HiLD 2024: 2nd Workshop on High-dimensional Learning Dynamics

Provable Benefit of Cutout and CutMix for Feature Learning

Junsoo Oh JUNSOO.OH@KAIST.AC.KR
KAIST, Seoul, Republic of Korea

Chulhee Yun CHULHEE.YUN@KAIST.AC.KR
KAIST, Seoul, Republic of Korea

Abstract

Patch-level data augmentation such as Cutout and CutMix have shown significant efficacy in enhancing the performance of image-based tasks. However, a comprehensive theoretical understanding of these methods remains elusive. In this paper, we study two-layer neural networks trained using three distinct methods: vanilla training without augmentation, Cutout training, and CutMix training. Our analysis focuses on a feature-noise data model, which consists of several label-dependent features of varying rarity and label-independent noises of differing strengths. Our theorems demonstrate that Cutout training can learn features with low frequencies that vanilla training cannot, while CutMix training can even learn rarer features that Cutout cannot capture. From this, we establish that CutMix yields the highest test accuracy among the three. Our novel analysis reveals that CutMix training makes the network learn all features and noise vectors "evenly" regardless of the rarity and strength, which provides an interesting insight into understanding patch-level augmentation.

1. Introduction

Data augmentation is a crucial technique in deep learning, particularly in the image domain. It involves creating additional training examples by applying various transformations to the original data, thereby enhancing the generalization performance of deep learning models. Traditional data augmentation techniques typically focus on geometric transformations such as random rotations, horizontal and vertical flips, and cropping [25], or color-based adjustments such as color jittering [31].

In recent years, several new data augmentation techniques have appeared. Among them, patch-level data augmentation techniques like Cutout [13] and CutMix [35] have received considerable attention for their effectiveness in improving generalization. Cutout is a straightforward method where random rectangular regions of an image are removed during training. In comparison, CutMix adopts a more complex strategy by cutting and pasting sections from different images and using mixed labels, encouraging the model to learn from blended contexts. The success of Cutout and CutMix has triggered the development of numerous variants including Random Erasing [39], GridMask [5], CutBlur [34], Puzzle Mix [21], and Co-Mixup [22]. However, a lack of comprehensive theoretical understanding persists: *why and how do they work?*

In this paper, we aim to address this gap by offering a theoretical analysis of two important patch-level data augmentation techniques: Cutout and CutMix. Our theoretical framework draws inspiration from a recent study by Shen et al. [30], which explores a data model comprising multiple label-dependent feature vectors and label-independent noises of varying frequencies and intensities. The key idea behind this work is that learning features with low frequency can be challenging due to strong noises (i.e., low signal-to-noise ratio). We focus on investigating how Cutout and CutMix can aid in learning such less common features.

2. Problem Setting

In this section, we introduce the data distribution, neural network architecture, and formal description of the training methods considered in this paper.

Data Distribution. We consider a structured data, consisting of patches of label-dependent vectors (referred to as *features*) and label-independent vectors (referred to as *noise*).

Definition 1 (Feature Noise Patch Data) *We define a data distribution $\mathcal{D}$ on $\mathbb{R}^{d\times P} \times \{\pm 1\}$ such that $(\boldsymbol{X}, y) \sim \mathcal{D}$ where $y \in \{\pm 1\}$ is uniformly sampled and $\boldsymbol{X} = \left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(P)}\right) \in \mathbb{R}^{d\times P}$ is constructed as follows.*

1. *Let $\{\boldsymbol{v}_{s,k}\}_{s\in\{\pm 1\}, k\in[K]} \subset \mathbb{R}^d$ be a set of orthonormal* feature vectors. *Choose the feature vector $\boldsymbol{v} \in \mathbb{R}^d$ for data point $\boldsymbol{X}$ as $\boldsymbol{v} = \boldsymbol{v}_{y,k}$ with probability ρ_k from $\{\boldsymbol{v}_{y,k}\}_{k\in[K]} \subset \mathbb{R}^d$, where $\rho_1 + \cdots + \rho_K = 1$, and $\rho_1 \geq \dots \rho_K$. In our setting, there are three types of features with significantly different frequencies:* common features, rare features, and extremely rare features. *The indices of these features partition $[K]$ into $(\mathcal{K}_C, \mathcal{K}_R, \mathcal{K}_E)$.*
2. *We construct P patches of $\boldsymbol{X}$ as follows.*
   - ***Feature Patch**: Choose p^* uniformly from $[P]$ and we set $\boldsymbol{x}^{(p^*)} = \boldsymbol{v}$.*
   - ***Dominant Noise Patch**: Choose $\tilde{p}$ uniformly from $[P] \setminus \{p^*\}$. We construct $\boldsymbol{x}^{(\tilde{p})} = \alpha\boldsymbol{u} + \xi^{(\tilde{p})}$ where $\alpha\boldsymbol{u}$ is* feature noise *drawn uniformly from $\{\alpha\boldsymbol{v}_{1,1}, \alpha\boldsymbol{v}_{-1,1}\}$ and $\xi^{(\tilde{p})}$ is Gaussian* dominant noise *drawn from $N(\boldsymbol{0}, \sigma_{\mathrm{d}}^2\boldsymbol{\Lambda})$.*
   - ***Background Noise Patch**: The remaining patches $p \in [P] \setminus \{p^*, \tilde{p}\}$ consist of Gaussian* background noise*, i.e., we set $\boldsymbol{x}^{(p)} = \xi^{(p)}$ where $\xi^{(p)} \sim N(\boldsymbol{0}, \sigma_{\mathrm{b}}^2\boldsymbol{\Lambda})$.*

   Here, the noise covariance matrix $\boldsymbol{\Lambda} = \boldsymbol{I} - \sum_{s,k} \boldsymbol{v}_{s,k}\boldsymbol{v}_{s,k}^\top$ and the dominant noise is stronger than the background noise, i.e., $\sigma_{\mathrm{b}} < \sigma_{\mathrm{d}}$.

We refer to data with common, rare, and extremely rare features as *common, rare, and extremely rare data*, respectively. Our data distribution captures characteristics of image data, where the input consists of several patches. Some patches contain information relevant to the image labels, such as cat faces, while other patches contain information irrelevant to the labels, such as the background. Intuitively, there are two ways to fit the given data: learning features or memorizing noise. If a model fits the data by learning features, it can correctly classify test data having the same features. However, if a model fits the data by memorizing noise, it cannot generalize to unseen data. Thus, learning more features is crucial for achieving better generalization performance. We refer to data with common, rare, extremely rare features as common, rare, extremely rare data, respectively.

Neural Network Architecture. We focus on the following two-layer convolutional neural network.

Definition 2 (2-Layer CNN) *We define 2-layer CNN $f_{\boldsymbol{W}} : \mathbb{R}^{d\times P} \to \mathbb{R}$ parameterized by $\boldsymbol{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_{-1}\} \in \mathbb{R}^{d\times 2}$. For each input $\boldsymbol{X} = \left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(P)}\right) \in \mathbb{R}^{d\times P}$, we define*

$$f_{\boldsymbol{W}}(\boldsymbol{X}) = \sum_{p\in[P]} \phi\left(\left\langle \boldsymbol{w}_1, \boldsymbol{x}^{(p)} \right\rangle\right) - \sum_{p\in[P]} \phi\left(\left\langle \boldsymbol{w}_{-1}, \boldsymbol{x}^{(p)} \right\rangle\right),$$

where $\phi(\cdot)$ is a smoothed version of leaky ReLU activation, defined as follows.

$$\phi(z) = \begin{cases} z - \frac{(1-\beta)r}{2} & z \geq r \\ \frac{1-\beta}{2r} z^2 + \beta z & 0 \leq z \leq r \\ \beta z & z \leq 0 \end{cases}, \quad \text{where } 0 < \beta \leq 1 \text{ and } r > 0.$$

Previous works on the theory of feature learning often consider neural networks with (smoothed) ReLU or polynomial activation functions. However, we adopt a smoothed leaky ReLU activation, which always has a positive slope, to exclude the possibility of neurons "dying" during the complex optimization trajectory. Using smoothed leaky ReLU to analyze the learning dynamics of networks is not entirely new; there is a body of work that studies phenomena such as benign overfitting [14] and implicit bias [15, 24] by analyzing networks with (smoothed) leaky ReLU activation. We provide further discussions on our neural network architecture in Appendix B.1.

We aim to learn a distribution $\mathcal{D}$ from the training set using three distinct learning methods: vanilla training without any augmentation, Cutout, and CutMix. We first introduce necessary terminology for our data and architecture and next, formalize training methods within our framework.

Training Data. We consider a training set $\mathcal{Z} = \{(\boldsymbol{X}_i, y_i)\}_{i\in[n]}$ comprising n data points, each independently drawn from $\mathcal{D}$. For each $i \in [n]$, let $\boldsymbol{X}_i = \left(\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^{(P)}\right)$, and denote p_i^* and $\tilde{p}_i$ as the indices of a feature patch and dominant noise patch, respectively. For each feature vector $\boldsymbol{v}_{s,k}$ with $s \in \{\pm 1\}$ and $k \in [K]$, let $\mathcal{V}_{s,k}$ represent the set of indices of data points having the feature vector $\boldsymbol{v}_{s,k}$ and $\mathcal{V}_s$ denotes the set of indices of data with label s. For each data point $i \in [n]$ and dominant or background noise patch $p \in [P] \setminus \{p_i^*\}$, we refer to the Gaussian noise in $\boldsymbol{x}_i^{(p)}$ as $\xi_i^{(p)}$.

Initialization. We initialize the model parameters in our neural network using random initialization. Specifically, we initialize the model parameter $\boldsymbol{W}^{(0)} = \{\boldsymbol{w}_1^{(0)}, \boldsymbol{w}_{-1}^{(0)}\}$, where $\boldsymbol{w}_1^{(0)}, \boldsymbol{w}_{-1}^{(0)} \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{0}, \sigma_0^2 \boldsymbol{I}_d)$. Let us denote updated model parameters at iteration t as $\boldsymbol{W}^{(t)} = \{\boldsymbol{w}_1^{(t)}, \boldsymbol{w}_{-1}^{(t)}\}$.

Vanilla Training. The vanilla approach to training a model $f_{\boldsymbol{W}}$ is solving the empirical risk minimization problem using gradient descent on the ERM training loss $\mathcal{L}_{\mathrm{ERM}}(\cdot)$ with a constant learning rate η, where $\mathcal{L}_{\mathrm{ERM}}(\cdot)$ is defined as

$$\mathcal{L}_{\mathrm{ERM}}(\boldsymbol{W}) := \frac{1}{n}\sum_{i\in[n]} \ell(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_i)). \tag{1}$$

Here, $\ell(\cdot)$ is the logistic loss $\ell(z) = \log(1 + e^{-z})$, and We refer to this method as ERM.

Cutout Training. Cutout is a data augmentation technique that randomly cuts out rectangular regions of image inputs. In our patch-wise data, we regard Cutout training as using inputs with masked patches from the original data. For each subset $\mathcal{C}$ of $[P]$ and $i \in [n]$, we define augmented data $\boldsymbol{X}_{i,\mathcal{C}} \in \mathbb{R}^{d\times P}$ as a data generated by cutting $\mathcal{C}$ part of data $\boldsymbol{X}_i$. We can represent $\boldsymbol{X}_{i,\mathcal{C}}$ as

$$\boldsymbol{X}_{i,\mathcal{C}} = \left(\boldsymbol{x}_{i,\mathcal{C}}^{(1)}, \ldots, \boldsymbol{x}_{i,\mathcal{C}}^{(P)}\right), \text{ where } \boldsymbol{x}_{i,\mathcal{C}}^{(p)} = \begin{cases} \boldsymbol{x}_i^{(p)} & \text{if } p \notin \mathcal{C} \\ \boldsymbol{0} & \text{otherwise} \end{cases}.$$

The objective function for Cutout training can be defined as

$$\mathcal{L}_{\mathrm{Cutout}}(\boldsymbol{W}) := \frac{1}{n}\sum_{i\in[n]} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_{\mathcal{C}}}[\ell(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_{i,\mathcal{C}}))],$$

where $\mathcal{D}_{\mathcal{C}}$ is a uniform distribution on the set of subsets of $[P]$ with size C, where C is a hyperparameter satisfying $1 \le C < \frac{P}{2}$.[1] We refer to the process of training our model using gradient descent on Cutout loss $\mathcal{L}_{\mathrm{Cutout}}(\boldsymbol{W})$ with constant learning rate η as Cutout.

[1] DeVries and Taylor [13] also employ a moderate size of cutting, such as cutting 16 × 16 pixels on CIFAR-10 data, which originally has 32 × 32 pixels.

CutMix Training. CutMix involves not only cutting parts of images, but also pasting them into different images as well as assigning them mixed labels. For each subset $\mathcal{S}$ of $[P]$ and $i, j \in [n]$, we define augmented data $\boldsymbol{X}_{i,j,\mathcal{S}} \in \mathbb{R}^{d \times P}$ as the data obtained by cutting patches with indices in $\mathcal{S}$ from data $\boldsymbol{X}_i$ and pasting them into $\boldsymbol{X}_j$ at the same indices $\mathcal{S}$. We can represent $\boldsymbol{X}_{i,j,\mathcal{S}}$ as

$$\boldsymbol{X}_{i,j,\mathcal{S}} = \left(\boldsymbol{x}_{i,j,\mathcal{S}}^{(1)}, \dots, \boldsymbol{x}_{i,j,\mathcal{S}}^{(P)}\right), \text{ where } \boldsymbol{x}_{i,j,\mathcal{S}}^{(p)} = \begin{cases} \boldsymbol{x}_i^{(p)} & \text{if } p \in \mathcal{S}, \\ \boldsymbol{x}_j^{(p)} & \text{otherwise.} \end{cases}$$

The one-hot encoding of the labels y_i and y_j are also mixed with proportions $\frac{|\mathcal{S}|}{P}$ and $1 - \frac{|\mathcal{S}|}{P}$, respectively. This mixed label results in the CutMix training loss $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$, the objective function of CutMix training, which can be defined as

$$\mathcal{L}_{\text{CutMix}}(\boldsymbol{W}) := \frac{1}{n^2} \sum_{i,j \in [n]} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}} \left[\frac{|\mathcal{S}|}{P} \ell(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_{i,j,\mathcal{S}})) + \left(1 - \frac{|\mathcal{S}|}{P}\right) \ell(y_j f_{\boldsymbol{W}}(\boldsymbol{X}_{i,j,\mathcal{S}})) \right].$$

Here, $\mathcal{D}_{\mathcal{S}}$ is a probability distribution on the set of subsets of $[P]$ which samples $\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}$ as follows.[2]

1. Choose the size s of $\mathcal{S}$ uniformly random from $\{0, 1, \dots, P\}$, and
2. Choose $\mathcal{S}$ uniformly at random from the set of subsets of $[P]$ with size s.

We refer to the process of training our network using gradient descent on CutMix loss $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$ with constant learning rate η as CutMix.

Assumptions on the Choice of Hyperparameters. To control the quantities that appear in the analysis of training dynamics, we make assumptions on several quantities in our problem setting. For simplicity, we use choices of hyperparameters as a function of the dimension of patches d and consider sufficiently large d. We use the standard asymptotic notation $\mathcal{O}(\cdot), \Omega(\cdot), \Theta(\cdot), o(\cdot), \omega(\cdot)$ to express the dependency on d. We also use $\widetilde{\mathcal{O}}(\cdot), \widetilde{\Omega}(\cdot), \widetilde{\Theta}(\cdot)$ to hide logarithmic factors of d. Additionally, $\text{poly}(d)$ (or $\text{polylog}(d)$) represents quantities that increase faster than d^{c_1}(or $(\log d)^{c_1}$) and slower than d^{c_2} (or $(\log d)^{c_2}$) for some constant $0 < c_1 < c_2$. Finally, $o(1/\text{poly}(d))$ denotes some quantities that decrease faster than $1/d^c$ for any constant c. We provide discussions on the choice of hyperparameters in Appendix B.2 and list our assumptions in Assumption 8. There are many choices of parameters satisfying the set of assumptions, including:

$$P = 6, C = 2, n = \Theta\left(d^{0.4}\right), \alpha = \Theta\left(d^{-0.02}\right), \beta = \frac{1}{\text{polylog}(d)}, \sigma_0 = \Theta(d^{-0.2}), r = \Theta(d^{-0.2}),$$
$$\sigma_{\text{d}} = \Theta\left(d^{-0.305}\right), \sigma_{\text{b}} = \Theta\left(d^{-0.375}\right), \rho_R = \Theta\left(d^{-0.1}\right), \rho_E = \Theta\left(d^{-0.195}\right), \eta = \Theta(d^{-0.5}).$$

3. Main Results

In this section, we provide a characterization of the high probability guarantees for the behavior of models trained using three distinct methods we have introduced. We denote by T^* the maximum admissible training iterates and we assume $T^* = \frac{\text{poly}(d)}{\eta}$ with a sufficiently large polynomial in

[2] Other types of distributions, such as those considered in Yun et al. [35], make the same conclusion. We adopt this distribution for ease of presentation.

d. In all of our theorem statements, the randomness is over the sampling of training data and the initialization of models. We provide overviews of analysis in Appendix D.

The following theorem characterizes training accuracy and test accuracy achieved by ERM.

Theorem 3 *Let $\boldsymbol{W}^{(t)}$ be iterates of ERM. Then with probability at least $1 - o\left(\frac{1}{\mathrm{poly}(d)}\right)$, there exists T_{ERM} such that any $T \in [T_{\mathrm{ERM}}, T^*]$ satisfies the following:*

- *(Perfectly fits training set): For all $i \in [n], y_i f_{\boldsymbol{W}^{(T)}}(\boldsymbol{X}_i) > 0$.*
- *(Random on (extremely) rare data): $\mathbb{P}_{(\boldsymbol{X},y)\sim\mathcal{D}}\left[y f_{\boldsymbol{W}^{(T)}}(\boldsymbol{X}) > 0\right] = 1 - \frac{1}{2}\sum_{k \in \mathcal{K}_R \cup \mathcal{K}_E} \rho_k \pm o\left(\frac{1}{\mathrm{poly}(d)}\right)$.*

Theorem 3 demonstrates that ERM achieves perfect training accuracy; however, it performs almost like random guessing on unseen data points with rare and extremely rare features. This is because ERM can only learn common features and overfit rare or extremely rare data in the training set by memorizing noises to achieve perfect training accuracy. The formal proof is in Appendix F.2.

In comparison, we show that Cutout can perfectly fit both augmented training data and original training data, and it can also learn rare features that ERM cannot. However, Cutout still makes random guesses on test data with extremely rare features. We state these in the following theorem, with the proof in Appendix G.2:

Theorem 4 *Let $\boldsymbol{W}^{(t)}$ be iterates of Cutout training. Then with probability at least $1 - o\left(\frac{1}{\mathrm{poly}(d)}\right)$, there exists T_{Cutout} such that any $T \in [T_{\mathrm{Cutout}}, T^*]$ satisfies the following:*

- *(Perfectly fits augmented data): For all $i \in [n]$ and $\mathcal{C} \subset [P]$ with $|\mathcal{C}| = C$, $y_i f_{\boldsymbol{W}^{(T)}}(\boldsymbol{X}_{i,\mathcal{C}}) > 0$.*
- *(Perfectly fits original training data): For all $i \in [n], y_i f_{\boldsymbol{W}^{(T)}}(\boldsymbol{X}_i) > 0$.*
- *(Random on extremely rare data): $\mathbb{P}_{(\boldsymbol{X},y)\sim\mathcal{D}}\left[y f_{\boldsymbol{W}^{(T)}}(\boldsymbol{X}) > 0\right] = 1 - \frac{1}{2}\sum_{k \in \mathcal{K}_E} \rho_k \pm o\left(\frac{1}{\mathrm{poly}(d)}\right)$.*

In the case of CutMix, it is challenging to discuss train accuracy directly because the augmented data have soft labels generated by mixing pairs of labels. Instead, we prove that CutMix achieves a sufficiently small gradient norm of loss, and the training accuracy on the original training data is perfect. We also demonstrate that CutMix achieves almost perfect test accuracy by learning all kinds of features.

Theorem 5 *Let $\boldsymbol{W}^{(t)}$ be iterates of CutMix training. Then with probability at least $1 - o\left(\frac{1}{\mathrm{poly}(d)}\right)$, there exists some $T_{\mathrm{CutMix}} \in [0, T^*]$ that satisfies the following:*

- *(Finds a near stationary point): $\left\|\nabla_{\boldsymbol{W}} \mathcal{L}_{\mathrm{CutMix}}\left(\boldsymbol{W}^{(T_{\mathrm{CutMix}})}\right)\right\| = \frac{1}{\mathrm{poly}(d)}$.*
- *(Perfectly fits original training data): For all $i \in [n]$, $y_i f_{\boldsymbol{W}^{(T_{\mathrm{CutMix}})}}(\boldsymbol{X}_i) > 0$.*
- *(Almost perfectly classifies test data): $\mathbb{P}_{(\boldsymbol{X},y)\sim\mathcal{D}}\left[y f_{\boldsymbol{W}^{(T_{\mathrm{CutMix}})}}(\boldsymbol{X}) > 0\right] = 1 - o\left(\frac{1}{\mathrm{poly}(d)}\right)$.*

To prove Theorem 5, we characterize the global minimum of objective loss of CutMix. Surprisingly, at the global minimum, the model has even outputs for all features and noise vectors regardless of the frequency and strength. The complete proof appears in Appendix H.2. We believe that our approach can be applied to the analysis of other patch-level techniques using mixed labels such as Puzzle-Mix [21] and Co-Mixup [22].

Our three main theorems elucidate the benefits of Cutout and CutMix. Cutout enables a model to learn rarer features than ERM, while CutMix can outperform even Cutout. These advantages in learning features lead to improvements in generalization performance.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

[2] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[3] Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.

[4] Luigi Carratino, Moustapha Cissé, Rodolphe Jenatton, and Jean-Philippe Vert. On mixup regularization. *arXiv preprint arXiv:2006.06049*, 2020.

[5] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020.

[6] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 21321–21333, 2020.

[7] Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding the mixture-of-experts layer in deep learning. *Advances in neural information processing systems*, 35:23049–23062, 2022.

[8] Zixiang Chen, Junkai Zhang, Yiwen Kou, Xiangning Chen, Cho-Jui Hsieh, and Quanquan Gu. Why does sharpness-aware minimization generalize better than sgd? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[9] Muthu Chidambaram and Rong Ge. For better or for worse? learning minimum variance features with label augmentation. *arXiv preprint arXiv:2402.06855*, 2024.

[10] Muthu Chidambaram, Xiang Wang, Yuzheng Hu, Chenwei Wu, and Rong Ge. Towards understanding the data dependency of mixup-style training. *arXiv preprint arXiv:2110.07647*, 2021.

[11] Muthu Chidambaram, Xiang Wang, Chenwei Wu, and Rong Ge. Provably learning diverse features in multi-view data with midpoint mixup. In *International Conference on Machine Learning*, pages 5563–5599. PMLR, 2023.

[12] Tri Dao, Albert Gu, Alexander Ratner, Virginia Smith, Chris De Sa, and Christopher Ré. A kernel theory of modern data augmentation. In *International conference on machine learning*, pages 1528–1537. PMLR, 2019.

[13] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[14] Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pages 2668–2703. PMLR, 2022.

[15] Spencer Frei, Gal Vardi, Peter L Bartlett, Nathan Srebro, and Wei Hu. Implicit bias in leaky relu networks trained on high-dimensional data. *arXiv preprint arXiv:2210.07082*, 2022.

[16] Boris Hanin and Yi Sun. How data augmentation affects optimization for linear regression. *Advances in Neural Information Processing Systems*, 34:8095–8105, 2021.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Wei Huang, Yuan Cao, Haonan Wang, Xin Cao, and Taiji Suzuki. Graph neural networks provably benefit from structural information: A feature learning perspective. *arXiv preprint arXiv:2306.13926*, 2023.

[19] Wei Huang, Ye Shi, Zhongyi Cai, and Taiji Suzuki. Understanding convergence and generalization in federated learning through feature learning theory. In *The Twelfth International Conference on Learning Representations*, 2023.

[20] Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, pages 9965–10040. PMLR, 2022.

[21] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*, pages 5275–5285. PMLR, 2020.

[22] Jang-Hyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. *arXiv preprint arXiv:2102.03065*, 2021.

[23] Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer relu convolutional neural networks. In *International Conference on Machine Learning*, pages 17615–17659. PMLR, 2023.

[24] Yiwen Kou, Zixiang Chen, and Quanquan Gu. Implicit bias of gradient descent for two-layer relu and leaky relu networks on nearly-orthogonal data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[26] Binghui Li and Yuanzhi Li. Why clean generalization and robust overfitting both happen in adversarial training. *arXiv preprint arXiv:2306.01271*, 2023.

[27] Junsoo Oh and Chulhee Yun. Provable benefit of mixup for finding optimal decision boundaries. In *International Conference on Machine Learning*, pages 26403–26450. PMLR, 2023.

[28] Chanwoo Park, Sangdoo Yun, and Sanghyuk Chun. A unified analysis of mixed sample data augmentation: A loss function perspective. *Advances in Neural Information Processing Systems*, 35:35504–35518, 2022.

[29] Shashank Rajput, Zhili Feng, Zachary Charles, Po-Ling Loh, and Dimitris Papailiopoulos. Does data augmentation lead to positive margin? In *International Conference on Machine Learning*, pages 5321–5330. PMLR, 2019.

[30] Ruoqi Shen, Sébastien Bubeck, and Suriya Gunasekar. Data augmentation as feature manipulation. In *International conference on machine learning*, pages 19773–19808. PMLR, 2022.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[32] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[33] Sen Wu, Hongyang Zhang, Gregory Valiant, and Christopher Ré. On the generalization effects of linear transformations in data augmentation. In *International Conference on Machine Learning*, pages 10410–10420. PMLR, 2020.

[34] Jaejun Yoo, Namhyuk Ahn, and Kyung-Ah Sohn. Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8375–8384, 2020.

[35] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[36] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[37] Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? *arXiv preprint arXiv:2010.04819*, 2020.

[38] Linjun Zhang, Zhun Deng, Kenji Kawaguchi, and James Zou. When and how mixup improves calibration. In *International Conference on Machine Learning*, pages 26135–26160. PMLR, 2022.

[39] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[40] Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of adam in learning neural networks with proper regularization. *arXiv preprint arXiv:2108.11371*, 2021.

[41] Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. The benefits of mixup for feature learning. In *International Conference on Machine Learning*, pages 43423–43479. PMLR, 2023.

Contents

Appendix A. Related Works

Feature Learning Theory. Our work aligns with a recent line of studies investigating how training methods and neural network architectures influence feature learning. These studies focus on a specific data distribution composed of two components: label-dependent features and label-independent noise. The key contribution of this body of work is the exploration of which training methods or neural networks are most effective at learning meaningful features and achieving good generalization performance. Allen-Zhu and Li [1] demonstrate that an ensemble model can achieve near-perfect performance by learning diverse features, while a single model tends to learn only certain parts of the feature space, leading to lower test accuracy. In other works, Cao et al. [3], Kou et al. [23] explore the phenomenon of benign overfitting when training a two-layer convolutional neural network. The authors identify the specific conditions under which benign overfitting occurs, providing valuable insights into how these networks behave during training. Several other studies seek to understand various aspects of deep learning through the lens of feature learning [7, 8, 18–20, 26, 40].

Theoretical Analysis of Data Augmentation. Several works aim to analyze traditional data augmentation from different perspectives, including kernel theory [12], margin-based approach [29], regularization effects [33], group invariance [6], and impact on optimization [16]. Moreover, many papers have explored various aspects of a recent technique called Mixup [36]. For example, studies have explored its regularization effects [4, 37], its role in improving calibration [38], its ability to find optimal decision boundaries [27] and its potential negative effects [9, 10]. Some works investigate the broader framework of Mixup, including CutMix, which aligns with the scope of our work. Park et al. [28] study the regularization effect of mixed-sample data augmentation within a unified framework that contains both Mixup and CutMix. In Oh and Yun [27], the authors analyze masking-based Mixup, which is a class of Mixup variants that also includes CutMix. In their context, they show that masking-based Mixup can deviate from the Bayes optimal classifier but require less training sample complexity. However, neither work provides a rigorous explanation for why CutMix has been successful. The studies most closely related to our work include Chidambaram et al. [11], Shen et al. [30], Zou et al. [41]. Shen et al. [30] regard traditional data augmentation as a form of feature manipulation and investigate its advantages from a feature learning perspective. Both Chidambaram et al. [11] and Zou et al. [41] analyze Mixup within a feature learning framework. However, patch-level data augmentation such as Cutout and CutMix, which are the focus of our work, have not yet been explored within this context.

Comparison to Previous Work. Our data distribution is similar to those considered in Shen et al. [30] and Zou et al. [41], which investigate the benefits of standard data augmentation methods and Mixup by comparing them to vanilla training without any augmentation. These results consider two types of features—common and rare—with different levels of rarity, along with two types of noise: feature noise and dominant noise. However, we consider three types of features: common, rare, and extremely rare, and three types of noise: feature noise, dominant noise, and background noise. this distinction allows us to compare three distinct methods and demonstrate the differences between them, whereas Shen et al. [30] and Zou et al. [41] compared only two methods.

Appendix B. Further Discussion on Problem Setting

B.1. Discussion on the Choice of Neural Network Architecture

A key difference between ReLU and leaky ReLU lies in the possibility of ReLU neurons "dying" in the negative region, where some negatively initialized neurons remain unchanged throughout training. As a result, using ReLU activation requires multiple neurons to ensure the survival of neurons at initialization, which becomes increasingly probable as the number of neurons increases. In contrast, the derivative of leaky ReLU is always positive, ensuring that a single neuron is often sufficient. Therefore, for mathematical simplicity, we consider the case where the network has a single neuron for each positive and negative output. We believe that our analysis can be extended to the multi-neuron case as we validate numerically in Appendix C.1.

B.2. Discussion on the Choice of Hyper parameters

We assume that $P = \Theta(1)$ for simplicity. Additionally, we consider a high-dimensional regime where the number of data points is much smaller than the dimension d, which is expressed as $n = o\left(\beta \sigma_{\mathrm{d}}^{-1} \sigma_{\mathrm{b}} d^{\frac{1}{2}}\right)$. We also assume that $\rho_k n = \omega\left(n^{\frac{1}{2}} \log d\right)$, which ensures the sufficiency of data with each feature.

In addition, as we will describe in Appendix D, the relative scales between the frequencies of features and the strengths of noises play crucial roles in our analysis, as they serve as a proxy for the "learning speed" in the initial phase. For common features $k \in \mathcal{K}_C$, we assume $\rho_k = \Theta(1)$ and the learning speed of common features is much faster than that of dominant noise which can be formulated as $\sigma_{\mathrm{d}}^2 d = o(\alpha^2 \beta^2 n)$. For rare features $k \in \mathcal{K}_R$, we assume $\rho_k = \Theta(\rho_R)$ for some ρ_R, and we consider the case where the learning speed of rare features is much slower than that of dominant noise but faster than background noise, which is expressed as $\rho_R n = o(\alpha^2 \beta^2 \sigma_{\mathrm{d}}^2 d), \omega(\alpha^{-2} \beta^{-2} \sigma_{\mathrm{b}}^2 d)$. Finally, for extremely rare features $k \in \mathcal{K}_E$, we say $\rho_k = \Theta(\rho_E)$ for some ρ_E and their learning is even slower than that of background noises, which can be expressed as $\rho_E n = o(\alpha^2 \beta^2 \sigma_{\mathrm{b}}^2 d)$.

Lastly, we assume the strength of feature noise satisfies $\alpha = o\left(\max\left\{\beta, \frac{\sigma_{\mathrm{d}}^2 d}{n}\right\}\right), \omega\left(n \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right)$, and $r, \sigma_0, \eta > 0$ are sufficiently small so that $\sigma_0, r = o\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right)$, $\eta = o(\sigma_{\mathrm{d}}^{-2} d^{-1})$.

Appendix C. Experiments

We conduct experiments both in our setting and real-world data CIFAR-10 to support our theoretical findings.

C.1. Numerical Experiments on Our Data

For the numerical experiments on our setting, we set the number of patches $P = 3$, dimension $d = 2000$, number of data points $n = 300$, dominant noise strength $\sigma_\mathrm{d} = 0.25$, background noise strength $\sigma_\mathrm{b} = 0.15$, and feature noise strength $\alpha = 0.005$. The feature vectors are given by the standard basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \boldsymbol{e}_4, \boldsymbol{e}_5, \boldsymbol{e}_6 \in \mathbb{R}^d$, where $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ are features for the positive label $y = 1$ and $\boldsymbol{e}_4, \boldsymbol{e}_5, \boldsymbol{e}_6$ are features for the negative label $y = -1$. We categorize $\boldsymbol{e}_1$ and $\boldsymbol{e}_4$ as common features with a frequency of 0.8, $\boldsymbol{e}_2$ and $\boldsymbol{e}_5$ as rare features with a frequency of 0.15, and lastly,$\boldsymbol{e}_3$ and $\boldsymbol{e}_6$ as extremely rare features with a frequency of 0.05. For the learner network, we set the slope of negative regime $\beta = 0.1$ and the length of the smoothed part $r = 1$. We train models using three methods: ERM, Cutout, and CutMix with a learning rate $\eta = 1$. For Cutout, we cut a single patch of data. We apply full-batch gradient descent and for Cutout and CutMix, we utilize all possible augmented data.[3]

For each feature vector $\boldsymbol{v}$ for a positive label , we plot the output of feature vector $\phi(\langle \boldsymbol{w}_1^{(t)}, \boldsymbol{v}\rangle) - \phi(\langle \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{v}\rangle)$ throughout training. Our numerical findings confirm that ERM can only learn common features, Cutout can learn common and rare features but cannot learn extremely rare features, and CutMix can learn all types of features. Especially, CutMix learn common features, rare features, and extremely rare features almost evenly. Also, we observed non-monotone behavior of the output in case of CutMix, which motivated our novel proof technique.

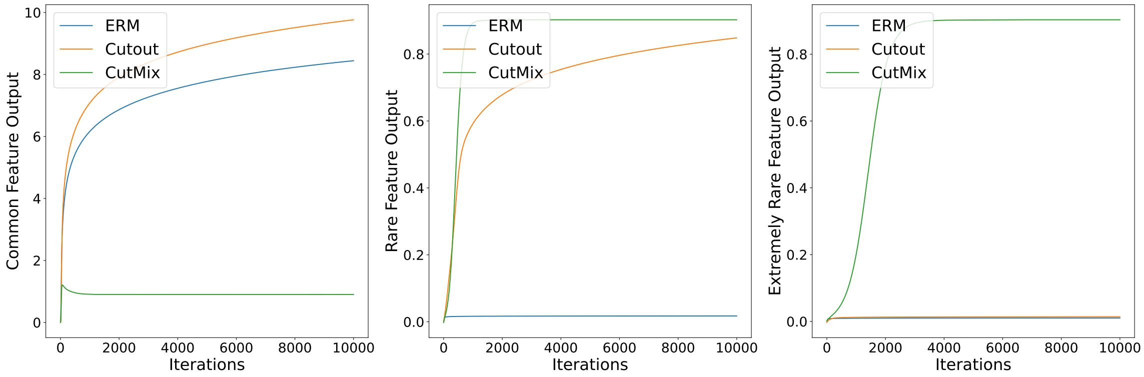

Figure 1: Numerical results on our problem setting. We validate our findings on the trends of ERM, Cutout, and CutMix in learning common feature (Left), rare feature (Center), and extremely rare feature (Right). The output of a common feature trained by CutMix shows non-monotone behavior.

The same trends are observed with different architectures, such as a smoothed (leaky) ReLU network with multiple neurons . We further conducted numerical experiments on our data distribution

[3] For CutMix, this may induce different choices of $\mathcal{D}_\mathcal{S}$ from those assumed in our analysis, but we mention that other general choices of $\mathcal{D}_\mathcal{S}$ do not affect trends in analysis.

by applying two variations to our architecture: increasing the number of neurons, and increasing the number of neurons with a smoothed ReLU activation (instead of smoothed leaky ReLU). We follow the same setting as previously introduced, except for the neural network architecture, the strength of dominant/background noise $\sigma_{\mathrm{d}}, \sigma_{\mathrm{b}}$, and the frequencies of features. We observed the same trends as predicted by our theoretical findings and shown in Figure 1.

For the multi-neuron with smoothed Leaky ReLU case (Figure 2), we use 10 neurons for each positive/negative output with the slope of negative regime $\beta = 0.1$ and the length of polynomial regime $r = 1$. We set the strength of dominant noise $\sigma_{\mathrm{d}} = 0.25$, and the strength of background noise $\sigma_{\mathrm{b}} = 0.12$. In addition, frequencies of common features, rare features, and extremely rare features are set to 0.72, 0.15, and 0.03, respectively.

For the multi-neuron with smoothed ReLU case i.e., $\beta = 0$ (Figure 3), we set the length of the polynomial regime as $r = 1$, and we use 10 neurons for each positive/negative output. We set the remaining hyperparameters as follows: the strength of dominant noise $\sigma_{\mathrm{d}} = 0.25$, and the strength of background noise $\sigma_{\mathrm{b}} = 0.12$. In addition, frequencies of common features, rare features, and extremely rare features are set to 0.75, 0.2, and 0.05, respectively.

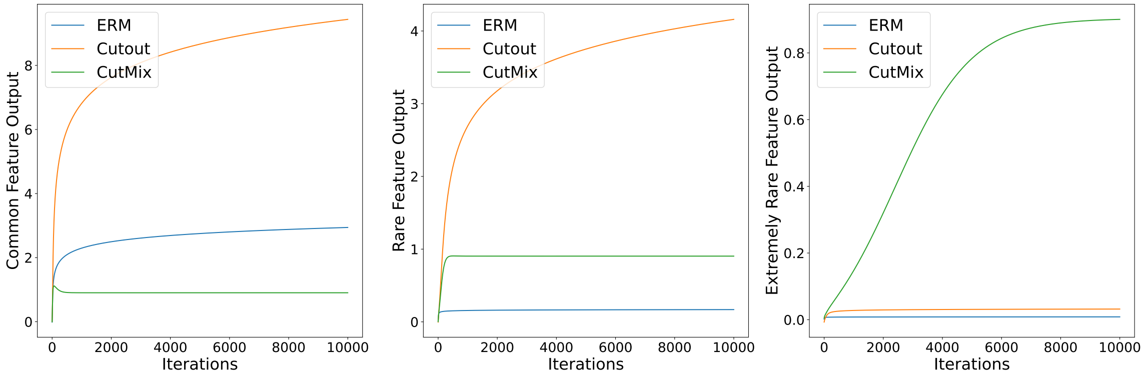

Figure 2: Multi-neuron with a smoothed leaky ReLU actiation

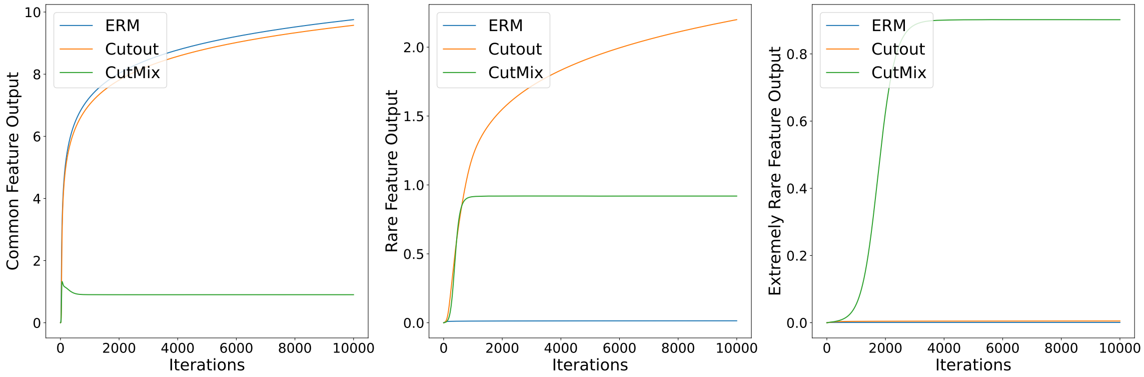

Figure 3: Multi-neuron with a smoothed ReLU

C.2. Experiments on CIFAR-10 Dataset

We compare three methods, ERM training, Cutout training, and CutMix training on CIFAR-10 classification. For ERM training, we apply only random cropping and random horizontal flipping on train dataset. In comparison, for Cutout training and CutMix training, we additionally apply Cutout and CutMix, respectively, on training data. For Cutout training, we randomly cut 16×16 pixels of input images, and for CutMix training, we sample the mixing ratio from a beta distribution $\mathrm{Beta}(0.5, 0.5)$. We train ResNet-18 [17] for 200 epochs with a batch size of 128 using SGD with a learning rate 0.1, momentum 0.9, and weight decay 5×10^{-4}. Trained models using ERM, Cutout, and CutMix achieve test accuracy 95.16%, 96.05%, and 96.29%, respectively.

We randomly generate augmented data using CutMix from pairs of cat images and dog images in CIFAR-10 with varying mixing ratios $\lambda = 1, 0.8, 0.6$ (Dog:Cat = $\lambda : 1 - \lambda$). We randomly make $5,000$ (cat, dot)-pairs in CIFAR-10 training set and apply CutMix randomly 10 times. By repeating this procedure 10 times, we generate total $5,000 \times 10 \times 10 = 500,000$ augmented samples for each mixing ratio λ. We plot a histogram of dog prediction output subtracted by cat prediction output (before applying the softmax function), evaluated on $500,000$ augmented data in Figure 4.

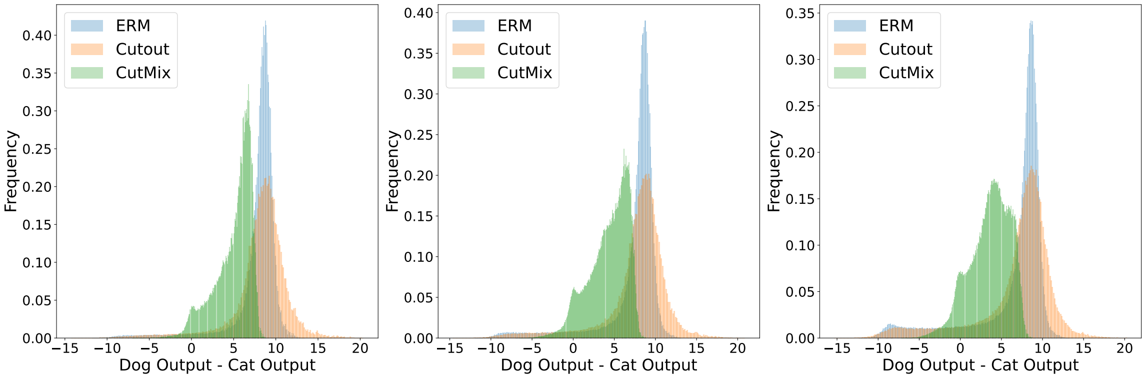

Figure 4: Histogram of dog prediction output subtracted by cat prediction output evaluated on data points augmented by CutMix data using cat data and dog data with varying mixing ratio λ (Dog : Cat $= \lambda : 1 - \lambda$) (Left) $\lambda = 1$, (Center) $\lambda = 0.8$, (Right) $\lambda = 0.6$

The leftmost plot represents the evaluation results for original dog images, as it uses a mixing ratio of $\lambda = 1$. We can observe that the output of the model trained using Cutout is skewed toward higher values compared to the output of the model trained using other methods. We believe this aligns with the theoretical intuition that Cutout learns more information from the original image using augmented data.

The remaining two plots show the output for randomly augmented data using CutMix. We observe that the models trained with CutMix exhibit a shorter tail, supporting our intuition from the CutMix analysis that the models learn uniformly across all patches.

Appendix D. Overview of Analysis

In this section, we discuss key proof ideas and the main challenges in our analysis. For ease of presentation, we consider the case $\alpha = 0$. Although we assume lower bounds on α, this is to show guarantees on the test accuracy and does not significantly affect the feature learning procedure.

D.1. Vanilla Training and Cutout Training

We will explain why ERM fails to learn (extremely) rare features, while Cutout can learn rare features but not extremely rare features. The formal proof of Theorem 3 and Theorem 4 appear in Appendix F.2 and Appendix G.2. Let us consider ERM. From (1), for $s, s' \in \{\pm 1\}, k \in [K], i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$, the component of $\boldsymbol{w}_s$ in the feature vector $\boldsymbol{v}_{s',k}$'s direction is updated as

$$\left\langle \boldsymbol{w}_s^{(t+1)}, \boldsymbol{v}_{s',k} \right\rangle = \left\langle \boldsymbol{w}_s^{(t+1)}, \boldsymbol{v}_{s',k} \right\rangle - \frac{ss'\eta}{n} \sum_{j \in \mathcal{V}_{s',k}} \ell'(y_j f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_j))\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s',k} \right\rangle\right), \qquad (2)$$

and similarly, the "update" of inner product of $\boldsymbol{w}_s$ with a noise patch $\xi_i^{(p)}$ can be written as

$$\left\langle \boldsymbol{w}_s^{(t+1)}, \xi_i^{(p)} \right\rangle \approx \left\langle \boldsymbol{w}_s^{(t)}, \xi_i^{(p)} \right\rangle - \frac{sy_i\eta}{n} \ell'(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i))\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \xi_i^{(p)} \right\rangle\right) \left\| \xi_i^{(p)} \right\|^2, \qquad (3)$$

where the approximation is due to the near-orthogonality of Gaussian random vectors in the high-dimensional regime. This approximation shows that $\langle \boldsymbol{w}_s^{(t+1)}, \boldsymbol{v}_{s',k} \rangle$'s and $\langle \boldsymbol{w}_s^{(t)}, \xi_i^{(p)} \rangle$'s are almost monotonically increasing or decreasing. We address the approximation errors using a variant of the technique introduced by Cao et al. [3], as detailed in Appendix E.2.

From (2) and (3), we can observe that in the early phase of training satisfying $-\ell'(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)) = \Theta(1)$, the main factor for the speed of learning features and noises are number of feature occurrence $|\mathcal{V}_{s',k}|$ and the strength of noises $\|\xi_i^{(p)}\|^2$. From our assumption we have introduced in Section B.2, if we compare the learning speed of each component, we have

common features $\gg$ dominant noises $\gg$ rare features $\gg$ background noises $\gg$ extremely rare features.

Based on this observation, we conduct a three-phase analysis for ERM.

- **Phase 1**: Learning common features quickly.
- **Phase 2**: Fitting (extremely) rare data by memorizing dominant noises instead of learning features.
- **Phase 3**: A model cannot learn (extremely) rare features since gradients of all data are small.

The main intuition behind why ERM cannot learn (extremely) rare features is that the gradients of all data containing these features become small after quickly memorizing dominant noise patches. In contrast, since Cutout randomly cuts some patches out, there exists augmented data that does not contain dominant noise and includes only features and background noise. This allows Cutout to learn rare features, thanks to these augmented data. However, extremely rare features cannot be learned since the learning speed of background noise is much faster and there are too many background noise patches to cut them all out.

Remark 6 *Shen et al. [30] conduct analysis on vanilla training and training using standard data augmentation, sharing the same intuition in similar but different data models and neural networks. Also, we emphasize that we proved the inability to learn (extremely) rare features within* $\frac{\text{poly}(d)}{\eta}$ *iterations, whereas Shen et al. [30] consider only the first iteration achieving perfect training accuracy.*

D.2. CutMix Training

In learning dynamics of ERM and Cutout, inner products between weight and data patches evolve (approximately) monotonically, which makes the analysis much more feasible. However, analyzing the learning dynamics of CutMix involves non-monotone inner products which are inevitable since CutMix uses mixed labels, as demonstrated in the experimental results in our setting (Section C.1, Figure 1). Non-monotonicity and non-convexity of the problem necessitates novel proof strategies.

Let us define $\boldsymbol{Z} := \{z_{s,k}\}_{s\in\{\pm 1\},k\in[K]} \cup \{z_i^{(p)}\}_{i\in[n],p\in[P]\setminus\{p_i^*\}}$ as a function of $\boldsymbol{W}$ as follows,

$$z_i^{(p)} := \phi\left(\left\langle \boldsymbol{w}_1, \xi_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-1}, \xi_i^{(p)}\right\rangle\right), \quad z_{s,k} := \phi(\langle \boldsymbol{w}_1, \boldsymbol{v}_{s,k}\rangle) - \phi(\langle \boldsymbol{w}_{-1}, \boldsymbol{v}_{s,k}\rangle).$$

Then, $\boldsymbol{Z}$ represents the contribution of each noise patch and feature vector to the neural network output, and the nonconvex function $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$ can be viewed as the composition of $\boldsymbol{Z}(\boldsymbol{W})$ and a convex function $h(\boldsymbol{Z})$. By using the convexity of $h(\boldsymbol{Z})$, we can characterize the global minimum of $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$. Surprisingly, we show that any global minimizer $\boldsymbol{W}^* = \{\boldsymbol{w}_1^*, \boldsymbol{w}_{-1}^*\}$ satisfies

$$\phi\left(\left\langle \boldsymbol{w}_s^*, \boldsymbol{x}_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^*, \boldsymbol{x}_i^{(p)}\right\rangle\right) = C_s,$$

for all $s \in \{\pm 1\}, i \in \mathcal{V}_s$, and $p \in [P]$, with some constants $C_1, C_{-1} = \Theta(1)$. In other words, at the global minimum, the output of model on each patch of the training data is uniform across the set of data with the same labels. We also prove that CutMix can achieve a point close to the global minimum within $\frac{\text{poly}(d)}{\eta}$ iterations. As a result, the model trained by CutMix can learn all features including extremely rare features. The complete proof of Theorem 5 appears in Appendix H.2.

Remark 7 *Zou et al. [41] investigate Mixup in a similar feature-noise model and show that Mixup can learn rarer features than vanilla training, with its benefits emerging from the early dynamics of training. However, our characterization of the global minimum of $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$ and experimental results in our setting (Section C.1, Figure 1) suggest that the benefits of CutMix, especially for learning extremely rare features, arise from the later stages of training. This suggests that Mixup and CutMix have different underlying mechanisms.*

Appendix E. Preliminary

In our analysis, we consider the choice of hyperparameters as a function of the dimension of patches d and consider sufficiently large d. Let us summarize the assumptions on the parameters for the problem setting and assume they hold.

Assumption 8 *The following conditions hold.*

A1 (The number of patches) $P = \Theta(1)$ *and* $P \geq 3$.

A2 (Overparameterized regime): $n = o\left(\beta \sigma_{\mathrm{d}}^{-1} \sigma_{\mathrm{b}} d^{\frac{1}{2}}\right)$.

A3 (Sufficient feature data): $\rho_k n = \omega\left(n^{\frac{1}{2}} \log d\right)$.

A4 (Common feature vs dominant noise): $\alpha^{-2}\beta^{-2}\sigma_{\mathrm{d}}^2 d = o(n)$.

A5 (Rare feature vs dominant/background noise): $\rho_R n = o(\alpha^2\beta^2\sigma_{\mathrm{d}}^2 d)$ *and* $\rho_R n = \omega(\alpha^{-2}\beta^{-2}\sigma_{\mathrm{b}}^2 d)$.

A6 (Extremely rare feature vs background noise) $\rho_E n = o\left(\alpha^2\beta^2\sigma_{\mathrm{b}}^2 d\right)$.

A7 (Strength of feature noise) $\alpha = o\left(\max\left\{\beta, \frac{\sigma_{\mathrm{d}}^2 d}{n}\right\}\right)$ *and* $\alpha = \omega\left(n\sigma_d \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right)$.

A8 $r, \sigma_0 = o\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right), \eta = o(\sigma_{\mathrm{d}}^{-2} d^{-1})$

E.1. Quantities at the Beginning

We characterize some quantities at the beginning of training.

Lemma 9 *Let* E_{init} *the event such that all the following holds:*

- $\frac{25}{52}n \leq |\mathcal{V}_1|, |\mathcal{V}_{-1}| \leq \frac{27}{52}n$
- *For each* $s \in \{\pm 1\}$ *and* $k \in [K]$, $\frac{\rho_k n}{4} \leq |\mathcal{V}_{s,k}| \leq \frac{3\rho_k n}{4}$
- $\cup_{i \in \mathcal{V}_{1,1}} \{p_i^*\} = [P]$
- *For any* $s, s' \in \{\pm 1\}$ *and* $k \in [K]$, $\left|\left\langle \boldsymbol{w}_s^{(0)}, \boldsymbol{v}_{s',k}\right\rangle\right| \leq \sigma_0 \log d$.
- *For any* $s \in \{\pm 1\}$ *and* $i \in [n]$, $\left|\left\langle \boldsymbol{w}_s^{(0)}, \xi_i^{(\tilde{p}_i)}\right\rangle\right| \leq \sigma_0 \sigma_{\mathrm{d}} d^{\frac{1}{2}} \log d$.
- *For any* $s \in \{\pm 1\}, i \in [n]$ *and* $p \in [P] \setminus \{p_i^*, \tilde{p}_i\}$, $\left|\left\langle \boldsymbol{w}_s^{(0)}, \xi_i^{(p)}\right\rangle\right| \leq \sigma_0 \sigma_{\mathrm{b}} d^{\frac{1}{2}} \log d$.
- *For any* $i, j \in [n]$ *with* $i \neq j$, $\frac{1}{2}\sigma_{\mathrm{d}}^2 d \leq \left\|\xi_i^{(\tilde{p}_i)}\right\|^2 \leq \frac{3}{2}\sigma_{\mathrm{d}}^2 d$ *and* $\left|\left\langle \xi_i^{(\tilde{p}_i)}, \xi_j^{(\tilde{p}_j)}\right\rangle\right| \leq \sigma_{\mathrm{d}}^2 d^{\frac{1}{2}} \log d$.
- *For any* $i, j \in [n]$ *and* $p \in [P] \setminus \{p_j^*, \tilde{p}_j\}$, $\left|\left\langle \xi_i^{(\tilde{p}_i)}, \xi_j^{(p)}\right\rangle\right| \leq \sigma_{\mathrm{d}} \sigma_{\mathrm{b}} d^{\frac{1}{2}} \log d$.
- *For any* $i, j \in [n]$ *and* $p \in [P] \setminus \{p_i^*, \tilde{p}_i\}, q \in [P] \setminus \{p_j^*, \tilde{p}_j\}$ *with* $(i, p) \neq (j, q)$, $\frac{1}{2}\sigma_{\mathrm{b}}^2 d \leq \left\|\xi_i^{(p)}\right\|^2 \leq \frac{3}{2}\sigma_{\mathrm{b}}^2 d$ *and* $\left|\left\langle \xi_i^{(p)}, \xi_j^{(q)}\right\rangle\right| \leq \sigma_{\mathrm{b}}^2 d^{\frac{1}{2}} \log d$.

- $\{\boldsymbol{v}_{s,k}\}_{s\in\{\pm 1\},k\in[K]} \cup \{\boldsymbol{x}_i^{(p)}\}_{i\in[n],p\in[P]\setminus\{p_i^*\}}$ *is linearly independent.*

Then, the event E_{init} *holds with probability at least* $1 - o\left(\frac{1}{\text{poly}(d)}\right)$*. Furthermore, if* $\xi \sim N(\mathbf{0}, \sigma^2\Lambda)$ *is independent of* $\boldsymbol{w}_1^{(0)}, \boldsymbol{w}_{-1}^{(0)}$ *and* $\{(\boldsymbol{X}_i, y_i)\}_{i\in[n]}$*, we have*

$$\left|\left\langle \boldsymbol{w}_1^{(0)}, \xi\right\rangle\right|, \left|\left\langle \boldsymbol{w}_{-1}^{(0)}, \xi\right\rangle\right| \leq \sigma_0\sigma d^{\frac{1}{2}}\log d, \text{ and } \left|\left\langle \xi, \xi_i^{(p)}\right\rangle\right| \leq \sigma\sigma_{\text{d}} d^{\frac{1}{2}}\log d,$$

for all $i \in [n]$ *and* $p \in [P]\setminus\{p_i^*\}$*, with probability at least* $1 - o\left(\frac{1}{\text{poly}(d)}\right)$.

Proof [Proof of Lemma 9] Let us prove the first three bullet points hold with probability at least $1 - o\left(\frac{1}{\text{poly}(d)}\right)$. By Höeffding's inequality,

$$\begin{aligned}\mathbb{P}\left[\left||\mathcal{V}_1| - \frac{n}{2}\right| > \frac{n}{52}\right] &= \mathbb{P}\left[\left|\sum_{i\in[n]}(\mathbb{1}_{y_i=1} - \mathbb{E}[\mathbb{1}_{y_i=1}])\right| > \frac{n}{52}\right]\\ &\leq 2\exp\left(-\frac{2}{52^2}n\right) = o\left(\frac{1}{\text{poly}(d)}\right),\end{aligned}$$

where the last equality is due to A3. In addition, for each $s \in \{\pm 1\}, k \in [K]$, by Höeffding's inequality

$$\begin{aligned}\mathbb{P}\left[\left||\mathcal{V}_{s,k}| - \frac{\rho_k}{2}n\right| > \frac{\rho_k}{4}n\right] &= \mathbb{P}\left[\left|\sum_{i\in[n]}\left(\mathbb{1}_{i\in\mathcal{V}_{s,k}} - \mathbb{E}[\mathbb{1}_{i\in\mathcal{V}_{s,k}}]\right)\right| > \frac{\rho_k}{4}n\right]\\ &\leq 2\exp\left(-\frac{\rho_k^2}{8}n\right) = o\left(\frac{1}{\text{poly}(d)}\right),\end{aligned}$$

where the last equality is due to A3. Also, for each $i \in [n]$ and $p \in [P]$,

$$\mathbb{P}[\{i \in \mathcal{V}_{1,1}\} \cap \{p_i^* = p\}] = \frac{\rho_1}{P}.$$

Hence,

$$\begin{aligned}\mathbb{P}\left[\cup_{i\in\mathcal{V}_{1,1}}\{p_i^*\} \neq [P]\right] &\leq \sum_{p\in[P]}\mathbb{P}\left[\cap_{i\in[n]}\left((\{i \in \mathcal{V}_{1,1}\} \cap \{p_i^* = p\})^{\complement}\right)\right]\\ &= P\left(1 - \frac{\rho_1}{P}\right)^n \leq P\exp\left(-\frac{\rho_1}{P}n\right)\\ &= o\left(\frac{1}{\text{poly}(d)}\right).\end{aligned}$$

Next, we will prove the remaining. Let us refer to the standard deviation of the Gaussian noise vector in p-th patch of i-th data as $\sigma_{i,p}$. In other words, for each $i \in [n]$ and $p \in [P]\setminus\{p_i^*\}$,

$$\sigma_{i,p} = \begin{cases}\sigma_{\text{d}} & \text{if } p = \tilde{p}_i,\\ \sigma_{\text{b}} & \text{otherwise.}\end{cases}$$

For each $s, s' \in \{\pm 1\}$ and $k \in [K]$, $\left\langle \boldsymbol{w}_s^{(0)}, \boldsymbol{v}_{s',k} \right\rangle \sim N(0, \sigma_0)$. Hence, we have

$$\mathbb{P}\left[\left|\left\langle \boldsymbol{w}_s^{(0)}, \boldsymbol{v}_{s',k} \right\rangle\right| > \sigma_0 \log d\right] \leq 2\exp\left(-\frac{(\sigma_0 \log d)^2}{2\sigma_0^2}\right) = o\left(\frac{1}{\operatorname{poly}(d)}\right).$$

Let $\{\boldsymbol{u}_l\}_{l \in [d-2K]}$ be an orthonormal basis of the orthogonal complement of $\operatorname{Span}(\{\boldsymbol{v}_{s,k}\}_{s \in \{\pm 1\}, k \in [K]})$. Note that for each $s \in \{\pm 1\}, i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$, we can write $\xi_i^{(p)}$ and ξ as

$$\boldsymbol{w}_s(0) = \sigma_0 \sum_{l \in [d-2K]} \mathbf{z}_{s,l} \boldsymbol{u}_l, \quad \xi_i^{(p)} = \sigma_{i,p} \sum_{l \in [d-2K]} \mathbf{z}_{i,l}^{(p)} \boldsymbol{u}_l, \quad \xi = \sigma \sum_{l \in [d-2K]} \mathbf{z}_l \boldsymbol{u}_l$$

where $\mathbf{z}_{s,l}, \mathbf{z}_{i,l}^{(p)}, \mathbf{z}_l \overset{i.i.d.}{\sim} N(0,1)$. The sub-gaussian norm of standard normal distribution $N(0,1)$ is $\sqrt{\frac{8}{3}}$. Then $\left(\mathbf{z}_{i,l}^{(p)}\right)^2 - 1$'s are mean zero sub-exponential with sub-exponential norm $\frac{8}{3}$ (Lemma 2.7.6 in Vershynin [32]). In addition, $\mathbf{z}_{s,l}\mathbf{z}_{i,l}^{(p)}$'s, $\mathbf{z}_{i,l}^{(p)}\mathbf{z}_{j,l}^{(q)}$'s and $\mathbf{z}_{i,l}^{(p)}\mathbf{z}_l$'s are mean zero sub-exponential with sub-exponential norm less than or equal to $\frac{8}{3}$ (Lemma 2.7.7 in Vershynin [32]). Using Bernstein's inequality (Theorem 2.8.1 in Vershynin [32]), let c be the absolute constant stated therein. We then have the following:

$$\begin{aligned}
1 - \mathbb{P}\left[\frac{1}{2}\sigma_{i,p}^2 d \leq \left\|\xi_i^{(p)}\right\|^2 \leq \frac{3}{2}\sigma_{i,p}^2 d\right] &\leq \mathbb{P}\left[\left|\left\|\xi_i^{(p)}\right\|^2 - \sigma_{i,p}^2(d-2K)\right| \geq \sigma_{i,p}^2 d^{\frac{1}{2}} \log d\right] \\
&= \mathbb{P}\left[\left|\sum_{l \in [d-2K]} \left(\left(\mathbf{z}_{i,l}^{(p)}\right)^2 - 1\right)\right| \geq d^{\frac{1}{2}} \log d\right] \\
&\leq 2\exp\left(-\frac{9cd\log^2 d}{64(d-2K)}\right) \\
&\leq 2\exp\left(-\frac{9c\log^2 d}{64}\right) = o\left(\frac{1}{\operatorname{poly}(d)}\right),
\end{aligned}$$

in addition,

$$\begin{aligned}
\mathbb{P}\left[\left|\left\langle \xi_i^{(p)}, \xi_j^{(q)} \right\rangle\right| \geq \sigma_{i,p}\sigma_{j,q} d^{\frac{1}{2}} \log d\right] &= \mathbb{P}\left[\left|\sum_{l \in [d-2K]} \mathbf{z}_{i,l}^{(p)} \mathbf{z}_{j,l}^{(q)}\right| \geq d^{\frac{1}{2}} \log d\right] \\
&\leq 2\exp\left(-\frac{9cd\log^2 d}{64(d-2K)}\right) \\
&\leq 2\exp\left(-\frac{9c\log^2 d}{64}\right) = o\left(\frac{1}{\operatorname{poly}(d)}\right).
\end{aligned}$$

Similarly, we have

$$\mathbb{P}\left[\left|\left\langle \boldsymbol{w}_s^{(0)}, \xi_i^{(p)} \right\rangle\right| \geq \sigma_0 \sigma_{i,p} d^{\frac{1}{2}} \log d\right] \leq 2\exp\left(-\frac{9c\log^2 d}{64}\right) = o\left(\frac{1}{\operatorname{poly}(d)}\right).$$

Applying the union bound to all events, each of which is at most $\operatorname{poly}(d)$, leads us to our first conclusion.

In addition, for each $s \in \{\pm 1\}, i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$,

$$\mathbb{P}\left[\left|\left\langle \boldsymbol{w}_s^{(0)}, \xi \right\rangle\right| \geq \sigma_0 \sigma d^{\frac{1}{2}} \log d\right] \leq 2\exp\left(-\frac{9c\log^2 d}{64}\right) = o\left(\frac{1}{\mathrm{poly}(d)}\right),$$

and

$$\mathbb{P}\left[\left|\left\langle \xi_i^{(p)}, \xi \right\rangle\right| \geq \sigma_{i,p} \sigma d^{\frac{1}{2}} \log d\right] \leq 2\exp\left(-\frac{9c\log^2 d}{64}\right) = o\left(\frac{1}{\mathrm{poly}(d)}\right).$$

Lastly, the last results holds almost surely since $d > nP$.

Applying the union bound to all events, each of which is at most $\mathrm{poly}(d)$, leads us to our second conclusion. ■

E.2. Feature Noise Decomposition

In our analysis, we use a technique that analyzes the coefficients of linear combinations of feature and noise vectors. A similar technique in a different data and network setting is introduced by Cao et al. [3].

Lemma 10 *If we run one of* ERM*,* Cutout*, and* CutMix *training to update parameters $\boldsymbol{W}^{(t)}$ of a model $f_{\boldsymbol{W}^{(t)}}$, then we can write $\boldsymbol{W}^{(t)} = \{\boldsymbol{w}_1^{(t)}, \boldsymbol{w}_{-1}^{(t)}\}$ as*

$$\begin{aligned}\boldsymbol{w}_s^{(t)} = \boldsymbol{w}_s^{(0)} &+ \sum_{k\in[K]} \gamma_s^{(t)}(s,k)\boldsymbol{v}_{s,k} - \sum_{k\in[K]} \gamma_s^{(t)}(-s,k)\boldsymbol{v}_{-s,k} \\ &+ \sum_{i\in\mathcal{V}_s, p\in[P]\setminus\{p_i^*\}} \rho_s^{(t)}(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2} - \sum_{i\in\mathcal{V}_{-s}, p\in[P]\setminus\{p_i^*\}} \rho_s^{(t)}(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2} \\ &+ \alpha\left(\sum_{i\in\mathcal{F}_s} sy_i\rho_s^{(t)}(i,\tilde{p}_i)\frac{\boldsymbol{v}_{s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2} + \sum_{i\in\mathcal{F}_{-s}} sy_i\rho_s^{(t)}(i,\tilde{p}_i)\frac{\boldsymbol{v}_{-s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2}\right)\end{aligned}$$

where $\mathcal{F}_s$ denotes the set of indices of data with feature noise $\boldsymbol{v}_{s,1}$. Furthermore, if we run one of ERM *and* Cutout*, the coefficients $\gamma_s^{(t)}(s',k)$'s and $\rho_s^{(t)}(i,p)$'s are monotone increasing.*

We provide proof of Lemma 10 for ERM in Appendix F.1, for Cutout in Appendix G.1 and for CutMix in Appendix H.1.

Since Gaussian vectors in a high-dimensional regime are nearly orthogonal, we can use the coefficients to approximate inner products or outputs of neurons. The following lemma quantifies the approximation error.

Lemma 11 *Suppose the event E_{init} occurs and $0 \leq \gamma_s^{(t)}(s',k), \rho_s^{(t)}(i,p) \leq \widetilde{\mathcal{O}}(\beta^{-1})$ for all $s, s' \in \{\pm 1\}, k \in [K], i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$ at iteration t. Then, for each $s \in \{\pm 1\}, k \in [K], i \in [n]$, and $p \in [P] \setminus \{p_i^*\}$, the following holds:*

- $\left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k}\right\rangle - \gamma_s^{(t)}(s,k)\right|, \left|\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) - \gamma_s^{(t)}(s,k)\right| = \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right).$
- $\left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{-s,k}\right\rangle + \gamma_s^{(t)}(-s,k)\right|, \left|\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{-s,k}\right\rangle\right) + \beta\gamma_s^{(t)}(-s,k)\right| = \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right).$

- $\left|\left\langle \boldsymbol{w}_{y_i}^{(t)}, \xi_i^{(p)} \right\rangle - \rho_{y_i}^{(t)}(i,p)\right|, \left|\phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \xi_i^{(p)} \right\rangle\right) - \rho_{y_i}^{(t)}(i,p)\right| = \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right)$,
- $\left|\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \xi_i^{(p)} \right\rangle + \rho_{-y_i}^{(t)}(i,p)\right|, \left|\phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \xi_i^{(p)} \right\rangle\right) + \beta\rho_{-y_i}^{(t)}(i,p)\right| = \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right)$,
- $\left|\phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \boldsymbol{x}_i^{(\tilde{p}_i)} \right\rangle\right) - \rho_{y_i}^{(t)}(i,\tilde{p}_i)\right|, \left|\phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \boldsymbol{x}_i^{(\tilde{p}_i)} \right\rangle\right) + \beta\rho_{-y_i}^{(t)}(i,\tilde{p}_i)\right| = \widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right)$.

Proof [Proof of Lemma 11] Note that from our Assumption 8, the following hold:

1. $\sigma_{\mathrm{b}} d^{\frac{1}{2}} = \omega(1)$.
2. $\sigma_0\sigma_{\mathrm{d}} d^{\frac{1}{2}}, r = o\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right)$,
3. $\alpha n\beta^{-1}\sigma_{\mathrm{d}}^{-2}d^{-1} = o\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right)$,
4. $n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}} = o(\alpha\beta^{-1})$.

For each $s \in \{\pm 1\}, k \in [K] \setminus \{1\}$, we have

$$\left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k} \right\rangle - \gamma_s^{(t)}(s,k)\right| = \left|\left\langle \boldsymbol{w}_s^{(0)}, \boldsymbol{v}_{s,k} \right\rangle\right| = \widetilde{\mathcal{O}}(\sigma_0) = \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right),$$

and

$$\begin{aligned}
&\left|\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k} \right\rangle\right) - \gamma_s^{(t)}(s,k)\right| \\
&\le \left|\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k} \right\rangle\right) - \phi\left(\gamma_s^{(t)}(s,k)\right)\right| + \left|\phi\left(\gamma_s^{(t)}(s,k)\right) - \gamma_s^{(t)}(s,k)\right| \\
&\le \left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k} \right\rangle - \gamma_s^{(t)}(s,k)\right| + \frac{(1-\beta)r}{2} \\
&= \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right).
\end{aligned}$$

Similarly,

$$\left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{-s,k} \right\rangle + \gamma_s^{(t)}(-s,k)\right| = \left|\left\langle \boldsymbol{w}_s^{(0)}, \boldsymbol{v}_{-s,k} \right\rangle\right| = \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right),$$

$$\begin{aligned}
\left|\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{-s,k} \right\rangle\right) + \beta\gamma_s^{(t)}(-s,k)\right| &= \left|\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{-s,k} \right\rangle\right) - \phi\left(-\gamma_s^{(t)}(-s,k)\right)\right| \\
&\le \left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{-s,k} \right\rangle + \gamma_s^{(t)}(-s,k)\right| \\
&= \widetilde{\mathcal{O}}(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}).
\end{aligned}$$

Next, we will consider the case of $\boldsymbol{v}_{1,1}$ and $\boldsymbol{v}_{-1,1}$. For each $s \in \{\pm 1\}$, we have

$$\begin{aligned}
&\left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,1} \right\rangle - \gamma_s^{(t)}(s,1)\right| \\
&\le \left|\left\langle \boldsymbol{w}_s^{(0)}, \boldsymbol{v}_{s,1} \right\rangle\right| + \alpha \sum_{i\in[n]} \rho_s^{(t)}(i,\tilde{p}_i) \left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2} \\
&\le \widetilde{\mathcal{O}}(\sigma_0) + \widetilde{\mathcal{O}}(\alpha n\beta^{-1}\sigma_{\mathrm{d}}^{-2}d^{-1}) \\
&= \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right),
\end{aligned}$$

and

$$\begin{aligned}
&\left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{-s,1}\right\rangle + \gamma_s^{(t)}(-s,1)\right|\\
&\leq \left|\left\langle \boldsymbol{w}_s^{(0)}, \boldsymbol{v}_{-s,1}\right\rangle\right| + \alpha \sum_{i\in[n]} \rho_s^{(t)}(i,\tilde{p}_i)\left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2}\\
&\leq \widetilde{\mathcal{O}}(\sigma_0) + \widetilde{\mathcal{O}}(\alpha n\beta^{-1}\sigma_\mathrm{d}^{-2}d^{-1})\\
&= \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_\mathrm{d}\sigma_\mathrm{b}^{-1}d^{-\frac{1}{2}}\right).
\end{aligned}$$

For each $i \in [n]$, and $p \in [P] \setminus \{p_i^*\}$, we have

$$\begin{aligned}
\left|\left\langle \boldsymbol{w}_{y_i}^{(t)}, \xi_i^{(p)}\right\rangle - \rho_{y_i}^{(t)}(i,p)\right| &\leq \left|\left\langle \boldsymbol{w}_{y_i}^{(0)}, \xi_i^{(p)}\right\rangle\right| + \sum_{\substack{j\in[n],q\in[P]\setminus\{p_i^*\}\\(i,p)\neq(j,q)}} \rho_{y_i}^{(t)}(j,q)\frac{\left|\left\langle \xi_i^{(p)}, \xi_j^{(q)}\right\rangle\right|}{\left\|\xi_j^{(q)}\right\|^2}\\
&\leq \widetilde{\mathcal{O}}\left(\sigma_0\sigma_\mathrm{d}d^{\frac{1}{2}}\right) + nP\widetilde{\mathcal{O}}\left(\beta^{-1}\sigma_\mathrm{d}\sigma_\mathrm{b}^{-1}d^{-\frac{1}{2}}\right)\\
&= \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_\mathrm{d}\sigma_\mathrm{b}^{-1}d^{-\frac{1}{2}}\right),
\end{aligned}$$

and

$$\begin{aligned}
&\left|\phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \xi_i^{(p)}\right\rangle\right) - \rho_{y_i}^{(t)}(i,p)\right|\\
&\leq \left|\phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \xi_i^{(p)}\right\rangle\right) - \phi\left(\rho_{y_i}^{(t)}(i,p)\right)\right| + \left|\phi\left(\rho_{y_i}^{(t)}(i,p)\right) - \rho_{y_i}^{(t)}(i,p)\right|\\
&\leq \left|\left\langle \boldsymbol{w}_{y_i}^{(t)}, \xi_i^{(p)}\right\rangle - \rho_{y_i}^{(t)}(i,p)\right| + \frac{(1-\beta)r}{2}\\
&= \widetilde{\mathcal{O}}(n\beta^{-1}\sigma_\mathrm{d}\sigma_\mathrm{b}^{-1}d^{-\frac{1}{2}}).
\end{aligned}$$

Also, if $i \in \mathcal{F}_s$ for some $s \in \{\pm 1\}$,

$$\begin{aligned}
&\left|\phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \boldsymbol{x}_i^{(\tilde{p}_i)}\right\rangle\right) - \rho_{y_i}^{(t)}(i,\tilde{p}_i)\right|\\
&\leq \left|\phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \xi_i^{(\tilde{p}_i)}\right\rangle\right) - \rho_{y_i}^{(t)}(i,\tilde{p}_i)\right| + \left|\phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \boldsymbol{x}_i^{(\tilde{p}_i)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \xi_i^{(\tilde{p}_i)}\right\rangle\right)\right|\\
&\leq \widetilde{\mathcal{O}}(n\beta^{-1}\sigma_\mathrm{d}\sigma_\mathrm{b}^{-1}d^{-\frac{1}{2}}) + \alpha\left|\left\langle \boldsymbol{w}_{y_i}^{(t)}, \boldsymbol{v}_{s,1}\right\rangle\right|\\
&\leq \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_\mathrm{d}\sigma_\mathrm{b}^{-1}d^{-\frac{1}{2}}\right) + \widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) + \widetilde{\mathcal{O}}(\alpha\sigma_0)\\
&= \widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\left|\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \xi_i^{(p)}\right\rangle + \rho_{-y_i}^{(t)}(i,p)\right| &\leq \left|\left\langle \boldsymbol{w}_{-y_i}^{(0)}, \xi_i^{(p)}\right\rangle\right| + \sum_{\substack{j\in[n],q\in[P]\setminus\{p_i^*\}\\(i,p)\neq(j,q)}} \rho_{-y_i}^{(t)}(j,q)\frac{\left|\left\langle \xi_i^{(p)}, \xi_j^{(q)}\right\rangle\right|}{\left\|\xi_j^{(q)}\right\|^2}\\
&\leq \widetilde{\mathcal{O}}(\sigma_0) + nP\widetilde{\mathcal{O}}\left(\beta^{-1}\sigma_\mathrm{d}\sigma_\mathrm{b}^{-1}d^{-\frac{1}{2}}\right)\\
&= \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_\mathrm{d}\sigma_\mathrm{b}^{-1}d^{-\frac{1}{2}}\right),
\end{aligned}$$

and

$$\begin{aligned}\left|\phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \xi_i^{(p)}\right\rangle\right) + \beta\rho_{-y_i}^{(t)}(i,p)\right| &= \left|\phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \xi_i^{(p)}\right\rangle\right) - \phi\left(-\rho_{-y_i}^{(t)}(i,p)\right)\right| \\ &\le \left|\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \xi_i^{(p)}\right\rangle + \rho_{-y_i}^{(t)}(i,p)\right| \\ &= \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right),\end{aligned}$$

Also, if $i \in \mathcal{F}_s$ for some $s \in \{\pm 1\}$,

$$\begin{aligned}&\left|\phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \boldsymbol{x}_i^{(\tilde{p}_i)}\right\rangle\right) + \beta\rho_{-y_i}^{(t)}(i,\tilde{p}_i)\right| \\ &= \left|\phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \xi_i^{(\tilde{p}_i)}\right\rangle\right) + \beta\rho_{-y_i}^{(t)}(i,\tilde{p}_i)\right| + \left|\phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \boldsymbol{x}_i^{(\tilde{p}_i)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \xi_i^{(\tilde{p}_i)}\right\rangle\right)\right| \\ &\le \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right) + \alpha\left|\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \boldsymbol{v}_{s,1}\right\rangle\right| \\ &\le \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right) + \widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) + \widetilde{\mathcal{O}}(\alpha\sigma_0) \\ &= \widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right).\end{aligned}$$

■

We define the set $\mathcal{W}$ as the collection of $\boldsymbol{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_{-1}\}$, where $\boldsymbol{w}_1 - \boldsymbol{w}_0^{(0)}, \boldsymbol{w}_{-1} - \boldsymbol{w}_0^{(0)}$ are elements of the subspace spanned by $\{\boldsymbol{v}_{s,k}\}_{s\in\{\pm1\},k\in[K]} \cup \left\{\boldsymbol{x}_i^{(p)}\right\}_{i\in[n],p\in[P]\setminus\{p_i^*\}}$. The following lemma guarantees the unique expression of any $\boldsymbol{W} \in \mathcal{W}$ in the form of the feature noise decomposition.

Lemma 12 *Suppose the event E_{init} occurs. Each element $\boldsymbol{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_{-1}\} \in \mathcal{W}$ is uniquely expressed as:*

$$\begin{aligned}\boldsymbol{w}_s = \boldsymbol{w}_s^{(0)} &+ \sum_{k\in[K]} \gamma_s(s,k)\boldsymbol{v}_{s,k} - \sum_{k\in[K]} \gamma_s(-s,k)\boldsymbol{v}_{-s,k} \\ &+ \sum_{\substack{i\in\mathcal{V}_s \\ p\in[P]\setminus\{p_i^*\}}} \rho_s(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2} - \sum_{\substack{i\in\mathcal{V}_{-s} \\ p\in[P]\setminus\{p_i^*\}}} \rho_s(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2} \\ &+ \alpha\left(\sum_{i\in\mathcal{F}_s} sy_i\rho_s(i,\tilde{p}_i)\frac{\boldsymbol{v}_{s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2} + \sum_{i\in\mathcal{F}_{-s}} sy_i\rho_s(i,\tilde{p}_i)\frac{\boldsymbol{v}_{-s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2}\right)\end{aligned}$$

for each $s \in \{\pm 1\}$. Hence, for each $s^* \in \{\pm 1\}$ and $k^* \in [K]$, we can introduce the function $Q^{(s^*,k^*)} : \mathcal{W} \to \mathbb{R}^{d\times 2}$ such that for each $\boldsymbol{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_{-1}\} \in \mathcal{W}$,

$$Q^{(s^*,k^*)}(\boldsymbol{W}) = \left\{Q_1^{(s^*,k^*)}(\boldsymbol{w}_1), Q_{-1}^{(s^*,k^*)}(\boldsymbol{w}_{-1})\right\}$$

is given by:

$$Q_s^{(s^*,k^*)}(\boldsymbol{w}_s) = ss^*\gamma_s(s^*,k^*)\boldsymbol{v}_{s^*,k^*} + ss^* \sum_{i\in\mathcal{V}_{s^*,k^*}, p\in[P]\setminus\{p_i^*\}} \rho_s(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2}$$
$$+ \alpha\left(\sum_{i\in\mathcal{F}_s\cap\mathcal{V}_{s^*,k^*}} ss^*\rho_s(i,\tilde{p}_i)\frac{\boldsymbol{v}_{s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2} + \sum_{i\in\mathcal{F}_{-s}\cap\mathcal{V}_{s^*,k^*}} ss^*\rho_s(i,\tilde{p}_i)\frac{\boldsymbol{v}_{-s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2}\right).$$

Proof From liner independency of $\{\boldsymbol{v}_{s,k}\}_{s\in\{\pm1\},k\in[K]} \cup \left\{\boldsymbol{x}_i^{(p)}\right\}_{i\in[n],p\in[P]\setminus\{p_i^*\}}$, we can express any element $\boldsymbol{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_{-1}\} \in \mathcal{W}$ as

$$\boldsymbol{w}_s = \boldsymbol{w}_s^{(0)} + \sum_{k\in[K]}\tilde{\gamma}_s(s,k)\boldsymbol{v}_{s,k} - \sum_{k\in[K]}\tilde{\gamma}_s(-s,k)\boldsymbol{v}_{-s,k}$$
$$+ \sum_{\substack{i\in\mathcal{V}_s,\\ p\in[P]\setminus\{p_i^*\}}} \rho_s(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|} - \sum_{\substack{i\in\mathcal{V}_{-s},\\ p\in[P]\setminus\{p_i^*\}}} \rho_s(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|} \tag{4}$$

with unique $\{\tilde{\gamma}_s(s,k), \tilde{\gamma}_s(-s,k)\}_{s\in\{\pm1\},k\in[K]}$ and $\{\rho_s(i,p)\}_{s\in\{\pm1\},i\in[n],p\in[P]\setminus\{i^*\}}$. If we define $\gamma_s(s,k)$ and $\gamma_s(-s,k)$ as $\gamma_s(s,k) = \tilde{\gamma}_s(s,k)$, $\gamma_s(-s,k) = \tilde{\gamma}_s(-s,k)$ for $k \neq 1$, and

$$\gamma_s(s,1) = \tilde{\gamma}_s(s,1) - \alpha\sum_{i\in\mathcal{F}_s} sy_i\rho_s(i,\tilde{p}_i)\left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2},$$
$$\gamma_s(-s,1) = \tilde{\gamma}_s(-s,1) + \alpha\sum_{i\in\mathcal{F}_s} sy_i\rho_s(i,\tilde{p}_i)\left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2},$$

then we have

$$\boldsymbol{w}_s = \boldsymbol{w}_s^{(0)} + \sum_{k\in[K]}\gamma_s(s,k)\boldsymbol{v}_{s,k} - \sum_{k\in[K]}\gamma_s(-s,k)\boldsymbol{v}_{-s,k}$$
$$+ \sum_{\substack{i\in\mathcal{V}_s\\ p\in[P]\setminus\{p_i^*\}}} \rho_s(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2} - \sum_{\substack{i\in\mathcal{V}_{-s}\\ p\in[P]\setminus\{p_i^*\}}} \rho_s(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2}$$
$$+ \alpha\left(\sum_{i\in\mathcal{F}_s} sy_i\rho_s(i,\tilde{p}_i)\frac{\boldsymbol{v}_{s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2} + \sum_{i\in\mathcal{F}_{-s}} sy_i\rho_s(i,\tilde{p}_i)\frac{\boldsymbol{v}_{-s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2}\right).$$

To show uniqueness, suppose $\{\hat{\gamma}_s(s,k), \hat{\gamma}_s(-s,k)\}_{s\in\{\pm1\},k\in[K]}$ and $\{\hat{\rho}_s(i,p)\}_{s\in\{\pm1\},i\in[n],p\in[P]\setminus\{i^*\}}$ satisfies

$$\begin{aligned}\boldsymbol{w}_s = \boldsymbol{w}_s^{(0)} &+ \sum_{k\in[K]} \hat{\gamma}_s(s,k)\boldsymbol{v}_{s,k} - \sum_{k\in[K]} \hat{\gamma}_s(-s,k)\boldsymbol{v}_{-s,k} \\ &+ \sum_{\substack{i\in\mathcal{V}_s \\ p\in[P]\setminus\{p_i^*\}}} \hat{\rho}_s(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2} - \sum_{\substack{i\in\mathcal{V}_{-s} \\ p\in[P]\setminus\{p_i^*\}}} \hat{\rho}_s(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2} \\ &+ \alpha\left(\sum_{i\in\mathcal{F}_s} sy_i\hat{\rho}_s(i,\tilde{p}_i)\frac{\boldsymbol{v}_{s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2} + \sum_{i\in\mathcal{F}_{-s}} sy_i\hat{\rho}_s(i,\tilde{p}_i)\frac{\boldsymbol{v}_{-s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2}\right).\end{aligned}$$

We have

$$\begin{aligned}\boldsymbol{w}_s = \boldsymbol{w}_s^{(0)} &+ \sum_{k\in[K]\setminus\{1\}} \hat{\gamma}_s(s,k)\boldsymbol{v}_{s,k} - \sum_{k\in[K]\setminus\{\}} \hat{\gamma}_s(-s,k)\boldsymbol{v}_{-s,k} \\ &+ \left(\hat{\gamma}_s(s,1) + \alpha\sum_{i\in\mathcal{F}_s} sy_i\hat{\rho}_s(i,\tilde{p}_i)\left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2}\right)\boldsymbol{v}_{s,1} \\ &- \left(\hat{\gamma}_s(-s,1) - \alpha\sum_{i\in\mathcal{F}_{-s}} sy_i\hat{\rho}_s(i,\tilde{p}_i)\left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2}\right)\boldsymbol{v}_{-s,1} \\ &+ \sum_{\substack{i\in\mathcal{V}_s \\ p\in[P]\setminus\{p_i^*\}}} \hat{\rho}_s(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2} - \sum_{\substack{i\in\mathcal{V}_{-s} \\ p\in[P]\setminus\{p_i^*\}}} \hat{\rho}_s(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2}.\end{aligned}$$

From the uniqueness of (4), we have

$$\hat{\gamma}_s(s,k) = \tilde{\gamma}_s(s,k) = \gamma_s(s,k), \hat{\gamma}_s(-s,k) = \tilde{\gamma}_s(-s,k) = \gamma_s(-s,k),$$

for each $s\in\{\pm1\}, k\in[K]\setminus\{1\}$, $\hat{\rho}_s(i,p) = \rho_s(i,p)$ for each $i\in[n], p\in[P]\setminus\{p_i^*\}$,

$$\hat{\gamma}_s(s,1) + \alpha\sum_{i\in\mathcal{F}_s} sy_i\hat{\rho}_s(i,\tilde{p}_i)\left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2} = \tilde{\gamma}_s(s,1) = \gamma_s(s,1) + \alpha\sum_{i\in\mathcal{F}_s} sy_i\hat{\rho}_s(i,\tilde{p}_i)\left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2},$$

and

$$\hat{\gamma}_s(-s,1) - \alpha\sum_{i\in\mathcal{F}_{-s}} sy_i\hat{\rho}_s(i,\tilde{p}_i)\left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2} = \tilde{\gamma}_s(-s,1) = \gamma_s(-s,1) - \alpha\sum_{i\in\mathcal{F}_{-s}} sy_i\hat{\rho}_s(i,\tilde{p}_i)\left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2}.$$

Hence, we obtain the uniqueness of the expression and $Q^{(s^*,k^*)}$ is well defined for each $s^*\in\{\pm1\}$ and $k^*\in[K]$. ■

Appendix F. Proof for ERM

In this section, we use $g_i^{(t)} := \frac{1}{1+\exp\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)}$ for each data i and iteration t, for simplicity.

F.1. Proof of Lemma 10 for ERM

For $s \in \{\pm 1\}$ and iterate t,

$$\begin{aligned}
&\boldsymbol{w}_s^{(t+1)} - \boldsymbol{w}_s^{(t)} \\
&= -\eta \nabla_{\boldsymbol{w}_s} \mathcal{L}_{\mathrm{ERM}}\left(\boldsymbol{W}^{(t)}\right) \\
&= \frac{\eta}{n} \sum_{i\in[n]} s y_i g_i^{(t)} \sum_{p\in[P]} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)} \\
&= \frac{\eta}{n}\left(\sum_{i\in\mathcal{V}_s} g_i^{(t)} \sum_{p\in[P]} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)} - \sum_{i\in\mathcal{V}_{-s}} g_i^{(t)} \sum_{p\in[P]} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)} \right),
\end{aligned}$$

and we have

$$\begin{aligned}
&\sum_{i\in\mathcal{V}_s} g_i^{(t)} \sum_{p\in[P]} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)} \\
&= \sum_{k\in[K]} \sum_{i\in\mathcal{V}_{s,k}} g_i^{(t)} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k} \right\rangle\right) \boldsymbol{v}_{s,k} + \sum_{i\in\mathcal{V}_s} g_i^{(t)} \sum_{p\in[P]\setminus\{p_i^*, \tilde{p}_i\}} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \xi_i^{(p)} \right\rangle\right) \xi_i^{(p)} \\
&\quad + \sum_{i\in\mathcal{V}_s\cap\mathcal{F}_s} g_i^{(t)} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha\boldsymbol{v}_{s,1} + \xi_i^{(\tilde{p}_i)} \right\rangle\right) \left(\alpha\boldsymbol{v}_{s,1} + \xi_i^{(\tilde{p}_i)}\right) \\
&\quad + \sum_{i\in\mathcal{V}_s\cap\mathcal{F}_{-s}} g_i^{(t)} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha\boldsymbol{v}_{-s,1} + \xi_i^{(\tilde{p}_i)} \right\rangle\right) \left(\alpha\boldsymbol{v}_{-s,1} + \xi_i^{(\tilde{p}_i)}\right),
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{i\in\mathcal{V}_{-s}} g_i^{(t)} \sum_{p\in[P]} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)} \\
&= \sum_{k\in[K]} \sum_{i\in\mathcal{V}_{-s,k}} g_i^{(t)} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{-s,k} \right\rangle\right) \boldsymbol{v}_{-s,k} + \sum_{i\in\mathcal{V}_{-s}} g_i^{(t)} \sum_{p\in[P]\setminus\{p_i^*, \tilde{p}_i\}} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \xi_i^{(p)} \right\rangle\right) \xi_i^{(p)} \\
&\quad + \sum_{i\in\mathcal{V}_{-s}\cap\mathcal{F}_s} g_i^{(t)} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha\boldsymbol{v}_{s,1} + \xi_i^{(\tilde{p}_i)} \right\rangle\right) \left(\alpha\boldsymbol{v}_{s,1} + \xi_i^{(\tilde{p}_i)}\right) \\
&\quad + \sum_{i\in\mathcal{V}_{-s}\cap\mathcal{F}_{-s}} g_i^{(t)} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha\boldsymbol{v}_{-s,1} + \xi_i^{(\tilde{p}_i)} \right\rangle\right) \left(\alpha\boldsymbol{v}_{-s,1} + \xi_i^{(\tilde{p}_i)}\right).
\end{aligned}$$

Hence, if we define $\gamma_s^{(t)}(s', k)$'s and $\rho_s^{(t)}(i,p)$'s recursively by using the rule

$$\gamma_s^{(t+1)}(s',k) = \gamma_s^{(t)}(s',k) + \frac{\eta}{n} \sum_{i\in\mathcal{V}_{s',k}} g_i^{(t)} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s',k} \right\rangle\right), \tag{5}$$

$$\rho_s^{(t+1)}(i,p) = \rho_s^{(t)}(i,p) + \frac{\eta}{n} g_i^{(t)} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \left\lVert \xi_i^{(p)} \right\rVert^2, \tag{6}$$

starting from $\gamma_s^{(0)}(s',k) = \rho_s^{(0)}(i,p) = 0$ for each $s, s' \in \{\pm 1\}, k \in [K], i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$, then we have

$$\begin{aligned} \boldsymbol{w}_s^{(t)} = \boldsymbol{w}_s^{(0)} &+ \sum_{k\in[K]} \gamma_s^{(t)}(s,k)\boldsymbol{v}_{s,k} - \sum_{k \in [K]} \gamma_s^{(t)}(-s,k)\boldsymbol{v}_{-s,k} \\ &+ \sum_{i \in \mathcal{V}_s, p \in [P]\setminus\{p_i^*\}} \rho_s^{(t)}(i,p) \frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2} - \sum_{i \in \mathcal{V}_{-s}, p\in[P]\setminus\{p_i^*\}} \rho_s^{(t)}(i,p)\frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2} \\ &+ \alpha \left(\sum_{i\in\mathcal{F}_s} s y_i \rho_s^{(t)}(i, \tilde{p}_i) \frac{\boldsymbol{v}_{s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2} + \sum_{i \in \mathcal{F}_{-s}} s y_i \rho_s^{(t)}(i,\tilde{p}_i) \frac{\boldsymbol{v}_{-s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2}\right), \end{aligned}$$

for each $s \in \{\pm 1\}$. Furthermore, $\gamma_s^{(t)}(s',k)$'s and $\rho_s^{(t)}(i,p)$'s are monotone increasing. □

F.2. Proof of Theorem 3

To demonstrate Theorem 3, we present a structured proof comprising the following five steps:

1. Establish upper bounds on $\gamma_s^{(t)}(s',k)$'s and $\rho_s^{(t)}(i,p)$'s to apply Lemma 11 (Section F.2.1).
2. Demonstrate that the model learns common features quickly (Section F.2.2).
3. Show that the model overfits dominant noises in (extremely) rare data instead of learning its feature (Section F.2.3).
4. Confirm the persistence of this tendency until T^* iterates (Section F.2.4).
5. Characterize train accuracy and test accuracy (Section F.2.5).

F.2.1. BOUNDS ON THE COEFFICIENTS IN FEATURE NOISE DECOMPOSITION

The following lemma provides upper bounds on Lemma 10 during T^* iterations.

Lemma 13 *Suppose the event E_{init} occurs. For any $t \in [0, T^*]$, we have*

$$0 \le \gamma_s^{(t)}(s,k) + \beta\gamma_s^{(t)}(-s,k) \le 4\log(\eta T^*), \quad 0 \le \rho_{y_i}^{(t)}(i,p) + \beta \rho_{-y_i}^{(t)}(i,p) \le 4 \log\left(\eta T^*\right),$$

for all $s \in \{\pm1\}, k \in [K], i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$.

Proof [Proof of Lemma 13] We will prove this by using induction on t. The initial case $t = 0$ is trivial. Suppose the given statement holds at $t = T$ and consider the case $t = T + 1$.

Note that from our Assumption 8, the following hold:

- $\eta \le \frac{\log(\eta T^*))}{2}$,
- $\alpha\beta^{-1}, n\beta^{-1}\sigma_\text{d}\sigma_\text{b}^{-1}d^{-\frac{1}{2}} = o(1)$.

Let $\tilde{T}_{s,k} \leq T$ denote the smallest iteration where $\gamma_s^{(\tilde{T}_{s,k}+1)}(s,k) + \beta\gamma_{-s}^{(\tilde{T}_{s,k}+1)}(s,k) > 2\log(\eta T^*)$. We assume the existence of $\tilde{T}_{s,k}$, as its absence would directly lead to our conclusion due to our small choice of η.

By (5), we have

$$\begin{aligned}
&\gamma_s^{(T+1)}(s,k) + \beta\gamma_{-s}^{(T+1)}(s,k)\\
&= \gamma_s^{(\tilde{T}_{s,k})}(s,k) + \beta\gamma_{-s}^{(\tilde{T}_{s,k})}(s,k)\\
&\quad + \sum_{t=\tilde{T}_{s,k}}^{T} \left(\gamma_s^{(t+1)}(s,k) + \beta\gamma_{-s}^{(t+1)}(s,k) - \gamma_s^{(t)}(s,k) - \beta\gamma_{-s}^{(t)}(s,k)\right)\\
&\leq 2\log(\eta T^*) + \log(\eta T^*) + \frac{\eta}{n}\sum_{t=\tilde{T}_{s,k}+1}^{T}\sum_{i\in\mathcal{V}_{s,k}} g_i^{(t)}\left(\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right)\right).
\end{aligned}$$

The inequality is due to $\gamma_s^{(\tilde{T}_{s,k})}(s,k) + \beta\gamma_{-s}^{(\tilde{T}_{s,k})}(s,k) \leq 2\log(\eta T^*)$ and

$$\frac{\eta}{n}\sum_{i\in\mathcal{V}_{s,k}} g_i^{(\tilde{T}_{s,k})}\left(\phi'\left(\left\langle \boldsymbol{w}_s^{(\tilde{T}_{s,k})}, \boldsymbol{v}_{s,k}\right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-s}^{(\tilde{T}_{s,k})}, \boldsymbol{v}_{s,k}\right\rangle\right)\right) \leq 2\eta \leq \log(\eta T^*),$$

from our choice of $\tilde{T}_{s,k}$ and η.

For each $t = \tilde{T}_{s,k}+1, \dots T$, and $i \in \mathcal{V}_{s,k}$, we have

$$\begin{aligned}
&y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\\
&= \phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) + \sum_{p\in[P]\setminus\{p_i^*\}}\left(\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right)\\
&\geq \gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) + \sum_{p\in[P]\setminus\{p_i^*\}}\left(\rho_s^{(t)}(i,p) + \beta\rho_{-s}^{(t)}(i,p)\right)\\
&\quad - 2P\widetilde{\mathcal{O}}(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}) - 2\widetilde{\mathcal{O}}(\alpha\beta^{-1})\\
&\geq \frac{3}{2}\log(\eta T^*)
\end{aligned}$$

The first inequality is due to Lemma 11 and the second inequality holds because from our choice of t, $\gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) \geq 2\log(\eta T^*)$.

Hence, we obtain

$$\begin{aligned}
&\frac{\eta}{n}\sum_{t=\tilde{T}_{s,k}}^{T}\sum_{i\in\mathcal{V}_{s,k}} g_i^{(t)}\left(\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right)\right)\\
&\leq \frac{2\eta}{n}\sum_{t=\tilde{T}_{s,k}}^{T}\sum_{i\in\mathcal{V}_{s,k}} \exp\left(-y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)\\
&\leq \frac{2|\mathcal{V}_{s,k}|}{n}(\eta T^*)\exp\left(-\frac{3}{2}\log(\eta T^*)\right)\\
&\leq \log(\eta T^*).
\end{aligned}$$

Thus, we have $\gamma_s^{(T+1)}(s,k) + \beta\gamma_{-s}^{(T+1)}(s,k) \leq 4\log(\eta T^*)$.

Next, we will follow similar arguments to show that

$$\rho_{y_i}^{(T+1)}(i,p) + \beta\rho_{-y_i}^{(T+1)}(i,p) \leq 4\log(\eta T^*)$$

for each $i \in [n]$ and $p \in [P] \setminus \{\tilde{p}_i\}$.

Let $\tilde{T}_i^{(p)} \leq T$ be the smallest iteration such that $\rho_{y_i}^{(\tilde{T}_i^{(p)}+1)}(i,p) + \beta\rho_{-y_i}^{(\tilde{T}_i^{(p)}+1)}(i,p) > 2\log(\eta T^*)$. We assume the existence of $\tilde{T}_i^{(p)}$, as its absence would directly lead to our conclusion due to our small choice of η.

By (6), we have

$$\begin{aligned}
&\rho_{y_i}^{(T+1)}(i,p) + \beta\rho_{-y_i}^{(T+1)}(i,p)\\
&= \rho_{y_i}^{(\tilde{T}_i^{(p)})}(i,p) + \beta\rho_{-y_i}^{(\tilde{T}_i^{(p)})}(i,p)\\
&\quad + \sum_{t=\tilde{T}_i^{(p)}}^{T} \left(\rho_{y_i}^{(t+1)}(i,p) + \beta\rho_{-y_i}^{(t+1)}(i,p) - \rho_{y_i}^{(t)}(i,p) - \beta\rho_{-y_i}^{(t)}(i,p)\right)\\
&\leq 2\log(\eta T^*) + \log(\eta T^*) + \frac{\eta}{n}\sum_{t=\tilde{T}_i^{(p)}+1}^{T} g_i^{(t)}\left(\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right)
\end{aligned}$$

The inequality is due to $\rho_{y_i}^{(\tilde{T}_i^{(p)})}(i,p) + \beta\rho_{-y_i}^{(\tilde{T}_i^{(p)})}(i,p) \leq 2\log T^*$ and

$$\frac{\eta}{n} g_i^{(\tilde{T}_i^{(p)})}\left[\phi'\left(\left\langle \boldsymbol{w}_s^{(\tilde{T}_i^{(p)})}, \boldsymbol{x}_i^{(p)}\right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-s}^{(\tilde{T}_i^{(p)})}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right] \leq 2\eta \leq \log(\eta T^*),$$

from our choice of $\tilde{T}_i^{(p)}$ and η.

For each $t = \tilde{T}_i^{(p)}+1, \dots, T$, we have

$$\begin{aligned}
&y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\\
&= \phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) + \sum_{q\in[P]\setminus\{p\}}\left(\phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right)\\
&\geq \rho_{y_i}^{(t)}(i,p) + \beta\rho_{-y_i}^{(t)}(i,p) - 2P\widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right) - 2\widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right)\\
&\geq \frac{3}{2}\log(\eta T^*).
\end{aligned}$$

The first inequality is due to Lemma 11 and the second inequality holds because from our choice of t, $\rho_{y_i}^{(t)}(i,p) + \beta\rho_{-y_i}^{(t)}(i,p) \geq 2\log(\eta T^*)$.

Therefore, we have

$$\begin{aligned}&\frac{\eta}{n}\sum_{t=\tilde{T}_i^{(p)}+1}^{T} g_i^{(t)}\left(\phi'\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)+\beta\phi'\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right)\\ &\leq \frac{2\eta}{n}\sum_{t=\tilde{T}_i^{(p)}+1}^{T}\exp\left(-y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right) \leq \frac{2}{n}(\eta T^*)\exp\left(-\frac{3}{2}\log(\eta T^*)\right)\\ &\leq \log(\eta T^*),\end{aligned}$$

and we conclude $\rho_{y_i}^{(T+1)}(i,p)+\beta\rho_{-y_i}^{(T+1)}(i,p)\leq 4\log(\eta T^*)$. ■

F.2.2. LEARNING COMMON FEATURES

In the initial stages of training, the model quickly learns common features while exhibiting minimal overfitting to Gaussian noises.

First, we establish lower bounds on the number of iterations ensuring that noise coefficients $\rho_s^{(t)}(i,p)$ remain small, up to the order of α^2.

Lemma 14 *Suppose the event E_{init} occurs. There exists $\tilde{T} > \frac{2n\alpha^2}{3\eta\sigma_{\mathrm{d}}^2 d}$ such that $\rho_s^{(t)}(i,p)\leq\alpha^2$ for all $0\leq t<\tilde{T}, s\in\{\pm 1\}, i\in[n]$ and $p\in[P]\setminus\{p_i^*\}$.*

Proof [Proof of Lemma 14] Let $\tilde{T}$ be the smallest iteration such that $\rho_s^{(\tilde{T})}(i,p)\geq\alpha^2$ for some $s\in\{\pm1\}, i\in[n]$ and $p\in[P]\setminus\{p_i^*\}$. We assume the existence of $\tilde{T}$, as its absence would directly lead to our conclusion. Then, for any $0\leq t<\tilde{T}$, we have

$$\rho_s^{(t+1)}(i,p)=\rho_s^{(t)}(i,p)+\frac{\eta}{n}g_i^{(t)}\phi'\left(\left\langle \boldsymbol{w}_s^{(t)},\boldsymbol{x}_i^{(p)}\right\rangle\right)\left\|\xi_i^{(p)}\right\|^2\leq\rho_s^{(t)}(i,p)+\frac{3\eta\sigma_{\mathrm{d}}^2 d}{2n},$$

where the inequality is due to $g_i^{(t)}<1$, $\phi'\leq1$, and $\left\|\xi_i^{(p)}\right\|^2\leq\frac{3}{2}\sigma_{\mathrm{d}}^2 d$. Hence, we have

$$\alpha^2\leq\rho_s^{(\tilde{T})}(i,p)<\frac{3\eta\sigma_{\mathrm{d}}^2 d}{2n}\tilde{T},$$

and we conclude $\tilde{T}>\frac{2n\alpha^2}{3\eta\sigma_{\mathrm{d}}^2 d}$ which is the desired result. ■

Next, we will show that the model learns common features in at least constant order within $\tilde{T}$ iterates.

Lemma 15 *Suppose the event E_{init} occurs and $\beta^{-1}\rho_k^{-1}=o\left(\frac{n\alpha^2}{\sigma_{\mathrm{d}}^2 d}\right)$ for some $k\in[K]$. Then, for each $s\in\{\pm1\}$, there exists $T_{s,k}\leq\frac{9n}{\eta\beta|\mathcal{V}_{s,k}|}$ such that $\gamma_s^{(t)}(s,k)+\beta\gamma_{-s}^{(t)}(s,k)\geq1$ for any $t>T_{s,k}$.*

Proof [Proof of Lemma 15] Note that from our Assumption 8, $n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}},\alpha^2,\alpha\beta^{-1}=o(1)$.

Suppose $\gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) < 1$ for all $0 \leq t \leq \frac{2n\alpha^2}{3\eta\sigma_\mathrm{d}^2 d}$. For each $i \in \mathcal{V}_{s,k}$, we have

$$\begin{aligned}
&y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\\
&= \phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) + \sum_{p\in[P]\setminus\{p_i^*\}} \left(\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right)\\
&\leq \gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) + \sum_{p\in[P]\setminus\{p_i^*\}} \left(\rho_s^{(t)}(i,p) + \beta\rho_{-s}^{(t)}(i,p)\right)\\
&\quad + 2P\widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_\mathrm{d}\sigma_\mathrm{b}^{-1}d^{-\frac{1}{2}}\right) + 2\widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right)\\
&\leq 1 + 2P\alpha^2 + 2P\widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_\mathrm{d}\sigma_\mathrm{b}^{-1}d^{-\frac{1}{2}}\right) + 2\widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right)\\
&\leq 2.
\end{aligned}$$

The first inequality is due to Lemma 11 and the second inequality is due to Lemma 14. Thus, $g_i^{(t)} = \frac{1}{1+\exp\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)} > \frac{1}{9}$ and we have

$$\begin{aligned}
&\gamma_s^{(t+1)}(s,k) + \beta\gamma_{-s}^{(t+1)}(s,k)\\
&= \gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) + \frac{\eta}{n}\sum_{i\in\mathcal{V}_{s,k}} g_i^{(t)}\left(\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right)\right)\\
&\geq \gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) + \frac{\eta\beta|\mathcal{V}_{s,k}|}{9n}.
\end{aligned}$$

From the condition in the lemma statement, we have $\frac{9n}{\eta\beta|\mathcal{V}_{s,k}|} = o\left(\frac{2n\alpha^2}{3\eta\sigma_\mathrm{d}^2 d}\right)$. If we choose $t_0 \in \left[\frac{9n}{\eta\beta|\mathcal{V}_{s,k}|}, \frac{2n\alpha^2}{3\eta\sigma_\mathrm{d}^2 d}\right]$, then

$$1 > \gamma_s^{(t_0)}(s,k) + \beta\gamma_{-s}^{(t_0)}(s,k) \geq \frac{\eta\beta|\mathcal{V}_{s,k}|}{9n} t_0 \geq 1,$$

and this is contradictory. Hence, there exists $0 \leq T_{s,k} < \frac{2n\alpha^2}{3\eta\sigma_\mathrm{d}^2 d}$ such that $\gamma_s^{(T_{s,k}+1)}(s,k) + \beta\gamma_{-s}^{(T_{s,k}+1)}(s,k) \geq 1$ and choose the smallest one. Then we obtain

$$1 > \gamma_s^{(T_{s,k})}(s,k) + \beta\gamma_{-s}^{(T_{s,k})}(s,k) \geq \frac{\eta\beta|\mathcal{V}_{s,k}|}{9n} T_{s,k}.$$

Therefore, $T_{s,k} < \frac{9n}{\eta\beta|\mathcal{V}_{s,k}|}$ and this is what we desired. ■

What We Have So Far. For any common feature $\boldsymbol{v}_{s,k}$ with $s \in \{\pm 1\}$ and $k \in \mathcal{K}_C$, it satisfies $\beta^{-1}\rho_k^{-1} = o\left(\frac{n\alpha^2}{\sigma_\mathrm{d}^2 d}\right)$. By Lemma 15, at any iterate $t \in \left[\bar{T}_1, T^*\right]$ with $\bar{T}_1 := \max_{s\in\{\pm1\},k\in\mathcal{K}_C} T_{s,k}$, the following properties hold if the event E_init occurs:

- (Learn common features): For any $s \in \{\pm 1\}$ and $k \in \mathcal{K}_C$,

$$\gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) = \Omega(1),$$

- For any $s \in \{\pm 1\}, i \in [n]$, and $p \in [P] \setminus \{p_i^*\}$, $\rho_s^{(t)}(i,p) = \widetilde{\mathcal{O}}\left(\beta^{-1}\right)$.

F.2.3. OVERFITTING (EXTREMELY) RARE DATA

In the previous step, we have shown that common data can be well-classified by learning common features. In this step, we will show that the model correctly classifies (extremely) rare data by overfitting dominant noise instead of learning its features.

We first introduce lower bounds on the number of iterates such that feature coefficients $\gamma_s^{(t)}(s', k)$ remain small, up to the order of α^2. This lemma holds to any kind of features, but we will focus on (extremely) rare features. This does not contradict the results from Section F.2.2 for common features since the upper bound on the number of iterations in Lemma 15 is larger than the lower bound on the number of iterations in this lemma.

Lemma 16 *Suppose the event E_{init} occurs. For each $s \in \{\pm 1\}$ and $k \in [K]$, there exists $\tilde{T}_{s,k} > \frac{n\alpha^2}{\eta|\mathcal{V}_{s,k}|}$ such that $\gamma_{s'}^{(t)}(s,k) \le \alpha^2$ for any $0 \le t < \tilde{T}_{s,k}$ and $s' \in \{\pm 1\}$.*

Proof [Proof of Lemma 16] Let $\tilde{T}_{s,k}$ be the smallest iterate such that $\gamma_{s'}^{(\tilde{T}_{s,k})}(s,k) > \alpha^2$ for some $s' \in \{\pm 1\}$. We assume the existence of $\tilde{T}_{s,k}$, as its absence would directly lead to our conclusion.

For any $0 \le t < \tilde{T}_{s,k}$,

$$\gamma_{s'}^{(t+1)}(s,k) = \gamma_{s'}^{(t)}(s,k) + \frac{\eta}{n} \sum_{i \in \mathcal{V}_{s,k}} g_i^{(t)} \phi' \left(\left\langle \boldsymbol{w}_{s'}^{(t)}, \boldsymbol{v}_{s,k} \right\rangle \right) \le \gamma_{s'}^{(t)}(s,k) + \frac{\eta |\mathcal{V}_{s,k}|}{n},$$

and we have $\alpha^2 < \gamma_{s'}^{(\tilde{T}_{s,k})} \le \frac{\eta|\mathcal{V}_{s,k}|}{n} \tilde{T}_{s,k}$. We conclude $\tilde{T}_{s,k} > \frac{n\alpha^2}{\eta|\mathcal{V}_{s,k}|}$ which is the desired result. ■

Next, we will show that the model overfits (extremely) rare data by memorizing dominant noise patches in at least constant order within $\tilde{T}_{s,k}$ iterates.

Lemma 17 *Suppose the event E_{init} occurs and $\frac{n}{\beta\sigma_{\mathrm{d}}^2 d} = o\left(\alpha^2 \rho_k^{-1}\right)$. Then, for each $i \in \mathcal{V}_{s,k}$, there exists $T_i \le \frac{18n}{\eta P \beta \sigma_{\mathrm{d}}^2 d}$ such that*

$$\rho_s^{(t)}(i, \tilde{p}_i) + \beta \rho_{-s}^{(t)}(i, \tilde{p}_i) \ge \frac{1}{P},$$

for any $t > T_i$.

Proof [Proof of Lemma 17] Note that from our Assumption 8, the following hold:

- $\sigma_{\mathrm{b}}^2 = o\left(\sigma_{\mathrm{d}}^2 \beta\right)$,
- $\alpha^2, n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}, \alpha\beta^{-1} = o(1)$.

Suppose $\rho_s^{(t)}(i, \tilde{p}_i) + \beta\rho_{-s}^{(t)}(i, \tilde{p}_i) < \frac{1}{P}$ for all $0 \le t \le \frac{n\alpha^2}{\eta|\mathcal{V}_{s,k}|}$.

For any $p \in [P] \setminus \{p_i^*\}$, by (6) we have

$$\begin{aligned}
\rho_{s'}^{(t+1)}(i,p) - \rho_{s'}^{(t)}(i,p) &= \frac{\eta}{n} g_i^{(t)} \phi' \left(\left\langle \boldsymbol{w}_{s'}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right) \left\| \xi_i^{(p)} \right\|^2 \\
&\le \frac{\eta}{n} g_i^{(t)} \phi' \left(\left\langle \boldsymbol{w}_{s'}^{(t)}, \boldsymbol{x}_i^{(\tilde{p}_i)} \right\rangle \right) \left\| \xi_i^{(\tilde{p}_i)} \right\|^2 \\
&= \rho_s^{(t+1)}(i, \tilde{p}_i) - \rho_s^{(t)}(i, \tilde{p}_i)
\end{aligned}$$

for each $s' \in \{\pm 1\}$, where the inequality holds since $\sigma_{\mathrm{b}}^2 d = o(\beta \sigma_{\mathrm{d}}^2 d)$. Thus,

$$\rho_s^{(t)}(i,p) + \beta \rho_{-s}^{(t)}(i,p) < \frac{1}{P},$$

for all $0 \leq t \leq \frac{n\alpha^2}{\eta |\mathcal{V}_{s,k}|}$. Therefore, we have

$$\begin{aligned}
&y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i) \\
&= \phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) + \sum_{p \in [P]\setminus\{p_i^*\}} \left(\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right) \\
&\leq \gamma_s^{(t)}(s,k) + \beta \gamma_s^{(t)}(s,k) + \sum_{p \in [P]\setminus\{p_i^*\}} \left(\rho_s^{(t)}(i,p) + \beta \rho_{-s}^{(t)}(i,p)\right) \\
&\quad + 2P\widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right) + 2\widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) \\
&\leq (1+\beta)\alpha^2 + P \cdot \frac{1}{P} + 2P\widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right) + 2\widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) \\
&\leq 2,
\end{aligned}$$

and $g_i^{(t)} = \frac{1}{1+\exp\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)} \geq \frac{1}{9}$. Also,

$$\begin{aligned}
&\rho_s^{(t+1)}(i,\tilde{p}_i) + \beta \rho_{-s}^{(t+1)}(i,\tilde{p}_i) \\
&= \rho_s^{(t)}(i,\tilde{p}_i) + \beta \rho_{-s}^{(t)}(i,\tilde{p}_i) + \frac{\eta}{n} g_i^{(t)} \left(\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(\tilde{p}_i)}\right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(\tilde{p}_i)}\right\rangle\right)\right) \left\|\xi_i^{(\tilde{p}_i)}\right\|^2 \\
&\geq \rho_s^{(t)}(i,\tilde{p}_i) + \beta \rho_{-s}^{(t)}(i,\tilde{p}_i) + \frac{\eta\beta\sigma_{\mathrm{d}}^2 d}{18n},
\end{aligned}$$

where the last inequality is due to $\left\|\xi_i^{(\tilde{p}_i)}\right\|^2 \geq \frac{1}{2}\sigma_{\mathrm{d}}^2 d$ and $\phi' \geq \beta$.

From the given condition in the lemma statement, we have $\frac{18n}{\eta\beta\sigma_{\mathrm{d}}^2 d} = o\left(\frac{n\alpha^2}{\eta|\mathcal{V}_{s,k}|}\right)$. If we choose $t_0 \in \left[\frac{18n}{\eta P \beta \sigma_{\mathrm{d}}^2 d}, \frac{n\alpha^2}{\eta|\mathcal{V}_{s,k}|}\right]$, then we have

$$\frac{1}{P} > \rho_s^{(t)}(i,\tilde{p}_i) + \beta \rho_{-s}^{(t)}(i,\tilde{p}_i) \geq \frac{\eta\beta\sigma_{\mathrm{d}}^2 d}{18n} t_0 \geq \frac{1}{P}.$$

This is a contradiction and thus there exists $0 \leq T_i < \frac{n\alpha^2}{\eta|\mathcal{V}_{s,k}|}$ such that $\rho_s^{(T_i+1)}(i,\tilde{p}_i) + \beta\rho_{-s}^{(T_i+1)}(i,\tilde{p}_i) \geq \frac{1}{P}$ and let us choose the smallest one.

For any $0 \leq t < T_i$,

$$\frac{1}{P} \geq \rho_s^{(T_i)}(i,\tilde{p}_i) + \beta \rho_{-s}^{(T_i)}(i,\tilde{p}_i) \geq \frac{\eta\beta\sigma_{\mathrm{d}}^2 d}{18n} T_i,$$

and we conclude that $T_i \leq \frac{18n}{\eta P \beta \sigma_{\mathrm{d}}^2 d}$ and this is what we desired. ■

What We Have So Far. For any $s \in \{\pm 1\}$ and $k \in \mathcal{K}_R \cup \mathcal{K}_E$, it satisfies $\frac{n}{\beta \sigma_{\mathrm{d}}^2 d} = o\left(\alpha^2 \rho_k^{-1}\right)$. By Lemma 17 at iterate $t \in [T_{\mathrm{ERM}}, T^*]$ with

$$T_{\mathrm{ERM}} := \max_{\substack{s \in \{\pm 1\} \\ k \in \mathcal{K}_R \cup \mathcal{K}_E}} \max_{i \in \mathcal{V}_{s,k}} T_i \quad \in \left[\bar{T}_1, T^*\right]$$

the following properties hold if the event E_{init} occurs:

- (Learn common features): For $s \in \{\pm 1\}$ and $k \in \mathcal{K}_C$,
$$\gamma_s^{(t)}(s,k) + \beta \gamma_{-s}^{(t)}(s,k) = \Omega(1),$$
- (Overfit (extremely) rare data): For any $s \in \{\pm 1\}$, $k \in \mathcal{K}_R \cup \mathcal{K}_E$, and $i \in \mathcal{V}_{s,k}$,
$$\rho_s^{(t)}(i, \tilde{p}_i) + \beta \rho_{-s}^{(t)}(i, \tilde{p}_i) = \Omega(1),$$
- (Do not learn (extremely) rare features): For any $s, s' \in \{\pm 1\}$ and $k \in \mathcal{K}_R \cup \mathcal{K}_E$, $\gamma_{s'}^{(T_{\mathrm{ERM}})}(s,k) \leq \alpha^2$.
- For any $s \in \{\pm 1\}, i \in [n]$, and $p \in [P] \setminus \{p_i^*\}$, $\rho_s^{(t)}(i,p) = \widetilde{\mathcal{O}}\left(\beta^{-1}\right)$.

F.2.4. ERM CANNOT LEARN (EXTREMELY) RARE FEATURES WITHIN POLYNOMIAL TIMES

In this step, we will show that ERM cannot learn (extremely) rare features within the maximum admissible iterations $T^* = \frac{\mathrm{poly}(d)}{\eta}$.

From now on, we fix any $s^* \in \{\pm 1\}$ and $k^* \in \mathcal{K}_R \cup \mathcal{K}_E$. Recall the function $Q^{(s^*,k^*)} : \mathcal{W} \to \mathbb{R}^{d \times 2}$, defined in Lemma 12 and omit supscripts for simplicity. For each iteration t, $Q(\boldsymbol{W}^{(t)})$ represents quantities updates by data with feature vector $\boldsymbol{v}_{s^*,k^*}$ until t-th iteration. We will sequentially introduce several technical lemmas and by combining these lemmas, quantify update by data with feature vector $\boldsymbol{v}_{s^*,k^*}$ after T_{ERM} and derive our conclusion.

Let us define $\boldsymbol{W}^* = \{\boldsymbol{w}_1^*, \boldsymbol{w}_{-1}^*\}$, where

$$\boldsymbol{w}_1^* = \boldsymbol{w}_1^{(T_{\mathrm{ERM}})} + s^* M \sum_{i \in \mathcal{V}_{s^*,k^*}} \frac{\xi_i^{(\tilde{p}_i)}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2}, \quad \boldsymbol{w}_{-1}^* = \boldsymbol{w}_{-1}^{(T_{\mathrm{ERM}})} - s^* M \sum_{i \in \mathcal{V}_{s^*,k^*}} \frac{\xi_i^{(\tilde{p}_i)}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2},$$

where $M = \beta^{-1} \log\left(\frac{2\eta T^*}{\alpha^2}\right) = \widetilde{\mathcal{O}}\left(\beta^{-1}\right)$.

Note that $\boldsymbol{W}^{(t)}, \boldsymbol{W}^* \in \mathcal{W}$ for any $t \geq 0$.

Lemma 18 *Suppose the event E_{init} occurs. Then,*

$$\left\|Q\left(\boldsymbol{W}^{(T_{\mathrm{ERM}})}\right) - Q(\boldsymbol{W}^*)\right\|^2 \leq 24 M^2 |\mathcal{V}_{s^*,k^*}| \sigma_{\mathrm{d}}^{-2} d^{-1}.$$

Proof [Proof of Lemma 18] For each $s \in \{\pm 1\}$,

$$\begin{aligned}
&Q_s\left(\boldsymbol{w}_s^*\right) - Q_s\left(\boldsymbol{w}_s^{(T_{\mathrm{ERM}})}\right) \\
&= M s s^* \sum_{i \in \mathcal{V}_{s^*,k^*}} \frac{\xi_i^{(\tilde{p}_i)}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2} + \alpha M s^* \left(\sum_{i \in \mathcal{F}_s \cap \mathcal{V}_{s^*,k^*}} \frac{\boldsymbol{v}_{s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2} + \sum_{i \in \mathcal{F}_{-s} \cap \mathcal{V}_{s^*,k^*}} \frac{\boldsymbol{v}_{-s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2} \right),
\end{aligned}$$

and we have

$$
\begin{aligned}
&\left\| Q\left(\boldsymbol{W}^{(T_{\mathrm{ERM}})}\right) - Q(\boldsymbol{W}^*) \right\|^2 \\
&\leq 2M^2 \left(\sum_{i \in \mathcal{V}_{s^*,k^*}} \left\| \xi_i^{(\tilde{p}_i)} \right\|^{-2} + \sum_{i,j \in \mathcal{V}_{s^*,k^*}, i \neq j} \frac{\left| \left\langle \xi_i^{(\tilde{p}_i)}, \xi_j^{(\tilde{p}_j)} \right\rangle \right|}{\left\| \xi_i^{(\tilde{p}_i)} \right\|^2 \left\| \xi_j^{(\tilde{p}_j)} \right\|^2} \right) \\
&\quad + 2M^2 \left(\alpha^2 \left(\sum_{i \in \mathcal{F}_s \cap \mathcal{V}_{s^*,k^*}} \left\| \xi_i^{(\tilde{p}_i)} \right\|^{-2} \right)^2 + \alpha^2 \left(\sum_{i \in \mathcal{F}_{-s'} \cap \mathcal{V}_{s^*,k^*}} \left\| \xi_i^{(\tilde{p}_i)} \right\|^{-2} \right)^2 \right).
\end{aligned}
$$

From E_{init} and $nd^{-\frac{1}{2}} = o(1)$, we have

$$
\sum_{i,j \in \mathcal{V}_{s^*,k^*}, i \neq j} \frac{\left| \left\langle \xi_i^{(\tilde{p}_i)}, \xi_j^{(\tilde{p}_j)} \right\rangle \right|}{\left\| \xi_i^{(\tilde{p}_i)} \right\|^2 \left\| \xi_j^{(\tilde{p}_j)} \right\|^2} \leq \sum_{i \in \mathcal{V}_{s^*,k^*}} \left\| \xi_i^{(\tilde{p}_i)} \right\|^{-2}
$$

In addition, $\rho_k^* n = o(\sigma_{\mathrm{d}}^2 d)$ and $\alpha = o(1)$, we have

$$
\begin{aligned}
&\alpha^2 \left(\sum_{i \in \mathcal{F}_s \cap \mathcal{V}_{s^*,k^*}} \left\| \xi_i^{(\tilde{p}_i)} \right\|^{-2} \right)^2 + \alpha^2 \left(\sum_{i \in \mathcal{F}_{-s'} \cap \mathcal{V}_{s^*,k^*}} \left\| \xi_i^{(\tilde{p}_i)} \right\|^{-2} \right)^2 \\
&\leq \left(\sum_{i \in \mathcal{F}_s \cap \mathcal{V}_{s^*,k^*}} \left\| \xi_i^{(\tilde{p}_i)} \right\|^{-2} \right)^2 + \left(\sum_{i \in \mathcal{F}_{-s'} \cap \mathcal{V}_{s^*,k^*}} \left\| \xi_i^{(\tilde{p}_i)} \right\|^{-2} \right)^2 \\
&\leq \sum_{i \in \mathcal{F}_s \cap \mathcal{V}_{s^*,k^*}} \left\| \xi_i^{(\tilde{p}_i)} \right\|^{-2} + \sum_{i \in \mathcal{F}_{-s'} \cap \mathcal{V}_{s^*,k^*}} \left\| \xi_i^{(\tilde{p}_i)} \right\|^{-2} \\
&= \sum_{i \in \mathcal{V}_{s^*,k^*}} \left\| \xi_i^{(\tilde{p}_i} \right\|^{-2}.
\end{aligned}
$$

Hence, from E_{init}, we obtain

$$
\left\| Q\left(\boldsymbol{W}^{(T_{\mathrm{ERM}})}\right) - Q(\boldsymbol{W}^*) \right\|^2 \leq 6M \sum_{i \in \mathcal{V}_{s^*,k^*}} \left\| \xi_i^{(\tilde{p}_i} \right\|^{-2} \leq 24M^2 |\mathcal{V}_{s^*,k^*}| \sigma_{\mathrm{d}}^{-2} d^{-1}.
$$

■

Lemma 19 *Suppose the E_{init} occurs. For any $t \geq T_{\mathrm{ERM}}$ and $i \in \mathcal{V}_{s^*,k^*}$, it holds that*

$$
\langle y_i \nabla_{\boldsymbol{W}} f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i), Q(\boldsymbol{W}^*) \rangle \geq \frac{M\beta}{2}.
$$

Proof [Proof of Lemma 19] We have

$$
\begin{aligned}
&\langle y_i \nabla_{\boldsymbol{W}} f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i), Q(\boldsymbol{W}^*) \rangle \\
&= \sum_{p \in [P]} \left(\phi'\left(\left\langle \boldsymbol{w}_{s^*}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right) \left\langle Q_{s^*}(\boldsymbol{w}_{s^*}^*), \boldsymbol{x}_i^{(p)} \right\rangle - \phi'\left(\left\langle \boldsymbol{w}_{-s^*}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right) \left\langle Q_{-s^*}(\boldsymbol{w}_{-s^*}^*), \boldsymbol{x}_i^{(p)} \right\rangle \right).
\end{aligned}
$$

Note that $n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}} = o(1)$ by Assumption 8. For any $s \in \{\pm 1\}$ and $p \in [P] \setminus \{p_i^*, \tilde{p}_i\}$,

$$\begin{aligned}
&ss^* \left\langle Q_s(\boldsymbol{w}_s^*), \xi_i^{(p)} \right\rangle \\
&= \rho_s^{(T_{\mathrm{ERM}})}(i,p) + \sum_{\substack{j \in \mathcal{V}_{s^*,k^*}, q \in [P]\setminus\{p_j^*\} \\ (i,p)\neq(j,q)}} \rho_s^{(T_{\mathrm{ERM}})}(j,q) \frac{\left\langle \xi_i^{(p)}, \xi_j^{(q)} \right\rangle}{\left\| \xi_j^{(q)} \right\|^2} + \sum_{j \in \mathcal{V}_{s^*,k^*}\setminus\{i\}} M \frac{\left\langle \xi_i^{(p)}, \xi_j^{(\tilde{p}_j)} \right\rangle}{\left\| \xi_j^{(\tilde{p}_j)} \right\|^2} \\
&\geq -\widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}} \right).
\end{aligned}$$

Also, for any $s \in \{\pm 1\}$, $ss^* \langle Q_s(\boldsymbol{w}_s^*), \boldsymbol{v}_{s^*,k^*} \rangle = \gamma_s^{(T_{\mathrm{ERM}})}(s^*,k^*) \geq 0$. In addition,

$$\begin{aligned}
&ss^* \left\langle Q_s(\boldsymbol{w}_s^*), \boldsymbol{x}_i^{(\tilde{p}_i)} \right\rangle \\
&\geq ss^* \left\langle Q_s(\boldsymbol{w}_s^*), \xi_i^{(\tilde{p}_i)} \right\rangle - \widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\rho_{k^*} n \sigma_{\mathrm{d}}^{-2} d^{-1} \right) \\
&= M + \rho_s^{(T_{\mathrm{ERM}})}(i,\tilde{p}_i) + \sum_{\substack{j \in \mathcal{V}_{s^*,k^*}, q \in [P]\setminus\{p_i^*\} \\ (i,\tilde{p}_i)\neq(j,q)}} \rho_s^{(T_{\mathrm{ERM}})}(j,q) \frac{\left\langle \xi_i^{(\tilde{p}_i)}, \xi_j^{(q)} \right\rangle}{\left\| \xi_j^{(q)} \right\|^2} \\
&\quad + \sum_{j \in \mathcal{V}_{s^*,k^*}\setminus\{i\}} M \frac{\left\langle \xi_i^{(\tilde{p}_i)}, \xi_j^{(\tilde{p}_j)} \right\rangle}{\left\| \xi_j^{(\tilde{p}_j)} \right\|^2} - \widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) \\
&\geq M - \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}} \right) - \widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) \\
&\geq \frac{M}{2}.
\end{aligned}$$

Hence, combining with $\phi' \geq \beta$, we have $\langle y_i \nabla_{\boldsymbol{W}} f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i), Q(\boldsymbol{W}^*) \rangle \geq \frac{M\beta}{2}$. ■

By combining Lemma 18 and Lemma 19, we can obtain the following result.

Lemma 20 *Suppose the event E_{init} occurs.*

$$\frac{\eta}{n} \sum_{t=T_{\mathrm{ERM}}}^{T^*} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right) \leq \left\| Q\left(\boldsymbol{W}^{(T_{\mathrm{ERM}})}\right) - Q(\boldsymbol{W}^*) \right\|^2 + 2\eta T^* e^{-M\beta}$$

Proof [Proof of Lemma 20] Note that for any $T_{\mathrm{ERM}} \leq t < T^*$,

$$Q\left(\boldsymbol{W}^{(t+1)}\right) = Q\left(\boldsymbol{W}^{(t)}\right) - \frac{\eta}{n} \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right),$$

and thus

$$
\begin{aligned}
&\left\| Q\left(\boldsymbol{W}^{(t)}\right) - Q\left(\boldsymbol{W}^*\right)\right\|^2 - \left\| Q\left(\boldsymbol{W}^{(t+1)}\right) - Q\left(\boldsymbol{W}^*\right)\right\|^2 \\
&= \frac{2\eta}{n}\left\langle \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right), Q\left(\boldsymbol{W}^{(t)}\right) - Q\left(\boldsymbol{W}^*\right)\right\rangle - \frac{\eta^2}{n^2}\left\| \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)\right\|^2 \\
&= \frac{2\eta}{n}\left\langle \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)), Q\left(\boldsymbol{W}^{(t)}\right)\right\rangle \\
&\quad - \frac{2\eta}{n} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell'(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)) \left\langle \nabla_{\boldsymbol{W}} y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i), Q\left(\boldsymbol{W}^*\right)\right\rangle - \frac{\eta^2}{n^2}\left\| \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)\right\|^2 \\
&\geq \frac{2\eta}{n}\left\langle \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)), Q\left(\boldsymbol{W}^{(t)}\right)\right\rangle \\
&\quad - \frac{M\beta\eta}{n} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell'(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)) - \frac{\eta^2}{n^2}\left\| \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)\right\|^2,
\end{aligned}
$$

where the last inequality is due to Lemma 19. By the chain rule, we have

$$
\begin{aligned}
&\left\langle \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)), Q\left(\boldsymbol{W}^{(t)}\right)\right\rangle \\
&= \sum_{i \in \mathcal{V}_{s^*,k^*}} \Bigg[\ell'(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)) \\
&\quad \times \sum_{p \in [P]} \left(\phi'\left(\left\langle \boldsymbol{w}_{s^*}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) \left\langle Q_{s^*}\left(\boldsymbol{w}_{s^*}^{(t)}\right), \boldsymbol{x}_i^{(p)}\right\rangle - \phi'\left(\left\langle \boldsymbol{w}_{-s^*}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) \left\langle Q_{-s^*}\left(\boldsymbol{w}_{-s^*}^{(t)}\right), \boldsymbol{x}_i^{(p)}\right\rangle \right) \Bigg].
\end{aligned}
$$

For each $s \in \{\pm 1\}$ and $i \in \mathcal{V}_{s^*,k^*}$,

$$
\begin{aligned}
&\left| \left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle - \left\langle Q_s\left(\boldsymbol{w}_s^{(t)}\right), \boldsymbol{x}_i^{(p)}\right\rangle \right| \\
&= \left| \left\langle \boldsymbol{w}_s^{(t)} - Q_s\left(\boldsymbol{w}_s^{(t)}\right), \boldsymbol{x}_i^{(p)}\right\rangle \right| \\
&\leq \sum_{j \in [n] \setminus \mathcal{V}_{s^*,k^*}, q \in [P] \setminus \{p_i^*\}} \left| \left\langle \rho_s^{(t)}(j,q) \frac{\xi_j^{(q)}}{\left\| \xi_j^{(q)} \right\|^2}, \boldsymbol{x}_i^{(p)} \right\rangle \right| \\
&\quad + \alpha \sum_{j \in [n] \setminus \mathcal{V}_{s^*,k^*}} \rho_s^{(t)}(j, \tilde{p}_j) \left\| \xi_j^{(\tilde{p}_j)} \right\|^{-2} \left| \left\langle \boldsymbol{v}_{s',1}, \boldsymbol{x}_i^{(p)} \right\rangle \right| \\
&\leq \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right) + \widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\rho_{k^*} n \sigma_{\mathrm{d}}^{-2} d^{-1} \right) \\
&= \widetilde{\mathcal{O}}(\alpha\beta^{-1}),
\end{aligned}
$$

where the last inequality is due to Lemma 13 and the event E_{init}. By Lemma 34,

$$\begin{aligned}
&\sum_{p\in[P]}\left(\phi'\left(\left\langle \boldsymbol{w}_{s^*}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\left\langle Q_{s^*}\left(\boldsymbol{w}_{s^*}^{(t)}\right), \boldsymbol{x}_i^{(p)}\right\rangle - \phi'\left(\left\langle \boldsymbol{w}_{-s^*}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\left\langle Q_{-s^*}\left(\boldsymbol{w}_{s^*}^{(t)}\right), \boldsymbol{x}_i^{(p)}\right\rangle\right)\\
&\geq \sum_{p\in[P]}\left(\phi\left(\left\langle \boldsymbol{w}_{s^*}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s^*}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right) - rP - \widetilde{\mathcal{O}}(\alpha\beta^{-1})\\
&= y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i) - \widetilde{\mathcal{O}}(\alpha\beta^{-1})
\end{aligned}$$

where the last equality is due to $r = o(\alpha\beta^{-1})$. Therefore, we have

$$\begin{aligned}
&\left\|Q\left(\boldsymbol{W}^{(t)}\right) - Q\left(\boldsymbol{W}^*\right)\right\|^2 - \left\|Q\left(\boldsymbol{W}^{(t+1)}\right) - Q(\boldsymbol{W}^*)\right\|^2\\
&\geq \frac{2\eta}{n}\sum_{i\in\mathcal{V}_{s^*,k^*}} \ell'\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i) - \widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) - \frac{M\beta}{2}\right)\\
&\quad - \frac{\eta^2}{n^2}\left\|\nabla_{\boldsymbol{W}} \sum_{i\in\mathcal{V}_{s^*,k^*}} \ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i))\right\|^2\\
&\geq \frac{2\eta}{n}\sum_{i\in\mathcal{V}_{s^*,k^*}} \ell'(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i))\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i) - M\beta\right)\\
&\quad - \frac{\eta^2}{n^2}\left\|\nabla_{\boldsymbol{W}} \sum_{i\in\mathcal{V}_{s^*,k^*}} \ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i))\right\|^2.
\end{aligned}$$

From the convexity of $\ell(\cdot)$,

$$\begin{aligned}
\sum_{i\in\mathcal{V}_{s^*,k^*}} \ell'(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i))\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i) - M\beta\right) &\geq \sum_{i\in\mathcal{V}_{s^*,k^*}} \left(\ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)) - \ell\left(M\beta\right)\right)\\
&\geq \sum_{i\in\mathcal{V}_{s^*,k^*}} \ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)) - ne^{-M\beta}.
\end{aligned}$$

In addition, by Lemma 35,

$$\begin{aligned}
\frac{\eta^2}{n^2}\left\|\nabla \sum_{i\in\mathcal{V}_{s^*,k^*}} \ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)\right\|^2 &\leq \frac{8\eta^2 P^2 \sigma_{\mathrm{d}}^2 d |\mathcal{V}_{s^*,k^*}|}{n^2} \sum_{i\in\mathcal{V}_{s^*,k^*}} \ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i))\\
&\leq \frac{\eta}{n}\sum_{i\in\mathcal{V}_{s^*,k^*}} \ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)),
\end{aligned}$$

where the last inequality is due to A8, and we have

$$\begin{aligned}
&\left\|Q\left(\boldsymbol{W}^{(t)}\right) - Q(\boldsymbol{W}^*)\right\|^2 - \left\|Q\left(\boldsymbol{W}^{(t+1)}\right) - Q(\boldsymbol{W}^*)\right\|^2\\
&\geq \frac{\eta}{n}\sum_{i\in\mathcal{V}_{s^*,k^*}} \ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)) - 2\eta e^{-M\beta}.
\end{aligned}$$

From telescoping summation, we have

$$\frac{\eta}{n}\sum_{t=T_{\text{ERM}}}^{T^*}\sum_{i\in\mathcal{V}_{s^*,k^*}}\ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)\le\left\|Q\left(\boldsymbol{W}^{(T_{\text{ERM}})}\right)-Q\left(\boldsymbol{W}^*\right)\right\|^2+2\eta T^* e^{-M\beta}.$$

■

Finally, we can prove that the model cannot learn (extremely) rare features within T^* iterations.

Lemma 21 *Suppose the event E_{init} occurs. For any $T\in[T_{\text{ERM}},T^*]$, we have $\gamma_s^{(T)}(s^*,k^*)\le 3\alpha^2$ for each $s'\in\{\pm1\}$.*

Proof [Proof of Lemma 21] For any $T\in[T_{\text{ERM}},T^*]$, we have

$$\begin{aligned}\gamma_s^{(T)}(s,k)&=\gamma_s^{(T_{\text{ERM}})}(s^*,k^*)+\frac{\eta}{n}\sum_{t=T_{\text{ERM}}}^{T-1}\sum_{i\in\mathcal{V}_{s^*,k^*}}g_i^{(t)}\phi'\left(\left\langle\boldsymbol{w}_s^{(t)},\boldsymbol{v}_{s^*,k^*}\right\rangle\right)\\&\le\gamma_s^{(T_{\text{ERM}})}(s^*,k^*)+\frac{\eta}{n}\sum_{t=T_{\text{ERM}}}^{T-1}\sum_{i\in\mathcal{V}_{s^*,k^*}}g_i^{(t)}\\&\le\gamma_s^{(T_{\text{ERM}})}(s^*,k^*)+\frac{\eta}{n}\sum_{t=T_{\text{ERM}}}^{T-1}\sum_{i\in\mathcal{V}_{s^*,k^*}}\ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right),\end{aligned}$$

where the first inequality is due to $\phi'\le1$ and the second inequality is due to $-\ell'\le\ell$. From the result of Section F.2.3 we know $\gamma_s^{(T_{\text{ERM}})}(s^*,k^*)\le\alpha^2$. Additionally, by Lemma 20 and Lemma 18, we have

$$\begin{aligned}\frac{\eta}{n}\sum_{t=T_{\text{ERM}}}^{(T-1)}\sum_{i\in\mathcal{V}_{s^*,k^*}}\ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)&\le\frac{\eta}{n}\sum_{t=T_{\text{ERM}}}^{(T^*)}\sum_{i\in\mathcal{V}_{s^*,k^*}}\ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)\\&\le\left\|Q\left(\boldsymbol{W}^{(T_{\text{ERM}})}\right)-Q(\boldsymbol{W}^*)\right\|^2+2\eta T^* e^{-M\beta}\\&\le 24M^2|\mathcal{V}_{s^*,k^*}|\sigma_{\text{d}}^{-2}d^{-1}+2\eta T^* e^{-M\beta}\\&\le\alpha^2+2\eta T^* e^{-M\beta}.\end{aligned}$$

The last inequality is due to $\rho_R n=o(\alpha^2\beta^2\sigma_{\text{d}}^2 d)$ by Assumption 8. Since $M=\beta^{-1}\log\left(\frac{2\eta T^*}{\alpha^2}\right)$, we have our conclusion. ■

What We Have So Far. Suppose the event E_{init} occurs. For any $t\in[T_{\text{ERM}},T^*]$, we have

- (Learn common features): For each $s\in\{\pm1\}$ and $k\in\mathcal{K}_C$,

$$\gamma_s^{(t)}(s,k)+\beta\gamma_{-s}^{(t)}(s,k)=\Omega(1).$$

- (Overfit (extremely) rare data): For each $s\in\{\pm1\}$, $k\in\mathcal{K}_R\cup\mathcal{K}_C$ and $i\in\mathcal{V}_{s,k}$,

$$\rho_s^{(t)}(i,\tilde{p}_i)+\beta\rho_{-s}^{(t)}(i,\tilde{p}_i)=\Omega(1).$$

- (Cannot learn (extremely) rare features): For each $s \in \{\pm 1\}$ and $k \in \mathcal{K}_R \cup \mathcal{K}_E$,

$$\gamma_s^{(t)}(s,k), \gamma_{-s}^{(t)}(s,k) = \mathcal{O}(\alpha^2).$$

- For any $s \in \{\pm 1\}, i \in [n]$, and $p \in [P] \setminus \{p_i^*\}$, $\rho_s^{(t)}(i,p) = \widetilde{\mathcal{O}}\left(\beta^{-1}\right)$,

F.2.5. TRAIN AND TEST ACCURACY

In this step, we will prove that the model trained by ERM has perfect training accuracy but has near-random guesses on (extremely) rare data.

For any $i \in \mathcal{V}_{s,k}$ with $s \in \{\pm 1\}$ and $k \in \mathcal{K}_C$, by Lemma 11, we have

$$\begin{aligned}
& y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i) \\
&= \sum_{p \in [P]} \left(\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right)\right) \\
&= \gamma_s^{(t)}(s,k) + \beta \gamma_{-s}^{(t)}(s,k) + \sum_{p \in [P] \setminus \{p_i^*\}} \left(\rho_s^{(t)}(i,p) + \beta \rho_{-s}^{(t)}(i,p) \right) \\
&\quad - \widetilde{\mathcal{O}}\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right) - \widetilde{\mathcal{O}}(\alpha \beta^{-1}) \\
&\geq \gamma_s^{(t)}(s,k) + \beta \gamma_{-s}^{(t)}(s,k) - \widetilde{\mathcal{O}}\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right) - \widetilde{\mathcal{O}}(\alpha \beta^{-1}) \\
&= \Omega(1) - \widetilde{\mathcal{O}}\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right) - \widetilde{\mathcal{O}}(\alpha \beta^{-1}) \\
&> 0,
\end{aligned}$$

for any $t \in [T_{\mathrm{ERM}}, T^*]$. In addition, for any $i \in \mathcal{V}_{s,k}$ with $s \in \{\pm 1\}$ and $k \in \mathcal{K}_R \cup \mathcal{K}_E$, we have

$$\begin{aligned}
& y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i) \\
&= \sum_{p \in [P]} \left(\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right)\right) \\
&= \gamma_s^{(t)}(s,k) + \beta \gamma_{-s}^{(t)}(s,k) + \sum_{p \in [P] \setminus \{p_i^*\}} \left(\rho_s^{(t)}(i,p) + \beta \rho_{-s}^{(t)}(i,p) \right) \\
&\quad - \widetilde{\mathcal{O}}\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right) - \widetilde{\mathcal{O}}(\alpha \beta^{-1}) \\
&\geq \rho_s^{(t)}(i,\tilde{p}_i) + \beta \rho_{-s}^{(t)}(i,\tilde{p}_i) - \widetilde{\mathcal{O}}\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right) - \widetilde{\mathcal{O}}(\alpha \beta^{-1}) \\
&= \Omega(1) - \widetilde{\mathcal{O}}\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right) - \widetilde{\mathcal{O}}(\alpha \beta^{-1}) \\
&> 0,
\end{aligned}$$

for any $t \in [T_{\mathrm{ERM}}, T^*]$. We can conclude that ERM with $t \in [T_{\mathrm{ERM}}, T^*]$ iterates achieve perfect training accuracy.

Next, let us move on to the test accuracy part. Let $(\boldsymbol{X}, y) \sim \mathcal{D}$ be a test data with $\boldsymbol{X} = \left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(P)}\right) \in \mathbb{R}^{d \times P}$ having feature patch p^*, dominant noise patch $\tilde{p}$, and feature vector $\boldsymbol{v}_{y,k}$.

We have $\boldsymbol{x}^{(p)} \sim N(\mathbf{0}, \sigma_{\mathrm{b}}^2 \boldsymbol{\Lambda})$ for each $p \in [P] \setminus \{p^*, \tilde{p}\}$ and $\boldsymbol{x}^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1} \sim N(\mathbf{0}, \sigma_{\mathrm{d}}^2 \boldsymbol{\Lambda})$ for some $s \in \{\pm 1\}$. Therefore, for all $t \in [T_{\mathrm{ERM}}, T^*]$ and $p \in [P] \setminus \{p^*, \tilde{p}\}$,

$$\begin{aligned}
&\left|\phi\left(\left\langle \boldsymbol{w}_1^{(t)}, \boldsymbol{x}^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(p)}\right\rangle\right)\right| \\
&\leq \left|\left\langle \boldsymbol{w}_1^{(t)} - \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(p)}\right\rangle\right| \\
&= \left|\left\langle \boldsymbol{w}_1^{(0)} - \boldsymbol{w}_{-1}^{(0)}, \boldsymbol{x}^{(p)}\right\rangle\right| + \sum_{i \in [n], q \in [P] \setminus \{p_i^*\}} \rho(i,q) \frac{\left|\left\langle \xi_i^{(q)}, \boldsymbol{x}^{(p)}\right\rangle\right|}{\left\|\xi_i^{(q)}\right\|^2} \\
&\leq \widetilde{\mathcal{O}}\left(\sigma_0 \sigma_{\mathrm{b}} d^{\frac{1}{2}}\right) + \widetilde{\mathcal{O}}\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right) \\
&= \widetilde{\mathcal{O}}\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right),
\end{aligned} \tag{7}$$

with probability at least $1 - o\left(\frac{1}{\mathrm{poly}(d)}\right)$ due to Lemma 9. Similarly, we have

$$\begin{aligned}
&\left|\phi\left(\left\langle \boldsymbol{w}_1^{(t)}, \boldsymbol{x}^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right)\right| \\
&\leq \left|\left\langle \boldsymbol{w}_1^{(t)} - \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right| = \widetilde{\mathcal{O}}\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right),
\end{aligned} \tag{8}$$

with probability at least $1 - o\left(\frac{1}{\mathrm{poly}(d)}\right)$.

Case 1: $k \in \mathcal{K}_C$

By Lemma 9

$$\begin{aligned}
&\left|\phi\left(\left\langle \boldsymbol{w}_1^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right)\right| \\
&\leq \left|\left\langle \boldsymbol{w}_1^{(t)} - \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right| \\
&\leq \alpha \left|\left\langle \boldsymbol{w}_1^{(t)} - \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{v}_{s,1}\right\rangle\right| + \left|\left\langle \boldsymbol{w}_1^{(t)} - \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(p)} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right| \\
&\leq \alpha \beta^{-1} \left|\phi\left(\left\langle \boldsymbol{w}_1^{(t)}, \boldsymbol{v}_{s,1}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{v}_{s,1}\right\rangle\right)\right| \\
&\quad + \left|\left\langle \boldsymbol{w}_1^{(0)} - \boldsymbol{w}_{-1}^{(0)}, \boldsymbol{x}^{(p)} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right| + \sum_{i \in [n], q \in [P] \setminus \{p_i^*\}} \rho(i,q) \frac{\left|\left\langle \xi_i^{(q)}, \boldsymbol{x}^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right|}{\left\|\xi_i^{(q)}\right\|^2} \\
&\leq \widetilde{\mathcal{O}}\left(\alpha \beta^{-2}\right) + \widetilde{\mathcal{O}}\left(\sigma_0 \sigma_{\mathrm{d}} d^{\frac{1}{2}}\right) + \widetilde{\mathcal{O}}\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right) \\
&= \widetilde{\mathcal{O}}\left(\alpha \beta^{-2}\right),
\end{aligned} \tag{9}$$

with probability at least $1 - o\left(\frac{1}{\mathrm{poly}(d)}\right)$. Suppose (7) and (9) holds. By Lemma 11, we have

$$\begin{aligned}
&y f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}) \\
&= \left(\phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{v}_{y,k} \right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{v}_{y,k} \right\rangle\right)\right) \\
&\quad + \sum_{p \in [P] \setminus \{p^*\}} \left(\phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right)\right) \\
&= \gamma_y^{(t)}(y,k) + \beta \gamma_{-y}^{(t)}(y,k) - \widetilde{\mathcal{O}}(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}) - \widetilde{\mathcal{O}}(\alpha\beta^{-1}) \\
&= \Omega(1) - \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right) - \widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) \\
&> 0.
\end{aligned}$$

Therefore, we have

$$\mathbb{P}_{(\boldsymbol{X},y)\sim\mathcal{D}}\left[y f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}) > 0 \mid \boldsymbol{x}^{(p^*)} = \boldsymbol{v}_{y,k}, k \in \mathcal{K}_C\right] \geq 1 - o\left(\frac{1}{\mathrm{poly}(d)}\right). \tag{10}$$

Case 2: $k \in \mathcal{K}_R \cup \mathcal{K}_E$ By triangular inequality and $\phi' \leq 1$, we have

$$\begin{aligned}
&\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}^{(\tilde{p})} \right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}^{(\tilde{p})} \right\rangle\right) \\
&= \phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{s,1} \right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \alpha \boldsymbol{v}_{s,1} \right\rangle\right) \\
&\quad + \left(\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(\tilde{p})} \right\rangle\right) - \phi\left(\gamma_s^{(t)}(s,1)\right)\right) - \left(\phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(\tilde{p})} \right\rangle\right) - \phi\left(\gamma_{-s}^{(t)}(s,1)\right)\right) \\
&\geq \phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{s,1} \right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \alpha \boldsymbol{v}_{s,1} \right\rangle\right) \\
&\quad - \alpha \left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,1} \right\rangle - \gamma_s^{(t)}(s,1)\right| - \alpha \left|\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,1} \right\rangle + \gamma_{-s}^{(t)}(s,1)\right| \\
&\quad - \left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1} \right\rangle\right| - \left|\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1} \right\rangle\right|.
\end{aligned}$$

In addition,

$$\begin{aligned}
&\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{s,1} \right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \alpha \boldsymbol{v}_{s,1} \right\rangle\right) \\
&= \left(\phi\left(\alpha\gamma_s^{(t)}(s,1)\right) - \phi\left(\alpha\gamma_{-s}^{(t)}(s,1)\right)\right) \\
&\quad + \left(\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{s,1} \right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha\gamma_s^{(t)}(s,1) \right\rangle\right)\right) \\
&\quad - \left(\phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \alpha \boldsymbol{v}_{s,1} \right\rangle\right) - \phi\left(\alpha\gamma_{-s}^{(t)}(s,1)\right)\right) \\
&\geq \left(\phi\left(\alpha\gamma_s^{(t)}(s,1)\right) - \phi\left(\alpha\gamma_{-s}^{(t)}(s,1)\right)\right) \\
&\quad - \alpha \left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,1} \right\rangle - \gamma_s^{(t)}(s,1)\right| - \alpha \left|\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,1} \right\rangle + \gamma_{-s}^{(t)}(s,1)\right|.
\end{aligned}$$

If (8) holds,then by Lemma 11, we have

$$
\begin{aligned}
&\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{s,1}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \alpha \boldsymbol{v}_{s,1}\right\rangle\right)\\
&\geq \alpha\left(\gamma_s^{(t)}(s,1) + \beta\gamma_{-s}^{(t)}(s,1)\right) - \mathcal{O}\left(\alpha n \beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right) - \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right)\\
&\geq \Omega(\alpha),
\end{aligned}
$$

where the last inequality is due to As a result,

$$
\begin{aligned}
&y f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X})\\
&= \phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{v}_{y,k}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{v}_{y,k}\right\rangle\right) + \phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right)\\
&\quad + \sum_{p \in [P]\setminus\{p^*, \tilde{p}\}} \left(\phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{x}^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{x}^{(p)}\right\rangle\right)\right).
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\left|\phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{v}_{y,k}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{v}_{y,k}\right\rangle\right)\right|\\
&\quad + \left|\sum_{p \in [P]\setminus\{p^*, \tilde{p}\}} \left(\phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{x}^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{x}^{(p)}\right\rangle\right)\right)\right|\\
&\leq \mathcal{O}(\alpha^2) + \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right)\\
&< \phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{w}^{(\tilde{p})}\right\rangle\right).
\end{aligned}
$$

Hence, $y f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}) > 0$ if $y = s$. Otherwise, $y f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}) < 0$. Therefore, we have

$$
\mathbb{P}_{(\boldsymbol{X},y)\sim\mathcal{D}}\left[y f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}) > 0 \mid \boldsymbol{x}^{(p^*)} = \boldsymbol{v}_{y,k} \in \mathcal{K}_R \cup \mathcal{K}_E\right] = \frac{1}{2} \pm o\left(\frac{1}{\mathrm{poly}(d)}\right)
$$

and we conclude

$$
\begin{aligned}
\mathbb{P}_{(\boldsymbol{X},y)\sim\mathcal{D}}\left[y f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}) > 0\right] &= \sum_{k \in \mathcal{K}_C} \rho_k + \frac{1}{2}\left(1 - \sum_{k\in\mathcal{K}_C}\rho_k\right) \pm o\left(\frac{1}{\mathrm{poly}(d)}\right)\\
&= 1 - \frac{1}{2}\sum_{k \in \mathcal{K}_R \cup \mathcal{K}_E} \rho_k \pm o\left(\frac{1}{\mathrm{poly}(d)}\right).
\end{aligned}
$$

□

Appendix G. Proof for Cutout

In this section, we use $g_{i,\mathcal{C}}^{(t)} := \frac{1}{1+\exp\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})\right)}$ for each data i, $\mathcal{C} \subset [P]$ with $|\mathcal{C}| = C$ and iteration t, for simplicity.

G.1. Proof of Lemma 10: Cutout Training Case

For $s \in \{\pm 1\}$ and iterate t,

$$
\begin{aligned}
&\boldsymbol{w}_s^{(t+1)} - \boldsymbol{w}_s^{(t)} \\
&= -\eta \nabla_{\boldsymbol{w}_s} \mathcal{L}_{\text{Cutout}}\left(\boldsymbol{W}^{(t)}\right) \\
&= \frac{\eta}{n} \sum_{i \in [n]} s y_i \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}} \left[g_{i,\mathcal{C}}^{(t)} \sum_{p \notin \mathcal{C}} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)} \right] \\
&= \frac{\eta}{n} \left(\sum_{i \in \mathcal{V}_s} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}} \left[g_{i,\mathcal{C}}^{(t)} \sum_{p \notin \mathcal{C}} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)} \right] \right. \\
&\qquad \left. - \sum_{i \in \mathcal{V}_{-s}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}} \left[g_{i,\mathcal{C}}^{(t)} \sum_{p \notin \mathcal{C}} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)} \right] \right),
\end{aligned}
$$

and we have

$$
\begin{aligned}
&\sum_{i \in \mathcal{V}_s} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}} \left[g_{i,\mathcal{C}}^{(t)} \sum_{p \notin \mathcal{C}} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)} \right] \\
&= \sum_{k \in [K]} \sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}} \left[g_{i,\mathcal{C}}^{(t)} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k} \right\rangle\right) \cdot \mathbb{1}_{p_i^* \notin \mathcal{C}} \right] \boldsymbol{v}_{s,k} \\
&\quad + \sum_{i \in \mathcal{V}_s} \sum_{p \in [P] \setminus \{p_i^*, \tilde{p}_i\}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}} \left[g_{i,\mathcal{C}}^{(t)} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \xi_i^{(p)} \right\rangle\right) \cdot \mathbb{1}_{p \notin \mathcal{C}} \right] \xi_i^{(p)} \\
&\quad + \sum_{i \in \mathcal{V}_s \cap \mathcal{F}_s} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}} \left[g_{i,\mathcal{C}}^{(t)} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{s,1} + \xi_i^{(\tilde{p}_i)} \right\rangle\right) \cdot \mathbb{1}_{\tilde{p}_i \notin \mathcal{C}} \right] \left(\alpha \boldsymbol{v}_{s,1} + \xi_i^{(\tilde{p}_i)} \right) \\
&\quad + \sum_{i \in \mathcal{V}_s \cap \mathcal{F}_{-s}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}} \left[g_{i,\mathcal{C}}^{(t)} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{-s,1} + \xi_i^{(\tilde{p}_i)} \right\rangle\right) \cdot \mathbb{1}_{\tilde{p}_i \notin \mathcal{C}} \right] \left(\alpha \boldsymbol{v}_{-s,1} + \xi_i^{(\tilde{p}_i)} \right),
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{i \in \mathcal{V}_{-s}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[g_{i,\mathcal{C}}^{(t)} \sum_{p \notin \mathcal{C}} \phi' \left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right) \boldsymbol{x}_i \right] \\
&= \sum_{k \in [K]} \sum_{i \in \mathcal{V}_{-s,k}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[g_{i,\mathcal{C}}^{(t)} \phi' \left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k} \right\rangle \right) \cdot \mathbb{1}_{p_i^* \notin \mathcal{C}} \right] \boldsymbol{v}_{s,k} \\
&\quad + \sum_{i \in \mathcal{V}_{-s}} \sum_{p \in [P] \setminus \{p_i^*, \tilde{p}_i\}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[g_{i,\mathcal{C}}^{(t)} \phi' \left(\left\langle \boldsymbol{w}_s^{(t)}, \xi_i^{(p)} \right\rangle \right) \cdot \mathbb{1}_{p \notin \mathcal{C}} \right] \xi_i^{(p)} \\
&\quad + \sum_{i \in \mathcal{V}_{-s} \cap \mathcal{F}_s} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[g_{i,\mathcal{C}}^{(t)} \phi' \left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{s,1} + \xi_i^{(\tilde{p}_i)} \right\rangle \right) \cdot \mathbb{1}_{\tilde{p}_i \notin \mathcal{C}} \right] \left(\alpha \boldsymbol{v}_{s,1} + \xi_i^{(\tilde{p}_i)} \right) \\
&\quad + \sum_{i \in \mathcal{V}_{-s} \cap \mathcal{F}_{-s}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[g_{i,\mathcal{C}}^{(t)} \phi' \left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{-s,1} + \xi_i^{(\tilde{p}_i)} \right\rangle \right) \cdot \mathbb{1}_{\tilde{p}_i \notin \mathcal{C}} \right] \left(\alpha \boldsymbol{v}_{-s,1} + \xi_i^{(\tilde{p}_i)} \right).
\end{aligned}
$$

Hence, if we define $\gamma_s^{(t)}(s',k)$'s and $\rho_s^{(t)}(i,p)$'s recursively by using the rule

$$
\gamma_s^{(t+1)}(s',k) = \gamma_s^{(t)}(s',k) + \frac{\eta}{n} \sum_{i \in \mathcal{V}_{s',k}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[g_{i,\mathcal{C}}^{(t)} \phi' \left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k} \right\rangle \right) \cdot \mathbb{1}_{p_i^* \notin \mathcal{C}} \right], \tag{11}
$$

$$
\rho_s^{(t+1)}(i,p) = \rho_s^{(t)}(i,p) + \frac{\eta}{n} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[g_{i,\mathcal{C}}^{(t)} \phi' \left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right) \left\| \xi_i^{(p)} \right\|^2 \cdot \mathbb{1}_{p \notin \mathcal{C}} \right], \tag{12}
$$

starting from $\gamma_s^{(0)}(s',k) = \rho_s^{(0)}(i,p) = 0$ for each $s, s' \in \{\pm 1\}, k \in [K], i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$, then we have

$$
\begin{aligned}
\boldsymbol{w}_s^{(t)} &= \boldsymbol{w}_s^{(0)} + \sum_{k \in [K]} \gamma_s^{(t)}(s,k) \boldsymbol{v}_{s,k} - \sum_{k \in [K]} \gamma_s^{(t)}(-s,k) \boldsymbol{v}_{-s,k} \\
&\quad + \sum_{i \in \mathcal{V}_s, p \in [P] \setminus \{p_i^*\}} \rho_s^{(t)}(i,p) \frac{\xi_i^{(p)}}{\left\| \xi_i^{(p)} \right\|^2} - \sum_{i \in \mathcal{V}_{-s}, p \in [P] \setminus \{p_i^*\}} \rho_s^{(t)}(i,p) \frac{\xi_i^{(p)}}{\left\| \xi_i^{(p)} \right\|^2} \\
&\quad + \alpha \left(\sum_{i \in \mathcal{F}_s} s y_i \rho_s^{(t)}(i,\tilde{p}_i) \frac{\boldsymbol{v}_{s,1}}{\left\| \xi_i^{(\tilde{p}_i)} \right\|^2} + \sum_{i \in \mathcal{F}_{-s}} s y_i \rho_s^{(t)}(i,\tilde{p}_i) \frac{\boldsymbol{v}_{-s,1}}{\left\| \xi_i^{(\tilde{p}_i)} \right\|^2} \right),
\end{aligned}
$$

for each $s \in \{\pm 1\}$. Furthermore, $\gamma_s^{(t)}(s',k)$'s and $\rho_s^{(t)}(i,p)$'s are monotone increasing. □

G.2. Proof of Theorem 4

To demonstrate Theorem 4, we present a structured proof comprising the following seven steps:

1. Establish upper bounds on $\gamma_s^{(t)}(s',k)$'s and $\rho_s^{(t)}(i,p)$'s to apply Lemma 11 (Section G.2.1).
2. Demonstrate that the model quickly learns common and rare features (Section G.2.2).
3. Show that the model overfits augmented data if it does not contain common or rare features (Section G.2.3).

4. Confirm the persistence of this tendency until T^* iterates (Section G.2.4).
5. Characterize train accuracy and test accuracy (Section G.2.5).

G.2.1. Bounds on the Coefficients in Feature Noise Decomposition

The following lemma provides upper bounds on Lemma 10 during T^* iterations.

Lemma 22 *Suppose the event E_{init} occurs. For any $0 \le t \le T^*$, we have*

$$0 \le \gamma_s^{(t)}(s,k) + \beta\gamma_s^{(t)}(-s,k) \le 4\log(\eta T^*), \quad 0 \le \rho_{y_i}^{(t)}(i,p) + \beta\rho_{-y_i}^{(t)}(i,p) \le 4\log\left(\eta T^*\right),$$

for all $s \in \{\pm 1\}, k \in [K], i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$.

Proof [Proof of Lemma 22] We will prove this by using induction on t. The initial case $t = 0$ is trivial. Suppose the statement holds at $t = T$ and consider the case $t = T + 1$.

Note that from our Assumption 8, the following hold:

- $\eta \le \frac{\log(\eta T^*))}{2}$,
- $\alpha\beta^{-1}, n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}} = o(1)$.

Let $\tilde{T}_{s,k} \le T$ denote the smallest iteration where $\gamma_s^{(\tilde{T}_{s,k}+1)}(s,k) + \beta\gamma_{-s}^{(\tilde{T}_{s,k}+1)}(s,k) > 2\log(\eta T^*)$. We assume the existence of $\tilde{T}_{s,k}$, as its absence would directly lead to our conclusion due to our small choice of η.

By (11), we have

$$\begin{aligned}
&\gamma_s^{(T+1)}(s,k) + \beta\gamma_{-s}^{(T+1)}(s,k)\\
&= \gamma_s^{(\tilde{T}_{s,k})}(s,k) + \beta\gamma_{-s}^{(\tilde{T}_{s,k})}(s,k)\\
&\quad + \sum_{t=\tilde{T}_{s,k}}^{T} \left(\gamma_s^{(t+1)}(s,k) + \beta\gamma_{-s}^{(t+1)}(s,k) - \gamma_s^{(t)}(s,k) - \beta\gamma_{-s}^{(t)}(s,k)\right)\\
&\le 2\log(\eta T^*) + \log(\eta T^*)\\
&\quad + \frac{\eta}{n}\sum_{t=\tilde{T}_{s,k}+1}^{T}\sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_{\mathcal{C}}}\left[g_{i,\mathcal{C}}^{(t)}\left(\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right)\right)\cdot \mathbb{1}_{p_i^*\notin\mathcal{C}}\right].
\end{aligned}$$

The inequality is due to $\gamma_s^{(\tilde{T}_{s,k})}(s,k) + \beta\gamma_{-s}^{(\tilde{T}_{s,k})}(s,k) \le 2\log(\eta T^*)$ and

$$\begin{aligned}
&\frac{\eta}{n}\sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_{\mathcal{C}}}\left[g_{i,\mathcal{C}}^{(\tilde{T}_{s,k})}\left(\phi'\left(\left\langle \boldsymbol{w}_s^{(\tilde{T}_{s,k})}, \boldsymbol{v}_{s,k}\right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-s}^{(\tilde{T}_{s,k})}, \boldsymbol{v}_{s,k}\right\rangle\right)\right)\cdot \mathbb{1}_{p_i^*\notin\mathcal{C}}\right]\\
&\le 2\eta \le \log(\eta T^*),
\end{aligned}$$

from our choice of $\tilde{T}_{s,k}$ and η.

For each $t = \tilde{T}_{s,k} + 1, \dots T$, $i \in \mathcal{V}_{s,k}$, and $\mathcal{C} \subset [P]$ such that $|\mathcal{C}| = C$ and $p_i^* \notin \mathcal{C}$, we have

$$\begin{aligned}
&y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}) \\
&= \phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k} \right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,k} \right\rangle\right) + \sum_{p \notin \mathcal{C} \cup \{p_i^*\}} \left(\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \right) \\
&\geq \gamma_s^{(t)}(s,k) + \beta \gamma_{-s}^{(t)}(s,k) + \sum_{p \in [P] \setminus \{p_i^*\}} \left(\rho_s^{(t)}(i,p) + \beta \rho_{-s}^{(t)}(i,p) \right) \\
&\quad - 2P\widetilde{\mathcal{O}}(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}) - 2\widetilde{\mathcal{O}}(\alpha\beta^{-1}) \\
&\geq \frac{3}{2}\log(\eta T^*)
\end{aligned}$$

The first inequality is due to Lemma 11 and the second inequality holds since from our choice of t, $\gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) \geq 2\log(\eta T^*)$.

Hence, we obtain

$$\begin{aligned}
&\frac{\eta}{n} \sum_{t=\tilde{T}_{s,k}}^{T} \sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[g_{i,\mathcal{C}}^{(t)} \left(\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k} \right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,k} \right\rangle\right) \right) \cdot \mathbb{1}_{p_i^* \notin \mathcal{C}} \right] \\
&\leq \frac{2\eta}{n} \sum_{t=\tilde{T}_{s,k}}^{T} \sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\exp\left(-y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})\right) \cdot \mathbb{1}_{p_i^* \notin \mathcal{C}} \right] \\
&\leq \frac{2|\mathcal{V}_{s,k}|}{n} (\eta T^*) \exp\left(-\frac{3}{2}\log(\eta T^*) \right) \\
&\leq \log(\eta T^*).
\end{aligned}$$

Thus, we have $\gamma_s^{(T+1)}(s,k) + \beta\gamma_{-s}^{(T+1)}(s,k) \leq 4\log(\eta T^*)$.

Next, we will follow similar arguments to show that

$$\rho_{y_i}^{(T+1)}(i,p) + \beta\rho_{-y_i}^{(T+1)}(i,p) \leq 4\log(\eta T^*)$$

for each $i \in [n]$ and $p \in [P] \setminus \{\tilde{p}_i\}$.

Let $\tilde{T}_i^{(p)} \leq T$ be the smallest iteration such that $\rho_{y_i}^{(\tilde{T}_i^{(p)}+1)}(i,p) + \beta\rho_{-y_i}^{(\tilde{T}_i^{(p)}+1)}(i,p) > 2\log(\eta T^*)$. We assume the existence of $\tilde{T}_i^{(p)}$, as its absence would directly lead to our conclusion due to our small choice of η.

By (12), we have

$$\begin{aligned}
&\rho_{y_i}^{(T+1)}(i,p) + \beta\rho_{-y_i}^{(T+1)}(i,p)\\
&= \rho_{y_i}^{(\tilde{T}_i^{(p)})}(i,p) + \beta\rho_{-y_i}^{(\tilde{T}_i^{(p)})}(i,p)\\
&\quad + \sum_{t=\tilde{T}_i^{(p)}}^{T} \left(\rho_{y_i}^{(t+1)}(i,p) + \beta\rho_{-y_i}^{(t+1)}(i,p) - \rho_{y_i}^{(t)}(i,p) - \beta\rho_{-y_i}^{(t)}(i,p)\right)\\
&\le 2\log(\eta T^*) + \log(\eta T^*)\\
&\quad + \frac{\eta}{n}\sum_{t=\tilde{T}_i^{(p)}+1}^{T} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}\left[g_{i,\mathcal{C}}^{(t)}\left(\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right)\cdot \mathbb{1}_{p\notin\mathcal{C}}\right]
\end{aligned}$$

The inequality is due to $\rho_{y_i}^{(t)}(i,p) + \beta\rho_{-y_i}^{(t)}(i,p) \le 2\log T^*$ and

$$\frac{\eta}{n} g_i^{(\tilde{T}_i^{(p)})}\left[\phi'\left(\left\langle \boldsymbol{w}_s^{(\tilde{T}_i^{(p)})}, \boldsymbol{x}_i^{(p)}\right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-s}^{(\tilde{T}_i^{(p)})}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right] \le 2\eta \le \log(\eta T^*),$$

from our choice of $\tilde{T}_i^{(p)}$ and η.

For each $t = \tilde{T}_i^{(p)} + 1, \dots, T$, and $\mathcal{C} \subset [P]$ such that $|\mathcal{C}| = C$ and $p \notin \mathcal{C}$, we have

$$\begin{aligned}
&y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})\\
&= \phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) + \sum_{q\notin\mathcal{C}\cup\{p\}}\left(\phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \boldsymbol{x}_i^{(q)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \boldsymbol{x}_i^{(q)}\right\rangle\right)\right)\\
&\ge \rho_{y_i}^{(t)}(i,p) + \beta\rho_{-y_i}^{(t)}(i,p) - 2P\widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_\mathrm{d}\sigma_\mathrm{b}^{-1}d^{-\frac{1}{2}}\right) - 2\widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right)\\
&\ge \frac{3}{2}\log(\eta T^*).
\end{aligned}$$

The first inequality is due to Lemma 11 and the second inequality holds since from our choice of t, $\rho_{y_i}^{(t)}(i,p) + \beta\rho_{-y_i}^{(t)}(i,p) \ge 2\log(\eta T^*)$.

Therefore, we have

$$\begin{aligned}
&\frac{\eta}{n}\sum_{t=\tilde{T}_i^{(p)}}^{T} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}\left[g_{i,\mathcal{C}}^{(t)}\left(\phi'\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right)\cdot \mathbb{1}_{p\notin\mathcal{C}}\right]\\
&\le \frac{2\eta}{n}\sum_{t=\tilde{T}_i^{(p)}}^{T} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}\left[\exp\left(-y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})\right)\mathbb{1}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}\right] \le \frac{2}{n}(\eta T^*)\exp\left(-\frac{3}{2}\log(\eta T^*)\right)\\
&\le \log(\eta T^*),
\end{aligned}$$

and we conclude $\rho_{y_i}^{(T+1)}(i,p) + \beta\rho_{-y_i}^{(T+1)}(i,p) \le 4\log(\eta T^*)$. ■

G.2.2. LEARNING COMMON FEATURES AND RARE FEATURES

In the initial stages of training, the model quickly learns common features while exhibiting minimal overfitting to Gaussian noises.

First, we establish lower bounds on the number of iterations, ensuring that background noise coefficients $\rho_s^{(t)}(i,p)$ for $p \neq p_i^*, \tilde{p}_i$ remain small, up to the order of α^2.

Lemma 23 *Suppose the event* E_{init} *occurs. There exists* $\tilde{T} > \frac{2n\alpha^2}{3\eta\sigma_{\text{b}}^2 d}$ *such that* $\rho_s^{(t)}(i,p) \leq \alpha^2$ *for all* $0 \leq t < \tilde{T}, s \in \{\pm 1\}, i \in [n]$ *and* $p \in [P] \setminus \{p_i^*, \tilde{p}_i\}$.

Proof [Proof of Lemma 23] Let $\tilde{T}$ be the smallest iteration such that $\rho_s^{(\tilde{T})}(i,p) \geq \alpha^2$ for some $s \in \{\pm 1\}, i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$. We assume the existence of $\tilde{T}$, as its absence would directly lead to our conclusion. Then, for any $0 \leq t < \tilde{T}$, we have

$$\rho_s^{(t+1)}(i,p) = \rho_s^{(t)}(i,p) + \frac{\eta}{n}\mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}\left[g_{i,\mathcal{C}}^{(t)}\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) \cdot \mathbb{1}_{p\notin\mathcal{C}}\right]\left\|\xi_i^{(p)}\right\|^2 \leq \rho_s^{(t)}(i,p) + \frac{3\eta\sigma_{\text{b}}^2 d}{2n},$$

where the inequality is due to $g_{i,\mathcal{C}}^{(t)} < 1$, $\phi' \leq 1$, and $\left\|\xi_i^{(p)}\right\|^2 \leq \frac{3}{2}\sigma_{\text{b}}^2 d$. Hence, we have

$$\alpha^2 \leq \rho_s^{(\tilde{T})}(i,p) < \frac{3\eta\sigma_{\text{b}}^2 d}{2n}\tilde{T},$$

and we conclude $\tilde{T} > \frac{2n\alpha^2}{3\eta\sigma_{\text{b}}^2 d}$ which is the desired result. ■

Next, we will show that the model learns common features in at least constant order within $\tilde{T}$ iterates.

Lemma 24 *Suppose the event* E_{init} *occurs and* $\beta^{-1}\rho_k^{-1} = o\left(\frac{n\alpha^2}{\sigma_{\text{b}}^2 d}\right)$ *for some* $k \in [K]$*. Then, for each* $s \in \{\pm 1\}$*. there exists* $T_{s,k} \leq \frac{9nP}{\eta\beta|\mathcal{V}_{s,k}|}$ *such that* $\gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) \geq 1$ *for any* $t > T_{s,k}$.

Proof [Proof of Lemma 24] Note that from our Assumption 8, $n\beta^{-1}\sigma_{\text{d}}\sigma_{\text{b}}^{-1}d^{-\frac{1}{2}}, \alpha^2, \alpha\beta^{-1} = o(1)$.

Suppose $\gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) < 1$ for all $0 \leq t \leq \frac{2n\alpha^2}{3\eta\sigma_{\text{b}}^2 d}$. For each $i \in \mathcal{V}_{s,k}$ and $\mathcal{C} \subset [P]$ with $|\mathcal{C}| = C$ such that $p_i^* \notin \mathcal{C}$ and $\tilde{p}_i \in \mathcal{C}$, we have

$$\begin{aligned}
&y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}) \\
&= \phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) + \sum_{p\notin\mathcal{C}\cup\{p_i^*\}}\left(\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right) \\
&\leq \gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) + \sum_{p\notin\mathcal{C}\cup\{p_i^*\}}\left(\rho_s^{(t)}(i,p) + \beta\rho_{-s}^{(t)}(i,p)\right) \\
&\quad + 2P\widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\text{d}}\sigma_{\text{b}}^{-1}d^{-\frac{1}{2}}\right) + 2\widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) \\
&\leq 1 + 2P\alpha^2 + 2P\widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\text{d}}\sigma_{\text{b}}^{-1}d^{-\frac{1}{2}}\right) + 2\widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) \\
&\leq 2.
\end{aligned}$$

The first inequality is due to Lemma 11 and the second inequality is due to Lemma 23. Thus, $g_{i,\mathcal{C}}^{(t)} = \frac{1}{1+\exp\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})\right)} > \frac{1}{9}$ and we have

$$\begin{aligned}
&\gamma_s^{(t+1)}(s,k) + \beta\gamma_{-s}^{(t+1)}(s,k)\\
&= \gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k)\\
&\quad + \frac{\eta}{n}\sum_{i\in\mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_{\mathcal{C}}}\left[g_{i,\mathcal{C}}^{(t)}\left(\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right) + \beta\phi'\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,k}\right\rangle\right)\right)\cdot \mathbb{1}_{p_i^*\notin\mathcal{C}}\right]\\
&\geq \gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) + \frac{\eta\beta}{9n}\sum_{i\in\mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_{\mathcal{C}}}[\mathbb{1}_{p_i^*\notin\mathcal{C}\wedge\tilde{p}_i\in\mathcal{C}}]\\
&= \gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) + \frac{\eta\beta C(P-C)}{9nP(P-1)}\\
&\geq \gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) + \frac{\eta\beta}{9nP}.
\end{aligned}$$

From the given condition in the lemma statement, we have $\frac{9nP}{\eta\beta|\mathcal{V}_{s,k}|} = o\left(\frac{2n\alpha^2}{3\eta\sigma_{\mathrm{b}}^2 d}\right)$. If we choose $t_0 \in \left[\frac{9nP}{\eta\beta|\mathcal{V}_{s,k}|}, \frac{2n\alpha^2}{3\eta\sigma_{\mathrm{b}}^2 d}\right]$, then

$$1 > \gamma_s^{(t_0)}(s,k) + \beta\gamma_{-s}^{(t_0)}(s,k) \geq \frac{\eta\beta|\mathcal{V}_{s,k}|}{9nP}t_0 \geq 1,$$

and this is contradictory. Hence, there exists $0 \leq T_{s,k} < \frac{2n\alpha^2}{3\eta\sigma_{\mathrm{b}}^2 d}$ such that $\gamma_s^{(T_{s,k}+1)}(s,k) + \beta\gamma_{-s}^{(T_{s,k}+1)}(s,k) \geq 1$ and choose the smallest one. Then we obtain

$$1 \geq \gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) \geq \frac{\eta\beta|\mathcal{V}_{s,k}|}{9nP}T_{s,k}.$$

Therefore, $T_{s,k} \leq \frac{9nP}{\eta\beta|\mathcal{V}_{s,k}|}$ and this is what we desired. ■

What We Have So Far. For any common feature or rare feature $\boldsymbol{v}_{s,k}$ with $s \in \{\pm 1\}$ and $k \in \mathcal{K}_C \cup \mathcal{K}_R$, it satisfies $\beta^{-1}\rho_k^{-1} = o\left(\frac{n\alpha^2}{\sigma_{\mathrm{d}}^2 d}\right)$. By Lemma 24, at any iterate $t \in \left[\bar{T}_1, T^*\right]$ with $\bar{T}_1 := \max_{s\in\{\pm1\},k\in\mathcal{C}} T_{s,k}$, the following properties hold if the event E_{init} occurs:

- (Learn common/rare features): For $s \in \{\pm 1\}$ and $k \in \mathcal{K}_C \cup \mathcal{K}_R$, $\gamma_s^{(t)}(s,k) + \beta\gamma_{-s}^{(t)}(s,k) = \Omega(1)$,
- For any $s \in \{\pm 1\}, i \in [n]$, and $p \in [P] \setminus \{p_i^*\}$, $\rho_s^{(t)}(i,p) = \widetilde{\mathcal{O}}\left(\beta^{-1}\right)$.

G.2.3. OVERFITTING AUGMENTED DATA

In the previous step, we have shown that common data can be well-classified by learning common features. In this step, we will show that the model correctly classifies rare or extreme data by overfitting dominant noise instead of learning its features.

We first introduce lower bounds on the number of iterates such that feature coefficients $\gamma_s^{(t)}(s',k)$ remain small, up to the order of α^2. We first introduce lower bounds on the number of iterates such that feature coefficients $\gamma_s^{(t)}(s',k)$ remain small, up to the order of α^2. This lemma holds to any kind of features, but we will focus on extremely rare features. This does not contradict the results from Section G.2.2 for common features and rare features since the upper bound on the number of iterations in Lemma 24 is larger than the lower bound on the number of iterations in this lemma.

Lemma 25 *Suppose the event E_{init} occurs. For each $s \in \{\pm 1\}$ and $k \in [K]$, there exists $\tilde{T}_{s,k} \geq \frac{n\alpha^2}{\eta|\mathcal{V}_{s,k}|}$ such that $\gamma_{s'}^{(t)}(s,k) \leq \alpha^2$ for any $0 \leq t < \tilde{T}_{s,k}$ and $s' \in \{\pm 1\}$.*

Proof [Proo of Lemma 25] Let $\tilde{T}_{s,k}$ be the smallest iterate such that $\gamma_{s'}^{(t)}(s,k) > \alpha^2$ for some $s'\{\pm 1\}$. We assume the existence of $\tilde{T}_{s,k}$, as its absence would directly lead to our conclusion.

For any $0 \leq t < \tilde{T}_{s,k}$,

$$\gamma_{s'}^{(t+1)}(s,k) = \gamma_{s'}^{(t)}(s,k) + \frac{\eta}{n} \sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_{\mathcal{C}}} \left[g_{i,\mathcal{C}}^{(t)} \phi' \left(\left\langle \boldsymbol{w}_{s'}^{(t)}, \boldsymbol{v}_{s,k} \right\rangle \right) \cdot \mathbb{1}_{p_i^* \notin \mathcal{C}} \right] \leq \gamma_{s'}^{(t)}(s,k) + \frac{\eta|\mathcal{V}_{s,k}|}{n},$$

and we have $\alpha^2 \leq \gamma_{s'}^{(\tilde{T}_{s,k})} \leq \frac{\eta|\mathcal{V}_{s,k}|}{n}\tilde{T}_{s,k}$. We conclude $\tilde{T}_{s,k} \geq \frac{n\alpha^2}{\eta|\mathcal{V}_{s,k}|}$ which is the desired result. ■

Next, we will show that the model overfits data augmented not containing common or rare features in at least constant order within $\tilde{T}_{s,k}$ iterates.

Lemma 26 *Suppose the event E_{init} occurs and $\frac{n}{\beta\sigma_{\text{b}}^2 d} = o\left(\alpha^2\rho_k^{-1}\right)$. For each $i \in [n]$ and $\mathcal{C} \subset [P]$ with $|\mathcal{C}| = C$, if (1) $i \in \mathcal{V}_{s,k}$ and $p_i^* \notin \mathcal{C}$ or (2) $p_i^* \in \mathcal{C}$, then there exists $T_{i,\mathcal{C}} \leq \frac{18n}{\eta\beta\sigma_{\text{b}}^2 d}$ such that*

$$\sum_{p \notin \mathcal{C}\cup\{p_i^*\}} \left(\rho_{y_i}^{(t)}(i,p) + \beta\rho_{-y_i}^{(t)}(i,p) \right) \geq 1,$$

for any $t > T_{i,\mathcal{C}}$.

Proof [Proof of Lemma 26] Note that from our Assumption 8, the following holds:

- $\alpha^2, n\beta^{-1}\sigma_{\text{d}}\sigma_{\text{b}}^{-1}d^{-\frac{1}{2}}, \alpha\beta^{-1} = o(1)$.

Suppose $\sum_{p\notin\mathcal{C}} \left(\rho_s^{(t)}(i,p) + \beta\rho_{-s}^{(t)}(i,p) \right) < 1$, for all $0 \leq t \leq \frac{n\alpha^2}{\eta|\mathcal{V}_{s,k}|}$.

We have

$$\begin{aligned}
& y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}) \\
&= \sum_{p\notin\mathcal{C}} \left(\phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right) - \phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right) \right) \\
&\leq \gamma_{y_i}^{(t)}(y_i,k) + \beta\gamma_{-y_i}^{(t)}(y_i,k) + \sum_{p\notin\mathcal{C}\cup\{p_i^*\}} \left(\rho_{y_i}^{(t)}(i,p) + \beta\rho_{-y_i}^{(t)}(i,p) \right) \\
&\quad + 2P\widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\text{d}}\sigma_{\text{b}}^{-1}d^{-\frac{1}{2}} \right) + 2\widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) \\
&\leq (1+\beta)\alpha^2 + 1 + 2P\widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\text{d}}\sigma_{\text{b}}^{-1}d^{-\frac{1}{2}} \right) + 2\widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) \\
&\leq 2,
\end{aligned}$$

and $g_i^{(t)} = \frac{1}{1+\exp\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)} \geq \frac{1}{9}$. Also, for each $p \notin \mathcal{C} \cup \{p_i^*\}$, we have

$$\begin{aligned}
&\rho_s^{(t+1)}(i,p) + \beta \rho_{-s}^{(t+1)}(i,p) \\
&= \rho_s^{(t)}(i,p) + \beta \rho_{-s}^{(t)}(i,p) + \frac{\eta}{n} g_i^{(t)} \left(\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) + \beta \phi'\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \right) \left\| \xi_i^{(\tilde{p}_i)} \right\|^2 \\
&\geq \rho_s^{(t)}(i,p) + \beta \rho_{-s}^{(t)}(i,p) + \frac{\eta \beta \sigma_{\mathrm{b}}^2 d}{18n},
\end{aligned}$$

where the last inequality is due to $\left\| \xi_i^{(p)} \right\|^2 \geq \frac{1}{2}\sigma_{\mathrm{b}}^2 d$ and $\phi' \geq \beta$. We also have

$$\sum_{p \notin \mathcal{C} \cup \{p\}} \left(\rho_s^{(t+1)}(i,p) + \beta \rho_{-s}^{(t+1)}(i,p) \right) \geq \sum_{p \notin \mathcal{C} \cup \{p_i^*\}} \left(\rho_s^{(t)}(i,p) + \beta \rho_{-s}^{(t)}(i,p) \right) + \frac{\eta \beta \sigma_{\mathrm{b}}^2 d}{18n}$$

From the given condition in the lemma statement, we have $\frac{18n}{\eta \beta \sigma_{\mathrm{b}}^2 d} = o\left(\frac{n\alpha^2}{\eta |\mathcal{V}_{s,k}|}\right)$. If we choose $t_0 \in \left[\frac{18n}{\eta \beta \sigma_{\mathrm{b}}^2 d}, \frac{n\alpha^2}{\eta |\mathcal{V}_{s,k}|}\right]$, then we have

$$1 > \sum_{p \notin \mathcal{C} \cup \{p_i^*\}} \left(\rho_s^{(t)}(i,p) + \beta \rho_{-s}^{(t)}(i,p) \right) \geq \frac{\eta \beta \sigma_{\mathrm{b}}^2 d}{18n} t_0 \geq 1,$$

and this makes a contradiction. Therefore, there exists some $0 \leq T_{i,\mathcal{C}} < \frac{n\alpha^2}{\eta |\mathcal{V}_{s,k}|}$ such that $\sum_{p \notin \mathcal{C} \cup \{p_i^*\}} \left(\rho_s^{(T_{i,\mathcal{C}}+1)}(i,p) + \beta \rho_{-s}^{(T_{i,\mathcal{C}}+1)}(i,p) \right) \geq 1$ and let us choose the smallest one.

For any $0 \leq t < T_{i,\mathcal{C}}$, we have

$$1 \geq \sum_{p \notin \mathcal{C} \cup \{p_i^*\}} \left(\rho_s^{(T_{i,\mathcal{C}})}(i,p) + \beta \rho_{-s}^{(T_{i,\mathcal{C}})}(i,p) \right) \geq \frac{\eta \beta \sigma_{\mathrm{b}}^2 d}{18n} T_i,$$

and we conclude that $T_{i,\mathcal{C}} \leq \frac{18n}{\eta \beta \sigma_{\mathrm{d}}^2 d}$ and this is what we desired.

■

What We Have So Far. For any $s \in \{\pm 1\}$ and $k \in \mathcal{K}_E$, it satisfies $\frac{n}{\beta \sigma_{\mathrm{d}}^2 d} = o\left(\alpha^2 \rho_k^{-1}\right)$. By Lemma 26 at iterate $t \in [T_{\mathrm{Cutout}}, T^*]$ with

$$T_{\mathrm{Cutout}} := \max_{\substack{s \in \{\pm 1\} \\ k \in \mathcal{K}_E}} \max_{\substack{i \in \mathcal{V}_{s,k} \\ p_i^* \notin \mathcal{C} \wedge \tilde{p}_i \in \mathcal{C}}} T_{i,\mathcal{C}} \quad \in \left[\bar{T}_1, T^*\right]$$

the following properties hold if the event E_{init} occurs.:

- (Learn common/rare features): For any $s \in \{\pm 1\}$ and $k \in \mathcal{K}_C \cup \mathcal{K}_R$,

$$\gamma_s^{(t)}(s,k) + \beta \gamma_{-s}^{(t)}(s,k) = \Omega(1),$$

- (Overfit Data Augmented): For each $i \in [n]$, $\mathcal{C} \subset [P]$ with $|\mathcal{C}|$ such that (1) $i \in \mathcal{V}_{s,k}$ and $p_i^* \notin \mathcal{C}$ or (2) $p_i^* \in \mathcal{C}$
$$\sum_{p \notin \mathcal{C} \cup \{p_i^*\}} \left(\rho_{y_i}^{(t)}(i,p) + \beta \rho_{-y_i}^{(t)}(i,p)\right) \geq 1.$$
- (Do not learn extremely rare features): For any $s, s' \in \{\pm 1\}$ and $k \in \mathcal{K}_R \cup \mathcal{K}_E$,
$$\gamma_{s'}^{(T_{\text{Cutout}})}(s,k) \leq \alpha^2.$$
- For any $s \in \{\pm 1\}, i \in [n]$, and $p \in [P] \setminus \{p_i^*\}$, $\rho_s^{(t)}(i,p) = \widetilde{\mathcal{O}}\left(\beta^{-1}\right)$.

G.2.4. CUTOUT CANNOT LEARN EXTREMELY RARE FEATURES WITHIN POLYNOMIAL TIMES

In this step, We will show that Cutout cannot learn extremely rare features within the maximum admissible iterate $T^* = \frac{\text{poly}(d)}{\eta}$.

we fix any $s^* \in \{\pm 1\}$ and $k^* \in \mathcal{K}_E$. Recall the function $Q^{(s^*,k^*)} : \mathcal{W} \to \mathbb{R}^{d \times 2}$, defined in Lemma 12 and omit supscripts for simplicity. For each iteration t, $Q(\boldsymbol{W}^{(t)})$ represents quantities updates by data with feature vector $\boldsymbol{v}_{s^*,k^*}$ until t-th iteration. We will sequentially introduce several technical lemmas and by combining these lemmas, quantify update by data with feature vector $\boldsymbol{v}_{s^*,k^*}$ after T_{Cutout} and derive our conclusion.

Let us define $\boldsymbol{W}^* = \{\boldsymbol{w}_1^*, \boldsymbol{w}_{-1}^*\}$, where

$$\boldsymbol{w}_1^* = \boldsymbol{w}_1^{(T_{\text{Cutout}})} + s^* M \sum_{i \in \mathcal{V}_{s^*,k^*}} \sum_{p \in [P] \setminus \{p_i^*\}} \frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2},$$
$$\boldsymbol{w}_{-1}^* = \boldsymbol{w}_{-1}^{(T_{\text{Cutout}})} - s^* M \sum_{i \in \mathcal{V}_{s^*,k^*}} \sum_{p \in [P] \setminus \{p_i^*\}} \frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2},$$

where $M = \beta^{-1} \log\left(\frac{2\eta T^*}{\alpha^2}\right) = \widetilde{\mathcal{O}}\left(\beta^{-1}\right)$.

Note that $\boldsymbol{W}^{(t)}, \boldsymbol{W}^* \in \mathcal{W}$ for any $t \geq 0$.

Lemma 27 *Suppose the event E_{init} occurs. Then,*

$$\left\|Q\left(\boldsymbol{W}^{(T_{\text{Cutout}})}\right) - Q(\boldsymbol{W}^*)\right\|^2 \leq 24 M^2 P |\mathcal{V}_{s^*,k^*}| \sigma_{\text{b}}^{-2} d^{-1}.$$

Proof [Proof of Lemma 27] For each $s \in \{\pm 1\}$,

$$\begin{aligned} &Q_s\left(\boldsymbol{w}_s^*\right) - Q_s\left(\boldsymbol{w}_s^{(T_{\text{Cutout}})}\right) \\ &= Mss^* \sum_{i \in \mathcal{V}_{s^*,k^*}} \sum_{p \in [P] \setminus \{p_i^*\}} \frac{\xi_i^{(p)}}{\left\|\xi_i^{(p)}\right\|^2} + \alpha M s^* \left(\sum_{i \in \mathcal{F}_s \cap \mathcal{V}_{s^*,k^*}} \frac{\boldsymbol{v}_{s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2} + \sum_{i \in \mathcal{F}_{-s} \cap \mathcal{V}_{s^*,k^*}} \frac{\boldsymbol{v}_{-s,1}}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2} \right), \end{aligned}$$

and we have

$$\begin{aligned}
&\left\| Q\left(\boldsymbol{W}^{(T_{\mathrm{Cutout}})}\right) - Q(\boldsymbol{W}^*)\right\|^2 \\
&\le 2M^2 \left(\sum_{i \in \mathcal{V}_{s^*,k^*}, p \in [P]\setminus\{p_i^*\}} \left\|\xi_i^{(p)}\right\|^{-2} + \sum_{\substack{i,j \in \mathcal{V}_{s^*,k^*} \\ p \in [P]\setminus\{p_i^*\}, q \in [P]\setminus\{p_j^*\} \\ (i,p)\neq(j,q)}} \frac{\left|\left\langle \xi_i^{(p)}, \xi_j^{(q)}\right\rangle\right|}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2 \left\|\xi_j^{(\tilde{p}_j)}\right\|^2} \right) \\
&\quad + 2M^2 \left(\alpha^2 \left(\sum_{i \in \mathcal{F}_s \cap \mathcal{V}_{s^*,k^*}} \left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2} \right)^2 + \alpha^2 \left(\sum_{i \in \mathcal{F}_{-s'} \cap \mathcal{V}_{s^*,k^*}} \left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2} \right)^2 \right).
\end{aligned}$$

From E_{init} and $nd^{-\frac{1}{2}} = o(1)$, we have

$$\sum_{\substack{i,j \in \mathcal{V}_{s^*,k^*} \\ p \in [P]\setminus\{p_i^*\}, q \in [P]\setminus\{p_j^*\} \\ (i,p)\neq(j,q)}} \frac{\left|\left\langle \xi_i^{(p)}, \xi_j^{(q)}\right\rangle\right|}{\left\|\xi_i^{(\tilde{p}_i)}\right\|^2 \left\|\xi_j^{(\tilde{p}_j)}\right\|^2} \le \sum_{i \in \mathcal{V}_{s^*,k^*}, p \in [P]\setminus\{p_i^*\}} \left\|\xi_i^{(p)}\right\|^{-2}$$

In addition, $\rho_k^* n = o(\sigma_{\mathrm{d}}^2 d)$ and $\alpha = o(1)$, we have

$$\begin{aligned}
&\alpha^2 \left(\sum_{i \in \mathcal{F}_s \cap \mathcal{V}_{s^*,k^*}} \left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2} \right)^2 + \alpha^2 \left(\sum_{i \in \mathcal{F}_{-s'} \cap \mathcal{V}_{s^*,k^*}} \left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2} \right)^2 \\
&\le \left(\sum_{i \in \mathcal{F}_s \cap \mathcal{V}_{s^*,k^*}} \left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2} \right)^2 + \left(\sum_{i \in \mathcal{F}_{-s'} \cap \mathcal{V}_{s^*,k^*}} \left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2} \right)^2 \\
&\le \sum_{i \in \mathcal{F}_s \cap \mathcal{V}_{s^*,k^*}} \left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2} + \sum_{i \in \mathcal{F}_{-s'} \cap \mathcal{V}_{s^*,k^*}} \left\|\xi_i^{(\tilde{p}_i)}\right\|^{-2} \\
&\le \sum_{i \in \mathcal{V}_{s^*,k^*}, p \in [P]\setminus\{p_i^*\}} \left\|\xi_i^{(p)}\right\|^{-2}.
\end{aligned}$$

Hence, from E_{init}, we obtain

$$\left\| Q\left(\boldsymbol{W}^{(T_{\mathrm{ERM}})}\right) - Q(\boldsymbol{W}^*)\right\|^2 \le 6M \sum_{i \in \mathcal{V}_{s^*,k^*}, p \in [P]\setminus\{p_i^*\}} \left\|\xi_i^{(p)}\right\|^{-2} \le 24M^2 P |\mathcal{V}_{s^*,k^*}| \sigma_{\mathrm{d}}^{-2} d^{-1}.$$

■

Lemma 28 *Suppose the E_{init} occurs. For any $t \ge T_{\mathrm{Cutout}}$, $i \in \mathcal{V}_{s,k}$ and any $\mathcal{C} \subset [P]$ with $|\mathcal{C}| = C$, it holds that*

$$\langle y_i \nabla_{\boldsymbol{W}} f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}), Q(W^*)\rangle \ge \frac{M\beta}{2}.$$

Proof [Proof of Lemma 28] From the calculation, we have

$$
\begin{aligned}
&\langle y_i \nabla_{\boldsymbol{W}} f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}), Q(\boldsymbol{W}^*)\rangle \\
&= \sum_{p \notin \mathcal{C}} \left(\phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \left\langle Q_s(\boldsymbol{w}_s^*), \boldsymbol{x}_i^{(p)} \right\rangle - \phi'\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \left\langle Q_{-s}(\boldsymbol{w}_{-s}^*), \boldsymbol{x}_i^{(p)} \right\rangle \right).
\end{aligned}
$$

Note that $n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}} = o(1)$ by Assumption 8. For any $s' \in \{\pm 1\}$ and $p \in [P] \setminus \{p_i^*, \tilde{p}_i\}$,

$$
\begin{aligned}
&ss^* \left\langle Q_s(\boldsymbol{w}_s^*), \xi_i^{(p)} \right\rangle \\
&\geq M + \rho_{s'}^{(T_{\text{Cutout}})}(i,p) + \sum_{\substack{j \in [n], q \in [P] \setminus \{p_j^*\} \\ (i,p) \neq (j,q)}} \rho_{s'}^{(T_{\text{Cutout}})}(j,q) \frac{\left|\left\langle \xi_i^{(p)}, \xi_j^{(q)} \right\rangle\right|}{\left\|\xi_j^{(q)}\right\|^2} \\
&\quad + M \sum_{\substack{j \in \mathcal{V}_{s^*,k^*}, q \in [P] \setminus \{p_j^*\} \\ (i,p) \neq (j,q)}} \frac{\left|\left\langle \xi_i^{(p)}, \xi_j^{(q)} \right\rangle\right|}{\left\|\xi_j^{(q)}\right\|^2} \\
&\geq -\widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right)
\end{aligned}
$$

Also, for any $s \in \{\pm 1\}$, $ss^* \langle Q_s(\boldsymbol{w}_s^*), \boldsymbol{v}_{s^*,k^*}\rangle = \gamma_s^{(T_{\text{Cutout}})}(s^*,k^*) \geq 0$. In addition,

$$
\begin{aligned}
&ss^* \left\langle Q_s(\boldsymbol{w}_s^*), \boldsymbol{x}_i^{(\tilde{p}_i)} \right\rangle \\
&= ss^* \left\langle Q_s(\boldsymbol{w}_s^*), \xi_i^{(\tilde{p}_i)} \right\rangle \\
&= M + \rho_s^{(T_{\text{Cutout}})}(i,\tilde{p}_i) + \sum_{\substack{j \in [n], p \in [P] \setminus \{p_i^*\} \\ (i,p) \neq (j,q)}} \rho_s^{(T_{\text{Cutout}})}(j,q) \frac{\left\langle \xi_i^{(p)}, \xi_j^{(q)} \right\rangle}{\left\|\xi_j^{(q)}\right\|^2} \\
&\quad + \sum_{\substack{j \in \mathcal{V}_{s^*,k^*}, p \in [P] \setminus \{p_i^*\} \\ (i,\tilde{p}_i) \neq (j,q)}} M \frac{\left\langle \xi_i^{(\tilde{p}_i)}, \xi_j^{(q)} \right\rangle}{\left\|\xi_j^{(q)}\right\|^2} \\
&\geq M - \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right) - \widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) \\
&\geq \frac{M}{2}.
\end{aligned}
$$

Hence, combining with $\phi' \geq \beta$, we have $\langle y_i \nabla_{\boldsymbol{W}} f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i), Q(W^*)\rangle \geq \frac{M\beta}{2}$. ■

By combining Lemma 27 and Lemma 28, we can obtain the following result.

Lemma 29 *Suppose the event E_{init} occurs.*

$$
\frac{\eta}{n} \sum_{t=T_{\text{Cutout}}}^{T^*} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right) \leq \left\| Q\left(\boldsymbol{W}^{(T_{\text{Cutout}})}\right) - Q(\boldsymbol{W}^*) \right\|^2 + 2\eta T^* e^{-M\beta}
$$

Proof [Proof of Lemma 29] Note that for any $T_{\text{Cutout}} \leq t < T^*$,

$$Q\left(\boldsymbol{W}^{(t+1)}\right) = Q\left(\boldsymbol{W}^{(t)}\right) - \frac{\eta}{n} \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell \left(y_i f_{\boldsymbol{W}^{(t)}} (\boldsymbol{X}_{i,\mathcal{C}}) \right) \right],$$

and thus

$$\begin{aligned}
&\left\| Q\left(\boldsymbol{W}^{(t)}\right) - Q\left(\boldsymbol{W}^*\right) \right\|^2 - \left\| Q\left(\boldsymbol{W}^{(t+1)}\right) - Q\left(\boldsymbol{W}^*\right) \right\|^2 \\
&= \frac{2\eta}{n} \left\langle \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell \left(y_i f_{\boldsymbol{W}^{(t)}} (\boldsymbol{X}_{i,\mathcal{C}}) \right) \right], Q\left(\boldsymbol{W}^{(t)}\right) - Q\left(\boldsymbol{W}^*\right) \right\rangle \\
&\quad - \frac{\eta^2}{n^2} \left\| \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell \left(y_i f_{\boldsymbol{W}^{(t)}} (\boldsymbol{X}_{i,\mathcal{C}}) \right) \right] \right\|^2 \\
&= \frac{2\eta}{n} \left\langle \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell(y_i f_{\boldsymbol{W}^{(t)}} (\boldsymbol{X}_{i,\mathcal{C}})), Q\left(\boldsymbol{W}^{(t)}\right) \right\rangle \right] \\
&\quad - \frac{2\eta}{n} \sum_{i \in \mathcal{V}_{s^*,k^*}} \left\langle \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell'(y_i f_{\boldsymbol{W}^{(t)}} (\boldsymbol{X}_{i,\mathcal{C}})) \nabla_{\boldsymbol{W}} y_i f_{\boldsymbol{W}^{(t)}} (\boldsymbol{X}_{i,\mathcal{C}}) \right], Q\left(\boldsymbol{W}^*\right) \right\rangle \\
&\quad - \frac{\eta^2}{n^2} \left\| \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell \left(y_i f_{\boldsymbol{W}^{(t)}} (\boldsymbol{X}_{i,\mathcal{C}}) \right) \right] \right\|^2 \\
&\geq \frac{2\eta}{n} \left\langle \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell(y_i f_{\boldsymbol{W}^{(t)}} (\boldsymbol{X}_{i,\mathcal{C}})) \right], Q\left(\boldsymbol{W}^{(t)}\right) \right\rangle \\
&\quad - \frac{M\beta\eta}{n} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell'(y_i f_{\boldsymbol{W}^{(t)}} (\boldsymbol{X}_{i,\mathcal{C}})) - \frac{\eta^2}{n^2} \left\| \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell \left(y_i f_{\boldsymbol{W}^{(t)}} (\boldsymbol{X}_{i,\mathcal{C}}) \right) \right] \right\|^2,
\end{aligned}$$

where the last inequality is due to Lemma 28. By the chain rule, for each $\mathcal{C} \subset [P]$ with $|\mathcal{C}| = C$,

$$\begin{aligned}
&\left\langle \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \ell(y_i f_{\boldsymbol{W}^{(t)}} (\boldsymbol{X}_{i,\mathcal{C}})), Q\left(\boldsymbol{W}^{(t)}\right) \right\rangle \\
&= \sum_{i \in \mathcal{V}_{s^*,k^*}} \Bigg[\ell'(y_i f_{\boldsymbol{W}^{(t)}} (\boldsymbol{X}_{i,\mathcal{C}})) \\
&\quad \times \sum_{p \notin \mathcal{C}} \left(\phi' \left(\left\langle \boldsymbol{w}_{s^*}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right) \left\langle Q_{s^*} \left(\boldsymbol{w}_{s^*}^{(t)} \right), \boldsymbol{x}_i^{(p)} \right\rangle - \phi' \left(\left\langle \boldsymbol{w}_{-s^*}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right) \left\langle Q_{-s^*} \left(\boldsymbol{w}_{-s^*}^{(t)} \right), \boldsymbol{x}_i^{(p)} \right\rangle \right) \Bigg].
\end{aligned}$$

For each $s \in \{\pm 1\}$ and $i \in \mathcal{V}_{s^*,k^*}$,

$$\begin{aligned}
&\left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle - \left\langle Q_s\left(\boldsymbol{w}_s^{(t)}\right), \boldsymbol{x}_i^{(p)} \right\rangle\right| \\
&= \left|\left\langle \boldsymbol{w}_s^{(t)} - Q_s\left(\boldsymbol{w}_s^{(t)}\right), \boldsymbol{x}_i^{(p)} \right\rangle\right| \\
&\leq \sum_{i \in [n] \setminus \mathcal{V}_{s^*,k^*}, p \in [P] \setminus \{p_i^*\}} \left|\left\langle \rho_s^{(t)}(j,q) \frac{\xi_j^{(q)}}{\left\|\xi_j^{(q)}\right\|^2}, \boldsymbol{x}_i^{(p)} \right\rangle\right| \\
&\quad \alpha \sum_{s \in \{\pm 1\}, j \in \mathcal{F}_{s'} \setminus \mathcal{V}_{s^*,k^*}} \rho_s^{(t)}(j, \tilde{p}_j) \left\|\xi_j^{(\tilde{p}_j)}\right\|^{-2} \left|\left\langle \boldsymbol{v}_{s',1}, \boldsymbol{x}_i^{(p)} \right\rangle\right| \\
&\leq \widetilde{\mathcal{O}}\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right) + \widetilde{\mathcal{O}}\left(\alpha \beta^{-1} \rho_{k^*} n \sigma_{\mathrm{d}}^{-2} d^{-1}\right) \\
&= \widetilde{\mathcal{O}}\left(\alpha \beta^{-1}\right),
\end{aligned}$$

where the last inequality is due to Lemma 22 and the event E_{init}. By Lemma 34,

$$\begin{aligned}
&\sum_{p \notin \mathcal{C}} \left(\phi'\left(\left\langle \boldsymbol{w}_{s^*}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \left\langle Q_{s^*}\left(\boldsymbol{w}_{s^*}^{(t)}\right), \boldsymbol{x}_i^{(p)} \right\rangle - \phi'\left(\left\langle \boldsymbol{w}_{-s^*}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \left\langle Q_{-s^*}\left(\boldsymbol{w}_{s^*}^{(t)}\right), \boldsymbol{x}_i^{(p)} \right\rangle \right) \\
&\geq \sum_{p \notin \mathcal{C}} \left(\phi\left(\left\langle \boldsymbol{w}_{s^*}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s^*}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \right) - rP - \widetilde{\mathcal{O}}\left(\alpha \beta^{-1}\right) \\
&= y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}) - \widetilde{\mathcal{O}}\left(\alpha \beta^{-1}\right),
\end{aligned}$$

where the last equality is due to $r = o\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right)$. Therefore, we have

$$\begin{aligned}
&\left\|Q\left(\boldsymbol{W}^{(t)}\right) - Q\left(\boldsymbol{W}^*\right)\right\|^2 - \left\|Q\left(\boldsymbol{W}^{(t+1)}\right) - Q(\boldsymbol{W}^*)\right\|^2 \\
&\geq \frac{2\eta}{n} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell'\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})\right) \left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}) - \widetilde{\mathcal{O}}\left(\alpha \beta^{-1}\right) - \frac{M\beta}{2} \right) \right] \\
&\quad - \frac{\eta^2}{n^2} \left\| \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})) \right] \right\|^2 \\
&\geq \frac{2\eta}{n} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell'(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})) \left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}) - M\beta \right) \right] \\
&\quad - \frac{\eta^2}{n^2} \left\| \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})) \right] \right\|^2.
\end{aligned}$$

From the convexity of $\ell(\cdot)$,

$$\begin{aligned}
&\sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}\left[\ell'(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i))\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i) - M\beta\right)\right]\\
&\geq \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}\left[(\ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}) - \ell\left(M\beta\right))\right]\\
&\geq \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}\left[\ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}))\right] - ne^{-M\beta}.
\end{aligned}$$

In addition, by Lemma 36,

$$\begin{aligned}
&\frac{\eta^2}{n^2}\left\|\nabla \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}[\ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})\right)]\right\|^2\\
&\leq \frac{8\eta^2 P^2 \sigma_\mathrm{d}^2 d |\mathcal{V}_{s,k}|}{n^2} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}[\ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}))]\\
&\leq \frac{\eta}{n} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}[\ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}))],
\end{aligned}$$

from our choice of small enough η, and we have

$$\begin{aligned}
&\left\|Q\left(\boldsymbol{W}^{(t)}\right) - Q(\boldsymbol{W}^*)\right\|^2 - \left\|Q\left(\boldsymbol{W}^{(t+1)}\right) - Q(\boldsymbol{W}^*)\right\|^2\\
&\geq \frac{\eta}{n} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}[\ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}))] - 2\eta e^{-M\beta}.
\end{aligned}$$

From telescoping summation, we have

$$\frac{\eta}{n} \sum_{t=T_\text{Cutout}}^{T^*} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}[|\ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i)\right)] \leq \left\|Q\left(\boldsymbol{W}^{(T_\text{Cutout})}\right) - Q\left(\boldsymbol{W}^*\right)\right\|^2 + 2\eta T^* e^{-M\beta}.$$

■

Finally, we can prove that the model cannot learn extremely rare features within T^* iterations.

Lemma 30 *Suppose the event E_init occurs. For any $T \in [T_\text{Cutout}, T^*]$, we have $\gamma_s^{(T)}(s^*, k^*) \leq 3\alpha^2$ for each $s \in \{\pm 1\}$.*

Proof [Proof of Lemma 30] For any $T \in [T_\text{Cutout}, T^*]$, we have

$$\begin{aligned}
\gamma_s^{(T)}(s^*, k^*) &= \gamma_s^{(T_\text{Cutout})}(s^*, k^*) + \frac{\eta}{n} \sum_{t=T_\text{Cutout}}^{T-1} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}\left[g_{i,\mathcal{C}}^{(t)}\right] \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s^*,k^*}\right\rangle\right)\\
&\leq \gamma_s^{(T_\text{Cutout})}(s^*, k^*) + \frac{\eta}{n} \sum_{t=T_\text{Cutout}}^{T-1} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}\left[g_{i,\mathcal{C}}^{(t)}\right]\\
&\leq \gamma_s^{(T_\text{Cutout})}(s^*, k^*) + \frac{\eta}{n} \sum_{t=T_\text{Cutout}}^{T-1} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C}\sim\mathcal{D}_\mathcal{C}}\left[\ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})\right)\right],
\end{aligned}$$

where the first inequality is due to $\phi' \leq 1$ and the second inequality is due to $-\ell' \leq \ell$. From the result of Section G.2.3, $\gamma_s^{(T_{\text{Cutout}})}(s^*, k^*) \leq \alpha^2$ and by Lemma 29 and Lemma 27, we have

$$\begin{aligned}
\frac{\eta}{n} \sum_{t=T_{\text{Cutout}}}^{(T-1)} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}}[\ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})\right)] &\leq \frac{\eta}{n} \sum_{t=T_{\text{Cutout}}}^{(T^*)} \sum_{i \in \mathcal{V}_{s^*,k^*}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}}[\ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})\right)] \\
&\leq \left\| Q\left(\boldsymbol{W}^{(T_{\text{Cutout}}}\right) - Q(\boldsymbol{W}^*) \right\|^2 + 2\eta T^* e^{-M\beta} \\
&\leq 24 M^2 P |\mathcal{V}_{s^*,k^*}| \sigma_{\text{b}}^{-2} d^{-1} + 2\eta T^* e^{-M\beta} \\
&\leq \alpha^2 + 2\eta T^* e^{-M\beta}.
\end{aligned}$$

The last inequality is due to $\rho_E n = o(\alpha^2 \beta^2 \sigma_{\text{b}}^2 d)$ by Assumption 8. Since $M = \beta^{-1} \log\left(\frac{2\eta T^*}{\alpha^2}\right)$, we have our conclusion. ■

What We Have So Far. Suppose the event E_{init} occurs. For any $t \in [T_{\text{Cutout}}, T^*]$, we have

- (Learn common features): $\gamma_s^{(t)}(s,k) + \beta \gamma_{-s}^{(t)}(s,k) = \Omega(1)$ for each $s \in \{\pm 1\}$ and $k \in \mathcal{K}_C$
- (Overfit rare/extreme data): $\rho_s^{(t)}(i, \tilde{p}_i) + \beta \rho_{-s}^{(t)}(i, \tilde{p}_i) = \Omega(1)$ for each $s \in \{\pm 1\}, k \in \mathcal{K}_R \cup \mathcal{K}_C$ and $i \in \mathcal{V}_{s,k}$.
- (Cannot learn rare/extreme features): $\gamma_s^{(t)}(s,k), \gamma_{-s}^{(t)}(s,k) = \mathcal{O}(\alpha^2)$ for each $s \in \{\pm 1\}$ and $k \in \mathcal{K}_R \cup \mathcal{K}_E$.
- For any $s \in \{\pm 1\}, i \in [n]$, and $p \in [P] \setminus \{p_i^*\}$, $\rho_s^{(t)}(i,p) = \widetilde{\mathcal{O}}\left(\beta^{-1}\right)$,

G.2.5. TRAIN AND TEST ACCURACY

In this step, we will prove that the model trained by Cutout has perfect training accuracy on both augmented data and original data but has near-random guesses on test data with extremely rare data.

For any $i \in \mathcal{V}_{s,k}$ with $s \in \{\pm 1\}$ $k \in \mathcal{K}_C \cup \mathcal{K}_\mathcal{R}$ and $\mathcal{C} \subset [P]$ with $|\mathcal{C}| = C$ and $p_i^* \notin \mathcal{C}$,

$$\begin{aligned}
& y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}) \\
&= \sum_{p \notin \mathcal{C}} \left(\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right) \right) \\
&= \gamma_s^{(t)}(s,k) + \beta \gamma_{-s}^{(t)}(s,k) + \sum_{p \notin \mathcal{C} \cup \{p_i^*\}} \left(\rho_s^{(t)}(i,p) + \beta \rho_{-s}^{(t)}(i,p) \right) \\
&\quad - \widetilde{\mathcal{O}}\left(n\beta^{-1} \sigma_{\text{d}} \sigma_{\text{b}}^{-1} d^{-\frac{1}{2}} \right) - \widetilde{\mathcal{O}}(\alpha \beta^{-1}) \\
&\geq \gamma_s^{(t)}(s,k) + \beta \gamma_{-s}^{(t)}(s,k) - \widetilde{\mathcal{O}}\left(n\beta^{-1} \sigma_{\text{d}} \sigma_{\text{b}}^{-1} d^{-\frac{1}{2}} \right) - \widetilde{\mathcal{O}}(\alpha \beta^{-1}) \\
&= \Omega(1) - \widetilde{\mathcal{O}}\left(n\beta^{-1} \sigma_{\text{d}} \sigma_{\text{b}}^{-1} d^{-\frac{1}{2}} \right) - \widetilde{\mathcal{O}}(\alpha \beta^{-1}) \\
&= \Omega(1),
\end{aligned}$$

for any $t \in [T_{\text{Cutout}}, T^*]$. In addition, for any $i \in [n]$ and $\mathcal{C} \subset [P]$ with $|\mathcal{C}| = C$ does not correspond to the case above, by Lemma 26 and Lemma 11, we have

$$\begin{aligned}
&y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}) \\
&= \sum_{p \notin \mathcal{C}} \left(\phi\left(\left\langle \boldsymbol{w}_{y_i}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right) - \phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right) \right) \\
&\geq \sum_{p \notin \mathcal{C} \cup \{p_i^*\}} \left(\rho_{y_i}^{(t)}(i,p) + \beta \rho_{-y_i}^{(t)}(i,p) \right) - \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right) - \widetilde{\mathcal{O}}(\alpha\beta^{-1}) \\
&= \Omega(1) - \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right) - \widetilde{\mathcal{O}}(\alpha\beta^{-1}) \\
&= \Omega(1),
\end{aligned}$$

for any $t \in [T_{\text{Cutout}}, T^*]$. We can conclude that Cutout with $t \in [T_{\text{Cutout}}, T^*]$ iterates achieve perfect training accuracy on augmented data.

Next, we will show that Cutout achieves perfect training accuracy on the original data. For any $i \in [n]$, let us choose $\mathcal{C} \subset [P]$ with $|\mathcal{C}| = C$ such that $p_i^* \in \mathcal{C}$. Then, from the result above, we have

$$\begin{aligned}
y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_i) &= y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}) + \sum_{p \in \mathcal{C}} \left(\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right)) - \phi\left(\left\langle \boldsymbol{w}_S^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle \right)) \right) \\
&\geq y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}) + \sum_{p \in \mathcal{C} \setminus \{p_i^*\}} \left(\rho_{y_i}^{(t)}(i,p) + \beta \rho_{-y_i}^{(t)}(i,p) \right) \\
&\quad - \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right) - \widetilde{\mathcal{O}}(\alpha\beta^{-1}) \\
&\geq \Omega(1),
\end{aligned}$$

for any $t \in [T_{\text{Cutout}}, T^*]$ and we have our conclusion.

Lastly, let us move on to the test accuracy part. Let $(\boldsymbol{X}, y) \sim \mathcal{D}$ be a test data with $\boldsymbol{X} = \left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(P)} \right) \in \mathbb{R}^{d \times P}$ having feature patch p^*, dominant noise patch $\tilde{p}$, and feature vector $\boldsymbol{v}_{y,k}$. We have $\boldsymbol{x}^{(p)} \sim N(\boldsymbol{0}, \sigma_{\mathrm{b}}^2 \boldsymbol{\Lambda})$ for each $p \in [P] \setminus \{p^*, \tilde{p}\}$ and $\boldsymbol{x}^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1} \sim N(\boldsymbol{0}, \sigma_{\mathrm{d}}^2 \boldsymbol{\Lambda})$ for some $s \in \{\pm 1\}$. Therefore, for all $t \in [T_{\text{Cutout}}, T^*]$ and $p \in [P] \setminus \{p^*, \tilde{p}\}$,

$$\begin{aligned}
&\left| \phi\left(\left\langle \boldsymbol{w}_1^{(t)}, \boldsymbol{x}^{(p)} \right\rangle \right) - \phi\left(\left\langle \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(p)} \right\rangle \right) \right| \\
&\leq \left| \left\langle \boldsymbol{w}_1^{(t)} - \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(p)} \right\rangle \right| \\
&= \left| \left\langle \boldsymbol{w}_1^{(0)} - \boldsymbol{w}_{-1}^{(0)}, \boldsymbol{x}^{(p)} \right\rangle \right| + \sum_{i \in [n], q \in [P] \setminus \{p_i^*\}} \rho(i,q) \frac{\left| \left\langle \xi_i^{(q)}, \boldsymbol{x}^{(p)} \right\rangle \right|}{\left\| \xi_i^{(q)} \right\|^2} \\
&\leq \widetilde{\mathcal{O}}\left(\sigma_0 \sigma_{\mathrm{b}} d^{\frac{1}{2}} \right) + \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right) \\
&= \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right),
\end{aligned} \tag{13}$$

with probability at least $1 - o\left(\frac{1}{\mathrm{poly}(d)}\right)$ due to Lemma 9. Similarly, we have

$$\begin{aligned}
&\left|\phi\left(\left\langle \boldsymbol{w}_1^{(t)}, \boldsymbol{x}^{(p)} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(p)} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right)\right| \\
&\leq \left|\left\langle \boldsymbol{w}_1^{(t)} - \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right| = \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right),
\end{aligned} \tag{14}$$

with probability at least $1 - o\left(\frac{1}{\mathrm{poly}(d)}\right)$.

Case 1: $k \in \mathcal{K}_C \cup \mathcal{K}_R$

By Lemma 9

$$\begin{aligned}
&\left|\phi\left(\left\langle \boldsymbol{w}_1^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right)\right| \\
&\leq \left|\left\langle \boldsymbol{w}_1^{(t)} - \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right| \\
&\leq \alpha\left|\left\langle \boldsymbol{w}_1^{(t)} - \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{v}_{s,1}\right\rangle\right| + \left|\left\langle \boldsymbol{w}_1^{(t)} - \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{x}^{(p)} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right| \\
&\leq \alpha\beta^{-1}\left|\phi\left(\left\langle \boldsymbol{w}_1^{(t)}, \boldsymbol{v}_{s,1}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-1}^{(t)}, \boldsymbol{v}_{s,1}\right\rangle\right)\right| \\
&\quad + \left|\left\langle \boldsymbol{w}_1^{(0)} - \boldsymbol{w}_{-1}^{(0)}, \boldsymbol{x}^{(p)} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right| + \sum_{i \in [n], q \in [P] \setminus \{p_i^*\}} \rho(i,q) \frac{\left|\left\langle \xi_i^{(q)}, \boldsymbol{x}^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right|}{\left\|\xi_i^{(q)}\right\|^2} \\
&\leq \widetilde{\mathcal{O}}\left(\alpha\beta^{-2}\right) + \widetilde{\mathcal{O}}\left(\sigma_0\sigma_{\mathrm{d}}d^{\frac{1}{2}}\right) + \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right) \\
&= \widetilde{\mathcal{O}}\left(\alpha\beta^{-2}\right),
\end{aligned} \tag{15}$$

with probability at least $1 - o\left(\frac{1}{\mathrm{poly}(d)}\right)$. Suppose (13) and (15) holds. By Lemma 11, we have

$$\begin{aligned}
&y f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}) \\
&= \left(\phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{v}_{y,k}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{v}_{y,k}\right\rangle\right)\right) \\
&\quad + \sum_{p \in [P] \setminus \{p^*\}} \left(\phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right) \\
&= \gamma_y^{(t)}(y,k) + \beta\gamma_{-y}^{(t)}(y,k) - \widetilde{\mathcal{O}}(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}) - \widetilde{\mathcal{O}}(\alpha\beta^{-1}) \\
&= \Omega(1) - \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right) - \widetilde{\mathcal{O}}\left(\alpha\beta^{-1}\right) \\
&> 0.
\end{aligned}$$

Therefore, we have

$$\mathbb{P}_{(\boldsymbol{X},y)\sim\mathcal{D}}\left[y f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}) > 0 \mid \boldsymbol{x}^{(p^*)} = \boldsymbol{v}_{y,k}, k \in \mathcal{K}_C \cup \mathcal{K}_R\right] \geq 1 - o\left(\frac{1}{\mathrm{poly}(d)}\right). \tag{16}$$

Case 2: $k \in \mathcal{K}_E$

By triangular inequality and $\phi' \leq 1$, we have

$$\begin{aligned}
&\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right)\\
&= \phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{s,1}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \alpha \boldsymbol{v}_{s,1}\right\rangle\right)\\
&\quad + \left(\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(\tilde{p})}\right\rangle\right) - \phi\left(\gamma_s^{(t)}(s,1)\right)\right) - \left(\phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(\tilde{p})}\right\rangle\right) - \phi\left(\gamma_{-s}^{(t)}(s,1)\right)\right)\\
&\geq \phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{s,1}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \alpha \boldsymbol{v}_{s,1}\right\rangle\right)\\
&\quad - \alpha\left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,1}\right\rangle - \gamma_s^{(t)}(s,1)\right| - \alpha\left|\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,1}\right\rangle + \gamma_{-s}^{(t)}(s,1)\right|\\
&\quad - \left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right| - \left|\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{x}_i^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right|.
\end{aligned}$$

In addition,

$$\begin{aligned}
&\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{s,1}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \alpha \boldsymbol{v}_{s,1}\right\rangle\right)\\
&= \left(\phi\left(\alpha\gamma_s^{(t)}(s,1)\right) - \phi\left(\alpha\gamma_{-s}^{(t)}(s,1)\right)\right)\\
&\quad + \left(\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{s,1}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha\gamma_s^{(t)}(s,1)\right\rangle\right)\right)\\
&\quad - \left(\phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \alpha \boldsymbol{v}_{s,1}\right\rangle\right) - \phi\left(\alpha\gamma_{-s}^{(t)}(s,1)\right)\right)\\
&\geq \left(\phi\left(\alpha\gamma_s^{(t)}(s,1)\right) - \phi\left(\alpha\gamma_{-s}^{(t)}(s,1)\right)\right)\\
&\quad - \alpha\left|\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,1}\right\rangle - \gamma_s^{(t)}(s,1)\right| - \alpha\left|\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{v}_{s,1}\right\rangle + \gamma_{-s}^{(t)}(s,1)\right|.
\end{aligned}$$

If (14) holds,then by Lemma 11, we have

$$\begin{aligned}
&\phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \alpha \boldsymbol{v}_{s,1}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \alpha \boldsymbol{v}_{s,1}\right\rangle\right)\\
&\geq \alpha\left(\gamma_s^{(t)}(s,1) + \beta\gamma_{-s}^{(t)}(s,1)\right) - \mathcal{O}\left(\alpha n \beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right) - \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right)\\
&\geq \Omega(\alpha),
\end{aligned}$$

where the last inequality is due to As a result,

$$\begin{aligned}
&y f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X})\\
&= \phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{v}_{y,k}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{v}_{y,k}\right\rangle\right) + \phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right)\\
&\quad + \sum_{p\in[P]\setminus\{p^*,\tilde{p}\}}\left(\phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{x}^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{x}^{(p)}\right\rangle\right)\right).
\end{aligned}$$

Note that

$$
\begin{aligned}
&\left|\phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{v}_{y,k}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{v}_{y,k}\right\rangle\right)\right| \\
&\quad + \left|\sum_{p \in [P] \setminus \{p^*, \tilde{p}\}} \left(\phi\left(\left\langle \boldsymbol{w}_y^{(t)}, \boldsymbol{x}^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(t)}, \boldsymbol{x}^{(p)}\right\rangle\right)\right)\right| \\
&\leq \mathcal{O}(\alpha^2) + \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1}d^{-\frac{1}{2}}\right) \\
&< \phi\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(t)}, \boldsymbol{w}^{(\tilde{p})}\right\rangle\right).
\end{aligned}
$$

□

Appendix H. Proof for CutMix

H.1. Proof of Lemma 10 for CutMix

For each $i, j \in [n]$ and $\mathcal{S} \subset [P]$, let

$$g_{i,j,\mathcal{S}}^{(t)} := -\frac{|\mathcal{S}|}{P} y_i \ell'\big(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,j,\mathcal{S}})\big) - \left(1 - \frac{|\mathcal{S}|}{P}\right) y_j \ell'\big(y_j f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,j,\mathcal{S}})\big).$$

For $s \in \{\pm 1\}$ and iterate t,

$$\begin{aligned}
&\boldsymbol{w}_s^{(t+1)} - \boldsymbol{w}_s^{(t)} \\
&= -\eta \nabla_{\boldsymbol{w}_s} \mathcal{L}_{\text{CutMix}}\left(\boldsymbol{W}^{(t)}\right) \\
&= \frac{\eta}{n^2} \sum_{i,j \in [n]} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}} \left[s g_{i,j,\mathcal{S}}^{(t)} \left(\sum_{p \in \mathcal{S}} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)} + \sum_{p \notin \mathcal{S}} \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_j^{(p)} \right) \right] \\
&= \frac{\eta}{n^2} \sum_{k \in [K]} \sum_{i \in \mathcal{V}_{s,k}, j \in [n]} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}} \left[g_{i,j,\mathcal{S}}^{(t)} \mathbb{1}_{p_i^* \in \mathcal{S}} + g_{j,i,\mathcal{S}}^{(t)} \mathbb{1}_{p_i^* \notin \mathcal{S}} \right] \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s,k} \right\rangle\right) \boldsymbol{v}_{s,k} \\
&\quad - \frac{\eta}{n^2} \sum_{k \in [K]} \sum_{i \in \mathcal{V}_{-s,k}, j \in [n]} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}} \left[g_{i,j,\mathcal{S}}^{(t)} \mathbb{1}_{p_i^* \in \mathcal{S}} + g_{j,i,\mathcal{S}}^{(t)} \mathbb{1}_{p_i^* \notin \mathcal{S}} \right] \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{-s,k} \right\rangle\right) \boldsymbol{v}_{-s,k} \\
&\quad + \frac{s\eta}{n^2} \sum_{i,j \in [n], p \in [P] \setminus \{p_i^*\}} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}} \left[g_{i,j,\mathcal{S}}^{(t)} \mathbb{1}_{p \in \mathcal{S}} + g_{j,i,\mathcal{S}}^{(t)} \mathbb{1}_{p \notin \mathcal{S}} \right] \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)}.
\end{aligned}$$

Hence, if we define $\gamma_s^{(t)}(s', k)$'s and $\rho_s^{(t)}(i, p)$'s recursively by using the rule

$$\gamma_s^{(t+1)}(s', k) = \gamma_s^{(t)}(s', k) + \frac{\eta}{n^2} \sum_{i \in \mathcal{V}_{s',k}, j \in [n]} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}} \left[g_{i,j,\mathcal{S}}^{(t)} \mathbb{1}_{p_i^* \in \mathcal{S}} + g_{j,i,\mathcal{S}}^{(t)} \mathbb{1}_{p_i^* \notin \mathcal{S}} \right] \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{v}_{s',k} \right\rangle\right),$$

$$\rho_s^{(t+1)}(i, p) = \rho_s^{(t)}(i, p) + \frac{s y_i \eta}{n^2} \sum_{j \in [n]} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}} \left[g_{i,j,\mathcal{S}}^{(t)} \mathbb{1}_{p \in \mathcal{S}} + g_{j,i,\mathcal{S}}^{(t)} \mathbb{1}_{p \notin \mathcal{S}} \right] \phi'\left(\left\langle \boldsymbol{w}_s^{(t)}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \left\| \xi_i^{(p)} \right\|^2,$$

starting from $\gamma_s^{(0)}(s'k) = \rho_s^{(0)}(i, p) = 0$ for each $s, s' \in \{\pm 1\}, k \in [K], i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$, then we have

$$\begin{aligned}
\boldsymbol{w}_s^{(t)} = \boldsymbol{w}_s^{(0)} &+ \sum_{k \in [K]} \gamma_s^{(t)}(s, k) \boldsymbol{v}_{s,k} - \sum_{k \in [K]} \gamma_s^{(t)}(-s, k) \boldsymbol{v}_{-s,k} \\
&+ \sum_{\substack{i \in [n], y_i = s \\ p \in [P] \setminus \{\tilde{p}_i\}}} \rho_s^{(t)}(i, p) \frac{\xi_i^{(p)}}{\left\| \xi_i^{(p)} \right\|^2} - \sum_{\substack{i \in [n], y_i = -s \\ p \in [P] \setminus \{\tilde{p}_i\}}} \rho_s^{(t)}(i, p) \frac{\xi_i^{(p)}}{\left\| \xi_i^{(p)} \right\|^2} \\
&+ \alpha \left(\sum_{i \in \mathcal{F}_s} s y_i \rho_s^{(t)}(i, \tilde{p}_i) \frac{\boldsymbol{v}_{s,1}}{\left\| \xi_i^{(\tilde{p}_i)} \right\|^2} + \sum_{i \in \mathcal{F}_{-s}} s y_i \rho_{-s}^{(t)}(i, \tilde{p}_i) \frac{\boldsymbol{v}_{-s,1}}{\left\| \xi_i^{(\tilde{p}_i)} \right\|^2} \right),
\end{aligned}$$

for each $s \in \{\pm 1\}$. □

H.2. Proof of Theorem 5

We will prove that the conclusion of Theorem 5 holds when the event E_{init} occurs. The proof of Theorem 5 is structured into the following six steps:

1. Introduce a reparametrization of the CutMix loss $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$ to a convex function $h(\boldsymbol{Z})$ for ease of analysis (Section H.2.1).
2. Characterize a global minimum of $h(\boldsymbol{Z})$ (Section H.2.2).
3. Evaluate strong convexity constant in the region near the global minimum of $h(\boldsymbol{Z})$ (Section H.2.3).
4. Show that near stationary point of $h(\boldsymbol{Z})$ is close to a global minimum (Section H.2.4).
5. Prove that gradient descent on the CutMix loss $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$ achieves a near-stationary point of the reparametrized function $h(\boldsymbol{Z})$ and perfect accuracy on original training data (Section H.2.5).
6. Evaluate the test accuracy of a model in near-stationary point (Section H.2.6).

H.2.1. REPARAMETRIZATION OF CUTMIX LOSS $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$

It is complicated to characterize the stationary points of CutMix loss $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$ due to its non-convexity. We will overcome this problem by introducing reparameterization of the objective function. Let us define

$$z_i^{(p)} := \phi\left(\left\langle \boldsymbol{w}_1, \boldsymbol{x}_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-1}, \boldsymbol{x}_i^{(p)}\right\rangle\right),$$

for $i \in [n], p \in [P]$ and

$$z_{s,k} := \phi(\langle \boldsymbol{w}_1, \boldsymbol{v}_{s,k}\rangle) - \phi(\langle \boldsymbol{w}_{-1}, \boldsymbol{v}_{s,k}\rangle),$$

for each $s \in \{\pm 1\}, k \in [K]$. We can rewrite CutMix loss $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$ as a function $h(\boldsymbol{Z})$ of the defined variables $\boldsymbol{Z} := \{z_{s,k}\}_{s\in\{\pm1\},k\in[K]} \cup \{z_i^{(p)}\}_{i\in[n],p\in[P]\setminus\{p_i^*\}}$ as follows.

$$\begin{aligned} h(\boldsymbol{Z}) := \frac{1}{n^2}\sum_{i,j\in[n]} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}\Bigg[&\frac{|\mathcal{S}|}{P}\ell\left(y_i\left(\sum_{p\in\mathcal{S}} z_i^{(p)} + \sum_{p\notin\mathcal{S}} z_j^{(p)}\right)\right) \\ &+ \left(1 - \frac{|\mathcal{S}|}{P}\right)\ell\left(y_j\left(\sum_{p\in\mathcal{S}} z_i^{(p)} + \sum_{p\notin\mathcal{S}} z_j^{(p)}\right)\right)\Bigg], \end{aligned}$$

where we write $z_i^{(p_i^*)} = z_{s,k}$ if $i \in \mathcal{V}_{s,k}$. For notational simplicity, let us consider $\boldsymbol{Z}$ as vectors in $\mathbb{R}^{2K+n(P-1)}$ with the standard orthonormal basis $\{\boldsymbol{e}_{s,k}\}_{s\in\{\pm1\},k\in[K]} \cup \left\{\boldsymbol{e}_i^{(p)}\right\}_{i\in[n],p\in[P]\setminus\{p_i^*\}}$ which means

$$\begin{aligned} \boldsymbol{Z} &= \{z_{s,k}\}_{s\in\{\pm1\},k\in[K]} \cup \left\{z_i^{(p)}\right\}_{i\in[n],p\in[P]\setminus\{p_i^*\}} \\ &= \sum_{s\in\{\pm1\},k\in[K]} z_{s,k}\boldsymbol{e}_{s,k} + \sum_{i\in[n],p\in[P]\setminus\{p_i^*\}} z_i^{(p)}\boldsymbol{e}_i^{(p)}. \end{aligned}$$

If there is no confusion, we will use $\boldsymbol{e}_i^{(p_i^*)}$ to represent $\boldsymbol{e}_{s,k}$, for $i \in \mathcal{V}_{s,k}$.

By the chain rule,

$$\nabla_{\boldsymbol{W}} \mathcal{L}_{\text{CutMix}}(\boldsymbol{W}) = \boldsymbol{J}(\boldsymbol{W}) \nabla_{\boldsymbol{Z}} h(\boldsymbol{Z}),$$

where each column of Jacobian matrix $\boldsymbol{J}(\boldsymbol{W}) \in \mathbb{R}^{2d \times (n(P-1)+2K)}$ is

$$\nabla_{\boldsymbol{W}} z_{s,k} = \begin{pmatrix} \phi'(\langle \boldsymbol{w}_1, \boldsymbol{v}_{s,k} \rangle) \boldsymbol{v}_{s,k} \\ -\phi'(\langle \boldsymbol{w}_{-1}, \boldsymbol{v}_{s,k} \rangle) \boldsymbol{v}_{s,k} \end{pmatrix} \in \mathbb{R}^{2d}, \nabla_{\boldsymbol{W}} z_i^{(p)} = \begin{pmatrix} \phi'\left(\left\langle \boldsymbol{w}_1, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)} \\ -\phi'\left(\left\langle \boldsymbol{w}_{-1}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)} \end{pmatrix} \in \mathbb{R}^{2d}.$$

Let us characterize the smallest singular value $\sigma_{\min}(\boldsymbol{J}(\boldsymbol{W}))$ of the Jacobian matrix $\boldsymbol{J}(\boldsymbol{W})$. For any unit vector $\boldsymbol{c} = \{c_{s,k}\}_{s \in \{\pm 1\}, k \in [K]} \cup \left\{ c_i^{(p)} \right\}_{i \in [n], p \in [P] \setminus \{p_i^*\}} \in \mathbb{R}^{2K + n(P-1)}$, we have

$$\begin{aligned} \|\boldsymbol{J}(\boldsymbol{W}) \boldsymbol{c}\|^2 = & \sum_{s \in \{\pm 1\}, k \in [K]} c_{s,k}^2 \left\| \nabla_{\boldsymbol{W}} z_{s,k} \right\|^2 + \sum_{i \in [n], p \in [P] \setminus \{p_i^*\}} \left(c_i^{(p)} \right)^2 \left\| \nabla_{\boldsymbol{W}} z_i^{(p)} \right\|^2 \\ & + \sum_{\substack{s_1, s_2 \in \{\pm 1\}, k_1, k_2 \in [K] \\ (s_1, k_1) \neq (s_2, k_2)}} c_{s_1, k_1} c_{s_2, k_2} \langle \nabla_{\boldsymbol{W}} z_{s_1, k_1}, \nabla_{\boldsymbol{W}} z_{s_2, k_2} \rangle \\ & + \sum_{\substack{s \in \{\pm 1\}, k \in [K] \\ i \in [n], p \in [P] \setminus \{p_i^*\}}} c_{s,k} c_i^{(p)} \left\langle \nabla_{\boldsymbol{W}} z_{s,k}, \nabla_{\boldsymbol{W}} z_i^{(p)} \right\rangle \\ & + \sum_{\substack{i \in [n], p \in [P] \setminus \{p_i^*\} \\ j \in [n], q \in [P] \setminus \{p_j^*\}}} c_i^{(p)} c_j^{(q)} \left\langle \nabla_{\boldsymbol{W}} z_i^{(p)}, \nabla_{\boldsymbol{W}} z_j^{(q)} \right\rangle. \end{aligned}$$

For each $s_1, s_2 \in \{\pm 1\}, k_1, k_2 \in [K]$ such that $(s_1, k_1) \neq (s_2, k_2)$, and $i \in [n], p \in [P] \setminus \{p_i^*, \tilde{p}_i\}$,

$$\langle \nabla_{\boldsymbol{W}} z_{s_1, k_1}, \nabla_{\boldsymbol{W}} z_{s_2, k_2} \rangle = \left\langle \nabla_{\boldsymbol{W}} z_{s_1, k_1}, \nabla_{\boldsymbol{W}} z_i^{(p)} \right\rangle = 0,$$

since $\langle \boldsymbol{v}_{s_1, k_1}, \boldsymbol{v}_{s_2, k_2} \rangle = \left\langle \boldsymbol{v}_{s_1, k_1}, \xi_i^{(p)} \right\rangle = 0$. Also, for each $s \in \{\pm 1\}$ and $i \in \mathcal{F}_s$, then

$$\begin{aligned} & \left| c_{s,1} c_i^{(\tilde{p}_i)} \left\langle \nabla_{\boldsymbol{W}} z_{s,1}, \nabla_{\boldsymbol{W}} z_i^{(\tilde{p}_i)} \right\rangle \right| \\ = & \left| c_{s,1} c_i^{(\tilde{p}_i)} \right| \left(\phi'(\langle \boldsymbol{w}_1, \boldsymbol{v}_{s,1} \rangle) \phi'\left(\left\langle \boldsymbol{w}_1, \boldsymbol{x}_i^{(\tilde{p}_i)} \right\rangle\right) + \phi'(\langle \boldsymbol{w}_{-1}, \boldsymbol{v}_{s,1} \rangle) \phi'\left(\left\langle \boldsymbol{w}_{-1}, \boldsymbol{x}_i^{(\tilde{p}_i)} \right\rangle\right) \right) \alpha \\ \leq & \frac{1}{2} c_{s,1}^2 \left(\phi'(\langle \boldsymbol{w}_1, \boldsymbol{v}_{s,1} \rangle)^2 + \phi'(\langle \boldsymbol{w}_{-1}, \boldsymbol{v}_{s,1} \rangle)^2 \right) \frac{\alpha^2}{\left\| \boldsymbol{x}_i^{(\tilde{p}_i)} \right\|^2} \\ & + \frac{1}{2} \left(c_i^{(\tilde{p}_i)} \right)^2 \left(\phi'\left(\boldsymbol{w}_1, \boldsymbol{x}_i^{(\tilde{p}_i)} \right)^2 + \phi'\left(\boldsymbol{w}_{-1}, \boldsymbol{x}_i^{(\tilde{p}_i)} \right)^2 \right) \left\| \boldsymbol{x}_i^{(\tilde{p}_i)} \right\|^2 \\ < & \frac{1}{2n} c_{s,1}^2 \left(\phi'(\langle \boldsymbol{w}_1, \boldsymbol{v}_{s,1} \rangle)^2 + \phi'(\langle \boldsymbol{w}_{-1}, \boldsymbol{v}_{s,1} \rangle)^2 \right) \\ & + \frac{1}{2} \left(c_i^{(\tilde{p}_i)} \right)^2 \left(\phi'\left(\boldsymbol{w}_1, \boldsymbol{x}_i^{(\tilde{p}_i)} \right)^2 + \phi'\left(\boldsymbol{w}_{-1}, \boldsymbol{x}_i^{(\tilde{p}_i)} \right)^2 \right) \left\| \boldsymbol{x}_i^{(\tilde{p}_i)} \right\|^2, \end{aligned}$$

where we use the fact that $n\alpha^2 = o\left(\sigma_{\mathrm{d}}^2 d\right)$ for the last inequality. Also,

$$\left\langle \nabla_{\boldsymbol{W}} z_{-s,1}, \nabla_{\boldsymbol{W}} z_i^{(\tilde{p}_i)} \right\rangle = 0.$$

Furthermore, for each $i, j \in [n], p \in [P] \setminus \{p_i^*\}, q \in [P] \setminus \{p_j^*\}$ with $(i,p) \neq (j,q)$ satisfies

$$\begin{aligned}
&\left|c_i^{(p)} c_j^{(q)} \left\langle \nabla_{\boldsymbol{W}} z_i^{(p)}, \nabla_{\boldsymbol{W}} z_j^{(q)} \right\rangle\right| \\
&= \left|c_i^{(p)} c_j^{(q)}\right| \left(\phi'\left(\left\langle \boldsymbol{w}_1, \boldsymbol{x}_i^{(p)} \right\rangle\right) \phi'\left(\left\langle \boldsymbol{w}_1, \boldsymbol{x}_j^{(q)} \right\rangle\right) + \phi'\left(\left\langle \boldsymbol{w}_{-1}, \boldsymbol{x}_i^{(p)} \right\rangle\right) \phi'\left(\left\langle \boldsymbol{w}_{-1}, \boldsymbol{x}_j^{(q)} \right\rangle\right) \right) \left|\left\langle \boldsymbol{x}_i^{(p)}, \boldsymbol{x}_j^{(q)} \right\rangle\right| \\
&\le \frac{1}{4Pn} \left(c_i^{(p)}\right)^2 \left(\phi'\left(\left\langle \boldsymbol{w}_1, \boldsymbol{x}_i^{(p)} \right\rangle\right)^2 + \phi'\left(\left\langle \boldsymbol{w}_{-1}, \boldsymbol{x}_i^{(p)} \right\rangle\right)^2 \right) \left\|\boldsymbol{x}_i^{(p)}\right\|^2 \\
&\quad + \frac{1}{4Pn} \left(c_j^{(q)}\right)^2 \left(\phi'\left(\left\langle \boldsymbol{w}_1, \boldsymbol{x}_j^{(q)} \right\rangle\right)^2 + \phi'\left(\left\langle \boldsymbol{w}_{-1}, \boldsymbol{x}_j^{(q)} \right\rangle\right)^2 \right) \left\|\boldsymbol{x}_j^{(q)}\right\|^2 \\
&= \frac{1}{4Pn} \left(\left(c_i^{(p)}\right)^2 \left\|\nabla_{\boldsymbol{W}} z_i^{(p)}\right\|^2 + \left(c_j^{(q)}\right)^2 \left\|\nabla_{\boldsymbol{W}} z_j^{(q)}\right\|^2 \right)
\end{aligned}$$

where the last inequality is due to AM-GM inequality and

$$\left\|\boldsymbol{x}_i^{(p)}\right\| \cdot \left\|\boldsymbol{x}_j^{(q)}\right\| \ge \left\|\xi_i^{(p)}\right\| \cdot \left\|\xi_j^{(q)}\right\| \ge 2np \left|\left\langle \xi_i^{(p)}, \xi_j^{(q)} \right\rangle\right| + \alpha^2 \ge 2nP \left|\left\langle \boldsymbol{x}_i^{(p)}, \boldsymbol{x}_j^{(q)} \right\rangle\right|,$$

which is implied by the event E_{init} and $n = o\left(d^{\frac{1}{2}}\right)$ from Assumption 8.

For $s \in \{\pm 1\}, k \in [K]$ and $i \in [n], p \in [P] \setminus \{p_i^*\}$,

$$\|\nabla_{\boldsymbol{W}} z_{s,k}\|^2 = \phi'(\langle \boldsymbol{w}_1, \boldsymbol{v}_{s,k} \rangle)^2 + \phi'(\langle \boldsymbol{w}_{-1}, \boldsymbol{v}_{s,k} \rangle)^2 \ge 2\beta^2,$$

and

$$\begin{aligned}
\left\|\nabla_{\boldsymbol{W}} z_i^{(p)}\right\|^2 &= \left(\phi'\left(\left\langle \boldsymbol{w}_1, \boldsymbol{x}_i^{(p)} \right\rangle\right)^2 + \phi'\left(\left\langle \boldsymbol{w}_{-1}, \boldsymbol{x}_i^{(p)} \right\rangle\right)^2 \right) \left\|\boldsymbol{x}_i^{(p)}\right\|^2 \\
&\ge \beta^2 \sigma_{i,p}^2 d \\
&\ge \beta^2.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\|\boldsymbol{J}(\boldsymbol{W})\boldsymbol{c}\|^2 \\
&= \sum_{s \in \{\pm 1\}} c_{s,k}^2 \|\nabla_{\boldsymbol{W}} z_{s,k}\|^2 + \sum_{i \in [n], p \in [P] \setminus \{p_i^*\}} \left(c_i^{(p)}\right)^2 \left\|\nabla_{\boldsymbol{W}} z_i^{(p)}\right\|^2 \\
&\quad + \sum_{s \in \{\pm 1\}, i \in \mathcal{F}_s} c_{s,1} c_i^{(\tilde{p}_i)} \left\langle \nabla_{\boldsymbol{W}} z_{s,1}, \nabla_{\boldsymbol{W}} z_i^{(\tilde{p}_i)} \right\rangle + \sum_{\substack{i \in [n], p \in [P] \setminus \{p_i^*\} \\ j \in [n], q \in [P] \setminus \{p_j^*\}}} c_i^{(p)} c_j^{(q)} \left\langle \nabla_{\boldsymbol{W}} z_i^{(p)}, \nabla_{\boldsymbol{W}} z_j^{(q)} \right\rangle \\
&\ge \sum_{s \in \{\pm 1\}} c_{s,k}^2 \|\nabla_{\boldsymbol{W}} z_{s,k}\|^2 + \sum_{i \in [n], p \in [P] \setminus \{p_i^*\}} \left(c_i^{(p)}\right)^2 \left\|\nabla_{\boldsymbol{W}} z_i^{(p)}\right\|^2 \\
&\quad - \sum_{s \in \{\pm 1\}, i \in \mathcal{F}_s} \left(\frac{1}{2n} c_{s,1}^2 \|\nabla_{\boldsymbol{W}} z_{s,1}\|^2 + \frac{1}{2} \left(c_i^{(\tilde{p}_i)}\right)^2 \|\nabla_{\boldsymbol{W}} z_i^{(\tilde{p}_i} \| \right) \\
&\quad - \frac{1}{4Pn} \sum_{\substack{i \in [n], p \in [P] \setminus \{p_i^*\} \\ j \in [n], q \in [P] \setminus \{p_j^*\}}} \left(\left(c_i^{(p)}\right)^2 \left\|\nabla_{\boldsymbol{W}} z_i^{(p)}\right\|^2 + \left(c_j^{(q)}\right)^2 \left\|\nabla_{\boldsymbol{W}} z_j^{(q)}\right\|^2 \right) \\
&> \frac{1}{4} \sum_{s \in \{\pm 1\}, k \in [K]} c_{s,k}^2 \|\nabla_{\boldsymbol{W}} z_{s,k}\|^2 + \frac{1}{4} \sum_{i \in [n], p \in [P] \setminus \{p_i^*\}} \left(c_i^{(p)}\right)^2 \left\|\nabla_{\boldsymbol{W}} z_i^{(p)}\right\| \ge \frac{\beta^2}{4},
\end{aligned}$$

and we conclude $\sigma_{\min}(\boldsymbol{J}(\boldsymbol{W})) \ge \frac{\beta}{2}$ for any $\boldsymbol{W}$.

H.2.2. CHARACTERIZATION OF A GLOBAL MINIMUM OF $h(\boldsymbol{Z})$

In this section, we will check that $h(\boldsymbol{Z})$ is strictly convex and it has a global minimum.

For each $i, j \in [n]$ and $\mathcal{S} \subset [P]$ let us define $\boldsymbol{a}_{i,j,\mathcal{S}} \in \mathbb{R}^{2K+n(P-1)}$ as

$$\boldsymbol{a}_{i,j,\mathcal{S}} = \sum_{p \in \mathcal{S}} \boldsymbol{e}_i^{(p)} + \sum_{p \notin \mathcal{S}} \boldsymbol{e}_j^{(p)},$$

and then

$$h(\boldsymbol{Z}) = \frac{1}{n^2} \sum_{i,j \in [n]} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}} \left[\frac{|\mathcal{S}|}{P} \ell\left(y_i \langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z} \rangle\right) + \left(1 - \frac{|\mathcal{S}|}{P}\right) \ell\left(y_j \langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z} \rangle\right) \right].$$

Since $\ell(\cdot)$ is convex, $h(\boldsymbol{Z})$ is also convex. Note that

$$\nabla h(\boldsymbol{Z}) = \frac{1}{n^2} \sum_{i,j \in [n]} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}} \left[\left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i \langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z} \rangle) + \left(1 - \frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j \langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z} \rangle) \right) \boldsymbol{a}_{i,j,\mathcal{S}} \right],$$

and

$$\begin{aligned}
&\nabla^2 h(\boldsymbol{Z}) \\
&= \frac{1}{n^2} \sum_{i,j \in [n]} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}} \left[\left(\frac{|\mathcal{S}|}{P} \ell''(y_i \langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z} \rangle) + \left(1 - \frac{|\mathcal{S}|}{P}\right) \ell''(y_j \langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z} \rangle) \right) \boldsymbol{a}_{i,j,\mathcal{S}} \boldsymbol{a}_{i,j,\mathcal{S}}^\top \right] \\
&= \frac{1}{n^2} \sum_{i,j \in [n]} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}} \left[\ell''(\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z} \rangle) \boldsymbol{a}_{i,j,\mathcal{S}} \boldsymbol{a}_{i,j,\mathcal{S}}^\top \right],
\end{aligned}$$

where the last equality holds since $\ell''(z) = \ell''(-z)$ for any $z \in \mathbb{R}$. From the equation above, it suffices to show that $\{\boldsymbol{a}_{i,j,\mathcal{S}}\}_{i,j \in [n], \mathcal{S} \subset [P]}$ spans $\mathbb{R}^{2K+n(P-1)}$ to show strict convexity of $h(\boldsymbol{Z})$. We define a function $I : [P] \to [P]$ such that for each $p \in [P]$, $p^*_{I(p)} = p$, where the existence is guaranteed by Lemma 9. Then for any $i \in [n]$ and $p \in [p]$, we have

$$\begin{aligned}
&\boldsymbol{a}_{i,i,\emptyset} + \sum_{q \in [P] \setminus \{p\}} \boldsymbol{a}_{I(q),i,\{q\}} - (P-1) \boldsymbol{a}_{I(p),i,\{p\}} \\
&= \sum_{p' \in [P]} \boldsymbol{e}_i^{(p')} + \sum_{q \in [P] \setminus \{p\}} \left(\boldsymbol{v}_{1,1} + \sum_{p' \in [P] \setminus \{q\}} \boldsymbol{e}_i^{(p')} \right) - (P-1) \left(\boldsymbol{v}_{1,1} + \sum_{p' \in [P] \setminus \{p\}} \boldsymbol{e}_i^{(p')} \right) \\
&= \sum_{p' \in [P]} \boldsymbol{e}_i^{(p')} + \left((P-1) \boldsymbol{e}_i^{(p)} + (P-2) \sum_{p' \in [P] \setminus \{p\}} \boldsymbol{e}_i^{(p')} \right) - (P-1) \sum_{p' \in [P] \setminus \{p\}} \boldsymbol{e}_i^{(p')} \\
&= P \boldsymbol{e}_i^{(p)}.
\end{aligned} \tag{17}$$

Hence, $h(\boldsymbol{Z})$ is strictly convex and it can have at most one global minimum. We want to show the existence of the global minimum and characterize it.

$$
\begin{aligned}
&n^2 \nabla h(\boldsymbol{Z})\\
&= \sum_{i,j\in[n]} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}\left[\left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle) + \left(1-\frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle)\right)\boldsymbol{a}_{i,j,\mathcal{S}}\right]\\
&= 2\sum_{\substack{i,j\in[n]\\ p\in[P]}} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}\left[\left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle) + \left(1-\frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle)\right)\mathbb{1}_{p\in\mathcal{S}}\right]\boldsymbol{e}_i^{(p)}.
\end{aligned}
$$

We can simplify terms as

$$
\begin{aligned}
&\sum_{j\in[n]} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}\left[\left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle) + \left(1-\frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle)\right)\mathbb{1}_{p\in\mathcal{S}}\right]\\
&= \sum_{j\in\mathcal{V}_{y_i}} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}\left[\left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle) + \left(1-\frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle)\right)\mathbb{1}_{p\in\mathcal{S}}\right]\\
&\quad + \sum_{j\in\mathcal{V}_{-y_i}} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}\left[\left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle) + \left(1-\frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle)\right)\mathbb{1}_{p\in\mathcal{S}}\right]\\
&= y_i \sum_{j\in\mathcal{V}_{y_i}} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}[\ell'(y_i\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle)\mathbb{1}_{p\in\mathcal{S}}]\\
&\quad + y_i \sum_{j\in\mathcal{V}_{-y_i}} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}\left[\left(\ell'(y_i\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle) + \left(1-\frac{|\mathcal{S}|}{P}\right)\right)\mathbb{1}_{p\in\mathcal{S}}\right]\\
&= y_i|\mathcal{V}_{-y_i}|\mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}\left[\left(1-\frac{|\mathcal{S}|}{P}\right)\mathbb{1}_{p\in\mathcal{S}}\right] + y_i \sum_{j\in[n]} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}[\ell'(y_i\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle)\mathbb{1}_{p\in\mathcal{S}}],
\end{aligned}
$$

where the second equality holds since $\ell'(z) + \ell'(-z) = -1$. Also, for any $p \in [P]$,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}\left[\left(1-\frac{|\mathcal{S}|}{P}\right)\mathbb{1}_{p\in\mathcal{S}}\right] &= \frac{1}{P}\sum_{q\in[P]} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}\left[\left(1-\frac{|\mathcal{S}|}{P}\right)\mathbb{1}_{q\in\mathcal{S}}\right]\\
&= \frac{1}{P}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}\left[\left(1-\frac{|\mathcal{S}|}{P}\right)\sum_{q\in\mathcal{S}}\mathbb{1}_{q\in\mathcal{S}}\right]\\
&= \frac{1}{P}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}\left[\left(1-\frac{|\mathcal{S}|}{P}\right)|\mathcal{S}|\right] = \frac{P-1}{6P}.
\end{aligned}
$$

Hence, if

$$
\sum_{j\in[n]} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_{\mathcal{S}}}[\ell'(y_i\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle)\mathbb{1}_{p\in\mathcal{S}}] + \frac{P-1}{6P}|\mathcal{V}_{-y_i}| = 0,
$$

for all $i \in [n]$ and $p \in [P]$, then we have $\nabla h(\boldsymbol{Z}) = 0$. Let us consider a specific $\boldsymbol{Z}$ parameterized by z_1, z_{-1}, of the form $z_i^{(p)} = y_i z_{y_i}$ for all $i \in [n]$ and $p \in [P]$. We will find a stationary point with this specific form and then it should be the unique global minimum in the entire domain. Then we have for each $i \in [n]$, and $p \in [P] \setminus \{p_i^*\}$, we have

$$
\begin{aligned}
&\sum_{j\in[n]} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[\ell'(y_i\langle a_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle)\mathbb{1}_{p\in\mathcal{S}}] \\
&= \sum_{j\in\mathcal{V}_{y_i}} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[\ell'(y_i\,\langle a_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle)\mathbb{1}_{p\in\mathcal{S}}] + \sum_{j\in\mathcal{V}_{-y_i}} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[\ell'(y_i\langle a_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle)\mathbb{1}_{p\in\mathcal{S}}] \\
&= |\mathcal{V}_{y_i}| \cdot \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[\ell'(Pz_{y_i})\mathbb{1}_{p\in\mathcal{S}}] + |\mathcal{V}_{-y_i}| \cdot \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[\ell'(|\mathcal{S}|z_{y_i} - (P-|\mathcal{S}|)z_{-y_i})\mathbb{1}_{p\in\mathcal{S}}] \\
&= \frac{1}{P}\sum_{p\in[P]}\Big(|\mathcal{V}_{y_i}| \cdot \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[\ell'(Pz_{y_i})\mathbb{1}_{p\in\mathcal{S}}] + |\mathcal{V}_{-y_i}| \cdot \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[\ell'(|\mathcal{S}|z_{y_i} - (P-|\mathcal{S}|)z_{-y_i})\mathbb{1}_{p\in\mathcal{S}}]\Big) \\
&= \frac{1}{P}\left(|\mathcal{V}_{y_i}| \cdot \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}\left[\ell'(Pz_{y_i})\sum_{p\in\mathcal{S}}\mathbb{1}_{p\in\mathcal{S}}\right]\right. \\
&\qquad \left. + |\mathcal{V}_{-y_i}| \cdot \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}\left[\ell'(|\mathcal{S}|z_{y_i} - (P-|\mathcal{S}|)z_{-y_i})\sum_{p\in\mathcal{S}}\mathbb{1}_{p\in\mathcal{S}}\right]\right) \\
&= \frac{1}{P}\Big(|\mathcal{V}_{y_i}| \cdot \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[|\mathcal{S}|\ell'(Pz_{y_i})] + |\mathcal{V}_{-y_i}| \cdot \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[|\mathcal{S}|\ell'(|\mathcal{S}|z_{y_i} - (P-|\mathcal{S}|)z_{-y_i})]\Big) \\
&= \frac{|\mathcal{V}_{y_i}|}{2}\ell'(Pz_{y_i}) + \frac{|\mathcal{V}_{-y_i}|}{P}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[|\mathcal{S}|\ell'(|\mathcal{S}|z_{y_i} - (P-|\mathcal{S}|)z_{-y_i})].
\end{aligned}
$$

To show the existence of global minimum and characterize it, we will prove the following lemma.

Lemma 31 *Suppose the event E_{init} occurs. Let $g_1, g_{-1} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be defined as*

$$
g_s(z_1, z_{-1}) := \frac{|\mathcal{V}_s|}{|\mathcal{V}_{-s}|}\ell'(Pz_s) + \frac{2}{P}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}\left[|\mathcal{S}|\ell'(|\mathcal{S}|z_s - (P-|\mathcal{S}|)z_{-s})\right] + \frac{P-1}{3P},
$$

for each $s \in \{\pm 1\}$. There exist unique $z_1^*, z_{-1}^* > 0$ such that $g_1(z_1^*, z_{-1}^*) = g_{-1}(z_1^*, z_{-1}^*) = 0$. Furthermore,

$$
\frac{1}{P}\log\left(\frac{3P\left(1+\frac{|\mathcal{V}_1|}{|\mathcal{V}_{-1}|}\right)}{P-1} - 1\right) \le z_1^* \le \log\left(\frac{3P\left(1+\frac{|\mathcal{V}_{-1}|}{|\mathcal{V}_1|}\right)}{P-1} - 1\right) + \log 9,
$$

and

$$
\frac{1}{P}\log\left(\frac{3P\left(1+\frac{|\mathcal{V}_{-1}|}{|\mathcal{V}_1|}\right)}{P-1} - 1\right) \le z_{-1}^* \le \log\left(\frac{3P\left(1+\frac{|\mathcal{V}_1|}{|\mathcal{V}_{-1}|}\right)}{P-1} - 1\right) + \log 9.
$$

Proof [Proof of Lemma 31] For each $z_1 > 0$,

$$
\begin{aligned}
g_{-1}(z_1, 0) &= \left(\frac{|\mathcal{V}_{-1}|}{|\mathcal{V}_1|} + 1\right)\cdot\left(-\frac{1}{2}\right) + \frac{2}{P}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[|\mathcal{S}|\ell'(-(P-|\mathcal{S}|)z_1)] + \frac{P-1}{3P} \\
&< \left(\frac{|\mathcal{V}_{-1}|}{|\mathcal{V}_1|} + 1\right)\cdot\left(-\frac{1}{2}\right) + \frac{P-1}{3P} < 0,
\end{aligned}
$$

since $\ell'(z) \leq -\frac{1}{2}$ for any $z \leq 0$. In addition,

$$\begin{aligned}
&g_{-1}(z_1, Pz_1 + \log 9) \\
&= \frac{|\mathcal{V}_{-1}|}{|\mathcal{V}_1|}\ell'(P^2 z_1 + P\log 9) + \frac{2}{P}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[|\mathcal{S}|\ell'(|\mathcal{S}|Pz_1 + |\mathcal{S}|\log 9 - (P - |\mathcal{S}|)z_1)] + \frac{P-1}{3P} \\
&\geq \left(\frac{|\mathcal{V}_{-1}|}{|\mathcal{V}_1|} + 1\right)\ell'(\log 9) + \frac{P-1}{3P} \\
&> 0.
\end{aligned}$$

Since $z \mapsto g_{-1}(z_1, z)$ is strictly increasing and by intermediate value theorem, there exists $S : (0,\infty) \to (0,\infty)$ such that $z = S(z_1)$ is a unique solution of $g_{-1}(z_1, z) = 0$ and $S(z_1) < Pz_1 + \log 9$. Note that S is strictly decreasing since $g_{-1}(z_1, z_{-1})$ is strictly decreasing with respect to z_1 and strictly increasing with respect to z_{-1}.

Let us choose $\underline{z} > 0$ such that

$$\underline{z} = \frac{1}{P}\log\left(\frac{3P\left(1 + \frac{|\mathcal{V}_1|}{|\mathcal{V}_{-1}|}\right)}{P-1} - 1\right),$$

and thus

$$\ell'(P\underline{z}) = -\frac{P-1}{3P\left(1 + \frac{|\mathcal{V}_1|}{|\mathcal{V}_{-1}|}\right)}.$$

We have

$$\begin{aligned}
g_1(\underline{z}, S(\underline{z})) &= \frac{|\mathcal{V}_1|}{|\mathcal{V}_{-1}|}\ell'(P\underline{z}) + \frac{2}{P}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[|\mathcal{S}|\ell'(|\mathcal{S}|\underline{z} - (P - |\mathcal{S}|)S(\underline{z}))] + \frac{P-1}{3P} \\
&\leq \left(\frac{|\mathcal{V}_1|}{|\mathcal{V}_{-1}|} + 1\right)\ell'(P\underline{z}) + \frac{P-1}{3P} \\
&= 0,
\end{aligned}$$

and

$$\lim_{z\to\infty} g_1(z, S(z)) = \frac{P-1}{3P} > 0.$$

Hence, there exist unique $z_1^*, z_{-1}^* > 0$ such that $g_1(z_1^*, z_{-1}^*) = g_{-1}(z_1^*, z_{-1}^*) = 0$. In addition, $\underline{z} \leq z_1^*$ and $z_{-1}^* \leq S(\underline{z}) < P\underline{z} + \log 9$. Thus,

$$z_1^* \geq \frac{1}{P}\log\left(\frac{3P\left(1 + \frac{|\mathcal{V}_1|}{|\mathcal{V}_{-1}|}\right)}{P-1} - 1\right), \quad z_{-1}^* \leq \log\left(\frac{3P\left(1 + \frac{|\mathcal{V}_1|}{|\mathcal{V}_{-1}|}\right)}{P-1} - 1\right) + \log 9$$

By using a similar argument, we can show that

$$z_{-1}^* \geq \frac{1}{P}\log\left(\frac{3P\left(1 + \frac{|\mathcal{V}_{-1}|}{|\mathcal{V}_1|}\right)}{P-1} - 1\right), \quad z_1^* \leq \log\left(\frac{3P\left(1 + \frac{|\mathcal{V}_{-1}|}{|\mathcal{V}_1|}\right)}{P-1} - 1\right) + \log 9,$$

and we have our conclusion. ■

Therefore, there exists a unique minimizer $\hat{\boldsymbol{Z}} = \{\hat{z}_{s,k}\}_{s\in\{\pm1\},k\in[K]} \cup \left\{\hat{z}_i^{(p)}\right\}_{i\in[n],p\in[P]\setminus\{p_i^*\}}$ of $h(\boldsymbol{Z})$ and it satisfies $s\hat{z}_{s,k} = z_s^* = \Theta(1)$ for all $k \in [K]$ and $y_i\hat{z}_i^{(p)} = z_{y_i}^* = \Theta(1)$ for all $i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$.

H.2.3. STRONG CONVEXITY NEAR GLOBAL MINIMUM OF $h(\boldsymbol{Z})$

We will show that $h(\boldsymbol{Z})$ is strongly convex in a set $\mathcal{G}$ containing a global minimum $\hat{\boldsymbol{Z}}$ where $\mathcal{G}$ is defined as follows.

$$\mathcal{G} := \left\{\boldsymbol{Z} \in \mathbb{R}^{2K+n(P-1)} : \|\boldsymbol{Z} - \hat{\boldsymbol{Z}}\|_\infty < \|\hat{\boldsymbol{Z}}\|_\infty\right\},$$

here $\|\cdot\|_\infty$ is ℓ_∞ norm. For any $\boldsymbol{Z} \in \mathcal{G}$ and a unit vector $\boldsymbol{c} \in \mathbb{R}^{2K+n(P-1)}$ with $\boldsymbol{c} = \sum_{s\in\{\pm1\},k\in[K]} c_{s,k}\boldsymbol{e}_{s,k} + \sum_{i\in[n],p\in[P]\setminus\{p_i^*\}} c_i^{(p)}\boldsymbol{e}_i^{(p)}$, we have

$$\begin{aligned}
\boldsymbol{c}^\top\nabla^2 h(\boldsymbol{Z})\boldsymbol{c} &= \frac{1}{n^2}\sum_{i,j\in[n]} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}\left[\ell''(\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{Z}\rangle)\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{c}\rangle^2\right] \\
&\geq \frac{\ell''(2P\|\hat{\boldsymbol{Z}}\|_\infty)}{n^2}\sum_{i,j\in[n]} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[\langle \boldsymbol{a}_{i,j,\mathcal{S}}, \boldsymbol{c}\rangle^2].
\end{aligned}$$

Note that for each $i \in [n], p \in [P]$, from (17), we have

$$c_i^{(p)} = \left\langle \boldsymbol{c}, \boldsymbol{e}_i^{(p)}\right\rangle = \frac{1}{P}\langle \boldsymbol{c}, \boldsymbol{a}_{i,i,\emptyset}\rangle + \frac{1}{P}\sum_{q\in[P]\setminus\{p\}} \langle \boldsymbol{c}, \boldsymbol{a}_{I(q),i,\{q\}}\rangle - \frac{P-1}{P}\left\langle \boldsymbol{c}, \boldsymbol{a}_{I(p),i,\{p\}}\right\rangle,$$

where we use the notational convention $c_i^{(p_i^*)} = c_{s,k}$ for $s \in \{\pm1\}, k \in [K]$ and $i \in \mathcal{V}_{s,k}$. By Cauchy-Schwartz inequality and the fact that $\mathbb{P}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[\mathcal{S} = \emptyset], \mathbb{P}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[\mathcal{S} = \{q\}] \geq \frac{1}{P(P+1)}$ for all $q \in [P]$,

$$\begin{aligned}
&\left(c_i^{(p)}\right)^2 \\
&= \left(\frac{1}{P}\left\langle \boldsymbol{c}, \boldsymbol{a}_{i,i,\emptyset}\right\rangle + \frac{1}{P}\sum_{q\in[P]\setminus\{p\}}\left\langle \boldsymbol{c}, \boldsymbol{a}_{I(q),i,\{q\}}\right\rangle - \frac{P-1}{P}\left\langle \boldsymbol{c}, \boldsymbol{a}_{I(p),i,\{p\}}\right\rangle\right)^2 \\
&\leq \left(\frac{1}{P^2} + \frac{P-1}{P^2} + \left(-\frac{P-1}{P}\right)^2\right)\left(\left\langle \boldsymbol{c}, \boldsymbol{a}_{i,i,\emptyset}\right\rangle^2 + \sum_{q\in[P]\setminus\{p\}}\left\langle \boldsymbol{c}, \boldsymbol{a}_{I(q),i,\{q\}}\right\rangle^2 + \left\langle \boldsymbol{c}, \boldsymbol{a}_{I(p),i,\{p\}}\right\rangle^2\right) \\
&\leq \left(\frac{1}{P^2} + \frac{P-1}{P^2} + \left(-\frac{P-1}{P}\right)^2\right)P(P+1)\sum_{i,j\in[n]}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}[\langle \boldsymbol{c}, \boldsymbol{a}_{i,j,\mathcal{S}}\rangle^2] \\
&\leq 2P^2\sum_{i,j,\in[n]}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}\left[\langle \boldsymbol{c}, \boldsymbol{a}_{i,j,\mathcal{S}}\rangle^2\right].
\end{aligned}$$

Hence, we have

$$
\begin{aligned}
\boldsymbol{c}^\top \nabla^2 h(\boldsymbol{Z})\boldsymbol{c} &\geq \frac{\ell''(2P\|\hat{Z}\|_\infty)}{(4K+2n(P-1))P^2n^2}(4K+2n(P-1))P^2 \sum_{i,j\in[n]} \mathbb{E}_{\mathcal{S}\sim\mathcal{D}_\mathcal{S}}\left[\langle \boldsymbol{c}, \boldsymbol{a}_{i,j,\mathcal{S}}\rangle^2\right] \\
&\geq \frac{\ell''(2P\|\hat{Z}\|_\infty)}{(4K+2n(P-1))P^2n^2}\left(\sum_{s\in\{\pm1\},k\in[K]} c_{s,k}^2 + \sum_{i\in[n],q\in[P]\setminus\{p_i^*\}} \left(c_i^{(q)}\right)^2\right) \\
&= \frac{\ell''(2P\|\hat{Z}\|_\infty)}{(4K+2n(P-1))P^2n^2},
\end{aligned}
$$

and we conclude $h(\boldsymbol{Z})$ is μ-strongly convex in $\mathcal{G}$ where $\mu := \frac{\ell''(2P\|\hat{Z}\|_\infty)}{(4K+2n(P-1))P^2n^2} = \frac{1}{\text{poly}(d)}$.

H.2.4. ϵ-STATIONARY POINTS OF $h(Z)$ ARE CLOSE TO GLOBAL MINIMUM

In this step, we want to show that near stationary points of $h(\boldsymbol{Z})$ are close to a global minimum $\hat{\boldsymbol{Z}}$.

Lemma 32 *Suppose $\boldsymbol{Z} \in \mathbb{R}^{2K+n(P-1)}$ satisfies $\|\nabla h(\boldsymbol{Z})\| < \mu\epsilon$ with some $0 < \epsilon < \frac{\|\hat{\boldsymbol{Z}}\|_\infty}{2}$. Then, we have $\left\|\boldsymbol{Z} - \hat{\boldsymbol{Z}}\right\| < \epsilon$.*

Proof [Proof of Lemma 32] If $\boldsymbol{Z} = \hat{\boldsymbol{Z}}$, we immediately have our conclusion. We may assume $\boldsymbol{Z} \neq \hat{\boldsymbol{Z}}$.

Let us define a function $g : \mathbb{R} \to \mathbb{R}$ as $g(t) = h\left(\hat{\boldsymbol{Z}} + t(\boldsymbol{Z} - \hat{\boldsymbol{Z}})\right)$. Then g is convex and

$$
\begin{aligned}
g'(t) &= \left\langle \nabla h\left(\hat{\boldsymbol{Z}} + t(\boldsymbol{Z} - \hat{\boldsymbol{Z}})\right), \boldsymbol{Z} - \hat{\boldsymbol{Z}}\right\rangle, \\
g''(t) &= \left(\boldsymbol{Z} - \hat{\boldsymbol{Z}}\right)^\top \nabla^2 h\left(\hat{\boldsymbol{Z}} + t(\boldsymbol{Z} - \hat{\boldsymbol{Z}})\right)\left(\boldsymbol{Z} - \hat{\boldsymbol{Z}}\right).
\end{aligned}
$$

Furthermore, for $0 \leq t \leq t_0$ where $t_0 := \frac{\|\hat{\boldsymbol{Z}}\|_\infty}{2\left\|\boldsymbol{Z}-\hat{\boldsymbol{Z}}\right\|_\infty}$,

$$
\hat{\boldsymbol{Z}} + t(\boldsymbol{Z} - \hat{\boldsymbol{Z}}) \in \mathcal{G}, \qquad \therefore g''(t) \geq \mu\left\|\boldsymbol{Z} - \hat{\boldsymbol{Z}}\right\|^2.
$$

We can conclude g is $\mu\left\|\boldsymbol{Z} - \hat{\boldsymbol{Z}}\right\|^2$-strongly convex in $[0, t_0]$. From strong convexity in $[0, t_0]$ and convexity in $\mathbb{R}$, we have

$$
(g'(t_0) - g'(0))t_0 = g'(t_0)t_0 \geq \mu\left\|\boldsymbol{Z} - \hat{\boldsymbol{Z}}\right\|^2 t_0^2, \quad (g'(1) - g'(t_0))(1 - t_0) \geq 0.
$$

If $t_0 < 1$, we have

$$
\|\nabla h(\boldsymbol{Z})\|\left\|\boldsymbol{Z} - \hat{\boldsymbol{Z}}\right\| \geq \left\langle \nabla h(\boldsymbol{Z}), \boldsymbol{Z} - \hat{\boldsymbol{Z}}\right\rangle = g'(1) \geq g'(t_0) \geq \mu\left\|\boldsymbol{Z} - \hat{\boldsymbol{Z}}\right\|^2 t_0,
$$

and

$$
\|\nabla h(\boldsymbol{Z})\| \geq \mu\left\|\boldsymbol{Z} - \hat{\boldsymbol{Z}}\right\| t_0 = \frac{\mu\left\|\boldsymbol{Z} - \hat{\boldsymbol{Z}}\right\|\left\|\hat{\boldsymbol{Z}}\right\|_\infty}{\left\|\boldsymbol{Z} - \hat{\boldsymbol{Z}}\right\|_\infty} \geq \frac{\mu\left\|\hat{\boldsymbol{Z}}\right\|_\infty}{2},
$$

this is contradictory. Thus, we have $t_0 \geq 1$ and $\boldsymbol{Z} \in \mathcal{G}$. From the strong convexity of $h(\boldsymbol{Z})$ in $\mathcal{G}$, we have

$$\mu \left\| \boldsymbol{Z} - \hat{\boldsymbol{Z}} \right\| \leq \left\| \nabla h(\boldsymbol{Z}) - \nabla h(\hat{\boldsymbol{Z}}) \right\| = \| \nabla h(\boldsymbol{Z}) \| < \mu \epsilon,$$

and we have our conclusion $\left\| \boldsymbol{Z} - \hat{\boldsymbol{Z}} \right\| < \epsilon$. ■

H.2.5. Gradient Descent on $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$ Achieves ϵ-Stationary Point of $h(\boldsymbol{Z})$

We will show that $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$ is a smooth function.

Lemma 33 *Suppose the event E_{init} occurs. CutMix Loss $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$ is L-smooth with $L = 9r^{-1}P\sigma_{\text{d}}^2 d$.*

Proof [Proof of Lemma 33] Note that

$$\begin{aligned}
&\nabla_{\boldsymbol{w}_1} \mathcal{L}_{\text{CutMix}}(\boldsymbol{W}) \\
&= \frac{1}{n^2} \sum_{i,j \in [n]} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\mathcal{S}}} \Bigg[\left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_{i,j,\mathcal{S}})) + \left(1 - \frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j f_{\boldsymbol{W}}(\boldsymbol{X}_{i,j,\mathcal{S}})) \right) \\
&\quad \times \left(\sum_{p \in \mathcal{S}} \phi'\left(\left\langle \boldsymbol{w}_1, \boldsymbol{x}_i^{(p)} \right\rangle \right) \boldsymbol{x}_i^{(p)} + \sum_{p \notin \mathcal{S}} \phi'\left(\left\langle \boldsymbol{w}_1, \boldsymbol{x}_j^{(p)} \right\rangle \right) \boldsymbol{x}_j^{(p)} \right) \Bigg].
\end{aligned}$$

Let $\widetilde{\boldsymbol{W}} = \{\widetilde{\boldsymbol{w}}_1, \widetilde{\boldsymbol{w}}_{-1}\}$ and $\overline{\boldsymbol{W}} = \{\overline{\boldsymbol{w}}_1, \overline{\boldsymbol{w}}_{-1}\}$ be any parameters of the neural network $f_{\boldsymbol{W}}$. For any $i, j \in [n]$ and $\mathcal{S} \subset [P]$,

$$\begin{aligned}
&\left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i f_{\widetilde{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) + \left(1 - \frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j f_{\widetilde{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) \right) \left(\sum_{p \in \mathcal{S}} \phi'\left(\left\langle \widetilde{\boldsymbol{w}}_1, \boldsymbol{x}_i^{(p)} \right\rangle \right) \boldsymbol{x}_i^{(p)} + \sum_{p \notin \mathcal{S}} \phi'\left(\left\langle \widetilde{\boldsymbol{w}}_1, \boldsymbol{x}_j^{(p)} \right\rangle \right) \boldsymbol{x}_j^{(p)} \right) \\
&- \left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i f_{\overline{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) + \left(1 - \frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j f_{\overline{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) \right) \left(\sum_{p \in \mathcal{S}} \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_i^{(p)} \right\rangle \right) \boldsymbol{x}_i^{(p)} + \sum_{p \notin \mathcal{S}} \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_j^{(p)} \right\rangle \right) \boldsymbol{x}_j^{(p)} \right) \\
= &\left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i f_{\widetilde{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) + \left(1 - \frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j f_{\widetilde{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) \right) \left(\sum_{p \in \mathcal{S}} \phi'\left(\left\langle \widetilde{\boldsymbol{w}}_1, \boldsymbol{x}_i^{(p)} \right\rangle \right) \boldsymbol{x}_i^{(p)} + \sum_{p \notin \mathcal{S}} \phi'\left(\left\langle \widetilde{\boldsymbol{w}}_1, \boldsymbol{x}_j^{(p)} \right\rangle \right) \boldsymbol{x}_j^{(p)} \right) \\
&- \left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i f_{\widetilde{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) + \left(1 - \frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j f_{\widetilde{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) \right) \left(\sum_{p \in \mathcal{S}} \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_i^{(p)} \right\rangle \right) \boldsymbol{x}_i^{(p)} + \sum_{p \notin \mathcal{S}} \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_j^{(p)} \right\rangle \right) \boldsymbol{x}_j^{(p)} \right) \\
&+ \left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i f_{\widetilde{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) + \left(1 - \frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j f_{\widetilde{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) \right) \left(\sum_{p \in \mathcal{S}} \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_i^{(p)} \right\rangle \right) \boldsymbol{x}_i^{(p)} + \sum_{p \notin \mathcal{S}} \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_j^{(p)} \right\rangle \right) \boldsymbol{x}_j^{(p)} \right) \\
&- \left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i f_{\overline{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) + \left(1 - \frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j f_{\overline{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) \right) \left(\sum_{p \in \mathcal{S}} \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_i^{(p)} \right\rangle \right) \boldsymbol{x}_i^{(p)} + \sum_{p \notin \mathcal{S}} \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_j^{(p)} \right\rangle \right) \boldsymbol{x}_j^{(p)} \right).
\end{aligned}$$

Since $|\ell'| \leq 1$,

$$\left| \frac{|\mathcal{S}|}{P} y_i \ell'\left(y_i f_{\widetilde{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})\right) + \left(1 - \frac{|\mathcal{S}|}{P}\right) y_j \ell'\left(y_j f_{\widetilde{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})\right) \right| \leq 1,$$

and since $|\phi'| \leq 1$,

$$\left\| \sum_{p \in \mathcal{S}} \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_i^{(p)} \right\rangle \right) \boldsymbol{x}_i^{(p)} + \sum_{p \notin \mathcal{S}} \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_j^{(p)} \right\rangle \right) \boldsymbol{x}_j^{(p)} \right\| \leq P \max_{i \in [n], p \in [P]} \left\| \boldsymbol{x}_i^{(p)} \right\|.$$

In addition, since ϕ is r^{-1}-smooth,

$$\begin{aligned}
&\left\| \left(\sum_{p\in\mathcal{S}} \phi'\left(\left\langle \widetilde{\boldsymbol{w}}_1, \boldsymbol{x}_i^{(p)}\right\rangle\right)\boldsymbol{x}_i^{(p)} + \sum_{p\notin\mathcal{S}} \phi'\left(\left\langle \widetilde{\boldsymbol{w}}_1, \boldsymbol{x}_j^{(p)}\right\rangle\right)\boldsymbol{x}_j^{(p)}\right)\right. \\
&\quad \left. - \left(\sum_{p\in\mathcal{S}} \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_i^{(p)}\right\rangle\right)\boldsymbol{x}_i^{(p)} + \sum_{p\notin\mathcal{S}} \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_j^{(p)}\right\rangle\right)\boldsymbol{x}_j^{(p)}\right)\right\| \\
&\le \sum_{p\in\mathcal{S}} \left| \phi'\left(\left\langle \widetilde{\boldsymbol{w}}_1, \boldsymbol{x}_i^{(p)}\right\rangle\right) - \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_i^{(p)}\right\rangle\right)\right| \left\|\boldsymbol{x}_i^{(p)}\right\| \\
&\quad + \sum_{p\notin\mathcal{S}} \left| \phi'\left(\left\langle \widetilde{\boldsymbol{w}}_1, \boldsymbol{x}_j^{(p)}\right\rangle\right) - \phi'\left(\left\langle \overline{\boldsymbol{w}}_1, \boldsymbol{x}_j^{(p)}\right\rangle\right)\right| \left\|\boldsymbol{x}_j^{(p)}\right\| \\
&\le r^{-1} \sum_{p\in\mathcal{S}} \left|\left\langle \widetilde{\boldsymbol{w}}_1 - \overline{\boldsymbol{w}}_1, \boldsymbol{x}_i^{(p)}\right\rangle\right| \left\|\boldsymbol{x}_i^{(p)}\right\| + r^{-1} \sum_{p\notin\mathcal{S}} \left|\left\langle \widetilde{\boldsymbol{w}}_1 - \overline{\boldsymbol{w}}_1, \boldsymbol{x}_j^{(p)}\right\rangle\right| \left\|\boldsymbol{x}_j^{(p)}\right\| \\
&\le r^{-1} P \left(\max_{i\in[n], p\in[P]} \left\|\boldsymbol{x}_i^{(p)}\right\|\right)^2 \|\widetilde{\boldsymbol{w}}_1 - \overline{\boldsymbol{w}}_1\|,
\end{aligned}$$

and since ℓ' and ϕ are 1-Lipschitz, we have

$$\begin{aligned}
&\left| \left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i f_{\widetilde{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) + \left(1 - \frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j f_{\widetilde{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}}))\right)\right. \\
&\quad \left. - \left(\frac{|\mathcal{S}|}{P} y_i \ell'(y_i f_{\overline{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})) + \left(1 - \frac{|\mathcal{S}|}{P}\right) y_j \ell'(y_j f_{\overline{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}}))\right)\right| \\
&\le \left| f_{\widetilde{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}}) - f_{\overline{\boldsymbol{W}}}(\boldsymbol{X}_{i,j,\mathcal{S}})\right| \\
&\le \sum_{p\in\mathcal{S}} \left(\left|\left\langle \widetilde{\boldsymbol{w}}_1 - \overline{\boldsymbol{w}}_1, \boldsymbol{x}_i^{(p)}\right\rangle\right| + \left|\left\langle \widetilde{\boldsymbol{w}}_{-1} - \overline{\boldsymbol{w}}_{-1}, \boldsymbol{x}_i^{(p)}\right\rangle\right|\right) \\
&\quad + \sum_{p\notin\mathcal{S}} \left(\left|\left\langle \widetilde{\boldsymbol{w}}_1 - \overline{\boldsymbol{w}}_1, \boldsymbol{x}_j^{(p)}\right\rangle\right| + \left|\left\langle \widetilde{\boldsymbol{w}}_{-1} - \overline{\boldsymbol{w}}_{-1}, \boldsymbol{x}_j^{(p)}\right\rangle\right|\right) \\
&\le P \max_{i\in[n], j\in[P]} \left\|\boldsymbol{x}_i^{(p)}\right\| \left(\|\widetilde{\boldsymbol{w}}_1 - \overline{\boldsymbol{w}}_1\| + \|\widetilde{\boldsymbol{w}}_{-1} - \overline{\boldsymbol{w}}_{-1}\|\right) \\
&\le \sqrt{2} P \max_{i\in[n], j\in[P]} \left\|\boldsymbol{x}_i^{(p)}\right\| \left\|\widetilde{\boldsymbol{W}} - \overline{\boldsymbol{W}}\right\|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left\| \nabla_{\boldsymbol{w}_1} \mathcal{L}_{\mathrm{CutMix}}(\widetilde{\boldsymbol{W}}) - \nabla_{\boldsymbol{w}_1} \mathcal{L}_{\mathrm{CutMix}}(\overline{\boldsymbol{W}})\right\| \\
&\le r^{-1} P \left(\max_{i\in[n], p\in[P]} \left\|\boldsymbol{x}_i^{(p)}\right\|\right)^2 \|\widetilde{\boldsymbol{w}}_1 - \overline{\boldsymbol{w}}_1\| + \sqrt{2} P^2 \left(\max_{i\in[n], p\in[P]} \left\|\boldsymbol{x}_i^{(p)}\right\|\right)^2 \left\|\widetilde{\boldsymbol{W}} - \overline{\boldsymbol{W}}\right\| \\
&\le 2r^{-1} P \left(\max_{i\in[n], p\in[P]} \left\|\boldsymbol{x}_i^{(p)}\right\|\right)^2 \left\|\widetilde{\boldsymbol{W}} - \overline{\boldsymbol{W}}\right\|.
\end{aligned}$$

In the same way, we can obtain

$$\left\|\nabla_{\boldsymbol{w}_{-1}}\mathcal{L}_{\text{CutMix}}(\widetilde{\boldsymbol{W}}) - \nabla_{\boldsymbol{w}_{-1}}\mathcal{L}_{\text{CutMix}}(\overline{\boldsymbol{W}})\right\| \leq 2r^{-1}P\left(\max_{i\in[n],p\in[P]}\left\|\boldsymbol{x}_i^{(p)}\right\|\right)^2\left\|\widetilde{\boldsymbol{W}} - \overline{\boldsymbol{W}}\right\|,$$

and

$$\begin{aligned}\left\|\nabla\mathcal{L}_{\text{CutMix}}(\widetilde{\boldsymbol{W}}) - \nabla\mathcal{L}_{\text{CutMix}}(\overline{\boldsymbol{W}})\right\| &\leq 4r^{-1}P\left(\max_{i\in[n],p\in[P]}\left\|\boldsymbol{x}_i^{(p)}\right\|\right)^2\left\|\widetilde{\boldsymbol{W}} - \overline{\boldsymbol{W}}\right\| \\ &\leq 9r^{-1}P\sigma_{\mathrm{d}}^2 d\left\|\widetilde{\boldsymbol{W}} - \overline{\boldsymbol{W}}\right\|,\end{aligned}$$

where the last inequality holds since $\left\|\xi_i^{(p)}\right\|^2 < \frac{3}{2}\sigma_{\mathrm{d}}^2 d$ and $\alpha^2 \leq \frac{3}{4}\sigma_{\mathrm{d}}^2 d$. Hence, $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$ is L-smooth with $L := 9r^{-1}P\sigma_{\mathrm{d}}^2 d$. ■

Since our objective function $\mathcal{L}_{\text{CutMix}}(\boldsymbol{W})$ is L-smooth and $\eta \leq \frac{1}{L}$, descent lemma (see Lemma 3.4 in Bubeck et al. [2]) implies

$$\mathcal{L}_{\text{CutMix}}\left(\boldsymbol{W}^{(t+1)}\right) - \mathcal{L}_{\text{CutMix}}\left(\boldsymbol{W}^{(t)}\right) \leq -\frac{\eta}{2}\left\|\nabla\mathcal{L}_{\text{CutMix}}\left(\boldsymbol{W}^{(t)}\right)\right\|^2,$$

and by telescoping sum, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla\mathcal{L}_{\text{CutMix}}\left(\boldsymbol{W}^{(t)}\right)\right\|^2 \leq \frac{2\eta\mathcal{L}_{\text{CutMix}}\left(\boldsymbol{W}^{(0)}\right)}{T} = \frac{\Theta(1)}{\eta T}, \tag{18}$$

for any $T > 0$.

Choose $\epsilon = \frac{1}{\text{poly}(d)}$ so that $2\beta^{-1}\epsilon < \left\|\hat{\boldsymbol{Z}}\right\|_\infty$ and $\mu^{-1}\beta^{-1}\epsilon = \frac{1}{\text{poly}(d)}$. Then from (18), there exists $T_{\text{CutMix}} \leq \frac{\text{poly}(d)}{\eta}$ such that

$$\left\|\nabla\mathcal{L}_{\text{CutMix}}\left(\boldsymbol{W}^{(T_{\text{CutMix}})}\right)\right\| \leq \epsilon.$$

From characterization of $\sigma_{\min}(\boldsymbol{J}(\boldsymbol{W}))$ in Section H.2.1,

$$\epsilon \geq \left\|\nabla\mathcal{L}\left(\boldsymbol{W}^{(T_{\text{CutMix}})}\right)\right\| \geq \sigma_{\min}(\boldsymbol{J}(\boldsymbol{W}))\left\|\nabla h\left(\boldsymbol{Z}^{(T_{\text{CutMix}})}\right)\right\| \geq \frac{\beta}{2}\left\|\nabla h\left(\boldsymbol{Z}^{(T_{\text{CutMix}})}\right)\right\|,$$

and thus

$$\left\|\nabla h\left(\boldsymbol{Z}^{(T_{\text{CutMix}})}\right)\right\| \leq 2\beta^{-1}\epsilon.$$

By Lemma 32 we have seen in Section H.2.4,

$$\left\|\boldsymbol{Z}^{(T_{\text{CutMix}})} - \hat{\boldsymbol{Z}}\right\| \leq 2\mu^{-1}\beta^{-1}\epsilon = \frac{1}{\text{poly}(d)},$$

and thus

$$y_i f_{\boldsymbol{W}^{(T_{\text{CutMix}})}}\left(\boldsymbol{x}_i^{(p)}\right) = \Theta(1),$$

for all $i \in [n]$ and $p \in [P]$, and therefore it reaches perfect training accuracy.

H.2.6. TEST ACCURACY OF SOLUTION FOUND BY GRADIENT DESCENT

The final step is showing that $\boldsymbol{W}^{(T_{\text{CutMix}})}$ reaches almost perfect test accuracy.

From the results of Section H.2.5, we have

$$\begin{aligned}\phi\left(\left\langle \boldsymbol{w}_s^{(T_{\text{CutMix}})}, \boldsymbol{v}_{s,k}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(T_{\text{CutMix}})}, \boldsymbol{v}_{s,k}\right\rangle\right) &= \Theta(1),\\ \phi\left(\left\langle \boldsymbol{w}_{y_i}^{(T_{\text{CutMix}})}, \xi_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(T_{\text{CutMix}})}, \xi_i^{(p)}\right\rangle\right) &= \Theta(1),\end{aligned}$$

for each $s \in \{\pm 1\}, k \in [K], i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$.

For any $u > v$, by the mean value theorem, we have

$$\beta(u - v) \leq \phi(u) - \phi(v) = (u - v)\frac{\phi(u) - \phi(v)}{u - v} \leq (u - v).$$

Hence, we have

$$\begin{aligned}\phi\left(\left\langle \boldsymbol{w}_s^{(T_{\text{CutMix}})}, \boldsymbol{v}_{s,k}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(T_{\text{CutMix}})}, \boldsymbol{v}_{s,k}\right\rangle\right) &\leq \left\langle \boldsymbol{w}_s^{(T_{\text{CutMix}})} - \boldsymbol{w}_{-s}^{(T_{\text{CutMix}})}, \boldsymbol{v}_{s,k}\right\rangle,\\ \left\langle \boldsymbol{w}_s^{(T_{\text{CutMix}})} - \boldsymbol{w}_{-s}^{(T_{\text{CutMix}})}, \boldsymbol{v}_{s,k}\right\rangle &\leq \beta^{-1}\left(\phi\left(\left\langle \boldsymbol{w}_s^{(T_{\text{CutMix}})}, \boldsymbol{v}_{s,k}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-s}^{(T_{\text{CutMix}})}, \boldsymbol{v}_{s,k}\right\rangle\right)\right),\end{aligned}$$

and

$$\Omega(1) \leq \left\langle \boldsymbol{w}_s^{(T_{\text{CutMix}})} - \boldsymbol{w}_{-s}^{(T_{\text{CutMix}})}, \boldsymbol{v}_{s,k}\right\rangle \leq \mathcal{O}(\beta^{-1}),$$

for each $s \in \{\pm 1\}$ and $k \in [K]$. Similarly, for all $i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$,

$$\begin{aligned}\phi\left(\left\langle \boldsymbol{w}_{y_i}^{(T_{\text{CutMix}})}, \xi_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(T_{\text{CutMix}})}, \xi_i^{(p)}\right\rangle\right) &\leq \left\langle \boldsymbol{w}_{y_i} - \boldsymbol{w}_{-y_i}, \xi_i^{(p)}\right\rangle,\\ \left\langle \boldsymbol{w}_{y_i}^{(T_{\text{CutMix}})} - \boldsymbol{w}_{-y_i}^{(T_{\text{CutMix}})}, \xi_i^{(p)}\right\rangle &\leq \beta^{-1}\left(\phi\left(\left\langle \boldsymbol{w}_{y_i}^{(T_{\text{CutMix}})}, \xi_i^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y_i}^{(T_{\text{CutMix}})}, \xi_i^{(p)}\right\rangle\right)\right),\end{aligned}$$

and

$$\Omega(1) \leq \left\langle \boldsymbol{w}_{y_i}^{(T_{\text{CutMix}})} - \boldsymbol{w}_{-y_i}^{(T_{\text{CutMix}})}, \xi_i^{(p)}\right\rangle \leq \mathcal{O}(\beta^{-1}).$$

By Lemma 10,

$$\begin{aligned}&\boldsymbol{w}_1^{(T_{\text{CutMix}})} - \boldsymbol{w}_{-1}^{(T_{\text{CutMix}})}\\ &= \boldsymbol{w}_1^{(0)} - \boldsymbol{w}_{-1}^{(0)} + \sum_{s \in \{\pm 1\}, k \in [K]} s\gamma(s,k)\boldsymbol{v}_{s,k} + \sum_{i \in [n], p \in [P] \setminus \{p_i^*\}} y_i \rho(i,p)\frac{\xi_i^{(p)}}{\left\lVert \xi_i^{(p)} \right\rVert^2},\end{aligned}$$

where for each $s \in \{\pm 1\}$,

$$\begin{aligned}\gamma(s,1) = \gamma_1^{(T_{\text{CutMix}})}(s,1) &+ \gamma_{-1}^{(T_{\text{CutMix}})}(s,1)\\ &+ \sum_{i \in \mathcal{F}_s} y_i \left(\rho_1^{(T_{\text{CutMix}})}(i,\tilde{p}_i) + \rho_{-1}^{(T_{\text{CutMix}})}(i,\tilde{p}_i)\right)\left\lVert \xi_i^{(\tilde{p}_i)} \right\rVert^{-2},\end{aligned}$$

and

$$\begin{aligned}\gamma(s,k) &= \gamma_1^{(T_{\text{CutMix}})}(s,k) + \gamma_{-1}^{(T_{\text{CutMix}})}(s,k),\\ \rho(i,p) &= \rho_1^{(T_{\text{CutMix}})}(i,p) + \rho_{-1}^{(T_{\text{CutMix}})}(i,p),\end{aligned}$$

for each $s \in \{\pm 1\}, k \in [K] \setminus \{1\}, i \in [n]$ and $p \in [P] \setminus \{p_i^*\}$. If we choose $j \in [n], q \in [P] \setminus \{p_j^*\}$ such that $\rho(j,q) = \max_{i \in [n], p \in \setminus \{p_i^*\}} \rho(i,p)$, then we have

$$\left\langle \boldsymbol{w}_{y_j}^{(T_{\mathrm{CutMix}})} - \boldsymbol{w}_{-y_j}^{(T_{\mathrm{CutMix}})}, \xi_j^{(q)} \right\rangle = \left\langle \boldsymbol{w}_{y_j}^{(0)} - \boldsymbol{w}_{-y_j}^{(0)}, \xi_j^{(q)} \right\rangle + \rho(j,q) + \sum_{\substack{i \in [n], p \in [P] \setminus \{p_i^*\} \\ (i,p) \neq (j,q)}} y_i \rho(i,p) \frac{\left\langle \xi_i^{(p)}, \xi_j^{(q)} \right\rangle}{\left\| \xi_i^{(p)} \right\|^2}.$$

From Lemma 9,

$$\left| \left\langle \boldsymbol{w}_{y_j}^{(0)} - \boldsymbol{w}_{-y_j}^{(0)}, \xi_j^{(q)} \right\rangle \right| = \widetilde{\mathcal{O}}\left(\sigma_0 \sigma_{\mathrm{d}} d^{\frac{1}{2}} \right) \leq \frac{1}{2} \left\langle \boldsymbol{w}_{y_j}^{(T_{\mathrm{CutMix}})} - \boldsymbol{w}_{-y_j}^{(T_{\mathrm{CutMix}})}, \xi_j^{(q)} \right\rangle,$$

where the inequality holds since $\left\langle \boldsymbol{w}_{y_j}^{(T_{\mathrm{CutMix}})} - \boldsymbol{w}_{-y_j}^{(T_{\mathrm{CutMix}})}, \xi_j^{(q)} \right\rangle = \Omega(1)$. In addition, by triangular inequality, we have

$$\begin{aligned}
\left| \sum_{\substack{i \in [n], p \in [P] \setminus \{p_i^*\} \\ (i,p) \neq (j,q)}} y_i \rho(i,p) \frac{\left\langle \xi_i^{(p)}, \xi_j^{(q)} \right\rangle}{\left\| \xi_i^{(p)} \right\|^2} \right| &\leq \sum_{(i,p) \neq (j,q)} \rho(i,p) \frac{\left| \left\langle \xi_i^{(p)}, \xi_j^{(q)} \right\rangle \right|}{\left\| \xi_i^{(p)} \right\|^2} \\
&\leq \rho(j,q) \mathcal{O}(n \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}) \leq \frac{\rho(j,q)}{2}.
\end{aligned}$$

Hence,

$$\frac{1}{3} \rho(j,q) \leq \left| \left\langle \boldsymbol{w}_{y_j}^{(T_{\mathrm{CutMix}})} - \boldsymbol{w}_{-y_j}^{(T_{\mathrm{CutMix}})}, \xi_j^{(q)} \right\rangle \right| \leq 3 \rho(j,q)$$

and we have $\rho(j,q) = \mathcal{O}(\beta^{-1})$.

Let $(\boldsymbol{X}, y) \sim \mathcal{D}$ be a test data with $\boldsymbol{X} = \left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(P)} \right) \in \mathbb{R}^{d \times P}$ having feature patch p^*, dominant noise patch $\tilde{p}$, and feature vector $\boldsymbol{v}_{y,k}$. We have $\boldsymbol{x}^{(p)} \sim N(\boldsymbol{0}, \sigma_{\mathrm{b}}^2 \boldsymbol{\Lambda})$ for each $p \in [P] \setminus \{p^*, \tilde{p}\}$ and $\boldsymbol{x}^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1} \sim N(\boldsymbol{0}, \sigma_{\mathrm{d}}^2 \boldsymbol{\Lambda})$ for some $s \in \{\pm 1\}$. Therefore, for all $p \in [P] \setminus \{p^*, \tilde{p}\}$

$$\begin{aligned}
&\left| \phi\left(\left\langle \boldsymbol{w}_1^{(T_{\mathrm{CutMix}})}, \boldsymbol{x}^{(p)} \right\rangle \right) - \phi\left(\left\langle \boldsymbol{w}_{-1}^{(T_{\mathrm{CutMix}})}, \boldsymbol{x}^{(p)} \right\rangle \right) \right| \\
&\leq \left| \left\langle \boldsymbol{w}_1^{(T_{\mathrm{CutMix}})} - \boldsymbol{w}_{-1}^{(T_{\mathrm{CutMix}})}, \boldsymbol{x}^{(p)} \right\rangle \right| \\
&= \left| \left\langle \boldsymbol{w}_1^{(0)} - \boldsymbol{w}_{-1}^{(0)}, \boldsymbol{x}^{(p)} \right\rangle \right| + \sum_{i \in [n], q \in [P] \setminus \{p_i^*\}} \rho(i,q) \frac{\left| \left\langle \xi_i^{(q)}, \boldsymbol{x}^{(p)} \right\rangle \right|}{\left\| \xi_i^{(q)} \right\|^2} \\
&\leq \widetilde{\mathcal{O}}\left(\sigma_0 \sigma_{\mathrm{b}} d^{\frac{1}{2}} \right) + \widetilde{\mathcal{O}}\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right) \\
&= \widetilde{\mathcal{O}}\left(n \beta^{-1} \sigma_{\mathrm{d}} \sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}} \right),
\end{aligned} \tag{19}$$

with probability at least $1 - o\left(\frac{1}{\mathrm{poly}(d)}\right)$ due to Lemma 9. In addition,

$$\begin{aligned}
&\left|\phi\left(\left\langle \boldsymbol{w}_1^{(T_{\mathrm{CutMix}})}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-1}^{(T_{\mathrm{CutMix}})}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right)\right| \\
&\leq \left|\left\langle \boldsymbol{w}_1^{(T_{\mathrm{CutMix}})} - \boldsymbol{w}_{-1}^{(T_{\mathrm{CutMix}})}, \boldsymbol{x}^{(\tilde{p})}\right\rangle\right| \\
&\leq \alpha \left|\left\langle \boldsymbol{w}_1^{(T_{\mathrm{CutMix}})} - \boldsymbol{w}_{-1}^{(T_{\mathrm{CutMix}})}, \boldsymbol{v}_{s,1}\right\rangle\right| + \left|\left\langle \boldsymbol{w}_1^{(T_{\mathrm{CutMix}})} - \boldsymbol{w}_{-1}^{(T_{\mathrm{CutMix}})}, \boldsymbol{x}^{(p)} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right| \\
&\leq \alpha\beta^{-1} \left|\phi\left(\left\langle \boldsymbol{w}_1^{(T_{\mathrm{CutMix}})}, \boldsymbol{v}_{s,1}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-1}^{(T_{\mathrm{CutMix}})}, \boldsymbol{v}_{s,1}\right\rangle\right)\right| \\
&\quad + \left|\left\langle \boldsymbol{w}_1^{(0)} - \boldsymbol{w}_{-1}^{(0)}, \boldsymbol{x}^{(p)} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right| + \sum_{i\in[n], q \in [P]\setminus\{p_i^*\}} \rho(i,q) \frac{\left|\left\langle \xi_i^{(q)}, \boldsymbol{x}^{(\tilde{p})} - \alpha \boldsymbol{v}_{s,1}\right\rangle\right|}{\left\|\xi_i^{(q)}\right\|^2} \\
&\leq \mathcal{O}\left(\alpha\beta^{-1}\right) + \widetilde{\mathcal{O}}\left(\sigma_0\sigma_{\mathrm{d}} d^{\frac{1}{2}}\right) + \widetilde{\mathcal{O}}\left(n\beta^{-1}\sigma_{\mathrm{d}}\sigma_{\mathrm{b}}^{-1} d^{-\frac{1}{2}}\right) \\
&= \mathcal{O}(\alpha\beta^{-1}),
\end{aligned} \tag{20}$$

with probability at least $1 - o\left(\frac{1}{\mathrm{poly}(d)}\right)$.

Suppose (19) and (20) holds. Then,

$$\begin{aligned}
&y f_{\boldsymbol{W}^{(T_{\mathrm{CutMix}})}}(\boldsymbol{X}) \\
&= \left(\phi\left(\left\langle \boldsymbol{w}_y^{(T_{\mathrm{CutMix}})}, \boldsymbol{v}_{y,k}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(T_{\mathrm{CutMix}})}, \boldsymbol{v}_{y,k}\right\rangle\right)\right) \\
&\quad + \sum_{p\in[P]\setminus\{p^*\}} \left(\phi\left(\left\langle \boldsymbol{w}_y^{(T_{\mathrm{CutMix}})}, \boldsymbol{x}^{(p)}\right\rangle\right) - \phi\left(\left\langle \boldsymbol{w}_{-y}^{(T_{\mathrm{CutMix}})}, \boldsymbol{x}^{(p)}\right\rangle\right)\right) \\
&= \Omega(1) - \mathcal{O}\left(\alpha\beta^{-1}\right) \\
&> 0.
\end{aligned}$$

Hence, we have our conclusion. □

Appendix I. Technical Lemmas

In this section, we introduce technical lemmas that are used for proving the main theorems. We present their proofs here for better readability.

The following lemma is used in Section F.2.4:

Lemma 34 *For any $z, \delta \in \mathbb{R}$,*

$$\left|\phi(z) - (z+\delta)\phi'(z)\right| \leq r + |\delta|.$$

Proof [Proof of Lemma 34]

$$\phi(z) - z\phi'(z) = \begin{cases} z - \frac{1-\beta}{2}r - z = -\frac{1-\beta}{2}r = -\frac{1-\beta}{2}r & \text{if } z \geq r \\ \frac{1-\beta}{2r}z^2 + \beta z - \left(\frac{1-\beta}{r}z + \beta\right)z = \frac{1-\beta}{2r}z^2 & \text{if } 0 \leq z \leq r \\ \beta z - \beta z = 0 & \text{if } z < 0 \end{cases},$$

and we obtain

$$\left|\phi(z) - (z+\delta)\phi'(z)\right| \leq \left|\phi(z) - z\phi'(z)\right| + |\delta|\phi'(z) \leq \frac{1-\beta}{2}r + |\delta| \leq r + |\delta|.$$

■

The following lemma is used in Section F.2.4.

Lemma 35 *Suppose E_{init} occurs. Then, for any model parameter $\boldsymbol{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_{-1}\}$, we have*

$$\left\|\nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s,k}} \ell\left(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_i)\right)\right\|^2 \leq 8P^2\sigma_{\mathrm{d}}^2 d|\mathcal{V}_{s,k}| \sum_{i \in \mathcal{V}_{s,k}} \ell(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_i)),$$

for each $s \in \{\pm 1\}$ and $k \in [K]$.

Proof [Proof of Lemma 35] For each $s \in \{\pm 1\}$ and $i \in [n]$, we have

$$\|\nabla_{\boldsymbol{w}_s} f_{\boldsymbol{W}}(\boldsymbol{X}_i)\| = \left\|\sum_{p \in [P]} \phi'\left(\left\langle \boldsymbol{w}_s, \boldsymbol{x}_i^{(p)}\right\rangle\right) \boldsymbol{x}_i^{(p)}\right\| \leq P \max_{p \in [P]} \left\|\boldsymbol{x}_i^{(p)}\right\| \leq 2P\sigma_{\mathrm{d}} d^{\frac{1}{2}},$$

here the inequality is due to E_{init}. Therefore, we have

$$\begin{aligned}\left\| \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s,k}} \ell\left(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_i)\right) \right\|^2 &= \left\| \sum_{i \in \mathcal{V}_{s,k}} \ell'\left(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_i)\right) \nabla_{\boldsymbol{W}} f_{\boldsymbol{W}}(\boldsymbol{X}_i) \right\|^2 \\ &\leq \left(\sum_{i \in \mathcal{V}_{s,k}} \ell'\left(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_i)\right) \|\nabla_{\boldsymbol{W}} f_{\boldsymbol{W}}(\boldsymbol{X}_i)\| \right)^2 \\ &\leq 4P^2 \sigma_{\text{d}}^2 d \left(\sum_{i \in \mathcal{V}_{s,k}} \ell'\left(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_i)\right) \right)^2 \\ &\leq 4P^2 \sigma_{\text{d}}^2 d |\mathcal{V}_{s,k}| \sum_{i \in \mathcal{V}_{s,k}} \left(\ell'\left(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_i)\right)\right)^2 \\ &\leq 4P^2 \sigma_{\text{d}}^2 d |\mathcal{V}_{s,k}| \sum_{i \in \mathcal{V}_{s,k}} \ell\left(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_i)\right).\end{aligned}$$

The first inequality is due to triangular inequality, the third inequality is due to Cauchy-Schwartz inequality and the last inequality is due to $0 \leq -\ell' \leq 1$, which can be used to show $(\ell')^2 \leq -\ell' \leq \ell$. ■

The following lemma is used in Section G.2.4.

Lemma 36 *Suppose E_{init} occurs. Then, for any model parameter $\boldsymbol{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_{-1}\}$, we have*

$$\left\| \nabla \sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}}\left[\ell\left(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}})\right)\right] \right\|^2 \leq 8P^2 \sigma_{\text{d}}^2 d |\mathcal{V}_{s,k}| \sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}}[\ell(y_i f_{\boldsymbol{W}^{(t)}}(\boldsymbol{X}_{i,\mathcal{C}}))]$$

for each $s \in \{\pm 1\}$ and $k \in [K]$.

Proof [Proof of Lemma 36] For each $s \in \{\pm 1\}$, $i \in [n]$ and $\mathcal{C} \subset [P]$ with $|\mathcal{C}| = C$, we have

$$\|\nabla_{\boldsymbol{w}_s} f_{\boldsymbol{W}}(\boldsymbol{X}_{i,\mathcal{C}})\| = \left\| \sum_{p \notin \mathcal{C}} \phi'\left(\left\langle \boldsymbol{w}_s, \boldsymbol{x}_i^{(p)} \right\rangle\right) \boldsymbol{x}_i^{(p)} \right\| \leq P \max_{p \in [P]} \left\| \boldsymbol{x}_i^{(p)} \right\| \leq 2P\sigma_{\text{d}} d^{\frac{1}{2}},$$

here the inequality is due to E_{init}. Therefore, we have

$$\begin{aligned}
\left\| \nabla_{\boldsymbol{W}} \sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} [\ell\left(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_{i,\mathcal{C}})\right)] \right\|^2 &= \left\| \sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} [\ell'\left(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_{i,\mathcal{C}})\right) \nabla_{\boldsymbol{W}} f_{\boldsymbol{W}}(\boldsymbol{X}_{i,\mathcal{C}})] \right\|^2 \\
&\leq \left(\sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell'\left(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_{i,\mathcal{C}})\right) \| \nabla_{\boldsymbol{W}} f_{\boldsymbol{W}}(\boldsymbol{X}_{i,\mathcal{C}}) \| \right] \right)^2 \\
&\leq 4P^2 \sigma_{\mathrm{d}}^2 d \left(\sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\ell'\left(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_{i,\mathcal{C}})\right) \right] \right)^2 \\
&\leq 4P^2 \sigma_{\mathrm{d}}^2 d |\mathcal{V}_{s,k}| \sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} \left[\left(\ell'\left(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_i)\right) \right)^2 \right] \\
&\leq 4P^2 \sigma_{\mathrm{d}}^2 d |\mathcal{V}_{s,k}| \sum_{i \in \mathcal{V}_{s,k}} \mathbb{E}_{\mathcal{C} \sim \mathcal{D}_{\mathcal{C}}} [\ell\left(y_i f_{\boldsymbol{W}}(\boldsymbol{X}_{i,\mathcal{C}})\right)].
\end{aligned}$$

The first inequality is due to triangular inequality, the third inequality is due to Cauchy-Schwartz inequality and the last inequality is due to $0 \leq -\ell' \leq 1$, which can be used to show $(\ell')^2 \leq -\ell' \leq \ell$. ■