2)\Big],\\ b_{\tau+1} &\leq b_\tau\cdot\Big[1+\Theta(\eta)\cdot\exp(-b_\tau) + O(\eta\rho b\alpha^2)\Big],\end{aligned}\tag{B.14}$$

for all $\tau\in[T_0,t]$. Then by Lemma B.6, we can immediately get that

$$a_{t+1} \leq O\bigg(\log\bigg(\frac{1}{\rho b\alpha^2}\cdot e^{t\eta\rho b\alpha^2}\bigg)\bigg).$$

Then recall that $T_1' = O\big(1/(\eta\rho b\alpha^2)\big)$ and $t\leq T_1'$, we can further get $a_{t+1}, b_{t+1} = O\big(\log(\frac{1}{\rho b\alpha^2})\big)$, which verify the hypothesis that $a_{t+1}, b_{t+1} \leq o(1/(b\alpha^2))$. Moreover, recall the definitions of a_t and b_t: $a_t := \sum_{r=1}^{m}(\langle\mathbf{w}_{1,r}^{(t+1)},\mathbf{v}\rangle)^2$ and $b_t := \sum_{r=1}^{m}(\langle\mathbf{w}_{2,r}^{(t+1)},\mathbf{u}\rangle)^2$, we can further get that for all $r\in[m]$,

$$|\langle\mathbf{w}_{1,r}^{(t+1)},\mathbf{v}\rangle| = \widetilde{O}\bigg(\log^{1/2}\bigg(\frac{1}{\rho b\alpha^2}\bigg)\bigg) = O\big(\mathrm{polylog}(n)\big),$$

and $|\langle\mathbf{w}_{2,r}^{(t+1)},\mathbf{u}\rangle| = O\big(\mathrm{polylog}(n)\big)$. This completes the proof.

□

Lemma B.6. Let $\{a_t\}_{t=0,\ldots,}$ be a sequence with $a_0\in[0,1]$ that satisfies

$$a_{t+1} = a_t\cdot[1 + c_1\cdot e^{-a_t} + c_2],$$

where c_1 and c_2 are two constants satisfying $c_1, c_2\in[0,1]$ and $c_2\leq c_1$. Then it holds that

$$a_t \leq O\Big(\log(c_1/c_2)\cdot e^{2c_2 t}\Big).$$

Proof of Lemma B.6. Note that $c_2 \leq c_1$, we will then consider two cases: (1) $c_1 e^{-a_t} \geq c_2$ and (2) $c_1 e^{-a_t} < c_2$. Then case (2) will occur after case (1) since a_t is strictly increasing. Regarding case (1), it is easy to see that $a_t \leq \log(c_1/c_2)$ by the condition that $c_1 e^{-a_t} \geq c_2$. For case (2), let t_0 be the first iteration t that $c_1 e^{-a_t} < c_2$, we can get that $a_{t_0} = O\big(\log(c_1/c_2)\big)$ and then for all $t > t_0$

$$a_{t+1} \leq a_t \cdot [1 + 2c_2],$$

which implies that

$$a_t \leq a_{t_0} \cdot [1 + 2c_2]^{t - T_0} \leq O\Big(\log(c_1/c_2) \cdot e^{2c_2 t}\Big).$$

Combining the results for case (1) and case (2), we can complete the proof.

□

Then we will focus on the rare feature data. Note that in the early stage of the second phase, their corresponding loss derivatives $\ell_{k,i}^{(t)}$'s are still in the constant order. The following Lemma summarizes the learning of rare features and noises for the rare feature data.

Lemma B.7. Let $T_1 = O\big(\frac{n \log(1/(\sigma_0 d^{1/2} \sigma_p))}{d\sigma^2 \eta}\big)$ be a quantity that satisfies $T_0 < T_1 < T_1'$. Then for any $t \in [T_0, T_1]$, it holds that

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}' \rangle = \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}' \rangle \cdot \big[1 + \Theta(\rho\eta)\big],\ \langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{v}' \rangle = \langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}' \rangle \cdot \big[1 - \Theta(\rho\eta)\big]$$
$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}' \rangle = \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}' \rangle \cdot \big[1 - \Theta(\rho\eta)\big],\ \langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{u}' \rangle = \langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u}' \rangle \cdot \big[1 + \Theta(\rho\eta)\big]$$

Besides, for any $i \in \mathcal{S}_1^+ \cup \mathcal{S}_1^-$ and $k = y_s$, we have

$$\max_{r,p} |\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle| = \max_{r,p} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle| \cdot \bigg[1 + \frac{\eta}{n} \cdot \widetilde{\Theta}(d\sigma_p^2)\bigg];$$

for $k \neq y_s$,

$$\max_{r,p} |\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle| = \widetilde{O}(d^{-1/2} n).$$

Proof. The proof is similar to that of Lemma B.2, except the proof for the dynamics of $\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle$. First, by standard concentration argument, we can get with probability $1 - 1/\mathrm{poly}(n)$, for all $\boldsymbol{\xi}_i^{(p)} \in \{\boldsymbol{\xi}\}$, it holds that

$$d\sigma_p^2 - \widetilde{O}\big(d^{1/2}\sigma_p^2\big) \leq \|\boldsymbol{\xi}_i^{(p)}\|_2^2 \leq d\sigma_p^2 + \widetilde{O}\big(d^{1/2}\sigma_p^2\big).$$

Then by (B.2), we can get

$$\begin{aligned}
\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle &= \langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle \cdot \bigg[1 + \frac{2\eta}{n} \cdot \ell_{k,s}^{(t)} \cdot \|\boldsymbol{\xi}_s^{(q)}\|_2^2\bigg] \pm \frac{2\eta}{n} \cdot \sum_{i \neq s || p \neq q} |\ell_{k,i}^{(t)}| \cdot |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_i^{(q)} \rangle| \cdot \widetilde{O}\big(d^{1/2}\sigma_p^2\big) \\
&= \langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle \cdot \bigg[1 + \frac{2\eta}{n} \cdot \ell_{k,s}^{(t)} \cdot d\sigma_p^2\bigg] \pm \frac{2\eta}{n} \cdot \sum_{\boldsymbol{\xi}_i^{(q)} \in \{\boldsymbol{\xi}\}} |\ell_{k,i}^{(t)}| \cdot |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_i^{(q)} \rangle| \cdot \widetilde{O}\big(d^{1/2}\sigma_p^2\big).
\end{aligned}$$

Then let $\zeta = O\big(1/\mathrm{polylog}(n)\big)$ be some user-defined constant, then let T' be the smallest iteration number such that $\max_i |\ell_{k,i}^{(t)}| \in [0.5 - \zeta, 0.5 + \zeta]$. Then we can get for any $t \leq T'$ and any i, r,

$$\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle = \langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle \cdot \bigg[1 + \frac{2\eta}{n} \cdot (0.5 \pm \zeta) \cdot d\sigma_p^2\bigg] \pm \frac{2\eta}{n} \cdot \sum_{\boldsymbol{\xi}_i^{(q)} \in \{\boldsymbol{\xi}\}} |\ell_{k,i}^{(t)}| \cdot |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_i^{(q)} \rangle| \cdot \widetilde{O}\big(d^{1/2}\sigma_p^2\big). \tag{B.15}$$

Then we will prove the main arguments via mathematical induction, including the following hypothesis:

- For all $i \in \mathcal{S}_1^+ \cup \mathcal{S}_1^-$, it holds that $\max_{r,q} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_i^{(p)} \rangle| \geq \frac{1}{n^{0.1}} \cdot \max_{r,i',q} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_{i'}^{(q)} \rangle|$
- $\max_{r,q} |\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle| = \max_{r,q} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle| \cdot \Big[1 + \frac{2\eta}{n} \cdot (0.5 \pm 2\zeta) \cdot d\sigma_p^2\Big]$.

Then it is clear that the first argument holds for $t = T_0$ as with probability at least $1 - 1/\mathrm{poly}(n)$ we have $\max_{r,q} |\langle \mathbf{w}_{k,r}^{(T_0)}, \boldsymbol{\xi}_i^{(p)} \rangle| = \widetilde{\Theta}(\sigma_0 d^{1/2} \sigma_p)$ and $\max_{r,i',q} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_{i'}^{(q)} \rangle| = \widetilde{\Theta}(\sigma_0 d^{1/2} \sigma_p)$, which implies that $|\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_i^{(p)} \rangle| \geq \frac{1}{\mathrm{polylog}(n)} \cdot \max_{r,i',q} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_{i'}^{(q)} \rangle|$.

Besides, given the first argument, we have

$$\begin{aligned}
\sum_{\boldsymbol{\xi}_i^{(q)} \in \{\boldsymbol{\xi}\}} |\ell_{k,i}^{(t)}| \cdot |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_i^{(q)} \rangle| \cdot \widetilde{O}\big(d^{1/2}\sigma_p^2\big) &\leq \widetilde{O}(nPd^{1/2}\sigma_p^2) \cdot \max_{r,i',q} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_{i'}^{(q)} \rangle| \\
&\leq O(n^{1.1}Pd^{1/2}\sigma_p^2) \cdot \max_{r,q} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_i^{(p)} \rangle| \\
&\leq \zeta \cdot d\sigma_p^2 \cdot \max_{r,q} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_i^{(p)} \rangle|,
\end{aligned}$$

where we use the fact that $d^{-1/2}n^{1.1}P = o\big(1/\mathrm{polylog}(n)\big) = o(\zeta)$. Then by (B.15), we can directly obtain the second argument.

Now we will verify the hypotheses by induction. First, similar to the previous derivation, the first argument at the t-th iteration can directly imply the second argument at the $t+1$-th iteration. Then it remains to verify the first argument. In fact, given the second argument, we have for any i and i' and $\tau \leq t$,

$$\begin{aligned}
\frac{\max_{r,q} |\langle \mathbf{w}_{k,r}^{(\tau+1)}, \boldsymbol{\xi}_i^{(p)} \rangle|}{\max_{r,q} |\langle \mathbf{w}_{k,r}^{(\tau+1)}, \boldsymbol{\xi}_{i'}^{(p)} \rangle|} &\leq \frac{1 + \frac{2\eta}{n} \cdot (0.5 + 2\zeta) \cdot d\sigma_p^2}{1 + \frac{2\eta}{n} \cdot (0.5 - 2\zeta) \cdot d\sigma_p^2} \cdot \frac{\max_{r,q} |\langle \mathbf{w}_{k,r}^{(\tau)}, \boldsymbol{\xi}_i^{(p)} \rangle|}{\max_{r,q} |\langle \mathbf{w}_{k,r}^{(\tau)}, \boldsymbol{\xi}_{i'}^{(p)} \rangle|} \\
&\leq \bigg(1 + \frac{\eta\zeta d\sigma_p^2}{n}\bigg)^{T_1} \cdot \frac{\max_{r,q} |\langle \mathbf{w}_{k,r}^{(T_1)}, \boldsymbol{\xi}_i^{(p)} \rangle|}{\max_{r,q} |\langle \mathbf{w}_{k,r}^{(T_1)}, \boldsymbol{\xi}_{i'}^{(p)} \rangle|}.
\end{aligned}$$

Therefore, using the fact that $T_1 = O\big(\frac{n \log(1/(\sigma_0 d^{1/2}\sigma_p))}{d\sigma^2\eta}\big)$, setting $\zeta = 1/\log^2(1/(\sigma_0 d^{1/2}\sigma_p)$, we can directly get that

$$\frac{\max_{r,q} |\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_i^{(p)} \rangle|}{\max_{r,q} |\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_{i'}^{(p)} \rangle|} \leq \bigg(1 + \frac{\eta\zeta d\sigma_p^2}{n}\bigg)^{T_1} \cdot O(\mathrm{polylog}(n)) = o(n^{0.1}).$$

Note that the above holds for all i and i', taking $i' = \arg\max_i |\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_{i'}^{(p)} \rangle|$ directly completes the verification of the first argument.

The proof for $\max_{r,q} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle|$ with $k \neq y_s$, we have the following by (B.2),

$$\begin{aligned}
\max_{r,q} |\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle| &\leq \max_{r,q} |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle| + \frac{2\eta}{n} \cdot \sum_{i \neq s || p \neq q} |\ell_{k,i}^{(t)}| \cdot |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_i^{(q)} \rangle| \cdot \widetilde{O}\big(d^{1/2}\sigma_p^2\big) \\
&\leq \max_{r,q} |\langle \mathbf{w}_{k,r}^{(T_0)}, \boldsymbol{\xi}_s^{(q)} \rangle| + T_1 P\eta \cdot \widetilde{O}(d^{1/2}\sigma_p^2) \\
&= \widetilde{O}(d^{-1/2}n),
\end{aligned}$$

where we use the fact that for all $t \leq T_1$, it holds that $\max_{i,r,q} |\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle| = \widetilde{O}(1)$.

□

Lemma B.8 (End of Phase 2)**.** Let T_1 be the same quantity defined in Lemma B.7, we have for all $i \in \mathcal{S}_0^+ \cup \mathcal{S}_0^-$,

$$|\langle \mathbf{w}_{1,r}^{(T_1)}, \mathbf{v} \rangle|, |\langle \mathbf{w}_{2,r}^{(T_1)}, \mathbf{u} \rangle| = \widetilde{\Theta}(1), \quad |\langle \mathbf{w}_{2,r}^{(T_1)}, \mathbf{v} \rangle|, |\langle \mathbf{w}_{1,r}^{(T_1)}, \mathbf{u} \rangle| = \widetilde{O}(\sigma_0), \quad |\langle \mathbf{w}_{k,r}^{(T_1)}, \boldsymbol{\xi}_i^{(p)} \rangle| = \widetilde{O}\big(d^{-1/2}n\big);$$

for all $i \in \mathcal{S}_1^+ \cup \mathcal{S}_1^-$,

$$|\langle \mathbf{w}_{1,r}^{(T_1)}, \mathbf{v}' \rangle|, |\langle \mathbf{w}_{2,r}^{(T_1)}, \mathbf{u}' \rangle| = \widetilde{O}(\sigma_0), \quad |\langle \mathbf{w}_{2,r}^{(T_1)}, \mathbf{v}' \rangle|, |\langle \mathbf{w}_{1,r}^{(T_1)}, \mathbf{u}' \rangle| = \widetilde{O}(\sigma_0), \quad |\langle \mathbf{w}_{y_i,r}^{(T_1)}, \boldsymbol{\xi}_i^{(p)} \rangle| = \widetilde{\Theta}(1).$$

Proof. The proof of this lemma is simply a combination of Lemmas B.7 and B.7, where we only need to verify the bound for $|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{v}' \rangle|$ and $|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{u}' \rangle|$. This can be done as follows:

$$|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{v}' \rangle| \leq \big[1 + \Theta(\rho\eta)\big]^{T_1} \cdot |\langle \mathbf{w}_{k,r}^{(T_0)}, \mathbf{v}' \rangle| \leq \exp\big[\widetilde{O}(n\rho/(d\sigma^2))\big] \cdot |\langle \mathbf{w}_{k,r}^{(T_0)}, \mathbf{v}' \rangle| = \widetilde{O}(\sigma_0),$$

where we use the fact that $\rho n = o(d\sigma_p^2)$. The proof for $|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{v}' \rangle|$ will be similar and thus is ommited here.

□

Phase 3. Training until convergence. In this phase, we will show that the feature learning and noise learning in **Phase 2** will be maintained. Particularly, we first make the following hypothesis and then verify them via mathematical induction.

Hypothesis 1. For all $t = \mathrm{poly}(n)$ that is greater than T_1, it holds that

(a) We have $\sum_{r=1}^m (\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle)^2 = \widetilde{\Theta}(1)$ and $\sum_{r=1}^m \big(\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u} \rangle\big)^2 = \widetilde{\Theta}(1)$.

(b) We have $|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u} \rangle| = O\big(|\langle \mathbf{w}_{1,r}^{(T_1)}, \mathbf{u} \rangle|\big) = o\big(\frac{1}{\mathrm{polylog}(n)}\big)$ and $|\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v} \rangle| = O\big(|\langle \mathbf{w}_{2,r}^{(T_1)}, \mathbf{v} \rangle|\big) = o\big(\frac{1}{\mathrm{polylog}(n)}\big)$.

(c) We have $|\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{v}' \rangle| = O\big(|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{v}' \rangle|\big) = o\big(\frac{1}{\mathrm{polylog}(n)}\big)$ and $|\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{u}' \rangle| = O\big(|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{u}' \rangle|\big) = o\big(\frac{1}{\mathrm{polylog}(n)}\big)$.

(d) For all $i \in \mathcal{S}_0^+ \cup \mathcal{S}_0^-$, we have $|\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_i^{(p)} \rangle| = O\big(|\langle \mathbf{w}_{k,r}^{(T_1)}, \boldsymbol{\xi}_i^{(p)} \rangle|\big) = o\big(\frac{1}{\mathrm{polylog}(n)}\big)$.

(e) For all $i \in \mathcal{S}_1^+$, we have $\sum_{r=1}^m \sum_{p \in \mathcal{P}_i(\boldsymbol{\xi})} (\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)} \rangle)^2 = \widetilde{\Theta}(1)$; for all $i \in \mathcal{S}_1^-$, we have $\sum_{r=1}^m \sum_{p \in \mathcal{P}_i(\boldsymbol{\xi})} (\langle \mathbf{w}_{2,r}^{(t)}, \boldsymbol{\xi}_i^{(p)} \rangle)^2 = \widetilde{\Theta}(1)$.

(f) For all $i \in \mathcal{S}_1^+$, we have $|\langle \mathbf{w}_{2,r}^{(t)}, \boldsymbol{\xi}_i^{(p)} \rangle| = o\big(\frac{1}{\mathrm{polylog}(n)}\big)$; for all $i \in \mathcal{S}_1^-$, we have $|\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)} \rangle| = o\big(\frac{1}{\mathrm{polylog}(n)}\big)$.

The hypothesis will be verified via induction. First, it is clear that all hypothesis are satisfied at $t = T_1$ according to Lemma B.8. Then, the following lemma is useful in the entire proof.

Lemma B.9. Assuming all hypothesis in Hypothesis 1 hold for $\tau \in [0, t]$, then we have for all $k \in [2]$,

$$\sum_{\tau=T_1}^{t} \sum_{i \in \mathcal{S}_0^+ \cup \mathcal{S}_0^-} |\ell_{k,i}^{(\tau)}| = \widetilde{O}\bigg(\frac{n}{\eta}\bigg), \quad \text{and} \quad \sum_{\tau=T_1}^{t} \sum_{i \in \mathcal{S}_1^+ \cup \mathcal{S}_1^-} |\ell_{k,i}^{(\tau)}| \leq \widetilde{O}\bigg(\frac{\rho n^2}{d\sigma_p^2 \eta}\bigg),$$

moreover, for any $i \in \mathcal{S}_0^+ \cup \mathcal{S}_0^-$, we have

$$\sum_{t=T_1}^{t} |\ell_{k,i}^{(\tau)}| = \widetilde{O}(1/\eta).$$

Proof of Lemma B.9. By (B.1), we have

$$\begin{aligned}
\sum_{r=1}^m (\mathbf{w}_{1,r}^{(\tau+1)}, \mathbf{v})^2 &= \sum_{r=1}^m (\mathbf{w}_{1,r}^{(\tau)}, \mathbf{v})^2 \cdot \bigg[1 + \frac{2\eta}{n} \cdot \sum_{i \in [n]} \ell_{1,i}^{(\tau)} \sum_{p \in \mathcal{P}_i(\mathbf{v})} \alpha_{i,p}^2\bigg]^2 \\
&\geq \sum_{r=1}^m (\mathbf{w}_{1,r}^{(\tau)}, \mathbf{v})^2 \cdot \bigg[1 + \Theta\bigg(\frac{\eta}{n}\bigg) \cdot \sum_{i \in \mathcal{S}_0^+} |\ell_{1,i}^{(\tau)}| - \Theta\bigg(\frac{b\eta\alpha^2}{n}\bigg) \cdot \sum_{i \notin \mathcal{S}_0^+} |\ell_{1,i}^{(\tau)}|\bigg]; \\
\sum_{r=1}^m (\mathbf{w}_{2,r}^{(\tau+1)}, \mathbf{u})^2 &= \sum_{r=1}^m (\mathbf{w}_{2,r}^{(\tau)}, \mathbf{v})^2 \cdot \bigg[1 + \frac{2\eta}{n} \cdot \sum_{i \in [n]} \ell_{2,i}^{(\tau)} \sum_{p \in \mathcal{P}_i(\mathbf{v})} \langle \mathbf{w}_{2,r}^{(\tau)}, \mathbf{v} \rangle \cdot \alpha_{i,p}^2\bigg]^2
\end{aligned}$$

$$\geq \sum_{r=1}^{m} (\mathbf{w}_{2,r}^{(\tau)}, \mathbf{v})^2 \cdot \bigg[1 + \Theta\bigg(\frac{\eta}{n}\bigg) \cdot \sum_{i \in \mathcal{S}_0^-} |\ell_{2,i}^{(\tau)}| - \Theta\bigg(\frac{b\eta\alpha^2}{n}\bigg) \cdot \sum_{i \notin \mathcal{S}_0^-} |\ell_{2,i}^{(\tau)}|\bigg]. \tag{B.16}$$

where we use the fact that $|\ell_{1,i}| = |\ell_{2,i}|$ Summing them up and further taking a summation over $\tau \in [T_1, t-1]$, applying Hypothesis 1(a) gives

$$\Theta\bigg(\frac{\eta}{n}\bigg) \cdot \sum_{\tau=T_1}^{t-1} \sum_{i \in \mathcal{S}_0^+ \cup \mathcal{S}_0^-} |\ell_{1,i}^{(\tau)}| - \Theta\bigg(\frac{b\eta\alpha^2}{n}\bigg) \cdot \sum_{\tau=T_1}^{t-1} \sum_{i=1}^{n} |\ell_{1,i}^{(\tau)}| \leq \widetilde{O}(1). \tag{B.17}$$

where we use the fact that $|\ell_{1,i}^{(t)}| = |\ell_{2,i}^{(t)}|$ and $\sum_{r=1}^{m} (\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle)^2, \sum_{r=1}^{m} \big(\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u} \rangle\big)^2 = \widetilde{\Theta}(1)$. Besides, by (B.2) and Hypotheses (e) and (f), we know that the correct noise learning for different weak feature data will be different by at most $O(\mathrm{polylog}(n))$ factors, therefore, we can get that

$$\begin{aligned}
\sum_{i \in \mathcal{S}_1^+} \sum_{p \in [P]} (\langle \mathbf{w}_{1,r}^{(\tau+1)}, \boldsymbol{\xi}_i^{(p)} \rangle)^2 &\geq \sum_{i \in \mathcal{S}_1^+} \sum_{p \in [P]} (\langle \mathbf{w}_{k,r}^{(\tau)}, \boldsymbol{\xi}_i^{(p)} \rangle)^2 \cdot \bigg[1 + \frac{2\eta}{n} \cdot \ell_{1,i}^{(\tau)} \cdot \|\boldsymbol{\xi}_i^{(p)}\|_2^2 - \frac{2\eta}{n} \cdot \sum_{i=1}^{n} |\ell_{1,i}^{(\tau)}| \cdot \widetilde{O}\big(Pd^{1/2}\sigma_p^2\big)\bigg]^2 \\
&\geq \sum_{i \in \mathcal{S}_1^+} \sum_{p \in [P]} (\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_i^{(p)} \rangle)^2 \cdot \bigg[1 + \widetilde{\Theta}\bigg(\frac{\eta d\sigma_p^2}{n}\bigg) \cdot |\ell_{1,i}^{(\tau)}| - \widetilde{\Theta}\bigg(\frac{\eta d^{3/2}\sigma_p^4 P}{n}\bigg) \cdot \sum_{i=1}^{n} |\ell_{1,i}^{(\tau)}|\bigg],
\end{aligned} \tag{B.18}$$

and similarly,

$$\sum_{i \in \mathcal{S}_1^-} \sum_{p \in [P]} (\langle \mathbf{w}_{2,r}^{(\tau+1)}, \boldsymbol{\xi}_i^{(p)} \rangle)^2 \geq \sum_{i \in \mathcal{S}_1^-} \sum_{p \in [P]} (\langle \mathbf{w}_{2,r}^{(\tau)}, \boldsymbol{\xi}_i^{(p)} \rangle)^2 \cdot \bigg[1 + \widetilde{\Theta}\bigg(\frac{\eta d\sigma_p^2}{n}\bigg) \cdot |\ell_{2,i}^{(\tau)}| - \widetilde{\Theta}\bigg(\frac{\eta d^{3/2}\sigma_p^4 P}{n}\bigg) \cdot \sum_{i=1}^{n} |\ell_{2,i}^{(\tau)}|\bigg].$$

Therefore, taking a summation over $r \in [m]$ and $\tau \in [T_1, t-1]$, and using the Hypothesis 1(e), we have

$$\Theta\bigg(\frac{\eta d\sigma_p^2}{n}\bigg) \cdot \sum_{\tau=T_1}^{t-1} \sum_{i \in \mathcal{S}_1^+ \cup \mathcal{S}_1^-} |\ell_{1,i}^{(\tau)}| - \Theta\bigg(\frac{\eta d^{3/2}\sigma_p^4 P}{n}\bigg) \cdot \sum_{\tau=T_1}^{t-1} \sum_{i=1}^{n} |\ell_{1,i}^{(\tau)}| \leq \widetilde{O}(|\mathcal{S}_1^+ \cup \mathcal{S}_1^-|) = \widetilde{O}(\rho n). \tag{B.19}$$

Combining (B.17) and (B.19) and using the fact that $d\sigma_p^2 = \omega(1)$ and $b\alpha^2 = \omega(d^{3/2}\sigma_p^4 P)$, we can get that

$$\Theta\bigg(\frac{\eta}{n}\bigg) \cdot \sum_{\tau=T_1}^{t-1} \sum_{i=1}^{n} |\ell_{1,i}^{(\tau)}| - \Theta\bigg(\frac{b\eta\alpha^2}{n}\bigg) \cdot \sum_{\tau=T_1}^{t-1} \sum_{i=1}^{n} |\ell_{1,i}^{(\tau)}| \leq \widetilde{O}(\rho n).$$

Note that $b\alpha^2 = o(1)$, the above inequality immediately implies that

$$\sum_{\tau=T_1}^{t-1} \sum_{i=1}^{n} |\ell_{1,i}^{(\tau)}| \leq \widetilde{O}\bigg(\frac{\rho n^2}{\eta}\bigg).$$

We will further use this argument to sharpen our result. First, (B.19) directly leads to

$$\Theta\bigg(\frac{\eta d\sigma_p^2}{n}\bigg) \cdot \sum_{\tau=T_1}^{t-1} \sum_{i \in \mathcal{S}_1^+ \cup \mathcal{S}_1^-} |\ell_{1,i}^{(\tau)}| \leq \widetilde{O}(\rho n) + \widetilde{O}\big(d^{3/2}\sigma_p^4 P \rho n\big) = \widetilde{O}(\rho n),$$

which implies that

$$\sum_{\tau=T_1}^{t-1} \sum_{i \in \mathcal{S}_1^+ \cup \mathcal{S}_1^-} |\ell_{1,i}^{(\tau)}| \leq \widetilde{O}\bigg(\frac{\rho n^2}{d\sigma_p^2 \eta}\bigg).$$

Plugging the above inequality into (B.19) and using the fact that $b\alpha^2 = o(1)$ gives

$$\Theta\bigg(\frac{\eta}{n}\bigg) \cdot \sum_{\tau=T_1}^{t-1} \sum_{i \in \mathcal{S}_0^+ \cup \mathcal{S}_0^-} |\ell_{1,i}^{(\tau)}| \leq \widetilde{O}(1) + \Theta\bigg(\frac{b\eta\alpha^2}{n}\bigg) \cdot \sum_{\tau=T_1}^{t-1} \sum_{i \in \mathcal{S}_1^+ \cup \mathcal{S}_1^-} |\ell_{1,i}^{(\tau)}| \leq \widetilde{O}(1) + \widetilde{O}\bigg(\frac{\rho n b \alpha^2}{d\sigma_p^2}\bigg) = \widetilde{O}(1).$$

where the last inequality is due to $\rho n = o(d\sigma_p^2)$. Further note that $|\ell_{1,i}^{(t)}| \leq 1$ and $\eta = o(1)$, we have

$$\sum_{\tau=T_1}^{t} \sum_{i \in \mathcal{S}_1^+ \cup \mathcal{S}_1^-} |\ell_{1,i}^{(\tau)}| \leq \widetilde{O}\bigg(\frac{\rho n^2}{d\sigma_p^2 \eta}\bigg) + O(\rho n) = \widetilde{O}\bigg(\frac{\rho n^2}{d\sigma_p^2 \eta}\bigg);$$
$$\sum_{\tau=T_1}^{t} \sum_{i \in \mathcal{S}_0^+ \cup \mathcal{S}_0^-} |\ell_{1,i}^{(\tau)}| \leq \widetilde{O}\bigg(\frac{n}{\eta}\bigg) + O(n) = \widetilde{O}\bigg(\frac{n}{\eta}\bigg).$$

Moreover, by Hypothesis 1 for all $\tau \in [T_1, t]$, we also have for all $i \in \mathcal{S}_0^+$,

$$|\ell_{1,i}^{(\tau)}| = \frac{\exp\big[F_2(\mathbf{W}^{(\tau)}; \mathbf{x}_i) - F_1(\mathbf{W}^{(\tau)}; \mathbf{x}_i)\big]}{1 + \exp\big[F_2(\mathbf{W}^{(\tau)}; \mathbf{x}_i) - F_1(\mathbf{W}^{(\tau)}; \mathbf{x}_i)\big]}.$$

Moreover, we have

$$\begin{aligned} F_2(\mathbf{W}^{(\tau)}; \mathbf{x}_i) - F_1(\mathbf{W}^{(\tau)}; \mathbf{x}_i) &= \sum_{r=1}^{m} \sum_{p \in [P]} \langle \mathbf{w}_{2,r}^{(\tau)}, \mathbf{x}_i^{(p)} \rangle - \sum_{r=1}^{m} \sum_{p \in [P]} \langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{x}_i^{(p)} \rangle \\ &= -\sum_{r=1}^{m} \sum_{p \in \mathcal{P}_i(\mathbf{v})} (\langle \mathbf{w}_{k,r}^{(\tau)}, \mathbf{v} \rangle)^2 \pm o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg) \leq 0. \end{aligned}$$

This implies that for any $i, j \in \mathcal{S}_0^+$ with $|\mathcal{P}_i(\mathbf{v})| = |\mathcal{P}_j(\mathbf{v})|$, we have

$$\frac{|\ell_{1,i}^{(\tau)}|}{|\ell_{1,j}^{(\tau)}|} = \Theta\bigg(\frac{\exp\big[F_2(\mathbf{W}^{(\tau)}; \mathbf{x}_i) - F_1(\mathbf{W}^{(\tau)}; \mathbf{x}_i)\big]}{\exp\big[F_2(\mathbf{W}^{(\tau)}; \mathbf{x}_j) - F_1(\mathbf{W}^{(\tau)}; \mathbf{x}_j)\big]}\bigg) = \Theta\big(\exp[o(1/\mathrm{polylog}(n))]\big) = \Theta(1).$$

Further note that, by Definition 3.1, the number of feature patches are uniformly sampled from $[1, \Theta(1)]$, implying that with probability at least $1 - 1/\mathrm{poly}(n)$, for any $i \in \mathcal{S}_0^+$,

$$\#\big\{j : j \in \mathcal{S}_0^+, |\mathcal{P}_j(\mathbf{v})| = |\mathcal{P}_i(\mathbf{v})|\big\} = \Theta(n)$$

Therefore, let $\mathcal{S}'$ be the above set of data points, we have for any $s \in \mathcal{S}_0^+$ or $s \in \mathcal{S}_1^+$,

$$\sum_{\tau=T_1}^{t} |\ell_{1,s}^{(\tau)}| = \Theta\big(|\mathcal{S}'|^{-1}\big) \sum_{\tau=T_1}^{t} \sum_{i \in \mathcal{S}'} |\ell_{1,i}^{(\tau)}| \leq \Theta\big(|\mathcal{S}'|^{-1}\big) \sum_{\tau=T_1}^{t} \sum_{i \in \mathcal{S}_0^+ \cup \mathcal{S}_0^-} |\ell_{1,i}^{(\tau)}| \leq \widetilde{O}\bigg(\frac{1}{\eta}\bigg).$$

where the last inequality is due to $|\mathcal{S}'| = \Theta(n)$. This completes the proof. □

We will then verify Hypothesis 1(c), which is summarized in the following lemma.

Lemma B.10. Let Hypothesis 1 holds for all $\tau \leq t$, then we have $|\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{v}' \rangle| = O\big(|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{v}' \rangle|\big)$ and $|\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{u}' \rangle| = O\big(|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{u}' \rangle|\big)$.

Proof of Lemma B.10. Recall the update of rare features in (B.1), we have

$$\langle \mathbf{w}_{k,r}^{(\tau+1)}, \mathbf{v}' \rangle = \langle \mathbf{w}_{k,r}^{(\tau)}, \mathbf{v}' \rangle + \frac{2\eta}{n} \cdot \sum_{i \in \mathcal{S}_1^+} \ell_{k,i}^{(\tau)} \sum_{p \in \mathcal{P}_i(\mathbf{v}')} \langle \mathbf{w}_{k,r}^{(\tau)}, \mathbf{v}' \rangle$$

$$\langle \mathbf{w}_{k,r}^{(\tau+1)}, \mathbf{u}' \rangle = \langle \mathbf{w}_{k,r}^{(\tau)}, \mathbf{u}' \rangle + \frac{2\eta}{n} \cdot \sum_{i \in \mathcal{S}_1^-} \ell_{k,i}^{(\tau)} \sum_{p \in \mathcal{P}_i(\mathbf{u}')} \langle \mathbf{w}_{k,r}^{(\tau)}, \mathbf{u}' \rangle.$$

Then according to the Hypothesis 1(c) for all $\tau \in [T_1, t]$, we have

$$\begin{aligned}
|\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{v}' \rangle| &\leq |\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{v}' \rangle| + \frac{2\eta}{n} \cdot \sum_{\tau=T_1}^{t} \sum_{i \in \mathcal{S}_1^+} |\ell_{k,i}^{(\tau)}| \sum_{p \in \mathcal{P}_i(\mathbf{v}')} |\langle \mathbf{w}_{k,r}^{(\tau)}, \mathbf{v}' \rangle| \\
&\leq O\big(|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{v}' \rangle|\big) + \frac{\eta}{n} \cdot \sum_{\tau=T_1}^{t} \sum_{i \in \mathcal{S}_1^+} |\ell_{k,i}^{(\tau)}| \cdot O\big(|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{v}' \rangle|\big),
\end{aligned}$$

where the last inequality is due to the fact that $|\mathcal{P}_i(\mathbf{v}')| = \Theta(1)$. By Lemma B.9, it is clear that $\sum_{\tau=T_1}^{t} \sum_{i \in \mathcal{S}_1^+} |\ell_{k,i}^{(\tau)}| = \widetilde{O}\big(\frac{\rho n^2}{d\sigma_p^2 \eta}\big)$. Therefore,

$$|\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{v}' \rangle| \leq O\big(|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{v}' \rangle|\big) + \widetilde{O}\bigg(\frac{\rho n}{d\sigma^2}\bigg) \cdot O\big(|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{v}' \rangle|\big) = O\big(|\langle \mathbf{w}_{k,r}^{(T_1)}, \mathbf{v}' \rangle|\big).$$

The proof for $|\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{u}' \rangle|$ is similar so we omit it here. □

Using the similar proof technique, we are able to verify Hypothesis 1(b), 1(d), and 1(f), which are summarized in the following lemmas.

Lemma B.11. Let Hypothesis 1 holds for all $\tau \leq t$, then we have $|\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{v} \rangle| = O\big(|\langle \mathbf{w}_{2,r}^{(T_1)}, \mathbf{v} \rangle|\big)$ and $|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u} \rangle| = O\big(|\langle \mathbf{w}_{1,r}^{(T_1)}, \mathbf{u} \rangle|\big)$.

Proof of Lemma B.11. Since the proofs for $|\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{v} \rangle|$ and $|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u} \rangle|$ are basically identical, we will only provide the proof regarding $|\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{v} \rangle|$. By (B.1) and data distribution in Definition 3.1, we have

$$\begin{aligned}
\langle \mathbf{w}_{2,r}^{(\tau+1)}, \mathbf{v} \rangle &= \langle \mathbf{w}_{2,r}^{(\tau)}, \mathbf{v} \rangle \cdot \bigg[1 + \frac{2\eta}{n} \cdot \sum_{i \in \mathcal{S}_0^+} \ell_{2,i}^{(\tau)} \sum_{p \in \mathcal{P}_i(\mathbf{v})} \alpha_{i,p}^2 \|\mathbf{v}\|_2^2 + \frac{2\eta}{n} \cdot \sum_{i \in [n] \setminus \mathcal{S}_0^+} \ell_{2,i}^{(\tau)} \sum_{p \in \mathcal{P}_i(\mathbf{v})} \alpha_{i,p}^2 \|\mathbf{v}\|_2^2\bigg] \\
&\leq \langle \mathbf{w}_{2,r}^{(\tau)}, \mathbf{v} \rangle + \frac{2\eta b \alpha^2}{n} \cdot \langle \mathbf{w}_{2,r}^{(\tau)}, \mathbf{v} \rangle \cdot \sum_{i=1}^{n} |\ell_{2,i}^{(\tau)}|.
\end{aligned}$$

Taking an absolute value on both sides and then applying Hypothesis 1(b), we have

$$\begin{aligned}
|\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{v} \rangle| &\leq |\langle \mathbf{w}_{2,r}^{(T_1)}, \mathbf{v} \rangle| + \frac{2\eta b \alpha^2}{n} \cdot \sum_{\tau=T_1}^{t} \sum_{i=1}^{n} |\ell_{2,i}^{(\tau)}| |\langle \mathbf{w}_{2,r}^{(\tau)}, \mathbf{v} \rangle| \\
&\leq |\langle \mathbf{w}_{2,r}^{(T_1)}, \mathbf{v} \rangle| + O\big(|\langle \mathbf{w}_{2,r}^{(T_1)}, \mathbf{v} \rangle|\big) \cdot O\bigg(\frac{\eta b \alpha^2}{n}\bigg) \cdot \widetilde{O}\bigg(\frac{\rho n^2}{d\sigma_p^2 \eta} + \frac{n}{\eta}\bigg) \\
&= O\big(|\langle \mathbf{w}_{2,r}^{(T_1)}, \mathbf{v} \rangle|\big),
\end{aligned}$$

where the second inequality is by Lemma B.9 and the last inequality is due to the fact that $\rho n = o(d\sigma_p^2)$ and $b\alpha^2 = o(1)$. This completes the proof.

□

Lemma B.12. Let Hypothesis 1 holds for all $\tau \leq t$, then we have $|\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle| = o\big(1/\mathrm{polylog}(n)\big)$ for all $s \in \mathcal{S}_0^+ \cup \mathcal{S}_0^-$, $r \in [m]$, $k \in [2]$, and $q \in [P]$.

Proof of Lemma B.12. By (B.2), we have

$$
\begin{aligned}
|\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle| &\leq |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle| + |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle| \cdot \widetilde{O}\bigg(\frac{\eta d \sigma_p^2}{n}\bigg) \cdot |\ell_{k,s}^{(t)}| + \widetilde{O}\bigg(\frac{d^{1/2}\sigma_p^2 \eta}{n}\bigg) \cdot \sum_{p\in[P]} \sum_{i=1}^n |\ell_{k,i}^{(t)}| \cdot |\langle \mathbf{w}_{k,r}^{(t)}, \boldsymbol{\xi}_i^{(p)} \rangle| \\
&\leq |\langle \mathbf{w}_{k,r}^{(T_1)}, \boldsymbol{\xi}_s^{(q)} \rangle| + \sum_{\tau=T_1}^t |\langle \mathbf{w}_{k,r}^{(\tau)}, \boldsymbol{\xi}_s^{(q)} \rangle| \cdot \widetilde{O}\bigg(\frac{\eta d \sigma_p^2}{n}\bigg) \cdot |\ell_{k,s}^{(t)}| \\
&\quad + \widetilde{O}\bigg(\frac{d^{1/2}\sigma_p^2 \eta}{n}\bigg) \cdot \sum_{\tau=T_1}^t \sum_{p\in[P]} \sum_{i=1}^n |\ell_{k,i}^{(\tau)}| \cdot |\langle \mathbf{w}_{k,r}^{(\tau)}, \boldsymbol{\xi}_i^{(p)} \rangle|.
\end{aligned}
$$

Then by Hypotheses 1, we can further get

$$
|\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle| \leq |\langle \mathbf{w}_{k,r}^{(T_1)}, \boldsymbol{\xi}_s^{(q)} \rangle| + o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg) \cdot \widetilde{O}\bigg(\frac{\eta d \sigma_p^2}{n}\bigg) \cdot \sum_{\tau=T_1}^t |\ell_{k,s}^{(t)}| + \widetilde{O}\bigg(\frac{P d^{1/2}\sigma_p^2 \eta}{n}\bigg) \cdot \sum_{\tau=T_1}^t \sum_{i=1}^n |\ell_{k,i}^{(\tau)}|. \tag{B.20}
$$

Note that $s \in \mathcal{S}_0^+ \cup \mathcal{S}_0^-$, then by Lemma B.9, we have

$$
\sum_{t=T_1}^t |\ell_{k,s}^{(\tau)}| = \widetilde{O}\bigg(\frac{1}{\eta}\bigg), \quad \sum_{\tau=T_1}^t \sum_{i=1}^n |\ell_{k,i}^{(\tau)}| \leq \widetilde{O}\bigg(\frac{n}{\eta}\bigg).
$$

Therefore, plugging the above inequalities into (B.20) gives

$$
\begin{aligned}
|\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle| &\leq |\langle \mathbf{w}_{k,r}^{(T_1)}, \boldsymbol{\xi}_s^{(q)} \rangle| + o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg) \cdot \widetilde{O}\bigg(\frac{\eta d \sigma_p^2}{n}\bigg) \cdot \widetilde{O}\bigg(\frac{1}{\eta}\bigg) + \widetilde{O}\bigg(\frac{P d^{1/2}\sigma_p^2 \eta}{n}\bigg) \cdot \widetilde{O}\bigg(\frac{n}{\eta}\bigg) \\
&= o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg) + o\bigg(\frac{d\sigma_p^2}{n}\bigg) + \widetilde{O}\big(d^{1/2}\sigma_p^2\big) \\
&= o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg),
\end{aligned}
$$

where we use the fact that $d\sigma_p^2 = o(n)$ and $d^{1/2}\sigma_p^2 = o(1/\mathrm{polylog}(n))$. This completes the proof. □

Lemma B.13. Let Hypothesis 1 holds for all $\tau \leq t$, then we have $|\langle \mathbf{w}_{k,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle| = O\big(|\langle \mathbf{w}_{2,r}^{(T_1)}, \boldsymbol{\xi}_s^{(q)} \rangle|\big)$ for all $s \in \mathcal{S}_1^+ \cup \mathcal{S}_1^-$, $r \in [m]$, $k \neq y_s$, and $q \in [P]$.

Proof of Lemma B.13. Similar to the previous proof, we will only prove the argument for $s \in \mathcal{S}_1^+$, the proof for $s \in \mathcal{S}_1^-$ can be performed using exactly the same analysis. By (B.2), we have for $s \in \mathcal{S}_1^+$

$$
\begin{aligned}
|\langle \mathbf{w}_{2,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)} \rangle| &\leq |\langle \mathbf{w}_{2,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle| - |\langle \mathbf{w}_{2,r}^{(t)}, \boldsymbol{\xi}_s^{(q)} \rangle| \cdot \widetilde{O}\bigg(\frac{\eta d \sigma_p^2}{n}\bigg) \cdot |\ell_{2,s}^{(t)}| + \widetilde{O}\bigg(\frac{d^{1/2}\sigma_p^2 \eta}{n}\bigg) \cdot \sum_{p\in[P]} \sum_{i=1}^n |\ell_{2,i}^{(t)}| \cdot |\langle \mathbf{w}_{2,r}^{(t)}, \boldsymbol{\xi}_i^{(P)} \rangle| \\
&\leq |\langle \mathbf{w}_{2,r}^{(T_1)}, \boldsymbol{\xi}_s^{(q)} \rangle| + \sum_{\tau=T_1}^t \widetilde{O}\bigg(\frac{P d^{1/2}\sigma_p^2 \eta}{n}\bigg) \cdot \sum_{i=1}^n |\ell_{k,i}^{(\tau)}| \\
&= |\langle \mathbf{w}_{2,r}^{(T_1)}, \boldsymbol{\xi}_s^{(q)} \rangle| + \widetilde{O}\bigg(\frac{P d^{1/2}\sigma_p^2 \eta}{n}\bigg) \cdot \sum_{\tau=T_1}^t \sum_{i=1}^n |\ell_{2,i}^{(\tau)}| \\
&= o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg),
\end{aligned}
$$

where the last inequality is by Lemma B.9. This completes the proof.

□

Finally, we will verify the common features learning (Hypothesis 1(a)) and noise learning for rare feature data (Hypothesis 1(e)).

Lemma B.14. Let Hypothesis 1 holds for all $\tau \leq t$, then we have $\sum_{r=1}^{m}(\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle)^2 = \widetilde{\Theta}(1)$ and $\sum_{r=1}^{m}(\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{u}\rangle)^2 = \widetilde{\Theta}(1)$.

Proof of Lemma B.14. We first prove the upper bound: $\sum_{r=1}^{m}(\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle)^2 \leq \widetilde{\Theta}(1)$. Particularly, by (B.8), (B.9) and Definition 3.1, we have

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle \leq \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle \cdot \bigg[1 + \Theta\bigg(\frac{\eta}{n}\bigg) \cdot \sum_{i\in\mathcal{S}_0^+} |\ell_{1,i}^{(t)}| + \Theta\bigg(\frac{b\alpha^2\eta}{n}\bigg) \cdot \sum_{i=1}^{n} |\ell_{1,i}^{(t)}|\bigg].$$

Therefore, we can get that

$$\sum_{r=1}^{m}(\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle)^2 \leq \sum_{r=1}^{m}(\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle)^2 \cdot \bigg[1 + \Theta\bigg(\frac{\eta}{n}\bigg) \cdot \sum_{i\in\mathcal{S}_0^+} |\ell_{1,i}^{(t)}| + \Theta\bigg(\frac{b\alpha^2\eta}{n}\bigg) \cdot \sum_{i=1}^{n} |\ell_{1,i}^{(t)}|\bigg], \tag{B.21}$$

where we use the fact that $|\ell_{1,i}^{(t)}| = |\ell_{2,i}^{(t)}|$. By Hypothesis 1, we have for all $\tau \leq t$ and i,

$$\ell_{1,i}^{(\tau)} = \frac{\exp\big[F_2(\mathbf{W}^{(\tau)}; \mathbf{x}_i) - F_1(\mathbf{W}^{(\tau)}; \mathbf{x}_i)}{1 + \exp\big[F_2(\mathbf{W}^{(\tau)}; \mathbf{x}_i) - F_1(\mathbf{W}^{(\tau)}; \mathbf{x}_i)\big]} = \exp\bigg[-\Theta\bigg(\sum_{r=1}^{m}(\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}\rangle)^2\bigg)\bigg].$$

Therefore, let $a_\tau := \sum_{r=1}^{m}(\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}\rangle)^2$, we have the following according to (B.21)

$$a_{\tau+1} \leq a_\tau \cdot \bigg[1 + \Theta(\eta) \cdot e^{-ca_\tau} + \Theta\bigg(\frac{b\alpha^2\eta}{n}\bigg) \cdot \sum_{i=1}^{n} |\ell_{1,i}^{(\tau)}|\bigg], \tag{B.22}$$

where c is an absolute positive constant. Let $T = \mathrm{polylog}(n)$ be the total iteration number, then we will show that $a_t \leq 3c^{-1}\log(T)$ for all $\tau \leq t$. Particularly, we will prove that either (1) $a_\tau < 2c^{-1}\log(T)$ or (2) $a_\tau > 2\log(T) > a_{\tau-1}$ but it will not reach $3\log(T)$ as τ increases before it becomes less than $2c^{-1}\log(T)$ again. The first case immediately implies that $a_\tau < 3c^{-1}\log(T)$, so we will only need to focus on case (2). In this case, we have $a_\tau \leq a_{\tau-1} + \Theta(\eta) \leq 2.1c^{-1}\log(T)$. Then before a_τ becomes less than $2c^{-1}\log(T)$, we have for any $\tau' \in [\tau, t]$ that

$$a_{\tau'} \leq a_\tau + \sum_{s=\tau}^{\tau'-1} a_s \cdot \bigg[\Theta(\eta) \cdot e^{-ca_s} + \Theta\bigg(\frac{b\alpha^2\eta}{n}\bigg) \cdot \sum_{i=1}^{n} |\ell_{1,i}^{(s)}|\bigg].$$

Note that $a_s \cdot e^{-ca_s} \leq 2c^{-1}\log(T)/T^2 \leq 0.1c^{-1}/T$ if $T = \omega(1)$, then using the fact that $\eta = o(1)$,

$$a_{\tau'} \leq a_\tau + \sum_{s=\tau}^{\tau'-1}\bigg[\frac{\eta}{T} + \Theta\bigg(\frac{b\alpha^2\eta}{n}\bigg) \cdot a_s \sum_{i=1}^{n} |\ell_{1,i}^{(s)}|\bigg] \leq 2.2c^{-1}\log(T) + \Theta\bigg(\frac{b\alpha^2\eta}{n}\bigg) \cdot \sum_{s=\tau}^{\tau'-1} a_s \sum_{i=1}^{n} |\ell_{1,i}^{(s)}|.$$

Then as long as $a_s < 10c^{-1}\log(T)$ for $s \in [\tau, \tau']$, we have the following according to Lemma B.9,

$$\Theta\bigg(\frac{b\alpha^2\eta}{n}\bigg) \cdot \sum_{s=\tau}^{\tau'-1} a_s \sum_{i=1}^{n} |\ell_{1,i}^{(s)}| = \widetilde{O}\bigg(b\alpha^2 + \frac{\rho n b\alpha^2}{d\sigma_p^2}\bigg) = o(1) \leq 0.1c^{-1}\log(T),$$

where we use the fact that $b\alpha^2 = o(1)$ and $d\sigma_p^2 = \omega(\rho n)$. Therefore, we can conclude that before α_τ' reaches $10c^{-1}\log(T)$, it must satisfy

$$a_{\tau'} \leq 2.3c^{-1}\log(T),$$

for any $\tau' \leq t$. This further implies that

$$a_{t+1} \leq a_t + \widetilde{O}(\eta) \leq 3c^{-1}\log(T) = O\big(\mathrm{polylog}(n)\big),$$

which completes the proof of $\sum_{r=1}^m (\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle)^2 = \widetilde{O}(1)$.

The next step is to show that $\sum_{r=1}^m (\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle)^2 = \widetilde{\Omega}(1)$. Similar to (B.22), we can get that

$$a_{\tau+1} \geq a_\tau \cdot \bigg[1 + \Theta(\eta) \cdot e^{-Ca_\tau} - \Theta\bigg(\frac{b\alpha^2\eta}{n}\bigg) \cdot \sum_{i=1}^n |\ell_{1,i}^{(\tau)}|\bigg],$$

where C is an absolute positive constant. In fact, we must have $a_\tau \geq \frac{1}{\mathrm{polylog}(n)}$ since otherwise,

$$a_{\tau+1} \geq a_\tau \cdot \bigg[1 + \Theta(\eta) \cdot e^{-Ca_\tau} - \Theta\bigg(\frac{b\alpha^2\eta}{n}\bigg) \cdot \sum_{i=1}^n |\ell_{1,i}^{(\tau)}|\bigg] \geq a_\tau \cdot \big[1 + \Theta(\eta)\big],$$

where the first inequality is due to $e^{-ca_\tau} = \Theta(1)$ if $a_\tau = O(1)$ and the second inequality is due to $|\ell_{1,i}^{(\tau)}| \leq 1$ and $b\alpha^2 = o(1)$. This implies that $a_{\tau+1}$ will keep increase, which will at least continue to the case that $a_\tau > 1$. This completes the proof that $a_{t+1} = \widetilde{\Omega}(1)$.

The proof for $\sum_{r=1}^m (\langle \mathbf{w}_{2,r}^{(t+1)}, \mathbf{u}\rangle)^2 = \widetilde{\Theta}(1)$ will be basically the same so we omit it here.

□

Lemma B.15. Let Hypothesis 1 holds for all $\tau \leq t$, then we have $\sum_{r=1}^m \sum_{q\in\mathcal{P}_s(\boldsymbol{\xi})} (\langle \mathbf{w}_{1,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)}\rangle)^2 = \widetilde{\Theta}(1)$ for all $s \in \mathcal{S}_1^+$, and $\sum_{r=1}^m \sum_{q\in\mathcal{P}_s(\boldsymbol{\xi})} (\langle \mathbf{w}_{2,r}^{(t+1)}, \boldsymbol{\xi}_i^{(q)}\rangle)^2 = \widetilde{\Theta}(1)$ for all $s \in \mathcal{S}_1^-$.

Proof of Lemma B.15. Note that $P, m = \Theta(\mathrm{polylog}(n))$, it suffices to prove that $\max_{q,r}(\langle \mathbf{w}_{1,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)}\rangle)^2 = \widetilde{\Theta}(1)$ for all $s \in \mathcal{S}_1^+$ and $\max_{q,r}(\langle \mathbf{w}_{2,r}^{(t+1)}, \boldsymbol{\xi}_s^{(q)}\rangle)^2 = \widetilde{\Theta}(1)$ all $s \in \mathcal{S}_1^-$. In the following proof we will only consider $s \in \mathcal{S}_1^+$ as the proof for $s \in \mathcal{S}_1^-$ will exactly the same.

By (B.2), we have for all $s \in \mathcal{S}_1^+$,

$$\langle \mathbf{w}_{1,r}^{(\tau+1)}, \boldsymbol{\xi}_s^{(q)}\rangle = \langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_s^{(q)}\rangle \cdot \bigg[1 + \frac{2\eta}{n} \cdot \ell_{1,s}^{(\tau)} \cdot \|\boldsymbol{\xi}_s^{(q)}\|_2^2\bigg] \pm \frac{2\eta}{n} \cdot \widetilde{O}\big(d^{1/2}\sigma_p^2\big) \cdot \sum_{i\neq s || p\neq q} |\ell_{k,i}^{(\tau)}| \cdot |\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_i^{(p)}\rangle|. \tag{B.23}$$

We first prove the upper bound of $\sum_{r=1}^m \sum_{q\in\mathcal{P}_s(\boldsymbol{\xi})} (\langle \mathbf{w}_{2,r}^{(\tau)}, \boldsymbol{\xi}_s^{(q)}\rangle)^2$. Then, using the Hypothesis 1 (e), we have for any $i \in [n]$, $s \in \mathcal{S}_1^+$, $r \in [m]$, and $p \in [P]$

$$(\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_i^{(p)}\rangle)^2 \leq O\bigg(\mathrm{polylog}(n)\bigg) \cdot O(mP) \cdot \max_{r,q}(\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_s^{(q)}\rangle)^2 = O\bigg(\mathrm{polylog}(n)\bigg) \cdot \max_{r,q}(\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_s^{(q)}\rangle)^2.$$

Then (B.23) implies that

$$\max_{r,q}(\langle \mathbf{w}_{1,r}^{(\tau+1)}, \boldsymbol{\xi}_s^{(q)}\rangle)^2 \leq \max_{r,q}(\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_s^{(q)}\rangle)^2 \cdot \bigg[1 + \Theta\bigg(\frac{\eta d\sigma_p^2}{n}\bigg) \cdot \ell_{1,s}^{(\tau)} + \widetilde{O}\bigg(\frac{\eta P d^{1/2}\sigma_p^2}{n}\bigg) \cdot \sum_{i=1}^n |\ell_{1,i}^{(\tau)}|\bigg].$$

Then by Hypothesis 1, we can further get that the quantity $\sum_{r=1}^m \sum_{q\in\mathcal{P}_s(\boldsymbol{\xi})} (\langle \mathbf{w}_{1,r}^{(\tau+1)}, \boldsymbol{\xi}_s^{(q)}\rangle)^2$ will be the dominating term in the neural network output function, so that $\ell_{1,s}^{(\tau)} \geq e^{-c\max_{r,q}(\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_s^{(q)}\rangle)^2}$ for some constant c. Therefore, let $a_\tau = \max_{r,q}(\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_s^{(q)}\rangle)^2$, we can follow the similar derivation of (B.22). Thus, it follows that

$$a_{\tau+1} \leq a_\tau \cdot \bigg[1 + \Theta\bigg(\frac{\eta d\sigma_p^2}{n}\bigg) \cdot e^{-ca_\tau} + \widetilde{O}\bigg(\frac{\eta P d^{1/2}\sigma_p^2}{n}\bigg) \cdot \sum_{i=1}^n |\ell_{1,i}^{(\tau)}|\bigg]$$

Then we can follow the exact proof technique in Lemma B.14 to conclude that $a_{t+1} = \widetilde{O}(1)$, while it only requires to verify that

$$\widetilde{O}\bigg(\frac{\eta P d^{1/2}\sigma_p^2}{n}\bigg) \cdot \sum_{\tau=T_1}^t \sum_{i=1}^n |\ell_{1,i}^{(\tau)}| = o(1),$$

which clearly holds by Lemma B.9 and the fact that $Pd^{1/2}\sigma_p^2 = o(1)$.

The lower bound can be similarly obtained as the following can be deduced by (B.23):

$$\max_{r,q}(\langle \mathbf{w}_{1,r}^{(\tau+1)}, \boldsymbol{\xi}_s^{(q)}\rangle)^2 \geq \max_{r,q}(\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_s^{(q)}\rangle)^2 \cdot \bigg[1 + \Theta\bigg(\frac{\eta d\sigma_p^2}{n}\bigg) \cdot \ell_{1,s}^{(\tau)} - \widetilde{O}\bigg(\frac{\eta P d^{1/2}\sigma_p^2}{n}\bigg) \cdot \sum_{i=1}^n |\ell_{1,i}^{(\tau)}|\bigg],$$

which leads to

$$a_{\tau+1} \geq a_\tau \cdot \bigg[1 + \Theta\bigg(\frac{\eta d\sigma_p^2}{n}\bigg) \cdot e^{-Ca_\tau} - \widetilde{O}\bigg(\frac{\eta P d^{1/2}\sigma_p^2}{n}\bigg) \cdot \sum_{i=1}^n |\ell_{1,i}^{(\tau)}|\bigg]$$

for some absolute constant C. Then following the same proof of Lemma B.14, we can get that $a_{t+1} = \widetilde{\Omega}(1)$. This completes the proof. □

B.3 Proof of Theorem 4.1

Proof of Theorem 4.1. We first show that $\|\mathbf{w}_{k,r}^{(T)}\|_2 = \widetilde{O}(n)$ for all $k \in [2]$ and $r \in [m]$. In particular, note that the update of standard training is always the linear combination of all critical vectors, i.e., $\mathbf{v}$, $\mathbf{u}$, $\mathbf{v}'$, $\mathbf{u}'$, and $\boldsymbol{\xi}_i^{(p)}$'s. Therefore, we have

$$\mathbf{w}_{k,r}^{(t)} = \mathbf{w}_{k,r}^{(0)} + \rho_{k,r}^{(t)}(\mathbf{v}) \cdot \mathbf{v} + \rho_{k,r}^{(t)}(\mathbf{u}) \cdot \mathbf{u} + \rho_{k,r}^{(t)}(\mathbf{v}') \cdot \mathbf{v}' + \rho_{k,r}^{(t)}(\mathbf{u}') \cdot \mathbf{u}' + \sum_{i=1}^n \sum_{p\in\mathcal{P}_i(\boldsymbol{\xi})} \rho_{k,r}^{(t)}(\boldsymbol{\xi}_i^{(p)}) \cdot \boldsymbol{\xi}_i^{(p)}.$$

Here we use $\rho_{k,r}^{(t)}(\mathbf{a})$ to denote the coefficient of $\mathbf{a}$ for all $\mathbf{a} \in \{\mathbf{v}, \mathbf{u}, \mathbf{v}', \mathbf{u}'\} \cup \{\boldsymbol{\xi}\}$. Then by Lemma 5.3 and using the fact that $\|\mathbf{v}\|_2, \|\mathbf{u}\|_2, \|\mathbf{v}'\|_2, \|\mathbf{u}'\|_2 = 1$, we have

$$|\rho_{k,r}^{(t)}(\mathbf{v})|, |\rho_{k,r}^{(t)}(\mathbf{u})| = \widetilde{O}(1), \quad |\rho_{k,r}^{(t)}(\mathbf{v}')|, |\rho_{k,r}^{(t)}(\mathbf{u}')| = o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg).$$

Moreover, using the fact that $|\langle \boldsymbol{\xi}_i^{(p)}, \boldsymbol{\xi}_j^{(q)}\rangle| = o(1/\mathrm{polylog}(n))$ for any $i \neq j$ or $p \neq q$, applying Lemma 5.3 and the fact that $\|\boldsymbol{\xi}_i^{(p)}\|_2^2 = \Omega(1)$ for all $i \in [n]$ and $p \in [P]$, we have

$$\bigg\|\sum_{i=1}^n \sum_{p\in\mathcal{P}_i(\boldsymbol{\xi})} \rho_{k,r}^{(t)}(\boldsymbol{\xi}_i^{(p)}) \cdot \boldsymbol{\xi}_i^{(p)}\bigg\|_2^2 \leq \widetilde{O}(n^2).$$

Combining the above results, we can readily conclude that $\|\mathbf{w}_{k,r}^{(t)}\|_2 = \widetilde{O}(n)$.

Then we will characterize the test errors for common feature data and rare feature data separately. Regarding the common feature data, we can take a positive common feature data $(\mathbf{x}, 1)$ as an example and obtain the following by Lemma 5.3,

$$F_1(\mathbf{W}^{(t)}; \mathbf{x}) = \sum_{r=1}^m \sum_{p=1}^P \big(\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{x}^{(p)}\rangle\big)^2 \geq \sum_{r=1}^m \sum_{p:\mathbf{x}^{(p)}=\mathbf{v}} \big(\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle\big)^2 = \widetilde{\Theta}(1). \tag{B.24}$$

Besides, we have the following regarding $F_2(\mathbf{W}^{(t)}; \mathbf{x})$:

$$\begin{aligned} F_2(\mathbf{W}^{(t)}; \mathbf{x}) &= \sum_{r=1}^m \sum_{p:\mathbf{x}^{(p)}=\mathbf{v}} \big(\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}\rangle\big)^2 + \sum_{r=1}^m \sum_{p:\mathbf{x}^{(p)}\neq\mathbf{v}} \big(\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{x}^{(p)}\rangle\big)^2 \\ &= \sum_{r=1}^m \sum_{p:\mathbf{x}^{(p)}\neq\mathbf{v}} \big(\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{x}^{(p)}\rangle\big)^2 + o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg). \end{aligned} \tag{B.25}$$

where we use the result $|\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}\rangle| = o\big(1/\mathrm{polylog}(n)\big)$. Then, note that if $\mathbf{x}^{(p)} \neq \mathbf{v}$, it can be either feature noise (i.e., $\alpha\mathbf{u}$ or $\alpha\mathbf{v}$) or random noise $\boldsymbol{\zeta}_i^{(p)}$, which is independent of the random noise vectors in the training data points (i.e., $\{\boldsymbol{\xi}\}$).

Therefore, using the result that $\|\mathbf{w}_{k,r}^{(t)}\|_2 = \widetilde{O}(n)$, we can obtain with probability at least $1 - \exp(-\Omega(d^{1/2}))$, it holds that for all $r \in [m]$

$$(\langle \mathbf{w}_{2,r}^{(t)}, \boldsymbol{\zeta}_i^{(p)} \rangle)^2 = \widetilde{O}(\sigma_p^2 n^2). \tag{B.26}$$

Besides, note that there are at most b patches within the total P patches that are feature noise, we have

$$\sum_{r=1}^{m} \sum_{p:\mathbf{x}^{(p)} \neq \mathbf{v}} \big(\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{x}^{(p)} \rangle\big)^2 \leq O(mb\alpha^2) + \widetilde{O}(mP\sigma_p^2 n^2) = o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg),$$

where the last equality is by the data model in Definition 3.1: $b\alpha^2 = o\big(1/\mathrm{polylog}(n)\big)$ and $\sigma_p = o(d^{-1/2}n^{1/2})$. Therefore, comparing (B.24) and (B.25), we can get $F_1(\mathbf{W}^{(t)}; \mathbf{x}) > F_2(\mathbf{W}^{(t)}; \mathbf{x})$ with probability at least $1 - 1/\mathrm{poly}(n)$.

Then we will move on to study the rare feature data. In particular, we consider the rare feature data with incorrect feature noise. Without loss of generality, we take a positive data $(\mathbf{x}, 1)$ as an example, which contains rare feature $\mathbf{v}$ and incorrect feature noise $\alpha\mathbf{u}$. Then we can get the following results for $F_k(\mathbf{W}^{(t)}; \mathbf{x})$

$$F_k(\mathbf{W}^{(t)}; \mathbf{x}) = \sum_{r=1}^{m} \sum_{p:\mathbf{x}^{(p)} = \mathbf{v}'} \big(\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{v}' \rangle\big)^2 + \sum_{r=1}^{m} \sum_{p:\mathbf{x}^{(p)} = \alpha\mathbf{u}} \big(\langle \mathbf{w}_{k,r}^{(t)}, \alpha\mathbf{u} \rangle\big)^2 + \sum_{r=1}^{m} \sum_{p:\mathbf{x}^{(p)} \notin \{\mathbf{v}', \alpha\mathbf{u}\}} \big(\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{x}^{(p)} \rangle\big)^2.$$

Note that if $\mathbf{x}^{(p)} \notin \{\mathbf{v}', \alpha\mathbf{u}\}$, then $\mathbf{x}^{(p)}$ must be a random noise vector that is independent of $\mathbf{w}_{k,r}^{(t)}$. To begin with, the first two terms of the above equation for different k's can be bounded by applying Lemma 5.3 (particularly $\sum_{r=1}^{m} (\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u} \rangle)^2 = \widetilde{\Omega}(1)$), we have

$$\sum_{r=1}^{m} \sum_{p:\mathbf{x}^{(p)} = \mathbf{v}'} \big(\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}' \rangle\big)^2 = \widetilde{O}(\sigma_0^2), \quad \sum_{r=1}^{m} \sum_{p:\mathbf{x}^{(p)} = \alpha\mathbf{u}} \big(\langle \mathbf{w}_{1,r}^{(t)}, \alpha\mathbf{u} \rangle\big)^2 = \widetilde{O}(b\alpha^2\sigma_0^2),$$
$$\sum_{r=1}^{m} \sum_{p:\mathbf{x}^{(p)} = \mathbf{v}'} \big(\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}' \rangle\big)^2 = \widetilde{O}(\sigma_0^2), \quad \sum_{r=1}^{m} \sum_{p:\mathbf{x}^{(p)} = \alpha\mathbf{u}} \big(\langle \mathbf{w}_{2,r}^{(t)}, \alpha\mathbf{u} \rangle\big)^2 = \widetilde{\Omega}(\alpha^2).$$

Moreover, by (B.26), we can further get that with probability at least $1 - \exp(-\Omega(d^{1/2})) > 1 - 1/\mathrm{poly}(n)$, we have

$$\sum_{r=1}^{m} \sum_{p:\mathbf{x}^{(p)} \notin \{\mathbf{v}', \alpha\mathbf{u}\}} \big(\langle \mathbf{w}_{k,r}^{(t)}, \mathbf{x}^{(p)} \rangle\big)^2 = \widetilde{O}(mP\sigma_p^2 n^2) = o(\alpha^2).$$

where the last equality is by our data model in Definition 3.1. This further implies that conditioning on $\mathbf{W}^{(t)}$, with probability at least $1 - 1/\mathrm{poly}(n)$, we have

$$F_2(\mathbf{W}^{(t)}; \mathbf{x}) > F_1(\mathbf{W}^{(t)}; \mathbf{x})$$

on the positive rare feature data that has incorrect feature noise.

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathrm{rare}}}[\operatorname*{argmax}_k F_k(\mathbf{W}^{(t)}; \mathbf{x}) \neq y] \geq \frac{1}{2} - \frac{1}{\mathrm{poly}(n)} \geq \frac{1}{2.01}.$$

Therefore, combining the test error analysis for common feature data and rare feature data and using the fact that the fraction of rare feature data is ρ, we can finally obtain:

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}[\operatorname*{argmax}_k F_k(\mathbf{W}^{(t)}; \mathbf{x}) \neq y] \geq \rho \cdot \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathrm{rare}}}[\operatorname*{argmax}_k F_k(\mathbf{W}^{(t)}; \mathbf{x}) \neq y] \geq \frac{\rho}{2.01}.$$

This completes the proof.

□

C Mixup data

C.1 Characterization of the mixup dataset

Category of different Mixup data patches. First recall the category of different Mixup training data points:

- Mix between two common feature data points, including $\mathcal{S}_{0,0}^{+,+}$, $\mathcal{S}_{0,0}^{-,-}$, $\mathcal{S}_{0,0}^{+,-}$, $\mathcal{S}_{0,0}^{-,+}$, each of them is of size $\Theta(n^2)$.
- Mix between common feature and rare feature data points with the same label, including $\mathcal{S}_{0,1}^{+,+}$, $\mathcal{S}_{0,1}^{-,-}$, $\mathcal{S}_{1,0}^{+,+}$, and $\mathcal{S}_{1,0}^{-,-}$, each of them is of size $\Theta(\rho n^2)$.
- Mix between common feature and rare feature data points with different labels, including $\mathcal{S}_{0,1}^{+,-}$, $\mathcal{S}_{0,1}^{-,+}$, $\mathcal{S}_{1,0}^{+,-}$, and $\mathcal{S}_{1,0}^{-,+}$, each of them is of size $\Theta(\rho n^2)$.
- Mix between two rare feature data points, including $\mathcal{S}_{1,1}^{+,+}$,$\mathcal{S}_{1,1}^{-,-}$, $\mathcal{S}_{1,1}^{+,-}$ and $\mathcal{S}_{1,1}^{-,+}$, each of them is of size $\Theta(\rho^2 n^2)$.

Then, given n^2 mixed data points, we have in total $n^2 P$ data patches. Besides, note that in the original dataset that consists of n training data points, each data patch $\mathbf{x}_i^{(p)}$ satisfies

$$\mathbf{x}_i^{(p)} \in \big\{\mathbf{v}, \mathbf{u}, \alpha\mathbf{u}, \alpha\mathbf{v}, \mathbf{v}', \mathbf{u}', \boldsymbol{\xi}_i^{(p)}\big\}.$$

Moreover, by the data distribution defined in Definition 3.1, we have

- $\mathbf{v}$ and $\mathbf{u}$ will appear in $\Theta(n)$ data and $\Theta(n)$ data patches.
- $\alpha\mathbf{v}$ and $\alpha\mathbf{u}$ will appear in n data and $\Theta(bn)$ data patches.
- $\mathbf{v}'$ and $\mathbf{u}'$ will appear in $\Theta(\rho n)$ data and $\Theta(\rho n)$ data patches.
- $\boldsymbol{\xi}_i^{(p)}$, if it is not zero, will appear in one data and one data patch.

Then based on the above facts, we provide the following lemma that characterizes the number of different types of data patches on the mixup dataset.

Lemma C.1. Let $\mathcal{P} := \{\mathbf{x}_{i,j}^{(p)}\}_{i,j\in[n],p\in[P]}$ be the collection of all data patches of the mixup dataset, then among these $n^2 P$ data patches, with probability at least $1 - 1/\mathrm{poly}(n)$, let $\mathbf{x}_{i,j}^{(p)} = \lambda\mathbf{a} + (1-\lambda)\boldsymbol{b}$, we have

- The vector with $\mathbf{a} \in \{\mathbf{v}, \mathbf{u}\}$ and $\boldsymbol{b} \in \{\mathbf{v}, \mathbf{u}\}$ will appear in $\Theta(n^2/P)$ data patches.
- The vector with $\mathbf{a} \in \{\mathbf{v}, \mathbf{u}\}$ and $\boldsymbol{b} \in \{\mathbf{v}', \mathbf{u}'\}$ will appear in $\Theta(\rho n^2/P)$ patches.
- The vector with $\mathbf{a} \in \{\mathbf{v}, \mathbf{u}\}$ and $\boldsymbol{b} \in \{\alpha\mathbf{v}, \alpha\mathbf{u}\}$ will appear in $O(bn^2/P)$ patches.
- The vector with $\mathbf{a} \in \{\mathbf{v}, \mathbf{u}\}$ and $\boldsymbol{b} \in \{\boldsymbol{\xi}\}$ will appear in $\Theta(n^2)$ patches.
- The vector with $\mathbf{a} \in \{\mathbf{v}', \alpha\mathbf{u}'\}$ and $\boldsymbol{b} \in \{\mathbf{v}', \mathbf{u}'\}$ will appear in $\Theta(\rho^2 n^2/P)$ data patches.
- The vector with $\mathbf{a} \in \{\mathbf{v}', \alpha\mathbf{u}'\}$ and $\boldsymbol{b} \in \{\alpha\mathbf{v}, \alpha\mathbf{u}\}$ will appear in $O(\rho bn^2/P)$ patches.
- The vector with $\mathbf{a} \in \{\mathbf{v}', \alpha\mathbf{u}'\}$ and $\boldsymbol{b} \in \{\boldsymbol{\xi}\}$ will appear in $\Theta(\rho n^2)$ patches.
- The vector with $\mathbf{a} \in \{\alpha\mathbf{v}, \alpha\mathbf{u}\}$ and $\boldsymbol{b} \in \{\alpha\mathbf{v}, \alpha\mathbf{u}\}$ will appear in $O(b^2 n^2/P)$ patches.
- The vector with $\mathbf{a} \in \{\alpha\mathbf{v}, \alpha\mathbf{u}\}$ and $\boldsymbol{b} \in \{\boldsymbol{\xi}\}$ will appear in $O(bn^2)$ patches.

Besides, regarding any non-zero noise vector $\boldsymbol{\xi}_i^{(p)}$, we have, among the collection of data patches $\{\mathbf{x}_{i,j}^{(p)}\}_{j\in[n]}$, with probability at least $1 - 1/\mathrm{poly}(n)$,

- $\mathbf{x}_{i,j}^{(p)} = \lambda \boldsymbol{\xi}_{i,j}^{(p)} + (1-\lambda)\boldsymbol{b}$ with $\boldsymbol{b} \in \{\mathbf{v}, \mathbf{u}\}$ will appear in $\Theta(n/P)$ patches.
- $\mathbf{x}_{i,j}^{(p)} = \lambda \boldsymbol{\xi}_{i,j}^{(p)} + (1-\lambda)\boldsymbol{b}$ with $\boldsymbol{b} \in \{\alpha\mathbf{v}, \alpha\mathbf{u}\}$ will appear in $O(bn/P)$ patches.
- $\mathbf{x}_{i,j}^{(p)} = \lambda \boldsymbol{\xi}_{i,j}^{(p)} + (1-\lambda)\boldsymbol{b}$ with $\boldsymbol{b} \in \{\mathbf{v}', \mathbf{u}'\}$ will appear in $\Theta(\rho n/P)$ patches.
- $\mathbf{x}_{i,j}^{(p)} = \lambda \boldsymbol{\xi}_{i,j}^{(p)} + (1-\lambda)\boldsymbol{b}$ with $\boldsymbol{b} \in \{\boldsymbol{\xi}\}$ will appear in $\Theta(n)$ patches.

Proof of Lemma C.1. We first consider a fixed $\mathbf{x}_i$ and the corresponding collection of data patches $\{\mathbf{x}_{i,j}^{(p)}\}_{j\in[n],p\in[P]}$. Then by Definition 3.1, conditioning on $\mathbf{x}_i^{(p)} = \mathbf{v}$, we have for any $j \neq i$

$$\mathbb{P}[\mathbf{x}_j^{(p)} = \mathbf{v} | \mathbf{x}_i^{(p)} = \mathbf{v}] = \mathbb{P}[\mathbf{x}_j^{(p)} = \mathbf{v}] = \Theta\bigg(\frac{1}{P}\bigg).$$

Therefore, we can further get that conditioning on $\mathbf{x}_i^{(p)} = \mathbf{v}$, the summation $\sum_{j\neq i} \mathbb{1}[\mathbf{x}_{i,j}^{(p)} = \mathbf{v}]$ follows Binomial distribution $\mathrm{Binom}(n-1, p)$ with probability parameter $p = \Theta(1/P)$. Then by Hoeffding's inequality, we can get that with probability at least $1 - \exp(-n^2/P^2)$, it holds that

$$\sum_{j\in[n]} \mathbb{1}[\mathbf{x}_j^{(p)} = \mathbf{v} | \mathbf{x}_i^{(p)} = \mathbf{v}] = \Theta\bigg(\frac{n}{P}\bigg).$$

Note that we have at least $\Theta(n)$ number of $\mathbf{x}_i$'s that consist of the common feature vector $\mathbf{v}$, then applying union bound over these $\mathbf{x}_i$'s, we can further get with probability at least $1 - 1/\mathrm{poly}(n)$, it holds that

$$\begin{aligned}
\sum_{i,j\in[n]} \sum_{p\in[P]} \mathbb{1}[\mathbf{x}_{i,j}^{(p)} = \mathbf{v}] &\geq \sum_{i,j\in[n]} \mathbb{1}[\mathbf{x}_{i,j}^{(p_i)} = \mathbf{v} | \mathbf{x}_i^{(p_i)} = \mathbf{v}] \cdot \mathbb{1}[\mathbf{x}_i^{(p_i)} = \mathbf{v}] \\
&\geq \Theta(n) \cdot \Theta\bigg(\frac{n}{P}\bigg) \\
&= \Theta\bigg(\frac{n^2}{P}\bigg).
\end{aligned}$$

Here we define p_i as the index of the data patch that is $\mathbf{v}$ if the data $\mathbf{x}_i$ has such a common feature vector, otherwise, p_i is arbitrarily chosen. On the other hand, we can also get

$$\sum_{i,j\in[n]} \sum_{p\in[P]} \mathbb{1}[\mathbf{x}_{i,j}^{(p)} = \mathbf{v}] \leq \sum_{i,j\in[n]} \sum_{p\in[P]} \mathbb{1}[\mathbf{x}_{i,j}^{(p)} = \mathbf{v} | \mathbf{x}_i^{(p)} = \mathbf{v}] \cdot \mathbb{1}[\mathbf{x}_i^{(p)} = \mathbf{v}] \leq n \cdot \Theta(1) \cdot \Theta\bigg(\frac{n}{P}\bigg) = \Theta\bigg(\frac{n^2}{P}\bigg),$$

where the second inequality is due to that each data will have at most $\Theta(1)$ patches being $\mathbf{v}$. Similarly, we can also prove the same results for the case of $\mathbf{x}_{i,j}^{(p)} = \lambda\mathbf{a} + (1-\lambda)\boldsymbol{b}$ with $\mathbf{a}, \boldsymbol{b} \in \{\mathbf{u}, \mathbf{v}\}$.

The proof for the case of $\mathbf{x}_{i,j}^{(p)} = \lambda\mathbf{a} + (1-\lambda)\boldsymbol{b}$ with $\mathbf{a} \in \{\mathbf{u}, \mathbf{v}\}$ and $\boldsymbol{b} \in \{\alpha\mathbf{v}, \alpha\mathbf{u}\}$ will be also similar, the only difference is that conditioning on $\mathbf{x}_i^{(p)} = \mathbf{v}$, the probability of $\mathbf{x}_j^{(p)} = \alpha\mathbf{v}$ or $\mathbf{x}_j^{(p)} = \alpha\mathbf{u}$ will be $O(b/P)$. Finally, we can get that (here we take $\mathbf{a} = \mathbf{v}$ and $\boldsymbol{b} = \mathbf{v}$ as an example)

$$\sum_{i,j\in[n]} \sum_{p\in[P]} \mathbb{1}[\mathbf{x}_{i,j}^{(p)} = \lambda\mathbf{v} + \alpha(1-\lambda)\mathbf{v}] = \Theta(n) \cdot \Theta\bigg(\frac{bn}{P}\bigg) = \Theta\bigg(\frac{bn^2}{P}\bigg).$$

The proof for the case of $\mathbf{x}_{i,j}^{(p)} = \lambda\mathbf{a} + (1-\lambda)\boldsymbol{b}$ with $\mathbf{a} \in \{\mathbf{u}, \mathbf{v}\}$ and $\boldsymbol{b} \in \{\mathbf{u}', \mathbf{v}'\}$ will also be similar, where we only need to use the fact that $\mathbb{P}[\mathbf{x}_j^{(p)} = \mathbf{v}' | \mathbf{x}_i^{(p)} = \mathbf{v}] = \Theta(\rho/P)$. Here we take $\mathbf{a} = \mathbf{v}$ and $\boldsymbol{b} = \mathbf{v}'$ as an example.

Regarding the case of $\mathbf{x}_{i,j}^{(p)} = \lambda\mathbf{a} + (1-\lambda)\boldsymbol{b}$ with $\mathbf{a} \in \{\mathbf{u}, \mathbf{v}\}$ and $\boldsymbol{b} \in \{\boldsymbol{\xi}\}$, we only need to use the fact that $\mathbb{P}[\mathbf{x}_{i,j}^{(p)} = \lambda\mathbf{v} + (1-\lambda)\boldsymbol{\xi}_j^{(p)} | \mathbf{x}_i^{(p)} = \mathbf{v}] = \Theta(1)$, where we take $\mathbf{a} = \mathbf{v}$ as an example. Then the desired result can be proved in a similar way.

When $\mathbf{a} \in \{\mathbf{v}', \mathbf{u}'\}$ we will also need to use the fact that we have in total $\Theta(\rho n)$ number of $\mathbf{x}_i$'s that consist of $\mathbf{v}'$ or $\mathbf{u}'$. Then take $\mathbf{a} = \mathbf{v}'$ and $\boldsymbol{b} = \mathbf{v}'$ as an example, conditioning on $\mathbf{x}_i^{(p)} = \mathbf{v}'$, we have for any $j \neq i$

$$\mathbb{P}[\mathbf{x}_j^{(p)} = \mathbf{v}' | \mathbf{x}_i^{(p)} = \mathbf{v}'] = \mathbb{P}[\mathbf{x}_j^{(p)} = \mathbf{v}'] = \Theta\bigg(\frac{\rho}{P}\bigg).$$

Therefore, we can get that with probability at least $1 - 1/\mathrm{poly}(n)$,

$$\sum_{j \in [n]} \mathbb{1}[\mathbf{x}_j^{(p)} = \mathbf{v}' | \mathbf{x}_i^{(p)} = \mathbf{v}'] = \Theta\bigg(\frac{\rho n}{P}\bigg).$$

Accordingly, we can further obtain

$$\sum_{i,j \in [n]} \sum_{p \in [P]} \mathbb{1}[\mathbf{x}_{i,j}^{(p)} = \mathbf{v}'] = \sum_{i,j \in [n]} \sum_{p \in [P]} \mathbb{1}[\mathbf{x}_{i,j}^{(p)} = \mathbf{v}' | \mathbf{x}_i^{(p)} = \mathbf{v}'] \cdot \mathbb{1}[\mathbf{x}_i^{(p)} = \mathbf{v}'] = \Theta(\rho n) \cdot \Theta\bigg(\frac{\rho n}{P}\bigg) = \Theta\bigg(\frac{\rho^2 n^2}{P}\bigg).$$

The proof for the case of $\mathbf{x}_{i,j}^{(p)} = \lambda \mathbf{a} + (1-\lambda)\boldsymbol{b}$ with $\mathbf{a} \in \{\mathbf{u}', \mathbf{v}'\}$ and $\boldsymbol{b} \in \{\alpha \mathbf{v}, \alpha \mathbf{v}\}$ or $\boldsymbol{b} \in \{\boldsymbol{\xi}\}$ will also be similar, where we only need to use the fact that $\mathbb{P}[\mathbf{x}_j^{(p)} = \alpha \mathbf{v} | \mathbf{x}_i^{(p)} = \mathbf{v}'] = O(b/P)$ and $\mathbb{P}[\mathbf{x}_j^{(p)} = \alpha \boldsymbol{\xi}_j^{(p)} | \mathbf{x}_i^{(p)} = \mathbf{v}'] = \Theta(1)$.

When $\mathbf{a} \in \{\alpha \mathbf{v}, \alpha \mathbf{u}\}$ we only need to use the fact that we have in total $\Theta(n)$ number of $\mathbf{x}_i$'s that consist of $\Theta(b)$ number of $\mathbf{v}'$ or $\mathbf{u}'$. The remaining proof will be similar to previous ones based on the fact that $\mathbb{P}[\mathbf{x}_j^{(p)} = \alpha \mathbf{v} | \mathbf{x}_i^{(p)} = \alpha \mathbf{v}] = O(b/P)$ and $\mathbb{P}[\mathbf{x}_j^{(p)} = \alpha \boldsymbol{\xi}_j^{(p)} | \mathbf{x}_i^{(p)} = \alpha \mathbf{v}] = \Theta(1)$, where we take $\mathbf{a} = \alpha \mathbf{v}$ and $\boldsymbol{b} = \alpha \mathbf{v}$ as an example.

Lastly, we will move on to the case of $\mathbf{a} = \boldsymbol{\xi}_i^{(p)}$. In this case, we only need to use the facts that for any $j \neq i$,

$$\begin{aligned}
\mathbb{P}[\mathbf{x}_j^{(p)} = \mathbf{v} | \mathbf{x}_i^{(p)} = \boldsymbol{\xi}_i^{(p)}] &= \mathbb{P}[\mathbf{x}_j^{(p)} = \mathbf{v}] = \Theta(1/P) \\
\mathbb{P}[\mathbf{x}_j^{(p)} = \mathbf{u} | \mathbf{x}_i^{(p)} = \boldsymbol{\xi}_i^{(p)}] &= \mathbb{P}[\mathbf{x}_j^{(p)} = \mathbf{u}] = \Theta(1/P) \\
\mathbb{P}[\mathbf{x}_j^{(p)} = \mathbf{v}' | \mathbf{x}_i^{(p)} = \boldsymbol{\xi}_i^{(p)}] &= \mathbb{P}[\mathbf{x}_j^{(p)} = \mathbf{v}'] = \Theta(\rho/P) \\
\mathbb{P}[\mathbf{x}_j^{(p)} = \mathbf{u}' | \mathbf{x}_i^{(p)} = \boldsymbol{\xi}_i^{(p)}] &= \mathbb{P}[\mathbf{x}_j^{(p)} = \mathbf{u}'] = \Theta(\rho/P) \\
\mathbb{P}[\mathbf{x}_j^{(p)} = \alpha \mathbf{v} | \mathbf{x}_i^{(p)} = \boldsymbol{\xi}_i^{(p)}] &= \mathbb{P}[\mathbf{x}_j^{(p)} = \alpha \mathbf{v}] = O(b/P) \\
\mathbb{P}[\mathbf{x}_j^{(p)} = \alpha \mathbf{u} | \mathbf{x}_i^{(p)} = \boldsymbol{\xi}_i^{(p)}] &= \mathbb{P}[\mathbf{x}_j^{(p)} = \alpha \mathbf{u}] = O(b/P) \\
\mathbb{P}[\mathbf{x}_j^{(p)} \in \{\boldsymbol{\xi}\} | \mathbf{x}_i^{(p)} = \boldsymbol{\xi}_i^{(p)}] &= \mathbb{P}[\mathbf{x}_j^{(p)} \in \{\boldsymbol{\xi}\}] = \Theta(1).
\end{aligned}$$

Then applying the standard concentration argument for binomial distribution yields the desired results.

□

C.2 Learning Dynamics of Feature and Noise vectors

Now, we will seek to study the learning of feature and noise vectors. Particularly, the update formulas of all feature vectors are provided as follows: for any $\mathbf{a} \in \{\mathbf{u}, \mathbf{v}, \mathbf{u}', \mathbf{v}'\} \cup \{\boldsymbol{\xi}\}$, we have

$$\begin{aligned}
\langle \mathbf{w}_{k,r}^{(t+1)}, \mathbf{a} \rangle &= \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{a} \rangle - \eta \cdot \langle \nabla_{\mathbf{w}_{k,r}} L(\mathbf{W}^{(t)}), \mathbf{a} \rangle \\
&= \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{a} \rangle + \frac{\eta}{n^2} \cdot \sum_{i,j \in [n]} \ell_{k,(i,j)}^{(t)} \sum_{p \in [P]} \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{x}_{i,j}^{(p)} \rangle \cdot \langle \mathbf{x}_{i,j}^{(p)}, \mathbf{a} \rangle
\end{aligned} \tag{C.1}$$

More specifically, we summarize the update of all critical vectors (e.g., common features, rare features, and data noise vectors) in the following Proposition.

Proposition C.2. For any critical vector $\mathbf{a} \in \{\mathbf{v}, \mathbf{u}, \mathbf{v}', \mathbf{u}'\} \cup \{\boldsymbol{\xi}\}$, we have

$$-\langle \nabla_{\mathbf{w}_{k,r}} L_{\mathcal{S}}(\mathbf{W}^{(t)}), \mathbf{a} \rangle = \gamma_k^{(t)}(\mathbf{v}, \mathbf{a}) \cdot \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{v} \rangle + \gamma_k^{(t)}(\mathbf{u}, \mathbf{a}) \cdot \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{u} \rangle + \gamma_k^{(t)}(\mathbf{v}', \mathbf{a}) \cdot \langle \mathbf{w}_{k,r}^{(t)}, \mathbf{v}' \rangle$$

$$+\gamma_k^{(t)}(\mathbf{u}',\mathbf{a})\cdot\langle\mathbf{w}_{k,r}^{(t)},\mathbf{u}'\rangle+\sum_{i=1}^{n}\sum_{p\in[P]}\gamma_k^{(t)}(\boldsymbol{\xi}_i^{(p)},\mathbf{a})\cdot\langle\mathbf{w}_{k,r}^{(t)},\boldsymbol{\xi}_i^{(p)}\rangle,$$

where $\gamma_k^{(t)}(\boldsymbol{b},\mathbf{a})$ is a scalar output function that depends on $\boldsymbol{b},\mathbf{a}\in\{\mathbf{v},\mathbf{u},\mathbf{v}',\mathbf{u}'\}\cup\{\boldsymbol{\xi}\}$. More specifically, let

$$\mathbf{x}_{i,j}^{(p)}=\theta_{i,j}^{(p)}(\mathbf{v})\cdot\mathbf{v}+\theta_{i,j}^{(p)}(\mathbf{u})\cdot\mathbf{u}+\theta_{i,j}^{(p)}(\mathbf{v}')\cdot\mathbf{v}'+\theta_{i,j}^{(p)}(\mathbf{u}')\cdot\mathbf{u}'+\sum_{s=1}^{n}\sum_{q\in[P]}\theta_{i,j}^{(p)}(\boldsymbol{\xi}_s^{(q)})\cdot\boldsymbol{\xi}_s^{(q)} \tag{C.2}$$

be a linear expansion of $\mathbf{x}_{i,j}^{(p)}$ on the space spanned by $\{\mathbf{v},\mathbf{u},\mathbf{v}',\mathbf{u}'\}\cup\{\boldsymbol{\xi}\}$, we have

$$\gamma_k^{(t)}(\boldsymbol{b},\mathbf{a})=\frac{1}{n^2}\sum_{i,j\in[n]}\ell_{k,(i,j)}^{(t)}\sum_{p\in[P]}\theta_{i,j}^{(p)}(\boldsymbol{b})\cdot\langle\mathbf{x}_{i,j}^{(p)},\mathbf{a}\rangle. \tag{C.3}$$

Proof of Proposition C.2. Recall (C.1) and the decomposition of $\mathbf{x}_{i,j}^{(p)}$ in (C.2), we have

$$\begin{aligned}
-\langle\nabla_{\mathbf{w}_{k,r}}L_{\mathcal{S}}(\mathbf{W}^{(t)}),\mathbf{a}\rangle&=\frac{1}{n^2}\sum_{i,j\in[n]}\ell_{k,(i,j)}^{(t)}\sum_{p\in[P]}\langle\mathbf{w}_{k,r}^{(t)},\mathbf{x}_{i,j}^{(p)}\rangle\cdot\langle\mathbf{x}_{i,j}^{(p)},\mathbf{a}\rangle\\
&=\frac{1}{n^2}\sum_{i,j\in[n]}\ell_{k,(i,j)}^{(t)}\sum_{p\in[P]}\sum_{\boldsymbol{b}\in\{\mathbf{v},\mathbf{u},\mathbf{v}',\mathbf{u}'\}\cup\{\boldsymbol{\xi}\}}\theta_{i,j}^{(p)}(\boldsymbol{b})\cdot\langle\mathbf{w}_{k,r}^{(t)},\boldsymbol{b}\rangle\cdot\langle\mathbf{x}_{i,j}^{(p)},\mathbf{a}\rangle\\
&=\sum_{\boldsymbol{b}\in\{\mathbf{v},\mathbf{u},\mathbf{v}',\mathbf{u}'\}\cup\{\boldsymbol{\xi}\}}\left[\frac{1}{n^2}\sum_{i,j\in[n]}\ell_{k,(i,j)}^{(t)}\sum_{p\in[P]}\theta_{i,j}^{(p)}(\boldsymbol{b})\cdot\langle\mathbf{x}_{i,j}^{(p)},\mathbf{a}\rangle\right]\cdot\langle\mathbf{w}_{k,r}^{(t)},\boldsymbol{b}\rangle.
\end{aligned}$$

Therefore, it is easy to see that using the definition of $\gamma_k^{(t)}(\boldsymbol{b},\mathbf{a})$ in (C.3), we have

$$-\langle\nabla_{\mathbf{w}_{k,r}}L_{\mathcal{S}}(\mathbf{W}^{(t)}),\mathbf{a}\rangle=\sum_{\boldsymbol{b}\in\{\mathbf{v},\mathbf{u},\mathbf{v}',\mathbf{u}'\}\cup\{\boldsymbol{\xi}\}}\gamma_k^{(t)}(\boldsymbol{b},\mathbf{a})\cdot\langle\mathbf{w}_{k,r}^{(t)},\boldsymbol{b}\rangle,$$

which completes the proof.

□

Note that the neural network outputs are in the order of $o(1)$ in the first few iterations, which implies that the output logits are within the range $[0.5-o(1),0.5+o(1)]$. Further note that the loss derivatives $\ell_{k;(i,j)}^{(t)}$ satisfies

$$|\ell_{k;(i,j)}^{(t)}|\in\big\{1-\mathrm{Logit}_k(\mathbf{W}^{(t)};\mathbf{x}_{i,j}),\mathrm{Logit}_k(\mathbf{w}^{(t)};\mathbf{x}_{i,j}),\lambda-\mathrm{Logit}_k(\mathbf{W}^{(t)};\mathbf{x}_{i,j}),\mathrm{Logit}_k(\mathbf{W}^{(t)};\mathbf{x}_{i,j})+\lambda-1\big\},$$

which will also be in the constant order. Then similar to the previous analysis on the standard training, we will directly take $|\ell_{k,(i,j)}^{(t)}|=\Theta(1)$ when characterizing the learning of feature and noise vectors in the initial phase.

Then, the challenging part in the analysis is the characterization of the mixed data patches $\{\mathbf{x}_{i,j}^{(p)}\}_{p\in[P]}$, since it can be: mixture of common features, mixture of rare features, mixture of common and rare features, mixture of feature and noise, which will produce different gradients. For any mixed data $\mathbf{x}_{i,j}=\lambda\mathbf{x}_i+(1-\lambda)\mathbf{x}_j$, we will denote it as the positive mixed data if $y_i=1$ and the negative mixed data if $y_i=-1$. The following lemma gives the characterization of the data patch of all mixed data.

C.3 Characterizing the Coefficient $\gamma_k^{(t)}(\cdot,\cdot)$

C.3.1 Correct Common Feature Learning

Lemma C.3. Assume $\max_{k\in[2],(i,j)\in\mathcal{S}}|F_k(\mathbf{W}^{(t)};\mathbf{x}_{i,j})|\le\zeta\in\big[\omega(b\alpha),o\big(\frac{1}{\mathrm{polylog}(n)}\big)\big]$, then recalling the update form in Proposition C.2, we have

$$\begin{aligned}
&\gamma_1^{(t)}(\mathbf{v},\mathbf{v})=\Theta(1),\quad|\gamma_1^{(t)}(\mathbf{u},\mathbf{v})|=O(\zeta+\alpha),\quad|\gamma_1^{(t)}(\mathbf{v}',\mathbf{v})|=O(\rho/P),\\
&|\gamma_1(\mathbf{u}',\mathbf{v})|=O(\zeta\rho/P),\quad|\gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)},\mathbf{v})|=\widetilde{O}\big(1/(Pn)\big).
\end{aligned}$$

Proof of Lemma C.3. We will prove all the arguments in order.

Proof for $\gamma_1^{(t)}(\mathbf{v},\mathbf{v})$. We first prove the bound for $\gamma_1^{(t)}(\mathbf{v},\mathbf{v})$. By (C.3), we have

$$\gamma_k^{(t)}(\mathbf{v},\mathbf{v}) = \frac{1}{n^2}\sum_{i,j\in[n]} \ell_{k,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle, \tag{C.4}$$

where $\theta_{i,j}^{(p)}(\mathbf{v}) = \langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle$. Therefore, we only need to consider the data patches that contain $\mathbf{v}$ (including common feature $\mathbf{v}$ and feature noise $\alpha\mathbf{v}$). The regarding the mixed data $\mathbf{x}_{i,j}$, we consider the following cases

- $i\in\mathcal{S}_0^+$ and $j\in\mathcal{S}_0^+$;
- $i\in\mathcal{S}_0^+$ and $j\in\mathcal{S}_1^+$, and $i\in\mathcal{S}_1^+$ and $j\in\mathcal{S}_0^+$;
- $i\in\mathcal{S}_0^+$ and $j\in\mathcal{S}_0^-\cup\mathcal{S}_1^-$, and $i\in\mathcal{S}_0^-\cup\mathcal{S}_1^-$ and $j\in\mathcal{S}_0^+$
- $i\in\mathcal{S}_0^-\cup\mathcal{S}_1^+\cup\mathcal{S}_1^-$ and $j\in\mathcal{S}_0^-\cup\mathcal{S}_1^+\cup\mathcal{S}_1^-$.

Analysis on the data $i\in\mathcal{S}_0^+$ and $j\in\mathcal{S}_0^+$ In particular, note that before the mixup, both the data $\mathbf{x}_i$ and $\mathbf{x}_j$ have a constant number of common feature patches. Therefore, let $\mathcal{P}_{i,j}^*(\mathbf{v})$ denote the set of patches with the common feature $\mathbf{v}$ (which appears in either $\mathbf{x}_i$ or $\mathbf{x}_j$), we have

$$\sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle = \sum_{p\in\mathcal{P}_{i,j}^*(\mathbf{v})} \theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle + \sum_{p\in\mathcal{P}_{i,j}(\mathbf{v})\setminus\mathcal{P}_{i,j}^*(\mathbf{v})} \theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle. \tag{C.5}$$

Regarding the first term on the R.H.S. of the above equation, by Definition 3.1, we know that there exists at least one common feature patch in both $\mathbf{x}_i$ and $\mathbf{x}_j$, which leads to $\theta_{i,j}^{(p)}\geq\lambda$ for at least one $p\in\mathcal{P}_{i,j}^*(vb)$. This further gives

$$\sum_{p\in\mathcal{P}_{i,j}^*(\mathbf{v})} \theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle = \sum_{p\in\mathcal{P}_{i,j}^*(\mathbf{v})} [\theta_{i,j}^{(p)}(\mathbf{v})]^2 \geq \lambda^2.$$

Besides, we also have that the number of common feature patches are upper bounded by some constant (i.e., $|\mathcal{P}_{i,j}^*(\mathbf{v})| = \Theta(1)$), this further leads to

$$\sum_{p\in\mathcal{P}_{i,j}^*(\mathbf{v})} \theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle \leq \Theta(1).$$

Regarding the second term on the R.H.S. of (C.5), we have $\theta_{i,j}^{(p)}\leq\alpha$ since $\mathbf{v}$ can only appear in the form of feature noise. Besides, by Definition 3.1, we know that the number of patches containing feature noise is at most b, then

$$\sum_{p\in\mathcal{P}_{i,j}(\mathbf{v})\setminus\mathcal{P}_{i,j}^*(\mathbf{v})} \theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle = \sum_{p\in\mathcal{P}_{i,j}(\mathbf{v})\setminus\mathcal{P}_{i,j}^*(\mathbf{v})} [\theta_{i,j}^{(p)}(\mathbf{v})]^2 \leq b\alpha^2 = o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg).$$

Moreover, note that in the initial phase we have $\ell_{1,(i,j)}^{(t)} = \Theta(1)$ for $(i,j)\in\mathcal{S}_{0,0}^{+,+}$, we can further get that

$$\ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle = \Theta(1).$$

Analysis on the data $i\in\mathcal{S}_0^+$ and $j\in\mathcal{S}_1^+$. The analysis for this type of data will be similar. In fact, we will consider two types of data: $i\in\mathcal{S}_0^+$ and $j\in\mathcal{S}_1^+$, and $i\in\mathcal{S}_1^+$ and $j\in\mathcal{S}_0^+$ since two original training data will give two mixed data.

In particular, note that $\ell_{i,j}^{(t)} = \Theta(1)$ for these two types of data, we can immediately get that there is a constant number of patches that satisfy $\theta_{i,j}^{(p)}\geq 1-\lambda$, while the remaining patches $p\in\mathcal{P}_{i,j}(\mathbf{v})$ satisfy $\theta_{i,j}^{(p)}\leq\alpha$. Therefore, we can follow the same proof technique as that for the data $(i,j)\in\mathcal{S}_{0,0}^{+,+}$ and get that for all $(i,j)\in\mathcal{S}_{0,1}^{+,+}\cup\mathcal{S}_{1,0}^{+,+}$,

$$\ell_{k,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle = \ell_{k,(i,j)}^{(t)} \sum_{p\in\mathcal{P}_{i,j}^*(\mathbf{v})} [\theta_{i,j}^{(p)}(\mathbf{v})]^2 + \sum_{p\in\mathcal{P}_{i,j}(\mathbf{v})\setminus\mathcal{P}_{i,j}^*(\mathbf{v})} [\theta_{i,j}^{(p)}(\mathbf{v})]^2 = \Theta(1). \tag{C.6}$$

Analysis on the data $i \in \mathcal{S}_0^+$ and $j \in \mathcal{S}_0^- \cup \mathcal{S}_1^-$. In this part, we will handle data $\mathbf{x}_{i,j}$ and $\mathbf{x}_{j,i}$ together. Different from the previous cases where the loss derivatives $\ell_{1,(i,j)}^{(t)}$ are positive, here the loss derivative $\ell_{1,(i,j)}^{(t)}$ will become negative for $(i,j) \in \mathcal{S}_{0,0}^{-,+} \cup \mathcal{S}_{0,1}^{-,+}$. Particularly, for any $(i,j) \in \mathcal{S}_{0,0}^{+,-}$, we have $(j,i) \in \mathcal{S}_{0,0}^{-,+}$, then

$$\begin{aligned}
&\ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{v}) \cdot \langle \mathbf{x}_{i,j}^{(p)}, \mathbf{v}\rangle + \ell_{1,(j,i)}^{(t)} \sum_{p\in[P]} \theta_{j,i}^{(p)}(\mathbf{v}) \cdot \langle \mathbf{x}_{j,i}^{(p)}, \mathbf{v}\rangle \\
&= \ell_{1,(i,j)}^{(t)} \sum_{p\in\mathcal{P}_{i,j}(\mathbf{v})} \big[[\theta_{i,j}^{(p)}(\mathbf{v})]^2 - [\theta_{j,i}^{(p)}(\mathbf{v})]^2\big] + \big[\ell_{1,(i,j)}^{(t)} + \ell_{1,(j,i)}^{(t)}\big] \cdot \sum_{p\in\mathcal{P}_{i,j}(\mathbf{v})} \big[\theta_{j,i}^{(p)}(\mathbf{v})\big]^2,
\end{aligned} \tag{C.7}$$

where we use the fact that $\mathcal{P}_{i,j}(\mathbf{v}) = \mathcal{P}_{j,i}(\mathbf{v})$ and $\langle \mathbf{x}_{i,j}^{(p)}, \mathbf{v}\rangle = \theta_{i,j}^{(p)}(\mathbf{v})$. Recall that the neural network output is upper bounded by ζ, then it is easy to see

$$|\ell_{1,(i,j)}^{(t)} + \ell_{1,(j,i)}^{(t)}| = |\lambda - 0.5 \pm O(\zeta) + 0.5 - \lambda \pm O(\zeta)| = O(\zeta).$$

Besides, note that

$$\mathbf{x}_{i,j}^{(p)} = \lambda \mathbf{x}_i^{(p)} + (1-\lambda)\mathbf{x}_j^{(p)}, \quad \mathbf{x}_{j,i}^{(p)} + (1-\lambda)\mathbf{x}_i^{(p)} + \lambda \mathbf{x}_j^{(p)}.$$

Then we will also define $\mathcal{P}_{i,j}^*(\mathbf{v})$ as the set of patches with common feature. Note that $\mathbf{x}_j^{(p)}$ does not have the common feature patch since $j \in \mathcal{S}_0^- \cup \mathcal{S}_1^-$, we can immediately get that $\mathcal{P}_{i,j}^*(\mathbf{v}) = \mathcal{P}_i^*(\mathbf{v})$, where $\mathcal{P}_i^*(\mathbf{v})$ denotes the set of common feature patches of $\mathbf{x}_i$. Besides, it is also clear that all data patches in $\mathcal{P}_{i,j}(\mathbf{v})$ only contain the feature noise $\alpha\mathbf{v}$. Then it follows that

$$\begin{aligned}
&\ell_{1,(i,j)}^{(t)} \sum_{p\in\mathcal{P}_{i,j}(\mathbf{v})} \big[[\theta_{i,j}^{(p)}(\mathbf{v})]^2 - [\theta_{j,i}^{(p)}(\mathbf{v})]^2\big] \\
&= \Theta(1) \cdot \bigg[\sum_{p\in\mathcal{P}_{i,j}^*(\mathbf{v})} [\lambda^2 - (1-\lambda)^2] + \sum_{p\in\mathcal{P}_{i,j}(\mathbf{v})\setminus\mathcal{P}_{i,j}^*(\mathbf{v})} \big[[\theta_{i,j}^{(p)}(\mathbf{v})]^2 - [\theta_{j,i}^{(p)}(\mathbf{v})]^2\big] \bigg] \\
&= \Theta(1) \pm O(b\alpha^2) \\
&= \Theta(1).
\end{aligned}$$

Similarly, we can also get $\sum_{p\in\mathcal{P}_{i,j}(\mathbf{v})} [\theta_{j,i}^{(p)}]^2 = \Theta(1)$. Therefore, putting everything to (C.7), we can finally obtain the following

$$\ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{v}) \cdot \langle \mathbf{x}_{i,j}^{(p)}, \mathbf{v}\rangle + \ell_{1,(j,i)}^{(t)} \sum_{p\in[P]} \theta_{j,i}^{(p)}(\mathbf{v}) \cdot \langle \mathbf{x}_{j,i}^{(p)}, \mathbf{v}\rangle = \Theta(1) \pm \Theta(1) \cdot O(\zeta) = \Theta(1).$$

Analysis on the data $i, j \in \mathcal{S}_0^- \cup \mathcal{S}_1^+ \cup \mathcal{S}_1^-$ In this case, we can observe that there is no common feature patches in $\mathbf{x}_i$ and $\mathbf{x}_j$, while the vector $\mathbf{v}$ will only appear in at most $2b$ patches of $\mathbf{x}_{i,j}$ in the form of feature noise. Therefore, we have $|\Theta_{i,j}^{(p)}| \in [(1-\lambda)\alpha, \alpha]$ for at most $2b$ patches and the remaining patches will give $|\Theta_{i,j}^{(p)}| = 0$. Consequently, we have

$$\bigg|\ell_{k,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{v}) \cdot \langle \mathbf{x}_{i,j}^{(p)}, \mathbf{v}\rangle\bigg| = \bigg|\ell_{k,(i,j)}^{(t)} \sum_{p\in\mathcal{P}_{i,j}(\mathbf{v})} [\theta_{i,j}^{(p)}(\mathbf{v})]^2\bigg| = O(b\alpha^2) = o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg).$$

Completing the analysis for $\gamma_1^{(t)}(\mathbf{v},\mathbf{v})$. Now we are able to complete the analysis on $\gamma_1^{(t)}(\mathbf{v},\mathbf{v})$ based on (C.4):

$$\begin{aligned}
\gamma_1^{(t)}(\mathbf{v},\mathbf{v}) &= \frac{1}{n^2} \sum_{i,j\in[n]} \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{v}) \cdot \langle \mathbf{x}_{i,j}^{(p)}, \mathbf{v}\rangle \\
&= \frac{1}{n^2} \bigg[\sum_{(i,j)\in\mathcal{S}_{0,0}^{+,+}} \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{v}) \cdot \langle \mathbf{x}_{i,j}^{(p)}, \mathbf{v}\rangle
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{(i,j)\in\mathcal{S}_{0,1}^{+,+}\cup\mathcal{S}_{1,0}^{+,+}} \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle \\
&+ \sum_{(i,j)\in\mathcal{S}_{0,0}^{+,-}\cup\mathcal{S}_{0,1}^{+,-}\cup\mathcal{S}_{0,0}^{-,+}\cup\mathcal{S}_{0,1}^{-,+}} \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle \\
&+ \sum_{i,j\in\mathcal{S}_0^-\cup\mathcal{S}_1^+\cup\mathcal{S}_1^-} \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle\Bigg] \\
=& \frac{1}{n^2}\Bigg[\Theta(1)\cdot|\mathcal{S}_{0,0}^{+,+}| + \Theta(1)\cdot|\mathcal{S}_{0,1}^{+,+}\cup\mathcal{S}_{1,0}^{+,+}| + \Theta(1)\cdot|\mathcal{S}_{0,0}^{+,-}\cup\mathcal{S}_{0,1}^{+,-}\cup\mathcal{S}_{0,0}^{-,+}\cup\mathcal{S}_{0,1}^{-,+}| \\
&\pm o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg)\cdot|\mathcal{S}_{0,0}^{-,-}\cup\mathcal{S}_{0,1}^{-,+}\cup\mathcal{S}_{0,1}^{-,-}\cup\mathcal{S}_{1,0}^{+,-}\cup\mathcal{S}_{1,1}^{+,+}\cup\mathcal{S}_{1,1}^{+,-}\cup\mathcal{S}_{1,0}^{-,-}\cup\mathcal{S}_{1,1}^{-,+}\cup\mathcal{S}_{1,1}^{-,-}|\Bigg] \\
=& \frac{1}{n^2}\Bigg[\Theta(n^2)\pm o\bigg(\frac{n^2}{\mathrm{polylog}(n)}\bigg)\Bigg] \\
=& \Theta(1).
\end{aligned}
$$

Proof for $\gamma_k^{(t)}(\mathbf{u},\mathbf{v})$. The next step is to characterize $\gamma_k^{(t)}(\mathbf{u},\mathbf{v})$. We will split the entire mixed training dataset into the following classes:

- $i\in\mathcal{S}_0^+$ and $j\in\mathcal{S}_0^-$, and $i\in\mathcal{S}_0^-$ and $j\in\mathcal{S}_0^+$, i.e., $\mathcal{S}_{0,0}^{+,-}\cup\mathcal{S}_{0,0}^{-,+}$.
- all $(i,j)\notin\mathcal{S}_{0,0}^{+,-}\cup\mathcal{S}_{0,0}^{-,+}$.

We first recall the formula of $\gamma_1^{(t)}(\mathbf{u},\mathbf{v})$ (see Proposition C.2):

$$
\gamma_1^{(t)}(\mathbf{u},\mathbf{v}) = \frac{1}{n^2}\sum_{i,j\in[n]} \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{u})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle. \tag{C.8}
$$

Analysis on the data $(i,j)\in\mathcal{S}_{0,0}^{+,-}\cup\mathcal{S}_{0,0}^{-,+}$. Since $\mathcal{S}_{0,0}^{+,-}$ and $\mathcal{S}_{0,0}^{-,+}$ are symmetric: i.e., for any $(i,j)\in\mathcal{S}_{0,0}^{+,-}$, we have $(j,i)\in\mathcal{S}_{0,0}^{-,+}$ and vise versa. Then we will handle data $\mathbf{x}_{i,j}$ and $\mathbf{x}_{j,i}$ together by studying the following quantity:

$$
\begin{aligned}
* :=& \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{u})\cdot\langle \mathbf{x}_{i,j}^{(p)},\mathbf{v}\rangle + \ell_{1,(j,i)}^{(t)} \sum_{p\in[P]} \theta_{j,i}^{(p)}(\mathbf{u})\cdot\langle \mathbf{x}_{j,i}^{(p)},\mathbf{v}\rangle \\
=& \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{u})\cdot\theta_{i,j}^{(p)}(\mathbf{v}) + \ell_{1,(j,i)}^{(t)} \sum_{p\in[P]} \theta_{j,i}^{(p)}(\mathbf{u})\cdot\theta_{j,i}^{(p)}(\mathbf{v}).
\end{aligned}
$$

Note that we will only consider the patch that contains both $\mathbf{u}$ and $\mathbf{v}$. Then consider a data patch $\mathbf{x}_{i,j}^{(p)}$ satisfy this condition: $\mathbf{x}_i^{(p)}=\alpha_i\mathbf{v}$ and $\mathbf{x}_j^{(p)}=\alpha_j\mathbf{u}$, where $\alpha_i,\alpha_j\in\{\alpha,1\}$, which further leads to $\mathbf{x}_{i,j}^{(p)}=\lambda\alpha_i\mathbf{v}+(1-\lambda)\alpha_j\mathbf{u}$ and $\mathbf{x}_{j,i}^{(p)}=\lambda\alpha_j\mathbf{u}+(1-\lambda)\alpha_i\mathbf{v}$. Accordingly, it further gives

$$
\theta_{i,j}^{(p)}(\mathbf{u})\cdot\theta_{i,j}^{(p)}(\mathbf{v}) = (1-\lambda)\alpha_j\cdot\alpha_i\lambda_i = \lambda\alpha_j\cdot(1-\lambda)\alpha_i = \theta_{j,i}^{(p)}(\mathbf{u})\cdot\theta_{j,i}^{(p)}(\mathbf{v}).
$$

Additionally, for any $p\in\mathcal{P}_{i,j}(\mathbf{v})$, we have at most $\Theta(1)$ among them satisfy $\theta_{j,i}^{(p)}(\mathbf{u})=\Theta(1)$ and at most $\Theta(1)$ among them satisfy $\theta_{j,i}^{(p)}(\mathbf{v})=\Theta(1)$, while the remaining, with size at most $2b$, can only give $\theta_{j,i}^{(p)}(\mathbf{u}),\theta_{j,i}^{(p)}(\mathbf{v})=\Theta(\alpha)$. This implies that

$$
\sum_{p\in[P]} \theta_{j,i}^{(p)}(\mathbf{u})\cdot\theta_{j,i}^{(p)}(\mathbf{v}) = O(1)+O(\alpha)+O(b\alpha^2) = O(1).
$$

Therefore, applying the above equations, we can get that

$$
* = \ell_{1,(i,j)}^{(t)} \sum_{p\in\mathcal{P}_{i,j}(\mathbf{v})} \big[\theta_{i,j}^{(p)}(\mathbf{u})\cdot\theta_{i,j}^{(p)}(\mathbf{v}) - \theta_{j,i}^{(p)}(\mathbf{u})\cdot\theta_{j,i}^{(p)}(\mathbf{v})\big] + \big[\ell_{1,(i,j)}^{(t)}+\ell_{1,(j,i)}^{(t)}\big]\cdot\sum_{p\in[P]} \theta_{j,i}^{(p)}(\mathbf{u})\cdot\theta_{j,i}^{(p)}(\mathbf{u})
$$

$$= \big[\ell_{1,(i,j)}^{(t)} + \ell_{1,(j,i)}^{(t)}\big] \cdot O(1).$$

Further note that in the initial phase we have $\ell_{1,(i,j)}^{(t)} + \ell_{1,(j,i)}^{(t)} = O(\zeta)$, we consequently get

$$|*| = \bigg|\ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{u}) \cdot \langle \mathbf{x}_{i,j}^{(p)}, \mathbf{v}\rangle + \ell_{1,(j,i)}^{(t)} \sum_{p\in[P]} \theta_{j,i}^{(p)}(\mathbf{u}) \cdot \langle \mathbf{x}_{j,i}^{(p)}, \mathbf{v}\rangle\bigg| = O(\zeta).$$

Analysis on the remaining data $(i,j) \notin \mathcal{S}_{0,0}^{+,-} \cup \mathcal{S}_{0,0}^{-,+}$. In this case, we note that there are no data patches that satisfy $\theta_{j,i}^{(p)}(\mathbf{v}) = \Theta(1)$ and $\theta_{j,i}^{(p)}(\mathbf{u}) = \Theta(1)$ simultaneously. Therefore, for any data $\mathbf{x}_{i,j}$, there will exist at most $\Theta(1)$ patches that satisfy $\theta_{j,i}^{(p)}(\mathbf{v}) \cdot \theta_{j,i}^{(p)}(\mathbf{u}) = \alpha$ and at most $2b$ patches satisfying $\theta_{j,i}^{(p)}(\mathbf{v}) \cdot \theta_{j,i}^{(p)}(\mathbf{u}) = \alpha^2$, while the remaining patches will give $\theta_{j,i}^{(p)}(\mathbf{v}) \cdot \theta_{j,i}^{(p)}(\mathbf{u}) = 0$. Therefore, we can get that

$$\sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{u}) \cdot \langle \mathbf{x}_{i,j}^{(p)}, \mathbf{v}\rangle = \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{u}) \cdot \theta_{i,j}^{(p)}(\mathbf{u}) = \Theta(1)\cdot\alpha + 2b\alpha^2 = O(\alpha),$$

where the last equality follows from the setting of the data distribution that $b\alpha < 1$.

Completing the analysis for $\gamma_1^{(t)}(\mathbf{u},\mathbf{v})$. By (C.8) and using the fact that $|\ell_{1,(i,j)}^{(t)}| \le 1$, we have

$$\begin{aligned}
|\gamma_1^{(t)}(\mathbf{u},\mathbf{v})| &= \frac{1}{n^2}\bigg|\sum_{i,j\in[n]} \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{u}) \cdot \theta_{i,j}^{(p)}(\mathbf{v})\bigg| \\
&= \frac{1}{n^2} \cdot \big[|\mathcal{S}_{0,0}^{+,-} \cup \mathcal{S}_{0,0}^{-,+}| \cdot O(\zeta) + \big(n^2 - |\mathcal{S}_{0,0}^{+,-} \cup \mathcal{S}_{0,0}^{-,+}|\big) \cdot O(\alpha)\big] \\
&= O(\zeta + \alpha).
\end{aligned}$$

Proof for $\gamma_1^{(t)}(\mathbf{v}',\mathbf{v})$**.** We then tend to characterize $\gamma_1^{(t)}(\mathbf{v}',\mathbf{v})$. We will consider the following two classes of data:

- $(i,j) \in \mathcal{S}_{0,1}^{+,+} \cup \mathcal{S}_{1,0}^{+,+}$
- all $(i,j) \notin \mathcal{S}_{0,1}^{+,+} \cup \mathcal{S}_{1,0}^{+,+}$.

Analysis on the data $(i,j) \in \mathcal{S}_{0,1}^{+,+} \cup \mathcal{S}_{1,0}^{+,+}$ First, it is easy to see that with probability at least $1 - 1/\mathrm{poly}(n)$, we have $|\mathcal{S}_{0,1}^{+,+} \cup \mathcal{S}_{1,0}^{+,+}| = O(\rho n^2)$. For this class of data, with probability $\Theta(1/P)$ we have the data $\mathbf{x}_{i,j}^{(p)}$ has a constant number of patches that satisfy $\theta_{i,j}^{(p)}(\mathbf{v}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}') = \Theta(1)$. Besides, by Lemma C.1, we have with probability at least $1 - 1/\mathrm{poly}(n)$, there are $\Theta(b\rho n^2/P)$ patches are the mixture of $\alpha\mathbf{v}$ and $\mathbf{v}'$, leading to $\theta_{i,j}^{(p)}(\mathbf{v}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}') = \Theta(\alpha)$. The remaining patches will give $\theta_{i,j}^{(p)}(\mathbf{v}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}') = 0$. Combine the above results, we can get

$$\sum_{(i,j)\in\mathcal{S}_{0,1}^{+,+}\cup\mathcal{S}_{1,0}^{+,+}} \ell_{1,(i,j)}^{(t)} \sum_{p\in\mathcal{P}} \theta_{i,j}^{(p)}(\mathbf{v}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}') = \Theta\bigg(\frac{\rho n^2}{P}\bigg) + \Theta\bigg(\frac{b\alpha\rho n^2}{P}\bigg) = \Theta\bigg(\frac{\rho n^2}{P}\bigg)$$

where we use the fact that $b\alpha = o(1)$.

Analysis on the remaining data Particular, we will only consider the data $(i,j) \in \mathcal{S}_{0,1}^{-,+} \cup \mathcal{S}_{1,1}^{-,+} \cup \mathcal{S}_{1,0}^{+,-}$ since otherwise there is no data containing the rare feature vector $\mathbf{v}'$. Moreover, note that for this class of data we only have $\theta_{i,j}^{(p)}(\mathbf{v}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}') = O(\alpha)$ since there is no data consisting of common feature patch (but only contain feature noise $\alpha\mathbf{v}$). Therefore, similar to the previous analysis, we can get that, by Lemma C.1, with probability at least $1 - 1/\mathrm{poly}(n)$, there are $\Theta(b\rho n^2/P)$ patches that give $\theta_{i,j}^{(p)}(\mathbf{v}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}') = \Theta(\alpha)$, which consequently leads to

$$\sum_{(i,j)\in\mathcal{S}_{0,1}^{-,+}\cup\mathcal{S}_{1,1}^{-,+}\cup\mathcal{S}_{1,0}^{+,-}} \bigg|\ell_{1,(i,j)}^{(t)} \sum_{p\in\mathcal{P}} \theta_{i,j}^{(p)}(\mathbf{v}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}')\bigg| = O\bigg(\frac{b\alpha\rho n^2}{P}\bigg).$$

Completing the analysis for $\gamma_1^{(t)}(\mathbf{v}', \mathbf{v})$. Completing the previous analysis, we have

$$\gamma_1^{(t)}(\mathbf{v}', \mathbf{v}) = \frac{1}{n^2} \sum_{i,j \in [n]} \ell_{1,(i,j)}^{(t)} \sum_{p \in [P]} \theta_{i,j}^{(p)}(\mathbf{v}') \cdot \theta_{i,j}^{(p)}(\mathbf{v}) = \Theta\bigg(\frac{\rho}{P}\bigg) \pm O\bigg(\frac{b\alpha\rho}{P}\bigg) = \Theta\bigg(\frac{\rho}{P}\bigg).$$

Proof for $\gamma_1^{(t)}(\mathbf{u}', \mathbf{v})$**.** Regarding the coefficient $\gamma_1^{(t)}(\mathbf{u}', \mathbf{v})$, we consider two cases (1) mixup between $\mathbf{u}'$ and $\mathbf{v}$; (2) mixup between $\mathbf{u}'$ and $\alpha\mathbf{v}$. Then it can be seen that the first cases cover the data $(i,j) \in \mathcal{S}_{1,0}^{-,+}$ and $(i,j) \in \mathcal{S}_{0,1}^{+,-}$, which is equivalent to the dataset $\{(i,j),(j,i) : (i,j) \in \mathcal{S}_{1,0}^{+,-}\}$. Therefore, we will handle the data (i,j) and (j,i) together in this case. In particular, we have

$$\begin{aligned}
&\ell_{1,(i,j)}^{(t)} \sum_{p \in [P]} \theta_{i,j}^{(p)}(\mathbf{u}') \cdot \theta_{i,j}^{(p)}(\mathbf{v}) + \ell_{1,(j,i)}^{(t)} \sum_{p \in [P]} \theta_{j,i}^{(p)}(\mathbf{u}') \cdot \theta_{j,i}^{(p)}(\mathbf{v}) \\
&= \ell_{1,(i,j)}^{(t)} \sum_{p \in [P]} \big[\theta_{i,j}^{(p)}(\mathbf{u}') \cdot \theta_{i,j}^{(p)}(\mathbf{v}) - \theta_{j,i}^{(p)}(\mathbf{u}') \cdot \theta_{j,i}^{(p)}(\mathbf{v})\big] + \big[\ell_{1,(i,j)}^{(t)} + \ell_{1,(j,i)}^{(t)}\big] \cdot \sum_{p \in [P]} \theta_{j,i}^{(p)}(\mathbf{u}') \cdot \theta_{j,i}^{(p)}(\mathbf{v}).
\end{aligned}$$

It is clear that the first term on the R.H.S. of the above equation is zero since in case (1)

$$\theta_{i,j}^{(p)}(\mathbf{u}') \cdot \theta_{i,j}^{(p)}(\mathbf{v}) = \theta_{j,i}^{(p)}(\mathbf{u}') \cdot \theta_{j,i}^{(p)}(\mathbf{v}) = \lambda(1-\lambda).$$

Regarding the second term, we can use Lemma C.1 and get that the number of patches falling in case (1) is $\Theta(\rho n^2/P)$. Then using the fact that $|\ell_{1,(i,j)}^{(t)} + \ell_{1,(i,j)}^{(t)}| = O(\zeta)$ can lead to the final bound for case (1).

Regarding case (2), we can follow the analysis for $\gamma_1^{(t)}(\mathbf{v}', \mathbf{v})$, which relies on the fact that $\theta_{i,j}^{(p)}(\mathbf{u}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}') = \Theta(\alpha)$. Therefore, we can finally get

$$\begin{aligned}
|\gamma_1^{(t)}(\mathbf{u}, \mathbf{v}')| &= \bigg| \frac{1}{n^2} \sum_{(i,j) \in \mathcal{S}_{0,1}^{-,+} \cup \mathcal{S}_{1,0}^{+,-}} \ell_{1,(i,j)}^{(t)} \sum_{p \in [P]} \theta_{i,j}^{(p)}(\mathbf{u}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}') \bigg| \\
&= O\bigg(\frac{\rho}{P}\bigg) \cdot \Theta(\zeta) + O\bigg(\frac{b\alpha\rho}{P}\bigg) \\
&= \Theta\bigg(\frac{\zeta\rho}{P}\bigg),
\end{aligned}$$

where we use the fact that $\zeta = \omega(b\alpha)$.

Proof for $\gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)}, \mathbf{v})$**.** Finally, we will study $\gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)}, \mathbf{v})$. Recall its formula in (C.3) we can get

$$\gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)}, \mathbf{v}) = \frac{1}{n^2} \sum_{i,j \in [n]} \ell_{1,(i,j)}^{(t)} \sum_{p \in [P]} \theta_{i,j}^{(p)}(\boldsymbol{\xi}_s^{(q)}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}).$$

Then it can be seen that the noise vector $\boldsymbol{\xi}_s^{(p)}$ will appear in $2n-1$ mixup data patches. By Lemma C.1, we have with probability at least $1 - 1/\mathrm{poly}(n)$, $\Theta(1/P)$ fraction of them are mixed with $\mathbf{v}$ and $O(b/P)$ fraction of them are mixed with $\alpha\mathbf{v}$. Therefore, we can get that

$$\gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)}, \mathbf{v}) = \underbrace{\frac{1}{n^2} \sum_{p=q, i=s || p=q, j=s} \ell_{1,(i,j)}^{(t)} \theta_{i,j}^{(p)}(\boldsymbol{\xi}_s^{(q)}) \cdot \theta_{i,j}^{(p)}(\mathbf{v})}_{I_1} + \underbrace{\frac{1}{n^2} \sum_{p \neq q || i \neq s, j \neq s} \ell_{1,(i,j)}^{(t)} \sum_{p \in [P]} \theta_{i,j}^{(t)}(\boldsymbol{\xi}_s^{(q)}) \cdot \theta_{i,j}^{(p)}(\mathbf{v})}_{I_2},$$

where it holds that

$$|I_1| \leq \frac{1}{n^2} \cdot \big[\Theta(n/P) + \Theta(b\alpha/P)\big] = \Theta\bigg(\frac{1}{Pn}\bigg),$$

and

$$|I_2| \leq \widetilde{O}\bigg(\frac{P}{d^{1/2}}\bigg),$$

where we use the fact that $b\alpha = o(1)$ and $\theta_{i,j}^{(t)}(\boldsymbol{\xi}_s^{(q)}) = \widetilde{O}(d^{-1/2})$ for all $i \neq s$ and $j \neq s$. This further implies that

$$|\gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)}, \mathbf{v})| = O\bigg(\frac{1}{Pn}\bigg)$$

since we have assumed that $d \geq P^4 n^2$. □

We can also get a similar result for the learning of common feature $\mathbf{u}$.

Lemma C.4. Assume $\max_{k\in[2],(i,j)\in\mathcal{S}} |F_k(\mathbf{W}^{(t)}; \mathbf{x}_{i,j})| \leq \zeta = o\big(\frac{1}{\text{polylog}(n)}\big)$, then recalling the update form in Proposition C.2, we have for any $r \in [m]$, $q \in [P]$, and $s \in [n]$,

$$\gamma_2^{(t)}(\mathbf{u}, \mathbf{u}) = \Theta(1), \quad |\gamma_2^{(t)}(\mathbf{v}, \mathbf{u})| = O(\zeta + \alpha), \quad |\gamma_2^{(t)}(\mathbf{v}', \mathbf{u})| = O(\zeta\rho/P),$$
$$|\gamma_2(\mathbf{u}', \mathbf{u})| = O(\rho/P), \quad |\gamma_2^{(t)}(\boldsymbol{\xi}_s^{(q)}, \mathbf{u})| = \widetilde{O}\big(1/(Pn)\big).$$

C.3.2 Incorrect Common Feature Learning

In this part, we will study the incorrect common feature learning, i.e., quantifying the inner products $\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{v}\rangle$ and $\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle$.

Lemma C.5. Assume $\max_{k\in[2],(i,j)\in\mathcal{S}} |F_k(\mathbf{W}^{(t)}; \mathbf{x}_{i,j})| \leq \zeta = o\big(\frac{1}{\text{polylog}(n)}\big)$, then recalling the update form in Proposition C.2, we have

$$\gamma_2^{(t)}(\mathbf{v}, \mathbf{v}) = -\Theta(1), \quad |\gamma_2^{(t)}(\mathbf{u}, \mathbf{v})| = O(\zeta + \alpha), \quad |\gamma_2^{(t)}(\mathbf{v}', \mathbf{v})| = O(\rho/P),$$
$$|\gamma_2^{(t)}(\mathbf{u}', \mathbf{v})| = O(\zeta\rho/P), \quad |\gamma_2^{(t)}(\boldsymbol{\xi}_s^{(q)}, \mathbf{v})| = \widetilde{O}\big(1/(Pn)\big).$$

Proof of Lemma C.5. Recall the definition of $\gamma_2^{(t)}(\mathbf{v}, \mathbf{v})$, we have

$$\gamma_2^{(t)}(\mathbf{v}, \mathbf{v}) = \frac{1}{n^2} \sum_{i,j\in[n]} \ell_{2,(i,j)}^{(t)} \sum_{p\in[P]} [\theta_{i,j}^{(p)}(\mathbf{v})]^2.$$

Then comparing with the previous analysis on $\gamma_2^{(t)}(\mathbf{v}, \mathbf{v})$, the only difference is to replace $\ell_{1,(i,j)}^{(t)}$ to $\ell_{2,(i,j)}^{(t)} = -\ell_{2,(i,j)}^{(t)}$. Therefore, we can immediately get that $\gamma_2^{(t)}(\mathbf{v}, \mathbf{v}) = -\gamma_1^{(t)}(\mathbf{v}, \mathbf{v}) = -\Theta(1)$.

Regarding other terms that are bounded in terms of their absolute values, we can get the same results as in Theorem C.3. This completes the proof. □

Similarly, we can get the following results for $\mathbf{u}$.

Lemma C.6. Assume $\max_{k\in[2],(i,j)\in\mathcal{S}} |F_k(\mathbf{W}^{(t)}; \mathbf{x}_{i,j})| \leq \zeta = o\big(\frac{1}{\text{polylog}(n)}\big)$, then recalling the update form in Proposition C.2, we have

$$\gamma_1^{(t)}(\mathbf{u}, \mathbf{u}) = -\Theta(1), \quad |\gamma_1^{(t)}(\mathbf{v}, \mathbf{u})| = O(\zeta + \alpha), \quad |\gamma_1^{(t)}(\mathbf{v}', \mathbf{u})| = O(\rho/P),$$
$$|\gamma_1(\mathbf{u}', \mathbf{u})| = O(\zeta\rho/P), \quad |\gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)}, \mathbf{u})| = \widetilde{O}\big(1/(Pn)\big).$$

C.3.3 Rare Feature Learning

In this part, we will study the rare feature learning, i.e., quantifying the inner products $\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle$ and $\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u}'\rangle$.

Lemma C.7. Assume $\max_{k\in[2],(i,j)\in\mathcal{S}} |F_k(\mathbf{W}^{(t)};\mathbf{x}_{i,j})| \leq \zeta$ for some $\zeta = o\big(\frac{1}{\mathrm{polylog}(n)}\big)$ and $\zeta > b\alpha$, then recalling the update form in Proposition C.2, we have

$$\gamma_1^{(t)}(\mathbf{v}',\mathbf{v}') = \Theta(\rho), \quad \gamma_1^{(t)}(\mathbf{v},\mathbf{v}') = \Theta(\rho/P), \quad |\gamma_1^{(t)}(\mathbf{u},\mathbf{v}')| = O(\zeta\rho/P),$$
$$|\gamma_1^{(t)}(\mathbf{u}',\mathbf{v}')| = O(\zeta\rho^2/P), \quad |\gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)},\mathbf{v}')| = \widetilde{O}\big(\rho/(Pn)\big).$$

Proof of Lemma C.7. Recalling the definition of $\gamma_1^{(t)}(\mathbf{v}',\mathbf{v}')$:

$$\gamma_1^{(t)}(\mathbf{v}',\mathbf{v}') = \frac{1}{n^2}\sum_{i,j\in[n]} \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} [\theta_{i,j}^{(p)}(\mathbf{v}')]^2.$$

Note that the rare feature $\mathbf{v}'$ will not appear in the form of feature noise, then we will only need to focus on the mixed data (i,j) with either $i \in \mathcal{S}_1^+$ or $j \in \mathcal{S}_1^+$, where the rare feature can only appear in the form of $\mathbf{v}$, $\lambda\mathbf{v}$, or $(1-\lambda)\mathbf{v}$. Particularly, regarding the data $(i,j) \in \mathcal{S}_{1,1}^{+,+} \cup \mathcal{S}_{1,0}^{+,+} \cup_{0,1}^{+,+}$, let $\mathcal{P}_{i,j}^*(\mathbf{v}')$ be the set of patches that contain the feature $\mathbf{v}'$, we have $|\mathcal{P}_{i,j}^*(\mathbf{v}')| = \Theta(1)$ and then

$$\ell_{1,(i,j)}^{(t)} \cdot \sum_{p\in[P]} [\theta_{i,j}^{(p)}(\mathbf{v}')]^2 = \ell_{1,(i,j)}^{(t)} \cdot \sum_{p\in\mathcal{P}_{i,j}^*(\mathbf{v})} [\theta_{i,j}^{(p)}(\mathbf{v}')]^2 = \Theta(1),$$

where we use the fact that $\ell_{1,(i,j)}^{(t)} = \Theta(1)$ for any $(i,j) \in \mathcal{S}_{1,1}^{+,+}$.

Regarding the data $(i,j) \in \mathcal{S}_{1,0}^{+,-} \cup \mathcal{S}_{1,1}^{+,-}$, we will consider (i,j) and (j,i) together. Particularly, we have

$$\ell_{1,(i,j)}^{(t)} \cdot \sum_{p\in[P]} [\theta_{i,j}^{(p)}(\mathbf{v}')]^2 + \ell_{1,(j,i)}^{(t)} \cdot \sum_{p\in[P]} [\theta_{j,i}^{(p)}(\mathbf{v}')]^2 = \underbrace{\ell_{1,(i,j)}^{(t)} \cdot \sum_{p\in[P]} \Big[[\theta_{i,j}^{(p)}(\mathbf{v}')]^2 - [\theta_{j,i}^{(p)}(\mathbf{v}')]^2\Big]}_{I_1}$$
$$+ \underbrace{\big[\ell_{1,(i,j)}^{(t)} + \ell_{1,(j,i)}^{(t)}\big] \cdot \sum_{p\in[P]} [\theta_{j,i}^{(p)}(\mathbf{v}')]^2}_{I_2}.$$

Then using the same definition of $\mathcal{P}_{i,j}^*(\mathbf{v})$, we have for any $p \in \mathcal{P}_{i,j}^*(\mathbf{v})$, it holds that $\theta_{i,j}^{(p)}(\mathbf{v}') = \lambda$ and $\theta_{j,i}^{(p)}(\mathbf{v}') = 1-\lambda$, then

$$I_1 = \Theta(1) \cdot |\mathcal{P}_{i,j}^*(\mathbf{v}')| \cdot [\lambda^2 - (1-\lambda)^2] = \Theta(1).$$

Regarding I_2, we can use the condition that the neural network output is upper bounded by ζ, then

$$|I_2| = \Big|\big[\lambda - 0.5 + 0.5 - \lambda \pm O(\zeta)\big] \cdot \sum_{p\in\mathcal{P}_{i,j}^*(\mathbf{v}')} [\theta_{j,i}^{(p)}(\mathbf{v}')]^2\Big| = O(\zeta).$$

Therefore, combining these results for I_1 and I_2, we can get

$$\ell_{1,(i,j)}^{(t)} \cdot \sum_{p\in[P]} [\theta_{i,j}^{(p)}(\mathbf{v}')]^2 + \ell_{1,(j,i)}^{(t)} \cdot \sum_{p\in[P]} [\theta_{j,i}^{(p)}(\mathbf{v}')]^2 = I_1 + I_2 = \Theta(1).$$

To complete the analysis, we have

$$\begin{aligned}
\gamma_1^{(t)}(\mathbf{v}',\mathbf{v}') &= \frac{1}{n^2}\sum_{i,j\in[n]}\sum_{p\in[P]} [\theta_{i,j}^{(p)}(\mathbf{v}')]^2 \\
&= \frac{1}{n^2}\sum_{i\in\mathcal{S}_1^+, j\in[n] || i\in[n], j\in\mathcal{S}_1^+}\sum_{p\in[P]} [\theta_{i,j}^{(p)}(\mathbf{v}')]^2 \\
&= |\mathcal{S}_{1,1}^{+,+} \cup \mathcal{S}_{1,0}^{+,+} \cup_{0,1}^{+,+}| + |\mathcal{S}_{1,0}^{+,-} \cup \mathcal{S}_{1,1}^{+,-}|
\end{aligned}$$

$$= \Theta(\rho).$$

The characterization of $\gamma_1(\mathbf{v}, \mathbf{v}')$ and $\gamma_1(\mathbf{u}, \mathbf{v}')$ will be exactly the same as $\gamma_1(\mathbf{v}', \mathbf{v})$ and $\gamma_1(\mathbf{v}', \mathbf{u})$ due to the fact that $\gamma_1(\mathbf{a}, \boldsymbol{b}) = \gamma_1(\boldsymbol{b}, \mathbf{a})$. Therefore, we can apply Lemmas C.3 and C.6 to get the desired results.

Regarding the proof for $\gamma_1^{(t)}(\mathbf{u}', \mathbf{v}')$, we will follow a similar proof for $\gamma_1^{(t)}(\mathbf{u}', \mathbf{v})$ in Lemma C.3, while two differences need to be considered: (1) the rare feature vectors $\mathbf{u}'$ and $\mathbf{v}'$ will not appear in the form of feature noise, thus we only need to consider the data $(i, j) \in \mathcal{S}_{1,1}^{+,-} \cup \mathcal{S}_{1,1}^{-,+}$; (2) the cardinality of the critical subset of data satisfies $|\mathcal{S}_{1,1}^{+,-} \cup \mathcal{S}_{1,1}^{-,+}| = \rho^2 n^2$. Therefore, for any $(i, j) \in \mathcal{S}_{1,1}^{+,-}$, we have

$$\begin{aligned}
&\ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\mathbf{u}') \cdot \theta_{i,j}^{(p)}(\mathbf{v}') + \ell_{1,(j,i)}^{(t)} \sum_{p\in[P]} \theta_{j,i}^{(p)}(\mathbf{u}') \cdot \theta_{j,i}^{(p)}(\mathbf{v}') \\
&= \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \big[\theta_{i,j}^{(p)}(\mathbf{u}') \cdot \theta_{i,j}^{(p)}(\mathbf{v}') - \theta_{j,i}^{(p)}(\mathbf{u}') \cdot \theta_{j,i}^{(p)}(\mathbf{v}')\big] + \big[\ell_{1,(i,j)}^{(t)} + \ell_{1,(j,i)}^{(t)}\big] \cdot \sum_{p\in[P]} \theta_{j,i}^{(p)}(\mathbf{u}') \cdot \theta_{j,i}^{(p)}(\mathbf{v}').
\end{aligned}$$

It is easy to see that $\theta_{i,j}^{(p)}(\mathbf{u}') \cdot \theta_{i,j}^{(p)}(\mathbf{v}') = \theta_{j,i}^{(p)}(\mathbf{u}') \cdot \theta_{j,i}^{(p)}(\mathbf{v}') = \lambda(1-\lambda)$. Besides, we have in total $\rho^2 n^2/P$ patches that consist of both $\mathbf{u}'$ and $\mathbf{v}'$. This further implies that

$$\begin{aligned}
|\gamma_1^{(t)}(\mathbf{u}', \mathbf{v}')| &= \bigg| \frac{1}{n^2} \sum_{(i,j)\in\mathcal{S}_{1,1}^{+,-}\cup\mathcal{S}_{1,1}^{-,+}} \big[\ell_{1,(i,j)}^{(t)} + \ell_{1,(j,i)}^{(t)}\big] \cdot \sum_{p\in[P]} \theta_{j,i}^{(p)}(\mathbf{u}') \cdot \theta_{j,i}^{(p)}(\mathbf{v}') \bigg| \\
&= O\bigg(\frac{\rho^2}{P}\bigg) \cdot O(\zeta) \\
&= O\bigg(\frac{\zeta\rho^2}{P}\bigg),
\end{aligned}$$

where we use the fact that $|\ell_{1,(i,j)}^{(t)} + \ell_{1,(j,i)}^{(t)}| = O(\zeta)$.

Lastly, we will characterize $\gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)}, \mathbf{v}')$. First recall its definition:

$$\begin{aligned}
\gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)}, \mathbf{v}') &= \frac{1}{n^2} \sum_{i,j\in[n]} \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\boldsymbol{\xi}_s^{(q)}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}') \\
&= \underbrace{\frac{1}{n^2} \sum_{p=q, i=s || p=q, j=s} \ell_{1,(i,j)}^{(t)} \theta_{i,j}^{(p)}(\boldsymbol{\xi}_s^{(q)}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}')}_{I_1} + \underbrace{\frac{1}{n^2} \sum_{p\neq q || i\neq s, j\neq s} \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\boldsymbol{\xi}_s^{(q)}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}')}_{I_2}.
\end{aligned}$$

Note that for any fixed $\boldsymbol{\xi}_s^{(p)}$, it will be mixed with n data patches in total, while, by Lemma C.1, we know that there are only $\Theta(\rho/P)$ fraction among them are $\mathbf{v}'$. Using the fact that $|\ell_{1,(i,j)}^{(t)}| \leq 1$, we have

$$|I_1| \leq \frac{1}{n} \cdot \Theta(\rho/P) = O\bigg(\frac{\rho}{Pn}\bigg).$$

Besides, note that $|\theta_{i,j}^{(p)}(\boldsymbol{\xi}_s^{(q)})| = \widetilde{O}(d^{-1/2})$ if $i, j \neq s$ or $p \neq q$, we have

$$|I_2| \leq \widetilde{O}\bigg(\frac{\rho P}{d^{1/2}}\bigg) = O\bigg(\frac{\rho}{Pn}\bigg),$$

where the last equality is by the assumption that $d \geq P^4 n^2$. Combining the above results for I_1 and I_2, we can get

$$|\gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)}, \mathbf{v}')| = O\bigg(\frac{\rho}{Pn}\bigg).$$

□

Following the exactly same procedure, we can get the following results regarding the learning of $\mathbf{u}'$.

Lemma C.8. Assume $\max_{k\in[2],(i,j)\in\mathcal{S}} |F_k(\mathbf{W}^{(t)};\mathbf{x}_{i,j})| \leq \zeta$ for some $\zeta = o\big(\frac{1}{\mathrm{polylog}(n)}\big)$ and $\zeta > b\alpha$, then recalling the update form in Proposition C.2, we have

$$\gamma_2^{(t)}(\mathbf{u}',\mathbf{u}') = \Theta(\rho), \quad \gamma_2^{(t)}(\mathbf{u},\mathbf{u}') = \Theta(\rho/P), \quad |\gamma_2^{(t)}(\mathbf{v},\mathbf{u}')| = O(\zeta\rho/P),$$
$$|\gamma_2^{(t)}(\mathbf{v}',\mathbf{u}')| = O(\zeta\rho^2/P), \quad |\gamma_2^{(t)}(\boldsymbol{\xi}_s^{(q)},\mathbf{u}')| = \widetilde{O}\big(\rho/(Pn)\big).$$

C.3.4 Incorrect Rare Feature Learning

In contrast to the previous section that studies $\langle \mathbf{w}_{1,r}^{(t)},\mathbf{v}'\rangle$ and $\langle \mathbf{w}_{2,r}^{(t)},\mathbf{u}'\rangle$, the incorrect rare feature learning aims to characterize the quantities $\langle \mathbf{w}_{2,r}^{(t)},\mathbf{v}'\rangle$ and $\langle \mathbf{w}_{1,r}^{(t)},\mathbf{u}'\rangle$. Similar to the proof of Lemmas C.5 and C.6, we only need to replace $\ell_{1,(i,j)}^{(t)}$ with $\ell_{2,(i,j)}^{(t)} = -\ell_{1,(i,j)}^{(t)}$ or $\ell_{2,(i,j)}^{(t)}$ with $\ell_{1,(i,j)}^{(t)} = -\ell_{2,(i,j)}^{(t)}$. Based on this, the update of $\langle \mathbf{w}_{2,r}^{(t)},\mathbf{v}'\rangle$ and $\langle \mathbf{w}_{1,r}^{(t)},\mathbf{u}'\rangle$ in each iteration are characterized in the following lemmas.

Lemma C.9. Assume $\max_{k\in[2],(i,j)\in\mathcal{S}} |F_k(\mathbf{W}^{(t)};\mathbf{x}_{i,j})| \leq \zeta$ for some $\zeta = o\big(\frac{1}{\mathrm{polylog}(n)}\big)$ and $\zeta > b\alpha$, then recalling the update form in Proposition C.2, we have

$$\gamma_2^{(t)}(\mathbf{v}',\mathbf{v}') = -\Theta(\rho), \quad \gamma_2^{(t)}(\mathbf{v},\mathbf{v}') = -\Theta(\rho/P), \quad |\gamma_2^{(t)}(\mathbf{u},\mathbf{v}')| = O(\zeta\rho/P),$$
$$|\gamma_2^{(t)}(\mathbf{u}',\mathbf{v}')| = O(\zeta\rho^2/P), \quad |\gamma_2^{(t)}(\boldsymbol{\xi}_s^{(q)},\mathbf{v}')| = \widetilde{O}\big(\rho/(Pn)\big).$$

Lemma C.10. Assume $\max_{k\in[2],(i,j)\in\mathcal{S}} |F_k(\mathbf{W}^{(t)};\mathbf{x}_{i,j})| \leq \zeta$ for some $\zeta = o\big(\frac{1}{\mathrm{polylog}(n)}\big)$ and $\zeta > b\alpha$, then recalling the update form in Proposition C.2, we have

$$\gamma_1^{(t)}(\mathbf{u}',\mathbf{u}') = -\Theta(\rho), \quad \gamma_1^{(t)}(\mathbf{u},\mathbf{u}') = -\Theta(\rho/P), \quad |\gamma_1^{(t)}(\mathbf{v},\mathbf{u}')| = O(\zeta\rho/P),$$
$$|\gamma_1^{(t)}(\mathbf{v}',\mathbf{u}')| = O(\zeta\rho^2/P), \quad |\gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)},\mathbf{u}'| = \widetilde{O}\big(\rho/(Pn)\big).$$

C.3.5 Noise Learning

Lemma C.11. Assume $\max_{k\in[2],(i,j)\in\mathcal{S}} |F_k(\mathbf{W}^{(t)};\mathbf{x}_{i,j})| \leq \zeta$ for some $\zeta = o\big(\frac{1}{\mathrm{polylog}(n)}\big)$ and $\zeta > b\alpha$, then recalling the update form in Proposition C.2, for any $\boldsymbol{\xi}_s^{(q)}$ with $y_s = 1$, we have

$$\gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)},\boldsymbol{\xi}_s^{(q)}) = \frac{d\sigma_p^2\cdot[n\lambda^3-(2\lambda-1)(1-\lambda)^2]}{2n^2} \pm \widetilde{O}\bigg(\frac{\zeta d\sigma_p^2}{n}\bigg)$$
$$|\gamma_1^{(t)}(\mathbf{v},\boldsymbol{\xi}_s^{(q)})| = O\big(d\sigma_p^2/(Pn)\big), \quad |\gamma_1^{(t)}(\mathbf{u},\boldsymbol{\xi}_s^{(q)})| = O\big(d\sigma_p^2/(Pn)\big), \quad |\gamma_1^{(t)}(\mathbf{v}',\boldsymbol{\xi}_s^{(q)})| = O\big(d\sigma_p^2\rho/(Pn)\big),$$
$$|\gamma_1^{(t)}(\mathbf{u}',\boldsymbol{\xi}_s^{(q)})| = O\big(d\sigma_p^2\rho/(Pn)\big), \quad |\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(q)},\boldsymbol{\xi}_s^{(q)})| = \mathbb{1}[y_i=y_s]\cdot\frac{\lambda(1-\lambda)d\sigma_p^2}{n^2} \pm O\bigg(\frac{\zeta d\sigma_p^2}{n^2}\bigg).$$

Proof of Lemma C.11. Without loss of generality, we assume $y_s = 1$. According to the definition of $\gamma_1^{(t)}\big(\boldsymbol{\xi}_s^{(q)},\boldsymbol{\xi}_s^{(q)}\big)$, we have

$$\begin{aligned}
\gamma_1^{(t)}\big(\boldsymbol{\xi}_s^{(q)},\boldsymbol{\xi}_s^{(q)}\big) &= \frac{1}{n^2}\sum_{i,j\in[n]}\ell_{1,(i,j)}^{(t)}\sum_{p\in[P]}[\theta_{i,j}^{(p)}(\boldsymbol{\xi}_s^{(q)})]^2\cdot\|\boldsymbol{\xi}_s^{(q)}\|_2^2 \\
&= \frac{\|\boldsymbol{\xi}_s^{(q)}\|_2^2}{n^2}\cdot\bigg(\sum_{i\in[n]}\ell_{1,(s,i)}^{(t)}[\theta_{s,i}^{(q)}(\boldsymbol{\xi}_s^{(q)})]^2+\sum_{i\neq s}\ell_{1,(i,s)}^{(t)}[\theta_{i,s}^{(q)}(\boldsymbol{\xi}_s^{(q)})]^2\bigg) \\
&= \frac{\|\boldsymbol{\xi}_s^{(q)}\|_2^2}{n^2}\cdot\bigg(\lambda^2\cdot\sum_{i\in[n]}\ell_{1,(s,i)}^{(t)}+(1-\lambda)^2\cdot\sum_{i\neq s}\ell_{1,(i,s)}^{(t)}\bigg) \\
&= \frac{\|\boldsymbol{\xi}_s^{(q)}\|_2^2}{n^2}\cdot\big[0.5n\lambda^3-(\lambda-0.5)(1-\lambda)^2\pm O(n\zeta)\big]
\end{aligned}$$

$$= \frac{\|\boldsymbol{\xi}_s^{(q)}\|_2^2 \cdot [n\lambda^3 - (2\lambda-1)(1-\lambda)^2]}{2n^2} \pm O\bigg(\frac{\zeta\|\boldsymbol{\xi}_s^{(q)}\|_2^2}{n}\bigg),$$

where the second equation is due to the fact that only $\mathbf{x}_{i,s}$ or $\mathbf{x}_{s,i}$ will contain the component of $\boldsymbol{\xi}_s^{(q)}$, the fourth inequality holds since we assume there have $n/2$ positive samples and $n/2$ negative samples in the training data. Moreover, note that $\boldsymbol{\xi}_s^{(q)} \sim N(0, \sigma_p^2\mathbf{I})$, applying union bound over all $s \in [n]$ and $p \in [P]$, we can get that with probability at least $1 - 1/\mathrm{poly}(n)$, we have

$$\big|\|\boldsymbol{\xi}_s^{(q)}\|_2^2 - d\sigma_p^2\big| \leq \mathrm{polylog}(n) \cdot d^{1/2}\sigma_p^2.$$

Therefore, it follows that for all $s \in [n]$ and $p \in [P]$, with probability at least $1 - 1/\mathrm{poly}(n)$,

$$\begin{aligned}\gamma_1^{(t)}\big(\boldsymbol{\xi}_s^{(q)}, \boldsymbol{\xi}_s^{(q)}\big) &= \frac{d\sigma_p^2 \cdot [n\lambda^3 - (2\lambda-1)(1-\lambda)^2]}{2n^2} \pm \widetilde{O}\bigg(\frac{d^{1/2}\sigma_p^2}{n} + \frac{\zeta d\sigma_p^2}{n}\bigg)\\ &= \frac{d\sigma_p^2 \cdot [n\lambda^2 - (1-\lambda)^2]}{2n^2} \pm \widetilde{O}\bigg(\frac{\zeta d\sigma_p^2}{n}\bigg),\end{aligned}$$

where we use the fact that $\zeta = \omega(d^{-1/2})$.

Regarding $\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(q)}, \boldsymbol{\xi}_s^{(q)})$, we have

$$\begin{aligned}\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(q)}, \boldsymbol{\xi}_s^{(q)}) &= \frac{\lambda(1-\lambda)}{n^2} \cdot [\ell_{1,(i,s)}^{(t)} + \ell_{1,(s,i)}^{(t)}] \cdot \|\boldsymbol{\xi}_s^{(q)}\|_2^2\\ &= \mathbb{1}[y_i = y_s] \cdot \frac{\lambda(1-\lambda)d\sigma_p^2}{n^2} \pm O\bigg(\frac{\zeta d\sigma_p^2}{n^2}\bigg).\end{aligned}$$

Regarding the remaining quantities, we can directly apply the aforementioned lemmas on the learning of common and rare features, since the following holds

$$\gamma_1^{(t)}(\mathbf{a}, \boldsymbol{\xi}_s^{(q)}) = \gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)}, \mathbf{a}) \cdot \|\boldsymbol{\xi}_s^{(q)}\|_2^2 = \gamma_1^{(t)}(\boldsymbol{\xi}_s^{(q)}, \mathbf{a}) \cdot \widetilde{O}(d\sigma_p^2),$$

where $\mathbf{a} \in \{\mathbf{v}, \mathbf{u}, \mathbf{v}', \mathbf{u}'\}$. This completes the proof.

□

C.4 Outcome of Phase 1 Mixup Training.

In this part, we will provide the outcome of Phase 1 mixup training.

We first recall Proposition C.2 and Lemma C.3 to obtain the learning dynamics of the common feature vector $\mathbf{v}$.

$$\begin{aligned}\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle &= \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle - \eta \cdot \langle \nabla_{\mathbf{w}_{1,r}} L_{\mathcal{S}}(\mathbf{W}^{(t)}), \mathbf{v}\rangle\\ &= \big[1 + \eta\gamma_1^{(t)}(\mathbf{v}, \mathbf{v})\big] \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle + \eta\gamma_1^{(t)}(\mathbf{u}, \mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle + \eta\gamma_1^{(t)}(\mathbf{v}', \mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle\\ &\quad + \eta\gamma_1^{(t)}(\mathbf{u}', \mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle + \sum_{i=1}^{n}\sum_{p\in[P]} \eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle.\end{aligned}$$

Then it can be seen that the most complicated part in the above update form is the composition of noise learning, i.e., $\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle$.The following lemma provides an upper bound on the term $\sum_{i=1}^n \sum_{p\in[P]} \eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle$, which will leverage the randomness of $\boldsymbol{\xi}_i^{(p)}$ at the initialization.

Lemma C.12. Assume $\max_{k\in[2],(i,j)\in\mathcal{S}} |F_k(\mathbf{W}^{(t)}; \mathbf{x}_{i,j})| \leq \zeta$ for some $\zeta \in \big[\omega\big((nP)^{-1/2}\big), o\big(\frac{1}{\mathrm{polylog}(n)}\big)\big]$. Let $z_t := \sum_{i=1}^n \sum_{p\in[P]} \gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle$, then we have with probability at least $1 - 1/\mathrm{poly}(n)$, for all $t = O\big(n\eta^{-1}/(d\sigma_p^2)\big)$, we have

$$|z_t| \leq O\bigg(\frac{d^{1/2}\sigma_0\sigma_p}{P^{1/2}n^{1/2}}\bigg) + O\bigg(\frac{\zeta}{Pn}\bigg) \cdot \sum_{s=1}^{n}\sum_{p\in[P]} |\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_s^{(p)}\rangle| + O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg) \cdot \sum_{\tau=0}^{t-1} \big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}\rangle|\big]$$

$$+ O\bigg(\frac{\eta\rho d\sigma_p^2}{Pn}\bigg) \cdot \sum_{\tau=0}^{t-1} \big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}'\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}'\rangle|\big] + O\bigg(\frac{\eta\zeta d\sigma_p^2}{n^2}\bigg) \cdot \sum_{\tau=0}^{t-1} \sum_{p\in[P]} \sum_{s=1}^{n} |\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_s^{(p)}\rangle|.$$

Proof. Based on the definition of z_t, we can conduct the following decomposition:

$$\begin{aligned} z_t &= \sum_{i=1}^{n} \sum_{p\in[P]} \gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle \\ &= \frac{1}{n^2} \sum_{i'=1}^{n} \sum_{q\in[P]} \sum_{i,j\in[n]} \ell_{1,(i,j)}^{(t)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\boldsymbol{\xi}_{i'}^{(q)}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_{i'}^{(q)}\rangle. \end{aligned}$$

Note that during the initial training phase $\ell_{1,(i,j)}^{(t)}$ is close to the constant $l_{1,(i,j)} \in \{0.5, -0.5, 0.5-\lambda, \lambda-0.5\}$, which is independent of the random noise vectors $\{\boldsymbol{\xi}\}$ and random initial weights $\{\mathbf{w}_{1,r}^{(0)}\}_{r\in[m]}$. Then using the fact that $|\ell_{1,(i,j)}^{(t)} - l_{1,(i,j)}| = O(\zeta)$, we can get

$$\begin{aligned} |z_0| \le & \underbrace{\bigg| \frac{1}{n^2} \sum_{i'=1}^{n} \sum_{q\in[P]} \sum_{i,j\in[n]} l_{1,(i,j)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\boldsymbol{\xi}_{i'}^{(q)}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\xi}_{i'}^{(q)}\rangle \bigg|}_{I_1} \\ & + \underbrace{\bigg| \frac{1}{n^2} \sum_{i'=1}^{n} \sum_{q\in[P]} \sum_{i,j\in[n]} [\ell_{1,(i,j)}^{(0)} - l_{1,(i,j)}] \sum_{p\in[P]} \theta_{i,j}^{(p)}(\boldsymbol{\xi}_{i'}^{(q)}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\xi}_{i'}^{(q)}\rangle \bigg|}_{I_2}. \end{aligned}$$

Regarding I_1, note that $\ell_{1,(i,j)}$, $\theta_{i,j}^{(p)}(\boldsymbol{\xi}_{i'}^{(q)})$, and $\theta_{i,j}^{(p)}(\mathbf{v})$ are independent of the random noise vectors $\{\boldsymbol{\xi}\}$ and random initial weights $\{\mathbf{w}_{1,r}^{(0)}\}_{r\in[m]}$. Besides, note that the inner products $\{\langle \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\xi}_{i'}^{(q)}\rangle\}_{i'\in[n],q\in[P]}$ are independent conditioning on $\mathbf{w}_{1,r}^{(0)}$ and for all $i' \in [n]$ and $q \in [P]$, . We can apply standard concentration arguments to get the upper bound of I_1. Before approaching this, we first apply Lemma C.1 and follow the similar proof of Lemma C.3, and obtain that with probability at least $1 - 1/\mathrm{poly}(n)$

$$\bigg| \frac{1}{n^2} \sum_{i,j\in[n]} l_{1,(i,j)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\boldsymbol{\xi}_{i'}^{(q)}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}) \bigg| = O\bigg(\frac{1}{nP}\bigg). \tag{C.9}$$

Then performing the following decomposition on I_1 according to the value of $y_{i'}$:

$$\begin{aligned} I_1 = & \underbrace{\frac{1}{n^2} \sum_{i':y_{i'}=1} \sum_{q\in[P]} \sum_{i,j\in[n]} l_{1,(i,j)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\boldsymbol{\xi}_{i'}^{(q)}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\xi}_{i'}^{(q)}\rangle}_{I_1^{(1)}} \\ & + \underbrace{\frac{1}{n^2} \sum_{i':y_{i'}=2} \sum_{q\in[P]} \sum_{i,j\in[n]} l_{1,(i,j)} \sum_{p\in[P]} \theta_{i,j}^{(p)}(\boldsymbol{\xi}_{i'}^{(q)}) \cdot \theta_{i,j}^{(p)}(\mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\xi}_{i'}^{(q)}\rangle}_{I_1^{(2)}}. \end{aligned}$$

Therefore, note that conditioning on $\mathbf{w}_{1,r}^{(0)}$, the quantity $\langle \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\xi}_{i'}^{(q)}\rangle$ is $\|\mathbf{w}_{1,r}^{(0)}\|_2 \cdot \sigma_p$-subGaussian, by (C.9), we can immediately get that both $I_1^{(1)}$ and $I_1^{(2)}$ are $\|\mathbf{w}_{1,r}^{(0)}\|_2 \cdot \sigma_p \cdot (nP)^{-1/2}$-subGuassian. Then using the fact that $\mathbf{w}_{1,r}^{(0)} \in N(0, \sigma_0^2\mathbf{I})$, we can get that with probability at least $1 - 1/\mathrm{poly}(n)$,

$$|I_1^{(1)}|, |I_1^{(2)}| \le \widetilde{O}\bigg(\frac{d^{1/2}\sigma_0\sigma_p}{(nP)^{1/2}}\bigg). \tag{C.10}$$

Regarding I_2, we can also apply Lemma C.1 and follow the similar proof of Lemma C.3, then with probability at least $1-1/\text{poly}(n)$,

$$\bigg|\frac{1}{n^2}\sum_{i,j\in[n]}[\ell_{1,(i,j)}^{(0)}-l_{1,(i,j)}]\sum_{p\in[P]}\theta_{i,j}^{(p)}(\boldsymbol{\xi}_{i'}^{(q)})\cdot\theta_{i,j}^{(p)}(\mathbf{v})\bigg| = O\bigg(\frac{\max_{i,j}|\ell_{1,(i,j)}^{(0)}-l_{1,(i,j)}|}{nP}\bigg) = O\bigg(\frac{\zeta}{nP}\bigg).$$

This further implies that

$$I_2 \le O\bigg(\frac{\zeta}{nP}\bigg)\cdot\sum_{i'=1}^{n}\sum_{q\in[P]}|\langle\mathbf{w}_{1,r}^{(0)},\boldsymbol{\xi}_{i'}^{(q)}\rangle| \le \widetilde{O}\big(\zeta d^{1/2}\sigma_0\sigma_p\big). \tag{C.11}$$

where we use the fact that $\mathbf{w}_{1,r}^{(0)}\sim N(0,\sigma_0^2\mathbf{I})$ and $\boldsymbol{\xi}_{i'}^{(q)}\sim N(0,\sigma_p^2\mathbf{I})$. Combining (C.10) and (C.11) leads to

$$|z_0| \le \widetilde{O}\bigg(\frac{d^{1/2}\sigma_0\sigma_p}{(nP)^{1/2}}+\zeta d^{1/2}\sigma_0\sigma_p\bigg) = O\big(\zeta d^{1/2}\sigma_0\sigma_p\big).$$

where we use the condition that $\zeta=\omega\big((nP)^{-1/2}\big)$.

Next we will move on to study the update of z_t using the update results of $\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_i^{(p)}\rangle$ in Lemma C.11. Particularly, we can again use the quantities $l_{1,(i,j)}$'s and get the following decomposition

$$\begin{aligned} z_t = &\underbrace{\frac{1}{n^2}\sum_{i'=1}^{n}\sum_{q\in[P]}\sum_{i,j\in[n]}l_{1,(i,j)}\sum_{p\in[P]}\theta_{i,j}^{(p)}(\boldsymbol{\xi}_{i'}^{(q)})\cdot\theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_{i'}^{(q)}\rangle}_{I_3}\\ &+\underbrace{\frac{1}{n^2}\sum_{i'=1}^{n}\sum_{q\in[P]}\sum_{i,j\in[n]}\big[\ell_{1,(i,j)}^{(t)}-l_{1,(i,j)}\big]\sum_{p\in[P]}\theta_{i,j}^{(p)}(\boldsymbol{\xi}_{i'}^{(q)})\cdot\theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_{i'}^{(q)}\rangle}_{I_4}. \end{aligned}$$

Recall the update results of $\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_i^{(p)}\rangle$ in Lemma C.11: for any $y_i=1$,

$$\begin{aligned} \langle\mathbf{w}_{1,r}^{(t+1)},\boldsymbol{\xi}_i^{(p)}\rangle = &\bigg[1+\eta\cdot\bigg(\frac{d\sigma_p^2\cdot[n\lambda^3-(2\lambda-1)(1-\lambda)^2]}{2n^2}\pm\widetilde{O}\bigg(\frac{\zeta d\sigma_p^2}{n}\bigg)\bigg)\bigg]\cdot\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_i^{(p)}\rangle\pm O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg)\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle\\ &\pm O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg)\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{u}\rangle\pm O\bigg(\frac{\eta d\sigma_p^2\rho}{Pn}\bigg)\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}'\rangle\pm O\bigg(\frac{\eta d\sigma_p^2\rho}{Pn}\bigg)\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{u}'\rangle\\ &+O\bigg(\frac{\eta\lambda(1-\lambda)d\sigma_p^2}{n^2}\bigg)\sum_{s:y_s=1}\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle\pm O\bigg(\frac{\eta\zeta d\sigma_p^2}{n^2}\bigg)\cdot\sum_{s=1}^{n}|\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle|. \end{aligned}\tag{C.12}$$

For any $y_i=2$, we have

$$\begin{aligned} \langle\mathbf{w}_{1,r}^{(t+1)},\boldsymbol{\xi}_i^{(p)}\rangle = &\bigg[1-\eta\cdot\bigg(\frac{d\sigma_p^2\cdot[n\lambda^3-(2\lambda-1)(1-\lambda)^2]}{2n^2}\pm\widetilde{O}\bigg(\frac{\zeta d\sigma_p^2}{n}\bigg)\bigg)\bigg]\cdot\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_i^{(p)}\rangle\pm O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg)\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle\\ &\pm O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg)\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{u}\rangle\pm O\bigg(\frac{\eta d\sigma_p^2\rho}{Pn}\bigg)\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}'\rangle\pm O\bigg(\frac{\eta d\sigma_p^2\rho}{Pn}\bigg)\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{u}'\rangle\\ &-O\bigg(\frac{\eta\lambda(1-\lambda)d\sigma_p^2}{n^2}\bigg)\cdot\sum_{s:y_s=1}\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle\pm O\bigg(\frac{\eta\zeta d\sigma_p^2}{n^2}\bigg)\cdot\sum_{s=1}^{n}|\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle|. \end{aligned}\tag{C.13}$$

We first prove the bound of the quantity $\sum_{p\in[P]}\sum_{s:y_s=1}\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle$. First, using the standard concentration result gives $|\sum_{p\in[P]}\sum_{s:y_s=1}\langle\mathbf{w}_{1,r}^{(0)},\boldsymbol{\xi}_s^{(p)}\rangle| = \widetilde{O}\big(d^{1/2}\sigma_0\sigma_pP^{1/2}n^{1/2}\big)$. Then, by the above update rule, we can get

$$\sum_{p\in[P]}\sum_{s:y_s=1}\langle\mathbf{w}_{1,r}^{(t+1)},\boldsymbol{\xi}_s^{(p)}\rangle = \bigg[1+\Theta\bigg(\frac{\eta d\sigma_p^2}{n}\bigg)\bigg]\cdot\sum_{p\in[P]}\sum_{s:y_s=1}\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle$$

$$\pm O\big(\eta d\sigma_p^2\big)\cdot\langle \mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle \pm O\big(\eta d\sigma_p^2\big)\cdot\langle \mathbf{w}_{1,r}^{(t)},\mathbf{u}\rangle \pm O\big(\eta\rho d\sigma_p^2\big)\cdot\langle \mathbf{w}_{1,r}^{(t)},\mathbf{v}'\rangle$$
$$\pm O\big(\eta\rho d\sigma_p^2\big)\cdot\langle \mathbf{w}_{1,r}^{(t)},\mathbf{u}'\rangle \pm O\bigg(\frac{\eta\zeta d\sigma_p^2}{n}\bigg)\cdot\sum_{p\in[P]}\sum_{s=1}^{n}|\langle \mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle|.$$

Then we can get that for any $t=O\big(n\eta^{-1}/(d\sigma_p^2)\big)$, we have

$$\bigg|\sum_{p\in[P]}\sum_{s:y_s=1}\langle \mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle\bigg| \le O\big(n^{1/2}P^{1/2}d^{1/2}\sigma_0\sigma_p\big)+O(\eta d\sigma_p^2)\cdot\sum_{\tau=0}^{t-1}\big[|\langle \mathbf{w}_{1,r}^{(\tau)},\mathbf{v}\rangle|+|\langle \mathbf{w}_{1,r}^{(\tau)},\mathbf{u}\rangle|\big]$$
$$+O(\eta\rho d\sigma_p^2)\cdot\sum_{\tau=0}^{t-1}\big[|\langle \mathbf{w}_{1,r}^{(\tau)},\mathbf{v}'\rangle|+|\langle \mathbf{w}_{1,r}^{(\tau)},\mathbf{u}'\rangle|\big]$$
$$+O\bigg(\frac{\eta\zeta d\sigma_p^2}{n}\bigg)\cdot\sum_{\tau=0}^{t-1}\sum_{p\in[P]}\sum_{s=1}^{n}|\langle \mathbf{w}_{1,r}^{(\tau)},\boldsymbol{\xi}_s^{(p)}\rangle|.$$

Moreover, similar result can be obtained for $\sum_{s:y_s=2}\langle \mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle$ and we omit the proof here.

Now we are ready to upper bound I_3. Particularly, let $\alpha_1^{(t)}$ and $\alpha_2^{(t)}$ be denoted as follows:

$$\alpha_1^{(t)}=\frac{1}{n^2}\sum_{i':y_{i'}=1}^{n}\sum_{q\in[P]}\sum_{i,j\in[n]}l_{1,(i,j)}\sum_{p\in[P]}\theta_{i,j}^{(p)}(\boldsymbol{\xi}_{i'}^{(q)})\cdot\theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_{i'}^{(q)}\rangle$$
$$\alpha_2^{(t)}=\frac{1}{n^2}\sum_{i':y_{i'}=2}^{n}\sum_{q\in[P]}\sum_{i,j\in[n]}l_{1,(i,j)}\sum_{p\in[P]}\theta_{i,j}^{(p)}(\boldsymbol{\xi}_{i'}^{(q)})\cdot\theta_{i,j}^{(p)}(\mathbf{v})\cdot\langle \mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_{i'}^{(q)}\rangle.$$

Then it is clear that $I_3=\alpha_1^{(t)}+\alpha_2^{(t)}$. Then by (C.9) and (C.12), we can get

$$\alpha_1^{(t+1)}=\bigg[1+\Theta\bigg(\frac{\eta d\sigma_p^2}{n}\bigg)\bigg]\cdot\alpha_1^{(t)}\pm O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg)\cdot\langle \mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle\pm O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg)\cdot\langle \mathbf{w}_{1,r}^{(t)},\mathbf{u}\rangle$$
$$\pm O\bigg(\frac{\eta\rho d\sigma_p^2}{Pn}\bigg)\cdot\langle \mathbf{w}_{1,r}^{(t)},\mathbf{v}'\rangle\pm O\bigg(\frac{\eta\rho d\sigma_p^2}{Pn}\bigg)\cdot\langle \mathbf{w}_{1,r}^{(t)},\mathbf{u}'\rangle$$
$$\pm O\bigg(\frac{\eta d\sigma_p^2}{n^2P}\bigg)\cdot\bigg|\sum_{p\in[P]}\sum_{s:y_s=1}\langle \mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle\bigg|\pm O\bigg(\frac{\eta\zeta d\sigma_p^2}{n^2P}\bigg)\cdot\sum_{p\in[P]}\sum_{s=1}^{n}|\langle \mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle|.$$

Similarly, we can also obtain

$$\alpha_2^{(t+1)}=\bigg[1-\Theta\bigg(\frac{\eta d\sigma_p^2}{n}\bigg)\bigg]\cdot\alpha_2^{(t)}\pm O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg)\cdot\langle \mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle\pm O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg)\cdot\langle \mathbf{w}_{1,r}^{(t)},\mathbf{u}\rangle$$
$$\pm O\bigg(\frac{\eta\rho d\sigma_p^2}{Pn}\bigg)\cdot\langle \mathbf{w}_{1,r}^{(t)},\mathbf{v}'\rangle\pm O\bigg(\frac{\eta\rho d\sigma_p^2}{Pn}\bigg)\cdot\langle \mathbf{w}_{1,r}^{(t)},\mathbf{u}'\rangle$$
$$\pm O\bigg(\frac{\eta d\sigma_p^2}{n^2P}\bigg)\cdot\bigg|\sum_{p\in[P]}\sum_{s:y_s=2}\langle \mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle\bigg|\pm O\bigg(\frac{\eta\zeta d\sigma_p^2}{n^2P}\bigg)\cdot\sum_{p\in[P]}\sum_{s=1}^{n}|\langle \mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle|.$$

Then using the previous results on $\big|\sum_{p\in[P]}\sum_{s:y_s=1}\langle \mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle\big|$ and $\big|\sum_{p\in[P]}\sum_{s:y_s=2}\langle \mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_s^{(p)}\rangle\big|$ and (C.10), we can get that for any $t=O\big(n\eta^{-1}/(d\sigma_p^2)\big)$,

$$|\alpha_1^{(t)}|\le O\bigg(\frac{d^{1/2}\sigma_0\sigma_p}{P^{1/2}n^{1/2}}\bigg)+O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg)\cdot\sum_{\tau=0}^{t-1}\big[|\langle \mathbf{w}_{1,r}^{(\tau)},\mathbf{v}\rangle|+|\langle \mathbf{w}_{1,r}^{(\tau)},\mathbf{u}\rangle|\big]$$
$$+O\bigg(\frac{\eta\rho d\sigma_p^2}{Pn}\bigg)\cdot\sum_{\tau=0}^{t-1}\big[|\langle \mathbf{w}_{1,r}^{(\tau)},\mathbf{v}'\rangle|+|\langle \mathbf{w}_{1,r}^{(\tau)},\mathbf{u}'\rangle|\big]+O\bigg(\frac{\eta\zeta d\sigma_p^2}{n^2P}\bigg)\cdot\sum_{\tau=0}^{t-1}\sum_{p\in[P]}\sum_{s=1}^{n}|\langle \mathbf{w}_{1,r}^{(\tau)},\boldsymbol{\xi}_s^{(p)}\rangle|,$$

where we use the upper bound of $|\alpha_1^{(0)}|$ provided in Similarly, we can obtain the same results for $\alpha_2^{(t)}$ as follows:

$$|\alpha_2^{(t)}| \leq O\bigg(\frac{d^{1/2}\sigma_0\sigma_p}{P^{1/2}n^{1/2}}\bigg) + O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg)\cdot\sum_{\tau=0}^{t-1}\big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}\rangle|\big]$$
$$+ O\bigg(\frac{\eta\rho d\sigma_p^2}{Pn}\bigg)\cdot\sum_{\tau=0}^{t-1}\big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}'\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}'\rangle|\big] + O\bigg(\frac{\eta\zeta d\sigma_p^2}{n^2P}\bigg)\cdot\sum_{\tau=0}^{t-1}\sum_{p\in[P]}\sum_{s=1}^{n}|\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_s^{(p)}\rangle|.$$

Combining the above results leads to the bound of I_3.

We will finally bound I_4 as follows: using the fact that $|\ell_{1,(i,j)} - l_{1,(i,j)}| = O(\zeta)$ and a similar characterization of (C.9), we can get

$$I_4 \leq O\bigg(\frac{\zeta}{Pn}\bigg)\cdot\sum_{s=1}^{n}\sum_{p\in[P]}|\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_s^{(p)}\rangle|.$$

Combining the above bounds on I_3 and I_4, we can finally get

$$|z_t| \leq |I_3| + |I_4|$$
$$\leq O\bigg(\frac{d^{1/2}\sigma_0\sigma_p}{P^{1/2}n^{1/2}}\bigg) + O\bigg(\frac{\zeta}{Pn}\bigg)\cdot\sum_{s=1}^{n}\sum_{p\in[P]}|\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_s^{(p)}\rangle| + O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg)\cdot\sum_{\tau=0}^{t-1}\big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}\rangle|\big]$$
$$+ O\bigg(\frac{\eta\rho d\sigma_p^2}{Pn}\bigg)\cdot\sum_{\tau=0}^{t-1}\big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}'\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}'\rangle|\big] + O\bigg(\frac{\eta\zeta d\sigma_p^2}{n^2P}\bigg)\cdot\sum_{\tau=0}^{t-1}\sum_{p\in[P]}\sum_{s=1}^{n}|\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_s^{(p)}\rangle|.$$

This completes the proof.

□

Then the following lemma characterizes the growth of common feature learning.

Lemma C.13. Assume $\max_{k\in[2],(i,j)\in\mathcal{S}}|F_k(\mathbf{W}^{(t)};\mathbf{x}_{i,j})| \leq \zeta$ for some $\zeta = o\big(d^{-1/2}\sigma_p^{-1}\big)$. Then for any $t = O\big(\mathrm{polylog}(n)/\eta\big)$ that satisfies this condition, we have with probability at least $1 - 1/\mathrm{poly}(n)$, there exists at least one $r \in [m]$ such that

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle = \big[1 + \Theta(\eta)\big]\cdot\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle.$$

Proof. First, note that $\langle \mathbf{w}_{1,r}^{(0)}, \mathbf{v}\rangle$ follows $N(0,\sigma_0^2)$, then it is easy to get that

$$\mathbb{P}\bigg[\max_{r\in[m]}|\langle \mathbf{w}_{1,r}^{(0)}, \mathbf{v}\rangle| \geq \sigma_0\bigg] = 1 - \big(\mathbb{P}_{\xi\sim N(0,\sigma_0^2)}[|\xi| \leq \sigma_0]\big)^m \geq 1 - 0.7^m \geq 1 - 1/\mathrm{poly}(n), \tag{C.14}$$

where the last inequality is by our assumption that $m = \mathrm{polylog}(n) > C\log(n)$ for some sufficiently large constant C.

Recall the update rule of $\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle$:

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle = \big[1 + \eta\gamma_1^{(t)}(\mathbf{v},\mathbf{v})\big]\cdot\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle + \eta\gamma_1^{(t)}(\mathbf{u},\mathbf{v})\cdot\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle + \eta\gamma_1^{(t)}(\mathbf{v}',\mathbf{v})\cdot\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle$$
$$+ \eta\gamma_1^{(t)}(\mathbf{u}',\mathbf{v})\cdot\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle + \sum_{i=1}^{n}\sum_{p\in[P]}\eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)},\mathbf{v})\cdot\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle.$$

Taking absolute value on both sides leads to

$$|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle| \geq \big[1 + \eta\gamma_1^{(t)}(\mathbf{v},\mathbf{v})\big]\cdot|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle| - \eta\big|\gamma_1^{(t)}(\mathbf{u},\mathbf{v})\cdot\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle\big| - \eta\big|\gamma_1^{(t)}(\mathbf{v}',\mathbf{v})\cdot\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle\big|$$

$$- \eta\big|\gamma_1^{(t)}(\mathbf{u}',\mathbf{v})\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{u}'\rangle\big| - \bigg|\sum_{i=1}^{n}\sum_{p\in[P]}\eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)},\mathbf{v})\cdot\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_i^{(p)}\rangle\bigg|.$$

Therefore, the next step is to show that these "negative" terms in the above inequality are dominated by $\eta\gamma_1^{(t)}(\mathbf{v},\mathbf{v})\cdot|\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle|$, i.e., showing that

$$\begin{aligned}&\big|\gamma_1^{(t)}(\mathbf{u},\mathbf{v})\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{u}\rangle\big|,\big|\gamma_1^{(t)}(\mathbf{v}',\mathbf{v})\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}'\rangle\big|,\big|\gamma_1^{(t)}(\mathbf{u}',\mathbf{v})\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{u}'\rangle\big|\ll|\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle|;\\&\bigg|\sum_{i=1}^{n}\sum_{p\in[P]}\eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)},\mathbf{v})\cdot\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_i^{(p)}\rangle\bigg|\ll|\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle|,\end{aligned}$$

where we use our result in Lemma C.3 that $\gamma_1^{(t)}(\mathbf{v},\mathbf{v})=\Theta(1)$. Then we are able to get that

$$|\langle\mathbf{w}_{1,r}^{(t+1)},\mathbf{v}\rangle| = \big[1+\eta\gamma_1^{(t)}(\mathbf{v},\mathbf{v})\pm o\big(1/\mathrm{polylog}(n)\big)\big]\cdot|\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle|\geq\big[1+\Theta(\eta)\big]\cdot|\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle|. \tag{C.15}$$

Regarding the first three terms, we will prove them by mathematical induction on a stronger argument (recall that $\gamma_1^{(t)}(\mathbf{v},\mathbf{v})=\Theta(1)$, $|\gamma_1^{(t)}(\mathbf{u},\mathbf{v})|,|\gamma_1^{(t)}(\mathbf{v}',\mathbf{v})|,|\gamma_1^{(t)}(\mathbf{u}',\mathbf{v})|=o\big(1/\mathrm{polylog}(n)\big)$, according to Lemma C.3): we aim to verify the hypothesis

$$|\langle\mathbf{w}_{1,r}^{(t)},\mathbf{u}\rangle|,\quad|\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}'\rangle|,\quad|\langle\mathbf{w}_{1,r}^{(t)},\mathbf{u}'\rangle|\leq c\cdot\log^2(n)|\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle|, \tag{C.16}$$

where c is some sufficiently small constant.

In particular, we can first consider the initialization where $t=0$, then by (C.14) and standard concentration bound of Gaussian random variable, we have with probability at least $1-1/\mathrm{poly}(n)$,

$$\begin{aligned}&|\langle\mathbf{w}_{1,r}^{(0)},\mathbf{v}\rangle|=\Omega(\sigma_0),\quad|\langle\mathbf{w}_{1,r}^{(0)},\mathbf{u}\rangle|=O\big(\log(n)\sigma_0\big),\\&|\langle\mathbf{w}_{1,r}^{(0)},\mathbf{v}'\rangle|=O\big(\log(n)\sigma_0\big),\quad|\langle\mathbf{w}_{1,r}^{(0)},\mathbf{v}'\rangle|=O\big(\log(n)\sigma_0\big).\end{aligned}$$

Therefore, using the fact that $\zeta=o\big(d^{-1/2}\sigma_p^{-1}\big)$, it is easy to verify the hypothesis. We will then assume the hypothesis holds for all $\tau\leq t$ and aim to verify it for $t+1$. Particularly, recall the update rules of $\langle\mathbf{w}_{1,r}^{(t)},\mathbf{u}\rangle$, we have

$$\begin{aligned}|\langle\mathbf{w}_{1,r}^{(t+1)},\mathbf{u}\rangle| &\leq\big[1-\eta\gamma_1^{(t)}(\mathbf{u},\mathbf{u})\big]\cdot|\langle\mathbf{w}_{1,r}^{(t)},\mathbf{u}\rangle|+\eta\big|\gamma_1^{(t)}(\mathbf{v},\mathbf{u})\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}\rangle\big|+\eta\big|\gamma_1^{(t)}(\mathbf{v}',\mathbf{v})\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{v}'\rangle\big|\\&\quad+\eta\big|\gamma_1^{(t)}(\mathbf{u}',\mathbf{u})\cdot\langle\mathbf{w}_{1,r}^{(t)},\mathbf{u}'\rangle\big|+\bigg|\sum_{i=1}^{n}\sum_{p\in[P]}\eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)},\mathbf{u})\cdot\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_i^{(p)}\rangle\bigg|\\&\leq|\langle\mathbf{w}_{1,r}^{(0)},\mathbf{u}\rangle|+\eta\sum_{\tau=0}^{t}\big|\gamma_1^{(\tau)}(\mathbf{v},\mathbf{u})\cdot\langle\mathbf{w}_{1,r}^{(\tau)},\mathbf{v}\rangle\big|+\eta\sum_{\tau=0}^{t}\big|\gamma_1^{(\tau)}(\mathbf{v}',\mathbf{v})\cdot\langle\mathbf{w}_{1,r}^{(\tau)},\mathbf{v}'\rangle\big|\\&\quad+\eta\sum_{\tau=0}^{t}\big|\gamma_1^{(\tau)}(\mathbf{u}',\mathbf{u})\cdot\langle\mathbf{w}_{1,r}^{(\tau)},\mathbf{u}'\rangle\big|+\sum_{\tau=0}^{t}\bigg|\sum_{i=1}^{n}\sum_{p\in[P]}\eta\gamma_1^{(\tau)}(\boldsymbol{\xi}_i^{(p)},\mathbf{u})\cdot\langle\mathbf{w}_{1,r}^{(\tau)},\boldsymbol{\xi}_i^{(p)}\rangle\bigg|\\&\leq O\big(\log(n)\sigma_0\big)+\widetilde{O}\big(\eta(\zeta+\alpha)\big)\cdot\sum_{\tau=0}^{t}|\langle\mathbf{w}_{1,r}^{(\tau)},\mathbf{v}\rangle|+O\bigg(\frac{\eta\rho}{P}+\frac{t\rho\eta^2d\sigma_p^2}{Pn}\bigg)\sum_{\tau=0}^{t}\big[|\langle\mathbf{w}_{1,r}^{(\tau)},\mathbf{v}'\rangle|+|\langle\mathbf{w}_{1,r}^{(\tau)},\mathbf{u}'\rangle|\big]\\&\quad+O\bigg(\frac{t\eta d^{1/2}\sigma_0\sigma_p}{P^{1/2}n^{1/2}}\bigg)+O\bigg(\frac{t\eta^2d\sigma_p^2}{Pn}\bigg)\cdot\sum_{\tau=0}^{t}\big[|\langle\mathbf{w}_{1,r}^{(\tau)},\mathbf{v}\rangle|+|\langle\mathbf{w}_{1,r}^{(\tau)},\mathbf{u}\rangle|\big]\\&\quad+O\bigg(\frac{\eta\zeta}{Pn}+\frac{t\eta^2\zeta d\sigma_p^2}{n^2P}\bigg)\cdot\sum_{\tau=0}^{t}\sum_{s=1}^{n}\sum_{p\in[P]}|\langle\mathbf{w}_{1,r}^{(\tau)},\boldsymbol{\xi}_s^{(p)}\rangle|.\end{aligned} \tag{C.17}$$

where the last inequality is by Lemma C.12. Then by (C.12), we have the following results regarding $|\langle\mathbf{w}_{1,r}^{(\tau)},\boldsymbol{\xi}_s^{(p)}\rangle|$

$$\max_{i\in[n],p\in[P]}|\langle\mathbf{w}_{1,r}^{(\tau+1)},\boldsymbol{\xi}_i^{(p)}\rangle|\leq\bigg[1+O\bigg(\frac{d\sigma_p^2}{n}\bigg)\bigg]\cdot\max_{i\in[n],p\in[P]}|\langle\mathbf{w}_{1,r}^{(\tau)},\boldsymbol{\xi}_i^{(p)}\rangle|+O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg)\cdot\big[|\langle\mathbf{w}_{1,r}^{(\tau)},\mathbf{v}\rangle|+|\langle\mathbf{w}_{1,r}^{(\tau)},\mathbf{u}\rangle|\big]$$

$$
\begin{aligned}
&+ O\bigg(\frac{\eta\rho d\sigma_p^2}{Pn}\bigg) \cdot \big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}'\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}'\rangle|\big] \\
&= O\big(d^{1/2}\sigma_0\sigma_p\big) + O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg) \cdot \sum_{s=0}^{\tau} \big[|\langle \mathbf{w}_{1,r}^{(s)}, \mathbf{v}\rangle| + |\langle \mathbf{w}_{1,r}^{(s)}, \mathbf{u}\rangle|\big] \\
&+ O\bigg(\frac{\eta\rho d\sigma_p^2}{Pn}\bigg) \cdot \sum_{s=0}^{\tau} \big[|\langle \mathbf{w}_{1,r}^{(s)}, \mathbf{v}'\rangle| + |\langle \mathbf{w}_{1,r}^{(s)}, \mathbf{u}'\rangle|\big].
\end{aligned}
$$

Therefore, we can accordingly get the following upper bound regarding the last term in the RHS of (C.17),

$$
\begin{aligned}
\frac{1}{nP}\sum_{\tau=0}^{t}\sum_{s=1}^{n}\sum_{p\in[P]} |\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_s^{(p)}\rangle| &\leq \sum_{\tau=0}^{t} \max_{i\in[n], p\in[P]} |\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_i^{(p)}\rangle| \\
&\leq O\big(td^{1/2}\sigma_0\sigma_p\big) + O\bigg(\frac{t\eta d\sigma_p^2}{Pn}\bigg) \cdot \sum_{\tau=0}^{t-1} \big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}\rangle|\big] \\
&\quad + O\bigg(\frac{t\eta\rho d\sigma_p^2}{Pn}\bigg) \cdot \sum_{\tau=0}^{t-1} \big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}'\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}'\rangle|\big]. \qquad \text{(C.18)}
\end{aligned}
$$

Then using the fact that $t\eta = O\big(\mathrm{polylog}(n)\big)$ and $n = \omega(d\sigma_p^2)$, we can further get the following on (C.17)

$$
\begin{aligned}
|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle| &\leq O\big(\log(n)\sigma_0\big) + \widetilde{O}\big(\eta(\zeta+\alpha)\big) \cdot \sum_{\tau=0}^{t} |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}\rangle| + O\bigg(\frac{\eta\rho}{P}\bigg) \sum_{\tau=0}^{t} \big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}'\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}'\rangle|\big] \\
&\quad + O\bigg(\frac{t\eta d^{1/2}\sigma_0\sigma_p}{P^{1/2}n^{1/2}} + t\eta\zeta d^{1/2}\sigma_0\sigma_p\bigg) + O\bigg(\frac{\eta}{P}\bigg) \cdot \sum_{\tau=0}^{t} \big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}\rangle|\big].
\end{aligned}
$$

Then according to the Hypothesis C.16 for any $\tau \leq t$, it is easy to get that

$$
|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle| \geq |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle| \geq \ldots \geq |\langle \mathbf{w}_{1,r}^{(0)}, \mathbf{u}\rangle|.
$$

Then we can get $|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle| = \Omega(\sigma_0)$, applying the fact that $t\eta = O(\mathrm{polylog}(n))$ further gives

$$
O\big(\log(n)\sigma_0\big) + O\bigg(\frac{t\eta d^{1/2}\sigma_0\sigma_p}{P^{1/2}n^{1/2}} + t\eta\zeta d^{1/2}\sigma_0\sigma_p\bigg)\bigg)\bigg] = o(\log^2(n)\sigma_0) = o\big(\log^2(n)|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle|\big). \qquad \text{(C.19)}
$$

Besides, note that the Hypothesis C.16 holds for all $\tau \leq t$, we have

$$
|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}\rangle|, |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}'\rangle|, |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}'\rangle| \leq c \cdot \log^2(n) \cdot |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}\rangle| \leq \log^2(n) \cdot |\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle|, \qquad \text{(C.20)}
$$

we can immediately get that

$$
\begin{aligned}
&\widetilde{O}\big(\eta(\zeta+\alpha)\big) \cdot \sum_{\tau=0}^{t} |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}\rangle| \leq \widetilde{O}\big(t\eta(\zeta+\alpha)\big) \cdot |\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle| = o\big(\log^2(n) \cdot |\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle|\big) \\
&O\bigg(\frac{\eta\rho}{P}\bigg) \cdot \sum_{\tau=0}^{t} \big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}'\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}'\rangle|\big] \leq \widetilde{O}\bigg(\frac{t\eta\rho}{P}\bigg) \cdot |\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle| = o\big(\log^2(n) \cdot |\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle|\big) \\
&O\bigg(\frac{\eta}{P}\bigg) \cdot \sum_{\tau=0}^{t} \big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}\rangle|\big] \leq \widetilde{O}\bigg(\frac{t\eta}{P}\bigg) \cdot |\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle| = o\big(\log^2(n) \cdot |\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle|\big). \qquad \text{(C.21)}
\end{aligned}
$$

Putting the above results together, we can verify that

$$
|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle| = o\big(\log^2(n) \cdot |\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle|\big).
$$

We will then verify the Hypothesis for $\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}' \rangle$. By its update rule, Lemma C.7, and Lemma C.12, we have

$$
\begin{aligned}
|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}' \rangle| &\leq \big[1 + \eta \gamma_1^{(t)}(\mathbf{v}', \mathbf{v}')\big] \cdot |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}' \rangle| + \eta \big|\gamma_1^{(t)}(\mathbf{v}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle\big| + \eta \big|\gamma_1^{(t)}(\mathbf{u}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u} \rangle\big| \\
&\quad + \eta \big|\gamma_1^{(t)}(\mathbf{u}', \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}' \rangle\big| + \bigg| \sum_{i=1}^{n} \sum_{p \in [P]} \eta \gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)} \rangle \bigg| \\
&\leq 2|\langle \mathbf{w}_{1,r}^{(0)}, \mathbf{v}' \rangle| + \eta \sum_{\tau=0}^{t} \big|\gamma_1^{(\tau)}(\mathbf{v}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v} \rangle\big| + \eta \sum_{\tau=0}^{t} \big|\gamma_1^{(\tau)}(\mathbf{u}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u} \rangle\big| \\
&\quad + \eta \sum_{\tau=0}^{t} \big|\gamma_1^{(\tau)}(\mathbf{u}', \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}' \rangle\big| + \sum_{\tau=0}^{t} \bigg| \sum_{i=1}^{n} \sum_{p \in [P]} \eta \gamma_1^{(\tau)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_i^{(p)} \rangle \bigg| \\
&\leq O\big(\log(n)\sigma_0\big) + O\bigg(\frac{\eta\rho}{P} + \frac{t\rho\eta^2 d\sigma_p^2}{Pn}\bigg) \cdot \sum_{\tau=0}^{t} |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v} \rangle| + O\bigg(\frac{\eta\zeta\rho}{P} + \frac{t\eta^2\rho d\sigma_p^2}{Pn}\bigg) \cdot \sum_{\tau=0}^{t} |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u} \rangle| \\
&\quad + O\bigg(\frac{\eta\zeta\rho^2}{P} + \frac{t\rho^2\eta^2 d\sigma_p^2}{Pn}\bigg) \cdot \sum_{\tau=0}^{t} |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}' \rangle| + O\bigg(\frac{t\eta\rho d^{1/2}\sigma_0\sigma_p}{P^{1/2}n^{1/2}}\bigg) \\
&\quad + O\bigg(\frac{\eta\rho\zeta}{Pn} + \frac{t\eta^2\zeta\rho d\sigma_p^2}{n^2 P}\bigg) \cdot \sum_{\tau=0}^{t} \sum_{s=1}^{n} \sum_{p \in [P]} |\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_s^{(p)} \rangle|.
\end{aligned}
$$

Then by (C.18) and using the fact that $t\eta = O(\mathrm{polylog}(n))$ and $n = \omega(d\sigma_p^2)$, we can finally get

$$
\begin{aligned}
|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}' \rangle| &\leq O\big(\log(n)\sigma_0\big) + O\bigg(\frac{\eta\rho}{P}\bigg) \cdot \sum_{\tau=0}^{t} |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v} \rangle| + O\bigg(\frac{\eta\rho}{P}\bigg) \cdot \sum_{\tau=0}^{t} |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u} \rangle| \\
&\quad + O\bigg(\frac{\eta\rho^2}{P}\bigg) \cdot \sum_{\tau=0}^{t} |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}' \rangle| + O\bigg(\frac{t\eta\rho d^{1/2}\sigma_0\sigma_p}{P^{1/2}n^{1/2}} + t\eta\rho\zeta d^{1/2}\sigma_0\sigma_p\bigg).
\end{aligned}
$$

Then applying (C.19), (C.20), and (C.21), we can also verify that

$$
|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}' \rangle| = o\big(\log^2(n) \cdot |\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v} \rangle|\big).
$$

The using exactly the same proof, we are also able to verify that

$$
|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}' \rangle| = o\big(\log^2(n) \cdot |\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v} \rangle|\big).
$$

Lastly, we will prove that

$$
\bigg| \sum_{i=1}^{n} \sum_{p \in [P]} \gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)} \rangle \bigg| \leq c \cdot |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle| \tag{C.22}
$$

for some sufficiently small constant c and all $t = O\big(\mathrm{polylog}(n)/\eta\big)$. This can be proved by the combination of Lemma C.12, (C.18), and our previous characterizations (C.19), (C.20), (C.21). In particular, using the fact that $|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle| = \Omega(\sigma_0)$, we have

$$
\begin{aligned}
&\bigg| \sum_{i=1}^{n} \sum_{p \in [P]} \gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{v}) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)} \rangle \bigg| \\
&\leq O\bigg(\frac{d^{1/2}\sigma_0\sigma_p}{P^{1/2}n^{1/2}}\bigg) + O\bigg(\frac{\zeta}{Pn}\bigg) \cdot \sum_{s=1}^{n} \sum_{p \in [P]} |\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_s^{(p)} \rangle| + O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg) \cdot \sum_{\tau=0}^{t-1} \big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v} \rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u} \rangle|\big] \\
&\quad + O\bigg(\frac{\eta\rho d\sigma_p^2}{Pn}\bigg) \cdot \sum_{\tau=0}^{t-1} \big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}' \rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}' \rangle|\big] + O\bigg(\frac{\eta\zeta d\sigma_p^2}{n^2 P}\bigg) \cdot \sum_{\tau=0}^{t-1} \sum_{p \in [P]} \sum_{s=1}^{n} |\langle \mathbf{w}_{1,r}^{(\tau)}, \boldsymbol{\xi}_s^{(p)} \rangle|
\end{aligned}
$$

$$
\begin{aligned}
&\leq O\bigg(\frac{d^{1/2}\sigma_0\sigma_p}{P^{1/2}n^{1/2}} + \zeta d^{1/2}\sigma_0\sigma_p\bigg) + O\bigg(\frac{\eta d\sigma_p^2}{Pn}\bigg)\cdot\sum_{\tau=0}^{t-1}\big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}\rangle|\big] \\
&\quad + O\bigg(\frac{\eta\rho d\sigma_p^2}{Pn}\bigg)\cdot\sum_{\tau=0}^{t-1}\big[|\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{v}'\rangle| + |\langle \mathbf{w}_{1,r}^{(\tau)}, \mathbf{u}'\rangle|\big] \\
&\leq O\bigg(\frac{d^{1/2}\sigma_0\sigma_p}{P^{1/2}n^{1/2}} + \zeta d^{1/2}\sigma_0\sigma_p\bigg) + O\bigg(\frac{d\sigma_p^2}{Pn}\bigg)\cdot|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle| \\
&\leq O\bigg(\bigg(\frac{d^{1/2}\sigma_p}{P^{1/2}n^{1/2}} + \zeta d^{1/2}\sigma_p + \frac{d\sigma_p^2}{Pn}\bigg)\cdot|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle|\bigg). \qquad \text{(C.23)}
\end{aligned}
$$

Then using the facts that $\zeta = o(d^{-1/2}\sigma_p^{-1})$ and $d\sigma_p^2 = o(n)$, we are able to complete the proof of (C.22).

□

Lemma C.14. Assume $\max_{k\in[2],(i,j)\in\mathcal{S}}|F_k(\mathbf{W}^{(t)};\mathbf{x}_{i,j})| \leq \zeta$ for some $\zeta = o\big(d^{-1/2}\sigma_p^{-1}\big)$. Then for any $t = O\big(\mathrm{polylog}(n)/\eta\big)$ that satisfies this condition, we have with probability at least $1 - 1/\mathrm{poly}(n)$,

$$
|\gamma_1^{(t)}(\mathbf{v},\mathbf{v}) - \gamma_2^{(t)}(\mathbf{u},\mathbf{u})| \leq o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg).
$$

Proof. Recall $\gamma_1^{(t)}(\mathbf{v},\mathbf{v})$, we have

$$
\begin{aligned}
\gamma_1^{(t)}(\mathbf{v},\mathbf{v}) &= \frac{1}{n^2}\sum_{i,j\in[n]}\ell_{1,(i,j)}^{(t)}\sum_{p\in[P]}[\theta_{i,j}^{(p)}(\mathbf{v})]^2 \\
&= \underbrace{\frac{1}{n^2}\sum_{i\in\mathcal{S}_0^+ \text{ or } j\in\mathcal{S}_0^+}\ell_{1,(i,j)}^{(t)}\sum_{p\in[P]}[\theta_{i,j}^{(p)}(\mathbf{v})]^2}_{I_1} + \underbrace{\frac{1}{n^2}\sum_{i\notin\mathcal{S}_0^+ \text{ and } j\notin\mathcal{S}_0^+}\ell_{1,(i,j)}^{(t)}\sum_{p\in[P]}[\theta_{i,j}^{(p)}(\mathbf{v})]^2}_{I_2}.
\end{aligned}
$$

Regarding I_2, using the similar proof in Lemma C.3, we can obtain that $I_2 = o\big(1/\mathrm{polylog}(n)\big)$. For I_1, using the condition that $\max_{k\in[2],(i,j)\in\mathcal{S}}|F_k(\mathbf{W}^{(t)};\mathbf{x}_{i,j})| \leq \zeta$, we have

$$
I_1 = \frac{1}{n^2}\sum_{i\in\mathcal{S}_0^+ \text{ or } j\in\mathcal{S}_0^+} l_{1,(i,j)}\sum_{p\in[P]}[\theta_{i,j}^{(p)}(\mathbf{v})]^2 \pm O(\zeta),
$$

where $l_{1,(i,j)} \in \{0.5, -0.5, 0.5-\lambda, \lambda-0.5\}$ denotes the loss derivative of data $(\mathbf{x}_{i,j}, y_{i,j})$ when its neural network output is forced to be zero. To this end, using the similar decomposition for $\gamma_2^{(t)}(\mathbf{u},\mathbf{u})$ and noting $\zeta = o(1/\mathrm{polylog}(n))$, we can obtain

$$
\begin{aligned}
|\gamma_1^{(t)}(\mathbf{v},\mathbf{v}) - \gamma_2^{(t)}(\mathbf{u},\mathbf{u})| \leq &\bigg|\frac{1}{n^2}\sum_{i\in\mathcal{S}_0^+ \text{ or } j\in\mathcal{S}_0^+} l_{1,(i,j)}\sum_{p\in[P]}[\theta_{i,j}^{(p)}(\mathbf{v})]^2 - \frac{1}{n^2}\sum_{i\in\mathcal{S}_0^- \text{ or } j\in\mathcal{S}_0^-} l_{2,(i,j)}\sum_{p\in[P]}[\theta_{i,j}^{(p)}(\mathbf{u})]^2\bigg| \\
&+ o\big(1/\mathrm{polylog}(n)\big). \qquad \text{(C.24)}
\end{aligned}
$$

Moreover, for any $i \in \mathcal{S}_0^+$, note that

$$
\begin{aligned}
\sum_{j\in[n]} l_{1,(i,j)}\sum_{p\in[P]}[\Theta_{i,j}^{(p)}(\mathbf{v})]^2 &= \sum_{j\in[n]} l_{1,(i,j)}\sum_{p\in\mathcal{P}_{i,j}^*(\mathbf{v})}[\Theta_{i,j}^{(p)}(\mathbf{v})]^2 + \sum_{j\in[n]} l_{1,(i,j)}\sum_{p\notin\mathcal{P}_{i,j}^*(\mathbf{v})}[\Theta_{i,j}^{(p)}(\mathbf{v})]^2 \\
&= \ell_{1,(i,i)} + \sum_{j\neq i} l_{1,(i,j)}\cdot z_{i,j}^2 \pm o\big(n/\mathrm{polylog}(n)\big),
\end{aligned}
$$

where $z_{i,j} = (1-\lambda)^2$ if $j \in \mathcal{S}_1^+ \cup \mathcal{S}_0^- \cup \mathcal{S}_1^-$ and

$$
z_{i,j} = \begin{cases} 1 & \text{with probability } 1/P; \\ (1-\lambda)^2 + \lambda^2 & \text{with probability } (P-1)/P, \end{cases}
$$

if $j \in \mathcal{S}_0^+$. Consequently, applying Hoeffeding's inequality regarding the random variable $z_{i,j}$ (when $j \in \mathcal{S}_0^+$), we have with probability at least $1 - 1/\mathrm{poly}(n)$,

$$\frac{1}{n^2} \sum_{i \in \mathcal{S}_0^+, j \notin \mathcal{S}_0^+} l_{1,(i,j)} \sum_{p \in [P]} [\theta_{i,j}^{(p)}(\mathbf{v})]^2 = \frac{(1-\lambda)^2}{n^2} \sum_{i \in \mathcal{S}_0^+, j \in \mathcal{S}_1^+ \cup \mathcal{S}_0^- \cup \mathcal{S}_1^-} l_{1,(i,j)} \pm o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg)$$
$$\frac{1}{n^2} \sum_{i \in \mathcal{S}_0^+, j \in \mathcal{S}_0^+} l_{1,(i,j)} \sum_{p \in [P]} [\theta_{i,j}^{(p)}(\mathbf{v})]^2 = \frac{\ell_{1,(i,i)}}{n^2} + \frac{1 + [(1-\lambda)^2 + \lambda^2](P-1)}{Pn^2} \cdot \sum_{i \in \mathcal{S}_0^+, j \in \mathcal{S}_0^+, j \neq i} l_{1,(i,j)} \pm o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg).$$

Similarly, we can also obtain

$$\frac{1}{n^2} \sum_{i \notin \mathcal{S}_0^+, j \in \mathcal{S}_0^+} l_{1,(i,j)} \sum_{p \in [P]} [\theta_{i,j}^{(p)} = \frac{(1-\lambda)^2}{n^2} \sum_{i \in \mathcal{S}_1^+ \cup \mathcal{S}_0^- \cup \mathcal{S}_1^-, j \in \mathcal{S}_0^+} l_{1,(i,j)} \pm o\big(1/\mathrm{polylog}(n)\big)$$

Therefore, combining the above results, we can get

$$\begin{aligned}
&\frac{1}{n^2} \sum_{i \in \mathcal{S}_0^+ \text{ or } j \in \mathcal{S}_0^+} l_{1,(i,j)} \sum_{p \in [P]} [\theta_{i,j}^{(p)}(\mathbf{v})]^2 \\
&= \frac{1}{n^2} \sum_{i \in \mathcal{S}_0^+, j \notin \mathcal{S}_0^+} l_{1,(i,j)} \sum_{p \in [P]} [\theta_{i,j}^{(p)}(\mathbf{v})]^2 + \frac{1}{n^2} \sum_{i \in \mathcal{S}_0^+, j \in \mathcal{S}_0^+} l_{1,(i,j)} \sum_{p \in [P]} [\theta_{i,j}^{(p)}(\mathbf{v})]^2 + \frac{1}{n^2} \sum_{i \notin \mathcal{S}_0^+, j \in \mathcal{S}_0^+} l_{1,(i,j)} \sum_{p \in [P]} [\theta_{i,j}^{(p)}(\mathbf{v})]^2 \\
&= \frac{(1-\lambda)^2}{n^2} \sum_{i \in \mathcal{S}_0^+, j \notin \mathcal{S}_0^+ \text{ or } i \notin \mathcal{S}_0^+, j \in \mathcal{S}_0^+} l_{1,(i,j)} + \frac{\ell_{1,(i,i)}}{n^2} + \frac{1 + [(1-\lambda)^2 + \lambda^2](P-1)}{Pn^2} \cdot \sum_{i \in \mathcal{S}_0^+, j \in \mathcal{S}_0^+, j \neq i} l_{1,(i,j)} \\
&\quad \pm o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg).
\end{aligned}$$

Similarly, we can get

$$\begin{aligned}
&\frac{1}{n^2} \sum_{i \in \mathcal{S}_0^- \text{ or } j \in \mathcal{S}_0^-} l_{1,(i,j)} \sum_{p \in [P]} [\theta_{i,j}^{(p)}(\mathbf{u})]^2 \\
&= \frac{1}{n^2} \sum_{i \in \mathcal{S}_0^-, j \notin \mathcal{S}_0^-} l_{1,(i,j)} \sum_{p \in [P]} [\theta_{i,j}^{(p)}(\mathbf{u})]^2 + \frac{1}{n^2} \sum_{i \in \mathcal{S}_0^-, j \in \mathcal{S}_0^-} l_{1,(i,j)} \sum_{p \in [P]} [\theta_{i,j}^{(p)}(\mathbf{u})]^2 + \frac{1}{n^2} \sum_{i \notin \mathcal{S}_0^-, j \in \mathcal{S}_0^-} l_{1,(i,j)} \sum_{p \in [P]} [\theta_{i,j}^{(p)}(\mathbf{u})]^2 \\
&= \frac{(1-\lambda)^2}{n^2} \sum_{i \in \mathcal{S}_0^-, j \notin \mathcal{S}_0- \text{ or } i \notin \mathcal{S}_0^-, j \in \mathcal{S}_0^-} l_{1,(i,j)} + \frac{\ell_{1,(i,i)}}{n^2} + \frac{1 + [(1-\lambda)^2 + \lambda^2](P-1)}{Pn^2} \cdot \sum_{i \in \mathcal{S}_0^-, j \in \mathcal{S}_0^-, j \neq i} l_{1,(i,j)} \\
&\quad \pm o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg).
\end{aligned}$$

Then note that the positive and negative data are generated with equal probability, we have $|\mathcal{S}_0^-|$ and $|\mathcal{S}_0^+|$ are different by at most $o\big(1/\mathrm{polylog}(n)\big)$, therefore, it is easy to get that

$$\bigg| \frac{1}{n^2} \sum_{i \in \mathcal{S}_0^+ \text{ or } j \in \mathcal{S}_0^+} l_{1,(i,j)} \sum_{p \in [P]} [\theta_{i,j}^{(p)}(\mathbf{v})]^2 - \frac{1}{n^2} \sum_{i \in \mathcal{S}_0^- \text{ or } j \in \mathcal{S}_0^-} l_{2,(i,j)} \sum_{p \in [P]} [\theta_{i,j}^{(p)}(\mathbf{u})]^2 \bigg| \leq o\big(1/\mathrm{polylog}(n)\big).$$

Plugging the above inequality into (C.24) we can conclude that

$$|\gamma_1^{(t)}(\mathbf{v}, \mathbf{v}) - \gamma_2^{(t)}(\mathbf{u}, \mathbf{u})| \leq o\bigg(\frac{1}{\mathrm{polylog}(n)}\bigg).$$

This completes the proof. □

Finally, we state the outcome of noise learning, common feature learning, and rare feature learning in the following Lemma.

Lemma C.15. Let ζ be a preset quantity satisfying $\zeta = [\omega(d\sigma_p^2/(Pn)), o(d^{-1/2}\sigma_p^{-1})]$ and T be the smallest iteration number such that $\max_{k\in[2],(i,j)\in\mathcal{S}} |F_k(\mathbf{W}^{(T)}; \mathbf{x}_{i,j})| \geq \zeta/2$, then with probability at least $1 - 1/\mathrm{poly}(n)$, it holds that

$$\max_r |\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{v}\rangle|, \max_r |\langle \mathbf{w}_{2,r}^{(T)}, \mathbf{u}\rangle| = \Omega\bigg(\frac{\zeta^{1/2}}{m^{1/2}}\bigg), \quad \max_r |\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{v}'\rangle|, \max_r |\langle \mathbf{w}_{2,r}^{(T)}, \mathbf{u}'\rangle| = \Omega\bigg(\frac{\rho\zeta^{1/2}}{Pm^{1/2}}\bigg)$$
$$\max_r |\langle \mathbf{w}_{2,r}^{(T)}, \mathbf{v}\rangle|, \max_r |\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{u}\rangle| = \widetilde{O}(\zeta^{3/2}), \quad \max_r |\langle \mathbf{w}_{2,r}^{(T)}, \mathbf{v}'\rangle|, \max_r |\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{u}'\rangle| = \widetilde{O}(\zeta^{3/2})$$

Proof. We will only prove the results for the inner products $\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle$, $\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{u}\rangle$, $\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle$, $\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle$, and $\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle$, as the proof for the remaining inner products will be exactly the same.

We first recall the update of $\langle \mathbf{w}_{1,r}, \mathbf{v}'\rangle$:

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}'\rangle = \big[1 + \eta\gamma_1^{(t)}(\mathbf{v}', \mathbf{v}')\big] \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle + \eta\gamma_1^{(t)}(\mathbf{v}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle + \eta\gamma_1^{(t)}(\mathbf{u}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle$$
$$+ \eta\gamma_1^{(t)}(\mathbf{u}', \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle + \sum_{i=1}^n \sum_{p\in[P]} \eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle.$$

The using Lemma C.7 and the similar proof of Lemma C.13, we can get

$$\big|\gamma_1^{(t)}(\mathbf{u}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle\big| = O\bigg(\frac{\zeta\rho}{P}\bigg) \cdot |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle| = O\bigg(\frac{\zeta\rho\log^2(n)}{P}\bigg) \cdot |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle|$$
$$\big|\gamma_1^{(t)}(\mathbf{u}', \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle\big| = O\bigg(\frac{\zeta\rho^2}{P}\bigg) \cdot |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle| = O\bigg(\frac{\zeta\rho^2\log^2(n)}{P}\bigg) \cdot |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle|$$
$$\bigg|\sum_{i=1}^n \sum_{p\in[P]} \gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle\bigg| \leq O\bigg(\bigg(\frac{\rho d^{1/2}\sigma_p}{P^{1/2}n^{1/2}} + \rho\zeta d^{1/2}\sigma_p\bigg) \cdot |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle|\bigg).$$

Therefore, noting that we have assumed $d\sigma_p = o(n/P)$ and $\zeta = o\big(\frac{1}{Pd^{1/2}\sigma_p}\big)$,

$$\big|\gamma_1^{(t)}(\mathbf{u}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle\big|, \big|\gamma_1^{(t)}(\mathbf{u}', \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle\big|, \bigg|\sum_{i=1}^n \sum_{p\in[P]} \gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle\bigg| \leq c \cdot |\gamma_1^{(t)}(\mathbf{v}, \mathbf{v}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle|$$

for some sufficiently small constant $c < 0.5$. Therefore, further applying Lemma C.7, we can get that

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}'\rangle = \big[1 + \Theta(\eta\rho)\big] \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle + \Theta(\eta\rho/P) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle \tag{C.25}$$

Given the above equation, we are able to complete the proof by combining it with Lemma C.13:

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{v}\rangle = \big[1 + \Theta(\eta)\big] \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle. \tag{C.26}$$

In particular, given the fact that $|\langle \mathbf{w}_{1,r}^{(0)}, \mathbf{v}\rangle| = \Omega(\sigma_0)$, we can get the following

$$|\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{v}\rangle| = \Omega\bigg(\frac{\zeta^{1/2}}{m^{1/2}}\bigg) \tag{C.27}$$

for some $T = O\big(\frac{\log(\zeta/(m\sigma_0))}{\eta}\big)$. Besides, by Lemma C.14 and (C.15), we have for any $r' \in [m]$,

$$\frac{|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle|}{|\langle \mathbf{w}_{2,r'}^{(t+1)}, \mathbf{u}\rangle|} = \frac{|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle|}{|\langle \mathbf{w}_{2,r'}^{(t)}, \mathbf{u}\rangle|} \cdot \bigg(\frac{1 + \eta\gamma_1^{(t)}(\mathbf{v}, \mathbf{v}) \pm o\big(\eta/\mathrm{polylog}(n)\big)}{1 + \eta\gamma_2^{(t)}(\mathbf{u}, \mathbf{u}) \pm o\big(\eta/\mathrm{polylog}(n)\big)}\bigg)$$
$$= \frac{|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle|}{|\langle \mathbf{w}_{2,r'}^{(t)}, \mathbf{u}\rangle|} \cdot \big[1 + \eta \cdot \big(\gamma_1^{(t)}(\mathbf{v}, \mathbf{v}) - \gamma_2^{(t)}(\mathbf{u}, \mathbf{u})\big) \pm o\big(\eta/\mathrm{polylog}(n)\big)\big]$$

$$= \frac{|\langle \mathbf{w}_{1,r}^{(0)}, \mathbf{v}\rangle|}{|\langle \mathbf{w}_{2,r'}^{(0)}, \mathbf{u}\rangle|} \cdot \big[1 \pm o\big(\eta/\mathrm{polylog}(n)\big)\big]^t.$$

Note that $t \leq T = \widetilde{O}(1/\eta)$, we can further get $\frac{|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle|}{|\langle \mathbf{w}_{2,r'}^{(t+1)}, \mathbf{u}\rangle|} = \Theta(1) \cdot \frac{|\langle \mathbf{w}_{1,r}^{(0)}, \mathbf{v}\rangle|}{|\langle \mathbf{w}_{2,r'}^{(0)}, \mathbf{u}\rangle|}$. This immediately implies that $\max_r |\langle \mathbf{w}_{2,r}^{(T)}, \mathbf{u}\rangle| = \Theta(\max_r |\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{v}\rangle|) = \Omega\big(\zeta^{1/2}/m^{1/2}\big)$.

Moreover, (C.25) implies that

$$\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{v}'\rangle = \big[1 + \Theta(\eta\rho)\big]^T \cdot \langle \mathbf{w}_{1,r}^{(0)}, \mathbf{v}'\rangle + \Theta(\eta\rho/P) \cdot \sum_{t=0}^{T-1} \big[1 + \Theta(\eta\rho)\big]^t \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle.$$

Further note that $\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle$ has the same sign for all $t \leq T$ and $[1 + \Theta(\eta\rho)]^t = \Theta(1)$ for all $t \leq T$, then define $T' = T - \Theta(1/\eta)$, we have

$$\begin{aligned}
|\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{v}'\rangle| &= \bigg|\Theta(1) \cdot \langle \mathbf{w}_{1,r}^{(0)}, \mathbf{v}'\rangle + \Theta(\eta\rho/P) \cdot \sum_{t=0}^{T-1} \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle\bigg| \\
&\geq \Theta(\eta\rho/P) \cdot \bigg|\sum_{t=0}^{T-1} \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle\bigg| - \Theta(1) \cdot \big|\langle \mathbf{w}_{1,r}^{(0)}, \mathbf{v}'\rangle\big| \\
&\geq \Theta(\eta\rho/P) \cdot \sum_{t=T'}^{T-1} \big|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle\big| - \widetilde{O}(\sigma_0).
\end{aligned}$$

Then by (C.26) and (C.27), we have for all $t \in [T', T-1]$, it holds that

$$|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle| = \Theta\big(|\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{v}\rangle|\big) = \Omega\bigg(\frac{\zeta^{1/2}}{m^{1/2}}\bigg).$$

Therefore, we can finally get

$$\begin{aligned}
|\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{v}'\rangle| &\geq \Theta\bigg(\frac{(T-T')\eta\rho}{P} \cdot\bigg) \cdot \Omega\bigg(\frac{\zeta^{1/2}}{m^{1/2}}\bigg) - \widetilde{O}(\sigma_0) \\
&= \Omega\bigg(\frac{\rho\zeta^{1/2}}{Pm^{1/2}}\bigg).
\end{aligned}$$

The remaining part is to establish the upper bounds in terms of incorrect feature learning, i.e., $\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{u}\rangle$ and $\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{u}'\rangle$. Particularly, recall their update forms as follows:

$$\begin{aligned}
\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle &= \big[1 - \eta\gamma_1^{(t)}(\mathbf{u}, \mathbf{u})\big] \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle + \eta\gamma_1^{(t)}(\mathbf{v}, \mathbf{u}) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle + \eta\gamma_1^{(t)}(\mathbf{v}', \mathbf{u}) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle \\
&\quad + \eta\gamma_1^{(t)}(\mathbf{u}', \mathbf{u}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle + \sum_{i=1}^{n} \sum_{p\in[P]} \eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{u}) \cdot \langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle, \\
\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}'\rangle &= \big[1 - \eta\gamma_1^{(t)}(\mathbf{u}', \mathbf{u}')\big] \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle + \eta\gamma_1^{(t)}(\mathbf{v}, \mathbf{u}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle + \eta\gamma_1^{(t)}(\mathbf{u}, \mathbf{u}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle \\
&\quad + \eta\gamma_1^{(t)}(\mathbf{v}', \mathbf{u}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle + \sum_{i=1}^{n} \sum_{p\in[P]} \eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{u}') \cdot \langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle.
\end{aligned}$$

Then by Lemmas C.6 and C.10, we have

$$\max\big\{|\gamma_1^{(t)}(\mathbf{u}', \mathbf{u})|, \gamma_1^{(t)}(\mathbf{u}, \mathbf{u}')|\big\} \leq \min\big\{\gamma_1^{(t)}(\mathbf{u}, \mathbf{u}), \gamma_1^{(t)}(\mathbf{u}', \mathbf{u}')\big\},$$

the above equations further yield

$$|\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle| + |\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}'\rangle|$$

$$\leq |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle| + |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle| + O(\eta\zeta)\cdot|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle| + O(\eta\zeta\rho/P)\cdot|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle|$$
$$+ \bigg|\sum_{i=1}^{n}\sum_{p\in[P]}\eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{u})\cdot\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle\bigg| + \bigg|\sum_{i=1}^{n}\sum_{p\in[P]}\eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{u}')\cdot\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle\bigg|.$$

Then using the fact that $T\eta = O(\mathrm{polylog}(n))$, we can further obtain

$$|\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{u}\rangle| + |\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{u}'\rangle| \leq \widetilde{O}(\sigma_0) + \widetilde{O}(\zeta)\cdot\max_{t\in[T]}|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle| + \widetilde{O}(\zeta\rho/P)\cdot\max_{t\in[T]}|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle|$$
$$+ \sum_{t=0}^{T-1}\bigg[\bigg|\sum_{i=1}^{n}\sum_{p\in[P]}\eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{u})\cdot\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle\bigg| + \bigg|\sum_{i=1}^{n}\sum_{p\in[P]}\eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{u}')\cdot\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle\bigg|\bigg].$$

Moreover, following the same procedure of (C.23), we can get

$$\sum_{t=0}^{T-1}\bigg|\sum_{i=1}^{n}\sum_{p\in[P]}\eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{u})\cdot\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle\bigg|, \sum_{t=0}^{T-1}\bigg|\sum_{i=1}^{n}\sum_{p\in[P]}\eta\gamma_1^{(t)}(\boldsymbol{\xi}_i^{(p)}, \mathbf{u}')\cdot\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i^{(p)}\rangle\bigg|$$
$$\leq \widetilde{O}(\sigma_0) + \widetilde{O}\bigg(\frac{d\sigma_p^2}{Pn}\bigg)\cdot\max_{t\in[T]}\big[|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle| + |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle| + |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle| + |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle|\big]$$

Finally, using the assumption that $\zeta = \omega(d\sigma_p^2/(Pn))$, we can get that

$$|\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{u}\rangle| + |\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{u}'\rangle|$$
$$\leq \widetilde{O}(\sigma_0) + \widetilde{O}(\zeta)\cdot\bigg[\max_{t\in[T]}|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle| + \max_{t\in[T]}|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle| + \max_{t\in[T]}\big[|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle| + |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle|\big]\bigg].$$

Besides, note that the above inequality actually holds for any $T' \leq T$, thus

$$|\langle \mathbf{w}_{1,r}^{(T')}, \mathbf{u}\rangle| + |\langle \mathbf{w}_{1,r}^{(T')}, \mathbf{u}'\rangle|$$
$$\leq \widetilde{O}(\sigma_0) + \widetilde{O}(\zeta)\cdot\bigg[\max_{t\in[T']}|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle| + \max_{t\in[T']}|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle| + \max_{t\in[T']}\big[|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle| + |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle|\big]\bigg]$$
$$\leq \widetilde{O}(\sigma_0) + \widetilde{O}(\zeta)\cdot\bigg[\max_{t\in[T]}|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle| + \max_{t\in[T]}|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle| + \max_{t\in[T]}\big[|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle| + |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle|\big]\bigg].$$

This further implies that

$$\max_{t\in[T]}\big[|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle| + |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle|\big]$$
$$\leq \widetilde{O}(\sigma_0) + \widetilde{O}(\zeta)\cdot\bigg[\max_{t\in[T]}|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle| + \max_{t\in[T]}|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle| + \max_{t\in[T]}\big[|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle| + |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle|\big]\bigg].$$

Then, rearranging terms will readily give the following result:

$$|\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{u}\rangle| + |\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{u}'\rangle| \leq \max_{t\in[T]}\big[|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle| + |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}'\rangle|\big]$$
$$\leq \widetilde{O}(\sigma_0) + \widetilde{O}(\zeta)\cdot\big[\max_{t\in[T]}\big[|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle| + \max_{t\in[T]}\big[|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle|\big]$$
$$\leq \widetilde{O}(\zeta^{3/2}),$$

where the last inequality holds since we must have

$$\max_{t\in[T]}|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle|, \max_{t\in[T]}|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}'\rangle| = O\big(\log(n)\cdot\zeta^{1/2}\big)$$

as otherwise, we cannot have $\max_{k\in[2],(i,j)\in\mathcal{S}}|F_k(\mathbf{W}^{(T)}; \mathbf{x}_{i,j})| \leq \zeta/2$ for all $t \leq T$, which contradicts the condition made in this lemma. This completes the upper bounds of $|\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{u}\rangle|$ and $|\langle \mathbf{w}_{1,r}^{(T)}, \mathbf{u}'\rangle|$.

□

C.5 Proof of Theorem 4.2

Proof of Theorem 4.2. We will evaluate the test error for common feature data and rare feature data separately. In particular, take the positive data $(\mathbf{x}, 1)$ as an example. Then note that the data $\mathbf{x}$ consists of the common feature $\mathbf{v}$, we can obtain the following by Lemma 5.6:

$$F_1(\mathbf{W}^{(t)}; \mathbf{x}) = \sum_{r=1}^{m} \sum_{p=1}^{P} \big(\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{x}^{(p)} \rangle\big)^2 \geq \sum_{r=1}^{m} \sum_{p:\mathbf{x}^{(p)}=\mathbf{v}} \big(\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle\big)^2 = \widetilde{\Omega}(\zeta).$$

On the other hand, we can follow the similar proof of Theorem 4.1 to show that $|\langle \mathbf{w}_{k,r}^{(T)}, \boldsymbol{\zeta} \rangle|^2 = \widetilde{O}(\sigma_p^2 n^2)$ with probability at least $1 - 1/\mathrm{poly}(n)$, then it follows that

$$F_2(\mathbf{W}^{(t)}; \mathbf{x}) = \sum_{r=1}^{m} \sum_{p=1}^{P} \big(\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{x}^{(p)} \rangle\big)^2 \leq \widetilde{O}(b\alpha^2 \zeta^3) + \widetilde{O}(\sigma_p^2 n^2) < F_1(\mathbf{W}^{(t)}; \mathbf{x}).$$

where we use the fact that $b\alpha^2 = o(1/\mathrm{polylog}(n))$ and $d = \omega(n^3 P)$. This clearly suggests that

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathrm{common}}}[\operatorname*{argmax}_k F_k(\mathbf{W}^{(t)}, \mathbf{x}) \neq y] \leq \frac{1}{\mathrm{poly}(n)}.$$

Then let's move on to the rare feature data. In particular, consider the positive rare feature data $(\mathbf{x}, 1)$, which contains the rare feature $\mathbf{v}'$, we have

$$F_1(\mathbf{W}^{(t)}; \mathbf{x}) = \sum_{r=1}^{m} \sum_{p=1}^{P} \big(\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{x}^{(p)} \rangle\big)^2 \geq\geq \sum_{r=1}^{m} \sum_{p:\mathbf{x}^{(p)}=\mathbf{v}'} \big(\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle\big)^2 = \widetilde{\Omega}(\rho^2 \zeta).$$

On the other hand, it holds that

$$F_2(\mathbf{W}^{(t)}; \mathbf{x}) = \sum_{r=1}^{m} \sum_{p=1}^{P} \big(\langle \mathbf{w}_{2,r}^{(t)}, \mathbf{x}^{(p)} \rangle\big)^2 \leq \widetilde{O}(b\alpha^2 \zeta^3) + \widetilde{O}(\sigma_p^2 n^2) = o(\rho^2 \zeta) < F_1(\mathbf{W}^{(t)}; \mathbf{x}),$$

where we use the fact that $b\alpha^2 \zeta^2 = o(\rho)$ and $d = \omega(n^3 P^3/\rho^2)$. Therefore, this implies that

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathrm{rare}}}[\operatorname*{argmax}_k F_k(\mathbf{W}^{(t)}, \mathbf{x}) \neq y] \leq \frac{1}{\mathrm{poly}(n)}.$$

Putting the results for common feature data and rare feature data together, we are able to complete the proof.

□