# Why Does Pruning during Training Work? A Signal-to-Noise Analysis of Sparse Neural Network Training

Shiyuan Ren[1] 24S051009@stu.hit.edu.cn
Jiarui Jiang[1] jiaruij@outlook.com
Miao Zhang[1] zhangmiao@hit.edu.cn
Xiang Deng[1] dengxiang@hit.edu.cn
Min Zhang[1] zhangmin2021@hit.edu.cn
Liqiang Nie[1] nieliqiang@gmail.com
[1]Harbin Institute of Technology, Shenzhen

## Abstract

Network pruning is an effective strategy to alleviate the computational burden of large neural networks for both training and inference, substantially reducing storage and computational costs while preserving predictive capacity. Despite its strong empirical success, the theoretical understanding of pruning during training remains limited, even in simple settings such as two-layer $\mathrm{ReLU}$ convolutional neural networks. As a result, the lack of a sufficient understanding makes it unclear when and why training-time pruning is effective, obscuring their underlying value. In this work, we provide a theoretical analysis of pruning during training for practical two-layer $\mathrm{ReLU}$ CNNs using a signal-to-noise decomposition framework. We characterize the learning dynamics induced by dynamic pruning and establish conditions under which sparse pruned networks achieve generalization performance comparable to, or even strictly better than, their dense counterparts. To the best of our knowledge, this work presents the first theoretical study of training-time pruning in $\mathrm{ReLU}$ neural networks, and our theoretical predictions are further validated through experiments.

## 1. Introduction

Neural network pruning reduces model complexity by removing unimportant parameters (LeCun et al., 1989), and has been widely used to decrease inference and storage costs in over-parameterized deep models while largely preserving performance. According to when pruning is applied, existing methods are typically categorized into pruning before training (PBT) (Frankle & Carbin, 2019; Malach et al., 2020; da Cunha et al., 2022), pruning during training (PDT) (Guo et al., 2020; Shen et al., 2024), and pruning after training (PAT) (Benbaki et al., 2023; Zou et al., 2025; Chen et al., 2025).

Sparse training enforces sparsity throughout the entire optimization process, encompassing PBT with static sparsity and PDT with more complex dynamic sparsity strategies, in contrast to PAT, which is applied only after training is completed. This paradigm can substantially reduce computational costs during both training and inference, and prior studies have shown that it can achieve performance comparable to, or even exceeding, that of dense training in large-scale applications (Liu, 2020; Jin et al., 2022; Shen et al., 2024). Moreover, sparse training is increasingly being adopted in large language models (Xia et al., 2024; Mozaffari et al., 2025; Zhu et al., 2025).

A growing body of research has sought to theoretically characterize the properties of sparse neural networks and explain various phenomena induced by sparsity. For instance, Muthukumar & Sulam (2023) derived data-dependent PAC-Bayesian generalization bounds for deep $\mathrm{ReLU}$ networks by exploiting activation sparsity, while Galanti et al. (2023) established norm-based generalization bounds that explicitly capture the effects of sparse structures and convolutional architectures. As a complement to these approaches, other researchers have investigated pruning techniques from an optimization perspective, correlating sparsity with parameter sensitivity and the flatness of the loss landscape (Tartaglione et al., 2018; Lee et al., 2025). Despite these advances in the field, most existing theoretical frameworks remain weakly connected to the training dynamics of practical deep neural networks. In this work, we

*Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).*

precisely characterize the dynamic evolution of network parameters based on a more practically grounded theoretical framework—the signal-to-noise model (Cao et al., 2022)—without introducing overly strong assumptions.

The most closely related work is Yang et al. (2025), which theoretically studies two-layer convolutional networks pruned at initialization and then trained by gradient descent, showing that mild pruning can improve generalization while excessive pruning is detrimental. In contrast, their setting belongs to PBT-based sparse training with static sparsity patterns, whereas our work focuses on PDT-based dynamic sparse training, where sparsity evolves during optimization and interacts nontrivially with gradient descent. This dynamic setting is considerably more challenging to analyze and is also more representative of practical training. Moreover, Yang et al. (2025) relies on $\mathrm{ReLU}^q$ activations with $q > 2$, leaving open the role of the standard $\mathrm{ReLU}$ activation and the impact of training-time pruning on learning dynamics—both of which are central to our study.

Motivated by these gaps, we take a first step toward understanding dynamic pruning mechanisms during training in multi-class $\mathrm{ReLU}$ neural networks. In particular, we consider training-time TopK weight pruning and investigate the following question:

*When sparse training is performed via dynamic pruning, how do the network parameters evolve over time, and in what ways does pruning affect convergence and generalization performance?*

Our main contributions are summarized as follows:

- We characterize the learning dynamics of multi-class $\mathrm{ReLU}$ neural networks trained via gradient descent, and derive explicit conditions that distinguish successful signal learning from noise-dominated regimes. By capturing the competition between signal and noise, we provide a theoretical explanation for the boundary between these two behaviors during dynamic sparse training(see Theorem 4.2 and 4.3).
- We provide the first theoretical elucidation of the intrinsic mechanism by which dynamic weight pruning during training improves generalization. We show that under specific conditions, dynamic sparse training leads to better performance than dense training, providing a rigorous theoretical foundation for previous empirical observations (see Theorem 4.4 and 4.5).
- We validate our theoretical results through experiments. The empirical findings are highly consistent with theoretical predictions, further corroborating the soundness of the proposed theoretical analysis.

**Notation.** For two sequences $\{x_n\}$ and $\{y_n\}$, we use standard asymptotic notations $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$. Their logarithmic variants $\tilde{O}(\cdot)$, $\tilde{\Omega}(\cdot)$, and $\tilde{\Theta}(\cdot)$ hide polylogarithmic factors. We write $x_n = \mathrm{poly}(y_n)$ if $x_n = O(y_n^D)$ for some constant $D > 0$, and $x_n = \mathrm{polylog}(y_n)$ if $x_n = \mathrm{poly}(\log(y_n))$.

## 2. Related Work

**Sparse Training and Dynamic Pruning.**

Sparse training methods are broadly divided into PBT (Lee et al., 2018; Wang et al., 2019; Tanaka et al., 2020; Chen et al., 2021) and PDT. Representative PDT approaches include Network Slimming (Liu et al., 2017), which enforces channel-level sparsity in CNNs during training; Sparse Structure Selection (Huang & Wang, 2018), which learns structural scaling factors with sparsity regularization for adaptive pruning; Differentiable Sparsity Allocation (DSA) (Ning et al., 2020), which allocates cross-layer sparsity via differentiable pruning and gradient-based optimization; and GraNet (Liu et al., 2021), which introduces gradual magnitude pruning with neuroregeneration, achieving strong sparse-to-sparse performance without extra training cost.

Dynamic pruning methods are a common strategy within the PDT setting and have also been widely adopted in other domains, including deep reinforcement learning (Ceron et al., 2024), neural networks (Kim et al., 2024), general machine learning (Guyard et al., 2024), and large language models (Le et al., 2025; Yin et al., 2025). Our work primarily focuses on dynamic pruning in the PDT regime, with particular emphasis on TopK weight pruning.

**Theoretical Interpretation of Neural Networks.** A substantial body of work analyzes neural network training dynamics in the Neural Tangent Kernel (NTK) regime. Early studies show that infinitely wide networks converge to deterministic kernels and admit convergence and generalization guarantees under gradient descent (Jacot et al., 2018;

Allen-Zhu et al., 2019; Cao & Gu, 2019). However, NTK-based analyses characterize a lazy training regime in which parameters remain close to initialization, limiting feature learning (Chizat et al., 2019).

Motivated by this limitation, recent work has shifted toward understanding neural networks in the feature learning regime. Prior studies analyze ensemble effects and knowledge distillation (Allen-Zhu & Li, 2023), as well as benign overfitting phenomena in overparameterized CNNs, showing sharp phase transitions governed by the signal-to-noise ratio (Cao et al., 2022; Kou et al., 2023). Related to pruning, Yang et al. (2025) study pruning effects under random initialization for two-layer CNNs. In contrast, our work focuses on standard $\mathrm{ReLU}$ networks and provides a theoretical analysis of dynamic pruning during training.

## 3. Problem Setup

In this section, we introduce the data generation model and the convolutional neural network we consider in this paper.

**Definition 3.1.** Consider we are given the set of signal vectors $\{\mu \mathbf{e}_i\}_{i=1}^{\mathcal{T}}$, where $\mu > 0$ denotes the strength of the signal, and $\mathbf{e}_i$ denotes the $i$-th standard basis vector with its $i$-th entry being 1 and all other coordinates being 0. Each data point $(\mathbf{x}, y)$ with $\mathbf{x} = [\mathbf{x}_1^\top, \mathbf{x}_2^\top]^\top \in \mathbb{R}^{2d}$ and $y \in [\mathcal{T}]$ is generated from the following distribution $\mathcal{D}$:

1. The label $y$ is generated from a uniform distribution over $[\mathcal{T}]$.
2. A noise vector $\boldsymbol{\xi} \in \mathbb{R}^d$ is generated from the Gaussian distribution $\mathcal{N}(0, \sigma_p^2 I_{d-\mathcal{T}})$, where the first $\mathcal{T}$ coordinates of $\boldsymbol{\xi}$ are all zero.
3. With probability $1/2$, assign $\mathbf{x}_1 = \boldsymbol{\mu}_y$, $\mathbf{x}_2 = \boldsymbol{\xi}$; with probability $1/2$, assign $\mathbf{x}_2 = \boldsymbol{\mu}_y$, $\mathbf{x}_1 = \boldsymbol{\xi}$ where $\boldsymbol{\mu}_y = \mu \mathbf{e}_y$.

This sparse signal model is motivated by the framework established in Yang et al. (2025). By ensuring the first $\mathcal{T}$ entries of the noise vector $\boldsymbol{\xi}$ are zero, we maintain strict orthogonality between the noise $\boldsymbol{\xi}$ and the signal vector $\boldsymbol{\mu}_y$. This design choice aligns with the "benign overfitting" philosophy proposed by Cao et al. (2022), allowing for a clear separation between informative features and stochastic noise.

**Network architecture and Loss Function.** We consider a two-layer convolutional neural network (CNN) employing the $\mathrm{ReLU}$ activation function, $\sigma(z) = \max\{0, z\}$. This architecture is designed to process partitioned inputs, such as the patch-based data $(\mathbf{x}_1, \mathbf{x}_2)$ defined previously. For a given data point $(\mathbf{x}, y)$, the network produces a multi-class output vector:

$$F(\mathbf{W}, \mathbf{x}) = (F_1(\mathbf{W}_1, \mathbf{x}), F_2(\mathbf{W}_2, \mathbf{x}), \ldots, F_{\mathcal{T}}(\mathbf{W}_{\mathcal{T}}, \mathbf{x}))\,.$$

The $j$-th component of the output, $F_j(\mathbf{W}_j, \mathbf{x})$, represents the network's "score" or logit for class $j$, computed as:

$$\begin{aligned} F_j(\mathbf{W}_j, \mathbf{x}) &= \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}_1\rangle) + \sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}_2\rangle)] \\ &= \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{j,r}, \boldsymbol{\mu}_y\rangle) + \sigma(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}\rangle)]. \end{aligned}$$

In this formulation, $\mathbf{w}_{j,r} \in \mathbb{R}^d$ represents the $r$-th filter (or neuron) weight vector within the $j$-th class-specific weight set $\mathbf{W}_j$. The summation over patches $\mathbf{x}_1$ and $\mathbf{x}_2$ demonstrates that the network is weight-sharing across different locations of the input. This structure is equivalent to a CNN followed by Global Average Pooling (GAP) (Lin et al., 2013). By summing the activations of the signal patch and the noise patch, the network must learn to distinguish the invariant signal $\boldsymbol{\mu}_y$ from the stochastic noise $\boldsymbol{\xi}$, regardless of which patch they occupy. To train the model, we utilize the standard cross-entropy loss coupled with softmax normalization. The softmax function transforms the raw outputs into a probability distribution over the $\mathcal{T}$ classes. The probability assigned to the $i$-th class is given by:

$$\mathrm{logit}_i(F, \mathbf{x}) := \frac{\exp(F_i(\mathbf{W}_i, \mathbf{x}))}{\sum_{j=1}^{\mathcal{T}} \exp(F_j(\mathbf{W}_j, \mathbf{x}))}.$$

For a single training pair $(\mathbf{x}, y)$, the cross-entropy loss measures the discrepancy between the predicted distribution and the ground-truth label:

$$\ell(F(\mathbf{W}, \mathbf{x}), y) = -\log\left(\mathrm{logit}_y(F, \mathbf{x})\right).$$

The optimization objective is to minimize the Empirical Risk over a training dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$:

$$L_S(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(F(\mathbf{W}, \mathbf{x}_i), y_i\right).$$

While $L_S(\mathbf{W})$ guides the gradient descent process, our primary analytical interest lies in the Population Risk (or generalization loss), defined as the expected loss over the true distribution $\mathcal{D}$:

$$L_{\mathcal{D}}(\mathbf{W}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell\left(F(\mathbf{W}, \mathbf{x}), y\right)\right].$$

Minimizing $L_S$ allows the filters $\mathbf{w}_{j,r}$ to align with the signal $\boldsymbol{\mu}_y$. However, in high-dimensional settings, the network may also "overfit" to the noise $\boldsymbol{\xi}$ to reduce the empirical loss. The relationship between $L_S$ and $L_{\mathcal{D}}$ is central to understanding whether the CNN achieves harmful overfitting or truly learns the underlying signal.

**Training algorithm.** To optimize the empirical risk $L_S(\mathbf{W})$, we employ Gradient Descent (GD). The update rule for the $r$-th filter of the $j$-th class at iteration $t$ is given by:

$$\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} - \eta \nabla_{\mathbf{w}_{j,r}^{(t)}} L_S(\mathbf{W}^{(t)}).$$

By applying the chain rule to the composition of the cross-entropy loss and the CNN architecture, the update can be decomposed into signal-driven and noise-driven components:

$$\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} + \underbrace{\frac{\eta}{nm} \sum_{i=1}^{n} \left(-\ell_{j,i}^{'(t)}\right) \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle\right) \boldsymbol{\mu}_{y_i}}_{\text{Signal Learning Term}} + \underbrace{\frac{\eta}{nm} \sum_{i=1}^{n} \left(-\ell_{j,i}^{'(t)}\right) \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right) \boldsymbol{\xi}_i}_{\text{Noise Memorization Term}}.$$

The term $-\ell_{j,i}^{'(t)}$ represents the negative partial derivative of the loss with respect to the $j$-th logit for the $i$-th sample. Under the cross-entropy loss, this term simplifies to the residual error:

$$-\ell_{j,i}^{'(t)} = \mathbb{I}(j = y_i) - \text{logit}_j(F, \mathbf{x}_i).$$

When $j = y_i$: The term is $1 - \text{logit}_y$, which is always positive. This pushes the filter $\mathbf{w}_{y,r}$ to align with the signal $\boldsymbol{\mu}_y$ to increase the correct class logit.

When $j \neq y_i$: The term is $-\text{logit}_j$, which is negative. This acts as a competitive mechanism, pushing filters away from signals associated with noises. For the $\mathrm{ReLU}$ activation, we adopt the standard convention that the subgradient at the origin is $\sigma'(0) = 1$. This ensures that even neurons with zero initialization can potentially receive gradient updates. We initialize the network parameters using a Gaussian initialization scheme. Every entry of the initial weight tensor $\mathbf{W}^{(0)}$ is sampled independently from a normal distribution:

$$[\mathbf{w}_{j,r}^{(0)}]_k \sim \mathcal{N}(0, \sigma_0^2), \quad \forall j \in [\mathcal{T}], r \in [m], k \in [d],$$

where $\sigma_0^2$ is the initial variance.

**TopK Weight Pruning Strategy.** To reduce the computational complexity of the network and enhance its generalization ability by mitigating overfitting to noise, we adopt a TopK weight pruning strategy on the filter weight vectors $\mathbf{w}_{j,r}$ of the CNN. We denote the pruned weight vector of the $r$-th filter in the $j$-th class as $\widetilde{\mathbf{w}}_{j,r}$, and the complete set of pruned weights as $\widetilde{\mathbf{W}} = \{\widetilde{\mathbf{W}}_j\}_{j=1}^{\mathcal{T}}$ where $\widetilde{\mathbf{W}}_j = \{\widetilde{\mathbf{w}}_{j,r}\}_{r=1}^m$.

The pruning operation is performed on each weight vector $\mathbf{w}_{j,r} \in \mathbb{R}^d$ independently, following these steps:

1. Magnitude Evaluation: For the weight vector $\mathbf{w}_{j,r}$ at a given training iteration, we compute the magnitude of each of its entries to quantify the contribution of individual weight components to the filter response.
2. TopK Selection: We retain the TopK entries of $\mathbf{w}_{j,r}$ with the largest magnitudes, as these entries are considered to be the most critical for capturing the discriminative signal features $\boldsymbol{\mu}_y$.
3. Zero Masking: All remaining entries of $\mathbf{w}_{j,r}$ (i.e., those not in the TopK) are set to zero, yielding the pruned weight vector $\widetilde{\mathbf{w}}_{j,r}$.

Notably, our pruning strategy is seamlessly integrated with the gradient descent training process of the network. Specifically, we implement pruning in an iterative manner: after a certain number of gradient descent updates, we apply the TopK pruning to the current weight vectors $\mathbf{w}_{j,r}^{(t)}$, and then resume the gradient descent optimization using the pruned weights $\widetilde{\mathbf{w}}_{j,r}^{(t)}$ as the initial point for subsequent iterations. This iterative pruning-training scheme ensures that the network can adjust its weights to compensate for the pruned components, while the TopK selection criterion guarantees that only the non-critical weight entries (which are more likely to be associated with noise memorization) are removed.

After pruning, the output of the $j$-th class filter is modified to:

$$\begin{aligned} F_j(\widetilde{\mathbf{W}}_j, \mathbf{x}) &= \frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\langle\widetilde{\mathbf{w}}_{j,r}, \mathbf{x}_1\rangle\right) + \sigma\left(\langle\widetilde{\mathbf{w}}_{j,r}, \mathbf{x}_2\rangle\right)\right] \\ &= \frac{1}{m}\sum_{r=1}^{m}\left[\sigma\left(\langle\widetilde{\mathbf{w}}_{j,r}, \boldsymbol{\mu}_y\rangle\right) + \sigma\left(\langle\widetilde{\mathbf{w}}_{j,r}, \boldsymbol{\xi}\rangle\right)\right]. \end{aligned}$$

Correspondingly, the empirical risk and population risk are updated to $L_S(\widetilde{\mathbf{W}})$ and $L_{\mathcal{D}}(\widetilde{\mathbf{W}})$ respectively, which serve as the optimization objectives for the pruned network. By removing the weight entries with small magnitudes, our TopK pruning strategy effectively suppresses the noise memorization term in the gradient descent update rule, thereby promoting the network to focus on learning the invariant signal features $\boldsymbol{\mu}_y$ rather than overfitting to the stochastic noise $\boldsymbol{\xi}$.

## 4. Main Results

In this section, we present our main theoretical results, which systematically characterize the learning behaviors of $\mathrm{ReLU}$ CNNs trained with gradient descent (GD) and TopK pruning, including signal learning, noise memorization, and the regulatory effect of pruning on generalization performance. Our results are built on the following technical conditions that constrain key hyperparameters and model settings, including the weight dimension $d$, the standard deviation of Gaussian initialization $\sigma_0$, TopK pruning threshold $K$, and learning rate $\eta$. These conditions ensure the robustness and validity of our theoretical derivations under high confidence.

*Condition* 4.1. We have that

1. Dimension $d$ is sufficiently large: $d \geq \mathcal{T} + \Theta(512\mathcal{T}n^2\alpha^2\log\left(\frac{4n^2}{\delta}\right))$, and $d \geq \sigma_p^{-2}$.
2. The standard deviation of Gaussian initialization $\sigma_0$ is appropriately chosen such that $\sigma_0 \leq \frac{1}{16\sqrt{\log\left(\frac{8mn\mathcal{T}^2}{\delta}\right)}} \cdot \frac{1}{\mathcal{T}n^2} \cdot \min\left\{\frac{1}{\mu}, \frac{1}{\sigma_p\sqrt{d-\mathcal{T}}}\right\}$.
3. The learning rate $\eta$ satisfies $\eta \leq O\left(\frac{m\mathcal{T}}{\mu^2}\right)$.

Based on the above conditions, we first establish the core result for the signal learning regime of CNNs. This theorem clarifies the critical SNR threshold under which the model prioritizes learning discriminative signals over memorizing spurious noises, ultimately achieving favorable generalization performance.

**Theorem 4.2.** *For any $\epsilon > 0$, let $T = \widetilde{O}(m\mathcal{T}\eta^{-1}\mu^{-2} + \eta^{-1}\mathcal{T}mn\sigma_p^{-2}d^{-1}\epsilon^{-1})$. Under Condition 4.1, if* $\mathrm{SNR} \geq \Theta(\mathcal{T}n^{-1/4})$*, then with probability at least $1-\delta$, there exists $0 \leq t \leq T$ such that:*

1. *The CNN learns the signal:* $\max_{r,\tau} \gamma_{j,r,\tau}^{(t)} = \Omega(1)$ *for* $j \in [\mathcal{T}]$. *This indicates that at least one convolution filter in each signal channel* $j$ *successfully captures the core signal pattern, as the signal-aligned coefficient reaches a constant order of magnitude.*

2. *The CNN does not memorize the noises in the training data:* $\max_{j,r,i} |\rho_{j,r,i}| = \widetilde{O}(\mathcal{T} n^{-1} \mathrm{SNR}^{-2})$. *The noise-aligned coefficients are bounded by a negligible term, confirming that the model avoids overfitting to spurious noise components.*

3. *The training loss converges to* $\epsilon$, *i.e.,* $L_S(\mathbf{W}^{(t)}) \leq \epsilon$. *The model achieves sufficient fitting of the training data within a finite number of iterations* $T$.

4. *The trained CNN achieves a small test loss:* $L_{\mathcal{D}}(\mathbf{W}^{(t)}) \leq 8e^4 \mathcal{T} L_S(\mathbf{W}^{(t)}) + \exp(-n^2)$. *The test loss is bounded by the training loss (up to a constant factor) and an exponentially small term, demonstrating strong generalization ability.*

Theorem 4.2 delineates the sufficient condition for signal-dominated learning: when the SNR is sufficiently high ($\mathrm{SNR} \geq \Theta(\mathcal{T} n^{-1/4})$), the model's training dynamics are dominated by signal learning, with at least one filter per signal channel capturing meaningful patterns. To verify the sharpness of this SNR threshold—i.e., that the model's behavior undergoes a qualitative shift when the SNR falls below this bound—we next present a theorem characterizing the noise memorization regime, which describes the opposite learning behavior.

**Theorem 4.3.** *For any* $\epsilon > 0$, *let* $T = \widetilde{O}(mn\eta^{-1}\sigma_p^{-2} d^{-1} + \eta^{-1}\mathcal{T} m \mu^{-2} \epsilon^{-1})$. *Under Condition 4.1, if* $\mathrm{SNR} \leq \Theta(n^{-1})$, *then with probability at least* $1 - \delta$, *there exists* $0 \leq t \leq T$ *such that:*

1. *The CNN memorizes the noise:* $\max_{r,i} \rho_{j,r,i}^{(t)} = \Omega(1)$ *for* $j \in [\mathcal{T}]$. *Here, the noise-aligned coefficients grow to a constant order, indicating that the model prioritizes fitting spurious noise in the training data.*

2. *The CNN fails to learn the signal:* $\max_{j,r,\tau} |\gamma_{j,r,\tau}| = \widetilde{O}(\mathcal{T}^{-1} n \mathrm{SNR}^2)$. *The signal-aligned coefficients remain negligible, confirming that the model cannot capture meaningful signal patterns.*

3. *The training loss converges to* $\epsilon$, *i.e.,* $L_S(\mathbf{W}^{(t)}) \leq \epsilon$. *Despite memorizing noise, the over-parameterized model still achieves full fitting of the training data.*

4. *The trained CNN has a constant order test loss:* $L_{\mathcal{D}}(\mathbf{W}^{(t)}) \geq \Theta(1)$. *The model's generalization performance degrades severely, as noise memorization fails to generalize to unseen test data.*

Theorem 4.3 complements Theorem 4.2 by defining the noise-dominated regime ($\mathrm{SNR} \leq \Theta(n^{-1})$), which is fundamentally distinct from the signal learning regime. In this low-SNR scenario, gradient descent drives the model to memorize training noise rather than learn signals, leading to poor generalization—even though the training loss is minimized. This contrast highlights the critical role of SNR in shaping the model's learning behavior.

Given the poor generalization in the low-SNR noise memorization regime, we next explore whether TopK pruning can reverse this outcome by suppressing noise components and guiding signal learning. The following theorem establishes that, with appropriate selection of the pruning threshold $K$, TopK pruning can effectively mitigate noise memorization and restore strong generalization performance even when $\mathrm{SNR} \leq \Theta(n^{-1})$.

**Theorem 4.4.** *For any* $\epsilon > 0$, *let* $T = \widetilde{O}(m\mathcal{T}\eta^{-1}\mu^{-2} + \eta^{-1}\mathcal{T} mn \sigma_p^{-2} K^{-1} \epsilon^{-1})$. *Under Condition 4.1, if* $\mathrm{SNR} \leq \Theta(n^{-1})$ *and if we adopt the TopK pruning strategy, K satisfies* $\Theta(d)\exp\Big(-\frac{\mu}{2\sigma_p^2}\Big) + O(\sqrt{d\log(1/\delta)}) \leq K \leq \Theta(\mu^2 \sqrt{n} \mathcal{T}^{-1} \sigma_p^{-2})$ *then with probability at least* $1 - \delta$, *there exists* $0 \leq t \leq T$ *such that:*

1. *The training loss converges to* $\epsilon$, *i.e.,* $L_S(\mathbf{W}^{(t)}) \leq \epsilon$. *Pruning does not hinder the model's ability to fit the training data, as the remaining weight components still enable sufficient expressive capacity.*

2. *The trained CNN achieves a small test loss:* $L_{\mathcal{D}}(\mathbf{W}^{(t)}) \leq O(\mathcal{T}\epsilon) + \exp(-n^2)$. *By retaining signal-dominant coordinates and pruning noise components, the model avoids noise memorization and achieves favorable generalization, even in the low-*SNR *regime.*

The above theorem demonstrates the positive regulatory effect of TopK pruning when $K$ is within a suitable range: it balances the model's expressive capacity (to fit training data) and noise suppression (to generalize). To further clarify the sharpness of the pruning threshold $K$, we present the following theorem, which shows that excessive pruning (i.e., $K$ is too small) reverses this benefit and leads to poor generalization again—even with TopK pruning applied.

**Theorem 4.5.** *For any $\epsilon > 0$, let $T = \widetilde{O}(m\mathcal{T}\eta^{-1}\mu^{-2} + \eta^{-1}\mathcal{T}mn\sigma_p^{-2}K^{-1}\epsilon^{-1})$. Under Condition 4.1, if we adopt the TopK pruning strategy, K satisfies $K \leq d\left(1 - \sqrt{1 - \exp\left(-\frac{2\mu^2}{\pi\sigma_p^2}\right)}\right) - O(\sqrt{d\log(1/\delta)})$ then with probability at least $1-\delta$, there exists $0 \leq t \leq T$ such that:*

1. *The training loss converges to $\epsilon$, i.e., $L_S(\mathbf{W}^{(t)}) \leq \epsilon$. Even with excessive pruning, the over-parameterized model can still fit the training data by adjusting the remaining few weight components.*
2. *The trained CNN has a constant order test loss: $L_{\mathcal{D}}(\mathbf{W}^{(t)}) \geq \Omega(\log(\mathcal{T}))$. Excessive pruning removes not only noise components but also critical signal coordinates, leaving the model unable to capture meaningful patterns and resulting in degraded generalization performance (with a test loss lower bounded by $\Omega(\log(\mathcal{T}))$).*

Collectively, these four theorems form a complete characterization of CNN learning behavior under gradient descent and TopK pruning: Theorem 4.2 and 4.3 delineate the SNR-dependent trade-off between signal learning and noise memorization without pruning; the subsequent two theorems establish the pruning threshold's dual role—appropriate pruning mitigates noise memorization in low-SNR scenarios, while excessive pruning undermines signal learning—providing a rigorous theoretical foundation for the practical application of TopK pruning in $\mathrm{ReLU}$ CNNs.

## 5. Overview of Proof Technique

This section is devoted to clarifying the analytical difficulties inherent in our setting and to motivating the techniques employed in the subsequent proofs. Together, these elements enable a precise characterization of the training dynamics and lead to our main theoretical result. Complete proofs are provided in the appendix.

### 5.1. Key Technique 1: Signal-Noise Analysis

The core objective of the analysis in this paper is to investigate how model weights allocate their incremental updates between discriminative signal directions and spurious noise directions during the training process of the gradient descent algorithm. Cao et al. (2022) proposed a signal-noise decomposition method, which we briefly recapitulate below.

**Definition 5.1.** Let $\mathbf{w}_{j,r}^{(t)}$ for $j \in [\mathcal{T}]$, $r \in [m]$ denote the convolution filters at iteration $t$ of gradient descent. There exist unique coefficients $\gamma_{j,r,\tau}^{(t)}$ and $\rho_{j,r,i}^{(t)}$ such that $\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + \sum_{\tau=1}^{\mathcal{T}} \gamma_{j,r,\tau}^{(t)} \mu^{-2}\boldsymbol{\mu}_\tau + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \|\boldsymbol{\xi}_i\|_2^{-2}\boldsymbol{\xi}_i$.

Define $\bar{\rho}_{j,r,i}^{(t)} = \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \geq 0)$ and $\underline{\rho}_{j,r,i}^{(t)} = \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \leq 0)$. Then the above decomposition can be equivalently written as

$$\begin{aligned}\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} &+ \sum_{\tau=1}^{\mathcal{T}} \gamma_{j,r,\tau}^{(t)} \mu^{-2}\boldsymbol{\mu}_\tau \\ &+ \sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(t)} \|\boldsymbol{\xi}_i\|_2^{-2}\boldsymbol{\xi}_i + \sum_{i=1}^{n} \underline{\rho}_{j,r,i}^{(t)} \|\boldsymbol{\xi}_i\|_2^{-2}\boldsymbol{\xi}_i. \end{aligned} \tag{1}$$

The normalization ensures that $\gamma_{j,r,\tau}^{(t)}$ and $\rho_{j,r,i}^{(t)}$ closely track the projections of $\mathbf{w}_{j,r}^{(t)}$ onto the signal and noise directions, respectively. This decomposition reduces the analysis of CNN training to controlling the evolution of these coefficients over time, without relying on smoothness assumptions on the activation function.

However, the analysis in Cao et al. (2022) focuses on the leading coefficients and crucially exploits the heterogeneous update speeds induced by the $\mathrm{ReLU}^q$ activation with $q > 2$. Such an argument no longer applies to standard $\mathrm{ReLU}$ networks, where all active neurons share the same derivative $\sigma'(x) = 1$.

To address this issue, we adopt the *time-invariant coefficient ratio* analysis introduced by Kou et al. (2023), which extends the signal–noise framework to the $\mathrm{ReLU}$ setting. This technique shows that, up to a stopping time $T^*$, the

relative magnitudes of signal and noise coefficients remain stable throughout training. The key result is summarized below.

**Proposition 5.2.** *Under Condition 4.1, for all* $t \in [0, T^*]$,

$$\frac{\gamma_{\tau,r_1,\tau}^{(t)}}{\bar{\rho}_{y_i,r_2,i}^{(t)}} = \Theta(\mathcal{T}^{-1} n \mathrm{SNR}^2),$$

*where* $r_1 \in \{r : \langle \mathbf{w}_{\tau,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle > 0\}$, $r_2 \in \{r : \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle > 0\}$, $\tau \in \mathcal{T}$, *and* $i \in [n]$.

The time-invariance of this ratio allows us to sharply characterize whether the training dynamics amplify signal directions or become dominated by noise, depending on the relative strength of $\mu$ and $\sigma_p\sqrt{d}$. This observation plays a crucial role in our analysis of $\mathrm{ReLU}$ networks trained with pruning.

### 5.2. Key Technique 2: Weight Dynamic Pruning Analysis

This section presents a theoretical analysis of TopK weight pruning performed synchronously during training, with the goal of revealing how pruning regulates the model's signal learning and noise memorization behaviors through probability bounds and virtual scenario analysis. Our analysis begins with two classical Gaussian tail inequalities, which provide the foundation for characterizing the statistical properties of weight coordinates.

Let $x \sim \mathcal{N}(0, \sigma_p^2)$. For any $c > 0$, the following inequalities hold: $\mathbb{P}(|x| \geq c) \leq 2\exp\left(-\frac{c^2}{2\sigma_p^2}\right)$, $\mathbb{P}(|x| \leq c) \leq \sqrt{1 - \exp\left(-\frac{2c^2}{\pi\sigma_p^2}\right)}$. Applying these bounds to the randomly initialized convolutional weight vector $\mathbf{w}_{j,r}^{(1)}$, we can, with probability at least $1 - \delta$, derive upper and lower bounds on the number of coordinates whose magnitudes exceed a signal threshold $\mu$. The deviation is controlled by a term of order $O(\sqrt{d\log(1/\delta)})$, where $d$ denotes the dimensionality of the weight vector.

Based on this statistical characterization, we can precisely analyze the filtering effect of TopK pruning on signal and noise components. The pruning threshold $K$ determines whether signal coordinates are preserved, leading to two extreme regimes: mild pruning and over pruning. When $K$ is no smaller than the upper bound on the number of signal coordinates, all signal-related weights are retained. Since signal coordinates typically have larger magnitudes than noise coordinates, they are prioritized during TopK selection, ensuring stable learning of the true signal. When $K$ falls below the lower bound on the number of signal coordinates, signal components are entirely pruned away. In this regime, only a small number of low-magnitude noise coordinates remain, preventing the model from capturing meaningful signal information.

To further illustrate the regulatory role of pruning in training dynamics, we consider an unfavorable virtual scenario in which the retained noise coordinates continue to grow during training. Even in this setting—where noise memorization is most likely to occur—TopK pruning maintains a sufficiently high $\mathrm{SNR}$ by filtering weight components, thereby enabling effective signal learning.

This conclusion is further supported by the previously established time-invariant coefficient ratio analysis: the relative magnitudes of signal and noise coefficients remain stable throughout training. Signal-dominant components preserved by pruning continue to grow and are not overwhelmed by unselected noise components. Together, this time-invariant property and TopK pruning suppress noise memorization and guide the model toward signal learning.

In contrast, when pruning is excessively aggressive, signal coordinates are irreversibly removed, depriving the model of the capacity to represent core signal structures. Even with continued training, the model can only fit noise, leading to a severe degradation in generalization performance.

## 6. Conclusion and Future Work

We analyzed dynamic TopK pruning in $\mathrm{ReLU}$ CNNs, showing that appropriate sparsity improves generalization by suppressing noise, whereas excessive pruning hinders signal learning. This finding provides a theoretical basis for balancing sparsity and model performance in practical pruning strategies. We also verified the consistency of the theoretical conclusions through experiments on benchmark datasets. One limitation is the reliance on two-layer CNNs and simplified data models. The future direction is to extend our analysis to complex architectures such as Transformers.

## References


Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019.

Benbaki, R., Chen, W., Meng, X., Hazimeh, H., Ponomareva, N., Zhao, Z., and Mazumder, R. Fast as chita: Neural network pruning with combinatorial optimization. In *International Conference on Machine Learning*, pp. 2031–2049. PMLR, 2023.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Cao, Y., Chen, Z., Belkin, M., and Gu, Q. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.

Ceron, J. S. O., Courville, A., and Castro, P. S. In value-based deep reinforcement learning, a pruned network is a good network. In *Forty-first International Conference on Machine Learning*, 2024.

Chen, B., Lyu, X., Gao, L., Song, J., and Shen, H. T. Safeptr: Token-level jailbreak defense in multimodal llms via prune-then-restore mechanism. *Advances in Neural Information Processing Systems*, 2025.

Chen, T., Ji, B., Ding, T., Fang, B., Wang, G., Zhu, Z., Liang, L., Shi, Y., Yi, S., and Tu, X. Only train once: A one-shot neural network training and pruning framework. *Advances in Neural Information Processing Systems*, 34: 19637–19651, 2021.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

da Cunha, A., Natale, E., and Viennot, L. Proving the strong lottery ticket hypothesis for convolutional neural networks. In *ICLR 2022-10th International Conference on Learning Representations*, 2022.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

Galanti, T., Xu, M., Galanti, L., and Poggio, T. Norm-based generalization bounds for sparse neural networks. *Advances in Neural Information Processing Systems*, 36:42482–42501, 2023.

Guo, S., Wang, Y., Li, Q., and Yan, J. Dmcp: Differentiable markov channel pruning for neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1539–1547, 2020.

Guyard, T., Herzet, C., Elvira, C., and Arslan, A.-N. A new branch-and-bound pruning framework for $\ell_0$-regularized problems. In *International Conference on Machine Learning*, volume 235, pp. 48077–48096, 2024.

Huang, Z. and Wang, N. Data-driven sparse structure selection for deep neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 304–320, 2018.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jin, T., Carbin, M., Roy, D., Frankle, J., and Dziugaite, G. K. Pruning's effect on generalization through the lens of training and regularization. *Advances in Neural Information Processing Systems*, 35:37947–37961, 2022.

Kim, J., El Halabi, M., Ji, M., and Song, H. O. Layermerge: Neural network depth compression through layer pruning and merging. *Proceedings of Machine Learning Research*, 235:23825–23842, 2024.

Kou, Y., Chen, Z., Chen, Y., and Gu, Q. Benign overfitting in two-layer relu convolutional neural networks. In *International conference on machine learning*, pp. 17615–17659. PMLR, 2023.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Le, Q., Diao, E., Wang, Z., Wang, X., Ding, J., Yang, L., and Anwar, A. Probe pruning: Accelerating llms through dynamic pruning via model-probing. In *The Thirteenth International Conference on Learning Representations*, 2025.

LeCun, Y., Denker, J., and Solla, S. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

Lee, D., Lee, K., Chung, J., and Lee, N. Safe: Finding sparse and flat minima to improve pruning. In *Forty-second International Conference on Machine Learning*, 2025.

Lee, N., Ajanthan, T., and Torr, P. Snip: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*, 2018.

Lin, M., Chen, Q., and Yan, S. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

Liu, S. Learning sparse neural networks for better generalization. In *29th International Joint Conference on Artificial Intelligence-17th Pacific Rim International Conference on Artificial Intelligence, IJCAI 2020, PRICAI 2020*, pp. 5190–5191. International Joint Conferences on Artificial Intelligence (IJCAI), 2020.

Liu, S., Chen, T., Chen, X., Atashgahi, Z., Yin, L., Kou, H., Shen, L., Pechenizkiy, M., Wang, Z., and Mocanu, D. C. Sparse training via boosting pruning plasticity with neuroregeneration. *Advances in Neural Information Processing Systems*, 34:9908–9922, 2021.

Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., and Zhang, C. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE international conference on computer vision*, pp. 2736–2744, 2017.

Malach, E., Yehudai, G., Shalev-Schwartz, S., and Shamir, O. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, pp. 6682–6691. PMLR, 2020.

Mozaffari, M., Yazdanbakhsh, A., Zhang, Z., and Dehnavi, M. M. Slope: Double-pruned sparse plus lazy low-rank adapter pretraining of llms. In *The Thirteenth International Conference on Learning Representations*, 2025.

Muthukumar, R. and Sulam, J. Sparsity-aware generalization theory for deep neural networks. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 5311–5342. PMLR, 2023.

Ning, X., Zhao, T., Li, W., Lei, P., Wang, Y., and Yang, H. Dsa: More efficient budgeted pruning via differentiable sparsity allocation. In *European Conference on Computer Vision*, pp. 592–607. Springer, 2020.

Shen, F., Li, C., Geng, Y., Deng, Y., and Chen, H. Prune and repaint: Content-aware image retargeting for any ratio. *Advances in Neural Information Processing Systems*, 37:75177–75198, 2024.

Tanaka, H., Kunin, D., Yamins, D. L., and Ganguli, S. Pruning neural networks without any data by iteratively conserving synaptic flow. *Advances in neural information processing systems*, 33:6377–6389, 2020.

Tartaglione, E., Lepsøy, S., Fiandrotti, A., and Francini, G. Learning sparse neural networks via sensitivity-driven regularization. *Advances in neural information processing systems*, 31, 2018.

Wang, C., Zhang, G., and Grosse, R. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations*, 2019.

Xia, M., Gao, T., Zeng, Z., and Chen, D. Sheared llama: Accelerating language model pre-training via structured pruning. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

Yang, H., Liang, Y., Guo, X., Wu, L., and Wang, Z. Random pruning over-parameterized neural networks can improve generalization: A training dynamics analysis. *Journal of Machine Learning Research*, 26(84):1–51, 2025.

Yin, R., Li, Y., Lee, D., and Panda, P. Duogpt: Training-free dual sparsity through activation-aware pruning in llms. *Advances in neural information processing systems*, 2025.

Zhu, Q., Duan, J., Chen, C., Liu, S., Li, X., Feng, G., Lv, X., Chuanfu, X., Lin, D., and Yang, C. Sampleattention: Near-lossless acceleration of long context llm inference with adaptive structured sparse attention. *Proceedings of Machine Learning and Systems*, 7, 2025.

Zou, L., Yin, S., Pei, Z., Ho, T.-Y., Farnia, F., and Yu, B. Permllm: Learnable channel permutation for n: M sparse large language models. *Advances in Neural Information Processing Systems*, 2025.

## A. Experiments

In this section, we present simulations of synthetic data to back up our theoretical analysis in the previous section.

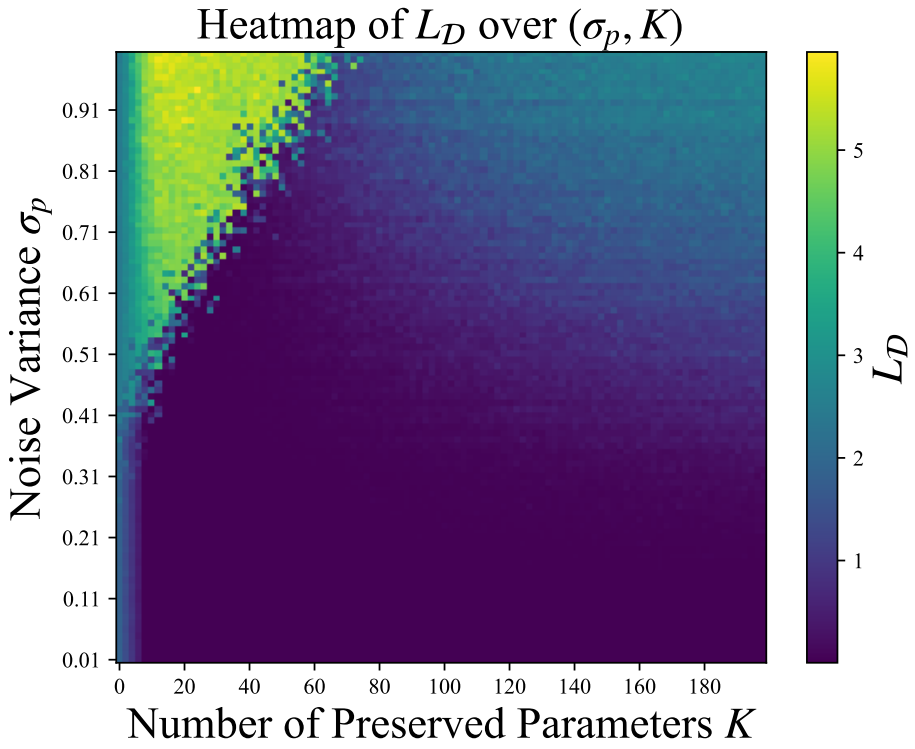


*Figure 1.* Heatmap of the test loss $L_D$ under different noise levels $\sigma_p$ and projection dimensions $K$.

### A.1. Synthetic data experiments.

Synthetic data in this study is generated in line with the specifications of Definition 3.1, with detailed experimental settings as follows: the training set consists of 200 samples, the convolutional neural network (CNN) is designed for $\mathcal{T} = 10$ classification categories, and the input data dimension is set to $d = 200$. To simplify the analysis without loss of generality, the norm of the signal vector is fixed to 1, i.e., $\mu_i = \mathbf{e}_i$ $(1 \leq i \leq \mathcal{T})$, where $\mathbf{e}_i$ denotes the standard basis vector. The noise vector $\boldsymbol{\xi}$ is sampled from a Gaussian distribution $\mathcal{N}(0, \sigma_p^2 I_{d-\mathcal{T}})$, with $I_{d-\mathcal{T}}$ being the $(d-\mathcal{T})$-dimensional identity matrix.

The model architecture adopts the two-layer CNN proposed in Section 3, with $\mathrm{ReLU}$ as the activation function and the number of filters set to $m = 20$. Network parameters are initialized using a Gaussian distribution $\mathcal{N}(0, \sigma_0^2)$ with $\sigma_0 = 0.01$; full-batch gradient descent is employed for training, with a learning rate of 0.05 and a total of 5000 training epochs to ensure model convergence.

Experiments focus on the interaction between pruning retention number $K$ (ranging from 2 to 200) and noise intensity $\sigma_p$ (varying from 0.01 to 1.00), covering all combinations for comparative analysis. The results indicate that under all $K$ and $\sigma_p$ configurations considered in this study, the proposed training scheme can constrain the model's training loss below 0.01. After training, 10,000 independent test samples are used to evaluate the test loss under each configuration, and the final experimental results are visualized as a heatmap with $K$ and $\sigma_p$ as axes (see Figure 1).

From Figure 2, we can directly draw a clear conclusion: with the increase of noise intensity $\sigma_p$, the valid range of $K$ for favorable model generalization gradually shrinks. This observation is consistent with our theoretical results: $\Theta(d)\exp\left(-\frac{\mu}{2\sigma_p^2}\right) + O\left(\sqrt{d\log(1/\delta)}\right) \leq K \leq \Theta\left(\mu^2\sqrt{n}\mathcal{T}^{-1}\sigma_p^{-2}\right)$. Correspondingly, the two boundary lines in Figure

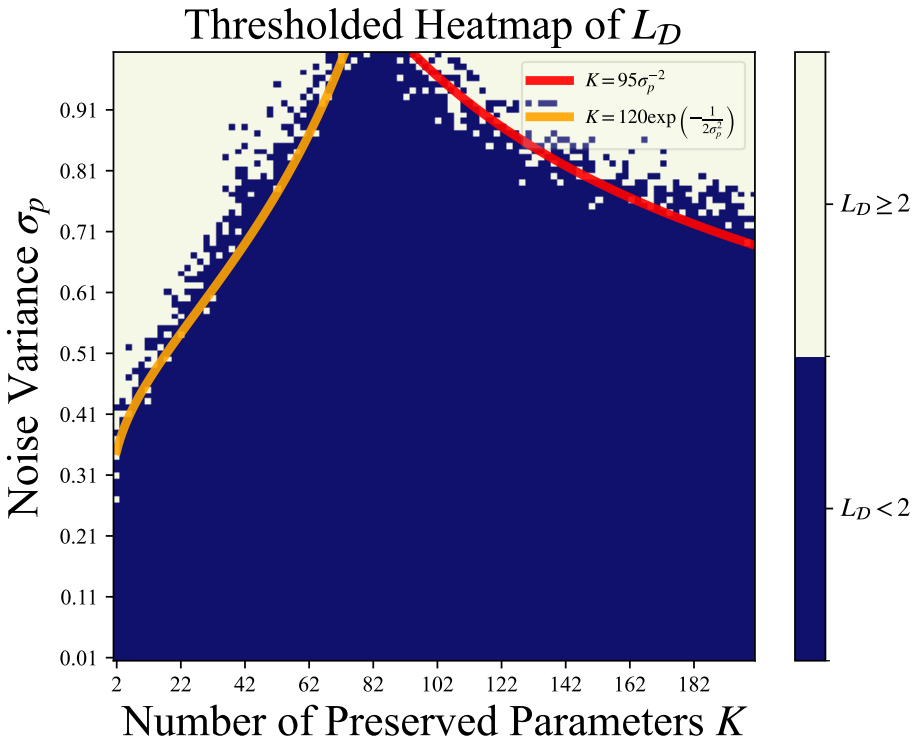


*Figure 2.* Thresholded heatmap of $L_D$, where values smaller than 2 are shown in blue and values greater than or equal to 2 are shown in light yellow. The expressions for the red and orange curves are $K = 95\sigma_p^{-2}$ and $K = 120\exp(-\frac{1}{2\sigma_p^2})$, respectively.

2 also match the theoretical expectations, as formulated by $120\exp\left(-\frac{1}{2\sigma_p^2}\right) \leq K \leq 95\sigma_p^{-2}$.

### A.2. Real data experiments.

We systematically evaluate Dynamic TopK pruning on the MNIST (Figure 3) and CIFAR-10 (Figure 4) datasets. To avoid trivial overfitting on relatively simple datasets, we introduce mild random perturbations during training.

As the retained weight ratio $K$ increases, the training accuracy on both datasets generally improves, indicating that higher parameter density facilitates fitting the training data. However, improved training performance does not necessarily translate into better test performance. On the MNIST dataset, test accuracy reaches its best performance at an intermediate sparsity level and does not consistently improve as $K$ increases, suggesting that excessive model parameters may harm generalization. On the CIFAR-10 dataset, test accuracy also peaks at moderate values of $K$ and degrades as the model approaches the dense setting, reflecting a trade-off between representational capacity and generalization.

Overall, these results indicate that an appropriate level of parameter sparsity can lead to improved generalization performance.

## B. Additional Experiments

In this section, we present comprehensive empirical results to further validate our theoretical findings. We conduct experiments on two typical architectures: a CNN on CIFAR-100 and a Vision Transformer on MNIST, under various noise levels and sparsity ratios. The key observation is consistent with our theory: *moderate sparsity improves generalization by suppressing noise fitting, while excessive sparsity degrades performance under low SNR.*

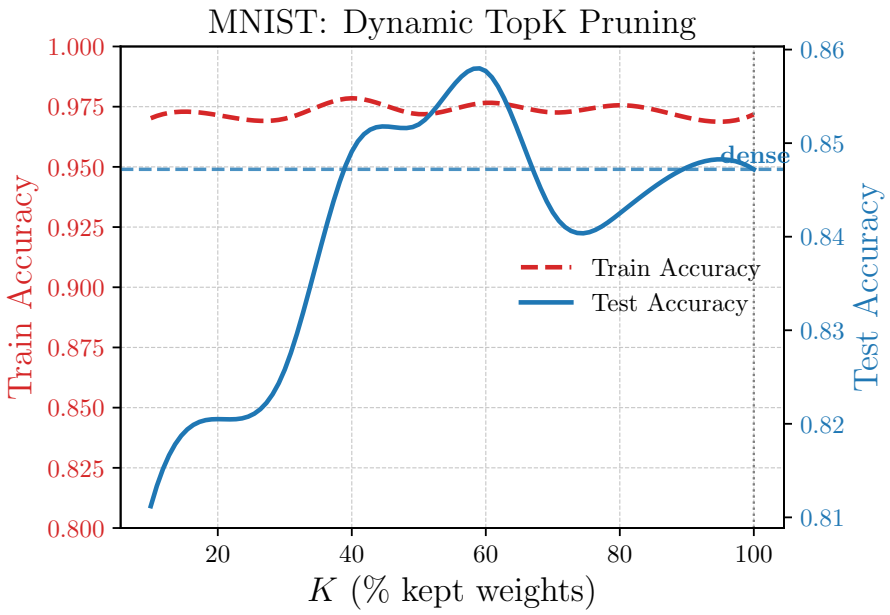


*Figure 3.* Test accuracy and training loss of a CNN on MNIST under dynamic TopK pruning with varying sparsity levels.

### B.1. CNN on CIFAR-100

We evaluate a convolutional neural network on the CIFAR-100 dataset with injected Gaussian noise of varying standard deviations. We test different sparsity regimes by keeping 80%, 50%, 30%, and 10% of weights, respectively. As shown in Table 1, under moderate noise, sparse models consistently outperform the dense model, validating that sparsity acts as a regularizer against label noise. Under extreme noise (i.e., large noise std), aggressive pruning becomes harmful, which aligns with our theoretical bound that excessive sparsity loses necessary signal capacity when SNR is extremely low.

*Table 1.* CNN Performance on CIFAR-100 under Different Noise Levels and Sparsity Ratios

| Noise Std | Dense Acc | Keep 80% Acc | Diff | Keep 50% Acc | Diff | Keep 30% Acc | Diff | Keep 10% Acc | Diff |
|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 66.84% | 67.62% | +0.78 | 67.64% | +0.80 | 67.38% | +0.54 | 66.73% | -0.11 |
| 0.10 | 63.96% | 64.81% | +0.85 | 64.56% | +0.60 | 64.49% | +0.53 | 63.97% | +0.01 |
| 0.20 | 57.38% | 58.47% | +1.09 | 58.23% | +0.85 | 57.96% | +0.58 | 56.62% | -0.76 |
| 0.30 | 51.93% | 52.47% | +0.54 | 52.74% | +0.81 | 52.64% | +0.71 | 51.48% | -0.45 |
| 0.40 | 47.41% | 48.89% | +1.48 | 48.53% | +1.12 | 48.30% | +0.89 | 45.82% | -1.59 |
| 0.50 | 44.16% | 45.33% | +1.17 | 45.19% | +1.03 | 44.36% | +0.20 | 43.32% | -0.84 |
| 5.00 | 14.98% | 15.14% | +0.16 | 15.21% | +0.23 | 15.38% | +0.40 | 16.05% | +1.07 |
| 6.00 | 13.43% | 14.00% | +0.57 | 13.88% | +0.45 | 13.22% | -0.21 | 14.24% | +0.81 |
| 7.00 | 11.83% | 11.45% | -0.38 | 11.88% | +0.05 | 11.18% | -0.65 | 11.59% | -0.24 |
| 8.00 | 10.91% | 11.07% | +0.16 | 10.85% | -0.06 | 11.45% | +0.54 | 11.47% | +0.56 |
| 9.00 | 10.44% | 10.41% | -0.03 | 10.00% | -0.44 | 10.35% | -0.09 | 10.20% | -0.24 |
| 10.00 | 9.43% | 9.23% | -0.20 | 9.41% | -0.02 | 9.25% | -0.18 | 9.02% | -0.41 |

### B.2. Vision Transformer on MNIST

To verify that our conclusion is architecture-agnostic, we further conduct experiments on a Vision Transformer (ViT) on MNIST. As shown in Table 2, the trend is highly consistent: light pruning improves accuracy in low-to-moderate noise, while aggressive pruning (e.g., keep 10%) leads to significant performance degradation, especially under strong noise.

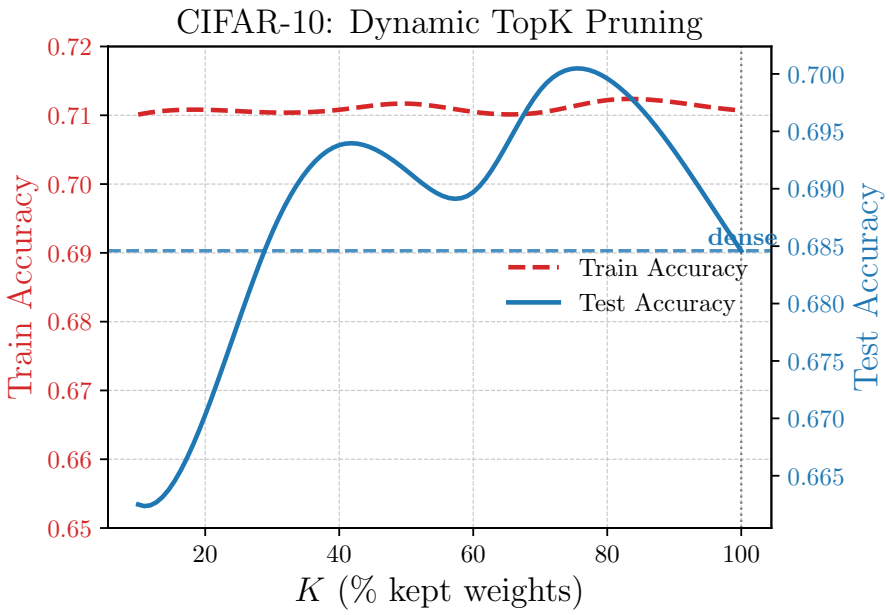


*Figure 4.* Test accuracy and training loss of a CNN on CIFAR-10 under dynamic TopK pruning with varying sparsity levels.

This confirms that the interplay between sparsity, noise, and model capacity is universal across CNNs and Transformers.

*Table 2.* Transformer Performance on MNIST under Different Noise Levels and Sparsity Ratios

| Noise Std | Dense Acc | Keep 80% Acc | Diff | Keep 50% Acc | Diff | Keep 30% Acc | Diff | Keep 10% Acc | Diff |
|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 98.16% | 98.19% | +0.03 | 98.21% | +0.05 | 98.15% | -0.01 | 97.19% | -0.97 |
| 0.10 | 98.17% | 98.21% | +0.04 | 98.14% | -0.03 | 98.19% | +0.02 | 96.97% | -1.20 |
| 0.20 | 98.46% | 98.49% | +0.03 | 98.44% | -0.02 | 98.35% | -0.11 | 97.55% | -0.91 |
| 0.30 | 98.56% | 98.57% | +0.01 | 98.57% | +0.01 | 98.40% | -0.16 | 97.39% | -1.17 |
| 0.40 | 98.52% | 98.60% | +0.08 | 98.55% | +0.03 | 98.42% | -0.10 | 97.37% | -1.15 |
| 0.50 | 98.43% | 98.46% | +0.03 | 98.40% | -0.03 | 98.40% | -0.03 | 96.85% | -1.58 |
| 5.00 | 93.08% | 93.32% | +0.24 | 93.26% | +0.18 | 92.23% | -0.85 | 88.28% | -4.80 |
| 6.00 | 90.49% | 90.82% | +0.33 | 90.53% | +0.04 | 89.47% | -1.02 | 83.96% | -6.53 |
| 7.00 | 88.74% | 88.70% | -0.04 | 88.63% | -0.11 | 87.24% | -1.50 | 82.21% | -6.53 |
| 8.00 | 85.58% | 85.57% | -0.01 | 85.55% | -0.03 | 84.72% | -0.86 | 77.76% | -7.82 |
| 9.00 | 84.65% | 84.52% | -0.13 | 84.10% | -0.55 | 82.50% | -2.15 | 76.85% | -7.80 |
| 10.00 | 82.32% | 82.06% | -0.26 | 81.94% | -0.38 | 80.70% | -1.62 | 76.91% | -5.41 |

## C. Preliminary Lemmas

In this section, we introduce several key lemmas that capture crucial properties of the data and the neural network parameters when randomly initialized.

**Lemma C.1.** *Suppose that $\delta > 0$ and $n \geq 13\mathcal{T}\log(2/\delta)$, then with probability at least $1-\delta$, we have that*

$$|\{i \in [n] : y_i = \tau\}| \in [\frac{4n}{5\mathcal{T}}, \frac{6n}{5\mathcal{T}}],$$

*for all $\tau \in [\mathcal{T}]$.*

*Proof of Lemma C.1.* By Hoeffding's inequality, the following holds with probability no less than $1-\delta$:

$$|\sum_{i=1}^{n}\mathbb{1}(y_i=\tau)-\frac{n}{\mathcal{T}}|\leq\sqrt{\frac{n}{2}\log(\frac{2}{\delta})}.$$

Then by $n\geq 12\log(2/\delta)$, we can get that

$$|\sum_{i=1}^{n}\mathbb{1}(y_i=\tau)-\frac{n}{\mathcal{T}}|\leq\sqrt{\frac{n}{2}\log(\frac{2}{\delta})}\leq\frac{n}{5\mathcal{T}},$$

which indicates that $|\{i\in[n]:y_i=\tau\}|\in[\frac{4n}{5\mathcal{T}},\frac{6n}{5\mathcal{T}}]$. Here the proof is completed. □

Denote $M_{j,\tau}$ as $\{r\in[m]:\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_\tau\rangle>0\}$. Then we have the following lemma:

**Lemma C.2.** *Suppose that $\delta>0$ and $n\geq 13\mathcal{T}\log(2/\delta)$, then with probability at least $1-\delta$, we have that*

$$|M_{j,\tau}|\in[\frac{1}{4}m,\frac{3}{4}m].$$

*Proof of Lemma C.2.* The method is similar with Lemma C.1. □

Denote $N_{j,i}$ as $\{r\in[m]:\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\xi}_i\rangle>0\}$. Then we have the following lemma:

**Lemma C.3.** *Suppose that $\delta>0$ and $n\geq 13\mathcal{T}\log(2/\delta)$, then with probability at least $1-\delta$, we have that*

$$|N_{j,i}|\in[\frac{1}{4}m,\frac{3}{4}m].$$

*Proof of Lemma C.3.* The method is similar with Lemma C.1. □

**Lemma C.4** (Concentration Inequalities)**.** *Suppose that $\delta>0$ and with probability at least $1-\delta$,*

$$\sigma_p^2(d-\mathcal{T})/2\leq\|\boldsymbol{\xi}_i\|_2^2\leq 3\sigma_p^2(d-\mathcal{T})/2,$$
$$|\langle\boldsymbol{\xi}_i,\boldsymbol{\xi}_{i'}\rangle|\leq 2\sigma_p^2\sqrt{(d-\mathcal{T})\log(4n^2/\delta)},i\neq i',$$

*for all $i,i'\in[n]$.*

*Proof of Lemma C.4.* By Bernstein's inequality, with probability at least $1-\delta/(2n)$, the following holds:

$$\left|\|\boldsymbol{\xi}_i\|_2^2-\sigma_p^2(d-\mathcal{T})\right|\leq O\left(\sigma_p^2\cdot\sqrt{(d-\mathcal{T})\log(4n/\delta)}\right).$$

Thus, imposing $d=\mathcal{T}+\Omega\left(\log(4n/\delta)\right)$ ensures

$$\left|\|\boldsymbol{\xi}_i\|_2^2-\sigma_p^2(d-\mathcal{T})\right|\leq\frac{1}{2}\sigma_p^2(d-\mathcal{T}).$$

An analogous argument applies to $\langle\boldsymbol{\xi}_i,\boldsymbol{\xi}_{i'}\rangle$. Specifically, by Bernstein's inequality, with probability at least $1-\delta/(2n)$,

$$|\langle\boldsymbol{\xi}_i,\boldsymbol{\xi}_{i'}\rangle|\leq 2\sigma_p^2\sqrt{(d-\mathcal{T})\log(4n^2/\delta)}$$

for all $i\neq i'\in[n]$.

Combining these results via the union bound completes the proof.

□

**Lemma C.5.** *Suppose $d \geq \mathcal{T} + \Omega(\log(mn/\delta))$ and $m = \Omega(\log(1/\delta))$. Then, with probability at least $1-\delta$, the following hold:*

$$\left|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle\right| \leq \sigma_0 \mu \cdot \sqrt{2\log\left(8m\mathcal{T}^2/\delta\right)}, \tag{2}$$

$$\left|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle\right| \leq 2\sigma_0\sigma_p\sqrt{d-\mathcal{T}} \cdot \sqrt{\log\left(8mn\mathcal{T}/\delta\right)} \tag{3}$$

*for all $r \in [m]$, $j, \tau \in [\mathcal{T}]$, and $i \in [n]$. Further,*

$$\frac{\sigma_0\mu}{2} \leq \max_{r\in[m]} \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle \leq \sigma_0 \mu \cdot \sqrt{2\log\left(8m\mathcal{T}^2/\delta\right)}, \tag{4}$$

$$\frac{\sigma_0\sigma_p\sqrt{d-\mathcal{T}}}{4} \leq \max_{r\in[m]} \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle \leq 2\sigma_0\sigma_p\sqrt{d-\mathcal{T}} \cdot \sqrt{\log\left(8mn\mathcal{T}/\delta\right)} \tag{5}$$

*for all $j, \tau \in [\mathcal{T}]$ and $i \in [n]$.*

*Proof of Lemma C.5.* For each $r \in [m]$, the inner product $\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle$ is a mean-zero Gaussian random variable with variance $\sigma_0^2\mu^2$. By the Gaussian tail bound, for each tuple $(j, \tau, r)$,

$$\mathbb{P}\left(\left|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle\right| > \sigma_0\mu \cdot \sqrt{2\log\left(8m\mathcal{T}^2/\delta\right)}\right) \leq \frac{\delta}{4m\mathcal{T}^2}.$$

Applying the union bound over all $j, \tau \in [\mathcal{T}]$ and $r \in [m]$, we obtain that (2) holds with probability at least $1 - \delta/4$.

Next, note that $\mathbb{P}\left(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle < \sigma_0\mu/2\right) = c$ for some constant $c \in (0,1)$ (independent of $j, \tau, r$). By the assumption $m = \Omega(\log(1/\delta))$, there exists a constant $C > 0$ such that $m \geq C\log(4\mathcal{T}^2/\delta)$, which implies

$$\mathbb{P}\left(\max_{r\in[m]} \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle < \sigma_0\mu/2\right) = \left[\mathbb{P}\left(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle < \sigma_0\mu/2\right)\right]^m \leq c^m \leq \frac{\delta}{4\mathcal{T}^2}.$$

A union bound over all $j, \tau \in [\mathcal{T}]$ shows that the lower bound in (4) holds with probability at least $1 - \delta/4$; the upper bound follows directly from (2).

For the inner products involving $\boldsymbol{\xi}_i$, Lemma C.4 guarantees that with probability at least $1 - \delta/4$,

$$\frac{\sigma_p\sqrt{d-\mathcal{T}}}{\sqrt{2}} \leq \|\boldsymbol{\xi}_i\|_2 \leq \sqrt{\frac{3}{2}}\sigma_p\sqrt{d-\mathcal{T}}, \quad \forall i \in [n].$$

Since $\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle$ is a mean-zero Gaussian random variable with variance $\sigma_0^2\|\boldsymbol{\xi}_i\|_2^2$, analogous arguments to those for $\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle$ establish (3) and (5), with the remaining probability budget $\delta/4$ absorbed via the union bound. Combining all cases completes the proof. □

## D. Gradient Calculation and Update Rule

We begin by analyzing the parameter update dynamics of the network. For each weight vector $\mathbf{w}_{j,r}^{(t)}$ at iteration $t$, the update rule follows gradient descent on the loss $L_S(\mathbf{W}^{(t)})$, where $S$ denotes the training set. Expanding the gradient step, we derive the explicit update formula:

$$\begin{aligned}
\mathbf{w}_{j,r}^{(t+1)} &= \mathbf{w}_{j,r}^{(t)} - \eta \nabla_{\mathbf{w}_{j,r}^{(t)}} L_S(\mathbf{W}^{(t)}) \\
&= \mathbf{w}_{j,r}^{(t)} + \frac{\eta}{n}\sum_{i=1}^{n}\left(-\ell_{j,i}^{\prime(t)}\right)\nabla_{\mathbf{w}_{j,r}^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) \\
&= \mathbf{w}_{j,r}^{(t)} + \frac{\eta}{nm}\sum_{i=1}^{n}\left(-\ell_{j,i}^{\prime(t)}\right)\sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right)\boldsymbol{\xi}_i \\
&\quad + \frac{\eta}{nm}\sum_{i=1}^{n}\left(-\ell_{j,i}^{\prime(t)}\right)\sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle\right)\boldsymbol{\mu}_{y_i},
\end{aligned} \tag{6}$$

where:

- $\ell_{j,i}^{'(t)}$ denotes the partial derivative of the loss with respect to the $j$-th logit for sample $i$ at iteration $t$, satisfying $-\ell_{j,i}^{'(t)} = \mathbb{I}(j = y_i) - \text{logit}_j(F, \mathbf{x}_i)$.
- $\sigma'$ is the derivative of the activation function $\sigma$.
- The final equality follows from differentiating $F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)$ with respect to $\mathbf{w}_{j,r}^{(t)}$, using the chain rule.

To simplify the analysis of (6), we introduce two sets of auxiliary coefficients to decouple the contributions of the basis vectors $\boldsymbol{\mu}_\tau$ and noise terms $\boldsymbol{\xi}_i$.

**Definition D.1.** For all $j, \tau \in [\mathcal{T}]$, $r \in [m]$, and $i \in [n]$, define the coefficients $\gamma_{j,r,\tau}^{(t)}$ and $\rho_{j,r,i}^{(t)}$ via the recursive relations:

$$\gamma_{j,r,\tau}^{(0)} = 0, \quad \rho_{j,r,i}^{(0)} = 0, \tag{7a}$$

$$\gamma_{j,r,\tau}^{(t+1)} - \gamma_{j,r,\tau}^{(t)} = \frac{\eta}{nm} \sum_{y_i=\tau} \left(-\ell_{j,i}^{'(t)}\right) \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle\right) \mu^2, \tag{7b}$$

$$\rho_{j,r,i}^{(t+1)} - \rho_{j,r,i}^{(t)} = \frac{\eta}{nm} \left(-\ell_{j,i}^{'(t)}\right) \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right) \|\boldsymbol{\xi}_i\|_2^2. \tag{7c}$$

Here, (7b) captures the cumulative update contribution from samples labeled $\tau$ via the basis $\boldsymbol{\mu}_\tau$, while (7c) quantifies the contribution from the noise term $\boldsymbol{\xi}_i$ for sample $i$. The scaling factors $\mu^2$ and $\|\boldsymbol{\xi}_i\|_2^2$ ensure consistency with the inner product terms in (6).

Substituting (7b) and (7c) into (6) and unrolling the recursion from $t = 0$ to $t$, we obtain a closed-form expression for $\mathbf{w}_{j,r}^{(t)}$ in terms of the initial weight $\mathbf{w}_{j,r}^{(0)}$ and the auxiliary coefficients:

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + \sum_{\tau=1}^{\mathcal{T}} \gamma_{j,r,\tau}^{(t)} \cdot \mu^{-2} \boldsymbol{\mu}_\tau + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \boldsymbol{\xi}_i. \tag{8}$$

This decomposition explicitly separates the weight vector into its initial component and the cumulative updates from the basis vectors $\boldsymbol{\mu}_\tau$ and noise terms $\boldsymbol{\xi}_i$, facilitating subsequent analysis of their respective roles in training.

## E. Signal Learning

Let's calculate the two inner product expressions $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle$ and $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$:

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle &= \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle + \sum_{\tau'=1}^{\mathcal{T}} \gamma_{j,r,\tau'}^{(t)} \mu^{-2} \langle \boldsymbol{\mu}_{\tau'}, \boldsymbol{\mu}_\tau \rangle + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \|\boldsymbol{\xi}_i\|_2^{-2} \langle \boldsymbol{\xi}_i, \boldsymbol{\mu}_\tau \rangle \\
&= \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle + \gamma_{j,r,\tau}^{(t)},
\end{aligned} \tag{9}$$

where the last equality is due to $\boldsymbol{\mu}_\tau$ is orthogonal to $\boldsymbol{\mu}_{\tau'} (\tau' \neq \tau)$ and $\boldsymbol{\xi}_i (i \in [n])$. As for $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$, we can obtain

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle &= \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle + \sum_{\tau'=1}^{\mathcal{T}} \gamma_{j,r,\tau'}^{(t)} \mu^{-2} \langle \boldsymbol{\mu}_\tau', \boldsymbol{\xi}_i \rangle + \sum_{i'=1}^{n} \rho_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \langle \boldsymbol{\xi}_i', \boldsymbol{\xi}_i \rangle \\
&= \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle + \rho_{j,r,i}^{(t)} + \sum_{i' \neq i} \rho_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle,
\end{aligned} \tag{10}$$

where the last equality is due to $\boldsymbol{\mu}_\tau$ is orthogonal to $\boldsymbol{\xi}_i$ for all $i \in [n]$. By substituting equalitys 9 and 10 into the update formulas for $\gamma_{j,r,\tau}^{(t)}$ and $\rho_{j,r,i}^{(t)}$, we can obtain:

$$\gamma_{j,r,\tau}^{(t+1)} = \gamma_{j,r,\tau}^{(t)} + \frac{\eta}{nm}\sum_{y_i=\tau}(-\ell_{j,i}^{'(t)})\sigma'(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau\rangle + \gamma_{j,r,\tau}^{(t)})\mu^2, \tag{11}$$

$$\rho_{j,r,i}^{(t+1)} = \rho_{j,r,i}^{(t)} + \frac{\eta}{nm}(-\ell_{j,i}^{'(t)})\sigma'(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \rho_{j,r,i}^{(t)} + \sum_{i'\neq i}\rho_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2. \tag{12}$$

With these two formulas, we can better study the variation dynamics of $\gamma_{j,r,\tau}^{(t)}$ and $\rho_{j,r,i}^{(t)}$. We give the notation that $\bar{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)}\mathbb{1}(\rho_{j,r,i}^{(t)} \geq 0)$, $\underline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)}\mathbb{1}(\rho_{j,r,i}^{(t)} \leq 0)$.

We denote

$$\alpha := 4\log(T^*)$$
$$\beta := \max_{j,r,\tau,i}\{|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau\rangle|, |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle|\}$$
$$\mathrm{SNR} := \frac{\mu}{\sigma_p\sqrt{d}}$$

By Condition 4.1, we have that $\beta \leq 0.25$.

**Proposition E.1.** *For $0 \leq t \leq T^*$, we have that*

$$\gamma_{j,r,\tau}^{(0)}, \bar{\rho}_{j,r,i}^{(0)}, \underline{\rho}_{j,r,i}^{(0)} = 0,$$
$$0 \leq \gamma_{\tau,r,\tau}^{(t)} \leq \alpha, \tag{13}$$
$$0 \geq \gamma_{j,r,\tau}^{(t)} \geq -\alpha, j \neq \tau, \tag{14}$$
$$0 \geq \underline{\rho}_{j,r,i}^{(t)} \geq -\alpha, \tag{15}$$

*and there exists a positive constant $C'$ such that*

$$0 \leq \bar{\rho}_{j,r,i}^{(t)} \leq C'\widetilde{\rho}\alpha, \tag{16}$$

*for all $r \in [m]$, $j \in [\mathcal{T}]$ and $i \in [n]$, where $\widetilde{\rho} := \mathcal{T}n^{-1}\cdot \mathrm{SNR}^{-2}$.*

We will use induction to prove (13) and (16) in Proposition E.1.

Firstly we will introduce several technical lemmas that will be used for the proof of Proposition E.1.

**Lemma E.2.** *Under Condition 4.1, assume that (13), (14), (15), and (16) hold at iteration $t$. Then for all $j \in [\mathcal{T}]$, $r \in [m]$, and $i \in [n]$, the following bounds hold:*

$$\left|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle - \bar{\rho}_{j,r,i}^{(t)}\right| \leq 4n\alpha\sqrt{\frac{\log(4n^2/\delta)}{d-\mathcal{T}}} \quad \text{if } j = y_i,$$
$$\left|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle - \underline{\rho}_{j,r,i}^{(t)}\right| \leq 4n\alpha\sqrt{\frac{\log(4n^2/\delta)}{d-\mathcal{T}}} \quad \text{if } j \neq y_i.$$

*Proof of Lemma E.2.* From the update rule (10), we decompose the inner product difference as:

$$\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle - \rho_{j,r,i}^{(t)} = \sum_{i'\neq i}\rho_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle.$$

Taking absolute values and applying the triangle inequality for sums, we obtain:

$$\left|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle - \rho_{j,r,i}^{(t)}\right| = \left|\sum_{i'\neq i}\rho_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle\right| \leq \sum_{i'\neq i}\left|\rho_{j,r,i'}^{(t)}\right|\cdot\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot|\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle|.$$

The final bound

$$\sum_{i' \neq i} \left| \rho_{j,r,i'}^{(t)} \right| \cdot \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle| \leq 4n\alpha \sqrt{\frac{\log(4n^2/\delta)}{d - \mathcal{T}}}$$

follows directly from (15), (16), and Lemma C.4. This completes the proof. □

**Lemma E.3.** *Under Condition 4.1, assume that (13), (14), (15), and (16) hold at iteration $t$. Then for all $j \in [\mathcal{T}]$, $r \in [m]$, $i \in [n]$ with $j \neq y_i$, the following inequality holds:*

$$F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) \leq 1.$$

*Proof of Lemma E.3.* We begin by bounding the activated inner products. By the monotonicity of the activation function $\sigma$ and Lemma E.2, we have:

$$\begin{aligned}
\sigma\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle\right) &\leq \sigma\left(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{y_i} \rangle\right) \leq \left|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{y_i} \rangle\right|, \\
\sigma\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right) &\leq \sigma\left(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle + 4n\alpha \sqrt{\frac{\log(4n^2/\delta)}{d - \mathcal{T}}}\right) \\
&\leq 2 \max\left\{ \left|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle\right|, \, 4n\alpha \sqrt{\frac{\log(4n^2/\delta)}{d - \mathcal{T}}} \right\}.
\end{aligned}$$

Substituting these bounds into the definition of $F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)$, we obtain:

$$\begin{aligned}
F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) &= \frac{1}{m} \sum_{r=1}^{m} \left[ \sigma\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle\right) + \sigma\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right) \right] \\
&\leq 3 \max\left\{ \left|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{y_i} \rangle\right|, \, \left|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle\right|, \, 4n\alpha \sqrt{\frac{2\log(4n^2/\delta)}{d - \mathcal{T}}} \right\}.
\end{aligned}$$

The final inequality $F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) \leq 1$ follows directly from Condition 4.1, completing the proof. □

**Lemma E.4.** *Under Condition 4.1, assume that (13), (14), (15), and (16) hold at iteration $t$. Then for all $\tau \in [\mathcal{T}]$ and $r \in [m]$, the following inequalities hold:*

$$\begin{aligned}
&\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle \geq -0.25, \\
&\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle \leq \sigma\left(\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle\right) \leq \langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle + 0.25.
\end{aligned}$$

*Proof of Lemma E.4.* By the update rule (9) and the non-negativity of $\gamma_{\tau,r,\tau}^{(t)}$, we have:

$$\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle \geq \langle \mathbf{w}_{\tau,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle \geq -\beta \geq -0.25.$$

Next, the leftmost inequality $\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle \leq \sigma\left(\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle\right)$ is immediate. For the rightmost bound, we consider two cases: If $\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle \leq 0$, then $\sigma\left(\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle\right) = 0 \leq \langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle + 0.25$ (since $\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle \geq -0.25$). If $\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle > 0$, then $\sigma\left(\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle\right) = \langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle \leq \langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle + 0.25$.

This completes the proof. □

By an analogous argument, we obtain the following result for $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle$:

$$\begin{aligned}
&\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle \geq -0.5, \\
&\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle \leq \sigma\left(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right) \leq \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle + 0.5
\end{aligned}$$

for all $i \in [n]$ and $r \in [m]$.

To further quantify $\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle$ and $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle$ in terms of $\gamma_{\tau,r,\tau}^{(t)}$ and $\bar{\rho}_{y_i,r,i}^{(t)}$, we present the following lemma:

**Lemma E.5.** *Under Condition 4.1, assume that (13), (14), (15), and (16) hold at iteration $t$. Then for all $\tau \in [\mathcal{T}]$, $r \in [m]$, and $i \in [n]$,*

$$\left|\sigma\left(\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle\right) - \gamma_{\tau,r,\tau}^{(t)}\right| \leq 1,$$
$$\left|\sigma\left(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right) - \bar{\rho}_{y_i,r,i}^{(t)}\right| \leq 1.$$

*Proof of Lemma E.5.* By the update rules (9) and (10), we have:

$$\left|\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle - \gamma_{\tau,r,\tau}^{(t)}\right| \leq 0.25 \quad \text{and} \quad \left|\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle - \bar{\rho}_{y_i,r,i}^{(t)}\right| \leq 0.5.$$

From the analysis in Lemma E.4 and its analogous result for $\boldsymbol{\xi}_i$, it follows that:

$$\left|\sigma\left(\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle\right) - \langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle\right| \leq 0.25$$

and

$$\left|\sigma\left(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right) - \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right| \leq 0.5.$$

Applying the triangle inequality to these bounds yields:

$$\left|\sigma\left(\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle\right) - \gamma_{\tau,r,\tau}^{(t)}\right| \leq \left|\sigma\left(\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle\right) - \langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle\right| + \left|\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle - \gamma_{\tau,r,\tau}^{(t)}\right| \leq 0.25 + 0.25 \leq 1,$$
$$\left|\sigma\left(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right) - \bar{\rho}_{y_i,r,i}^{(t)}\right| \leq \left|\sigma\left(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right) - \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right| + \left|\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle - \bar{\rho}_{y_i,r,i}^{(t)}\right| \leq 0.5 + 0.5 = 1,$$

completing the proof. □

**Lemma E.6.** *For $\mathcal{T} - 1 \leq u, v \leq (\mathcal{T} - 1)e$ and $|F_1 - F_2| \leq C$, the following inequality holds:*

$$e^{-(1+C)} \leq \frac{u\left(v + \exp\{F_1\}\right)}{v\left(u + \exp\{F_2\}\right)} \leq e^{1+C}.$$

*Proof of Lemma E.6.* We start with the well-known inequality for fractions: for non-negative $A, B, C, D$,

$$\min\left\{\frac{A}{C}, \frac{B}{D}\right\} \leq \frac{A+B}{C+D} \leq \max\left\{\frac{A}{C}, \frac{B}{D}\right\}.$$

By the condition $\mathcal{T} - 1 \leq u, v \leq (\mathcal{T} - 1)e$, we derive $e^{-1} \leq \frac{v}{u} \leq e$, or equivalently $e^{-1} \leq \frac{u}{v} \leq e$.

Additionally, since $|F_1 - F_2| \leq C$, it follows that $e^{-C} \leq \frac{\exp\{F_1\}}{\exp\{F_2\}} \leq e^C$.

Applying the fraction inequality with $A = uv$, $B = u\exp\{F_1\}$, $C = vu$, $D = v\exp\{F_2\}$, we obtain:

$$\min\left\{\frac{uv}{vu}, \frac{u\exp\{F_1\}}{v\exp\{F_2\}}\right\} \leq \frac{uv + u\exp\{F_1\}}{vu + v\exp\{F_2\}} = \frac{u\left(v + \exp\{F_1\}\right)}{v\left(u + \exp\{F_2\}\right)} \leq \max\left\{\frac{uv}{vu}, \frac{u\exp\{F_1\}}{v\exp\{F_2\}}\right\}.$$

Substituting the bounds for $\frac{u}{v}$ and $\frac{\exp\{F_1\}}{\exp\{F_2\}}$, the left-hand side simplifies to $e^{-(1+C)}$. A similar argument for the upper bound yields $e^{1+C}$, completing the proof. □

**Lemma E.7.** *Under Condition 4.1, assume that (13), (14), (15), and (16) hold at any iteration $t' \leq t$. If*

$$\text{SNR} \geq \Theta(\mathcal{T} n^{-1/4}) \tag{17}$$

*holds, then the following conditions hold for any iteration $t' \leq t$:*

1. $|\sum_{r=1}^{m}[\gamma_{\tau,r,\tau}^{(t')}-\gamma_{\kappa,r,\kappa}^{(t')}]|\leq\vartheta$ *for all* $\tau,\kappa\in[\mathcal{T}]$.

2. $|F_{y_i}(\mathbf{W}_{y_i}^{(t')},\mathbf{x}_i)-F_{y_j}(\mathbf{W}_{y_j}^{(t')},\mathbf{x}_j)|\leq C_1$ *for all* $i,j\in[n]$.

3. $\ell_{y_i,i}^{'(t')}/\ell_{y_j,j}^{'(t')}\leq\exp\{1+C_1\}=C_2$ *for all* $i,j\in[n]$.

*Proof of Lemma E.7.* We will use induction to prove Lemma E.7. When $t'=0$, we have that $|\sum_{r=1}^{m}[\gamma_{\tau,r,\tau}^{(0)}-\gamma_{\kappa,r,\kappa}^{(0)}]|=0\leq\vartheta$. As we know, $F_{y_i}(\mathbf{W}^{(t)},\mathbf{x}_i)=\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle\mathbf{w}_{y_i,r}^{(t)},\boldsymbol{\xi}_i\rangle)+\sigma(\langle\mathbf{w}_{y_i,r}^{(t)},\boldsymbol{\mu}_{y_i}\rangle)]$, so we have

$$\begin{aligned}|F_{y_i}(\mathbf{W}_{y_i}^{(0)},\mathbf{x}_i)|&=|\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle\mathbf{w}_{y_i,r}^{(0)},\boldsymbol{\xi}_i\rangle)+\sigma(\langle\mathbf{w}_{y_i,r}^{(0)},\boldsymbol{\mu}_{y_i}\rangle)]|\\&\leq\frac{1}{m}\sum_{r=1}^{m}[|\sigma(\langle\mathbf{w}_{y_i,r}^{(0)},\boldsymbol{\xi}_i\rangle)|+|\sigma(\langle\mathbf{w}_{y_i,r}^{(0)},\boldsymbol{\mu}_{y_i}\rangle)|]\\&\leq\frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(0)}+1+\gamma_{y_i,r,y_i}^{(0)}+1]\\&\leq 2,\end{aligned}\tag{18}$$

where the fist inequality is by absolute value inequality, the second inequality is by Lemma E.5. Then we can get that $|F_{y_i}(\mathbf{W}_{y_i}^{(0)},\mathbf{x}_i)-F_{y_j}(\mathbf{W}_{y_j}^{(0)},\mathbf{x}_j)|\leq 4$. We also have that $\ell_{y_i,i}^{'(0)}=1-\mathbf{logit}_i(F,\mathbf{x}_i)$. By Lemma E.3 and (18), we have $-2\leq F_j(\mathbf{W}_j^{(0)},\mathbf{x}_i)-F_{y_i}(\mathbf{W}_{y_i}^{(0)},\mathbf{x}_i)\leq 1$. So $1-1/(1+(\mathcal{T}-1)e^{-2})\leq\ell_{y_i,i}^{'(0)}\leq 1-1/(1+(\mathcal{T}-1)e)$. We have verified the correctness of the three formulas at $t=0$.

Now suppose there exists $\widetilde{t}\leq t$ such that these five conditions hold for any $0\leq t'\leq\widetilde{t}-1$. We aim to prove that these conditions also hold for $t'=\widetilde{t}$.

Firstly we will show that $F_{y_i}(\mathbf{W}_{y_i}^{(t')},\mathbf{x}_i)-F_{y_j}(\mathbf{W}_{y_j}^{(t')},\mathbf{x}_j)$ is dominated by $\frac{1}{m}\sum_{r=1}^{m}[\gamma_{y_i,r,y_i}^{(t')}-\gamma_{y_j,r,y_j}^{(t')}]$. We have that

$$\begin{aligned}&F_{y_i}(\mathbf{W}_{y_i}^{(t')},\mathbf{x}_i)-F_{y_j}(\mathbf{W}_{y_j}^{(t')},\mathbf{x}_j)\\&=\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle\mathbf{w}_{y_i,r}^{(t')},\boldsymbol{\xi}_i\rangle)+\sigma(\langle\mathbf{w}_{y_i,r}^{(t')},\boldsymbol{\mu}_{y_i}\rangle)]-\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle\mathbf{w}_{y_j,r}^{(t')},\boldsymbol{\xi}_j\rangle)+\sigma(\langle\mathbf{w}_{y_j,r}^{(t')},\boldsymbol{\mu}_{y_j}\rangle)]\\&=\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle\mathbf{w}_{y_i,r}^{(t')},\boldsymbol{\xi}_i\rangle)-\sigma(\langle\mathbf{w}_{y_j,r}^{(t')},\boldsymbol{\xi}_j\rangle)]+\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle\mathbf{w}_{y_i,r}^{(t')},\boldsymbol{\mu}_{y_i}\rangle)-\sigma(\langle\mathbf{w}_{y_j,r}^{(t')},\boldsymbol{\mu}_{y_j}\rangle)].\end{aligned}$$

By Lemma E.5, we have that

$$\begin{aligned}|\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle\mathbf{w}_{y_i,r}^{(t')},\boldsymbol{\xi}_i\rangle)-\sigma(\langle\mathbf{w}_{y_j,r}^{(t')},\boldsymbol{\xi}_j\rangle)]|&\leq\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle\mathbf{w}_{y_i,r}^{(t')},\boldsymbol{\xi}_i\rangle)+\sigma(\langle\mathbf{w}_{y_j,r}^{(t')},\boldsymbol{\xi}_j\rangle)]\\&\leq\frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t)}+\bar{\rho}_{y_j,r,j}^{(t)}+2]\\&\leq\frac{1}{m}\sum_{r=1}^{m}[2C'\widetilde{\rho}\alpha+2]\\&\leq 4,\end{aligned}\tag{19}$$

where the third inequality is by (16), and the last inequality is by (17).

By Lemma E.5, we also have

$$
\begin{aligned}
&|\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\mu}_{y_i}\rangle) - \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\mu}_{y_j}\rangle)] - \sum_{r=1}^{m}[\gamma_{y_i,r,y_i}^{(t')} - \gamma_{y_j,r,y_j}^{(t')}]| \\
&= |\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\mu}_{y_i}\rangle) - \gamma_{y_i,r,y_i}^{(t')}] - \sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\mu}_{y_j}\rangle) - \gamma_{y_j,r,y_j}^{(t')}]| \\
&\leq \sum_{r=1}^{m}|[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\mu}_{y_i}\rangle) - \gamma_{y_i,r,y_i}^{(t')}]| + \sum_{r=1}^{m}|[\sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\mu}_{y_j}\rangle) - \gamma_{y_j,r,y_j}^{(t')}]| \\
&\leq 2m. \qquad (20)
\end{aligned}
$$

So we have

$$
\begin{aligned}
&|F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - F_{y_j}(\mathbf{W}_{y_j}^{(t')}, \mathbf{x}_j) - \frac{1}{m}\sum_{r=1}^{m}[\gamma_{y_i,r,y_i}^{(t')} - \gamma_{y_j,r,y_j}^{(t')}]| \\
&= |F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - F_{y_j}(\mathbf{W}_{y_j}^{(t')}, \mathbf{x}_j) - \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\mu}_{y_i}\rangle) - \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\mu}_{y_j}\rangle)] \\
&+ \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\mu}_{y_i}\rangle) - \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\mu}_{y_j}\rangle)] - \frac{1}{m}\sum_{r=1}^{m}[\gamma_{y_i,r,y_i}^{(t')} - \gamma_{y_j,r,y_j}^{(t')}]| \\
&\leq |F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - F_{y_j}(\mathbf{W}_{y_j}^{(t')}, \mathbf{x}_j) - \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\mu}_{y_i}\rangle) - \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\mu}_{y_j}\rangle)]| \\
&+ |\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\mu}_{y_i}\rangle) - \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\mu}_{y_j}\rangle)] - \frac{1}{m}\sum_{r=1}^{m}[\gamma_{y_i,r,y_i}^{(t')} - \gamma_{y_j,r,y_j}^{(t')}]| \\
&\leq 4 + 2 = 6,
\end{aligned}
$$

where the last inequality is by (19) and (20).

By the same analysis, we can also get that

$$
|F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - \frac{1}{m}\sum_{r=1}^{m}\gamma_{y_i,r,y_i}^{(t')}| \leq 3. \qquad (21)
$$

By the update rule (11), we have that

$$
\begin{aligned}
&\frac{1}{m}\sum_{r=1}^{m}[\gamma_{j,r,j}^{(t')} - \gamma_{p,r,p}^{(t')}] = \frac{1}{m}\sum_{r=1}^{m}[\gamma_{j,r,j}^{(t'-1)} - \gamma_{p,r,p}^{(t'-1)}] \\
&+ \frac{\eta}{nm^2}|M_{j,j}|\sum_{y_i=j}(-\ell_{j,i}^{'(t'-1)})\sigma'(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_j\rangle + \gamma_{j,r,j}^{(t'-1)})\mu^2 - \frac{\eta}{nm^2}|M_{p,p}|\sum_{y_i=p}(-\ell_{p,i}^{'(t'-1)})\sigma'(\langle \mathbf{w}_{p,r}^{(0)}, \boldsymbol{\mu}_p\rangle + \gamma_{p,r,p}^{(t'-1)})\mu^2
\end{aligned}
$$

Firstly, We need to explain the ratio of derivatives when the labels are the same. When $y_i = y_j = \tau$, we have that

$$
\begin{aligned}
&\ell_{\tau,i}^{'(t'-1)}/\ell_{\tau,j}^{'(t'-1)} \\
&= \frac{\sum_{p\neq\tau}\exp\{F_p(\mathbf{W}_p^{(t'-1)}, \mathbf{x}_i)\}}{\sum_{p\neq\tau}\exp\{F_p(\mathbf{W}_p^{(t'-1)}, \mathbf{x}_i)\} + \exp\{F_\tau(\mathbf{W}_\tau^{(t'-1)}, \mathbf{x}_i)\}} \cdot \frac{\sum_{p\neq\tau}\exp\{F_p(\mathbf{W}_p^{(t'-1)}, \mathbf{x}_j)\} + \exp\{F_\tau(\mathbf{W}_\tau^{(t'-1)}, \mathbf{x}_j)\}}{\sum_{p\neq\tau}\exp\{F_p(\mathbf{W}_p^{(t'-1)}, \mathbf{x}_j)\}} \\
&\leq \frac{(\mathcal{T}-1)e}{(\mathcal{T}-1)e + \exp\{\frac{1}{m}\sum_{r=1}^{m}\gamma_{\tau,r,\tau}^{(t'-1)} - 3\}} \cdot \frac{\mathcal{T} - 1 + \exp\{\frac{1}{m}\sum_{r=1}^{m}\gamma_{\tau,r,\tau}^{(t'-1)} + 3\}}{\mathcal{T}-1} \\
&\leq e \cdot \max\{e^{-1}, e^6\} = e^7.
\end{aligned}
$$

If $\frac{1}{m}\sum_{r=1}^{m}[\gamma_{j,r,j}^{(t'-1)}-\gamma_{p,r,p}^{(t'-1)}]\leq 0.9\vartheta$, we have that

$$\begin{aligned}\frac{1}{m}\sum_{r=1}^{m}[\gamma_{j,r,j}^{(t')}-\gamma_{p,r,p}^{(t')}]&\leq\frac{1}{m}\sum_{r=1}^{m}[\gamma_{j,r,j}^{(t'-1)}-\gamma_{p,r,p}^{(t'-1)}]+\frac{\eta}{nm^2}|M_{j,j}|\sum_{y_i=j}(-\ell_{j,i}^{'(t'-1)})\sigma'(\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_j\rangle+\gamma_{j,r,j}^{(t'-1)})\mu^2\\&\leq 0.9\vartheta+\frac{\eta}{nm^2}\frac{3m}{4}\frac{6n}{5\mathcal{T}}\mu^2\\&\leq\vartheta,\end{aligned}$$

where the last inequality is by Condition 4.1.

If $\frac{1}{m}\sum_{r=1}^{m}[\gamma_{j,r,j}^{(t'-1)}-\gamma_{p,r,p}^{(t'-1)}]\geq 0.9\vartheta$, we have that

$$F_{y_i}(\mathbf{W}_{y_i}^{(t')},\mathbf{x}_i)-F_{y_j}(\mathbf{W}_{y_j}^{(t')},\mathbf{x}_j)\geq\frac{1}{m}\sum_{r=1}^{m}[\gamma_{y_i,r,y_i}^{(t')}-\gamma_{y_j,r,y_j}^{(t')}]-6\geq 0.9\vartheta-6,$$

so we have

$$\begin{aligned}&\ell_{p,i}^{'(t'-1)}/\ell_{j,q}^{'(t'-1)}\\&=\frac{\sum_{j\neq p}\exp\{F_j(\mathbf{W}_j^{(t'-1)},\mathbf{x}_i)\}}{\sum_{j\neq p}\exp\{F_j(\mathbf{W}_j^{(t'-1)},\mathbf{x}_i)\}+\exp\{F_p(\mathbf{W}_p^{(t'-1)},\mathbf{x}_i)\}}\cdot\frac{\sum_{l\neq j}\exp\{F_l(\mathbf{W}_l^{(t'-1)},\mathbf{x}_q)\}+\exp\{F_j(\mathbf{W}_j^{(t'-1)},\mathbf{x}_q)\}}{\sum_{l\neq j}\exp\{F_l(\mathbf{W}_l^{(t'-1)},\mathbf{x}_q)\}}\\&\geq\frac{(\mathcal{T}-1)}{(\mathcal{T}-1)+\exp\{\frac{1}{m}\sum_{r=1}^{m}\gamma_{p,r,p}^{(t'-1)}+3\}}\cdot\frac{(\mathcal{T}-1)e+\exp\{\frac{1}{m}\sum_{r=1}^{m}\gamma_{j,r,j}^{(t'-1)}-3\}}{(\mathcal{T}-1)e}\\&\geq 0.5\exp\{0.9\vartheta-12\},\end{aligned}$$

So we have that

$$\begin{aligned}&\frac{\eta}{nm^2}|M_{j,j}|\sum_{y_i=j}(-\ell_{j,i}^{'(t'-1)})\sigma'(\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_j\rangle+\gamma_{j,r,j}^{(t'-1)})\mu^2\leq\frac{\eta}{nm^2}m\frac{6n}{5\mathcal{T}}e^7(-\ell_{j,i'}^{'(t'-1)})\mu^2,\\&\frac{\eta}{nm^2}|M_{p,p}|\sum_{y_i=p}(-\ell_{p,i}^{'(t'-1)})\sigma'(\langle\mathbf{w}_{p,r}^{(0)},\boldsymbol{\mu}_p\rangle+\gamma_{p,r,p}^{(t'-1)})\mu^2\\&\geq\frac{\eta}{nm^2}\frac{m}{4}\frac{4n}{5\mathcal{T}}e^{-7}(-\ell_{p,i'}^{'(t'-1)})\mu^2\geq\frac{\eta}{nm^2}\frac{m}{4}\frac{4n}{5\mathcal{T}}e^{-7}0.5\exp\{0.9\vartheta-12\}(-\ell_{j,i'}^{'(t'-1)})\mu^2,\end{aligned}$$

So if we choose $\vartheta=43$, then we have $\frac{\eta}{nm^2}|M_{j,j}|\sum_{y_i=j}(-\ell_{j,i}^{'(t'-1)})\sigma'(\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_j\rangle+\gamma_{j,r,j}^{(t'-1)})\mu^2\leq\frac{\eta}{nm^2}|M_{p,p}|\sum_{y_i=p}(-\ell_{p,i}^{'(t'-1)})\sigma'(\langle\mathbf{w}_{p,r}^{(0)},\boldsymbol{\mu}_p\rangle+\gamma_{p,r,p}^{(t'-1)})\mu^2$, then $\frac{1}{m}\sum_{r=1}^{m}[\gamma_{j,r,j}^{(t')}-\gamma_{p,r,p}^{(t')}]\leq\frac{1}{m}\sum_{r=1}^{m}[\gamma_{j,r,j}^{(t'-1)}-\gamma_{p,r,p}^{(t'-1)}]\leq\vartheta$.

Then we have

$$|F_{y_i}(\mathbf{W}_{y_i}^{(t')},\mathbf{x}_i)-F_{y_j}(\mathbf{W}_{y_j}^{(t')},\mathbf{x}_j)|\leq|\frac{1}{m}\sum_{r=1}^{m}[\gamma_{y_i,r,y_i}^{(t')}-\gamma_{y_j,r,y_j}^{(t')}]|+6\leq\vartheta+6.$$

$$\ell_{y_i,i}^{'(t')}/\ell_{y_j,j}^{'(t')}\leq\exp\{\vartheta+7\}.$$

□

Now we are ready to prove Proposition E.1.

*Proof of Proposition E.1.* Our proof is based on induction. The results are obvious at $t=0$ as all the coefficients are zero. Suppose that there exists $\widetilde{T}\leq T^*$ such that the results in Proposition E.1 hold for all time $0\leq t\leq\widetilde{T}-1$. We aim to prove that they also hold for $t=\widetilde{T}$. Note that according to Lemma E.7, we also have for any $0\leq t\leq\widetilde{T}-1$ that

1. $|\sum_{r=1}^{m}[\gamma_{\tau,r,\tau}^{(t')}-\gamma_{\kappa,r,\kappa}^{(t')}]|\leq\vartheta$ for all $\tau,\kappa\in[\mathcal{T}]$.
2. $|F_{y_i}(\mathbf{W}_{y_i}^{(t')},\mathbf{x}_i)-F_{y_j}(\mathbf{W}_{y_j}^{(t')},\mathbf{x}_j)|\leq C_1$ for all $i,j\in[n]$.
3. $\ell_{y_i,i}^{'(t')}/\ell_{y_j,j}^{'(t')}\leq\exp\{1+C_1\}=C_2$ for all $i,j\in[n]$.

We now prove (14) and (15). First, we note that $\gamma_{j,r,\tau}^{(t)}$ and $\underline{\rho}_{j,r,i}^{(t)}$ are non-increasing sequences. By the convergence theorem for sequences, (9) and (10), if $\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\mu}_\tau\rangle(\tau\neq j)$ and $\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}_i\rangle$ are positive, then they converge to 0. So we have

$$\begin{aligned}
&0\geq\gamma_{j,r,\tau}^{(t)}\geq-\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_\tau\rangle\geq-\beta\geq-\alpha,\\
&0\geq\underline{\rho}_{j,r,i}^{(t)}\geq-\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\xi}_i\rangle-\sum_{i'\neq i}\rho_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle\\
&\geq-\beta-4n\alpha\sqrt{\frac{\log(4n^2/\delta)}{d-\mathcal{T}}}\geq-\alpha,
\end{aligned}$$

where the inequalities are by Lemma C.4 and Condition 4.1.

Now we prove (13) holds for $t=\widetilde{T}$. Consider that for $j=y_i$, we have that

$$\begin{aligned}
-\ell_{j,i}^{'(t)}&\leq\frac{\sum_{p\neq\tau}\exp\{F_p(\mathbf{W}_p^{(t-1)},\mathbf{x}_i)\}}{\sum_{p\neq\tau}\exp\{F_p(\mathbf{W}_p^{(t-1)},\mathbf{x}_i)\}+\exp\{F_\tau(\mathbf{W}_\tau^{(t-1)},\mathbf{x}_i)\}}\\
&\leq\frac{(\mathcal{T}-1)e}{(\mathcal{T}-1)e+\exp\{\frac{1}{m}\sum_{r=1}^{m}\gamma_{\tau,r,\tau}^{(t-1)}-3\}}.
\end{aligned}\tag{22}$$

Now recall the iterative update rule of $\gamma_{j,r,\tau}^{(t)}$:

$$\gamma_{j,r,\tau}^{(t+1)}=\gamma_{j,r,\tau}^{(t)}+\frac{\eta}{nm}\sum_{y_i=\tau}(-\ell_{j,i}^{'(t)})\sigma'(\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_\tau\rangle+\gamma_{j,r,\tau}^{(t)})\mu^2.$$

Let $t_{\tau,r}$ be the last time $t<T^*$ that $\gamma_{\tau,r,\tau}^{(t)}\leq0.5\alpha$. Then by iterating the update rule from $t=t_{\tau,r}$ to $t=\widetilde{T}-1$, we get

$$\begin{aligned}
\gamma_{\tau,r,\tau}^{(\widetilde{T})}=\gamma_{\tau,r,\tau}^{(t_{\tau,r})}&+\underbrace{\frac{\eta}{nm}\sum_{y_i=\tau}(-\ell_{\tau,i}^{'(t_{\tau,r})})\sigma'(\langle\mathbf{w}_{\tau,r}^{(0)},\boldsymbol{\mu}_\tau\rangle+\gamma_{\tau,r,\tau}^{(t_{\tau,r})})\mu^2}_{\spadesuit}\\
&+\underbrace{\sum_{t_{\tau,r}<t<\widetilde{T}}\frac{\eta}{nm}\sum_{y_i=\tau}(-\ell_{\tau,i}^{'(t)})\sigma'(\langle\mathbf{w}_{\tau,r}^{(0)},\boldsymbol{\mu}_\tau\rangle+\gamma_{\tau,r,\tau}^{(t)})\mu^2}_{\clubsuit}
\end{aligned}\tag{23}$$

We first bound ♠ as follows:

$$\begin{aligned}
\spadesuit&\leq\frac{\eta}{nm}\frac{6n}{5\mathcal{T}}(-\ell_{\tau,i}^{'(t_{\tau,r})})\sigma'(\langle\mathbf{w}_{\tau,r}^{(0)},\boldsymbol{\mu}_\tau\rangle+\gamma_{\tau,r,\tau}^{(t_{\tau,r})})\mu^2\\
&\leq\frac{\eta}{nm}\frac{6n}{5\mathcal{T}}\mu^2\\
&\leq0.25\alpha,
\end{aligned}\tag{24}$$

where the first inequality is by Lemma C.1, the second inequality is by $-\ell_{\tau,i}^{'(t_{\tau,r})},\sigma'(\cdot)\leq1$ and the last inequality follows by Condition 4.1.

Second, we bound ♣.

$$\begin{aligned}
\clubsuit &\leq \widetilde{T}\frac{\eta}{nm}\frac{6n}{5\mathcal{T}}(-\ell_{\tau,i}^{'(t-1)})\sigma'(\langle \mathbf{w}_{\tau,r}^{(0)}, \boldsymbol{\mu}_\tau\rangle + \gamma_{\tau,r,\tau}^{(t-1)})\mu^2 \\
&\leq \widetilde{T}\frac{\eta}{nm}\frac{6n}{5\mathcal{T}}\frac{(\mathcal{T}-1)e}{(\mathcal{T}-1)e + \exp\{\frac{1}{m}\sum_{r=1}^m \gamma_{\tau,r,\tau}^{(t-1)} - 3\}}\mu^2 \\
&\leq \frac{2\eta e m^{-1}\widetilde{T}\mu^2}{\exp\{\gamma_{\tau,r,\tau}^{(t-1)} - 3\}} \\
&\leq \frac{2\eta e^4 m^{-1}\widetilde{T}\mu^2}{\exp\{2\log(T^*)\}} \\
&\leq 2\eta e^4 m^{-1} T^{*-1}\mu^2 \\
&\leq 0.25\alpha, \qquad (25)
\end{aligned}$$

where the first inequality is by Lemma C.1, the second inequality is by 22, the fourth inequality is due to our assumption that $\gamma_{\tau,r,\tau}^{(t-1)} \geq 0.5\alpha$, the last inequality follows by Condition 4.1.

Plugging (24) and (25) into 23, we have that $\gamma_{\tau,r,\tau}^{(\widetilde{T})} = \gamma_{\tau,r,\tau}^{(t_{\tau,r})} + \spadesuit + \clubsuit \leq 0.5\alpha + 0.25\alpha + 0.25\alpha = \alpha$.

Then we will use induction to prove (16). First, we focus on $\bar{\rho}_{y_i,r,i}^{(t)}$ that can achieve sufficient growth. When $t = 1$, we have that

$$\begin{aligned}
\gamma_{\tau,r,\tau}^{(1)} &= \frac{\eta}{nm}\sum_{y_i=\tau}(-\ell_{\tau,i}^{'(0)})\sigma'(\langle \mathbf{w}_{\tau,r}^{(0)}, \boldsymbol{\mu}_\tau\rangle)\mu^2, \\
\bar{\rho}_{y_i,r,i}^{(1)} &= \frac{\eta}{nm}(-\ell_{y_i,i}^{'(0)})\sigma'(\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2,
\end{aligned}$$

where the $r$ in $\gamma_{\tau,r,\tau}^{(1)}$ satisfies that $r \in M_{\tau,\tau}$, so we have

$$\gamma_{\tau,r,\tau}^{(1)}/\bar{\rho}_{y_i,r,i}^{(1)} = |i \in [n], y_i = \tau|\frac{(-\ell_{\tau,i'}^{'(0)})\mu^2}{(-\ell_{y_i,i}^{'(0)})\|\boldsymbol{\xi}_i\|_2^2} = \Theta(n\cdot \mathrm{SNR}^2/\mathcal{T}),$$

where the last equation is by Lemma C.1, Lemma C.4 and Lemma E.7. Denote $I_1 = \frac{\eta}{nm}\sum_{y_i=\tau}(-\ell_{j,i}^{'(t)})\sigma'(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau\rangle + \gamma_{j,r,\tau}^{(t)})\mu^2$ and $I_2 = \frac{\eta}{nm}(-\ell_{j,i}^{'(t)})\sigma'(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \rho_{j,r,i}^{(t)} + \sum_{i'\neq i}\rho_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2$. By these lemmas, we can also get that

$$I_1/I_2 = \frac{\frac{\eta}{nm}\sum_{y_i=\tau}(-\ell_{j,i}^{'(t)})\sigma'(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau\rangle + \gamma_{j,r,\tau}^{(t)})\mu^2}{\frac{\eta}{nm}(-\ell_{j,i}^{'(t)})\sigma'(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \rho_{j,r,i}^{(t)} + \sum_{i'\neq i}\rho_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2} = \Theta(n\cdot \mathrm{SNR}^2/\mathcal{T}).$$

Suppose that when $t = t'-1$, $\gamma_{\tau,r,\tau}^{(t'-1)}/\bar{\rho}_{j,r,i}^{(t'-1)} = \Theta(n\cdot \mathrm{SNR}^2/\mathcal{T})$ holds, then we have

$$\Theta(n\cdot \mathrm{SNR}^2/\mathcal{T}) = \min\{\frac{\gamma_{\tau,r,\tau}^{(t'-1)}}{\bar{\rho}_{j,r,i}^{(t')}}, \frac{I_1}{I_2}\} \leq \gamma_{\tau,r,\tau}^{(t')}/\bar{\rho}_{j,r,i}^{(t')} = \frac{\gamma_{\tau,r,\tau}^{(t'-1)} + I_1}{\bar{\rho}_{j,r,i}^{(t')} + I_2} \leq \max\{\frac{\gamma_{\tau,r,\tau}^{(t'-1)}}{\bar{\rho}_{j,r,i}^{(t')}}, \frac{I_1}{I_2}\} = \Theta(n\cdot \mathrm{SNR}^2/\mathcal{T}),$$

where the inequalities are by the same method in Lemma E.6. By induction, we can get that $\gamma_{\tau,r,\tau}^{(t')}/\bar{\rho}_{j,r,i}^{(t')} = \Theta(n\cdot \mathrm{SNR}^2/\mathcal{T})$ for all $0 < t' \leq T^*$. As we have got that $0 \leq \gamma_{\tau,r,\tau}^{(t')} \leq \alpha$, so we can have that $0 \leq \bar{\rho}_{j,r,i}^{(t')} \leq C'\widetilde{\rho}\alpha$. Here the proof completes.

□

By the definition of $(-\ell_{j,i}^{'(t)})$ and Lemma E.3, we have that

$$\frac{\mathcal{T}-1}{e} \leq (-\ell_{y_i,i}^{'(t)})/\ell_{j,i}^{'(t)} \leq (\mathcal{T}-1)e. \qquad (26)$$

We also have that

**Lemma E.8.**

$$\gamma^{(t)}_{\tau,r,\tau}/\gamma^{(t)}_{\tau',r',\tau'} \leq C_3, \tag{27}$$
$$\bar{\rho}^{(t)}_{y_i,r_1,i}/\bar{\rho}^{(t)}_{y_{i'},r_2,i'} \leq C_4 \tag{28}$$

*for all* $0 < t \leq T^*$, $\tau, \tau' \in \mathcal{T}$, $r \in M_{\tau,\tau}$, $r' \in M_{\tau',\tau'}$, $i, i' \in [n]$ *and* $r_1, r_2$ *belong to the set of indices for which* $\bar{\rho}^{(t)}_{y_i,r_1,i}$, $\bar{\rho}^{(t)}_{y_{i'},r_2,i'}$ *can achieve sufficient growth.*

*Proof of Lemma E.8.* We use induction to prove this lemma. When $t = 1$, we have that

$$\begin{aligned}
\gamma^{(1)}_{\tau,r,\tau} &= \frac{\eta}{nm}\sum_{y_i=\tau}(-\ell'^{(0)}_{\tau,i})\sigma'(\langle \mathbf{w}^{(0)}_{\tau,r}, \boldsymbol{\mu}_\tau\rangle)\mu^2,\\
\gamma^{(1)}_{\tau',r',\tau'} &= \frac{\eta}{nm}\sum_{y_i=\tau'}(-\ell'^{(0)}_{\tau',i})\sigma'(\langle \mathbf{w}^{(0)}_{\tau',r'}, \boldsymbol{\mu}_{\tau'}\rangle)\mu^2,\\
\bar{\rho}^{(1)}_{y_i,r_1,i} &= \frac{\eta}{nm}(-\ell'^{(0)}_{y_i,i})\sigma'(\langle \mathbf{w}^{(0)}_{y_i,r_1}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2,\\
\bar{\rho}^{(1)}_{y_{i'},r_2,i'} &= \frac{\eta}{nm}(-\ell'^{(0)}_{y_{i'},i'})\sigma'(\langle \mathbf{w}^{(0)}_{y_{i'},r_2}, \boldsymbol{\xi}_{i'}\rangle)\|\boldsymbol{\xi}_{i'}\|_2^2,
\end{aligned}$$

Then we have that

$$\begin{aligned}
\gamma^{(1)}_{\tau,r,\tau}/\gamma^{(1)}_{\tau',r',\tau'} &\leq \frac{3}{2}C_2\\
\bar{\rho}^{(1)}_{y_i,r_1,i}/\bar{\rho}^{(1)}_{y_{i'},r_2,i'} &\leq 3C_2,
\end{aligned}$$

Suppose that (27) and (28) hold when $0 < t \leq \widetilde{T} - 1$, we prove that when $t = \widetilde{T}$, they still hold. We have that

$$\begin{aligned}
\gamma^{(\widetilde{T})}_{\tau,r,\tau}/\gamma^{(\widetilde{T})}_{\tau',r',\tau'} &\leq \max\{\gamma^{(\widetilde{T}-1)}_{\tau,r,\tau}/\gamma^{(\widetilde{T}-1)}_{\tau',r',\tau'}, \frac{\frac{\eta}{nm}\sum_{y_i=\tau}(-\ell'^{(\widetilde{T}-1)}_{j,i})\sigma'(\langle \mathbf{w}^{(0)}_{\tau,r}, \boldsymbol{\mu}_\tau\rangle + \gamma^{(\widetilde{T}-1)}_{\tau,r,\tau})\mu^2}{\frac{\eta}{nm}\sum_{y_i=\tau'}(-\ell'^{(\widetilde{T}-1)}_{j,i})\sigma'(\langle \mathbf{w}^{(0)}_{\tau',r}, \boldsymbol{\mu}_{\tau'}\rangle + \gamma^{(\widetilde{T}-1)}_{\tau',r',\tau})\mu^2}\}\\
&\leq \frac{3}{2}C_2,\\
\bar{\rho}^{(\widetilde{T})}_{y_i,r_1,i}/\bar{\rho}^{(\widetilde{T})}_{y_{i'},r_2,i'} &\leq \max\{\bar{\rho}^{(\widetilde{T}-1)}_{y_i,r_1,i}/\bar{\rho}^{(\widetilde{T}-1)}_{y_{i'},r_2,i'}, \frac{\frac{\eta}{nm}(-\ell'^{(\widetilde{T}-1)}_{y_i,i})\sigma'(\langle \mathbf{w}^{(\widetilde{T}-1)}_{y_i,r_1}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2}{\frac{\eta}{nm}(-\ell'^{(\widetilde{T}-1)}_{y_{i'},i'})\sigma'(\langle \mathbf{w}^{(\widetilde{T}-1)}_{y_{i'},r_2}, \boldsymbol{\xi}_{i'}\rangle)\|\boldsymbol{\xi}_{i'}\|_2^2}\}\\
&\leq 3C_2.
\end{aligned}$$

If we choose $C_3 = \frac{3}{2}C_2$, $C_4 = 3C_2$, then the proof completes.

□

**Lemma E.9.** *Under Condition 4.1, the following hold for all* $0 \leq t \leq T^*$, $j, \tau \in [\mathcal{T}]$, $r \in [m]$, *and* $i \in [n]$*:*

1. *If* $j \neq \tau$, $0 \leq \langle \mathbf{w}^{(t)}_{j,r}, \boldsymbol{\mu}_\tau\rangle \leq \beta$ *(converging to* $0$*);*
2. *If* $j = \tau$, $\langle \mathbf{w}^{(t)}_{j,r}, \boldsymbol{\mu}_\tau\rangle$ *is non-decreasing (non-zero iff* $\langle \mathbf{w}^{(0)}_{j,r}, \boldsymbol{\mu}_\tau\rangle > 0$*);*
3. *If* $\langle \mathbf{w}^{(0)}_{j,r}, \boldsymbol{\xi}_i\rangle > 0$*:*
   - *For* $j \neq y_i$, $\langle \mathbf{w}^{(t)}_{j,r}, \boldsymbol{\xi}_i\rangle \leq \beta$ *(non-increasing);*
   - *For* $j = y_i$, $\langle \mathbf{w}^{(t)}_{j,r}, \boldsymbol{\xi}_i\rangle > 0$ *(non-decreasing).*

*Proof of Lemma E.9.* From the weight update rule (6), we derive the recursions for inner products with signal and noise vectors:

$$\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\mu}_\tau \rangle = \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle + \frac{\eta\mu^2}{nm} \sum_{y_i=\tau} \left(-\ell_{j,i}^{'(t)}\right) \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle\right), \tag{29}$$

$$\begin{aligned}\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle = \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle &+ \frac{\eta}{nm} \left(-\ell_{j,i}^{'(t)}\right) \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right) \|\boldsymbol{\xi}_i\|_2^2 \\ &+ \frac{\eta}{nm} \sum_{s\neq i} \left(-\ell_{j,s}^{'(t)}\right) \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_s \rangle\right) \langle \boldsymbol{\xi}_s, \boldsymbol{\xi}_i \rangle. \end{aligned} \tag{30}$$

For the signal vector inner product $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle$, note that $\sigma'(\cdot) \geq 0$ and $\sum_{y_i=\tau} \left(-\ell_{j,i}^{'(t)}\right) \geq 0$, so the update term in (29) is non-negative. If $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle \leq 0$, the update term vanishes (via $\sigma'(x)x = \sigma(x) \geq 0$), implying $\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\mu}_\tau \rangle = \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle$. For $j \neq \tau$, Condition 4.1 ensures the update term is negligible, leading to convergence to 0 and $0 \leq \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle \leq \beta$. For $j = \tau$, the non-negative update term implies $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle$ is non-decreasing, and it remains non-zero iff initialized as such.

For the noise vector inner product $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$, assume $\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle > 0$. When $j \neq y_i$, we have $\left(-\ell_{j,i}^{'(t)}\right) < 0$. By triangle inequality and Lemma C.4, combined with $|-\ell_{j,s}^{'(t)}| \leq -\ell_{y_s,s}^{'(t)}$, the cross term satisfies:

$$\left| \sum_{s\neq i} \left(-\ell_{j,s}^{'(t)}\right) \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_s \rangle\right) \langle \boldsymbol{\xi}_s, \boldsymbol{\xi}_i \rangle \right| \leq \widetilde{O}\left(\sigma_p^2 \sqrt{d} \sum_{s\neq i} \left(-\ell_{y_s,s}^{'(t)}\right)\right).$$

By (26), Lemma E.7, and Condition 4.1:

$$\left(-\ell_{j,i}^{'(t)}\right) \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right) \|\boldsymbol{\xi}_i\|_2^2 \leq -\widetilde{O}\left(\sigma_p^2 \sqrt{d} \sum_{s\neq i} \left(-\ell_{y_s,s}^{'(t)}\right)\right),$$

so substituting into (30) gives $\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle \leq \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle \leq \beta$.

When $j = y_i$, $\left(-\ell_{j,i}^{'(t)}\right) > 0$. By Lemma C.4, Condition 4.1, and Lemma E.7:

$$\left(-\ell_{j,i}^{'(t)}\right) \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle\right) \|\boldsymbol{\xi}_i\|_2^2 \geq \widetilde{O}\left(\sigma_p^2 \sqrt{d} \sum_{s\neq i} \left(-\ell_{y_s,s}^{'(t)}\right)\right),$$

dominating the cross term in (30). Thus, $\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle \geq \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle > 0$.

This completes the proof. □

## F. Decoupling with a Two-Stage Analysis

### F.1. First Stage

**Lemma F.1.** *If we denote*

$$\widetilde{\rho} := \mathcal{T} n^{-1} \cdot \mathrm{SNR}^{-2},$$

then there exist

$$T_1 = C_5 m \mathcal{T} \eta^{-1} \mu^{-2}, T_2 = C_6 m \mathcal{T} \eta^{-1} \mu^{-2}$$

where $C_5 = \Theta(1)$ is a large constant and $C_6 = \Theta(1)$ is a small constant, such that

1. $\gamma_{\tau,r,\tau}^{(T_1)} \geq 3$ for any $r \in M_{\tau,\tau} = \{r \in [m] : \langle \mathbf{w}_{\tau,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle > 0\}$ and $\tau \in [\mathcal{T}]$.
2. $\bar{\rho}_{y_i,r,i}^{(t)} = \Theta(\widetilde{\rho})$ when $T_2 \leq t \leq T_1$ for $r \in N_{y_i,i}$.
3. $0 \leq |\gamma_{j,r,\tau}^{(t)}| \leq O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T}n^{1/4}\sigma_0\mu)$ for all $j \neq \tau$ and $0 \leq t \leq T^*$.
4. $0 \leq |\underline{\rho}_{j,r,i}^{(t)}| \leq \max\{O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T}n^{1/4}\sigma_0\mu), O(\sqrt{\log(n^2/\delta)}n\log(T^*)/\sqrt{d-\mathcal{T}}\}$ for all $0 \leq t \leq T^*$.

*Proof of Lemma F.1.* First, we note that $\gamma_{j,r,\tau}^{(t)}(j \neq \tau)$ and $\underline{\rho}_{j,r,i}^{(t)}$ are non-increasing sequences. By the convergence theorem for sequences, (9) and (10), if $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle (\tau \neq j)$ and $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$ are positive, then they converge to 0. So we have

$$\begin{aligned}
&0 \geq \gamma_{j,r,\tau}^{(t)} \geq -\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle \geq -\beta, \\
&0 \geq \underline{\rho}_{j,r,i}^{(t)} \geq -\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle - \sum_{i' \neq i} \rho_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle \\
&\geq -\beta - 4n\alpha\sqrt{\frac{\log(4n^2/\delta)}{d-\mathcal{T}}}.
\end{aligned}$$

By the definition of $\beta$ and Lemma C.5, we have that

$$\begin{aligned}
\beta &= \max_{j,r,\tau,i} \{|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle|, |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle|\} \\
&\leq \max\{\sigma_0\mu \cdot \sqrt{2\log(8m\mathcal{T}^2/\delta)}, 2\sigma_0\sigma_p\sqrt{d-\mathcal{T}} \cdot \sqrt{\log(8mn\mathcal{T}/\delta)}\} \\
&\leq O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T}n^{1/4}\sigma_0\mu).
\end{aligned}$$

So we have that

$$0 \leq |\gamma_{j,r,\tau}^{(t)}| \leq \beta \leq O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T}n^{1/4}\sigma_0\mu),$$

and

$$\begin{aligned}
0 \leq |\underline{\rho}_{j,r,i}^{(t)}| &\leq \beta + 4n\alpha\sqrt{\frac{\log(4n^2/\delta)}{d-\mathcal{T}}} \\
&\leq \max\{O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T}n^{1/4}\sigma_0\mu), O(\sqrt{\log(n^2/\delta)}n\log(T^*)/\sqrt{d-\mathcal{T}}\}.
\end{aligned}$$

By the update rule for $\bar{\rho}_{y_i,r,i}^{(t)}$, we have that

$$\begin{aligned}
\bar{\rho}_{y_i,r,i}^{(t+1)} &= \bar{\rho}_{y_i,r,i}^{(t)} + \frac{\eta}{nm}(-\ell_{y_i,i}^{'(t)})\sigma'(\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle + \bar{\rho}_{y_i,r,i}^{(t)} + \sum_{i' \neq i} \rho_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle) \|\boldsymbol{\xi}_i\|_2^2 \\
&\leq \bar{\rho}_{y_i,r,i}^{(t)} + \frac{\eta}{nm}\|\boldsymbol{\xi}_i\|_2^2 \\
&\leq \bar{\rho}_{y_i,r,i}^{(t)} + \frac{2\eta}{nm}\sigma_p^2 d,
\end{aligned}$$

where the first inequality is by that $-\ell_{y_i,i}^{'(t)} \leq 1$ and $\sigma'(\cdot) \leq 1$, and the last inequality follows by Lemma C.4. Note that $\bar{\rho}_{y_i,r,i}^{(0)} = 0$ and recursively use the inequality $t$ times we have $\bar{\rho}_{y_i,r,i}^{(t)} \leq \frac{2\eta t}{nm}\sigma_p^2 d$.

Since $\mathcal{T}n^{-1}\mathrm{SNR}^{-2} = \widetilde{\rho}$, we have that

$$T_1 = C_5 m\mathcal{T}\eta^{-1}\mu^{-2} = C_5 mn\eta^{-1}\sigma_p^{-2}d^{-1}\widetilde{\rho},$$

and it follows that

$$\bar{\rho}_{y_i,r,i}^{(t)} \leq \frac{2\eta t}{nm}\sigma_p^2 d \leq 2C_5\widetilde{\rho}$$

for all $t \leq T_1$.

For each $\tau, r$, denote by $T_1$ the last time in the period $[0, T_1]$ satisfying that $\max_{\tau,r} \gamma_{\tau,r,\tau}^{(t)} \leq 3$. Then for $0 \leq t \leq T_1$, $\max_{j,r,i} |\rho_{j,r,i}^{(t)}| = O(1)$ and $\max_{j,r,i} |\gamma_{j,r,i}^{(t)}| = O(1)$. Then we have that $F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) = O(1)$ for all $j \in [\mathcal{T}]$ and $i \in [n]$. Thus there exists a positive constant $C$ such that $-\ell_{\tau,i}^{'(t)} \geq C$ for $0 \leq t \leq T$. Then we have

$$\begin{aligned}\gamma_{\tau,r,\tau}^{(t+1)} &= \gamma_{\tau,r,\tau}^{(t)} + \frac{\eta}{nm} \sum_{y_i=\tau} (-\ell_{\tau,i}^{'(t)}) \sigma'(\langle \mathbf{w}_{\tau,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle + \gamma_{\tau,r,\tau}^{(t)}) \mu^2 \\ &\geq \gamma_{\tau,r,\tau}^{(t)} + \frac{4\eta C \mu^2}{5m\mathcal{T}}.\end{aligned}$$

Therefore, $\gamma_{\tau,r,\tau}^{(t)} \geq \frac{4\eta C\mu^2}{5m\mathcal{T}} t$ and $\gamma_{\tau,r,\tau}^{(t)}$ will reach 3 within

$$T_1 = C_5 m \mathcal{T} \eta^{-1} \mu^{-2}$$

iterations for any $\tau \in [\mathcal{T}]$ and $i \in [n]$, where $C_5$ can be taken as $4/C$.

Next, we will discuss the lower bound of the growth of $\bar{\rho}_{y_i,r,i}^{(t)}$. For $\gamma_{\tau,r,\tau}^{(t)}$, we have that

$$\gamma_{\tau,r,\tau}^{(t+1)} - \gamma_{\tau,r,\tau}^{(t)} = \frac{\eta}{nm} \sum_{y_i=\tau} \left(-\ell_{\tau,i}^{'(t)}\right) \sigma' \left(\langle \mathbf{w}_{\tau,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle\right) \mu^2 \leq \frac{2\eta\mu^2}{m\mathcal{T}},$$

where the inequality is by Lemma C.1 and $-\ell_{\tau,i}^{'(t)}, \sigma'(\cdot) \leq 1$.

So we have that $\gamma_{\tau,r,\tau}^{(t)} \leq \frac{2\eta\mu^2}{m\mathcal{T}} t$. Therefore, $\max_{\tau,r} \gamma_{\tau,r,\tau}^{(t)}$ will be smaller than 1 and $\max_{r,i} \bar{\rho}_{y_i,r,i}^{(t)} \leq O(\widetilde{\rho}) = O(1)$ within

$$T_2 = C_6 m \mathcal{T} \eta^{-1} \mu^{-2}$$

iterations, where $C_6$ can be taken as 0.5. Then we have that $F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) = O(1)$ for all $j \in [\mathcal{T}]$ and $i \in [n]$. Thus there exists a positive constant $C$ such that $-\ell_{y_i,i}^{'(t)} \geq C$ for $0 \leq t \leq T_2$. Then we have that

$$\begin{aligned}\bar{\rho}_{y_i,r,i}^{(t+1)} &= \bar{\rho}_{y_i,r,i}^{(t)} + \frac{\eta}{nm}(-\ell_{y_i,i}^{'(t)}) \sigma'(\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle + \bar{\rho}_{y_i,r,i}^{(t)} + \sum_{i' \neq i} \rho_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle) \|\boldsymbol{\xi}_i\|_2^2 \\ &\geq \bar{\rho}_{y_i,r,i}^{(t)} + \frac{\eta C}{nm} \|\boldsymbol{\xi}_i\|_2^2 \\ &\geq \bar{\rho}_{y_i,r,i}^{(t)} + \frac{\eta C}{4nm} \sigma_p^2 d,\end{aligned}$$

where the last inequality is by Lemma C.4. So that $\bar{\rho}_{y_i,r,i}^{(t)} \geq \frac{\eta C}{4nm} \sigma_p^2 dt$ and $\bar{\rho}_{y_i,r,i}^{(T_2)} \geq \frac{CC_6\mathcal{T}}{n\mathrm{SNR}^2} = \Theta(\widetilde{\rho})$. Here the proof completes.

□

### F.2. Second Stage

By the signal-noise decomposition, at the end of the first stage, we have

$$\mathbf{w}_{j,r}^{(T_1)} = \mathbf{w}_{j,r}^{(0)} + \sum_{\tau=1}^{\mathcal{T}} \gamma_{j,r,\tau}^{(T_1)} \cdot \mu^{-2} \boldsymbol{\mu}_\tau + \sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(T_1)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \boldsymbol{\xi}_i + \sum_{i=1}^{n} \underline{\rho}_{j,r,i}^{(T_1)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \boldsymbol{\xi}_i$$

for $j \in [\mathcal{T}]$ and $r \in [m]$. By the results we get in the first stage, we know that at the beginning of this stage, we have the following property holds:

1. $\gamma_{\tau,r,\tau}^{(T_1)} \geq 3$ for any $r \in M_{\tau,\tau} = \{r \in [m] : \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle > 0\}$ and $\tau \in [\mathcal{T}]$.

2. $\bar{\rho}_{y_i,r,i}^{(t)} = \Theta(\widetilde{\rho})$ when $T_2 \leq t \leq T_1$ for the $\bar{\rho}_{y_i,r,i}^{(t)}$ that can grow large.

3. $0 \leq |\gamma_{j,r,\tau}^{(t)}| \leq O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T}n^{1/4}\sigma_0\mu)$ for all $j \neq \tau$ and $0 \leq t \leq T^*$.

4. $0 \leq |\underline{\rho}_{j,r,i}^{(t)}| \leq \max\{O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T}n^{1/4}\sigma_0\mu), O(\sqrt{\log(n^2/\delta)}n\log(T^*)/\sqrt{d-\mathcal{T}}\}$ for all $0 \leq t \leq T^*$.

Now we choose $\mathbf{W}^*$ as follows:

$$\mathbf{w}_{j,r}^* = \mathbf{w}_{j,r}^{(0)} + 10\log(2\mathcal{T}/\epsilon) \cdot \mu^{-2}\boldsymbol{\mu}_j$$

**Lemma F.2.** *Under Condition 4.1, the following inequality holds:*

$$\|\nabla_{\mathbf{W}^{(t)}} L_S(\mathbf{W}^{(t)})\|_F^2 \leq O\left(\mathcal{T}^2 \max\left\{\sigma_p^2 d, \mu^2\right\}\right) L_S(\mathbf{W}^{(t)}) \tag{31}$$

*for all* $0 \leq t \leq T^*$.

*Proof of Lemma F.2.* First, we compute the Frobenius norm $\|\nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)\|_F$ as follows:

$$\begin{aligned}
\|\nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)\|_F &= \sqrt{\sum_{r=1}^m \|\nabla_{\mathbf{w}_{j,r}^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)\|_2^2} \\
&= \frac{1}{m}\sqrt{\sum_{r=1}^m \left\|\sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right)\boldsymbol{\xi}_i + \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i}\rangle\right)\boldsymbol{\mu}_{y_i}\right\|_2^2} \\
&\leq \frac{1}{m}\sum_{r=1}^m \left\|\sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right)\boldsymbol{\xi}_i + \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i}\rangle\right)\boldsymbol{\mu}_{y_i}\right\|_2 \\
&\leq \frac{1}{m}\sum_{r=1}^m \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right)\|\boldsymbol{\xi}_i\|_2 + \frac{1}{m}\sum_{r=1}^m \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i}\rangle\right)\|\boldsymbol{\mu}_{y_i}\|_2 \\
&\leq \|\boldsymbol{\xi}_i\|_2 + \mu \\
&\leq \sqrt{\frac{3}{2}}\sigma_p\sqrt{d-\mathcal{T}} + \mu,
\end{aligned} \tag{32}$$

where the first inequality follows from the Cauchy-Schwarz inequality (sum of squares inequality), the second from the triangle inequality for vector norms, the third from the fact that $\sigma'(\cdot) \leq 1$, and the last inequality is a consequence of Lemma C.4.

Next, consider the case $j \neq y_i$. We derive the following bound:

$$\begin{aligned}
|\ell_{j,i}^{'(t)}|\|\nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)\|_F^2 &= \frac{\exp\left\{F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)\right\}}{\exp\left\{F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)\right\} + \sum_{j\neq y_i}\exp\left\{F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)\right\}}\|\nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)\|_F^2 \\
&\leq \frac{e}{e+\mathcal{T}-2+\exp\left\{F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)\right\}}\left(\sqrt{\frac{3}{2}}\sigma_p\sqrt{d-\mathcal{T}}+\mu\right)^2 \\
&\leq \exp\left\{1 - F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)\right\}\left(\sqrt{\frac{3}{2}}\sigma_p\sqrt{d-\mathcal{T}}+\mu\right)^2,
\end{aligned} \tag{33}$$

where the first inequality uses Lemma E.3 and (32).

For the case $j = y_i$, we have:

$$
\begin{aligned}
|\ell_{j,i}^{'(t)}| \|\nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)\|_F^2 &= \frac{\exp\left\{F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)\right\}}{\exp\left\{F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)\right\} + \sum_{j \neq y_i} \exp\left\{F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)\right\}} \|\nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)\|_F^2 \\
&\leq \frac{(\mathcal{T}-1)e}{(\mathcal{T}-1)e + \exp\left\{F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)\right\}} \left(\sqrt{\frac{3}{2}}\sigma_p\sqrt{d-\mathcal{T}} + \mu\right)^2 \\
&\leq \mathcal{T} \exp\left\{1 - F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)\right\} \left(\sqrt{\frac{3}{2}}\sigma_p\sqrt{d-\mathcal{T}} + \mu\right)^2, \qquad (34)
\end{aligned}
$$

where the first inequality again invokes Lemma E.3 and (32).

Combining (33) and (34), we obtain:

$$
\begin{aligned}
\sum_{j=1}^{\mathcal{T}} |\ell_{j,i}^{'(t)}|^2 \|\nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)\|_F^2 &\leq |\ell_{y_i,i}^{'(t)}| \sum_{j=1}^{\mathcal{T}} |\ell_{j,i}^{'(t)}| \|\nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)\|_F^2 \\
&\leq |\ell_{y_i,i}^{'(t)}| \mathcal{T}^2 \exp\left\{1 - F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i)\right\} \left(\sqrt{\frac{3}{2}}\sigma_p\sqrt{d-\mathcal{T}} + \mu\right)^2 \\
&\leq |\ell_{y_i,i}^{'(t)}| O\left(\mathcal{T}^2 \max\left\{\sigma_p^2 d, \mu^2\right\}\right), \qquad (35)
\end{aligned}
$$

where the first inequality uses $|\ell_{j,i}^{'(t)}| \leq |\ell_{y_i,i}^{'(t)}|$ for all $j$, the second follows from (33) and (34), and the last inequality holds because $F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i) \geq 0$ (hence $\exp\{1 - F_{y_i}(\cdot)\} \leq e$) and $\left(\sqrt{\frac{3}{2}}\sigma_p\sqrt{d-\mathcal{T}} + \mu\right)^2 = O\left(\max\left\{\sigma_p^2 d, \mu^2\right\}\right)$.

Finally, we bound $\|\nabla_{\mathbf{W}^{(t)}} L_S(\mathbf{W}^{(t)})\|_F^2$:

$$
\begin{aligned}
\|\nabla_{\mathbf{W}^{(t)}} L_S(\mathbf{W}^{(t)})\|_F^2 &= \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_{\mathbf{W}^{(t)}} L(\mathbf{W}^{(t)}, \mathbf{x}_i) \right\|_F^2 \\
&\leq \left( \frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\mathbf{W}^{(t)}} L(\mathbf{W}^{(t)}, \mathbf{x}_i)\|_F \right)^2 \\
&= \left( \frac{1}{n} \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{\mathcal{T}} |\ell_{j,i}^{'(t)}|^2 \|\nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)\|_F^2} \right)^2 \\
&\leq \left( \frac{1}{n} \sum_{i=1}^{n} \sqrt{|\ell_{y_i,i}^{'(t)}| O\left(\mathcal{T}^2 \max\left\{\sigma_p^2 d, \mu^2\right\}\right)} \right)^2 \\
&\leq O\left(\mathcal{T}^2 \max\left\{\sigma_p^2 d, \mu^2\right\}\right) \frac{1}{n} \sum_{i=1}^{n} |\ell_{y_i,i}^{'(t)}| \\
&\leq O\left(\mathcal{T}^2 \max\left\{\sigma_p^2 d, \mu^2\right\}\right) \frac{1}{n} \sum_{i=1}^{n} \ell^{(t)}(F(\mathbf{x}_i, y_i)) \\
&= O\left(\mathcal{T}^2 \max\left\{\sigma_p^2 d, \mu^2\right\}\right) L_S(\mathbf{W}^{(t)}), \qquad (36)
\end{aligned}
$$

where the first inequality is the triangle inequality for the Frobenius norm, the second uses (35), the third applies Hölder's inequality, and the last inequality follows from the fact that $\frac{1}{x+1} \leq \log\left(1 + \frac{1}{x}\right)$ for all $x > 0$ (which relates the derivative of the cross-entropy loss to the loss itself).This completes the proof. □

**Lemma F.3.** *Under Condition 4.1, the following inequality holds:*

$$
\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \geq \eta L_S(\mathbf{W}^{(t)}) - \eta\epsilon
$$

*for all $T_1 \leq t \leq T^*$.*

*Proof of Lemma F.3.* To establish the desired inequality, we first derive a series of foundational inner product identities and bounds that will support the subsequent analysis. We begin with a key identity involving the gradient of $F_j$ with respect to the weight matrix $\mathbf{W}_j^{(t)}$. Expanding the Frobenius inner product $\langle \nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{W}_j^{(t)} \rangle$ as a sum over the individual weight vectors $\{\mathbf{w}_{j,r}^{(t)}\}_{r=1}^m$, we leverage the fundamental property of the activation function derivative $\sigma'(x)x = \sigma(x)$ for all $x \in \mathbb{R}$. This substitution simplifies the sum to a direct expression of $F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)$, yielding the identity:

$$\begin{aligned}\langle \nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{W}_j^{(t)} \rangle &= \sum_{r=1}^m \langle \nabla_{\mathbf{w}_{j,r}^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{w}_{j,r}^{(t)} \rangle \\ &= \sum_{r=1}^m \frac{1}{m} \left[ \sigma'\left( \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle \right) \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle + \sigma'\left( \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle \right) \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle \right] \\ &= F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i). \end{aligned} \tag{37}$$

Next, we decompose the inner product $\langle \nabla_{\mathbf{w}_{j,r}^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{w}_{j,r}^* \rangle$ by considering the structure of the optimal weight matrix $\mathbf{W}^*$ and the orthogonality between the noise vector $\boldsymbol{\xi}_i$ and the signal vectors $\{\boldsymbol{\mu}_\tau\}_{\tau \in [\mathcal{T}]}$. Two distinct cases arise based on whether the class index $j$ matches the true label $y_i$ of the input $\mathbf{x}_i$. When $j = y_i$, the optimal weight vector $\mathbf{w}_{y_i,r}^*$ includes an additional signal component, leading to the decomposition:

$$\begin{aligned}\langle \nabla_{\mathbf{w}_{y_i,r}^{(t)}} F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i), \mathbf{w}_{y_i,r}^* \rangle &= \frac{1}{m}\sigma'\left( \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle \right) \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\mu}_{y_i} \rangle + \frac{1}{m}\sigma'\left( \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle \right) \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle \\ &\quad + \frac{1}{m}\sigma'\left( \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle \right) \cdot 10 \log(2\mathcal{T}/\epsilon). \end{aligned} \tag{38}$$

In contrast, when $j \neq y_i$, the optimal weight vector $\mathbf{w}_{j,r}^*$ lacks this additional component, resulting in the simpler decomposition:

$$\langle \nabla_{\mathbf{w}_{j,r}^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{w}_{j,r}^* \rangle = \frac{1}{m}\sigma'\left( \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle \right) \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{y_i} \rangle + \frac{1}{m}\sigma'\left( \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle \right) \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle. \tag{39}$$

Building on these decompositions, we derive bounds for the aggregated inner product $A_{j,i} := \langle \nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{W}_j^* \rangle$ by summing the component-wise results over $r \in [m]$. For the case $j = y_i$, summing the expression in (38) and applying Lemma C.2, along with the definition of $\beta := \max_{j,r,\tau,i}\{|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle|, |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle|\}$, we obtain a lower bound. Condition 4.1 ensures that this bound simplifies to $A_{y_i,i} \geq 2\log(2\mathcal{T}/\epsilon)$, as shown below:

$$\begin{aligned} A_{y_i,i} &= \sum_{r=1}^m \langle \nabla_{\mathbf{w}_{y_i,r}^{(t)}} F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i), \mathbf{w}_{y_i,r}^* \rangle \\ &= \sum_{r=1}^m \frac{1}{m}\left[ \sigma'\left( \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle \right) \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\mu}_{y_i} \rangle + \sigma'\left( \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle \right) \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle + \sigma'\left( \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle \right) \cdot 10 \log(2\mathcal{T}/\epsilon) \right] \\ &\geq 0.25 \cdot 10 \log(2\mathcal{T}/\epsilon) - \beta - \beta \geq 2\log(2\mathcal{T}/\epsilon). \end{aligned} \tag{40}$$

For $j \neq y_i$, summing the expression in (39) and applying the triangle inequality to bound the absolute value of the sum, we find that $|A_{j,i}| \leq 2\beta$, a result that follows directly from the definition of $\beta$ and the fact that $\sigma'(\cdot) \leq 1$:

$$\begin{aligned} |A_{j,i}| &= \left| \sum_{r=1}^m \langle \nabla_{\mathbf{w}_{j,r}^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{w}_{j,r}^* \rangle \right| \\ &= \left| \sum_{r=1}^m \frac{1}{m}\left[ \sigma'\left( \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle \right) \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{y_i} \rangle + \sigma'\left( \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle \right) \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle \right] \right| \\ &\leq \sum_{r=1}^m \frac{1}{m}\left[ \sigma'\left( \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle \right) |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{y_i} \rangle| + \sigma'\left( \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle \right) |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle| \right] \\ &\leq 2\beta. \end{aligned} \tag{41}$$

With these bounds in place, we now establish a lower bound for the critical inner product $\langle \nabla L_S(\mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \rangle$, which links the gradient of the loss to the difference between the current and optimal weight matrices. Expanding this inner product by first decomposing over the class index $j$ and then over the training samples $i$, we substitute the identity for $\langle \nabla_{\mathbf{W}_j^{(t)}} F_j, \mathbf{W}_j^{(t)} \rangle$ and the definition of $A_{j,i}$ to rewrite the expression in terms of $F_j$ and $A_{j,i}$. Leveraging the convexity of the cross-entropy loss $\ell$, we relate the linear term involving the loss derivative to the loss itself, introducing a logarithmic correction term. The bounds on $A_{j,i}$ derived earlier, specifically $A_{y_i,i} \geq 2\log(2\mathcal{T}/\epsilon)$ and $|A_{j,i}| \leq 2\beta$ for $j \neq y_i$, ensure that this logarithmic term is bounded by $0.5\epsilon$ under Condition 4.1, leading to the lower bound:

$$\begin{aligned}
\langle \nabla L_S(\mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \rangle &= \sum_{j=1}^{\mathcal{T}} \langle \nabla_{\mathbf{W}_j^{(t)}} L_S(\mathbf{W}^{(t)}), \mathbf{W}_j^{(t)} - \mathbf{W}_j^* \rangle \\
&= \sum_{j=1}^{\mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} \ell_{j,i}^{'(t)} \langle \nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{W}_j^{(t)} - \mathbf{W}_j^* \rangle \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{\mathcal{T}} \ell_{j,i}^{'(t)} \left( F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) - A_{j,i} \right) \\
&\geq \frac{1}{n} \sum_{i=1}^{n} \left( \ell(\mathbf{W}^{(t)}, \mathbf{x}_i, y_i) - \log \left( 1 + \sum_{j \neq y_i} \exp\{A_{j,i} - A_{y_i,i}\} \right) \right) \\
&\geq L_S(\mathbf{W}^{(t)}) - 0.5\epsilon. \qquad (42)
\end{aligned}$$

To complete the proof, we use the weight update rule $\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \nabla L_S(\mathbf{W}^{(t)})$ and the algebraic identity for the difference of squared vector norms, which states that $\|\mathbf{a} - \mathbf{b}\|^2 - \|\mathbf{a} - \eta\mathbf{g} - \mathbf{b}\|^2 = 2\eta\langle \mathbf{g}, \mathbf{a} - \mathbf{b} \rangle - \eta^2 \|\mathbf{g}\|^2$ for any vectors $\mathbf{a}, \mathbf{b}, \mathbf{g}$ and scalar $\eta$. Applying this identity with $\mathbf{a} = \mathbf{W}^{(t)}$, $\mathbf{b} = \mathbf{W}^*$, and $\mathbf{g} = \nabla L_S(\mathbf{W}^{(t)})$, we rewrite the difference of Frobenius norms as:

$$\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 = 2\eta \langle \nabla L_S(\mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \rangle - \eta^2 \|\nabla L_S(\mathbf{W}^{(t)})\|_F^2.$$

Substituting the lower bound from (42) and the gradient bound from Lemma F.2—which asserts $\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \leq O\left(\mathcal{T}^2 \max\{\sigma_p^2 d, \mu^2\}\right) L_S(\mathbf{W}^{(t)})$—into this expression, we obtain:

$$\begin{aligned}
&\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \\
&\geq 2\eta \left( L_S(\mathbf{W}^{(t)}) - 0.5\epsilon \right) - \eta^2 O\left(\mathcal{T}^2 \max\{\sigma_p^2 d, \mu^2\}\right) L_S(\mathbf{W}^{(t)}).
\end{aligned}$$

Condition 4.1 ensures that the step size $\eta$ is sufficiently small such that $\eta O\left(\mathcal{T}^2 \max\{\sigma_p^2 d, \mu^2\}\right) \leq 1$. This guarantee allows us to simplify the right-hand side by bounding the coefficient of $L_S(\mathbf{W}^{(t)})$ below by $\eta$, leading to the final result:

$$\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \geq \eta L_S(\mathbf{W}^{(t)}) - \eta\epsilon.$$

This completes the proof. □

**Lemma F.4.** *Under Condition 4.1, we have that* $\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F \leq \widetilde{O}(\sqrt{\mathcal{T}mn}\sigma_p^{-1} d^{-1/2})$.

*Proof of Lemma F.4.* We have

$$\begin{aligned}\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F &\le \|\mathbf{W}^{(T_1)} - \mathbf{W}^{(0)}\|_F + \|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F \\ &= \sqrt{\sum_{j,r} \|\mathbf{w}_{j,r}^{(T_1)} - \mathbf{w}_{j,r}^{(0)}\|_F^2} + \sqrt{\sum_{j,r} \|\mathbf{w}_{j,r}^{*} - \mathbf{w}_{j,r}^{(0)}\|_F^2} \\ &\le O(\sqrt{\mathcal{T}m}) \max_{j,r} \|\sum_{\tau=1}^{\mathcal{T}} \gamma_{j,r,\tau}^{(T_1)} \cdot \mu^{-2}\boldsymbol{\mu}_\tau\|_2 + O(\sqrt{\mathcal{T}m}) \max_{j,r} \|\sum_{i=1}^{n} \rho_{j,r,i}^{(T_1)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2}\boldsymbol{\xi}_i\|_2 + \widetilde{O}(\sqrt{\mathcal{T}m}\mu^{-1}) \\ &\le \widetilde{O}(\sqrt{\mathcal{T}m}\mu^{-1}) + \widetilde{O}(\sqrt{\mathcal{T}mn}\sigma_p^{-1}d^{-1/2}) \\ &\le \widetilde{O}(\sqrt{\mathcal{T}mn}\sigma_p^{-1}d^{-1/2}),\end{aligned}$$

where the first inequality is by triangle inequality, the second inequality is by our decomposition of $\mathbf{W}^{(T_1)}$ and $\mathbf{W}^*$, the third inequality is by Lemma C.4 and Proposition E.1, and the last inequality follows by (17). Here the proof completes. □

**Lemma F.5.** *Under Condition 4.1, let* $T^* = T_1 + \left\lfloor \frac{\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F^2}{\eta\epsilon} \right\rfloor = T_1 + \widetilde{O}\left(\eta^{-1}\mathcal{T}mn\sigma_p^{-2}d^{-1}\epsilon^{-1}\right)$. *Then the following statements hold:*

1. *For all* $j \neq \tau$ *and* $T_1 \le t \le T^*$, $\max_{j,r,i} |\rho_{j,r,i}^{(t)}| \le \widetilde{O}(1)$ *and* $\max_{j,r,\tau} |\gamma_{j,r,\tau}^{(t)}| \le \widetilde{O}(1)$;
2. *For all* $T_1 \le t \le T^*$, *the average training loss satisfies:*

$$\frac{1}{t - T_1 + 1} \sum_{t'=T_1}^{t} L_S(\mathbf{W}^{(t')}) \le \epsilon + \frac{\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F^2}{(t - T_1 + 1)\eta};$$

3. *There exists an iterate* $\mathbf{W}^{(t)}$ *with* $T_1 \le t \le T^*$ *such that* $L_S(\mathbf{W}^{(t)}) < 2\epsilon$.

*Proof of Lemma F.5.* We begin by analyzing the boundedness of $\rho_{j,r,i}^{(t)}$ and $\gamma_{j,r,\tau}^{(t)}$. From Lemma F.1, we know that for all $j \neq \tau$ and $0 \le t \le T^*$:

$$\begin{aligned}0 \le |\gamma_{j,r,\tau}^{(t)}| &\le O\left(\sqrt{\log\left(mn\mathcal{T}^2/\delta\right)}\mathcal{T}n^{1/4}\sigma_0\mu\right), \\ 0 \le |\underline{\rho}_{j,r,i}^{(t)}| &\le \max\left\{O\left(\sqrt{\log\left(mn\mathcal{T}^2/\delta\right)}\mathcal{T}n^{1/4}\sigma_0\mu\right), O\left(\sqrt{\log(n^2/\delta)}\frac{n\log(T^*)}{\sqrt{d-\mathcal{T}}}\right)\right\}.\end{aligned}$$

Condition 4.1 ensures that the right-hand sides of these inequalities are bounded by $\widetilde{O}(1)$, so we immediately obtain $\max_{j,r,i} |\underline{\rho}_{j,r,i}^{(t)}| \le \widetilde{O}(1)$ and $\max_{j,r,\tau} |\gamma_{j,r,\tau}^{(t)}| \le \widetilde{O}(1)$ for all $j \neq \tau$ and $T_1 \le t \le T^*$.

Next, we derive the average training loss bound. By Lemma F.3, the difference of squared Frobenius norms satisfies:

$$\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \ge \eta L_S(\mathbf{W}^{(t)}) - \eta\epsilon$$

for all $T_1 \le t \le T^*$. Rearranging this inequality isolates the training loss:

$$L_S(\mathbf{W}^{(t)}) \le \epsilon + \frac{\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2}{\eta}.$$

Summing both sides from $t' = T_1$ to $t' = t$ (where $T_1 \le t \le T^*$) yields a telescoping sum on the right-hand side:

$$\sum_{t'=T_1}^{t} L_S(\mathbf{W}^{(t')}) \le \epsilon(t - T_1 + 1) + \frac{1}{\eta}\left(\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2\right).$$

Since the squared Frobenius norm is non-negative, $\|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \geq 0$, so we can drop this term to get a simpler upper bound:

$$\sum_{t'=T_1}^{t} L_S(\mathbf{W}^{(t')}) \leq \epsilon(t - T_1 + 1) + \frac{\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F^2}{\eta}.$$

Dividing both sides by $t - T_1 + 1$ gives the average loss bound:

$$\frac{1}{t - T_1 + 1} \sum_{t'=T_1}^{t} L_S(\mathbf{W}^{(t')}) \leq \epsilon + \frac{\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F^2}{(t - T_1 + 1)\eta}.$$

Using Lemma F.4, which bounds $\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F^2 \leq \widetilde{O}(\eta^{-1}\mathcal{T}mn\sigma_p^{-2}d^{-1})$, we can rewrite this as:

$$\frac{1}{t - T_1 + 1} \sum_{t'=T_1}^{t} L_S(\mathbf{W}^{(t')}) \leq \epsilon + \frac{\widetilde{O}(\eta^{-1}\mathcal{T}mn\sigma_p^{-2}d^{-1})}{t - T_1 + 1}.$$

We now show the existence of an iterate with training loss less than $2\epsilon$. By the definition of $T^* = T_1 + \widetilde{O}\left(\eta^{-1}\mathcal{T}mn\sigma_p^{-2}d^{-1}\epsilon^{-1}\right)$, substituting $t = T^*$ into the average loss bound gives:

$$\frac{1}{T^* - T_1 + 1} \sum_{t'=T_1}^{T^*} L_S(\mathbf{W}^{(t')}) \leq 2\epsilon. \tag{43}$$

If the average of a set of non-negative values is bounded by $2\epsilon$, at least one value in the set must be bounded by $2\epsilon$. Thus, there exists some $t$ with $T_1 \leq t \leq T^*$ such that $L_S(\mathbf{W}^{(t)}) \leq 2\epsilon$.

Finally, we confirm the boundedness of $\bar{\rho}_{y_i,r,i}^{(t)}$. Using the update rule for $\bar{\rho}_{y_i,r,i}^{(t)}$ from (12) and applying Lemma C.4 (which gives $\|\boldsymbol{\xi}_i\|_2^2 \leq 2\sigma_p^2 d$) and $\sigma'(\cdot) \leq 1$, we get:

$$\bar{\rho}_{y_i,r,i}^{(t+1)} \leq \bar{\rho}_{y_i,r,i}^{(t)} + \frac{2\eta\sigma_p^2 d}{nm}(-\ell_{y_i,i}^{'(t)}).$$

Summing this recurrence from $s = T_1$ to $s = t - 1$ (with $\bar{\rho}_{y_i,r,i}^{(T_1)} = \widetilde{O}(1)$ from Lemma F.1) yields:

$$\bar{\rho}_{y_i,r,i}^{(t)} \leq \bar{\rho}_{y_i,r,i}^{(T_1)} + \frac{2\eta\sigma_p^2 d}{nm} \sum_{s=T_1}^{t-1} (-\ell_{y_i,i}^{'(s)}).$$

The inequality $\frac{1}{x+1} \leq \log\left(1 + \frac{1}{x}\right)$ for all $x > 0$ implies $-\ell_{y_i,i}^{'(s)} \leq \ell^{(s)}(F(\mathbf{x}_i, y_i))$, so:

$$\bar{\rho}_{y_i,r,i}^{(t)} \leq \widetilde{O}(1) + \frac{2\eta\sigma_p^2 d}{nm} \sum_{s=T_1}^{t-1} \ell^{(s)}(F(\mathbf{x}_i, y_i)).$$

Since $\ell^{(s)}(F(\mathbf{x}_i, y_i)) \leq O(nL_S(\mathbf{W}^{(s)}))$ (by the definition of $L_S$ as the average loss over $n$ samples), we use (43) to bound the sum:

$$\sum_{s=T_1}^{t-1} L_S(\mathbf{W}^{(s)}) \leq 2\epsilon(t - T_1) \leq \widetilde{O}\left(\eta^{-1}\mathcal{T}mn\sigma_p^{-2}d^{-1}\right).$$

Substituting this into the expression for $\bar{\rho}_{y_i,r,i}^{(t)}$ and using Proposition E.1 to simplify the constants, we find $\bar{\rho}_{y_i,r,i}^{(t)} \leq \widetilde{O}(1)$ for all $T_1 \leq t \leq T^*$.

Combining these results confirms all claims of the lemma. This completes the proof. □

### F.3. Population Loss

Consider a new data point $(\mathbf{x}, y)$ drawn from the previously defined distribution. Without loss of generality, we assume the first patch corresponds to the signal and the second to the noise, i.e., $\mathbf{x} = [\boldsymbol{\mu}_y, \boldsymbol{\xi}]$. By the signal-noise decomposition, the parameters of the learned neural network satisfy:

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + \sum_{\tau=1}^{\mathcal{T}} \gamma_{j,r,\tau}^{(t)} \cdot \mu^{-2} \boldsymbol{\mu}_\tau + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \boldsymbol{\xi}_i$$

for all $j \in [\mathcal{T}]$ and $r \in [m]$.

**Lemma F.6.** *Under Condition 4.1,* $\max_{j,r,i} |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle| \leq \widetilde{O}(1)$ *holds for all* $0 \leq t \leq T^*$.

*Proof of Lemma F.6.* We derive an upper bound for the inner product between the network parameter and the noise vector by expanding the expression and applying key inequalities:

$$\begin{aligned} |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle| &= \left| \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle + \rho_{j,r,i}^{(t)} + \sum_{i' \neq i} \rho_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle \right| \\ &\leq \left| \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle \right| + |\rho_{j,r,i}^{(t)}| + \sum_{i' \neq i} |\rho_{j,r,i'}^{(t)}| \cdot \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle|. \end{aligned}$$

By the definition of $\beta = \max_{j,r,i} |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle|$, Lemma C.4 (which bounds $\sum_{i' \neq i} \|\boldsymbol{\xi}_{i'}\|_2^{-2} |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle| \leq 4n\alpha \sqrt{\log(4n^2/\delta)/(d - \mathcal{T})}$), and the boundedness of $\rho_{j,r,i}^{(t)} \leq \widetilde{O}(1)$ from prior results, this simplifies to:

$$|\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle| \leq \beta + \widetilde{O}(1) + 4n\alpha \sqrt{\frac{\log(4n^2/\delta)}{d - \mathcal{T}}}.$$

Condition 4.1 ensures all terms on the right-hand side are bounded by $\widetilde{O}(1)$, so we conclude $\max_{j,r,i} |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle| \leq \widetilde{O}(1)$. □

**Lemma F.7.** *Under Condition 4.1, with probability at least* $1 - 2\mathcal{T}mT^* \cdot \exp\left(-C_2^{-1} \sigma_0^{-2} \sigma_p^{-2} n^{-1} d^{-1}\right)$, *we have* $\max_{j,r} |\langle \widetilde{\mathbf{w}}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle| \leq 1/2$ *for all* $0 \leq t \leq T^*$, *where* $C_2 = \widetilde{O}(1)$.

*Proof of Lemma F.7.* Define the noise-projected parameter $\widetilde{\mathbf{w}}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(t)} - \sum_{\tau=1}^{\mathcal{T}} \gamma_{j,r,\tau}^{(t)} \cdot \mu^{-2} \boldsymbol{\mu}_\tau$. Due to the orthogonality between the signal vectors $\{\boldsymbol{\mu}_\tau\}$ and the noise vector $\boldsymbol{\xi}$, the inner product $\langle \widetilde{\mathbf{w}}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle$ equals $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle$ (the signal component vanishes).

We first bound the Euclidean norm of $\widetilde{\mathbf{w}}_{j,r}^{(t)}$. By the signal-noise decomposition and Condition 4.1, which relates SNR to $\mu$ and $\sigma_p$, we have:

$$\|\widetilde{\mathbf{w}}_{j,r}^{(t)}\|_2 \leq \widetilde{O}\left(\sigma_0 \sqrt{d} + \sqrt{n} \mathrm{SNR} \mu^{-1}\right) = \widetilde{O}(\sigma_0 \sqrt{nd}), \tag{44}$$

where the equality follows from $d$ in Condition 4.1. Let $C_1 = \widetilde{O}(1)$ denote the constant hidden in the $\widetilde{O}(\cdot)$ notation, so $\max_{j,r} \|\widetilde{\mathbf{w}}_{j,r}^{(t)}\|_2 \leq C_1 \sigma_0 \sqrt{nd}$.

Since $\boldsymbol{\xi}$ is a Gaussian random vector with i.i.d. components of variance $\sigma_p^2$, the inner product $\langle \widetilde{\mathbf{w}}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle$ follows a zero-mean Gaussian distribution. Its standard deviation is bounded by $\|\widetilde{\mathbf{w}}_{j,r}^{(t)}\|_2 \cdot \sigma_p \leq C_1 \sigma_0 \sigma_p \sqrt{nd}$.

Applying the Gaussian tail bound (for a zero-mean Gaussian $Z$ with standard deviation $\sigma$, $\mathbb{P}(|Z| \geq a) \leq 2\exp(-a^2/(2\sigma^2))$) with $a = 1/2$, we get:

$$\mathbb{P}\left(|\langle \widetilde{\mathbf{w}}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle| \geq 1/2\right) \leq 2\exp\left(-\frac{1}{8 C_1^2 \sigma_0^2 \sigma_p^2 nd}\right). \tag{45}$$

To account for all parameters and iterations, we apply a union bound over $j \in [\mathcal{T}]$, $r \in [m]$, and $t \in [0, T^*]$. This gives the total failure probability bound, and the result follows by setting $C_2 = 8C_1^2 = \widetilde{O}(1)$. □

**Lemma F.8.** *Under Condition 4.1, for any $0 \le t \le T^*$, it holds that $L_{\mathcal{D}}(\mathbf{W}^{(t)}) \le 8e^4 \mathcal{T} L_S(\mathbf{W}^{(t)}) + \exp(-n^2)$.*

*Proof of Lemma F.8.* Let event $\mathcal{E}$ to be the event that Lemma F.7 holds. Then we can divide $L_{\mathcal{D}}(\mathbf{W}^{(t)})$ into two parts:

$$\mathbb{E}[\ell(F(\mathbf{W}, \mathbf{x}))] = \underbrace{\mathbb{E}[\mathbb{I}(\mathcal{E})\ell(F(\mathbf{W}, \mathbf{x}))]}_{I_1} + \underbrace{\mathbb{E}[\mathbb{I}(\mathcal{E}^c)\ell(F(\mathbf{W}, \mathbf{x}))]}_{I_2}$$

In the following, we bound $I_1$ and $I_2$ respectively.

Bounding $I_1$: Since $L_S(\mathbf{W}^{(t)}) = \frac{1}{n}\sum_{i=1}^n \ell(F(\mathbf{x}_i, y_i))$, we have

$$L_S(\mathbf{W}^{(t)}) = \frac{1}{n}\sum_{i=1}^n \ell(F(\mathbf{x}_i, y_i)) \ge \frac{1}{n}\sum_{y_i = y} \ell(F(\mathbf{x}_i, y_i)) \ge \frac{4}{5\mathcal{T}} \min_{y_i = y} \ell(F(\mathbf{x}_i, y_i)), \tag{46}$$

so there must exist an $i$ that $\ell(F(\mathbf{x}_i, y_i)) \le 2\mathcal{T} L_S(\mathbf{W}^{(t)})$ and $y_i = y$. Recall that $F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i) = \frac{1}{m}\sum_{r=1}^m [\sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle) + \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\mu}_{y_i}\rangle)]$, then we have

$$F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}) = \frac{1}{m}\sum_{r=1}^m [\sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}\rangle) + \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\mu}_{y_i}\rangle)]$$

and

$$\begin{aligned}
|F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i) - F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}) - \frac{1}{m}\sum_{r=1}^m \bar{\rho}_{y_i,r,y_i}^{(t)}| &= |\frac{1}{m}\sum_{r=1}^m \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle) - \frac{1}{m}\sum_{r=1}^m \bar{\rho}_{y_i,r,y_i}^{(t)} - \frac{1}{m}\sum_{r=1}^m \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}\rangle)| \\
&\le |\frac{1}{m}\sum_{r=1}^m \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle) - \frac{1}{m}\sum_{r=1}^m \bar{\rho}_{y_i,r,y_i}^{(t)}| + |\frac{1}{m}\sum_{r=1}^m \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}\rangle)| \\
&\le 1 + 1 \\
&= 2,
\end{aligned} \tag{47}$$

where the first inequality is by triangle inequality and the last inequality is by Lemma E.5 and the definition of event $\mathcal{E}$, so we have

$$|F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i) - F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x})| \le |F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i) - F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}) - \frac{1}{m}\sum_{r=1}^m \bar{\rho}_{y_i,r,y_i}^{(t)}| + \frac{1}{m}\sum_{r=1}^m \bar{\rho}_{y_i,r,y_i}^{(t)} \tag{48}$$

$$\le 2 + \widetilde{\rho}\frac{1}{m}\sum_{r=1}^m \gamma_{y_i,r,y_i}^{(t)} \tag{49}$$

$$\le 2 + \widetilde{O}(\widetilde{\rho}) \le 3, \tag{50}$$

where the first inequality is by triangle inequality, the second inequality is by (47) and Proposition E.1, and the last inequality is by (17). Besides, we have that

$$F_j(\mathbf{W}_j^{(t)}, \mathbf{x}) = \frac{1}{m}\sum_{r=1}^m [\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle) + \sigma(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{y_i}\rangle)] \le 0.5 + 0.5 = 1, \tag{51}$$

where the inequality is by the definition of event $\mathcal{E}$. Then we have

$$\begin{aligned}
\ell(F(\mathbf{W},\mathbf{x})) &= \log(1+\frac{\sum_{j\neq y_i}\exp\{F_j(\mathbf{W}_j^{(t)},\mathbf{x})\}}{\exp\{F_{y_i}(\mathbf{W}_{y_i}^{(t)},\mathbf{x})\}}) \\
&\leq \log(1+\frac{\mathcal{T}e}{\exp\{F_{y_i}(\mathbf{W}_{y_i}^{(t)},\mathbf{x})\}}) \\
&\leq \log(1+\frac{\mathcal{T}e^4}{\exp\{F_{y_i}(\mathbf{W}_{y_i}^{(t)},\mathbf{x}_i)\}}) \\
&\leq 2e^4\log(1+\frac{\mathcal{T}-1}{\exp\{F_{y_i}(\mathbf{W}_{y_i}^{(t)},\mathbf{x}_i)\}}) \\
&\leq 2e^4\ell(F(\mathbf{x}_i,y_i)) \\
&\leq 8e^4\mathcal{T}L_S(\mathbf{W}^{(t)}),
\end{aligned}$$

where the first inequality is by 51, the second inequality is by 50, and the last inequality is due to the definition of $L_S(\mathbf{W}^{(t)})$. So we can get that

$$I_1 = \mathbb{E}[\mathbb{I}(\mathcal{E})\ell(F(\mathbf{W},\mathbf{x}))] \leq 8e^4\mathcal{T}L_S(\mathbf{W}^{(t)}).$$

Bounding $I_2$: Next we bound the second term $I_2$. We choose an arbitrary training data $(\mathbf{x}_{i'},y_{i'})$ such that $y_{i'}=y$. Then we have

$$\begin{aligned}
\ell(F(\mathbf{W},\mathbf{x})) &= \log(1+\frac{\sum_{j\neq y}\exp\{F_j(\mathbf{W}_j^{(t)},\mathbf{x})\}}{\exp\{F_y(\mathbf{W}_y^{(t)},\mathbf{x})\}}) \\
&\leq \log(1+\sum_{j\neq y}\exp\{F_j(\mathbf{W}_j^{(t)},\mathbf{x})\}) \\
&\leq \sum_{j\neq y}\log(1+\exp\{F_j(\mathbf{W}_j^{(t)},\mathbf{x})\}) \\
&\leq \mathcal{T}+\sum_{j\neq y}F_j(\mathbf{W}_j^{(t)},\mathbf{x}) \\
&\leq \mathcal{T}+\sum_{j\neq y}\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}\rangle)+\sigma(\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}_y\rangle)] \\
&\leq 2\mathcal{T}+\mathcal{T}\cdot\widetilde{O}(\sigma_0\sqrt{nd})\|\boldsymbol{\xi}\|_2,
\end{aligned}$$

where the second inequality is by Bernoulli's Inequality, the third inequality is by $\log(1+e^x)\leq x+1$ for any $x\geq 0$, and the last inequality is by 44 and Lemma C.5. Then we have

$$\begin{aligned}
I_2 &\leq \sqrt{\mathbb{E}[\mathbb{I}(\mathcal{E}^c)]}\cdot\sqrt{\mathbb{E}[\ell(F(\mathbf{W},\mathbf{x}))^2]} \\
&\leq \sqrt{\mathbb{P}(\mathcal{E}^c)}\cdot\sqrt{4\mathcal{T}^2+\mathcal{T}^2\widetilde{O}(\sigma_0^2nd)\mathbb{E}[\|\boldsymbol{\xi}\|_2^2]} \\
&\leq \exp[\widetilde{O}(1)-\widetilde{O}(\sigma_0^2\sigma_p^2nd)^{-1}+\mathrm{polylog}(d)] \\
&\leq \exp(-n^2),
\end{aligned}$$

where the first inequality is by Cauchy–Schwarz Inequality, and the last inequality is by the $\sigma_0$'s condition in Condition 4.1. Here the proof completes. □

## G. Noise Memorization

**Proposition G.1.** *For $0 \leq t \leq T^*$, we have that*

$$\gamma_{j,r,\tau}^{(0)}, \bar{\rho}_{j,r,i}^{(0)}, \underline{\rho}_{j,r,i}^{(0)} = 0,$$
$$0 \leq \bar{\rho}_{j,r,i}^{(t)} \leq \alpha, \tag{52}$$
$$0 \geq \gamma_{j,r,\tau}^{(t)} \geq -\alpha, j \neq \tau, \tag{53}$$
$$0 \geq \underline{\rho}_{j,r,i}^{(t)} \geq -\alpha, \tag{54}$$

*and there exists a positive constant $C'$ such that*

$$0 \leq \gamma_{\tau,r,\tau}^{(t)} \leq C'\widetilde{\gamma}\alpha, \tag{55}$$

*for all $r \in [m]$, $j \in [\mathcal{T}]$ and $i \in [n]$, where $\widetilde{\gamma} := \mathcal{T}^{-1} n \cdot \mathrm{SNR}^2$.*

*Proof of Proposition G.1.* The proof methods for (52), (53), (54) and (55) are analogous to that of Proposition E.1. □

Here, Lemma E.2, Lemma E.3, Lemma E.4, Lemma E.5, Lemma E.6 and Lemma E.8 still hold. We will prove the following lemma:

**Lemma G.2.** *Under Condition 4.1, assume that (52), (53), (54), and (55) hold at any iteration $t' \leq t$. If*

$$\mathrm{SNR} \leq \Theta(n^{-1}) \tag{56}$$

*holds, then the following conditions hold for any iteration $t' \leq t$:*

1. $|\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t')} - \bar{\rho}_{y_{i'},r,i'}^{(t')}]| \leq \varrho$ *for all* $i, i' \in [\mathcal{T}]$.
2. $|F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - F_{y_j}(\mathbf{W}_{y_j}^{(t')}, \mathbf{x}_j)| \leq C_1$ *for all* $i, j \in [n]$.
3. $\ell_{y_i,i}^{'(t')} / \ell_{y_j,j}^{'(t')} \leq \exp\{1 + C_1\} = C_2$ *for all* $i, j \in [n]$.

*Proof of Lemma G.2.* We will use induction to prove Lemma G.2. When $t' = 0$, we have that $|\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t')} - \bar{\rho}_{y_{i'},r,i'}^{(t')}]| = 0 \leq \varrho$. As we know, $F_{y_i}(\mathbf{W}^{(t)}, \mathbf{x}_i) = \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle) + \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle)]$, so we have

$$\begin{aligned}
|F_{y_i}(\mathbf{W}_{y_i}^{(0)}, \mathbf{x}_i)| &= |\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle) + \sigma(\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\mu}_{y_i} \rangle)]| \\
&\leq \frac{1}{m}\sum_{r=1}^{m}[|\sigma(\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle)| + |\sigma(\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\mu}_{y_i} \rangle)|] \\
&\leq \frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(0)} + 1 + \gamma_{y_i,r,y_i}^{(0)} + 1] \\
&\leq 2,
\end{aligned} \tag{57}$$

where the fist inequality is by absolute value inequality, the second inequality is by Lemma E.5. Then we can get that $|F_{y_i}(\mathbf{W}_{y_i}^{(0)}, \mathbf{x}_i) - F_{y_j}(\mathbf{W}_{y_j}^{(0)}, \mathbf{x}_j)| \leq 4$. We also have that $\ell_{y_i,i}^{'(0)} = 1 - \mathbf{logit}_i(F, \mathbf{x}_i)$. By Lemma E.3 and (57), we have $-2 \leq F_j(\mathbf{W}_j^{(0)}, \mathbf{x}_i) - F_{y_i}(\mathbf{W}_{y_i}^{(0)}, \mathbf{x}_i) \leq 1$. So $1 - 1/(1 + (\mathcal{T} - 1)e^{-2}) \leq \ell_{y_i,i}^{'(0)} \leq 1 - 1/(1 + (\mathcal{T} - 1)e)$. We have verified the correctness of the three formulas at $t = 0$.

Now suppose there exists $\widetilde{t} \leq t$ such that these five conditions hold for any $0 \leq t' \leq \widetilde{t} - 1$. We aim to prove that these conditions also hold for $t' = \widetilde{t}$.

Firstly we will show that $F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - F_{y_j}(\mathbf{W}_{y_j}^{(t')}, \mathbf{x}_j)$ is dominated by $\frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t')} - \bar{\rho}_{y_j,r,j}^{(t')}]$. We have that

$$
\begin{aligned}
&F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - F_{y_j}(\mathbf{W}_{y_j}^{(t')}, \mathbf{x}_j) \\
&= \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) + \sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\mu}_{y_i}\rangle)] - \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\xi}_j\rangle) + \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\mu}_{y_j}\rangle)] \\
&= \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) - \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\xi}_j\rangle)] + \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\mu}_{y_i}\rangle) - \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\mu}_{y_j}\rangle)].
\end{aligned}
$$

By Lemma E.5, we have that

$$
\begin{aligned}
|\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\mu}_{y_i}\rangle) - \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\mu}_{y_j}\rangle)]| &\leq \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\mu}_{y_i}\rangle) + \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\mu}_{y_j}\rangle)] \\
&\leq \frac{1}{m}\sum_{r=1}^{m}[\gamma_{y_i,r,y_i}^{(t)} + \gamma_{y_j,r,y_j}^{(t)} + 2] \\
&\leq \frac{1}{m}\sum_{r=1}^{m}[2C'\widetilde{\gamma}\alpha + 2] \\
&\leq 4, \qquad (58)
\end{aligned}
$$

where the third inequality is by (55), and the last inequality is by (56).

By Lemma E.5, we also have

$$
\begin{aligned}
&|\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) - \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\xi}_j\rangle)] - \sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t')} - \bar{\rho}_{y_j,r,j}^{(t')}]| \\
&= |\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) - \bar{\rho}_{y_i,r,i}^{(t')}] - \sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\xi}_j\rangle) - \bar{\rho}_{y_j,r,j}^{(t')}]| \\
&\leq \sum_{r=1}^{m}|[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) - \bar{\rho}_{y_i,r,i}^{(t')}]| + \sum_{r=1}^{m}|\sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\xi}_j\rangle) - \bar{\rho}_{y_j,r,j}^{(t')}| \\
&\leq 2m. \qquad (59)
\end{aligned}
$$

So we have

$$
\begin{aligned}
&|F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - F_{y_j}(\mathbf{W}_{y_j}^{(t')}, \mathbf{x}_j) - \frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t')} - \bar{\rho}_{y_j,r,j}^{(t')}]| \\
&= |F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - F_{y_j}(\mathbf{W}_{y_j}^{(t')}, \mathbf{x}_j) - \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) - \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\xi}_j\rangle)] \\
&+ \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) - \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\xi}_j\rangle)] - \frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t')} - \bar{\rho}_{y_j,r,j}^{(t')}]| \\
&\leq |F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - F_{y_j}(\mathbf{W}_{y_j}^{(t')}, \mathbf{x}_j) - \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) - \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\xi}_j\rangle)]| \\
&+ |\frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) - \sigma(\langle \mathbf{w}_{y_j,r}^{(t')}, \boldsymbol{\xi}_j\rangle)] - \frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t')} - \bar{\rho}_{y_j,r,j}^{(t')}]| \\
&\leq 4 + 2 = 6,
\end{aligned}
$$

where the last inequality is by (58) and (59).

By the same analysis, we can also get that

$$|F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - \frac{1}{m}\sum_{r=1}^{m}\bar{\rho}_{y_i,r,i}^{(t')}| \leq 3. \tag{60}$$

By the update rule (12), we have that

$$\begin{aligned}\frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t')} - \bar{\rho}_{y_j,r,j}^{(t')}] &= \frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t'-1)} - \bar{\rho}_{y_j,r,j}^{(t'-1)}] \\ &+ \frac{\eta}{nm^2}|N_{y_i,i}|(-\ell_{y_i,i}^{'(t'-1)})\|\boldsymbol{\xi}_i\|_2^2 - \frac{\eta}{nm^2}|N_{y_j,j}|(-\ell_{y_j,j}^{'(t'-1)})\|\boldsymbol{\xi}_j\|_2^2.\end{aligned}$$

If $\frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t'-1)} - \bar{\rho}_{y_j,r,j}^{(t'-1)}] \leq 0.9\varrho$, we have that

$$\begin{aligned}\frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t')} - \bar{\rho}_{y_j,r,j}^{(t')}] &\leq \frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t'-1)} - \bar{\rho}_{y_j,r,j}^{(t'-1)}] + \frac{\eta}{nm^2}|N_{y_i,i}|(-\ell_{y_i,i}^{'(t'-1)})\|\boldsymbol{\xi}_i\|_2^2 \\ &\leq 0.9\varrho + \frac{\eta}{nm^2}\cdot\frac{3m}{4}\cdot 2\sigma_p^2 d \\ &\leq \varrho,\end{aligned}$$

where the last inequality is by Condition 4.1.

If $\frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t'-1)} - \bar{\rho}_{y_j,r,j}^{(t'-1)}] \geq 0.9\varrho$, we have that

$$F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - F_{y_j}(\mathbf{W}_{y_j}^{(t')}, \mathbf{x}_j) \geq \frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t'-1)} - \bar{\rho}_{y_j,r,j}^{(t'-1)}] - 6 \geq 0.9\varrho - 6,$$

so we have

$$\begin{aligned}&\ell_{p,i}^{'(t'-1)}/\ell_{j,q}^{'(t'-1)} \\ &= \frac{\sum_{j\neq p}\exp\{F_j(\mathbf{W}_j^{(t'-1)}, \mathbf{x}_i)\}}{\sum_{j\neq p}\exp\{F_j(\mathbf{W}_j^{(t'-1)}, \mathbf{x}_i)\} + \exp\{F_p(\mathbf{W}_p^{(t'-1)}, \mathbf{x}_i)\}} \cdot \frac{\sum_{l\neq j}\exp\{F_l(\mathbf{W}_l^{(t'-1)}, \mathbf{x}_q)\} + \exp\{F_j(\mathbf{W}_j^{(t'-1)}, \mathbf{x}_q)\}}{\sum_{l\neq j}\exp\{F_l(\mathbf{W}_l^{(t'-1)}, \mathbf{x}_q)\}} \\ &\geq \frac{(\mathcal{T}-1)}{(\mathcal{T}-1) + \exp\{\frac{1}{m}\sum_{r=1}^{m}\bar{\rho}_{p,r,i}^{(t'-1)} + 3\}} \cdot \frac{(\mathcal{T}-1)e + \exp\{\frac{1}{m}\sum_{r=1}^{m}\bar{\rho}_{j,r,q}^{(t'-1)} - 3\}}{(\mathcal{T}-1)e} \\ &\geq 0.5\exp\{0.9\varrho - 12\},\end{aligned}$$

So we have that

$$\begin{aligned}\frac{\eta}{nm^2}|N_{y_i,i}|(-\ell_{y_i,i}^{'(t'-1)})\|\boldsymbol{\xi}_i\|_2^2 &\leq \frac{\eta}{nm^2}m(-\ell_{y_i,i}^{'(t'-1)})\cdot 2\sigma_p^2(d-\mathcal{T}), \\ \frac{\eta}{nm^2}|N_{y_j,j}|(-\ell_{y_j,j}^{'(t'-1)})\|\boldsymbol{\xi}_j\|_2^2 &\geq \frac{\eta}{nm^2}\frac{m}{4}0.5\exp\{0.9\varrho - 12\}(-\ell_{y_i,i}^{'(t'-1)})\frac{1}{2}\sigma_p^2(d-\mathcal{T})\end{aligned}$$

So if we choose $\varrho = 18$, then we have $\frac{\eta}{nm^2}|N_{y_i,i}|(-\ell_{y_i,i}^{'(t'-1)})\|\boldsymbol{\xi}_i\|_2^2 \leq \frac{\eta}{nm^2}|N_{y_j,j}|(-\ell_{y_j,j}^{'(t'-1)})\|\boldsymbol{\xi}_j\|_2^2$, then $\frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t')} - \bar{\rho}_{y_j,r,j}^{(t')}] \leq \frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t'-1)} - \bar{\rho}_{y_j,r,j}^{(t'-1)}] \leq \varrho$.

Then we have

$$|F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - F_{y_j}(\mathbf{W}_{y_j}^{(t')}, \mathbf{x}_j)| \leq |\frac{1}{m}\sum_{r=1}^{m}[\bar{\rho}_{y_i,r,i}^{(t')} - \bar{\rho}_{y_j,r,j}^{(t')}]| + 6 \leq \varrho + 6.$$

$$\ell_{y_i,i}^{'(t')}/\ell_{y_j,j}^{'(t')} \leq \exp\{\varrho + 7\}.$$

□

Then Lemma E.9 still holds.

# H. Decoupling with a Two-Stage Analysis

## H.1. First Stage

**Lemma H.1.** *If we denote*

$$\widetilde{\gamma} := \mathcal{T}^{-1} n \cdot \mathrm{SNR}^2,$$

then there exist

$$T_3 = C_7 mn\eta^{-1}\sigma_p^{-2} d^{-1}, T_4 = C_8 mn\eta^{-1}\sigma_p^{-2} d^{-1}$$

where $C_5 = \Theta(1)$ is a large constant and $C_6 = \Theta(1)$ is a small constant, such that

1. $\bar{\rho}_{y_i,r,i}^{(T_3)} \geq 3$ for any $r \in N_{y_i,i} = \{r \in [m] : \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle > 0\}$ and $i \in [n]$.
2. $\gamma_{\tau,r,\tau}^{(t)} = \Theta(\widetilde{\gamma})$ when $T_4 \leq t \leq T_3$ for any $r \in M_{\tau,\tau}$ and $\tau \in [\mathcal{T}]$.
3. $0 \leq |\gamma_{j,r,\tau}^{(t)}| \leq O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T} n^{1/4}\sigma_0\mu)$ for all $j \neq \tau$ and $0 \leq t \leq T^*$.
4. $0 \leq |\underline{\rho}_{j,r,i}^{(t)}| \leq \max\{O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T} n^{1/4}\sigma_0\mu), O(\sqrt{\log(n^2/\delta)} n \log(T^*)/\sqrt{d-\mathcal{T}}\}$ for all $0 \leq t \leq T^*$.

*Proof of Lemma H.1.* First, we note that $\gamma_{j,r,\tau}^{(t)} (j \neq \tau)$ and $\underline{\rho}_{j,r,i}^{(t)}$ are non-increasing sequences. By the convergence theorem for sequences, (9) and (10), if $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_\tau \rangle (\tau \neq j)$ and $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$ are positive, then they converge to 0. So we have

$$\begin{aligned}
&0 \geq \gamma_{j,r,\tau}^{(t)} \geq -\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle \geq -\beta, \\
&0 \geq \underline{\rho}_{j,r,i}^{(t)} \geq -\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle - \sum_{i' \neq i} \rho_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle \\
&\geq -\beta - 4n\alpha\sqrt{\frac{\log(4n^2/\delta)}{d-\mathcal{T}}}.
\end{aligned}$$

By the definition of $\beta$ and Lemma C.5, we have that

$$\begin{aligned}
\beta &= \max_{j,r,\tau,i}\{|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle|, |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle|\} \\
&\leq \max\{\sigma_0\mu \cdot \sqrt{2\log(8m\mathcal{T}^2/\delta)}, 2\sigma_0\sigma_p\sqrt{d-\mathcal{T}} \cdot \sqrt{\log(8mn\mathcal{T}/\delta)}\} \\
&\leq O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T} n^{1/4}\sigma_0\mu).
\end{aligned}$$

So we have that

$$0 \leq |\gamma_{j,r,\tau}^{(t)}| \leq \beta \leq O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T} n^{1/4}\sigma_0\mu),$$

and

$$\begin{aligned}
0 \leq |\underline{\rho}_{j,r,i}^{(t)}| &\leq \beta + 4n\alpha\sqrt{\frac{\log(4n^2/\delta)}{d-\mathcal{T}}} \\
&\leq \max\{O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T} n^{1/4}\sigma_0\mu), O(\sqrt{\log(n^2/\delta)} n \log(T^*)/\sqrt{d-\mathcal{T}}\}.
\end{aligned}$$

By the update rule for $\gamma_{\tau,r,\tau}^{(t)}$, we have

$$\begin{aligned}
\gamma_{\tau,r,\tau}^{(t+1)} &= \gamma_{\tau,r,\tau}^{(t)} + \frac{\eta}{nm}\sum_{y_i=\tau}(-\ell_{\tau,i}^{'(t)})\sigma'(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau \rangle + \gamma_{\tau,r,\tau}^{(t)})\mu^2 \\
&\leq \gamma_{\tau,r,\tau}^{(t)} + \frac{2\eta\mu^2}{\mathcal{T}m},
\end{aligned}$$

where the inequality is by $-\ell_{y_i,i}^{'(t)} \leq 1$. $\sigma'(\cdot) \leq 1$ and Lemma C.1. Note that $\gamma_{\tau,r,\tau}^{(t)} = 0$ and recursively use the inequality $t$ times we have $\gamma_{\tau,r,\tau}^{(t)} \leq \frac{2\eta\mu^2}{\mathcal{T}m}t$.

Since $\mathcal{T}^{-1}n\mathrm{SNR}^2 = \widetilde{\gamma}$, we have that

$$T_3 = C_7 mn\eta^{-1}\sigma_p^{-2}d^{-1} = C_7 m\eta^{-1}\mu^{-2}\mathcal{T}\widetilde{\gamma},$$

and it follows that

$$\gamma_{\tau,r,\tau}^{(t)} \leq \frac{2\eta\mu^2}{\mathcal{T}m}t \leq 2C_7\widetilde{\gamma}$$

for all $t \leq T_3$.

For each $i$, denote by $T_3^{(i)}$ the last time in the period $[0, T_3]$ satisfying that $\max_r \bar{\rho}_{y_i,r,i}^{(t)} \leq 3$. Then for $0 \leq t \leq T_3$, $\max_{j,r,i} |\rho_{j,r,i}^{(t)}| = O(1)$ and $\max_{j,r,i} |\gamma_{j,r,i}^{(t)}| = O(1)$. Then we have that $F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) = O(1)$ for all $j \in [\mathcal{T}]$ and $i \in [n]$. Thus there exists a positive constant $C$ such that $-\ell_{y_i,i}^{'(t)} \geq C$ for $0 \leq t \leq T$. Then we have

$$\begin{aligned}
\bar{\rho}_{y_i,r,i}^{(t+1)} &= \bar{\rho}_{y_i,r,i}^{(t)} + \frac{\eta}{nm}(-\ell_{y_i,i}^{'(t)})\sigma'(\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \bar{\rho}_{y_i,r,i}^{(t)} + \sum_{i'\neq i}\rho_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\langle\boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2 \\
&\geq \bar{\rho}_{y_i,r,i}^{(t)} + \frac{\eta C\sigma_p^2 d}{2nm},
\end{aligned}$$

where the inequality is by Lemma C.4.

Therefore, $\bar{\rho}_{y_i,r,i}^{(t)} \geq \frac{\eta C\sigma_p^2 d}{2nm}t$ and $\bar{\rho}_{y_i,r,i}^{(t)}$ will reach 3 within

$$T_3 = C_7 mn\eta^{-1}\sigma_p^{-2}d^{-1}$$

iterations for any $r \in N_{y_i,i}$ and $i \in [n]$, where $C_7$ can be taken as $6/C$.

Next, we will discuss the lower bound of the growth of $\gamma_{\tau,r,\tau}^{(t)}$. For $\bar{\rho}_{y_i,r,i}^{(t)}$, we have that

$$\begin{aligned}
\bar{\rho}_{y_i,r,i}^{(t+1)} &= \bar{\rho}_{y_i,r,i}^{(t)} + \frac{\eta}{nm}(-\ell_{y_i,i}^{'(t)})\sigma'(\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \bar{\rho}_{y_i,r,i}^{(t)} + \sum_{i'\neq i}\rho_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\langle\boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2 \\
&\leq \bar{\rho}_{y_i,r,i}^{(t)} + \frac{2\eta\sigma_p^2 d}{nm},
\end{aligned}$$

where the inequality is by Lemma C.4 and $-\ell_{\tau,i}^{'(t)}, \sigma'(\cdot) \leq 1$.

So we have that $\bar{\rho}_{y_i,r,i}^{(t)} \leq \frac{2\eta\sigma_p^2 d}{nm}t$. Therefore, $\max_{i,r} \bar{\rho}_{y_i,r,i}^{(t)}$ will be smaller than 1 and $\max_{\tau,i} \gamma_{\tau,r,\tau}^{(t)} \leq O(\widetilde{\gamma}) = O(1)$ within

$$T_4 = C_8\eta^{-1}nm\sigma_p^{-2}d^{-1}$$

iterations, where $C_8$ can be taken as 0.5. Then we have that $F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) = O(1)$ for all $j \in [\mathcal{T}]$ and $i \in [n]$. Thus there exists a positive constant $C$ such that $-\ell_{\tau,i}^{'(t)} \geq C$ for $0 \leq t \leq T_2$. Then we have that

$$\begin{aligned}
\gamma_{\tau,r,\tau}^{(t+1)} &= \gamma_{\tau,r,\tau}^{(t)} + \frac{\eta}{nm}\sum_{y_i=\tau}(-\ell_{\tau,i}^{'(t)})\sigma'(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau\rangle + \gamma_{\tau,r,\tau}^{(t)})\mu^2 \\
&\geq \gamma_{\tau,r,\tau}^{(t)} + \frac{4\eta C\mu^2}{5\mathcal{T}m},
\end{aligned}$$

where the last inequality is by Lemma C.1. So that $\gamma_{\tau,r,\tau}^{(t)} \geq \frac{4\eta C\mu^2}{5\mathcal{T}m}t$ and $\gamma_{\tau,r,\tau}^{(T_4)} \geq 0.8CC_8 n\mathrm{SNR}^2/\mathcal{T} = \Theta(\widetilde{\gamma})$. Here the proof completes.

□

### H.2. Second Stage

By the signal-noise decomposition, at the end of the first stage, we have

$$\mathbf{w}_{j,r}^{(T_3)} = \mathbf{w}_{j,r}^{(0)} + \sum_{\tau=1}^{\mathcal{T}} \gamma_{j,r,\tau}^{(T_3)} \cdot \mu^{-2}\boldsymbol{\mu}_\tau + \sum_{i=1}^{n} \bar{\rho}_{j,r,i}^{(T_3)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2}\boldsymbol{\xi}_i + \sum_{i=1}^{n} \underline{\rho}_{j,r,i}^{(T_3)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2}\boldsymbol{\xi}_i$$

for $j \in [\mathcal{T}]$ and $r \in [m]$. By the results we get in the first stage, we know that at the beginning of this stage, we have the following property holds:

1. $\bar{\rho}_{y_i,r,i}^{(T_3)} \geq 3$ for any $r \in N_{y_i,i} = \{r \in [m] : \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle > 0\}$ and $i \in [n]$.
2. $\gamma_{\tau,r,\tau}^{(t)} = \Theta(\widetilde{\gamma})$ when $T_4 \leq t \leq T_3$ for any $r \in M_{\tau,\tau}$ and $\tau \in [\mathcal{T}]$.
3. $0 \leq |\gamma_{j,r,\tau}^{(t)}| \leq O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T}n^{1/4}\sigma_0\mu)$ for all $j \neq \tau$ and $0 \leq t \leq T^*$.
4. $0 \leq |\underline{\rho}_{j,r,i}^{(t)}| \leq \max\{O(\sqrt{\log(mn\mathcal{T}^2/\delta)}\mathcal{T}n^{1/4}\sigma_0\mu), O(\sqrt{\log(n^2/\delta)}n\log(T^*)/\sqrt{d-\mathcal{T}}\}$ for all $0 \leq t \leq T^*$.

Now we choose $\mathbf{W}^*$ as follows:

$$\mathbf{w}_{j,r}^* = \mathbf{w}_{j,r}^{(0)} + 10\log(2\mathcal{T}/\epsilon) \cdot [\sum_{i=1}^{n} \mathbb{1}(y_i = j) \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2}].$$

**Lemma H.2.** *Under Condition 4.1, the following inequality holds:*

$$\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \geq \eta L_S(\mathbf{W}^{(t)}) - \eta\epsilon$$

*for all $T_1 \leq t \leq T^*$.*

*Proof of Lemma H.2.* To establish the desired inequality, we first derive a series of foundational inner product identities and bounds that will support the subsequent analysis. We begin with a key identity involving the gradient of $F_j$ with respect to the weight matrix $\mathbf{W}_j^{(t)}$. Expanding the Frobenius inner product $\langle \nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{W}_j^{(t)} \rangle$ as a sum over the individual weight vectors $\{\mathbf{w}_{j,r}^{(t)}\}_{r=1}^m$, we leverage the fundamental property of the activation function derivative $\sigma'(x)x = \sigma(x)$ for all $x \in \mathbb{R}$. This substitution simplifies the sum to a direct expression of $F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i)$, yielding the identity:

$$\begin{aligned}
\langle \nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{W}_j^{(t)} \rangle &= \sum_{r=1}^{m} \langle \nabla_{\mathbf{w}_{j,r}^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{w}_{j,r}^{(t)} \rangle \\
&= \sum_{r=1}^{m} \frac{1}{m} \left[ \sigma'\left( \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle \right) \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle + \sigma'\left( \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle \right) \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle \right] \\
&= F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i). \qquad (61)
\end{aligned}$$

Next, we decompose the inner product $\langle \nabla_{\mathbf{w}_{j,r}^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{w}_{j,r}^* \rangle$ by considering the structure of the optimal weight matrix $\mathbf{W}^*$ and the orthogonality between the noise vector $\boldsymbol{\xi}_i$ and the signal vectors $\{\boldsymbol{\mu}_\tau\}_{\tau \in [\mathcal{T}]}$. Two distinct cases arise based on whether the class index $j$ matches the true label $y_i$ of the input $\mathbf{x}_i$. When $j = y_i$, the optimal weight vector $\mathbf{w}_{y_i,r}^*$ includes an additional signal component, leading to the decomposition:

$$\begin{aligned}
\langle \nabla_{\mathbf{w}_{y_i,r}^{(t)}} F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i), \mathbf{w}_{y_i,r}^* \rangle &= \frac{1}{m}\sigma'\left( \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\mu}_{y_i} \rangle \right) \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\mu}_{y_i} \rangle + \frac{1}{m}\sigma'\left( \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle \right) \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle \\
&\quad + \frac{1}{m}\sigma'\left( \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle \right) \cdot 10\log(2\mathcal{T}/\epsilon) \\
&\quad + \frac{1}{m}\sigma'\left( \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle \right) \cdot 10\log(2\mathcal{T}/\epsilon) \cdot \Big( \sum_{s \neq i, y_s = y_i} \frac{\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_s \rangle}{\|\boldsymbol{\xi}_s\|_2^2} \Big). \qquad (62)
\end{aligned}$$

In contrast, when $j \neq y_i$, the optimal weight vector $\mathbf{w}_{j,r}^*$ lacks this additional component, resulting in the simpler decomposition:

$$\begin{aligned}\langle \nabla_{\mathbf{w}_{j,r}^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{w}_{j,r}^* \rangle &= \frac{1}{m}\sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i}\rangle\right)\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{y_i}\rangle + \frac{1}{m}\sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right)\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle \\ &\quad + \frac{1}{m}\sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right)\cdot 10\log(2\mathcal{T}/\epsilon)\cdot\Big(\sum_{y_s=j}\frac{\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_s\rangle}{\|\boldsymbol{\xi}_s\|_2^2}\Big). \end{aligned} \tag{63}$$

Building on these decompositions, we derive bounds for the aggregated inner product $A_{j,i} := \langle \nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{W}_j^* \rangle$ by summing the component-wise results over $r \in [m]$. For the case $j = y_i$, summing the expression in (62) and applying Lemma C.2, along with the definition of $\beta := \max_{j,r,\tau,i}\{|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_\tau\rangle|, |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle|\}$, we obtain a lower bound. Condition 4.1 ensures that this bound simplifies to $A_{y_i,i} \geq 2\log(2\mathcal{T}/\epsilon)$, as shown below:

$$\begin{aligned} A_{y_i,i} &= \sum_{r=1}^m \langle \nabla_{\mathbf{w}_{y_i,r}^{(t)}} F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i), \mathbf{w}_{y_i,r}^* \rangle \\ &= \sum_{r=1}^m \frac{1}{m}\Big[\sigma'\left(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\mu}_{y_i}\rangle\right)\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\mu}_{y_i}\rangle + \sigma'\left(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right)\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle \\ &\quad + \sigma'\left(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right)\cdot 10\log(2\mathcal{T}/\epsilon) + \sigma'\left(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right)\cdot 10\log(2\mathcal{T}/\epsilon)\Big(\sum_{s\neq i, y_s=y_i}\frac{\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_s\rangle}{\|\boldsymbol{\xi}_s\|_2^2}\Big)\Big] \\ &\geq 2.5\log(2\mathcal{T}/\epsilon) - \beta - \beta - \frac{\widetilde{O}(n)}{\sqrt{d-\mathcal{T}}} \geq 2\log(2\mathcal{T}/\epsilon). \end{aligned} \tag{64}$$

For $j \neq y_i$, summing the expression in (63) and applying the triangle inequality to bound the absolute value of the sum, we find that $|A_{j,i}| \leq 2\beta$, a result that follows directly from the definition of $\beta$ and the fact that $\sigma'(\cdot) \leq 1$:

$$\begin{aligned} |A_{j,i}| &= \Big|\sum_{r=1}^m \langle \nabla_{\mathbf{w}_{j,r}^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{w}_{j,r}^* \rangle\Big| \\ &= \Big|\sum_{r=1}^m \frac{1}{m}\Big[\sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i}\rangle\right)\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{y_i}\rangle + \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right)\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right)\cdot A\Big(\sum_{y_s=j}\frac{\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_s\rangle}{\|\boldsymbol{\xi}_s\|_2^2}\Big)\Big]\Big| \\ &\leq \sum_{r=1}^m \frac{1}{m}\Big[\sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i}\rangle\right)|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{y_i}\rangle| + \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right)|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle| + \sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right)\cdot A\Big(\sum_{y_s=j}\frac{|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_s\rangle|}{\|\boldsymbol{\xi}_s\|_2^2}\Big)\Big] \\ &\leq 2\beta + \frac{\widetilde{O}(n)}{\sqrt{d-\mathcal{T}}}. \end{aligned} \tag{65}$$

With these bounds in place, we now establish a lower bound for the critical inner product $\langle \nabla L_S(\mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^*\rangle$, which links the gradient of the loss to the difference between the current and optimal weight matrices. Expanding this inner product by first decomposing over the class index $j$ and then over the training samples $i$, we substitute the identity for $\langle \nabla_{\mathbf{W}_j^{(t)}} F_j, \mathbf{W}_j^{(t)}\rangle$ and the definition of $A_{j,i}$ to rewrite the expression in terms of $F_j$ and $A_{j,i}$. Leveraging the convexity of the cross-entropy loss $\ell$, we relate the linear term involving the loss derivative to the loss itself, introducing a logarithmic correction term. The bounds on $A_{j,i}$ derived earlier, specifically $A_{y_i,i} \geq 2\log(2\mathcal{T}/\epsilon)$ and $|A_{j,i}| \leq 2\beta + \frac{\widetilde{O}(n)}{\sqrt{d-\mathcal{T}}}$ for $j \neq y_i$, ensure that this logarithmic term is bounded by $0.5\epsilon$ under Condition 4.1, leading to

the lower bound:

$$\begin{aligned}
\langle \nabla L_S(\mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \rangle &= \sum_{j=1}^{\mathcal{T}} \langle \nabla_{\mathbf{W}_j^{(t)}} L_S(\mathbf{W}^{(t)}), \mathbf{W}_j^{(t)} - \mathbf{W}_j^* \rangle \\
&= \sum_{j=1}^{\mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} \ell_{j,i}^{'(t)} \langle \nabla_{\mathbf{W}_j^{(t)}} F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i), \mathbf{W}_j^{(t)} - \mathbf{W}_j^* \rangle \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{\mathcal{T}} \ell_{j,i}^{'(t)} \left( F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) - A_{j,i} \right) \\
&\geq \frac{1}{n} \sum_{i=1}^{n} \left( \ell(\mathbf{W}^{(t)}, \mathbf{x}_i, y_i) - \log \left( 1 + \sum_{j \neq y_i} \exp\{A_{j,i} - A_{y_i,i}\} \right) \right) \\
&\geq L_S(\mathbf{W}^{(t)}) - 0.5\epsilon. \qquad (66)
\end{aligned}$$

To complete the proof, we use the weight update rule $\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \nabla L_S(\mathbf{W}^{(t)})$ and the algebraic identity for the difference of squared vector norms, which states that $\|\mathbf{a} - \mathbf{b}\|^2 - \|\mathbf{a} - \eta\mathbf{g} - \mathbf{b}\|^2 = 2\eta\langle \mathbf{g}, \mathbf{a} - \mathbf{b} \rangle - \eta^2 \|\mathbf{g}\|^2$ for any vectors $\mathbf{a}, \mathbf{b}, \mathbf{g}$ and scalar $\eta$. Applying this identity with $\mathbf{a} = \mathbf{W}^{(t)}$, $\mathbf{b} = \mathbf{W}^*$, and $\mathbf{g} = \nabla L_S(\mathbf{W}^{(t)})$, we rewrite the difference of Frobenius norms as:

$$\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 = 2\eta \langle \nabla L_S(\mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \rangle - \eta^2 \|\nabla L_S(\mathbf{W}^{(t)})\|_F^2.$$

Substituting the lower bound from (66) and the gradient bound from Lemma F.2—which asserts $\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \leq O\left(\mathcal{T}^2 \max\{\sigma_p^2 d, \mu^2\}\right) L_S(\mathbf{W}^{(t)})$—into this expression, we obtain:

$$\begin{aligned}
&\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \\
&\geq 2\eta \left( L_S(\mathbf{W}^{(t)}) - 0.5\epsilon \right) - \eta^2 O\left(\mathcal{T}^2 \max\{\sigma_p^2 d, \mu^2\}\right) L_S(\mathbf{W}^{(t)}).
\end{aligned}$$

Condition 4.1 ensures that the step size $\eta$ is sufficiently small such that $\eta O\left(\mathcal{T}^2 \max\{\sigma_p^2 d, \mu^2\}\right) \leq 1$. This guarantee allows us to simplify the right-hand side by bounding the coefficient of $L_S(\mathbf{W}^{(t)})$ below by $\eta$, leading to the final result:

$$\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \geq \eta L_S(\mathbf{W}^{(t)}) - \eta\epsilon.$$

This completes the proof. □

**Lemma H.3.** *Under Condition 4.1, we have that* $\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F \leq \widetilde{O}(\sqrt{\mathcal{T}mn}\sigma_p^{-1} d^{-1/2})$.

*Proof of Lemma H.3.* We have

$$\begin{aligned}
\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F &\leq \|\mathbf{W}^{(T_1)} - \mathbf{W}^{(0)}\|_F + \|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F \\
&= \sqrt{\sum_{j,r} \|\mathbf{w}_{j,r}^{(T_1)} - \mathbf{w}_{j,r}^{(0)}\|_F^2} + \sqrt{\sum_{j,r} \|\mathbf{w}_{j,r}^{*} - \mathbf{w}_{j,r}^{(0)}\|_F^2} \\
&\leq O(\sqrt{\mathcal{T}m}) \max_{j,r} \| \sum_{\tau=1}^{\mathcal{T}} \gamma_{j,r,\tau}^{(T_1)} \cdot \mu^{-2} \boldsymbol{\mu}_\tau \|_2 + O(\sqrt{\mathcal{T}m}) \max_{j,r} \| \sum_{i=1}^{n} \rho_{j,r,i}^{(T_1)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \boldsymbol{\xi}_i \|_2 + \widetilde{O}(\sqrt{\mathcal{T}mn/\mathcal{T}}\sigma_p^{-1} d^{-1/2}) \\
&\leq \widetilde{O}(\sqrt{\mathcal{T}m}\mu^{-1}) + \widetilde{O}(\sqrt{\mathcal{T}mn}\sigma_p^{-1} d^{-1/2}) \\
&\leq \widetilde{O}(\sqrt{\mathcal{T}m}\mu^{-1}),
\end{aligned}$$

where the first inequality is by triangle inequality, the second inequality is by our decomposition of $\mathbf{W}^{(T_1)}$ and $\mathbf{W}^*$, the third inequality is by Lemma C.4 and Proposition E.1, and the last inequality follows by (17). Here the proof completes. □

### H.3. Population Loss

**Lemma H.4.** *Under Condition 4.1, we can have that* $L_{\mathcal{D}}(\mathbf{W}^{(t)}) \geq 0.5 \cdot \log(1 + \frac{\mathcal{T}-1}{e})$ *for any* $0 \leq t \leq T^*$.

*Proof of Lemma H.4.* Given a new example $(\mathbf{x}, y)$, we have that

$$\begin{aligned}
\|\mathbf{w}_{j,r}^{(t)}\|_2 &= \|\mathbf{w}_{j,r}^{(0)} + \sum_{\tau=1}^{\mathcal{T}} \gamma_{j,r,\tau}^{(t)} \cdot \mu^{-2}\boldsymbol{\mu}_\tau + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2}\boldsymbol{\xi}_i\|_2 \\
&\leq \|\mathbf{w}_{j,r}^{(0)}\|_2 + \|\sum_{\tau=1}^{\mathcal{T}} \gamma_{j,r,\tau}^{(t)} \cdot \mu^{-2}\boldsymbol{\mu}_\tau\|_2 + \|\sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2}\boldsymbol{\xi}_i\|_2 \\
&= O(\sigma_0\sqrt{d}) + \widetilde{O}(n^{-1}\mu^{-1}) + \widetilde{O}(\sigma_p^{-1}\sqrt{nd^{-1}}) \\
&= \widetilde{O}(\sigma_0\sqrt{d}) + \widetilde{O}(\sigma_p^{-1}\sqrt{nd^{-1}}),
\end{aligned} \tag{67}$$

where the first inequality is by triangle inequality. Therefore, we have that $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle \sim \mathcal{N}(0, \sigma_p^2\|\mathbf{w}_{j,r}^{(t)}\|_2^2)$. So with probability at least $1 - 1/2$

$$|\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle| \leq \widetilde{O}(\sigma_0\sigma_p\sqrt{d} + \sqrt{nd^{-1}}),$$

for all $j \in [\mathcal{T}]$ and $r \in [m]$. Since the signal vectors are orthogonal to noises, by $\max_{\tau,r} \gamma_{\tau,r,\tau}^{(t)} \leq \widetilde{O}(\widetilde{\gamma}) = \widetilde{O}(n^{-1})$, we also have that $|\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i}\rangle| \leq |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{y_i}\rangle| + |\gamma_{j,r,y_i}^{(t)}| = \widetilde{O}(\sigma_o\mu + n^{-1})$. Then with probability at least $0.5$, we have that

$$\begin{aligned}
F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) &= \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i}\rangle) + \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle)] \\
&\leq \max_r |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}_{y_i}\rangle| + \max_r |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle| \\
&\leq \widetilde{O}(\sigma_o\mu + n^{-1}) + \widetilde{O}(\sigma_0\sigma_p\sqrt{d} + \sqrt{nd^{-1}}) \\
&\leq 1,
\end{aligned}$$

where the last inequality is by Condition 4.1. Therefore, with probability at least $0.5$, we have that

$$\ell(F(\mathbf{W}, \mathbf{x})) = \log(1 + \frac{\sum_{j \neq y_i} \exp\{F_j(\mathbf{W}_j^{(t)}, \mathbf{x})\}}{\exp\{F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x})\}}) \geq \log(1 + \frac{\mathcal{T}-1}{e}).$$

Thus $L_{\mathcal{D}}(\mathbf{W}^{(t)}) \geq 0.5 \cdot \log(1 + \frac{\mathcal{T}-1}{e})$. This completes the proof.

□

## I. TopK Pruning

By the update rule for $\mathbf{w}_{j,r}^{(t)}$, we have

$$\mathbf{w}_{j,r}^{(1)} = \mathbf{w}_{j,r}^{(0)} + \frac{\eta}{nm}\sum_{i=1}^{n}\left(-\ell_{j,i}^{'(0)}\right)\sigma'\left(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle\right)\boldsymbol{\xi}_i + \frac{\eta}{nm}\sum_{i=1}^{n}\left(-\ell_{j,i}^{'(0)}\right)\sigma'\left(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{y_i}\rangle\right)\boldsymbol{\mu}_{y_i}.$$

When $t = 0$, we can have that

$$-\ell_{j,i}^{'(t)} = \mathbb{I}(j = y_i) - \text{logit}_j(F, \mathbf{x}_i) = \mathbb{I}(j = y_i) - \Theta(\frac{1}{\mathcal{T}}),$$

Each term of $\frac{\eta}{nm}\sum_{i=1}^{n}\left(-\ell_{j,i}^{'(0)}\right)\sigma'\left(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle\right)\boldsymbol{\xi}_i$ can be approximated as $\frac{\eta}{2\mathcal{T}^2m}x_1$, where $x_1 \sim \mathcal{N}(0, \mathcal{T}(\mathcal{T}-1)\sigma_p^2)$. $\frac{\eta}{nm}\sum_{i=1}^{n}\left(-\ell_{j,i}^{'(0)}\right)\sigma'\left(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}_{y_i}\rangle\right)\boldsymbol{\mu}_{y_i}$ at position $j$ can be approximated as $\frac{\eta(\mathcal{T}-1)\mu}{2\mathcal{T}^2m}$, and other positions can be

approximated as $-\frac{\eta\mu}{2\mathcal{T}^2 m}$. By Gauss tail bound, with probability at least $1-\delta$, we can have that $|x_1| \leq (\mathcal{T}-1)\mu$ for all $j \in [\mathcal{T}]$, $r \in [m]$ and at least $d-k-\mathcal{T}$ positions in every $\mathbf{w}_{j,r}^{(1)}$. We can derive that $k \geq d-\mathcal{T}-\frac{\delta}{2\mathcal{T}m}\exp\{\frac{(\mathcal{T}-1)\mu^2}{2\mathcal{T}\sigma_p^2}\}$. If we have that $k \leq \Theta(\mu^2\sqrt{n}\mathcal{T}^{-1}\sigma_p^{-2})$, we have $p := \mathbb{P}(|\mathcal{N}(0,\sigma_p^2)| \leq \mu)$

$$1-2\exp\Big(-\frac{c^2}{2\sigma_p^2}\Big) \leq \mathbb{P}(|x| \leq c) \leq \sqrt{1-\exp\Big(-\frac{2c^2}{\pi\sigma_p^2}\Big)}.$$

$$1-2\exp\Big(-\frac{\mu^2}{2\sigma_p^2}\Big) \leq p \leq \sqrt{1-\exp\Big(-\frac{2\mu^2}{\pi\sigma_p^2}\Big)}.$$

Let $S := \sum_{k=1}^{d} \mathbb{1}\{|\mathbf{w}_k| \leq \mu\}$. By Hoeffding, we have that $dp - O(\sqrt{d\log(1/\delta)}) \leq S \leq dp + O(\sqrt{d\log(1/\delta)})$. So we can have that when $K \geq \Theta(d)\exp\Big(-\frac{\mu}{2\sigma_p^2}\Big) + O(\sqrt{d\log(1/\delta)})$, the signal part of $\mathbf{w}_{j,r}^{(1)}$ will be selected, and the left part of $\mathbf{w}_{j,r}^{(1)}$ is no more than $\sqrt{\mu}$. If we compare the incremental changes between $\gamma_{j,r,\tau}^{(t)}$ and $\bar{\rho}_{j,r,i}^{(t)}$, we can have that

$$\frac{\gamma_{j,r,\tau}^{(t)}\mu^{-1}}{\sum_i \mathbb{1}(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle > 0)\bar{\rho}_{j,r,i}^{(t)}\sigma_p^2 d\sqrt{\mu}} = \frac{\mathcal{T}^{-1}n\mathrm{SNR}^2\mu^{-1}}{\Theta(n/\mathcal{T})\sigma_p^2 d\sqrt{\mu}} = \Theta(\sqrt{\mu}),$$

so we can get that the select part will always be selected, and the network will learn the signals. If $K \leq d(1-\sqrt{1-\exp\Big(-\frac{2\mu^2}{\pi\sigma_p^2}\Big)}) - O(\sqrt{d\log(1/\delta)})$, then the some of the signal part will not be selected. Some of the noise part and some of the signal part will increase together, so the train loss $L_S(\mathbf{W}^{(t)}) \leq \epsilon$. However, the test loss $L_{\mathcal{D}}(\mathbf{W}^{(t)}) \geq \Theta(\log(\mathcal{T}))$ because the network can not learn some signals.