Benign Overfitting in Two-layer ReLU Convolutional Neural Networks

Yiwen Kou [*1] **Zixiang Chen** [*1] **Yuanzhou Chen** [1] **Quanquan Gu** [1]

Abstract

Modern deep learning models with great expressive power can be trained to overfit the training data but still generalize well. This phenomenon is referred to as *benign overfitting*. Recently, a few studies have attempted to theoretically understand benign overfitting in neural networks. However, these works are either limited to neural networks with smooth activation functions or to the neural tangent kernel regime. How and when benign overfitting can occur in ReLU neural networks remains an open problem. In this work, we seek to answer this question by establishing algorithm-dependent risk bounds for learning two-layer ReLU convolutional neural networks with label-flipping noise. We show that, under mild conditions, the neural network trained by gradient descent can achieve near-zero training loss and Bayes optimal test risk. Our result also reveals a sharp transition between benign and harmful overfitting under different conditions on data distribution in terms of test risk. Experiments on synthetic data back up our theory.

1. Introduction

Modern deep learning models have a large number of parameters, often exceeding the number of training data points. Despite being over-parameterized and overfitting the training data, these models can still make accurate predictions on the unseen test data (Zhang et al., 2017; Neyshabur et al., 2018b). This phenomenon, often referred to as *benign overfitting* (Bartlett et al., 2020), has revolutionized traditional theories of statistical learning and attracted great attention from the statistics and machine learning communities (Belkin et al., 2018; 2019; 2020; Hastie et al., 2022).

[*]Equal contribution [1]Department of Computer Science, University of California, Los Angeles. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

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

There has been a line of work in recent years studying benign overfitting from the theoretical perspective. Despite their contributions and insights into the benign overfitting phenomenon, most of these works focus on linear models (Belkin et al., 2020; Bartlett et al., 2020; Hastie et al., 2022; Wu & Xu, 2020; Chatterji & Long, 2021; Zou et al., 2021b; Cao et al., 2021) or kernel/random features models (Belkin et al., 2018; Liang & Rakhlin, 2020; Montanari & Zhong, 2022). Adlam & Pennington (2020) and Li et al. (2021) focused on benign overfitting in neural network models, yet their results are limited to the neural tangent kernel (NTK) regime (Jacot et al., 2018), where the neural network learning is essentially equivalent to kernel regression.

Understanding benign overfitting in neural networks beyond the NTK regime is much more challenging because of the non-convexity of the problem. Recently, Frei et al. (2022) studied the problem of learning log-concave mixture data with label-flipping noise, using fully-connected two-layer neural networks with smoothed leaky ReLU activation. They proved the risk upper bound under certain regularity conditions, which matches the lower bound given in Cao et al. (2021) when the label-flipping noise is zero. Cao et al. (2022) provided an analysis for learning two-layer convolutional neural networks (CNNs) with polynomial ReLU activation function (ReLU^q, $q > 2$). Their analysis also identifies a condition that controls the phase transition between benign and harmful overfitting. The analyses in both Frei et al. (2022) and Cao et al. (2021) highly rely on smooth activation functions and cannot deal with the most widely used ReLU activation function. Thus, there remains an open question:

How and when does benign overfitting occur in ReLU neural networks?

In this paper, we seek to answer the above question by establishing risk bounds for learning two-layer CNNs with ReLU activation function.

1.1. Problem Setup

We consider a similar data distribution that was explored in Cao et al. (2022). In this particular distribution, the input

data consists of two types of components: *label dependent signals* and *label independent noises*. This data generation model takes inspiration from image data, where the inputs are composed of various patches, and only certain patches are relevant to the class label of the image. Similar models have also been investigated in recent works by Li et al. (2019); Allen-Zhu & Li (2020a;b); Zou et al. (2021a); Shen & Bubeck (2022).

Definition 1.1. *Let $\boldsymbol{\mu} \in \mathbb{R}^d$ be a fixed vector representing the signal contained in each data point. Each data point $(\mathbf{x}, y)$ with predictor $\mathbf{x} = [\mathbf{x}^{(1)\top}, \mathbf{x}^{(2)\top}]^\top \in \mathbb{R}^{2d}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)} \in \mathbb{R}^d$ and label $y \in \{-1, 1\}$ is generated from a distribution $\mathcal{D}$, which we specify as follows:*

1. *The true label $\widehat{y}$ is generated as a Rademacher random variable, i.e. $\mathbb{P}[\widehat{y} = 1] = \mathbb{P}[\widehat{y} = -1] = 1/2$. The observed label y is then generated by flipping $\widehat{y}$ with probability p where $p < 1/2$, i.e. $\mathbb{P}[y = \widehat{y}] = 1 - p$ and $\mathbb{P}[y = -\widehat{y}] = p$.*
2. *A noise vector $\boldsymbol{\xi}$ is generated from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma_p^2 \mathbf{I})$.*
3. *One of $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ is randomly selected and then assigned as $\widehat{y} \cdot \boldsymbol{\mu}$, which represents the signal, while the other is given by $\boldsymbol{\xi}$, which represents noises.*

Definition 1.1 strictly generalizes the data distribution in Cao et al. (2022), in the sense that it introduces label-flipping noise to the true label $\widehat{y}$, and relaxes the orthogonal condition between the signal vector $\boldsymbol{\mu}$ and the noise vectors $\boldsymbol{\xi}$ (See Definition 3.1 in Cao et al. (2022) for a comparison).

Given a training data set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ drawn from some unknown joint distribution $\mathcal{D}$ over $\mathbf{x} \times y$, we train a two-layer CNN with ReLU activation by minimizing the following empirical risk measured by logistic loss

$$L_S(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^{n} \ell[y_i \cdot f(\mathbf{W}, \mathbf{x}_i)], \tag{1.1}$$

where $\ell(z) = \log(1 + \exp(-z))$, and $f(\mathbf{W}, \mathbf{x})$ is the two-layer CNN (See Section 3 for the detailed definition). We will use gradient descent to minimize the training loss $L_S(\mathbf{W})$, and we are interested in characterizing the test error (i.e., true error) defined by

$$L_{\mathcal{D}}^{0-1}(\mathbf{W}) := \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[y \neq \operatorname{sign}\big(f(\mathbf{W}, \mathbf{x})\big)\big]. \tag{1.2}$$

1.2. Main Contributions

We prove the following main result, which characterizes the training loss and test error of the two-layer ReLU CNN trained by gradient descent.

Theorem 1.2 (Informal)**.** *For any $\epsilon > 0$, under certain regularity conditions, with probability at least $1 - \delta$, there exists $0 \leq t \leq T$ such that:*

1. *The training loss converges to ϵ, i.e., $L_S(\mathbf{W}^{(t)}) \leq \epsilon$.*
2. *If $n\|\boldsymbol{\mu}\|_2^4 \geq \Omega(\sigma_p^4 d)$, we have $L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) \leq p + \exp\Big(-n\|\boldsymbol{\mu}\|_2^4/(C_2\sigma_p^4 d)\Big)$.*
3. *If $n\|\boldsymbol{\mu}\|_2^4 \leq O(\sigma_p^4 d)$, we have $L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) \geq p + 0.1$.*

The significance of Theorem 1.2 is highlighted as follows:

- The ReLU CNN trained by standard gradient descent on the logistic loss can interpolate the noisy training data and achieve near-zero training loss.
- Under the condition on the data distribution and the training sample size that $n\|\boldsymbol{\mu}\|_2^4 \geq \Omega(\sigma_p^4 d)$, the learned CNN can achieve nearly optimal test error (i.e., Bayes risk p).
- On the flip side, if $n\|\boldsymbol{\mu}\|_2^4 \leq O(\sigma_p^4 d)$, the interpolating CNN model will suffer a test error that is at least a constant worse than the Bayes risk. This together with the positive result reveals a sharp phase transition between benign and harmful overfitting.

Our analysis relies on several new proof techniques that significantly generalize the signal-noise decomposition technique (Cao et al., 2022). More specifically, to handle ReLU activation, we directly use the activation pattern and data structure to characterize the loss of each training data point rather than using the smoothness condition. To deal with the label-flipping noise, we show that the loss of each training data point decreases at roughly the same rate throughout training, which ensures signal learning even in the presence of label noise.

2. Related Work

In this section, we will discuss in detail some of the related work briefly mentioned before.

Benign overfitting of linear models. One line of research sought a theoretical understanding of the benign overfitting phenomenon in linear models. Some of these works focused on linear regression problems. Belkin et al. (2020) provided a precise analysis for the shape of the risk curve in Gaussian and Fourier series models with the least squares predictor. Hastie et al. (2022); Wu & Xu (2020) studied the setting where both the dimension and sample size grow but their ratio is fixed and demonstrated a double descent risk

curve with respect to this ratio. Bartlett et al. (2020) established matching upper and lower risk bounds for the overparameterized minimum norm interpolator and showed that benign overfitting can occur under certain conditions on the spectrum of the data covariance. Zou et al. (2021b) studied how well constant stepsize stochastic gradient descent with iterate averaging or tail averaging generalizes in the overparameterized regime. Several other works studied benign overfitting of maximum margin linear classifiers. Muthukumar et al. (2021) showed that the max-margin predictor and the least square predict coincide in the overparametrized regime, and generalize differently when using 0-1 loss and square loss functions. Wang & Thrampoulidis (2021); Cao et al. (2021) respectively studied Gaussian and sub-Gaussian mixtures data models without label noise and characterized the condition under which benign overfitting can occur. Chatterji & Long (2021) showed that the maximum margin algorithm trained on noisy data can achieve nearly optimal risk with sufficient overparameterization. Shamir (2022) studied both minimum-norm interpolating predictors for linear regression and max-margin predictors for classification and discussed the conditions under which benign overfitting can or cannot occur.

Benign overfitting of neural networks. A series of recent works studied benign overfitting of neural networks. Liang et al. (2020) showed that kernel "ridgeless" regression can lead to a multiple-descent risk curve for various scaling of input dimension and sample size. Adlam & Pennington (2020) provided a precise analysis of generalization under kernel regression and revealed non-monotonic behavior for the test error. Li et al. (2021) examined benign overfitting in random feature models defined as two-layer neural networks. Montanari & Zhong (2022) studied two-layer neural networks in the NTK regime, focusing on its generalization properties when dimension, sample size, and the number of neurons are overparametrized and polynomially related. Chatterji & Long (2022) bounded the excess risk of interpolating deep linear networks trained by gradient flow and showed that randomly initialized deep linear networks can closely approximate the risk bounds for the minimum norm interpolator.

3. Preliminaries

In this section, we introduce the notation, two-layer CNN models, and the gradient descent-based training algorithm.

Notation. We use lower case letters, lower case bold face letters, and upper case bold face letters to denote scalars, vectors, and matrices respectively. For a vector $\mathbf{v} = (v_1, \cdots, v_d)^\top$, we denote by $\|\mathbf{v}\|_2 := \big(\sum_{j=1}^d v_j^2\big)^{1/2}$ its l_2 norm. For two sequence $\{a_k\}$ and $\{b_k\}$, we denote $a_k = O(b_k)$ if $|a_k| \le C|b_k|$ for some absolute constant C, denote $a_k = \Omega(b_k)$ if $b_k = O(a_k)$, and denote $a_k = \Theta(b_k)$ if $a_k = O(b_k)$ and $a_k = \Omega(b_k)$. We also denote $a_k = o(b_k)$ if $\lim |a_k/b_k| = 0$. Finally, we use $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to omit logarithmic terms in the notation.

Two-layer CNNs. We consider a two-layer convolutional neural network described in the following: its first layer consists of m positive filters and m negative filters, with each filter applying to the two patches $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ separately; its second layer parameters are fixed as $+1/m$ and $-1/m$ respectively for positive and negative convolutional filters. Then the network can be written as $f(\mathbf{W}, \mathbf{x}) = F_{+1}(\mathbf{W}_{+1}, \mathbf{x}) - F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$, where the partial network function of positive and negative filters $F_{+1}(\mathbf{W}_{+1}, \mathbf{x})$, $F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$ are defined as:

$$\begin{aligned} F_j(\mathbf{W}_j, \mathbf{x}) &= \frac{1}{m}\sum_{r=1}^m \big[\sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}^{(1)}\rangle) + \sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}^{(2)}\rangle)\big] \\ &= \frac{1}{m}\sum_{r=1}^m \big[\sigma(\langle \mathbf{w}_{j,r}, \widehat{y}\cdot\boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}\rangle)\big] \end{aligned}$$

for $j \in \{\pm 1\}$. Here $\sigma(z) = \max\{0, z\}$ is the ReLU activation function, $\mathbf{W}_j$ is the collection of model weights associated with F_j (positive/negative filters), and $\mathbf{w}_{j,r} \in \mathbb{R}^d$ denotes the weight vector for the r-th filter / neuron in $\mathbf{W}_j$. We use $\mathbf{W}$ to denote the collection of all model weights. We note that our CNN model can also be viewed as a CNN with average global pooling (Lin et al., 2013). Besides the training loss and test error defined in (1.1) and (1.2), we also define the true loss (test loss) as $L_{\mathcal{D}}(\mathbf{W}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\ell[y\cdot f(\mathbf{W}, \mathbf{x})]$.

Training algorithm. We use gradient descent to optimize (1.1). The gradient descent update of the filters in the CNN can be written as

$$\begin{aligned} \mathbf{w}_{j,r}^{(t+1)} &= \mathbf{w}_{j,r}^{(t)} - \eta\cdot\nabla_{\mathbf{w}_{j,r}} L_S(\mathbf{W}^{(t)}) \\ &= \mathbf{w}_{j,r}^{(t)} - \frac{\eta}{nm}\sum_{i=1}^n \ell_i'^{(t)}\cdot\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\cdot j y_i \boldsymbol{\xi}_i \\ &\quad - \frac{\eta}{nm}\sum_{i=1}^n \ell_i'^{(t)}\cdot\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\boldsymbol{\mu}\rangle)\cdot \widehat{y}_i y_i j\boldsymbol{\mu}. \end{aligned} \tag{3.1}$$

for all $j \in \{\pm 1\}$ and $r \in [m]$, where we introduce a shorthand notation $\ell_i'^{(t)} = \ell'[y_i\cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)]$ and assume the gradient of the ReLU activation function at 0 to be $\sigma'(0) = 1$ without losing generality. We initialize the gradient descent by Gaussian initialization, where all entries of $\mathbf{W}^{(0)}$ are sampled from i.i.d. Gaussian distributions $\mathcal{N}(0, \sigma_0^2)$, with

σ_0^2 as the variance.

4. Main Results

In this section, we present our main theoretical results. Our results are based on the following conditions on the dimension d, sample size n, neural network width m, initialization scale σ_0, signal norm $\|\boldsymbol{\mu}\|_2$, noise rate p, and learning rate η. In this paper, we consider the learning period $0 \leq t \leq T^*$, where $T^* = \eta^{-1}\text{poly}(\epsilon^{-1}, d, n, m)$ is the maximum admissible iterations. We can deal with any polynomial maximum admissible iterations T^* greater than $\widetilde{\Omega}(\eta^{-1}\epsilon^{-1}mnd^{-1}\sigma_p^{-2})$.

Condition 4.1. *Suppose there exists a sufficiently large constant C, such that the following hold:*

1. *Dimension d is sufficiently large: $d \geq C\max\{n\sigma_p^{-2}\|\boldsymbol{\mu}\|_2^2\log(T^*), n^2\log(nm/\delta)(\log(T^*))^2\}$.*
2. *Training sample size n and neural network width satisfy $m \geq C\log(n/\delta), n \geq C\log(m/\delta)$.*
3. *The norm of the signal satisfies $\|\boldsymbol{\mu}\|_2^2 \geq C \cdot \sigma_p^2\log(n/\delta)$.*
4. *The noise rate p satisfies $p \leq 1/C$.*
5. *The standard deviation of Gaussian initialization σ_0 is appropriately chosen such that $\sigma_0 \leq \big(C\max\big\{\sigma_p d/\sqrt{n}, \sqrt{\log(m/\delta)}\cdot\|\boldsymbol{\mu}\|_2\big\}\big)^{-1}$.*
6. *The learning rate η satisfies $\eta \leq \big(C\max\big\{\sigma_p^2 d^{3/2}/(n^2 m\sqrt{\log(n/\delta)}), \sigma_p^2 d/n\big\}\big)^{-1}$.*

The conditions on d, n, m are to ensure that the learning problem is in a sufficiently over-parameterized setting, and similar conditions have been made in Chatterji & Long (2021); Cao et al. (2022); Frei et al. (2022). The conditions on σ_0 and η are to ensure that gradient descent can effectively minimize the training loss. The difference between Condition 4.1 and Assumption (A1)-(A6) in Frei et al. (2022) is that our setting assumes a milder condition of order $O(d^{-3/2})$ on learning rate η rather than $O(d^{-2})$ ((A5) in Frei et al. (2022)), as well as a milder condition of order $O(d^{-1}n^{1/2})$ on initialization σ_0 rather than $O(d^{-5/2}m^{-1/2})$ ((A6) in Frei et al. (2022)). Another difference is that Frei et al. (2022) allows neural networks of arbitrary width m, but our condition requires a mild assumption that m should be no more than an exponential order of dimension d. We also require another mild condition that m and n cannot exceed the exponential order of each other. Besides, in contrast to Cao et al. (2022), our Condition 4.1 relaxes the dependency of m and d in that, we do not require any polynomial upper bound of the neural network width m, whereas Condition 4.2 in Cao et al. (2022) requires that m is upper bounded by a certain fractional order of d. Another improvement to Cao et al. (2022) is that we add label-flipping noise p to the problem, but this is also included in Frei et al. (2022). Detailed comparisons are shown in Table 1 and Table 2.

Based on these conditions, we give our main result in the following theorem.

Theorem 4.2. *For any $\epsilon > 0$, under Condition 4.1, with probability at least $1-\delta$ there exists $t = \widetilde{O}(\eta^{-1}\epsilon^{-1}mnd^{-1}\sigma_p^{-2})$ such that:*

1. *The training loss converges to ϵ, i.e., $L_S(\mathbf{W}^{(t)}) \leq \epsilon$.*
2. *When $n\|\boldsymbol{\mu}\|_2^4 \geq C_1\sigma_p^4 d$, the trained CNN will generalize with classification error close to the noise rate p: $L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) \leq p + \exp\Big(-n\|\boldsymbol{\mu}\|_2^4/(C_2\sigma_p^4 d)\Big)$.*
3. *When $n\|\boldsymbol{\mu}\|_2^4 \leq C_3\sigma_p^4 d$, the test error $L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) \geq p + 0.1$.*

Here C_1, C_2, C_3 are some absolute constants.

Remark 4.3. *Theorem 4.2 demonstrates that the training loss converges to ϵ within $\widetilde{O}(\eta^{-1}\epsilon^{-1}mnd^{-1}\sigma_p^{-2})$ iterations. Moreover, when the training loss converges, the model can achieve optimal test error if the signal-to-noise ratio is large. However, if the signal-to-noise ratio is small, the model will experience a test error that is at least a constant worse than the Bayes risk. The threshold for this distinction is determined by the condition $n\|\boldsymbol{\mu}\|_2^4 = \Theta(\sigma_p^4 d)$. In addition to the results mentioned in Theorem 4.2, it is important to emphasize that the second and third bullet points regarding the test error also hold true for training time t that is greater than $\widetilde{O}(\eta^{-1}\epsilon^{-1}mnd^{-1}\sigma_p^{-2})$, but smaller than the maximum allowable iterations $T^* = \eta^{-1}\text{poly}(\epsilon^{-1}, d, n, m)$.*

Comparison with prior works. Although Theorem 4.2 and Theorem 3.1 in Frei et al. (2022) both show that the network achieves arbitrarily small logistic loss, and simultaneously achieves test error close to the noise rate, our results differ from Frei et al. (2022) since Frei et al. (2022) considered a neural network with *smoothed leaky ReLU* activation, while we consider the ReLU activation which is not smooth. Besides, to obtain a training error smaller than ϵ, Frei et al. (2022) needed $O(\epsilon^{-2})$ iterations, whereas our results only require $O(\epsilon^{-1})$ iterations. In contrast to Cao et al. (2022) which studied a CNN model with ReLU^q $(q > 2)$ activation function and without label noise, our setting is more practical as we work with ReLU activation, take label-flipping noise into consideration, and also remove the orthogonal

assumption between the signal patch and the noise patch. Because of the label-flipping noise, it is more natural to evaluate generalization performance by comparing the test error with the Bayes optimal classifier. This is why our Theorem 4.2 provides test error bounds while Cao et al. (2022) provided test loss bounds and despite the difference both our results and theirs present exact phase transition conditions.

5. Overview of Proof Techniques

In this section, we discuss the main challenges in studying benign overfitting under our setting, and explain some key techniques we implement in our proofs to overcome these challenges. Based on these techniques, the proof of our main Theorem 4.2 will follow naturally. The complete proofs of all the results are given in the appendix.

5.1. Key Technique 1: Time-invariant Coefficient Ratio

Our first main challenge is dealing with the ReLU activation function, i.e., $\sigma(z) = \max\{0, z\}$. This is one of the most common and widely used activation functions, but as we explain below, it is also hard to analyze. The key difficulty in establishing benign overfitting guarantees is demonstrating that the neural network can interpolate the data. Frei et al. (2022) adopted the smoothness-based convergence proof technique proposed in Frei & Gu (2021). This technique requires the activation function to be strictly increasing and smooth, therefore it cannot be applied to the ReLU activation function. Cao et al. (2022) provides an iterative analysis of the coefficients in the signal-noise decomposition, which is given in the following definition.

Definition 5.1. *Let $\mathbf{w}_{j,r}^{(t)}$ for $j \in \{\pm 1\}$, $r \in [m]$ be the convolution filters of the CNN at the t-th iteration of gradient descent. Then there exist unique coefficients $\gamma_{j,r}^{(t)}$ and $\rho_{j,r,i}^{(t)}$ such that*

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

Further denote $\overline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \geq 0)$, $\underline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \leq 0)$. Then

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \cdot \gamma_{j,r}^{(t)} \cdot \|\boldsymbol{\mu}\|_2^{-2} \cdot \boldsymbol{\mu} + \sum_{i=1}^{n} \overline{\rho}_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i + \sum_{i=1}^{n} \underline{\rho}_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i. \tag{5.1}$$

(5.1) is called the *signal-noise decomposition* of $\mathbf{w}_{j,r}^{(t)}$ where the normalization factors $\|\boldsymbol{\mu}\|_2^{-2}, \|\boldsymbol{\xi}_i\|_2^{-2}$ are to ensure that $\gamma_{j,r}^{(t)} \approx \langle \mathbf{w}_{j,r}^{(t)}, j\boldsymbol{\mu} \rangle$, $\rho_{j,r,i}^{(t)} \approx \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$. With Definition 5.1, one can reduce the study of the CNN learning process to a careful assessment of the coefficients $\gamma_{j,r}^{(t)}$, $\overline{\rho}_{j,r,i}^{(t)}$, $\underline{\rho}_{j,r,i}^{(t)}$ throughout training. This technique does not rely on the strictly increasing and smoothness properties of the activation function and will act as the basis of our analysis. However, Cao et al. (2022) only characterized the behavior of the leading neurons by studying $\max_r \gamma_{j,r}^{(t)}$, $\max_r \overline{\rho}_{j,r,i}^{(t)}$. To guarantee that the leading neuron can dominate other neurons after training, they require neurons with different initial weights to have different update speeds, which is guaranteed thanks to the activation function ReLU^q with $q > 2$. But the ReLU function is piece-wise linear, and every activated neuron has the same learning speed $\sigma'(x) = 1$. Therefore dealing with ReLU requires new techniques.

To overcome this difficulty, we propose a *time-invariant coefficient ratio* analysis which generalizes Cao et al. (2022)'s technique. The key lemma is presented as follows, which characterizes the coefficient orders at any time $t \leq T^*$ and helps derive the second and third parts of Theorem 4.2 on the upper and lower bounds of test error.

Proposition 5.2. *Under Condition 4.1, the following bounds hold for $t \in [0, T^*]$:*

- *$\overline{\rho}_{j,r,i}^{(t)}$ is an increasing sequence. Besides, $0 \leq \overline{\rho}_{j,r,i}^{(t)} \leq 4\log(T^*)$ for all $j \in \{\pm 1\}$, $r \in [m]$ and $i \in [n]$.*
- *$\underline{\rho}_{j,r,i}^{(t)}$ is a decreasing sequence. Besides, $-4\log(T^*) \leq -2\max_{i,j,r}\{|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle|, |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle|\} - 10n\sqrt{\log(6n^2/\delta)/d} \cdot 4\log(T^*) \leq \underline{\rho}_{j,r,i}^{(t)} \leq 0$ for all $j \in \{\pm 1\}$, $r \in [m]$ and $i \in [n]$.*
- *$\gamma_{j,r}^{(t)}$ is a strictly increasing sequence. Besides, $\gamma_{j,r}^{(t)} = \Theta(\|\boldsymbol{\mu}\|_2^2/(d\sigma_p^2)) \sum_{i=1}^{n} \overline{\rho}_{j,r,i}^{(t)}$ for all $j \in \{\pm 1\}$ and $r \in [m]$.*

In Definition 5.1, $\gamma_{j,r}^{(t)}$ characterizes the progress of learning the signal vector $\boldsymbol{\mu}$, and $\rho_{j,r,i}^{(t)}$ characterizes the degree of noise memorization by the filter. The first and second bullets in Proposition 5.2 tell us that for any iteration t, the degree of noise memorization $\overline{\rho}_{j,r,i}^{(t)}, \underline{\rho}_{j,r,i}^{(t)}$ are bounded by a logarithmic order of total epochs T^*. In particular, when $T^* = \eta^{-1}\mathrm{poly}(\epsilon^{-1}, d, n, m)$, $\overline{\rho}_{j,r,i}^{(t)}, \underline{\rho}_{j,r,i}^{(t)} = \widetilde{O}(1)$. The third bullet in Proposition 5.2 is the major improvement of our technique compared to Cao et al. (2022). It shows that $\gamma_{j,r}^{(t)}$ is strictly increasing, indicating that the CNN will learn the signal $\boldsymbol{\mu}$ despite label-flipping noise. Besides, the order of the coefficient ratio $\gamma_{j,r}^{(t)}/(\sum_{i=1}^{n} \overline{\rho}_{j,r,i}^{(t)})$ is time-invariant. When the signal strength $\|\boldsymbol{\mu}\|_2$ is large compared to the noise variance $\sqrt{d}\sigma_p$, the neurons tend to learn the signal.

When $\|\boldsymbol{\mu}\|_2$ is small compared to $\sqrt{d}\sigma_p$, the neurons tend to learn the noises. By the time-invariant coefficient ratio technique, we can characterize the behavior of all the neurons during training, which enables us to deal with the ReLU activation function.

To prove the third bullet, we need to characterize the activation pattern of $\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle)$. Observing that the increment of $\sum_i \overline{\rho}_{i,j,r}^{(t)}$ is scaled by $\sum_i \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle)$, for any weight $\mathbf{w}_{j,r}^{(t)}$, we consider the set sequence $\{S_{j,r}^{(t)}\}_{t=0}^{T^*}$, where $S_{j,r}^{(t)}$ is defined as $\{i | y_i = j, \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle > 0\}$. We show that this is an increasing set sequence throughout the training, leading to $|S_{j,r}^{(t)}| = \Theta(n)$. This intuitively means that for a given sample, once a neuron is activated by the noise patch, it will remain activated throughout training even though the weights of the neuron are updated by gradient descent. Applying this finding to (5.3) and (5.4), it follows directly that the increment ratio of $\gamma_{j,r}^{(t)}$ and $\sum_i \overline{\rho}_{j,r,i}^{(t)}$ will always remain $\Theta(\|\boldsymbol{\mu}\|_2^2/(\sigma_p^2 d))$.

5.2. Key Technique 2: Automatic Balance of Coefficient Updates

Our second main challenge is dealing with label-flipping noise. Empirical studies found that over-parameterized neural networks can generalize well when trained on data with label noise (Belkin et al., 2019; Zhang et al., 2021), which is in conflict with the long-standing theories of statistical learning. To fit corrupted data with signal $-y\boldsymbol{\mu}^\top$ and noise $\boldsymbol{\xi}$, the neural network weights must capture the random noise $\boldsymbol{\xi}$, which harms generalization. Even worse, label-flipping noise may trick the learner into capturing the adversarial signal $-\boldsymbol{\mu}$ rather than $\boldsymbol{\mu}$. Let us investigate the update rule of the coefficient $\gamma_{j,r}, \overline{\rho}_{j,r,i}, \underline{\rho}_{j,r,i}$.

Lemma 5.3. *The coefficients $\gamma_{j,r}^{(t)}, \overline{\rho}_{j,r,i}^{(t)}, \underline{\rho}_{j,r,i}^{(t)}$ defined in Definition 5.1 satisfy the following iterative equations:*

$$\gamma_{j,r}^{(0)}, \overline{\rho}_{j,r,i}^{(0)}, \underline{\rho}_{j,r,i}^{(0)} = 0, \tag{5.2}$$

$$\gamma_{j,r}^{(t+1)} = \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \Bigg[\sum_{i \in S_+} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) - \sum_{i \in S_-} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) \Bigg] \cdot \|\boldsymbol{\mu}\|_2^2, \tag{5.3}$$

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}(y_i = j), \tag{5.4}$$

$$\underline{\rho}_{j,r,i}^{(t+1)} = \underline{\rho}_{j,r,i}^{(t)} + \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}(y_i = -j), \tag{5.5}$$

for all $r \in [m]$, $j \in \{\pm 1\}$ and $i \in [n]$, where $S_+ := \{i \in [n] | y_i = \widehat{y}_i\}$ and $S_- := \{i \in [n] | y_i \neq \widehat{y}_i\}$.

When there is no label-flipping noise, we can conclude that $S_- = \varnothing$ and the signal coefficient $\gamma_{j,r}^{(t)}$ is strictly increasing since $\ell_i'^{(t)}$ is strictly negative. This key observation plays an important role in the proof of Cao et al. (2022). Unfortunately, the presence of noisy labels introduces the presence of a negative term $\sum_{i \in S_-} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle)$. Therefore, we cannot conclude directly from formula (5.2) whether $\gamma_{j,r}^{(t)}$ is increasing or not. If the gradient of losses $\ell_i'^{(t)}$ for (noisy) samples $i \in S_-$ are particularly large relative to the gradient of losses $\ell_i'^{(t)}$ for (clean) samples $i \in S_+$, then indeed (5.2) may fail to guarantee an increase of $\gamma_{j,r}^{(t)}$. In order to show that the neural networks can still learn signals while interpolating the noisy data S_-, we need more advanced and careful characterization of the learning process.

To overcome the difficulty in dealing with label-flipping noise, we apply a key technique called *automatic balance of coefficient updates*. As indicated in (5.3), if we can show that the loss gradients $\ell_i'^{(t)}$ are essentially 'balanced' across all samples, i.e., $\ell_i'^{(t)}/\ell_k'^{(t)} \leq C, \forall i, j \in [n]$, then provided that the fraction of noisy labels is not too large, the effect of the noisy labels will be countered by clean labels, and one can eventually show that $\gamma_{j,r}^{(t)}$ is increasing. This provides motivation for our next lemma.

Lemma 5.4. *Under Condition 4.1, the following bounds hold for any $t \in [0, T^*]$:*

$$y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_k) \leq C_4, \tag{5.6}$$

$$\ell_i'^{(t)}/\ell_k'^{(t)} \leq C_5, \tag{5.7}$$

for any $i, k \in [n]$, where $C_4 = \Theta(1)$ is a positive constant, $C_5 = \exp(C_4)$, and $\ell_i'^{(t)} = \ell'(y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i))$, $\ell_k'^{(t)} = \ell'(y_k f(\mathbf{W}^{(t)}, \mathbf{x}_k))$.

The strategy of bounding $\ell_i'^{(t)}/\ell_k'^{(t)}$ is first proposed by Chatterji & Long (2021) in studying linear classification and has later been extended to neural networks with smoothed leaky ReLU activation function (Frei et al., 2021; 2022). The main idea is that according to the property of logit function $\ell'(z) = -1/(1 + \exp(z))$ that $\ell'(z_1)/\ell'(z_2) \approx \exp(z_2 - z_1)$, to upper bound the ratio of ℓ_i' and ℓ_k', one only needs to bound the difference between $y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)$ and $y_k f(\mathbf{W}^{(t)}, \mathbf{x}_k)$. To further characterize this difference, the works above utilize the smoothness property, translating the function difference to the gradient difference $\nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i)$ and $\nabla f(\mathbf{W}^{(t)}, \mathbf{x}_k)$. However, such a smoothness-based technique cannot be directly applied to ReLU neural networks.

In this paper, we apply signal-noise decomposition and approximate $y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)$ by $\sum_r \overline{\rho}_{y_i,r,i}^{(t)}$ with a small approximation error for any $i \in [n]$. Therefore, Lemma 5.4 can be further simplified into proving the following intermediate result.

Lemma 5.5. *Under Condition 4.1, the following bounds hold for $t \in [0, T^*]$:*

$$\sum_{r=1}^{m} \overline{\rho}_{y_i,r,i}^{(t)} - \sum_{r=1}^{m} \overline{\rho}_{y_k,r,k}^{(t)} \leq \kappa, \tag{5.8}$$

for any $i, k \in [n]$, where $\kappa = \Theta(1)$ is a positive constant.

Note that (5.8) is much easier to deal with than (5.6) because we can directly use the iterative analysis of (5.4), which leads to the update rule:

$$\begin{aligned}\sum_{r=1}^{m} [\overline{\rho}_{y_i,r,i}^{(t+1)} - \overline{\rho}_{y_k,r,k}^{(t+1)}] &= \sum_{r=1}^{m} [\overline{\rho}_{y_i,r,i}^{(t)} - \overline{\rho}_{y_k,r,k}^{(t)}] \\ &- \frac{\eta}{nm} \cdot \big(|S_i^{(t)}| \ell_i'^{(t)} \|\boldsymbol{\xi}_i\|_2^2 - |S_k^{(t)}| \ell_k'^{(t)} \|\boldsymbol{\xi}_i\|_2^2\big),\end{aligned} \tag{5.9}$$

where $S_i^{(t)} = \{r \in [m] : \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle \geq 0\}, i \in [n]$. Now, we consider two cases:

- If $\sum_{r=1}^{m} \overline{\rho}_{y_i,r,i}^{(t)} - \sum_{r=1}^{m} \overline{\rho}_{y_k,r,k}^{(t)}$ is relatively small, we will show that $\sum_{r=1}^{m} \overline{\rho}_{y_i,r,i}^{(t+1)} - \sum_{r=1}^{m} \overline{\rho}_{y_k,r,k}^{(t+1)}$ will not grow too much for small enough step-size η.
- If $\sum_{r=1}^{m} \overline{\rho}_{y_i,r,i}^{(t)} - \sum_{r=1}^{m} \overline{\rho}_{y_k,r,k}^{(t)}$ is relatively large, then it will cause $\ell_i^{(t)}/\ell_k^{(t)}$ to contract because $\ell_i^{(t)}/\ell_k^{(t)}$ can be approximated by $\exp(\sum_{r=1}^{m} \overline{\rho}_{y_k,r,k}^{(t)} - \sum_{r=1}^{m} \overline{\rho}_{y_i,r,i}^{(t)})$. Moreover, since we can prove that $\|\boldsymbol{\xi}_i\|_2^2 \approx \|\boldsymbol{\xi}_k\|_2^2$ and $|S_i^{(t)}|/|S_k^{(t)}| = \Theta(1)$, we have $\sum_{r=1}^{m} \overline{\rho}_{y_i,r,i}^{(t+1)} - \sum_{r=1}^{m} \overline{\rho}_{y_k,r,k}^{(t+1)}$ will decrease according to (5.9).

Combining the two cases, $\sum_{r=1}^{m} \overline{\rho}_{y_i,r,i}^{(t)} - \sum_{r=1}^{m} \overline{\rho}_{y_k,r,k}^{(t)}$ can be upper bounded by a constant, which completes the proof of Lemma 5.5, and the proof of Lemma 5.4 directly follows.

5.3. Key Technique 3: Algorithm-dependent Test Error Analysis

By choosing $\epsilon = 1/(4n)$, Theorem 4.2 gives that $L_S(\mathbf{W}^{(t)}) \leq 1/(4n)$ which further implies that the training error is 0. On the other hand, we know that the Bayes optimal test error is at least p due to the presence of the label-flipping noise. Thus the gap between the test error and training error is at least p, which prevents us from applying commonly-used standard uniform convergence-based bounds (Bartlett et al., 2017; Neyshabur et al., 2018a) or stability-based bounds (Hardt et al., 2016; Mou et al., 2017; Chen et al., 2018). In this paper, we will give an algorithm-dependent test error analysis. First, we can decompose the test error as follows

$$\begin{aligned}&\mathbb{P}\big(y \neq \mathrm{sign}(f(\mathbf{W}^{(t)}, \mathbf{x}))\big) \\ &= p + (1-2p)\mathbb{P}\big(\widehat{y} f(\mathbf{W}^{(t)}, \mathbf{x}) \leq 0\big).\end{aligned} \tag{5.10}$$

With (5.10), the analysis of test error can be reduced to bounding the wrong prediction probability $\mathbb{P}\big(\widehat{y} f(\mathbf{W}^{(t)}, \mathbf{x}) \leq 0\big)$. To achieve this, we need to bound the coefficient order when the training loss converges to ϵ. The following result demonstrates that a constant proportion of $\overline{\rho}_{y_i,r,i}^{(t)}$ will reach constant order at time $T_1 < T^*$.

Lemma 5.6. *Under Condition 4.1, there exists $T_1 = \Theta(\eta^{-1} nm\sigma_p^{-2} d^{-1})$ such that $\overline{\rho}_{y_i,r,i}^{(T_1)} \geq 2$ for all $r \in S_i^{(0)} := \{r \in [m] : \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle > 0\}$ and $i \in [n]$.*

The main idea in proving this lemma is that $\ell_i'^{(t)}$ remain $\Theta(1)$ before time T_1, and the dynamics of the coefficients in (5.4) can be greatly simplified by replacing the $\ell_i'^{(t)}$ factors by their constant lower bounds. After time T_1, by the monotonicity and order of coefficients in Proposition 5.2, we can describe the orders of the coefficients in the following lemma.

Lemma 5.7. *Under Condition 4.1, the following coefficient orders hold for $t \in [T_1, T^*]$:*

- $\sum_{i=1}^{n} \overline{\rho}_{j,r,i}^{(t)} = \Omega(n) = O(n \log(T^*))$ *for any $j \in \{\pm 1\}$ and $r \in [m]$.*
- $\sum_{i=1}^{n} \overline{\rho}_{j,r,i}^{(t)} / \gamma_{j',r'}^{(t)} = \Theta(\mathrm{SNR}^{-2})$ *for any $j, j' \in \{\pm 1\}$ and $r, r' \in [m]$.*
- $\max_{j,r,i} |\underline{\rho}_{j,r,i}^{(t)}| = \max\big\{O\big(\sqrt{\log(mn/\delta)} \cdot \sigma_0 \sigma_p \sqrt{d}\big), O\big(\sqrt{\log(n/\delta)} \log(T^*) \cdot n/\sqrt{d}\big)\big\}$.

By applying the scale of $\gamma_{j,r}^{(t)}, \overline{\rho}_{j,r,i}^{(t)}, \underline{\rho}_{j,r,i}^{(t)}$ given in Lemma 5.7 and Gaussian concentration of Lipschitz function, we can directly get the test error upper bound (the second part of Theorem 4.2) using a similar idea as the proofs of Theorem 1 in Chatterji & Long (2021) and Lemma 3 in Frei et al. (2022). To prove the test error lower bound (the third part of Theorem 4.2), we first lower bound wrong prediction probability term $\mathbb{P}\big(\widehat{y} f(\mathbf{W}^{(t)}, \mathbf{x}) \leq 0\big)$ by

$$0.5\mathbb{P}\Bigg(\underbrace{\Big|\sum_{j,r} j\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle)\Big| \geq C_6 \max_j \Big\{\sum_r \gamma_{j,r}^{(t)}\Big\}}_{\text{event } \Omega}\Bigg),$$

where all the randomness is on the left-hand side, which can be treated as a function of Gaussian random vector $\boldsymbol{\xi}$.

Next, we give our key lemma, which can be proved by leveraging decomposition of $\mathbf{w}_{j,r}^{(t)}$ and scale of decomposition coefficients given in Lemma 5.7.

Lemma 5.8. *For $t \in [T_1, T^*]$, denote $g(\boldsymbol{\xi}) = \sum_{j,r} j\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle)$. There exists a fixed vector $\mathbf{v}$ with $\|\mathbf{v}\|_2 \leq 0.06\sigma_p$ such that*

$$\sum_{j'\in\{\pm 1\}} [g(j'\boldsymbol{\xi} + \mathbf{v}) - g(j'\boldsymbol{\xi})] \geq 4C_6 \max_{j\in\{\pm 1\}} \Big\{ \sum_r \gamma_{j,r}^{(t)} \Big\}, \tag{5.11}$$

for all $\boldsymbol{\xi} \in \mathbb{R}^d$.

Based on Lemma 5.8, by the pigeonhole principle, there must exist one among $\boldsymbol{\xi}$, $\boldsymbol{\xi} + \mathbf{v}$, $-\boldsymbol{\xi}$, $-\boldsymbol{\xi} + \mathbf{v}$ that belongs to Ω, that is, $\Omega \cup (-\Omega) \cup (\Omega - \mathbf{v}) \cup (-\Omega - \mathbf{v}) = \mathbb{R}^d$. By union bound, it follows that

$$\mathbb{P}(\Omega) + \mathbb{P}(-\Omega) + \mathbb{P}(\Omega - \mathbf{v}) + \mathbb{P}(-\Omega - \mathbf{v}) \geq 1. \tag{5.12}$$

Since the noise $\boldsymbol{\xi}$ follows symmetric distribution, we have that $\mathbb{P}(\Omega) = \mathbb{P}(-\Omega)$. We can use some techniques based on the total variation (TV) distance to show that

$$|\mathbb{P}(\Omega) - \mathbb{P}(\Omega - \mathbf{v})|, |\mathbb{P}(-\Omega) - \mathbb{P}(-\Omega - \mathbf{v})| \leq 0.03. \tag{5.13}$$

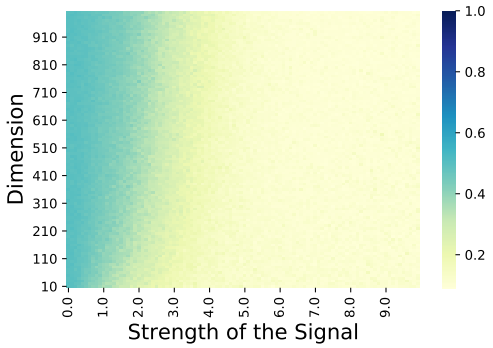
By (5.12) and (5.13), we have proved that $\mathbb{P}(\Omega) \geq 0.22$. By plugging $\mathbb{P}(\Omega) \geq 0.22$ into (5.10), we complete the proof of test error lower bound.

6. Experiments

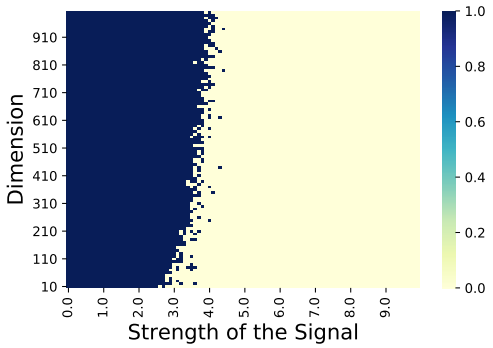
In this section, we present simulations of synthetic data to back up our theoretical analysis in the previous section. The code for our experiments can be found on Github [1].

Synthetic-data experiments. Here we generate synthetic data exactly following Definition 1.1. Specifically, we set training data size $n = 20$ and label-flipping noise to 0.1. Since the learning problem is rotation-invariant, without loss of generality, we set $\boldsymbol{\mu} = \|\boldsymbol{\mu}\|_2 \cdot [1, 0, \ldots, 0]^\top$. We then generate the noise vector $\boldsymbol{\xi}$ from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma_p^2\mathbf{I})$ with fixed standard deviation $\sigma_p = 1$.

We train a two-layer CNN model defined in Section 3 with ReLU activation function. The number of filters is set as $m = 10$. We use the default initialization method in PyTorch to initialize the CNN parameters and train the CNN with full-batch gradient descent with a learning rate of 0.1 for 100 iterations. We consider different dimensions d ranging from 100 to 1100, and different signal strengths $\|\boldsymbol{\mu}\|_2$ ranging from 1 to 11. Based on our results, for any dimension d and signal strength μ setting we consider, our training setup can guarantee a training loss smaller than 0.01. After training, we estimate the test error for each case using 1000 test data points. The results are given as a heatmap on parameters d and $\|\boldsymbol{\mu}\|_2$ in Figure 1.



(a) Original Test Error Heatmap



(b) Cutoff Test Error Heatmap

Figure 1. a) is a heatmap of test error on synthetic data under different dimensions d and signal strengths $\boldsymbol{\mu}$. High test errors are marked in blue, and low test errors are marked in yellow. b) is a cutoff value heatmap that sets the values smaller than 0.2 to be 0 (yellow) and the values greater than 0.2 to be 1 (blue).

For the specific case $\|\boldsymbol{\mu}\|_2 = 5$ and $d = 100$, we plot the training loss, test loss, and test error throughout training in Figure 2. As we can see from the figure, the test error reaches the Bayesian optimal error of 0.1, while the training loss converges to zero.

In Section 5, we directly used the activation pattern and data structure to characterize the loss of each sample, and proved that $y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_k) \leq C_4$ for $t \leq T^*$ and any $i, k \in [n]$ in Lemma 5.7. To demonstrate this, we conduct another experiment for the case $\|\boldsymbol{\mu}\|_2 = 5$, $d = 100$ and plot $\max y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)$ and $\min y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)$ (margin) for each iteration. As we can see from Figure 3, the difference between them never grows too large during training (bounded by 6).

[1] https://github.com/uclaml/Benign_ReLU_CNN

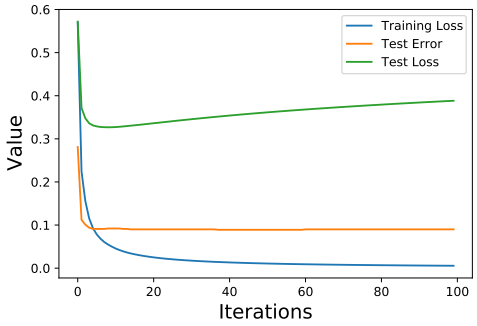


Figure 2. Training loss, test loss and test error throughout 100 iterations with $\|\boldsymbol{\mu}\|_2 = 5$ and $d = 100$.

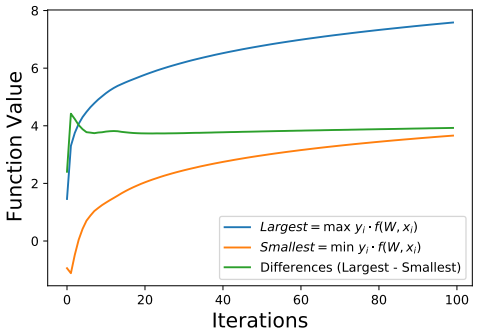


Figure 3. $\max y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)$ and $\min y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)$ (margin) throughout 100 iterations with $\|\boldsymbol{\mu}\|_2 = 5$ and $d = 100$.

7. Conclusion and Future Work

This paper studies benign overfitting in two-layer ReLU CNNs with label-flipping noise. We generalize the signal-noise decomposition technique first proposed by Cao et al. (2022) and propose three key techniques: *time-invariant coefficient ratio*, *automatic balance of coefficient updates* and *algorithm-dependent test error analysis*. With the help of these techniques, we prove the convergence of training loss, give exact conditions under which the CNN achieves test error close to the noise rate, and reveal a sharp phase transition between benign and harmful overfitting. Our results theoretically demonstrate how and when benign overfitting can happen in ReLU neural networks. An important future work direction is to generalize our analysis to deep ReLU neural networks in learning other data models.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. YK, ZC, YC and QG are supported in part by the National Science Foundation CAREER Award 1906169 and IIS-2008981, and the Sloan Research Fellowship. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Adlam, B. and Pennington, J. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pp. 74–84. PMLR, 2020.

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

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

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6240–6249, 2017.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pp. 540–548, 2018.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Cao, Y., Gu, Q., and Belkin, M. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Chatterji, N. S. and Long, P. M. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22:129–1, 2021.

Chatterji, N. S. and Long, P. M. Deep linear networks can benignly overfit when shallow ones do. *arXiv preprint arXiv:2209.09315*, 2022.

Chen, Y., Jin, C., and Yu, B. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

Devroye, L., Mehrabian, A., and Reddad, T. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

Frei, S. and Gu, Q. Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. *Advances in Neural Information Processing Systems*, 34:7937–7949, 2021.

Frei, S., Cao, Y., and Gu, Q. Provable generalization of sgd-trained neural networks of any width in the presence of adversarial label noise. In *International Conference on Machine Learning*, pp. 3427–3438. PMLR, 2021.

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

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pp. 1225–1234. JMLR. org, 2016.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Li, Y., Wei, C., and Ma, T. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, pp. 11669–11680, 2019.

Li, Z., Zhou, Z.-H., and Gretton, A. Towards an understanding of benign overfitting in neural networks. *arXiv preprint arXiv:2106.03212*, 2021.

Liang, T. and Rakhlin, A. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Liang, T., Rakhlin, A., and Zhai, X. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pp. 2683–2711. PMLR, 2020.

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

Montanari, A. and Zhong, Y. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816–2847, 2022.

Mou, W., Wang, L., Zhai, X., and Zheng, K. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. *arXiv preprint arXiv:1707.05947*, 2017.

Muthukumar, V., Narang, A., Subramanian, V., Belkin, M., Hsu, D., and Sahai, A. Classification vs regression in overparameterized regimes: Does the loss function matter? *The Journal of Machine Learning Research*, 22(1): 10104–10172, 2021.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representation*, 2018a.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018b.

Shamir, O. The implicit bias of benign overfitting. *arXiv preprint arXiv:2201.11489*, 2022.

Shen, R. and Bubeck, S. Data augmentation as feature manipulation: a story of desert cows and grass cows. *ArXivorg*, 2022.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Wang, K. and Thrampoulidis, C. Benign overfitting in binary classification of gaussian mixtures. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4030–4034. IEEE, 2021.

Wu, D. and Xu, J. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33, 2020.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

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

Zou, D., Wu, J., Braverman, V., Gu, Q., and Kakade, S. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on Learning Theory*, pp. 4633–4635. PMLR, 2021b.

A. Comparison of Conditions Made by Related Works

In this section, we present the difference between Condition 4.1 and the conditions on parameters made by two related works (Frei et al., 2022; Cao et al., 2022) in the following two tables (Tables 1 and 2).

Number of samples	Frei et al. (2022)	$n \geq C\log(1/\delta)$
	Ours	$n \geq C\log(m/\delta)$
Neural network width	Frei et al. (2022)	-
	Ours	$m \geq C\log(n/\delta)$
Dimension	Frei et al. (2022)	$d \geq \max\{n\lVert\boldsymbol{\mu}\rVert_2^2, n^2\log(n/\delta)\}$
	Ours	$d \geq C\max\{n\sigma_p^{-2}\lVert\boldsymbol{\mu}\rVert_2^2\log(T^*), n^2\log(nm/\delta)(\log(T^*))^2\}$
Norm of the signal	Frei et al. (2022)	$\lVert\boldsymbol{\mu}\rVert_2^2 \geq C\cdot\log(n/\delta)$
	Ours	$\lVert\boldsymbol{\mu}\rVert_2^2 \geq C\cdot\sigma_p^2\log(n/\delta)$
Noise rate	Frei et al. (2022)	$p \leq 1/C$
	Ours	$p \leq 1/C$
Learning rate	Frei et al. (2022)	$\eta \leq (C\max\{1, H/\sqrt{m}\}d^2)^{-1}$
	Ours	$\eta \leq \left(C\max\left\{\sigma_p^2 d/n, \sigma_p^2 d^{3/2}/\left(n^2 m\cdot\sqrt{\log(n/\delta)}\right)\right\}\right)^{-1}$
Initialization variance	Frei et al. (2022)	$\sigma_0 \leq \eta/\sqrt{md}$
	Ours	$\sigma_0 \leq \left(C\max\left\{\sigma_p d/\sqrt{n}, \sqrt{\log(m/\delta)}\cdot\lVert\boldsymbol{\mu}\rVert_2\right\}\right)^{-1}$

Table 1. Comparison of conditions with Frei et al. (2022). H is the smoothness of leaky ReLU activation under the setting of Frei et al. (2022). In our paper, σ_p is the noise scale that can be treated as a constant.

Number of samples	Cao et al. (2022)	$n = \Omega(\text{polylog}(d))$
	Ours	$n \geq C\log(m/\delta)$
Neural network width	Cao et al. (2022)	$m = \Omega(\text{polylog}(d))$
	Ours	$m \geq C\log(n/\delta)$
Dimension	Cao et al. (2022)	$d = \Omega(m^{2\vee[4/(q-2)]}n^{4\vee[(2q-2)/(q-2)]})$
	Ours	$d \geq C\max\{n\sigma_p^{-2}\lVert\boldsymbol{\mu}\rVert_2^2\log(T^*), n^2\log(nm/\delta)(\log(T^*))^2\}$
Norm of the signal	Cao et al. (2022)	-
	Ours	$\lVert\boldsymbol{\mu}\rVert_2^2 \geq C\cdot\sigma_p^2\log(n/\delta)$
Noise rate	Cao et al. (2022)	$p = 0$
	Ours	$p \leq 1/C$
Learning rate	Cao et al. (2022)	$\eta \leq \widetilde{O}(\min\{\lVert\boldsymbol{\mu}\rVert_2^{-2}, \sigma_p^{-2}d^{-1}\})$
	Ours	$\eta \leq \left(C\max\left\{\sigma_p^2 d/n, \sigma_p^2 d^{3/2}/\left(n^2 m\cdot\sqrt{\log(n/\delta)}\right)\right\}\right)^{-1}$
Initialization variance	Cao et al. (2022)	$\sigma_0 \leq \widetilde{O}(m^{-2/(q-2)}n^{-[1/(q-2)\vee 1]})\cdot\min\{(\sigma_p\sqrt{d})^{-1}, \lVert\boldsymbol{\mu}\rVert_2^{-1}\}$, $\sigma_0 \geq \widetilde{O}(nd^{-1/2})\cdot\min\{(\sigma_p\sqrt{d})^{-1}, \lVert\boldsymbol{\mu}\rVert_2^{-1}\}$
	Ours	$\sigma_0 \leq \left(C\max\left\{\sigma_p d/\sqrt{n}, \sqrt{\log(m/\delta)}\cdot\lVert\boldsymbol{\mu}\rVert_2\right\}\right)^{-1}$

Table 2. Comparison of conditions with Cao et al. (2022). q is the order of polynomial ReLU activation function under the setting of Cao et al. (2022).

B. Preliminary Lemmas

In this section, we present some pivotal lemmas that illustrate some important properties of the data and neural network parameters at their random initialization.

We first give some concentration lemmas regarding the data set S. Let $S_+ = \{i | y_i = \widehat{y}_i\}$ and $S_- = \{i | y_i \neq \widehat{y}_i\}$ denote index sets corresponding to data points with true and flipped labels, respectively. We first have the following lemma.

Lemma B.1. *Given $\delta > 0$, with probability at least $1-\delta$,*

$$\big||S_+| - (1-p)n\big| \leq \sqrt{\frac{n}{2}\log\left(\frac{4}{\delta}\right)},\ \big||S_-| - pn\big| \leq \sqrt{\frac{n}{2}\log\left(\frac{4}{\delta}\right)}.$$

Proof of Lemma B.1. Since $|S_+| = \sum_{i=1}^n \mathbb{1}[\widehat{y}_i = y_i]$, $|S_-| = \sum_{i=1}^n \mathbb{1}[\widehat{y}_i \neq y_i]$, according to Hoeffding's inequality, we have for arbitrary $t > 0$ that

$$\mathbb{P}\big(\big||S_+| - \mathbb{E}[|S_+|]\big| \geq t\big) \leq 2\exp\Big(-\frac{2t^2}{n}\Big), \mathbb{P}\big(\big||S_-| - \mathbb{E}[|S_-|]\big| \geq t\big) \leq 2\exp\Big(-\frac{2t^2}{n}\Big).$$

By the data distribution $\mathcal{D}$ defined in Definition 1.1, we have $\mathbb{E}[S_+] = (1-p)n$, $\mathbb{E}[S_-] = pn$. Setting $t = \sqrt{(n/2)\log(4/\delta)}$ and taking a union bound, it follows that with probability at least $1-\delta$,

$$\big||S_+| - (1-p)n\big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{4}{\delta}\Big)}, \big||S_-| - pn\big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{4}{\delta}\Big)},$$

which completes the proof. □

Next, let $S_1 = \{i | y_i = 1\}$ and $S_{-1} = \{i | y_i = -1\}$. We have the following lemmas characterizing their sizes.

Lemma B.2. *Suppose that $\delta > 0$ and $n \geq 8\log(4/\delta)$. Then with probability at least $1-\delta$,*

$$|S_1|, |S_{-1}| \in [n/4, 3n/4].$$

Proof of Lemma B.2. According the data distribution $\mathcal{D}$ defined in Definition 1.1, for $(\mathbf{x}, y) \sim \mathcal{D}$, we have

$$\begin{aligned}\mathbb{P}(y=1) &= \mathbb{P}(\widehat{y}=1) \times \mathbb{P}(y=\widehat{y}) + \mathbb{P}(\widehat{y}=-1) \times \mathbb{P}(y=-\widehat{y}) \\ &= \frac{1}{2}(1-p) + \frac{1}{2}p \\ &= \frac{1}{2},\end{aligned}$$

and hence $\mathbb{P}(y=-1) = 1/2$ as well. Since $|S_1| = \sum_{i=1}^n \mathbb{1}[y_i = 1]$, $|S_{-1}| = \sum_{i=1}^n \mathbb{1}[y_i = -1]$, we have $\mathbb{E}[|S_1|] = \mathbb{E}[|S_{-1}|] = n/2$. By Hoeffding's inequality, for arbitrary $t > 0$ the following holds:

$$\begin{aligned}&\mathbb{P}\big(\big||S_1| - \mathbb{E}[|S_1|]\big| \geq t\big) \leq 2\exp\Big(-\frac{2t^2}{n}\Big), \\ &\mathbb{P}\big(\big||S_{-1}| - \mathbb{E}[|S_{-1}|]\big| \geq t\big) \leq 2\exp\Big(-\frac{2t^2}{n}\Big).\end{aligned}$$

Setting $t = \sqrt{(n/2)\log(4/\delta)}$ and taking a union bound, it follows that with probability at least $1-\delta$,

$$\Big||S_1| - \frac{n}{2}\Big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{4}{\delta}\Big)}, \Big||S_{-1}| - \frac{n}{2}\Big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{4}{\delta}\Big)}.$$

Therefore, as long as $n \geq 8\log(4/\delta)$, we have $\sqrt{n\log(4/\delta)/2} \leq n/4$ and hence $3n/4 \geq |S_1|, |S_{-1}| \geq n/4$.

□

Lemma B.3. *For $|S_+ \cap S_y|$ and $|S_- \cap S_y|$ where $y \in \{\pm 1\}$, it holds with probability at least $1-\delta(\delta > 0)$ that*

$$\Big||S_+ \cap S_y| - \frac{(1-p)n}{2}\Big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{8}{\delta}\Big)}, \Big||S_- \cap S_y| - \frac{pn}{2}\Big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{8}{\delta}\Big)}, \forall\, y \in \{\pm 1\}.$$

Proof. Since $|S_+ \cap S_y| = \sum_{i=1}^n \mathbb{1}[\widehat{y}_i = y_i = y]$, $|S_- \cap S_y| = \sum_{i=1}^n \mathbb{1}[\widehat{y}_i \neq y_i, y_i = y]$, according to Hoeffding's inequality, we have

$$\mathbb{P}\big(\big||S_+ \cap S_y| - \mathbb{E}[|S_+ \cap S_y|]\big| \geq t\big) \leq 2\exp\Big(-\frac{2t^2}{n}\Big), \forall\, y \in \{\pm 1\},$$

$$\mathbb{P}\big(\big||S_-\cap S_y| - \mathbb{E}[|S_-\cap S_y|]\big| \geq t\big) \leq 2\exp\Big(-\frac{2t^2}{n}\Big),\ \forall\, y\in\{\pm 1\}.$$

According to the definition of $\mathcal{D}$ in Definition 1.1, we have $\mathbb{E}[|S_+\cap S_y|] = (1-p)n/2$, $\mathbb{E}[|S_-\cap S_y|] = pn/2$. It follows with probability at least $1-\delta$ that

$$\Big||S_+\cap S_y| - \frac{(1-p)n}{2}\Big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{8}{\delta}\Big)},\Big||S_-\cap S_y| - \frac{pn}{2}\Big| \leq \sqrt{\frac{n}{2}\log\Big(\frac{8}{\delta}\Big)},\forall\, y\in\{\pm 1\},$$

which completes the proof. □

The following lemma estimates the norms of the noise vectors $\boldsymbol{\xi}_i$, $i\in[n]$, and gives an upper bound of their inner products with each other and with the signal vector $\boldsymbol{\mu}$.

Lemma B.4. *Suppose that $\delta>0$ and $d=\Omega(\log(6n/\delta))$. Then with probability at least $1-\delta$,*

$$\begin{aligned}
&\sigma_p^2 d/2 \leq \|\boldsymbol{\xi}_i\|_2^2 \leq 3\sigma_p^2 d/2,\\
&|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle| \leq 2\sigma_p^2\cdot\sqrt{d\log(6n^2/\delta)},\\
&|\langle \boldsymbol{\xi}_i, \boldsymbol{\mu}\rangle| \leq \|\boldsymbol{\mu}\|_2\sigma_p\cdot\sqrt{2\log(6n/\delta)}
\end{aligned}$$

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

Proof of Lemma B.4. By Bernstein's inequality, with probability at least $1-\delta/(3n)$ we have

$$\big|\|\boldsymbol{\xi}_i\|_2^2 - \sigma_p^2 d\big| = O(\sigma_p^2\cdot\sqrt{d\log(6n/\delta)}).$$

Therefore, if we set appropriately $d=\Omega(\log(6n/\delta))$, we get

$$\sigma_p^2 d/2 \leq \|\boldsymbol{\xi}_i\|_2^2 \leq 3\sigma_p^2 d/2.$$

Moreover, clearly $\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle$ has mean zero. For any i,i' with $i\neq i'$, by Bernstein's inequality, with probability at least $1-\delta/(3n^2)$ we have

$$|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle| \leq 2\sigma_p^2\cdot\sqrt{d\log(6n^2/\delta)}.$$

Finally, note that $\langle \boldsymbol{\xi}_i, \boldsymbol{\mu}\rangle \sim \mathcal{N}(0, \|\boldsymbol{\mu}\|_2^2\sigma_p^2)$. By Gaussian tail bounds, with probability at least $1-\delta/3n$ we have

$$|\langle \boldsymbol{\xi}_i, \boldsymbol{\mu}\rangle| \leq \|\boldsymbol{\mu}\|_2\sigma_p\cdot\sqrt{2\log(6n/\delta)}.$$

Applying a union bound completes the proof. □

Now turning to network initialization, the following lemma studies the inner product between a randomly initialized CNN convolutional filter $\mathbf{w}_{j,r}^{(0)}$ ($j\in\{\pm 1\}$ and $r\in[m]$) and the signal/noise vectors in the training data. The calculations characterize how the neural network at initialization randomly captures signal and noise information.

Lemma B.5. *Suppose that $d=\Omega(\log(mn/\delta))$, $m=\Omega(\log(1/\delta))$. Then with probability at least $1-\delta$,*

$$\begin{aligned}
&\sigma_0^2 d/2 \leq \|\mathbf{w}_{j,r}^{(0)}\|_2^2 \leq 3\sigma_0^2 d/2,\\
&|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle| \leq \sqrt{2\log(12m/\delta)}\cdot\sigma_0\|\boldsymbol{\mu}\|_2,\\
&|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle| \leq 2\sqrt{\log(12mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}
\end{aligned}$$

for all $r\in[m]$, $j\in\{\pm 1\}$ and $i\in[n]$. Moreover,

$$\sigma_0\|\boldsymbol{\mu}\|_2/2 \leq \max_{r\in[m]} j\cdot\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle \leq \sqrt{2\log(12m/\delta)}\cdot\sigma_0\|\boldsymbol{\mu}\|_2,$$

$$\sigma_0\sigma_p\sqrt{d}/4 \leq \max_{r\in[m]} j\cdot\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle \leq 2\sqrt{\log(12mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}$$

for all $j\in\{\pm 1\}$ *and* $i\in[n]$.

Proof of Lemma B.5. First of all, the initial weights $\mathbf{w}_{j,r}^{(0)}\sim\mathcal{N}(\mathbf{0},\sigma_0\mathbf{I})$. By Bernstein's inequality, with probability at least $1-\delta/(6m)$ we have

$$\big|\|\mathbf{w}_{j,r}^{(0)}\|_2^2-\sigma_0^2 d\big| = O(\sigma_0^2\cdot\sqrt{d\log(12m/\delta)}).$$

Therefore, if we set appropriately $d=\Omega(\log(mn/\delta))$, we have with probability at least $1-\delta/3$, for all $j\in\{\pm1\}$ and $r\in[m]$,

$$\sigma_0^2 d/2 \leq \|\mathbf{w}_{j,r}^{(0)}\|_2^2 \leq 3\sigma_0^2 d/2.$$

Next, it is clear that for each $r\in[m]$, $j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}\rangle$ is a Gaussian random variable with mean zero and variance $\sigma_0^2\|\boldsymbol{\mu}\|_2^2$. Therefore, by Gaussian tail bound and union bound, with probability at least $1-\delta/6$, for all $j\in\{\pm1\}$ and $r\in[m]$,

$$j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}\rangle \leq |\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}\rangle| \leq \sqrt{2\log(12m/\delta)}\cdot\sigma_0\|\boldsymbol{\mu}\|_2.$$

Moreover, $\mathbb{P}(\sigma_0\|\boldsymbol{\mu}\|_2/2 > j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}\rangle)$ is an absolute constant, and therefore with the condition $m=\Omega(\log(1/\delta))$, we have

$$\begin{aligned}\mathbb{P}\big(\sigma_0\|\boldsymbol{\mu}\|_2/2 \leq \max_{r\in[m]} j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}\rangle\big) &= 1-\mathbb{P}\big(\sigma_0\|\boldsymbol{\mu}\|_2/2 > \max_{r\in[m]} j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}\rangle\big)\\ &= 1-\mathbb{P}\big(\sigma_0\|\boldsymbol{\mu}\|_2/2 > j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}\rangle\big)^{2m}\\ &\geq 1-\delta/6,\end{aligned}$$

hence with probability at least $1-\delta/3$, we have $\sigma_0\|\boldsymbol{\mu}\|_2/2 \leq \max_{r\in[m]} j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}\rangle \leq \sqrt{2\log(12m/\delta)}\cdot\sigma_0\|\boldsymbol{\mu}\|_2$.

Finally, under the results of Lemma B.4, we have $\sigma_p\sqrt{d}/\sqrt{2}\leq\|\boldsymbol{\xi}_i\|_2\leq\sqrt{3/2}\cdot\sigma_p\sqrt{d}$ for all $i\in[n]$. Therefore, we can get the result for $\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\xi}_i\rangle$ with probability at least $1-\delta/3$, following the same proof outline as $j\cdot\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}\rangle$. □

Next, we denote $S_i^{(0)}$ as $\{r\in[m]:\langle\mathbf{w}_{y_i,r}^{(0)},\boldsymbol{\xi}_i\rangle>0\}$ and $S_{j,r}^{(t)}$ as $\{i\in[n]:y_i=j,\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}_i\rangle>0\}$, $j\in\{\pm1\}$, $r\in[m]$. We give a lower bound of $|S_i^{(0)}|$ and $|S_{j,r}^{(0)}|$ in the following two lemmas.

Lemma B.6. *Suppose that* $\delta>0$ *and* $m\geq 50\log(2n/\delta)$. *Then with probability at least* $1-\delta$,

$$|S_i^{(0)}|\geq 0.4m,\ \forall i\in[n].$$

Proof of Lemma B.6. Note that $|S_i^{(0)}|=\sum_{r=1}^m \mathbb{1}[\langle\mathbf{w}_{y_i,r}^{(0)},\boldsymbol{\xi}_i\rangle>0]$ and $P(\langle\mathbf{w}_{y_i,r}^{(0)},\boldsymbol{\xi}_i\rangle>0)=1/2$, then by Hoeffding's inequality, with probability at least $1-\delta/n$, we have

$$\left|\frac{|S_i^{(0)}|}{m}-\frac{1}{2}\right|\leq\sqrt{\frac{\log(2n/\delta)}{2m}}.$$

Therefore, as long as $m\geq 50\log(2n/\delta)$, by applying union bound, with probability at least $1-\delta$, we have

$$|S_i^{(0)}|\geq 0.4m,\ \forall i\in[n].$$

□

Lemma B.7. *Suppose that $\delta > 0$ and $n \geq 32\log(4m/\delta)$. Then with probability at least $1-\delta$,*

$$|S_{j,r}^{(0)}| \geq n/8,\ \forall j \in \{\pm 1\}, r \in [m].$$

Proof of Lemma B.7. Note that $|S_{j,r}^{(0)}| = \sum_{i=1}^{n} \mathbb{1}[y_i = j]\,\mathbb{1}[\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle > 0]$ and $\mathbb{P}(y_i = j, \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle > 0) = 1/4$, then by Hoeffding's inequality, with probability at least $1-\delta/2m$, we have

$$\big||S_{j,r}^{(0)}|/n - 1/4\big| \leq \sqrt{\frac{\log(4m/\delta)}{2n}}.$$

Therefore, as long as $n \geq 32\log(4m/\delta)$, by applying union bound, we have with probability at least $1-\delta$,

$$|S_{j,r}^{(0)}| \geq n/8,\ \forall j \in \{\pm 1\}, r \in [m].$$

□

C. Signal-noise Decomposition Coefficient Analysis

In this section, we establish a series of results on the signal-noise decomposition. These results are based on the conclusions in Appendix B, which hold with high probability. Denote by $\mathcal{E}_{\text{prelim}}$ the event that all the results in Appendix B hold (for a given δ, we see $\mathbb{P}(\mathcal{E}_{\text{prelim}}) \geq 1 - 7\delta$ by a union bound). For simplicity and clarity, we state all the results in this and the following sections conditional on $\mathcal{E}_{\text{prelim}}$.

C.1. Iterative Expression for Decomposition Coefficients

We begin by analyzing the coefficients in the signal-noise decomposition in Definition 5.1. The first lemma presents an iterative expression for the coefficients.

Lemma C.1. *(Restatement of Lemma 5.3) The coefficients $\gamma_{j,r}^{(t)}, \overline{\rho}_{j,r,i}^{(t)}, \underline{\rho}_{j,r,i}^{(t)}$ defined in Definition 5.1 satisfy the following iterative equations:*

$$\begin{aligned}
&\gamma_{j,r}^{(0)}, \overline{\rho}_{j,r,i}^{(0)}, \underline{\rho}_{j,r,i}^{(0)} = 0,\\
&\gamma_{j,r}^{(t+1)} = \gamma_{j,r}^{(t)} - \frac{\eta}{nm}\cdot\bigg[\sum_{i\in S_+} \ell_i'^{(t)}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\cdot\boldsymbol{\mu}\rangle) - \sum_{i\in S_-} \ell_i'^{(t)}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\cdot\boldsymbol{\mu}\rangle)\bigg]\cdot\|\boldsymbol{\mu}\|_2^2,\\
&\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm}\cdot \ell_i'^{(t)}\cdot\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\cdot\|\boldsymbol{\xi}_i\|_2^2\cdot\mathbb{1}(y_i = j),\\
&\underline{\rho}_{j,r,i}^{(t+1)} = \underline{\rho}_{j,r,i}^{(t)} + \frac{\eta}{nm}\cdot \ell_i'^{(t)}\cdot\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\cdot\|\boldsymbol{\xi}_i\|_2^2\cdot\mathbb{1}(y_i = -j),
\end{aligned}$$

for all $r \in [m]$, $j \in \{\pm 1\}$ and $i \in [n]$.

Proof of Lemma C.1. First, we iterate the gradient descent update rule (3.1) t times and get

$$\begin{aligned}
\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(0)} &- \frac{\eta}{nm}\sum_{s=0}^{t}\sum_{i=1}^{n} \ell_i'^{(s)}\cdot\sigma'(\langle \mathbf{w}_{j,r}^{(s)}, \boldsymbol{\xi}_i\rangle)\cdot j y_i \boldsymbol{\xi}_i\\
&- \frac{\eta}{nm}\sum_{s=0}^{t}\sum_{i=1}^{n} \ell_i'^{(s)}\cdot\sigma'(\langle \mathbf{w}_{j,r}^{(s)}, \widehat{y}_i\boldsymbol{\mu}\rangle)\cdot \widehat{y}_i y_i j\boldsymbol{\mu}.
\end{aligned}$$

According to the definition of $\gamma_{j,r}^{(t)}$ and $\rho_{j,r,i}^{(t)}$,

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

Note that $\boldsymbol{\xi}_i$ and $\boldsymbol{\mu}$ are linearly independent with probability 1, under which condition we have the unique representation

$$\gamma_{j,r}^{(t)} = -\frac{\eta}{nm}\sum_{s=0}^{t}\sum_{i=1}^{n}\ell_i'^{(s)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(s)},\widehat{y}_i\boldsymbol{\mu}\rangle)\cdot\|\boldsymbol{\mu}\|_2^2\cdot\widehat{y}_i y_i,$$

$$\rho_{j,r,i}^{(t)} = -\frac{\eta}{nm}\sum_{s=0}^{t}\ell_i'^{(s)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(s)},\boldsymbol{\xi}_i\rangle)\cdot\|\boldsymbol{\xi}_i\|_2^2\cdot jy_i.$$

Recall $S_+ = \{i|y_i = \widehat{y}_i\}$, $S_- = \{i|y_i \neq \widehat{y}_i\}$, we can further write

$$\gamma_{j,r}^{(t)} = -\frac{\eta}{nm}\sum_{s=0}^{t}\sum_{i\in S_+}\ell_i'^{(s)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(s)},\widehat{y}_i\boldsymbol{\mu}\rangle)\cdot\|\boldsymbol{\mu}\|_2^2 + \frac{\eta}{nm}\sum_{s=0}^{t}\sum_{i\in S_-}\ell_i'^{(s)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(s)},\widehat{y}_i\boldsymbol{\mu}\rangle)\cdot\|\boldsymbol{\mu}\|_2^2. \tag{C.1}$$

Now with the notation $\overline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)}\mathbb{1}(\rho_{j,r,i}^{(t)} \geq 0)$, $\underline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)}\mathbb{1}(\rho_{j,r,i}^{(t)} \leq 0)$ and the fact $\ell_i'^{(s)} < 0$, we get

$$\overline{\rho}_{j,r,i}^{(t)} = -\frac{\eta}{nm}\sum_{s=0}^{t}\ell_i'^{(s)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(s)},\boldsymbol{\xi}_i\rangle)\cdot\|\boldsymbol{\xi}_i\|_2^2\cdot\mathbb{1}(y_i = j), \tag{C.2}$$

$$\underline{\rho}_{j,r,i}^{(t)} = \frac{\eta}{nm}\sum_{s=0}^{t}\ell_i'^{(s)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(s)},\boldsymbol{\xi}_i\rangle)\cdot\|\boldsymbol{\xi}_i\|_2^2\cdot\mathbb{1}(y_i = -j). \tag{C.3}$$

Writing out the iterative versions of (C.1), (C.2) and (C.3) completes the proof. □

C.2. Scale of Decomposition Coefficients

The rest of this section will be dedicated to the proof of the following Proposition C.2, which shows that the coefficients in the signal-noise decomposition will stay within a reasonable range for a considerable amount of time. Consider the training period $0 \leq t \leq T^*$, where $T^* = \eta^{-1}\mathrm{poly}(\epsilon^{-1}, d, n, m)$, as defined in Theorem 4.2, is the maximum admissible iteration. Now denote

$$\alpha := 4\log(T^*), \tag{C.4}$$

$$\beta := 2\max_{i,j,r}\{|\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}\rangle|, |\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\xi}_i\rangle|\}, \tag{C.5}$$

$$\mathrm{SNR} := \|\boldsymbol{\mu}\|_2/(\sigma_p\sqrt{d}). \tag{C.6}$$

By Lemma B.5, β can be bounded by $4\sigma_0\cdot\max\{\sqrt{\log(12mn/\delta)}\cdot\sigma_p\sqrt{d}, \sqrt{\log(12m/\delta)}\cdot\|\boldsymbol{\mu}\|_2\}$. Then, by Condition 4.1, by choosing a large constant C, it is straightforward to verify the following inequality:

$$\max\left\{\beta, \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha, 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha\right\} \leq \frac{1}{12}. \tag{C.7}$$

Proposition C.2. *(Partial restatement of Proposition 5.2) Under Condition 4.1, for $0 \leq t \leq T^*$, we have that*

$$\gamma_{j,r}^{(0)}, \overline{\rho}_{j,r,i}^{(0)}, \underline{\rho}_{j,r,i}^{(0)} = 0 \tag{C.8}$$

$$0 \leq \overline{\rho}_{j,r,i}^{(t)} \leq \alpha, \tag{C.9}$$

$$0 \geq \underline{\rho}_{j,r,i}^{(t)} \geq -\beta - 10\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha \geq -\alpha, \tag{C.10}$$

and there exists a positive constant C' such that

$$0 \leq \gamma_{j,r}^{(t)} \leq C'\widehat{\gamma}\alpha, \tag{C.11}$$

for all $r \in [m]$, $j \in \{\pm 1\}$ *and* $i \in [n]$, *where* $\widehat{\gamma} := n \cdot \mathrm{SNR}^2$. *Besides,* $\gamma_{j,r}^{(t)}$ *is non-decreasing for* $0 \le t \le T^*$.

We will use induction to prove Proposition C.2. We first introduce several technical lemmas (Lemmas C.3, C.4 and C.5) that will be used for the inductive proof of Proposition C.2.

Lemma C.3. *Under Condition 4.1, suppose* (C.9), (C.10) *and* (C.11) *hold at iteration* t. *Then, for all* $r \in [m]$, $j \in \{\pm 1\}$ *and* $i \in [n]$,

$$\big|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle - j \cdot \gamma_{j,r}^{(t)}\big| \le \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha, \tag{C.12}$$

$$\big|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle - \underline{\rho}_{j,r,i}^{(t)}\big| \le 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,\ j \ne y_i, \tag{C.13}$$

$$\big|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle - \overline{\rho}_{j,r,i}^{(t)}\big| \le 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,\ j = y_i. \tag{C.14}$$

Proof of Lemma C.3. First, for any time $t \ge 0$, we have from the signal-noise decomposition (5.1) that

$$\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle = j \cdot \gamma_{j,r}^{(t)} + \sum_{i'=1}^{n} \overline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\mu}\rangle + \sum_{i'=1}^{n} \underline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\mu}\rangle$$

According to Lemma B.4, we have

$$\begin{aligned}
&\bigg|\sum_{i'=1}^{n} \overline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\mu}\rangle + \sum_{i'=1}^{n} \underline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\mu}\rangle\bigg| \\
&\le \sum_{i'=1}^{n} |\overline{\rho}_{j,r,i'}^{(t)}| \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\mu}\rangle| + \sum_{i'=1}^{n} |\underline{\rho}_{j,r,i'}^{(t)}| \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\mu}\rangle| \\
&\le \frac{2\|\boldsymbol{\mu}\|_2\sqrt{2\log(6n/\delta)}}{\sigma_p d}\bigg(\sum_{i'=1}^{n} |\overline{\rho}_{j,r,i'}^{(t)}| + \sum_{i'=1}^{n} |\underline{\rho}_{j,r,i'}^{(t)}|\bigg) \\
&= \mathrm{SNR}\sqrt{\frac{8\log(6n/\delta)}{d}}\bigg(\sum_{i'=1}^{n} |\overline{\rho}_{j,r,i'}^{(t)}| + \sum_{i'=1}^{n} |\underline{\rho}_{j,r,i'}^{(t)}|\bigg) \\
&\le \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha,
\end{aligned}$$

where the first inequality is by triangle inequality, the second inequality is by Lemma B.4, the equality is by the definition of $\mathrm{SNR} = \|\boldsymbol{\mu}\|_2/(\sigma_p\sqrt{d})$, and the last inequality is by (C.9), (C.10). It follows that

$$\big|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle - j \cdot \gamma_{j,r}^{(t)}\big| \le \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha.$$

Second, for $j \ne y_i$ and any $t \ge 0$, we have $\overline{\rho}_{j,r,i}^{(t)} = 0$, and so

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle &= j \cdot \gamma_{j,r}^{(t)} \|\boldsymbol{\mu}\|_2^{-2} \cdot \langle \boldsymbol{\mu}, \boldsymbol{\xi}_i\rangle + \sum_{i'=1}^{n} \overline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle + \sum_{i'=1}^{n} \underline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle \\
&= \underline{\rho}_{j,r,i}^{(t)} + j \cdot \gamma_{j,r}^{(t)} \|\boldsymbol{\mu}\|_2^{-2} \cdot \langle \boldsymbol{\mu}, \boldsymbol{\xi}_i\rangle + \sum_{i' \ne i} \underline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle + \sum_{i' \ne i} \overline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle.
\end{aligned}$$

Now we look at

$$\bigg|j \cdot \gamma_{j,r}^{(t)} \|\boldsymbol{\mu}\|_2^{-2} \cdot \langle \boldsymbol{\mu}, \boldsymbol{\xi}_i\rangle + \sum_{i' \ne i} \underline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle + \sum_{i' \ne i} \overline{\rho}_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle\bigg|$$

$$
\begin{aligned}
&\leq \gamma_{j,r}^{(t)}\|\boldsymbol{\mu}\|_2^{-2}\cdot|\langle\boldsymbol{\mu},\boldsymbol{\xi}_i\rangle| + \sum_{i'\neq i}(|\underline{\rho}_{j,r,i'}^{(t)}| + |\overline{\rho}_{j,r,i'}^{(t)}|)\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot|\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle| \\
&\leq \gamma_{j,r}^{(t)}\|\boldsymbol{\mu}\|_2^{-1}\sigma_p\sqrt{2\log(6n/\delta)} + 4\sqrt{\frac{\log(6n^2/\delta)}{d}}\bigg(\sum_{i'\neq i}|\overline{\rho}_{j,r,i'}^{(t)}| + \sum_{i'\neq i}|\underline{\rho}_{j,r,i'}^{(t)}|\bigg) \\
&= \mathrm{SNR}^{-1}\sqrt{\frac{2\log(6n/\delta)}{d}}\gamma_{j,r}^{(t)} + 4\sqrt{\frac{\log(6n^2/\delta)}{d}}\bigg(\sum_{i'\neq i}|\overline{\rho}_{j,r,i'}^{(t)}| + \sum_{i'\neq i}|\underline{\rho}_{j,r,i'}^{(t)}|\bigg) \\
&\leq \mathrm{SNR}\sqrt{\frac{8C'^2\log(6n/\delta)}{d}}n\alpha + 4\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha \\
&\leq 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,
\end{aligned}
$$

where the first inequality is by triangle inequality and $\gamma_{j,r}^{(t)} \geq 0$; the second inequality is by Lemma B.4; the equality is by the definition of $\mathrm{SNR} = \|\boldsymbol{\mu}\|_2/\sigma_p\sqrt{d}$; the second last inequality is by (C.10) and (C.11); the last inequality is by $\mathrm{SNR} \leq 1/\sqrt{8C'^2}$. It follows that for $j \neq y_i$

$$
\big|\langle\mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)},\boldsymbol{\xi}_i\rangle - \underline{\rho}_{j,r,i}^{(t)}\big| \leq 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha.
$$

Similarly, for $y_i = j$, we have that $\underline{\rho}_{j,r,i}^{(t)} = 0$ and

$$
\begin{aligned}
\langle\mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)},\boldsymbol{\xi}_i\rangle &= j\cdot\gamma_{j,r}^{(t)}\|\boldsymbol{\mu}\|_2^{-2}\cdot\langle\boldsymbol{\mu},\boldsymbol{\xi}_i\rangle + \sum_{i'=1}^{n}\overline{\rho}_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle + \sum_{i'=1}^{n}\underline{\rho}_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle \\
&= \overline{\rho}_{j,r,i}^{(t)} + j\cdot\gamma_{j,r}^{(t)}\|\boldsymbol{\mu}\|_2^{-2}\cdot\langle\boldsymbol{\mu},\boldsymbol{\xi}_i\rangle + \sum_{i'\neq i}\overline{\rho}_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle + \sum_{i'\neq i}\underline{\rho}_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle,
\end{aligned}
$$

and also

$$
\begin{aligned}
&\bigg|j\cdot\gamma_{j,r}^{(t)}\|\boldsymbol{\mu}\|_2^{-2}\cdot\langle\boldsymbol{\mu},\boldsymbol{\xi}_i\rangle + \sum_{i'\neq i}\overline{\rho}_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle + \sum_{i'\neq i}\underline{\rho}_{j,r,i'}^{(t)}\|\boldsymbol{\xi}_{i'}\|_2^{-2}\cdot\langle\boldsymbol{\xi}_{i'},\boldsymbol{\xi}_i\rangle\bigg| \\
&\leq \mathrm{SNR}^{-1}\sqrt{\frac{2\log(6n/\delta)}{d}}\gamma_{j,r}^{(t)} + 4\sqrt{\frac{\log(6n^2/\delta)}{d}}\bigg(\sum_{i'\neq i}|\overline{\rho}_{j,r,i'}^{(t)}| + \sum_{i'\neq i}|\underline{\rho}_{j,r,i'}^{(t)}|\bigg) \\
&\leq \mathrm{SNR}\sqrt{\frac{8C'^2\log(6n/\delta)}{d}}n\alpha + 4\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha \\
&\leq 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,
\end{aligned}
$$

where the second last inequality is by (C.9), (C.11); the last inequality is by $\mathrm{SNR} \leq 1/\sqrt{8C'^2}$. It follows that for $j = y_i$

$$
\big|\langle\mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)},\boldsymbol{\xi}_i\rangle - \overline{\rho}_{j,r,i}^{(t)}\big| \leq 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,
$$

which completes the proof. □

Lemma C.4. *Under Condition 4.1, suppose* (C.9), (C.10) *and* (C.11) *hold at iteration t. Then, for all $j \neq y_i$, $j \in \{\pm 1\}$ and $i \in [n]$, $F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) \leq 0.5$.*

Proof of Lemma C.4. According to Lemma C.3, we have

$$\begin{aligned}
F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) &= \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)] \\
&\leq 5\max\left\{|\langle \mathbf{w}_{j,r}^{(0)}, \widehat{y}_i\boldsymbol{\mu}\rangle|, |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle|, \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha, 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha, C'\widehat{\gamma}\alpha\right\} \\
&\leq 5\max\left\{\beta, \mathrm{SNR}\sqrt{\frac{32\log(6n/\delta)}{d}}n\alpha, 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha, C'\widehat{\gamma}\alpha\right\} \\
&< 0.5,
\end{aligned}$$

where the first inequality is by (C.12), (C.13) and (C.14); the second inequality is due to the definition of β; the third inequality is by (C.7). □

Lemma C.5. *Under Condition 4.1, suppose* (C.9)*,* (C.10) *and* (C.11) *hold at iteration t. Then, it holds that*

$$\begin{aligned}
&\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle \geq -0.25, \\
&\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle \leq \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle) \leq \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle + 0.25,
\end{aligned}$$

for any $i \in [n]$.

Proof of Lemma C.5. According to (C.14) in Lemma C.3, we have

$$\begin{aligned}
\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle &\geq \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \overline{\rho}_{y_i,r,i}^{(t)} - 5n\sqrt{\frac{\log(4n^2/\delta)}{d}}\alpha \\
&\geq -\beta - 5n\sqrt{\frac{\log(4n^2/\delta)}{d}}\alpha \\
&\geq -0.25,
\end{aligned}$$

where the second inequality is due to $\overline{\rho}_{y_i,r,i}^{(t)} \geq 0$, the third inequality is due to $\beta < 1/8$ and $5n\sqrt{\log(4n^2/\delta)/d}\cdot\alpha < 1/8$.

For the second inequality, LHS holds naturally since $z \leq \sigma(z)$. For RHS, if $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle \leq 0$, then

$$\sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle) = 0 \leq \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle + 0.25.$$

If $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle > 0$, then

$$\sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle) = \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle < \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle + 0.25.$$

□

Next we present an important Lemma C.7, which ensures the logits $\ell_i'^{(t)}$ for different $i \in [n]$ are balanced. As we will see later, this guarantees the coefficients $\gamma_{j,r}^{(t)}$ are monotone with respect to t despite label-flipping noise, which is essential for the proof of Proposition C.2. In preparation, we first present a supplementary lemma.

Lemma C.6. *Let $g(z) = \ell'(z) = -1/(1+\exp(z))$, then for all $z_2 - c \geq z_1 \geq -1$ where $c \geq 0$ we have that*

$$\frac{\exp(c)}{4} \leq \frac{g(z_1)}{g(z_2)} \leq \exp(c).$$

Proof of Lemma C.6. On one hand, we have

$$\frac{1+\exp(z_2)}{1+\exp(z_1)} \leq \max\{1, \exp(z_2 - z_1)\} = \exp(c),$$

while on the other hand, we have

$$\frac{1+\exp(z_2)}{1+\exp(z_1)} = \frac{\exp(-z_1)+\exp(z_2-z_1)}{\exp(-z_1)+1} \geq \frac{\exp(-z_1)+\exp(c)}{\exp(-z_1)+1} \geq \frac{\exp(1)+\exp(c)}{\exp(1)+1} \geq \frac{\exp(c)}{4}.$$

□

Lemma C.7. *Under Condition 4.1, suppose (C.9), (C.10) and (C.11) hold for any iteration $t' \leq t$. Then, the following conditions hold for any iteration $t' \leq t$:*

1. $\sum_{r=1}^{m} \left[\overline{\rho}_{y_i,r,i}^{(t')} - \overline{\rho}_{y_k,r,k}^{(t')}\right] \leq \kappa$ *for all* $i, k \in [n]$.
2. $y_i \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_k) \leq C_1$ *for all* $i, k \in [n]$,
3. $\ell_i'^{(t')}/\ell_k'^{(t')} \leq C_2 = \exp(C_1)$ *for all* $i, k \in [n]$.
4. $S_i^{(0)} \subseteq S_i^{(t')}$, *where* $S_i^{(t')} := \{r \in [m] : \langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i \rangle > 0\}$, *and hence* $|S_i^{(t')}| \geq 0.4m$ *for all* $i \in [n]$.
5. $S_{j,r}^{(0)} \subseteq S_{j,r}^{(t')}$, *where* $S_{j,r}^{(t')} := \{i \in [n] : y_i = j, \langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi}_i \rangle > 0\}$, *and hence* $|S_{j,r}^{(t')}| \geq n/8$ *for all* $j \in \{\pm 1\}, r \in [m]$.

Here we take κ and C_1 as 3.25 and 5 respectively.

Proof of Lemma C.7. We prove this lemma by induction. When $t' = 0$, the fourth and fifth conditions hold naturally, so we only need to verify the first three hypotheses. Since according to (C.8) we have $\overline{\rho}_{j,r,i}^{(0)} = 0$ for any j, r, i, it follows that $\sum_{r=1}^{m} \left[\overline{\rho}_{y_i,r,i}^{(0)} - \overline{\rho}_{y_k,r,k}^{(0)}\right] = 0$ for all $i, k \in [n]$, and so the first condition holds for $t' = 0$. For the second condition, we have for any $i, k \in [n]$

$$\begin{aligned}
& y_i \cdot f(\mathbf{W}^{(0)}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(0)}, \mathbf{x}_k) \\
&= F_{y_i}(\mathbf{W}_{y_i}^{(0)}, \mathbf{x}_i) - F_{-y_i}(\mathbf{W}_{-y_i}^{(0)}, \mathbf{x}_i) + F_{-y_k}(\mathbf{W}_{-y_k}^{(0)}, \mathbf{x}_i) - F_{y_k}(\mathbf{W}_{y_k}^{(0)}, \mathbf{x}_i) \\
&\leq F_{y_i}(\mathbf{W}_{y_i}^{(0)}, \mathbf{x}_i) + F_{-y_k}(\mathbf{W}_{-y_k}^{(0)}, \mathbf{x}_i) \\
&= \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \widehat{y}_i \boldsymbol{\mu} \rangle) + \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle)] + \frac{1}{m}\sum_{r=1}^{m}[\sigma(\langle \mathbf{w}_{-y_k,r}^{(t)}, \widehat{y}_k \boldsymbol{\mu} \rangle) + \sigma(\langle \mathbf{w}_{-y_k,r}^{(t)}, \boldsymbol{\xi}_i \rangle)] \\
&\leq 2\beta \leq 1/3 \leq C_1,
\end{aligned}$$

where the first inequality is by the fact that $F_j(\mathbf{W}_j^{(0)}, \mathbf{x}_i) > 0$ for all $i \in [n], j \in [m]$, the second inequality is by the definition of β in (C.5), while the third inequality follows from (C.7). Finally, using the second condition, the third condition follows by

$$\frac{\ell_i'^{(0)}}{\ell_k'^{(0)}} \leq \exp\left(y_k \cdot f(\mathbf{W}^{(0)}, \mathbf{x}_k) - y_i \cdot f(\mathbf{W}^{(0)}, \mathbf{x}_i)\right) \leq \exp(C_1),$$

according to Lemma C.6.

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 $t' = \widetilde{t}$.

We first show that, for any $0 \leq t' \leq t$, $y_i \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_k)$ can be approximated by $\sum_{r=1}^{m} \left[\overline{\rho}_{y_i,r,i}^{(t')} - \overline{\rho}_{y_k,r,k}^{(t')}\right]$

with a small constant approximation error. We begin by writing out

$$
\begin{aligned}
& y_i \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_k) \\
&= y_i \sum_{j \in \{\pm 1\}} j \cdot F_j(\mathbf{W}_j^{(t')}, \mathbf{x}_i) - y_k \sum_{j \in \{\pm 1\}} j \cdot F_j(\mathbf{W}_j^{(t')}, \mathbf{x}_k) \\
&= F_{-y_k}(\mathbf{W}_{-y_k}^{(t')}, \mathbf{x}_k) - F_{-y_i}(\mathbf{W}_{-y_i}^{(t')}, \mathbf{x}_i) + F_{y_i}(\mathbf{W}_{y_i}^{(t')}, \mathbf{x}_i) - F_{y_k}(\mathbf{W}_{y_k}^{(t')}, \mathbf{x}_k) \\
&= F_{-y_k}(\mathbf{W}_{-y_k}^{(t')}, \mathbf{x}_k) - F_{-y_i}(\mathbf{W}_{-y_i}^{(t')}, \mathbf{x}_i) + \frac{1}{m} \sum_{r=1}^{m} [\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) + \sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i \rangle)] \\
&\quad - \frac{1}{m} \sum_{r=1}^{m} [\sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \widehat{y}_k \cdot \boldsymbol{\mu} \rangle) + \sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k \rangle)] \\
&= \underbrace{F_{-y_k}(\mathbf{W}_{-y_k}^{(t')}, \mathbf{x}_k) - F_{-y_i}(\mathbf{W}_{-y_i}^{(t')}, \mathbf{x}_i)}_{I_1} + \underbrace{\frac{1}{m} \sum_{r=1}^{m} [\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) - \sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \widehat{y}_k \cdot \boldsymbol{\mu} \rangle)]}_{I_2} \\
&\quad + \underbrace{\frac{1}{m} \sum_{r=1}^{m} [\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i \rangle) - \sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k \rangle)]}_{I_3},
\end{aligned} \tag{C.15}
$$

where all the equalities are due to the network definition. Next we estimate I_1, I_2 and I_3 one by one. For $|I_1|$, we have the following upper bound according to Lemma C.4:

$$
|I_1| \leq |F_{-y_k}(\mathbf{W}_{-y_k}^{(t')}, \mathbf{x}_k)| + |F_{-y_i}(\mathbf{W}_{-y_i}^{(t')}, \mathbf{x}_i)| = F_{-y_k}(\mathbf{W}_{-y_k}^{(t')}, \mathbf{x}_k) + F_{-y_i}(\mathbf{W}_{-y_i}^{(t')}, \mathbf{x}_i) \leq 1. \tag{C.16}
$$

For $|I_2|$, we have the following upper bound:

$$
\begin{aligned}
|I_2| &\leq \max \left\{ \frac{1}{m} \sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle), \frac{1}{m} \sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \widehat{y}_k \cdot \boldsymbol{\mu} \rangle) \right\} \\
&\leq 3 \max \left\{ |\langle \mathbf{w}_{y_i,r}^{(0)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle|, |\langle \mathbf{w}_{y_k,r}^{(0)}, \widehat{y}_k \cdot \boldsymbol{\mu} \rangle|, \gamma_{j,r}^{(t')}, \mathrm{SNR} \sqrt{\frac{32 \log(6n/\delta)}{d}} n\alpha \right\} \\
&\leq 3 \max \left\{ \beta, C' \widehat{\gamma} \alpha, \mathrm{SNR} \sqrt{\frac{32 \log(6n/\delta)}{d}} n\alpha \right\} \\
&\leq 0.25,
\end{aligned} \tag{C.17}
$$

where the second inequality is due to (C.12); the second inequality is due to the definition of β and (C.11); the last inequality is due to Condition 4.1 and (C.7).

For I_3, we have the following upper bound

$$
\begin{aligned}
I_3 &= \frac{1}{m} \sum_{r=1}^{m} \left[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i \rangle) - \sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k \rangle) \right] \\
&\leq \frac{1}{m} \sum_{r=1}^{m} \left[\langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_i \rangle - \langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k \rangle \right] + 0.25 \\
&\leq \frac{1}{m} \sum_{r=1}^{m} \left[\overline{\rho}_{y_i,r,i}^{(t')} - \overline{\rho}_{y_k,r,k}^{(t')} + 10 \sqrt{\frac{\log(6n^2/\delta)}{d}} n\alpha + 0.25 \right] \\
&\leq \frac{1}{m} \sum_{r=1}^{m} \left[\overline{\rho}_{y_i,r,i}^{(t')} - \overline{\rho}_{y_k,r,k}^{(t')} \right] + 0.5,
\end{aligned} \tag{C.18}
$$

where the first inequality is due to $\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) \leq \langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle + 0.25$ and $\sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k\rangle) \geq \langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k\rangle$ according to Lemma C.5; the second inequality is due to (C.14) in Lemma C.3; the last inequality is due to $5\sqrt{\log(6n^2/\delta)/d}n\alpha \leq 1/8$ according to Condition 4.1. Similarly, we have the following lower bound

$$
\begin{aligned}
I_3 &= \frac{1}{m}\sum_{r=1}^{m}\big[\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) - \sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k\rangle)\big] \\
&\geq \frac{1}{m}\sum_{r=1}^{m}\big[\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle - \langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k\rangle\big] - 0.25 \\
&\geq \frac{1}{m}\sum_{r=1}^{m}\bigg[\overline{\rho}_{y_i,r,i}^{(t')} - \overline{\rho}_{y_k,r,k}^{(t')} - 10\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha - 0.25\bigg] \\
&\geq \frac{1}{m}\sum_{r=1}^{m}\big[\overline{\rho}_{y_i,r,i}^{(t')} - \overline{\rho}_{y_k,r,k}^{(t')}\big] - 0.5,
\end{aligned}
\tag{C.19}
$$

where the first inequality is due to $\sigma(\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle) \geq \langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i\rangle$ and $\sigma(\langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k\rangle) \leq \langle \mathbf{w}_{y_k,r}^{(t')}, \boldsymbol{\xi}_k\rangle + 0.25$ according to Lemma C.5; the second inequality is due to (C.14) in Lemma C.3; the last inequality is due to $5\sqrt{\log(6n^2/\delta)/d}n\alpha \leq 1/8$ according to Condition 4.1. Now, by plugging (C.16)-(C.18) into (C.15), we get

$$
\begin{aligned}
y_i \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_k) &\leq |I_1| + |I_2| + I_3 \leq \frac{1}{m}\sum_{r=1}^{m}\big[\overline{\rho}_{y_i,r,i}^{(t')} - \overline{\rho}_{y_k,r,k}^{(t')}\big] + 1.75 \\
y_i \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_k) &\geq -|I_1| - |I_2| + I_3 \geq \frac{1}{m}\sum_{r=1}^{m}\big[\overline{\rho}_{y_i,r,i}^{(t')} - \overline{\rho}_{y_k,r,k}^{(t')}\big] - 1.75,
\end{aligned}
$$

which is equivalent to

$$
\bigg| y_i \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t')}, \mathbf{x}_k) - \frac{1}{m}\sum_{r=1}^{m}\big[\overline{\rho}_{y_i,r,i}^{(t')} - \overline{\rho}_{y_k,r,k}^{(t')}\big]\bigg| \leq 1.75. \tag{C.20}
$$

With this, we see that when the first condition holds for t', the second condition immediately follows for t'.

Next, we prove the first condition holds for $t' = \widetilde{t}$. We first write an iterative update rule for $\sum_{r=1}^{m}\big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t})} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t})}\big]$. Recall that from Lemma 5.3 that

$$
\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm}\cdot \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\cdot \mathbb{1}(y_i = j)\|\boldsymbol{\xi}_i\|_2^2
$$

for all $j \in \{\pm 1\}, r \in [m], i \in [n], t \in [0, T^*]$. Also recall the definition of $S_i^{(t)} = \{r \in [m] : \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle > 0\}$, it follows that

$$
\sum_{r=1}^{m}\big[\overline{\rho}_{y_i,r,i}^{(t+1)} - \overline{\rho}_{y_k,r,k}^{(t+1)}\big] = \sum_{r=1}^{m}\big[\overline{\rho}_{y_i,r,i}^{(t)} - \overline{\rho}_{y_k,r,k}^{(t)}\big] - \frac{\eta}{nm}\cdot\big(|S_i^{(t)}|\ell_i'^{(t)}\cdot\|\boldsymbol{\xi}_i\|_2^2 - |S_k^{(t)}|\ell_k'^{(t)}\cdot\|\boldsymbol{\xi}_k\|_2^2\big),
$$

for all $i, k \in [n]$ and $0 \leq t \leq T^*$. Now we consider two separate cases: $\sum_{r=1}^{m}\big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t}-1)} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t}-1)}\big] \leq 0.9\kappa$ and $\sum_{r=1}^{m}\big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t}-1)} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t}-1)}\big] > 0.9\kappa$.

For when $\sum_{r=1}^{m}\big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t}-1)} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t}-1)}\big] \leq 0.9\kappa$, we have

$$
\begin{aligned}
\sum_{r=1}^{m}\big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t})} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t})}\big] &= \sum_{r=1}^{m}\big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t}-1)} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t}-1)}\big] - \frac{\eta}{nm}\cdot\Big(|S_i^{(\widetilde{t}-1)}|\ell_i'^{(\widetilde{t}-1)}\cdot\|\boldsymbol{\xi}_i\|_2^2 - |S_k^{(\widetilde{t}-1)}|\ell_k'^{(\widetilde{t}-1)}\cdot\|\boldsymbol{\xi}_k\|_2^2\Big) \\
&\leq \sum_{r=1}^{m}\big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t}-1)} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t}-1)}\big] - \frac{\eta}{nm}\cdot|S_i^{(\widetilde{t}-1)}|\ell_i'^{(\widetilde{t}-1)}\cdot\|\boldsymbol{\xi}_i\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{r=1}^{m} \big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t}-1)} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t}-1)}\big] + \frac{\eta}{n} \cdot \|\boldsymbol{\xi}_i\|_2^2 \\
&\leq 0.9\kappa + 0.1\kappa \\
&= \kappa,
\end{aligned}
$$

where the first inequality is due to $\ell_i'^{(\widetilde{t}-1)} < 0$; the second inequality is due to $|S_i^{(\widetilde{t}-1)}| \leq m$ and $-\ell_i'^{(\widetilde{t}-1)} < 1$; the third inequality is due to the assumption in this case and $\eta \leq C^{-1} \cdot n\sigma_p^{-2} d^{-1}$ from Condition 4.1.

On the other hand, for when $\sum_{r=1}^{m} \big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t}-1)} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t}-1)}\big] > 0.9\kappa$, we have from the (C.20) that

$$
\begin{aligned}
y_i \cdot f(\mathbf{W}^{(\widetilde{t}-1)}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(\widetilde{t}-1)}, \mathbf{x}_k) &\geq \frac{1}{m} \sum_{r=1}^{m} \big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t}-1)} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t}-1)}\big] - 1.75 \\
&\geq 0.9\kappa - 0.54\kappa \\
&= 0.36\kappa,
\end{aligned} \tag{C.21}
$$

where the second inequality is due to $\kappa = 3.25$. Thus, according to Lemma C.6, we have

$$
\frac{\ell_i'^{(\widetilde{t}-1)}}{\ell_k'^{(\widetilde{t}-1)}} \leq \exp\big(y_k \cdot f(\mathbf{W}^{(\widetilde{t}-1)}, \mathbf{x}_k) - y_i \cdot f(\mathbf{W}^{(\widetilde{t}-1)}, \mathbf{x}_i)\big) \leq \exp(-0.36\kappa).
$$

Since we have $|S_i^{(\widetilde{t}-1)}| \leq m$ and $|S_k^{(\widetilde{t}-1)}| \geq 0.4m$ according to the fourth condition, it follows that

$$
\frac{|S_i^{(\widetilde{t}-1)}| \ell_i'^{(\widetilde{t}-1)}}{|S_k^{(\widetilde{t}-1)}| \ell_k'^{(\widetilde{t}-1)}} \leq 2.5 \exp(-0.36\kappa) < 0.8.
$$

According to Lemma B.4, under event $\mathcal{E}_{\text{prelim}}$, we have

$$
\big|\|\boldsymbol{\xi}_i\|_2^2 - d \cdot \sigma_p^2\big| = O\big(\sigma_p^2 \cdot \sqrt{d \log(6n/\delta)}\big), \forall i \in [n].
$$

Note that $d = \Omega(\log(6n/\delta))$ from Condition 4.1, it follows that

$$
|S_i^{(\widetilde{t}-1)}|(-\ell_i'^{(\widetilde{t}-1)}) \cdot \|\boldsymbol{\xi}_i\|_2^2 < |S_k^{(\widetilde{t}-1)}|(-\ell_k'^{(\widetilde{t}-1)}) \cdot \|\boldsymbol{\xi}_k\|_2^2.
$$

Then we have

$$
\sum_{r=1}^{m} \big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t})} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t})}\big] \leq \sum_{r=1}^{m} \big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t}-1)} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t}-1)}\big] \leq \kappa,
$$

which completes the proof of the first hypothesis at iteration $t' = \widetilde{t}$. Next, by applying the approximation in (C.20), we are ready to verify the second hypothesis at iteration $\widetilde{t}$. In fact, we have

$$
y_i \cdot f(\mathbf{W}^{(\widetilde{t})}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(\widetilde{t})}, \mathbf{x}_k) \leq \frac{1}{m} \sum_{r=1}^{m} \big[\overline{\rho}_{y_i,r,i}^{(\widetilde{t})} - \overline{\rho}_{y_k,r,k}^{(\widetilde{t})}\big] + 1.75 \leq C_1,
$$

where the first inequality is by (C.20); the last inequality is by induction hypothesis and taking κ as 3.25 and C_1 as 5. And the third hypothesis directly follows by noting that

$$
\frac{\ell_i'^{(\widetilde{t})}}{\ell_k'^{(\widetilde{t})}} \leq \exp\big(y_k \cdot f(\mathbf{W}^{(\widetilde{t})}, \mathbf{x}_k) - y_i \cdot f(\mathbf{W}^{(\widetilde{t})}, \mathbf{x}_i)\big) \leq \exp(C_1) = C_2.
$$

To verify the fourth hypothesis, according to the gradient descent rule, we have

$$\begin{aligned}
\langle \mathbf{w}_{y_i,r}^{(\widetilde{t})}, \boldsymbol{\xi}_i \rangle &= \langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i \rangle - \frac{\eta}{nm} \cdot \sum_{i'=1}^{n} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \widehat{y}_{i'} \boldsymbol{\mu} \rangle) \cdot \langle \widehat{y}_{i'} \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle \\
&\quad - \frac{\eta}{nm} \cdot \sum_{i'=1}^{n} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_{i'} \rangle) \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle \\
&= \langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i \rangle - \frac{\eta}{nm} \cdot \sum_{i'=1}^{n} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \widehat{y}_{i'} \boldsymbol{\mu} \rangle) \cdot \langle \widehat{y}_{i'} \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle \\
&\quad - \frac{\eta}{nm} \cdot \ell_{i}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i \rangle) \cdot \|\boldsymbol{\xi}_i\|_2^2 - \frac{\eta}{nm} \cdot \sum_{i' \neq i} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_{i'} \rangle) \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle \\
&= \langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i \rangle - \frac{\eta}{nm} \cdot \underbrace{\ell_{i}^{(\widetilde{t}-1)} \cdot \|\boldsymbol{\xi}_i\|_2^2}_{I_4} - \frac{\eta}{nm} \cdot \underbrace{\sum_{i' \neq i} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_{i'} \rangle) \cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle}_{I_5} \\
&\quad - \frac{\eta}{nm} \cdot \underbrace{\sum_{i'=1}^{n} \ell_{i'}^{(\widetilde{t}-1)} \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \widehat{y}_{i'} \boldsymbol{\mu} \rangle) \cdot \langle \widehat{y}_{i'} \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle}_{I_6},
\end{aligned}$$

for any $r \in S_i^{(\widetilde{t}-1)}$, where the last equality is by $\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i \rangle > 0$. Then we respectively estimate I_4, I_5, I_6. For I_4, according to Lemma B.4, we have

$$-I_4 \geq |\ell_i^{(\widetilde{t}-1)}| \cdot \sigma_p^2 d/2.$$

For I_5, we have following upper bound

$$\begin{aligned}
|I_5| &\leq \sum_{i' \neq i} |\ell_{i'}^{(\widetilde{t}-1)}| \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_{i'} \rangle) \cdot |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle| \\
&\leq \sum_{i' \neq i} |\ell_{i'}^{(\widetilde{t}-1)}| \cdot |\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle| \\
&\leq \sum_{i' \neq i} |\ell_{i'}^{(\widetilde{t}-1)}| \cdot 2\sigma_p^2 \cdot \sqrt{d \log(6n^2/\delta)} \\
&\leq nC_2 |\ell_i^{(\widetilde{t}-1)}| \cdot 2\sigma_p^2 \cdot \sqrt{d \log(6n^2/\delta)},
\end{aligned}$$

where the first inequality is due to triangle inequality; the second inequality is due to $\sigma'(z) \in \{0, 1\}$; the third inequality is due to Lemma B.4; the last inequality is due to the third hypothesis at iteration $\widetilde{t} - 1$.

For I_6, we have following upper bound

$$\begin{aligned}
|I_6| &\leq \sum_{i'=1}^{n} |\ell_{i'}^{(\widetilde{t}-1)}| \cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \widehat{y}_{i'} \boldsymbol{\mu} \rangle) \cdot |\langle \widehat{y}_{i'} \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle| \\
&\leq \sum_{i'=1}^{n} |\ell_{i'}^{(\widetilde{t}-1)}| \cdot |\langle \widehat{y}_{i'} \boldsymbol{\mu}, \boldsymbol{\xi}_i \rangle| \\
&\leq \sum_{i'=1}^{n} |\ell_{i'}^{(\widetilde{t}-1)}| \cdot \|\boldsymbol{\mu}\|_2 \sigma_p \sqrt{2 \log(6n/\delta)} \\
&\leq nC_2 |\ell_i^{(\widetilde{t}-1)}| \cdot \|\boldsymbol{\mu}\|_2 \sigma_p \sqrt{2 \log(6n/\delta)},
\end{aligned}$$

where the first inequality is by triangle inequality; the second inequality is due to $\sigma'(z) \in \{0, 1\}$; the third inequality is by Lemma B.4; the last inequality is due to the third hypothesis at iteration $\widetilde{t} - 1$. Since $d \geq \max\{32C_2^2 n^2 \cdot$

$\log(6n^2/\delta), 4C_2 n\|\boldsymbol{\mu}\|\sigma_p^{-1}\sqrt{2\log(6n/\delta)}\}$, we have $-I_4 \geq \max\{|I_5|/2, |I_6|/2\}$ and hence $-I_4 \geq |I_5| + |I_6|$. It follows that

$$\langle \mathbf{w}_{y_i,r}^{(\widetilde{t})}, \boldsymbol{\xi}_i\rangle \geq \langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i\rangle > 0,$$

for any $r \in S_i^{(\widetilde{t}-1)}$. Therefore, $S_i^{(0)} \subseteq S_i^{(\widetilde{t}-1)} \subseteq S_i^{(\widetilde{t})}$. And it directly follows by Lemma B.6 that $|S_i^{(\widetilde{t})}| \geq 0.4m$, $\forall i \in [n]$, which implies that the fourth hypothesis holds for $t' = \widetilde{t}$. For the fifth hypothesis, similar to the proof of the fourth hypothesis, we also have

$$\begin{aligned}\langle \mathbf{w}_{j,r}^{(\widetilde{t})}, \boldsymbol{\xi}_i\rangle = \langle \mathbf{w}_{j,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i\rangle &- \frac{\eta}{nm}\cdot \ell_i^{(\widetilde{t}-1)}\cdot \|\boldsymbol{\xi}_i\|_2^2 - \frac{\eta}{nm}\cdot \sum_{i'\neq i} \ell_{i'}^{(\widetilde{t}-1)}\cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_{i'}\rangle)\cdot \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle \\ &- \frac{\eta}{nm}\cdot \sum_{i'=1}^{n} \ell_{i'}^{(\widetilde{t}-1)}\cdot \sigma'(\langle \mathbf{w}_{y_i,r}^{(\widetilde{t}-1)}, \widehat{y}_{i'}\boldsymbol{\mu}\rangle)\cdot \langle \widehat{y}_{i'}\boldsymbol{\mu}, \boldsymbol{\xi}_i\rangle\end{aligned}$$

for any $i \in S_{j,r}^{(\widetilde{t}-1)}$, where the equality holds due to $\langle \mathbf{w}_{j,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i\rangle > 0$ and $y_i = j$. By applying the same technique used in the proof of the fourth hypothesis, it follows that

$$\langle \mathbf{w}_{j,r}^{(\widetilde{t})}, \boldsymbol{\xi}_i\rangle \geq \langle \mathbf{w}_{j,r}^{(\widetilde{t}-1)}, \boldsymbol{\xi}_i\rangle > 0,$$

for any $i \in S_{j,r}^{(\widetilde{t}-1)}$. Thus, we have $S_{j,r}^{(0)} \subseteq S_{j,r}^{(\widetilde{t}-1)} \subseteq S_{j,r}^{(\widetilde{t})}$. And it directly follows by Lemma B.7 that $|S_{j,r}^{(\widetilde{t})}| \geq n/8$, which implies that the fourth hypothesis holds for $t' = \widetilde{t}$. Therefore, the five hypotheses hold for $t' = \widetilde{t}$, which completes the induction.

□

Now we are ready to prove Proposition C.2.

Proof of Proposition C.2. 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 C.2 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 C.7, we also have for any $0 \leq t \leq \widetilde{T} - 1$ that

1. $\sum_{r=1}^{m} \left[\overline{\rho}_{y_i,r,i}^{(t)} - \overline{\rho}_{y_k,r,k}^{(t)}\right] \leq \kappa$ for all $i, k \in [n]$.
2. $y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i) - y_k \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_k) \leq C_1$ for all $i, k \in [n]$,
3. $\ell_i'^{(t)}/\ell_k'^{(t)} \leq C_2 = \exp(C_1)$ for all $i, k \in [n]$.
4. $S_i^{(0)} \subseteq S_i^{(t)}$ for all $i \in [n]$, where $S_i^{(t)} := \{r \in [m] : \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle > 0\}$, and hence $|S_i^{(t)}| \geq 0.4m$, for all $i \in [n]$.
5. $S_{j,r}^{(0)} \subseteq S_{j,r}^{(t)}$, where $S_{j,r}^{(t)} := \{i \in [n] : y_i = j, \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle > 0\}$, and hence $|S_{j,r}^{(t)}| \geq n/8$ for all $j \in \{\pm 1\}, r \in [m]$.

We first prove that (C.10) holds for $t = \widetilde{T}$, i.e., $\underline{\rho}_{j,r,i}^{(t)} \geq -\beta - 10\sqrt{\log(6n^2/\delta)/d}\cdot n\alpha$ for $t = \widetilde{T}$ and any $r \in [m]$, $j \in \{\pm 1\}$ and $i \in [n]$. Notice that $\underline{\rho}_{j,r,i}^{(t)} = 0$ for $j = y_i$, therefore we only need to consider the case that $j \neq y_i$. When $\underline{\rho}_{j,r,t}^{(\widetilde{T}-1)} < -0.5\beta - 5\sqrt{\log(6n^2/\delta)/d}\cdot n\alpha$, by (C.14) in Lemma C.3 we have that

$$\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\xi}_i\rangle \leq \underline{\rho}_{j,r,i}^{(\widetilde{T}-1)} + \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle + 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha < 0,$$

and thus

$$\begin{aligned}\underline{\rho}_{j,r,i}^{(\widetilde{T})} &= \underline{\rho}_{j,r,i}^{(\widetilde{T}-1)} + \frac{\eta}{nm}\cdot \ell_i'^{(\widetilde{T}-1)}\cdot \mathbb{1}(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\xi}_i\rangle \geq 0)\cdot \mathbb{1}(y_i = -j)\|\boldsymbol{\xi}_i\|_2^2 \\ &= \underline{\rho}_{j,r,i}^{(\widetilde{T}-1)}\end{aligned}$$

$$\geq -\beta - 10\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,$$

where the last inequality is by induction hypothesis. When $\underline{\rho}_{j,r,t}^{(\widetilde{T}-1)} \geq -0.5\beta - 5\sqrt{\log(6n^2/\delta)/d} \cdot n\alpha$, we have

$$\begin{aligned}
\underline{\rho}_{j,r,i}^{(\widetilde{T})} &= \underline{\rho}_{j,r,i}^{(\widetilde{T}-1)} + \frac{\eta}{nm} \cdot \ell_i'^{(\widetilde{T}-1)} \cdot \mathbb{1}(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\xi}_i \rangle \geq 0) \cdot \mathbb{1}(y_i = -j)\|\boldsymbol{\xi}_i\|_2^2 \\
&\geq -0.5\beta - 5\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha - \frac{3\eta\sigma_p^2 d}{2nm} \\
&\geq -0.5\beta - 10\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha \\
&\geq -\beta - 10\sqrt{\frac{\log(6n^2/\delta)}{d}}n\alpha,
\end{aligned}$$

where the first equality is by $\ell_i'^{(\widetilde{T}-1)} \in (-1, 0)$ and $\|\boldsymbol{\xi}_i\|_2^2 \leq (3/2)\sigma_p^2 d$ by Lemma B.4; the second inequality is due to $5\sqrt{\log(6n^2/\delta)/d} \cdot n\alpha \geq 3\eta\sigma_p^2 d/2nm$ by the condition for η in Condition 4.1.

Next we prove (C.9) holds for $t = \widetilde{T}$. Consider

$$\begin{aligned}
|\ell_i'^{(t)}| &= \frac{1}{1 + \exp\{y_i \cdot [F_{+1}(\mathbf{W}_{+1}^{(t)}, \mathbf{x}_i) - F_{-1}(\mathbf{W}_{-1}^{(t)}, \mathbf{x}_i)]\}} \\
&\leq \exp\{-y_i \cdot [F_{+1}(\mathbf{W}_{+1}^{(t)}, \mathbf{x}_i) - F_{-1}(\mathbf{W}_{-1}^{(t)}, \mathbf{x}_i)]\} \\
&\leq \exp\{-F_{y_i}(\mathbf{W}_{y_i}^{(t)}, \mathbf{x}_i) + 0.5\},
\end{aligned} \tag{C.22}$$

where the last inequality is by $F_j(\mathbf{W}_j^{(t)}, \mathbf{x}_i) \leq 0.5$ for $j \neq y_i$ according to Lemma C.4. Now recall the iterative update rule of $\overline{\rho}_{j,r,i}^{(t)}$:

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \mathbb{1}(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle \geq 0) \cdot \mathbb{1}(y_i = j)\|\boldsymbol{\xi}_i\|_2^2.$$

Let $t_{j,r,i}$ be the last time $t < T^*$ that $\overline{\rho}_{j,r,i}^{(t)} \leq 0.5\alpha$. Then by iterating the update rule from $t = t_{j,r,i}$ to $t = \widetilde{T} - 1$, we get

$$\begin{aligned}
\overline{\rho}_{j,r,i}^{(\widetilde{T})} = \overline{\rho}_{j,r,i}^{(t_{j,r,i})} &- \underbrace{\frac{\eta}{nm} \cdot \ell_i'^{(t_{j,r,i})} \cdot \mathbb{1}(\langle \mathbf{w}_{j,r}^{(t_{j,r,i})}, \boldsymbol{\xi}_i \rangle \geq 0) \cdot \mathbb{1}(y_i = j)\|\boldsymbol{\xi}_i\|_2^2}_{I_7} \\
&- \underbrace{\sum_{t_{j,r,i} < t < \widetilde{T}} \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \mathbb{1}(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle \geq 0) \cdot \mathbb{1}(y_i = j)\|\boldsymbol{\xi}_i\|_2^2}_{I_8}.
\end{aligned} \tag{C.23}$$

We first bound I_7 as follows:

$$|I_7| \leq (\eta/nm) \cdot \|\boldsymbol{\xi}_i\|_2^2 \leq (\eta/nm) \cdot 3\sigma_p^2 d/2 \leq 1 \leq 0.25\alpha,$$

where the first inequality is by $\ell_i'^{(t_{j,r,i})} \in (-1, 0)$; the second inequality is by Lemma B.4; the third inequality is by $\eta \leq C^{-1} \cdot n/(\sigma_p^2 d)$ from Condition 4.1; the last inequality is by our choice of $\alpha = 4\log(T^*)$ and $T^* \geq e$.

Second, we bound I_8. For $t_{j,r,i} < t < \widetilde{T}$ and $y_i = j$, we can lower bound the inner product $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$ as follows

$$\begin{aligned}\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle &\geq \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle + \overline{\rho}_{j,r,i}^{(t)} - 5\sqrt{\frac{\log(6n^2/\delta)}{d}} n\alpha \\ &\geq -0.5\beta + 0.5\alpha - 5\sqrt{\frac{\log(6n^2/\delta)}{d}} n\alpha \\ &\geq 0.25\alpha,\end{aligned} \tag{C.24}$$

where the first inequality is by (C.13) in Lemma C.3; the second inequality is by $\overline{\rho}_{j,r,i}^{(t)} > 0.5\alpha$ and $\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle \geq -0.5\beta$ due to the definition of $t_{j,r,i}$ and β; the last inequality is by $\beta \leq 1/8 \leq 0.1\alpha$ and $5\sqrt{\log(6n^2/\delta)/d} \cdot n\alpha \leq 0.2\alpha$ by $d \geq C \cdot n^2 \log(nm/\delta)(\log T^*)^2$ from Condition 4.1. Thus, plugging the lower bounds of $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$ into I_8 gives

$$\begin{aligned}|I_8| &\leq \sum_{t_{j,r,i} < t < \widetilde{T}} \frac{\eta}{nm} \cdot \exp(-\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) + 0.5) \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \cdot \mathbb{1}(y_i = j) \|\boldsymbol{\xi}_i\|_2^2 \\ &\leq \frac{2\eta(\widetilde{T} - t_{j,r,i} - 1)}{nm} \cdot \exp(-0.25\alpha) \cdot \frac{3\sigma_p^2 d}{2} \\ &\leq \frac{2\eta T^*}{nm} \cdot \exp(-\log(T^*)) \cdot \frac{3\sigma_p^2 d}{2} \\ &= \frac{2\eta}{nm} \cdot \frac{3\sigma_p^2 d}{2} \leq 1 \leq 0.25\alpha,\end{aligned}$$

where the first inequality is by (C.22); the second inequality is by (C.24); the third inequality is by $\alpha = 4\log(T^*)$; the fourth inequality is by $\eta \leq C^{-1} n^2 m \sqrt{\log(n/\delta)} \sigma_p^{-2} d^{-3/2} \leq nm/(3\sigma_p^2 d)$ based on the conditions for η and d in Condition 4.1; the last inequality is by $\log(T^*) \geq 1$ and $\alpha = 4\log(T^*)$. Plugging the bound of I_7, I_8 into (C.23) completes the proof for $\overline{\rho}$.

Next, we prove (C.11) holds for $t = \widetilde{T}$. Recall the iterative update rule of $\gamma_{j,r}^{(t)}$, we have

$$\gamma_{j,r}^{(\widetilde{T})} = \gamma_{j,r}^{(\widetilde{T}-1)} - \frac{\eta}{nm} \cdot \bigg[\sum_{i \in S_+} \ell_i'^{(\widetilde{T}-1)} \sigma'(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) - \sum_{i \in S_-} \ell_i'^{(\widetilde{T}-1)} \sigma'(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) \bigg] \cdot \|\boldsymbol{\mu}\|_2^2$$

We first prove that the coefficients $\gamma_{j,r}^{(\widetilde{T})} \geq \gamma_{j,r}^{(\widetilde{T}-1)}$ and hence $\gamma_{j,r}^{(\widetilde{T})} \geq \gamma_{j,r}^{(0)} = 0$ for any $j \in \{\pm 1\}, r \in [m]$. Recall the definition of $S_+ = \{i | y_i = \widehat{y}_i\}$, $S_- = \{i | y_i \neq \widehat{y}_i\}$, $S_1 = \{i | \widehat{y}_i = 1\}$ and $S_{-1} = \{i | \widehat{y}_i = -1\}$. We will consider the following two cases separately: $\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu} \rangle \geq 0$ and $\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu} \rangle < 0$. If $\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu} \rangle \geq 0$, then

$$\begin{aligned}&- \sum_{i \in S_+} \ell_i'^{(\widetilde{T}-1)} \sigma'(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) + \sum_{i \in S_-} \ell_i'^{(\widetilde{T}-1)} \sigma'(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) \\ &= - \sum_{i \in S_+} \ell_i'^{(\widetilde{T}-1)} \mathbb{1}(\widehat{y}_i \cdot \langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu} \rangle \geq 0) + \sum_{i \in S_-} \ell_i'^{(\widetilde{T}-1)} \mathbb{1}(\widehat{y}_i \cdot \langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu} \rangle \geq 0) \\ &= \sum_{i \in S_+ \cap S_1} |\ell_i'^{(\widetilde{T}-1)}| - \sum_{i \in S_- \cap S_{-1}} |\ell_i'^{(\widetilde{T}-1)}| \\ &\geq |S_+ \cap S_1| \cdot \min_{i \in S_+ \cap S_1} |\ell_i'^{(\widetilde{T}-1)}| - |S_- \cap S_{-1}| \cdot \max_{i \in S_- \cap S_{-1}} |\ell_i'^{(\widetilde{T}-1)}|,\end{aligned}$$

where the second equality is due to $\ell_i'^{(\widetilde{T}-1)} < 0$. If $\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu} \rangle < 0$, then with a similar reasoning we have

$$- \sum_{i \in S_+} \ell_i'^{(\widetilde{T}-1)} \sigma'(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) + \sum_{i \in S_-} \ell_i'^{(\widetilde{T}-1)} \sigma'(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle)$$

$$
\begin{aligned}
&= -\sum_{i\in S_+} \ell_i'^{(\widetilde{T}-1)}\, \mathbb{1}(\widehat{y}_i \cdot \langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu}\rangle \geq 0) + \sum_{i\in S_-} \ell_i'^{(\widetilde{T}-1)}\, \mathbb{1}(\widehat{y}_i \cdot \langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu}\rangle \geq 0) \\
&= \sum_{i\in S_+\cap S_{-1}} |\ell_i'^{(\widetilde{T}-1)}| - \sum_{i\in S_-\cap S_1} |\ell_i'^{(\widetilde{T}-1)}| \\
&\geq |S_+\cap S_{-1}| \cdot \min_{i\in S_+\cap S_{-1}} |\ell_i'^{(\widetilde{T}-1)}| - |S_-\cap S_1| \cdot \max_{i\in S_-\cap S_1} |\ell_i'^{(\widetilde{T}-1)}|.
\end{aligned}
$$

According to Lemma B.3 and the third statement from Lemma C.7 that $\ell_i'^{(\widetilde{T}-1)}/\ell_k'^{(\widetilde{T}-1)} \leq C_2, \forall i,k \in [n]$, under event $\mathcal{E}_{\text{prelim}}$, we have

$$
\begin{aligned}
&\frac{|S_+\cap S_1| \cdot \min_{i\in S_+\cap S_1} |\ell_i'^{(\widetilde{T}-1)}|}{|S_-\cap S_{-1}| \cdot \max_{i\in S_-\cap S_{-1}} |\ell_i'^{(\widetilde{T}-1)}|} \geq \frac{|S_+\cap S_1|}{C_2|S_-\cap S_{-1}|} \geq \frac{(1-p)n - \sqrt{2n\log(8/\delta)}}{C_2\cdot(pn+\sqrt{2n\log(8/\delta)})}, \\
&\frac{|S_+\cap S_{-1}| \cdot \min_{i\in S_+\cap S_{-1}} |\ell_i'^{(\widetilde{T}-1)}|}{|S_-\cap S_1| \cdot \max_{i\in S_-\cap S_1} |\ell_i'^{(\widetilde{T}-1)}|} \geq \frac{|S_+\cap S_{-1}|}{C_2|S_-\cap S_1|} \geq \frac{(1-p)n - \sqrt{2n\log(8/\delta)}}{C_2\cdot(pn+\sqrt{2n\log(8/\delta)})}.
\end{aligned}
$$

As long as $p < 1/[2(1+C_2)]$ and $n \geq 8(C_2+1)^2\log(8/\delta)$, we have

$$
\frac{|S_+\cap S_1| \cdot \min_{i\in S_+\cap S_1} |\ell_i'^{(\widetilde{T}-1)}|}{|S_-\cap S_{-1}| \cdot \max_{i\in S_-\cap S_{-1}} |\ell_i'^{(\widetilde{T}-1)}|} \geq 1,\ \frac{|S_+\cap S_{-1}| \cdot \min_{i\in S_+\cap S_{-1}} |\ell_i'^{(\widetilde{T}-1)}|}{|S_-\cap S_1| \cdot \max_{i\in S_-\cap S_1} |\ell_i'^{(\widetilde{T}-1)}|} \geq 1.
$$

And it follows for both cases $\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu}\rangle \geq 0$ and $\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\mu}\rangle < 0$ that

$$
\gamma_{j,r}^{(\widetilde{T})} \geq \gamma_{j,r}^{(\widetilde{T}-1)}, \tag{C.25}
$$

and hence

$$
\gamma_{j,r}^{(\widetilde{T})} \geq \gamma_{j,r}^{(0)} = 0.
$$

For the other part of (C.11), we prove a strengthened hypothesis that there exists a $i^* \in [n]$ with $y_{i^*} = j$ such that for $1 \leq t \leq T^*$ we have that

$$
\gamma_{j,r}^{(t)}/\overline{\rho}_{j,r,i^*}^{(t)} \leq C' n\|\boldsymbol{\mu}\|_2^2/\sigma_p^2 d,
$$

and i^* can be taken as any sample from set $S_{j,r}^{(0)}$ and C' can be taken as $2C_2$.

Recall the update rule of $\gamma_{j,r}^{(t)}$ and $\underline{\rho}_{j,r,i}^{(t)}$, we have

$$
\begin{aligned}
\gamma_{j,r}^{(\widetilde{T})} &= \gamma_{j,r}^{(\widetilde{T}-1)} - \frac{\eta}{nm}\cdot\Big[\sum_{i\in S_+} \ell_i'^{(\widetilde{T}-1)}\, \mathbb{1}(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i\cdot\boldsymbol{\mu}\rangle \geq 0) - \sum_{i\in S_-} \ell_i'^{(\widetilde{T}-1)}\, \mathbb{1}(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \widehat{y}_i\cdot\boldsymbol{\mu}\rangle \geq 0)\Big]\cdot\|\boldsymbol{\mu}\|_2^2, \\
\overline{\rho}_{j,r,i}^{(\widetilde{T})} &= \overline{\rho}_{j,r,i}^{(\widetilde{T}-1)} - \frac{\eta}{nm}\cdot \ell_i'^{(\widetilde{T}-1)}\cdot \mathbb{1}(\langle \mathbf{w}_{j,r}^{(\widetilde{T}-1)}, \boldsymbol{\xi}_i\rangle \geq 0)\cdot \mathbb{1}(y_i = j)\|\boldsymbol{\xi}_i\|_2^2.
\end{aligned}
$$

According to the fifth statement of Lemma C.7, for any $i^* \in S_{j,r}^{(0)}$ it holds that $j = y_{i^*}$ and $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_{i^*}\rangle \geq 0$ for any $0 \leq t \leq \widetilde{T}-1$. Thus, we have

$$
\overline{\rho}_{j,r,i^*}^{(\widetilde{T})} = \overline{\rho}_{j,r,i^*}^{(\widetilde{T}-1)} - \frac{\eta}{nm}\cdot \ell_{i^*}'^{(\widetilde{T}-1)}\cdot\|\boldsymbol{\xi}_{i^*}\|_2^2 \geq \overline{\rho}_{j,r,i^*}^{(\widetilde{T}-1)} - \frac{\eta}{nm}\cdot \ell_{i^*}'^{(\widetilde{T}-1)}\cdot\sigma_p^2 d/2.
$$

For the update rule of $\gamma_{j,r}^{(\widetilde{T})}$, we have

$$\left|\sum_{i\in S_+}\ell_i'^{(\widetilde{T}-1)}\mathbb{1}(\langle\mathbf{w}_{j,r}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle\geq 0)-\sum_{i\in S_-}\ell_i'^{(\widetilde{T}-1)}\mathbb{1}(\langle\mathbf{w}_{j,r}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle\geq 0)\right|\leq\sum_{i=1}^{n}|\ell_i'^{(\widetilde{T}-1)}|\cdot\|\boldsymbol{\mu}\|_2^2$$
$$\leq C_2 n\cdot|\ell_{i^*}'^{(\widetilde{T}-1)}|\cdot\|\boldsymbol{\mu}\|_2^2,$$

where the first inequality is due to triangle inequality; the second inequality is due to the third statement of Lemma C.7 where C_2 is a positive constant. Then, we have

$$\frac{\gamma_{j,r}^{(\widetilde{T})}}{\overline{\rho}_{j,r,i^*}^{(\widetilde{T})}}\leq\max\left\{\frac{\gamma_{j,r}^{(\widetilde{T}-1)}}{\overline{\rho}_{j,r,i^*}^{(\widetilde{T}-1)}},\frac{C_2 n\cdot|\ell_{i^*}'^{(\widetilde{T}-1)}|\cdot\|\boldsymbol{\mu}\|_2^2}{|\ell_{i^*}'^{(\widetilde{T}-1)}|\cdot\sigma_p^2 d/2}\right\}=\max\left\{\frac{\gamma_{j,r}^{(\widetilde{T}-1)}}{\overline{\rho}_{j,r,i^*}^{(\widetilde{T}-1)}},\frac{2C_2 n\|\boldsymbol{\mu}\|_2^2}{\sigma_p^2 d}\right\}\leq\frac{2C_2 n\|\boldsymbol{\mu}\|_2^2}{\sigma_p^2 d},$$

where the last inequality is by $\gamma_{j,r}^{(\widetilde{T}-1)}/\overline{\rho}_{j,r,i^*}^{(\widetilde{T}-1)}\leq C'\widehat{\gamma}=C'n\|\boldsymbol{\mu}\|_2^2/\sigma_p^2 d$ and C' can be taken as $2C_2$, which completes the induction. □

By then, we have already proved Proposition C.2. Then, according to Lemma C.7, next proposition directly follows.

Proposition C.8. *Under Condition 4.1, for $0\leq t\leq T^*$, we have that*

1. $\sum_{r=1}^m\left[\overline{\rho}_{y_i,r,i}^{(t)}-\overline{\rho}_{y_k,r,k}^{(t)}\right]\leq\kappa$ *for all* $i,k\in[n]$.
2. $y_i\cdot f(\mathbf{W}^{(t)},\mathbf{x}_i)-y_k\cdot f(\mathbf{W}^{(t)},\mathbf{x}_k)\leq C_1$ *for all* $i,k\in[n]$,
3. $\ell_i'^{(t)}/\ell_k'^{(t)}\leq C_2=\exp(C_1)$ *for all* $i,k\in[n]$.
4. $S_i^{(0)}\subseteq S_i^{(t)}$, *where* $S_i^{(t)}:=\{r\in[m]:\langle\mathbf{w}_{y_i,r}^{(t)},\boldsymbol{\xi}_i\rangle>0\}$, *and hence* $|S_i^{(t)}|\geq 0.4m$ *for all* $i\in[n]$.
5. $S_{j,r}^{(0)}\subseteq S_{j,r}^{(t)}$, *where* $S_{j,r}^{(t)}:=\{i\in[n]:y_i=j,\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}_i\rangle>0\}$, *and hence* $|S_{j,r}^{(t)}|\geq n/8$ *for all* $j\in\{\pm1\},r\in[m]$.

Here κ and C_1 can be taken as 3.25 and 5 respectively.

D. Decoupling with a Two-Stage Analysis

We utilize a two-stage analysis to decouple the complicated relations between the coefficients $\gamma_{j,r}^{(t)}$, $\overline{\rho}_{j,r,i}^{(t)}$ and $\underline{\rho}_{j,r,i}^{(t)}$. Intuitively, the initial neural network weights are small enough so that the neural network at initialization has constant level cross-entropy loss derivatives on all the training data: $\ell_i'^{(0)}=\ell'[y_i\cdot f(\mathbf{W}^{(0)},\mathbf{x}_i)]=\Theta(1)$ for all $i\in[n]$. Motivated by this, we can consider the first stage of the training process where $\ell_i'^{(0)}=\Theta(1)$, in which case we can show significant scale differences among $\gamma_{j,r}^{(t)}$, $\overline{\rho}_{j,r,i}^{(t)}$ and $\underline{\rho}_{j,r,i}^{(t)}$. Based on the result in the first stage, we then proceed to the second stage of the training process where the loss derivatives are no longer at a constant level and show that the training loss can be optimized to be arbitrarily small and meanwhile, the scale differences shown in the first learning stage remain the same throughout the training process. Recall that we denote $\alpha=4\log(T^*)$, $\beta=2\max_{i,j,r}\{|\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\mu}\rangle|,|\langle\mathbf{w}_{j,r}^{(0)},\boldsymbol{\xi}_i\rangle|\}$ and $\mathrm{SNR}=\|\boldsymbol{\mu}\|_2/(\sigma_p\sqrt{d})$. We remind the readers that the proofs in this section are based on the results in Section C, which hold with high probability.

D.1. First Stage

Lemma D.1. *If we denote*

$$n\cdot\mathrm{SNR}^2=\widehat{\gamma},$$

then there exist

$$T_1=C_3\eta^{-1}nm\sigma_p^{-2}d^{-1},T_2=C_4\eta^{-1}nm\sigma_p^{-2}d^{-1}$$

where $C_3=\Theta(1)$ is a large constant and $C_4=\Theta(1)$ is a small constant, such that

- $\overline{\rho}_{j,r^*,i}^{(T_1)} \geq 2$ *for any* $r^* \in S_i^{(0)} = \{r \in [m] : \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle > 0\}$, $j \in \{\pm 1\}$ *and* $i \in [n]$ *with* $y_i = j$.
- $\max_{j,r} \gamma_{j,r}^{(t)} = O(\widehat{\gamma})$ *for all* $0 \leq t \leq T_1$.
- $\max_{j,r,i} |\underline{\rho}_{j,r,i}^{(t)}| = \max\{O\big(\sqrt{\log(mn/\delta)} \cdot \sigma_0 \sigma_p \sqrt{d}\big), O\big(n\sqrt{\log(n/\delta)} \log(T^*)/\sqrt{d}\big)\}$ *for all* $0 \leq t \leq T_1$.
- $\min_{j,r} \gamma_{j,r}^{(t)} = \Omega(\widehat{\gamma})$ *for all* $t \geq T_2$.
- $\max_{j,r} \overline{\rho}_{j,r,i}^{(T_1)} = O(1)$ *for all* $i \in [n]$.

Proof of Lemma D.1. By Proposition C.2, we have that $\underline{\rho}_{j,r,i}^{(t)} \geq -\beta - 10n\sqrt{\frac{\log(6n^2/\delta)}{d}}\alpha$ for all $j \in \{\pm 1\}$, $r \in [m]$, $i \in [n]$ and $0 \leq t \leq T^*$. According to Lemma B.5, for β we have

$$\begin{aligned}
\beta &= 2\max_{i,j,r}\{|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|, |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle|\} \\
&\leq 2\max\{\sqrt{2\log(12m/\delta)} \cdot \sigma_0 \|\boldsymbol{\mu}\|_2, 2\sqrt{\log(12mn/\delta)} \cdot \sigma_0 \sigma_p \sqrt{d}\} \\
&= O\big(\sqrt{\log(mn/\delta)} \cdot \sigma_0 \sigma_p \sqrt{d}\big)
\end{aligned}$$

where the last equality is by the first condition of Condition 4.1. Since $\underline{\rho}_{j,r,i}^{(t)} \leq 0$, we have that

$$\begin{aligned}
\max_{j,r,i} |\underline{\rho}_{j,r,i}^{(t)}| &= \max_{j,r,i} -\underline{\rho}_{j,r,i}^{(t)} \\
&\leq \beta + 10\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha \\
&= \max\left\{O\big(\sqrt{\log(mn/\delta)} \cdot \sigma_0 \sigma_p \sqrt{d}\big), O\big(\sqrt{\log(n/\delta)} \log(T^*) \cdot n/\sqrt{d}\big)\right\}.
\end{aligned}$$

Next, for the growth of $\gamma_{j,r}^{(t)}$, we have following upper bound

$$\begin{aligned}
\gamma_{j,r}^{(t+1)} &= \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \left[\sum_{i \in S_+} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu}\rangle) - \sum_{i \in S_-} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu}\rangle)\right] \cdot \|\boldsymbol{\mu}\|_2^2 \\
&= \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \sum_{i=1}^{n} \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu}\rangle) \|\boldsymbol{\mu}\|_2^2 \\
&\leq \gamma_{j,r}^{(t)} + \frac{\eta}{m} \cdot \|\boldsymbol{\mu}\|_2^2,
\end{aligned}$$

where the inequality is by $|\ell'| \leq 1$. Note that $\gamma_{j,r}^{(0)} = 0$ and recursively use the inequality t times we have

$$\gamma_{j,r}^{(t)} \leq \frac{\eta t}{m} \cdot \|\boldsymbol{\mu}\|_2^2. \tag{D.1}$$

Since $n \cdot \mathrm{SNR}^2 = n\|\boldsymbol{\mu}\|_2^2/\sigma_p^2 d = \widehat{\gamma}$, we have

$$T_1 = C_3 \eta^{-1} nm\sigma_p^{-2} d^{-1} = C_3 \eta^{-1} m \|\boldsymbol{\mu}\|_2^{-2} \widehat{\gamma}.$$

And it follows that

$$\gamma_{j,r}^{(t)} \leq \frac{\eta t}{m} \cdot \|\boldsymbol{\mu}\|_2^2 \leq \frac{\eta T_1}{m} \cdot \|\boldsymbol{\mu}\|_2^2 \leq C_3 \widehat{\gamma},$$

for all $0 \leq t \leq T_1$.

For $\overline{\rho}_{j,r,i}^{(t)}$, recall from (5.4) that

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \cdot \mathbb{1}(y_i = j) \|\boldsymbol{\xi}_i\|_2^2.$$

According to Proposition C.8, for any $r^* \in S_i^{(0)} = \{r \in [m] : \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle > 0\}$, we have $\langle \mathbf{w}_{y_i,r^*}^{(t)}, \boldsymbol{\xi}_i \rangle > 0$ for all $0 \le t \le T^*$ and hence

$$\overline{\rho}_{y_i,r^*,i}^{(t+1)} = \overline{\rho}_{y_i,r^*,i}^{(t)} - \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^2.$$

Note that $\overline{\rho}_{y_i,r^*,i}^{(0)} = 0$ and recursively use the equation t times, we have

$$\overline{\rho}_{y_i,r^*,i}^{(t)} = -\frac{\eta}{nm} \cdot \sum_{s=0}^{t-1} \ell_i'^{(s)} \cdot \|\boldsymbol{\xi}_i\|_2^2.$$

For each i, denote by $T_1^{(i)}$ the last time in the period $[0, T_1]$ satisfying that $\max_{j,r} |\rho_{j,r,i}^{(t)}| \le 2$. Then for $0 \le t \le T_1^{(i)}$, $\max_{j,r}\{|\overline{\rho}_{j,r,i}^{(t)}|, |\underline{\rho}_{j,r,i}^{(t)}|\} = O(1)$ and $\max_{j,r} \gamma_{j,r}^{(t)} = O(1)$. Therefore, we know that $F_{-1}(\mathbf{W}^{(t)}, \mathbf{x}_i), F_{+1}(\mathbf{W}^{(t)}, \mathbf{x}_i) = O(1)$. Thus there exists a positive constant C such that $-\ell_i'^{(t)} \ge C$ for $0 \le t \le T_1^{(i)}$. Then we have

$$\overline{\rho}_{y_i,r^*,i}^{(t)} \ge \frac{C\eta\sigma_p^2 dt}{2nm}.$$

Therefore, $\overline{\rho}_{y_i,r^*,i}^{(t)}$ will reach 2 within

$$T_1 = C_3 \eta^{-1} nm \sigma_p^{-2} d^{-1}$$

iterations for any $r^* \in S_i^{(0)}$, where C_3 can be taken as $4/C$.

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

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \cdot \mathbb{1}(y_i = j) \|\boldsymbol{\xi}_i\|_2^2 \le \overline{\rho}_{j,r,i}^{(t)} + \frac{\eta}{nm} \|\boldsymbol{\xi}_i\|_2^2 \le \overline{\rho}_{j,r,i}^{(t)} + \frac{3\eta\sigma_p^2 d}{2nm},$$

where the first inequality is by $-\ell_i' \in (0,1)$ and $\sigma' \in \{0,1\}$; the second inequality is by Lemma B.4. According to (D.1) and $\overline{\rho}_{j,r,i}^{(0)} = 0$, it follows that

$$\overline{\rho}_{j,r,i}^{(t)} \le \frac{3\eta\sigma_p^2 dt}{2nm}, \gamma_{j,r}^{(t)} \le \frac{\eta t}{m} \cdot \|\boldsymbol{\mu}\|_2^2. \tag{D.2}$$

Therefore, $\max_{j,r,i} \overline{\rho}_{j,r,i}^{(t)}$ will be smaller than 1 and $\gamma_{j,r}^{(t)}$ smaller than $\Theta(n\|\boldsymbol{\mu}\|_2^2/\sigma_p^2 d) = \Theta(n \cdot \mathrm{SNR}^2) = \Theta(\widehat{\gamma}) = O(1)$ within

$$T_2 = C_4 \eta^{-1} nm \sigma_p^{-2} d^{-1}$$

iterations, where C_4 can be taken as $2/3$. Therefore, we know that $F_{-1}(\mathbf{W}^{(t)}, \mathbf{x}_i), F_{+1}(\mathbf{W}^{(t)}, \mathbf{x}_i) = O(1)$ in $[0, T_2]$. Thus there exists a positive constant C such that $-\ell_i'^{(t)} \ge C$ for $0 \le t \le T_2$.

Recall that we denote $\{i \in [n] | y_i = y\}$ as S_y. For the growth of $\gamma_{j,r}^{(t)}$, if $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle \ge 0$, we have

$$\begin{aligned}
\gamma_{j,r}^{(t+1)} &= \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \Bigg[\sum_{i \in S_+} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) - \sum_{i \in S_-} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \cdot \boldsymbol{\mu} \rangle) \Bigg] \cdot \|\boldsymbol{\mu}\|_2^2 \\
&= \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \Bigg[\sum_{i \in S_+ \cap S_1} \ell_i'^{(t)} - \sum_{i \in S_- \cap S_{-1}} \ell_i'^{(t)} \Bigg] \cdot \|\boldsymbol{\mu}\|_2^2 \\
&\ge \gamma_{j,r}^{(t)} + \frac{\eta}{nm} \cdot (C|S_+ \cap S_1| - |S_- \cap S_{-1}|) \cdot \|\boldsymbol{\mu}\|_2^2.
\end{aligned} \tag{D.3}$$

And if $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle < 0$, we have

$$\begin{aligned} \gamma_{j,r}^{(t+1)} &= \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \bigg[\sum_{i \in S_+ \cap S_{-1}} \ell_i'^{(t)} - \sum_{i \in S_- \cap S_1} \ell_i'^{(t)} \bigg] \cdot \|\boldsymbol{\mu}\|_2^2 \\ &\geq \gamma_{j,r}^{(t)} + \frac{\eta}{nm} \cdot (C|S_+ \cap S_{-1}| - |S_- \cap S_1|) \cdot \|\boldsymbol{\mu}\|_2^2. \end{aligned} \tag{D.4}$$

According to Lemma B.3, under event $\mathcal{E}_{\text{prelim}}$, we have

$$\begin{aligned} &\frac{|S_+ \cap S_1|}{|S_- \cap S_{-1}|}, \frac{|S_+ \cap S_{-1}|}{|S_- \cap S_1|} \geq \frac{(1-p)n - \sqrt{2n \log(8/\delta)}}{pn + \sqrt{2n \log(8/\delta)}}, \\ &|S_+ \cap S_1|, |S_+ \cap S_{-1}| \geq (1-p)n - \sqrt{2n \log(8/\delta)}. \end{aligned} \tag{D.5}$$

As long as $p < C/6$ and $n \geq 72C^{-2} \log(8/\delta)$, it follows that

$$\begin{aligned} &\frac{|S_+ \cap S_1|}{|S_- \cap S_{-1}|}, \frac{|S_+ \cap S_{-1}|}{|S_- \cap S_1|} \geq 2/C, \\ &|S_+ \cap S_1|, |S_+ \cap S_{-1}| \geq n/4. \end{aligned}$$

Therefore, we have

$$\begin{aligned} \gamma_{j,r}^{(t+1)} &\geq \gamma_{j,r}^{(t)} + \frac{C\eta}{2nm} \cdot |S_+ \cap S_{-1}| \cdot \|\boldsymbol{\mu}\|_2^2 \geq \gamma_{j,r}^{(t)} + \frac{C\eta}{8m} \cdot \|\boldsymbol{\mu}\|_2^2, \text{ if } \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle \geq 0, \\ \gamma_{j,r}^{(t+1)} &\geq \gamma_{j,r}^{(t)} + \frac{C\eta}{2nm} \cdot |S_+ \cap S_1| \cdot \|\boldsymbol{\mu}\|_2^2 \geq \gamma_{j,r}^{(t)} + \frac{C\eta}{8m} \cdot \|\boldsymbol{\mu}\|_2^2, \text{ if } \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle < 0. \end{aligned} \tag{D.6}$$

Note that $\gamma_{j,r}^{(0)} = 0$, it follows that

$$\gamma_{j,r}^{(t)} \geq \frac{C\|\boldsymbol{\mu}\|_2^2 \eta t}{8m}, \gamma_{j,r}^{(T_2)} \geq \frac{CC_4 n \|\boldsymbol{\mu}\|_2^2}{8\sigma_p^2 d} = \Theta(n \cdot \text{SNR}^2) = \Theta(\widehat{\gamma}).$$

Note that we have proved (C.25) in Lemma C.2 that $\gamma_{j,r}^{(t)}$ is increasing for $0 \leq t \leq T^*$, thus we have

$$\gamma_{j,r}^{(t)} = \Omega(\widehat{\gamma})$$

for $T_2 \leq t \leq T^*$. And it follows directly from (D.2) that

$$\overline{\rho}_{j,r,i}^{(T_1)} \leq \frac{3\eta\sigma_p^2 d T_1}{2nm} = \frac{3C_3}{2}, \overline{\rho}_{j,r,i}^{(T_1)} = O(1),$$

which completes the proof. □

D.2. Second Stage

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

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

for $j \in [\pm 1]$ 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:

- $\overline{\rho}_{j,r^*,i}^{(T_1)} \geq 2$ for any $r^* \in S_i^{(0)} = \{r \in [m] : \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i \rangle > 0\}$, $j \in \{\pm 1\}$ and $i \in [n]$ with $y_i = j$.
- $\max_{j,r,i} |\underline{\rho}_{j,r,i}^{(T_1)}| = \max\{O\big(\sqrt{\log(mn/\delta)} \cdot \sigma_0 \sigma_p \sqrt{d}\big), O\big(n\sqrt{\log(n/\delta)}\log(T^*)/\sqrt{d}\big)\}$.
- $\gamma_{j,r}^{(T_1)} = \Theta(\widehat{\gamma})$ for any $j \in \{\pm 1\}, r \in [m]$.

where $\widehat{\gamma} = n \cdot \mathrm{SNR}^2$. Now we choose $\mathbf{W}^*$ as follows

$$\mathbf{w}_{j,r}^* = \mathbf{w}_{j,r}^{(0)} + 5\log(2/\epsilon)\Big[\sum_{i=1}^n \mathbb{1}(j = y_i) \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2}\Big].$$

Lemma D.2. *Under the same conditions as Theorem 4.2, we have that* $\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F \leq \widetilde{O}(m^{1/2}n^{1/2}\sigma_p^{-1}d^{-1/2})$.

Proof of Lemma D.2. We have

$$\begin{aligned}
\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F &\leq \|\mathbf{W}^{(T_1)} - \mathbf{W}^{(0)}\|_F + \|\mathbf{W}^* - \mathbf{W}^{(0)}\|_F \\
&\leq O(\sqrt{m}) \max_{j,r} \gamma_{j,r}^{(T_1)} \|\boldsymbol{\mu}\|_2^{-1} + O(\sqrt{m}) \max_{j,r} \bigg\| \sum_{i=1}^n \overline{\rho}_{j,r,i}^{(T_1)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2} + \sum_{i=1}^n \underline{\rho}_{j,r,i}^{(T_1)} \cdot \frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|_2^2} \bigg\|_2 \\
&\quad + O(m^{1/2}n^{1/2}\log(1/\epsilon)\sigma_p^{-1}d^{-1/2}) \\
&= O(m^{1/2}\widehat{\gamma}\|\boldsymbol{\mu}\|_2^{-1}) + \widetilde{O}(m^{1/2}n^{1/2}\sigma_p^{-1}d^{-1/2}) + O(m^{1/2}n^{1/2}\log(1/\epsilon)\sigma_p^{-1}d^{-1/2}) \\
&= O(m^{1/2}n \cdot \mathrm{SNR} \cdot \sigma_p^{-1}d^{-1/2}) + \widetilde{O}(m^{1/2}n^{1/2}\log(1/\epsilon)\sigma_p^{-1}d^{-1/2}) \\
&= \widetilde{O}(m^{1/2}n^{1/2}\sigma_p^{-1}d^{-1/2}),
\end{aligned}$$

where the first inequality is by triangle inequality, the second inequality and the first equality are by our decomposition of $\mathbf{W}^{(T_1)}$, $\mathbf{W}^*$ and Lemma B.4; the second equality is by $n \cdot \mathrm{SNR}^2 = \Theta(\widehat{\gamma})$ and $\mathrm{SNR} = \|\boldsymbol{\mu}\|/\sigma_p d^{1/2}$; the third equality is by $n^{1/2} \cdot \mathrm{SNR} = O(1)$. □

Lemma D.3. *Under the same conditions as Theorem 4.2, we have that*

$$y_i \langle \nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^* \rangle \geq \log(2/\epsilon)$$

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

Proof of Lemma D.3. Recall that $f(\mathbf{W}^{(t)}) = (1/m)\sum_{j,r} j \cdot [\sigma(\langle \mathbf{w}_{j,r}, y_i \cdot \boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}_i\rangle)]$, thus we have

$$\begin{aligned}
&y_i \langle \nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^* \rangle \\
&= \frac{1}{m}\sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \boldsymbol{\mu}\rangle)\langle \boldsymbol{\mu}, j\mathbf{w}_{j,r}^*\rangle + \frac{1}{m}\sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\langle y_i\boldsymbol{\xi}_i, j\mathbf{w}_{j,r}^*\rangle \\
&= \frac{1}{m}\sum_{j,r}\sum_{i'=1}^n \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) 5\log(2/\epsilon)\, \mathbb{1}(j = y_{i'}) \cdot \frac{\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle}{\|\boldsymbol{\xi}_{i'}\|_2^2} \\
&\quad + \frac{1}{m}\sum_{j,r}\sum_{i'=1}^n \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \boldsymbol{\mu}\rangle) 5\log(2/\epsilon)\, \mathbb{1}(j = y_{i'}) \cdot \frac{\langle \boldsymbol{\mu}, \boldsymbol{\xi}_{i'}\rangle}{\|\boldsymbol{\xi}_{i'}\|_2^2} \\
&\quad + \frac{1}{m}\sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i \boldsymbol{\mu}\rangle)\langle \boldsymbol{\mu}, j\mathbf{w}_{j,r}^{(0)}\rangle + \frac{1}{m}\sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\langle y_i\boldsymbol{\xi}_i, j\mathbf{w}_{j,r}^{(0)}\rangle \\
&\geq \frac{1}{m}\sum_{j=y_i,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) 5\log(2/\epsilon) - \frac{1}{m}\sum_{j,r}\sum_{i' \neq i} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) 5\log(2/\epsilon) \cdot \frac{|\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle|}{\|\boldsymbol{\xi}_{i'}\|_2^2}
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{m}\sum_{j,r}\sum_{i'=1}^{n}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\boldsymbol{\mu}\rangle)5\log(2/\epsilon)\cdot\frac{|\langle\boldsymbol{\mu},\boldsymbol{\xi}_{i'}\rangle|}{\|\boldsymbol{\xi}_{i'}\|_2^2}\\
&- \frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\boldsymbol{\mu}\rangle)O\big(\sqrt{\log(m/\delta)}\cdot\sigma_0\|\boldsymbol{\mu}\|_2\big) - \frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)O\big(\sqrt{\log(mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}\big)\\
\geq\ & \underbrace{\frac{1}{m}\sum_{j=y_i,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)5\log(2/\epsilon)}_{I_9} - \underbrace{\frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)5\log(2/\epsilon)O\big(n\sqrt{\log(n/\delta)}/\sqrt{d}\big)}_{I_{10}}\\
&\underbrace{- \frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}_i\boldsymbol{\mu}\rangle)5\log(2/\epsilon)O\big(n\sqrt{\log(n/\delta)}\cdot\mathrm{SNR}\cdot d^{-1/2}\big)}_{I_{11}}\\
&\underbrace{- \frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i\boldsymbol{\mu}\rangle)O\big(\sqrt{\log(m/\delta)}\cdot\sigma_0\|\boldsymbol{\mu}\|_2\big)}_{I_{12}} - \underbrace{\frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)O\big(\sqrt{\log(mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}\big)}_{I_{14}},
\end{aligned}
\tag{D.7}
$$

where the first inequality is by Lemma B.5 and the last inequality is by Lemma B.4. Next, we will bound the inner-product terms in (D.7) respectively. For I_{10}, I_{11}, I_{12}, I_{14}, note that $\sigma' \in \{0,1\}$ we have that

$$
\begin{aligned}
&|I_{10}| \leq \log(2/\epsilon)O\big(n\sqrt{\log(n/\delta)}/\sqrt{d}\big),\ |I_{11}| \leq \log(2/\epsilon)O\big(n\sqrt{\log(n/\delta)}\cdot\mathrm{SNR}\cdot d^{-1/2}\big),\\
&|I_{12}| \leq O\big(\sqrt{\log(m/\delta)}\cdot\sigma_0\|\boldsymbol{\mu}\|_2\big),\ |I_{14}| \leq O\big(\sqrt{\log(mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}\big).
\end{aligned}
\tag{D.8}
$$

For $j = y_i$ and $r \in S_i^{(0)}$, according to Lemma C.3, we have

$$
\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle &\geq \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \overline{\rho}_{j,r,i}^{(t)} - 5n\sqrt{\frac{\log(4n^2/\delta)}{d}}\alpha\\
&\geq 2 - \beta - 5n\sqrt{\frac{\log(4n^2/\delta)}{d}}\alpha\\
&\geq 1
\end{aligned}
$$

where the first inequality is by Lemma C.3; the last inequality is by $\beta \leq 0.5$ and $5n\sqrt{\frac{\log(4n^2/\delta)}{d}} \leq 0.5$. Therefore, for I_9, according to the fourth statement of Proposition C.8, we have

$$
I_9 \geq \frac{1}{m}|S_i^{(t)}|5\log(2/\epsilon) \geq 2\log(2/\epsilon). \tag{D.9}
$$

By plugging (D.8) and (D.9) into (D.7) and according to triangle inequality we have

$$
y_i\langle\nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^*\rangle \geq I_9 - |I_{10}| - |I_{11}| - |I_{12}| - |I_{14}| \geq \log(2/\epsilon),
$$

which completes the proof. □

Lemma D.4. *Under Condition 4.1, for $0 \leq t \leq T^*$, the following result holds.*

$$
\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \leq O(\max\{\|\boldsymbol{\mu}\|_2^2, \sigma_p^2 d\})L_S(\mathbf{W}^{(t)}).
$$

Proof of Lemma D.4. We first prove that

$$
\|\nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i)\|_F = O(\max\{\|\boldsymbol{\mu}\|_2, \sigma_p\sqrt{d}\}). \tag{D.10}
$$

Without loss of generality, we suppose that $\widehat{y}_i = 1$ and $\mathbf{x}_i = [\boldsymbol{\mu}^\top, \boldsymbol{\xi}_i]$. Then we have that

$$\begin{aligned}
\|\nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i)\|_F &\leq \frac{1}{m}\sum_{j,r}\Big\|\big[\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}\rangle)\boldsymbol{\mu} + \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\boldsymbol{\xi}_i\big]\Big\|_2 \\
&\leq \frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}\rangle)\|\boldsymbol{\mu}\|_2 + \frac{1}{m}\sum_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2 \\
&\leq 4\max\{\|\boldsymbol{\mu}\|_2, 2\sigma_p\sqrt{d}\},
\end{aligned}$$

where the first and second inequalities are by triangle inequality, the third inequality is by Lemma B.4 and $\sigma' \leq 1$. Now we can upper bound the gradient norm $\|\nabla L_S(\mathbf{W}^{(t)})\|_F$ as follows,

$$\begin{aligned}
\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 &\leq \bigg[\frac{1}{n}\sum_{i=1}^n \ell'\big(y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)\big)\|\nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i)\|_F\bigg]^2 \\
&\leq \bigg[\frac{1}{n}\sum_{i=1}^n O(\max\{\|\boldsymbol{\mu}\|_2^2, \sigma_p^2 d\}) - \ell'\big(y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)\big)\bigg]^2 \\
&\leq O(\max\{\|\boldsymbol{\mu}\|_2^2, \sigma_p^2 d\}) \cdot \frac{1}{n}\sum_{i=1}^n -\ell'\big(y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)\big) \\
&\leq O(\max\{\|\boldsymbol{\mu}\|_2^2, \sigma_p^2 d\}) L_S(\mathbf{W}^{(t)}),
\end{aligned}$$

where the first inequality is by triangle inequality, the second inequality is by (D.10), the third inequality is by Cauchy-Schwartz inequality and the last inequality is due to the property of the cross entropy loss $-\ell' \leq \ell$. □

Lemma D.5. *Under the same conditions as Theorem 4.2, we have that*

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

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

Proof of Lemma D.5. We have

$$\begin{aligned}
&\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \\
&= 2\eta\langle \nabla L_S(\mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^*\rangle - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \\
&= \frac{2\eta}{n}\sum_{i=1}^n \ell_i'^{(t)}[y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) - \langle \nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^*\rangle] - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \\
&\geq \frac{2\eta}{n}\sum_{i=1}^n \ell_i'^{(t)}[y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) - \log(2/\epsilon)] - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \\
&\geq \frac{2\eta}{n}\sum_{i=1}^n [\ell\big(y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)\big) - \epsilon/2] - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \\
&\geq \eta L_S(\mathbf{W}^{(t)}) - \eta\epsilon,
\end{aligned}$$

where the first inequality is by Lemma D.3; the second inequality is due to the convexity of the cross entropy function; the last inequality is due to Lemma D.4. □

Lemma D.6. *Under the same conditions as Theorem 4.2, for all $T_1 \leq t \leq T^*$, we have* $\max_{j,r,i}|\underline{\rho}_{j,r,i}^{(t)}| =$

$\max\big\{O\big(\sqrt{\log(mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}\big), O\big(n\sqrt{\log(n/\delta)}\log(T^*)/\sqrt{d}\big)\big\}$. *Besides,*

$$\frac{1}{t-T_1+1}\sum_{s=T_1}^{t} L_S(\mathbf{W}^{(s)}) \le \frac{\|\mathbf{W}^{(T_1)}-\mathbf{W}^*\|_F^2}{\eta(t-T_1+1)}+\epsilon$$

for all $T_1 \le t \le T^*$. *Therefore, we can find an iterate with training loss smaller than* 2ϵ *within* $T = T_1 + \Big\lfloor \|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F^2/(\eta\epsilon)\Big\rfloor = T_1 + \widetilde{O}(\eta^{-1}\epsilon^{-1}mnd^{-1}\sigma_p^{-2})$ *iterations.*

Proof of Lemma D.6. Note that $\max_{j,r,i}|\underline{\rho}_{j,r,i}^{(t)}| = \max\big\{O\big(\sqrt{\log(mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}\big), O\big(n\sqrt{\log(n/\delta)}\log(T^*)/\sqrt{d}\big)\big\}$ can be proved in the same way as Lemma D.1, we eliminate the proof details here. For any $t\in[T_1,T]$, by taking a summation of the inequality in Lemma D.5 and dividing $(t-T_1+1)$ on both sides, we obtain that

$$\frac{1}{t-T_1+1}\sum_{s=T_1}^{t} L_S(\mathbf{W}^{(s)}) \le \frac{\|\mathbf{W}^{(T_1)}-\mathbf{W}^*\|_F^2}{\eta(t-T_1+1)}+\epsilon$$

for all $T_1 \le t \le T$. According to the definition of T, we have

$$\frac{1}{T-T_1+1}\sum_{s=T_1}^{T} L_S(\mathbf{W}^{(s)}) \le 2\epsilon.$$

Then there exists iteration $T_1 \le t \le T$ such that the training loss is smaller than ϵ. □

Besides, we have the following lemma about the order of $\overline{\rho}_{j,r,i}^{(t)}, \gamma_{j,r}^{(t)}$ ratio when training loss is smaller than ϵ. And this lemma will help us prove the theorem about test error.

Lemma D.7. *Under the same conditions as Theorem 4.2, we have*

$$\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}/\gamma_{j',r'}^{(t)} = \Theta(\mathrm{SNR}^{-2}) \tag{D.11}$$

for all $j,j'\in\{\pm1\}$, $r,r'\in[m]$ *and* $T_1\le t\le T^*$.

Proof of Lemma D.7. We will prove this lemma by using induction. We first verify that (D.11) holds for $t=T_1$. By Lemma D.1, we have $\gamma_{j',r'}^{(T_1)} = \Theta(\widehat{\gamma}) = \Theta(n\cdot\mathrm{SNR}^2)$ and $\sum_{i=1}^n \overline{\rho}_{j,r,i}^{(T_1)} = \Theta(n)$, and (D.11) follows directly. Now suppose that there exists $\widetilde{T}\in[T_1,T^*]$ such that $\sum_{i=1}^n \overline{\rho}_{j,r,i}^{(t)}/\gamma_{j',r'}^{(t)} = \Theta(\mathrm{SNR}^2)$ for all $t\in[T_1,\widetilde{T}-1]$. Then for $\overline{\rho}_{j,r,i}^{(t)}$, according to Lemma C.1, we have

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm}\cdot\ell_i'^{(t)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}_i\rangle)\cdot\mathbb{1}(y_i=j)\|\boldsymbol{\xi}_i\|_2^2.$$

$$\gamma_{j',r'}^{(t+1)} = \gamma_{j',r'}^{(t)} - \frac{\eta}{nm}\cdot\bigg[\sum_{i\in S_+}\ell_i'^{(t)}\sigma'(\langle\mathbf{w}_{j',r'}^{(t)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle) - \sum_{i\in S_-}\ell_i'^{(t)}\sigma'(\langle\mathbf{w}_{j',r'}^{(t)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle)\bigg]\cdot\|\boldsymbol{\mu}\|_2^2$$

It follows that

$$\begin{aligned}\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T})} = \sum_{i:y_i=j}\overline{\rho}_{j,r,i}^{(\widetilde{T})} &= \sum_{i:y_i=j}\overline{\rho}_{j,r,i}^{(\widetilde{T}-1)} - \frac{\eta}{nm}\cdot\sum_{i:y_i=j}\ell_i'^{(\widetilde{T}-1)}\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(\widetilde{T}-1)},\boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2\\ &= \sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T}-1)} - \frac{\eta}{nm}\cdot\sum_{i\in S_{j,r}^{(\widetilde{T}-1)}}\ell_i'^{(\widetilde{T}-1)}\|\boldsymbol{\xi}_i\|_2^2\\ &\geq \sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T}-1)} + \frac{\eta\sigma_p^2 d}{16m}\cdot\min_{i\in S_{j,r}^{(\widetilde{T}-1)}}|\ell_i'^{(\widetilde{T}-1)}|,\end{aligned} \tag{D.12}$$

where the last equality is by the definition of $S_{j,r}^{(\widetilde{T}-1)}$ as $\{i\in[n]: y_i=j, \langle\mathbf{w}_{j,r}^{(\widetilde{T}-1)},\boldsymbol{\xi}_i\rangle>0\}$; the last inequality is by Lemma B.4 and the fifth statement of Proposition C.8. And

$$\begin{aligned}\gamma_{j',r'}^{(\widetilde{T})} &\leq \gamma_{j',r'}^{(\widetilde{T}-1)} - \frac{\eta}{nm}\cdot\sum_{i\in S_+}\ell_i'^{(\widetilde{T}-1)}\sigma'(\langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle)\cdot\|\boldsymbol{\mu}\|_2^2\\ &\leq \gamma_{j',r'}^{(\widetilde{T}-1)} + \frac{\eta\|\boldsymbol{\mu}\|_2^2}{m}\cdot\max_{i\in S_+}|\ell_i'^{(\widetilde{T}-1)}|.\end{aligned} \tag{D.13}$$

According to the third statement of Proposition C.8, we have $\max_{i\in S_+}|\ell_i'^{(\widetilde{T}-1)}| \leq C_2\min_{i\in S_{j,r}^{(\widetilde{T}-1)}}|\ell_i'^{(\widetilde{T}-1)}|$. Then by combining (D.12) and (D.13), we have

$$\frac{\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T})}}{\gamma_{j',r'}^{(\widetilde{T})}} \geq \min\left\{\frac{\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T}-1)}}{\gamma_{j',r'}^{(\widetilde{T}-1)}}, \frac{\sigma_p^2 d}{16C_2\|\boldsymbol{\mu}\|_2^2}\right\} = \Theta(\mathrm{SNR}^{-2}). \tag{D.14}$$

On the other hand, according to (D.12) and by Lemma B.4, we have

$$\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T})} \leq \sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(\widetilde{T}-1)} + \frac{9\eta\sigma_p^2 d}{8m}\cdot\max_{i\in S_{j,r}^{(\widetilde{T}-1)}}|\ell_i'^{(\widetilde{T}-1)}|, \tag{D.15}$$

where the inequality is by $|S_{j,r}^{(\widetilde{T}-1)}| \leq |S_j| \leq 3n/4$. And by arguing in a similar way as (D.3), (D.4), (D.5) and (D.6), we can obtain that as long as $q < C_2/6$ and $n \geq 72C_2^{-2}\log(8/\delta)$, it holds that

$$\sum_{i\in S_+}|\ell_i'^{(\widetilde{T}-1)}|\sigma'(\langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle) \geq 2\sum_{i\in S_-}|\ell_i'^{(\widetilde{T}-1)}|\sigma'(\langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle)$$

and hence

$$\begin{aligned}\gamma_{j',r'}^{(\widetilde{T})} &= \gamma_{j',r'}^{(\widetilde{T}-1)} - \frac{\eta}{nm}\cdot\left[\sum_{i\in S_+}\ell_i'^{(\widetilde{T}-1)}\sigma'(\langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle) - \sum_{i\in S_-}\ell_i'^{(\widetilde{T}-1)}\sigma'(\langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle)\right]\cdot\|\boldsymbol{\mu}\|_2^2\\ &\geq \gamma_{j',r'}^{(\widetilde{T}-1)} - \frac{\eta}{2nm}\cdot\sum_{i\in S_+}\ell_i'^{(\widetilde{T}-1)}\sigma'(\langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\widehat{y}_i\cdot\boldsymbol{\mu}\rangle)\cdot\|\boldsymbol{\mu}\|_2^2.\end{aligned}$$

Then we have

$$\begin{aligned}\gamma_{j',r'}^{(\widetilde{T})} &\geq \gamma_{j',r'}^{(\widetilde{T}-1)} - \frac{\eta}{2nm}\sum_{i\in S_+\cap S_1}\ell_i'^{(\widetilde{T}-1)}\cdot\|\boldsymbol{\mu}\|_2^2 \geq \gamma_{j',r'}^{(\widetilde{T}-1)} + \frac{\eta\|\boldsymbol{\mu}\|_2^2}{8m}\min_{i\in S_+\cap S_1}\ell_i'^{(\widetilde{T}-1)}, \text{ if } \langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\boldsymbol{\mu}\rangle \geq 0,\\ \gamma_{j',r'}^{(\widetilde{T})} &\geq \gamma_{j',r'}^{(\widetilde{T}-1)} - \frac{\eta}{2nm}\sum_{i\in S_+\cap S_{-1}}\ell_i'^{(\widetilde{T}-1)}\cdot\|\boldsymbol{\mu}\|_2^2 \geq \gamma_{j',r'}^{(\widetilde{T}-1)} + \frac{\eta\|\boldsymbol{\mu}\|_2^2}{8m}\min_{i\in S_+\cap S_{-1}}\ell_i'^{(\widetilde{T}-1)}, \text{ if } \langle\mathbf{w}_{j',r'}^{(\widetilde{T}-1)},\boldsymbol{\mu}\rangle < 0,\end{aligned} \tag{D.16}$$

where the second inequality is by Lemma B.3. According to the fourth statement of Proposition C.8, we have $\max_{i\in S_{j,r}^{(\widetilde{T}-1)}} |\ell_i'^{(\widetilde{T}-1)}| \leq C_2 \min_{i\in S_+\cap S_1} \ell_i'^{(\widetilde{T}-1)}$ and $\max_{i\in S_{j,r}^{(\widetilde{T}-1)}} |\ell_i'^{(\widetilde{T}-1)}| \leq C_2 \min_{i\in S_+\cap S_{-1}} \ell_i'^{(\widetilde{T}-1)}$. Then by combining (D.15) and (D.16), we have

$$\frac{\sum_{i=1}^n \overline{\rho}_{j,r,i}^{(\widetilde{T})}}{\gamma_{j',r'}^{(\widetilde{T})}} \leq \max\left\{\frac{\sum_{i=1}^n \overline{\rho}_{j,r,i}^{(\widetilde{T}-1)}}{\gamma_{j',r'}^{(\widetilde{T}-1)}}, \frac{9C_2\sigma_p^2 d}{\|\boldsymbol{\mu}\|_2^2}\right\} = \Theta(\mathrm{SNR}^{-2}). \tag{D.17}$$

By (D.14) and (D.17), we have

$$\frac{\sum_{i=1}^n \overline{\rho}_{j,r,i}^{(\widetilde{T})}}{\gamma_{j',r'}^{(\widetilde{T})}} = \Theta(\mathrm{SNR}^{-2}),$$

which completes the induction.

□

Actually, the result in Lemma D.7 also holds for $0 \leq t \leq T^*$, that is,

Lemma D.8. *Under the same conditions as Theorem 4.2, we have*

$$\sum_{i=1}^n \overline{\rho}_{j,r,i}^{(t)}/\gamma_{j',r'}^{(t)} = \Theta(\mathrm{SNR}^{-2}) \tag{D.18}$$

for all $j, j' \in \{\pm 1\}$, $r, r' \in [m]$ and $0 \leq t \leq T^*$.

The proof argument is nearly the same as Lemma D.7, and we only need to use Lemma D.7 in later arguments, so we eliminate the proof details here.

E. Test Error Analysis

E.1. Test Error Upper Bound

Next, we give an upper bound for the test error at iteration t defined in Theorem 4.2 when the training loss converges to ϵ. First of all, notice that $T_1 \leq t \leq T^*$ by Lemma D.6, we can summarize previous results into the following:

- $\sum_{i=1}^n \overline{\rho}_{j,r,i}^{(t)}/\gamma_{j',r'}^{(t)} = \Theta(\mathrm{SNR}^{-2})$ (from Lemma D.7),
- $\sum_{i=1}^n \overline{\rho}_{j,r,i}^{(t)} = \Omega(n) = O(n\log(T^*)) = \widetilde{\Theta}(n)$ (from Proposition C.2 and Lemma D.1)
- $\max_{j,r,i} |\underline{\rho}_{j,r,i}^{(t)}| = \max\left\{O\big(\sqrt{\log(mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d}\big), O\big(\sqrt{\log(n/\delta)}\log(T^*)\cdot n/\sqrt{d}\big)\right\}$ (from Lemma D.6).

Additionally, recalling the definition $\widehat{\gamma} = n\cdot\mathrm{SNR}^2$, from the first two conclusions, we have $\gamma_{j,r}^{(t)} = \widetilde{\Theta}(\widehat{\gamma})$ for all j, r. Also note that from the third conclusion, since $\sigma_0\sigma_p\sqrt{d} = \widetilde{O}(\sqrt{n}/\sqrt{d}) = o(1)$ and $\sqrt{\log(n/\delta)}\log(T^*)\cdot n/\sqrt{d} = O(1)$ from Condition 4.1, we have $\max_{j,r,i} |\underline{\rho}_{j,r,i}^{(t)}| = O(1)$ and so $\sum_{i=1}^n |\underline{\rho}_{\widehat{y},r,i}^{(t)}| = O\big(\sum_{i=1}^n \overline{\rho}_{\widehat{y},r,i}^{(t)}\big)$, hence we can ignore the sum of $\underline{\rho}$ whenever it appears together with the sum of $\overline{\rho}$. We are now ready to analyze the test error in the following theorem.

Theorem E.1 (Second part of Theorem 4.2)**.** *Under the same conditions as Theorem 4.2, then there exists a large constant C_1 such that when $n\|\boldsymbol{\mu}\|_2^2 \geq C_1\sigma_p^4 d$, for time t defined in Lemma D.6, we have the test error*

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(y \neq \mathrm{sign}(f(\mathbf{W}^{(t)},\mathbf{x}))\big) \leq p + \exp\bigg(-n\|\boldsymbol{\mu}\|_2^4/(C_2\sigma_p^4 d)\bigg),$$

where $C_2 = O(1)$.

Proof. For the sake of convenience, we use $(\mathbf{x}, \widehat{y}, y) \sim \mathcal{D}$ to denote the following: data point $(\mathbf{x}, y)$ follows distribution $\mathcal{D}$ defined in Definition 1.1, and $\widehat{y}$ is its true label. We can write out the test error as

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(y \neq \mathrm{sign}(f(\mathbf{W}^{(t)}, \mathbf{x}))\big) \\
&= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(yf(\mathbf{W}^{(t)}, \mathbf{x}) \leq 0\big) \\
&= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(yf(\mathbf{W}^{(t)}, \mathbf{x}) \leq 0, y \neq \widehat{y}\big) + \mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(yf(\mathbf{W}^{(t)}, \mathbf{x}) \leq 0, y = \widehat{y}\big) \\
&= p \cdot \mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)}, \mathbf{x}) \geq 0\big) + (1-p) \cdot \mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)}, \mathbf{x}) \leq 0\big) \\
&\leq p + \mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)}, \mathbf{x}) \leq 0\big),
\end{aligned} \tag{E.1}$$

where in the second equation we used the definition of $\mathcal{D}$ in Definition 1.1. It therefore suffices to provide an upper bound for $\mathbb{P}_{(\mathbf{x},\widehat{y})\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)}, \mathbf{x}) \leq 0\big)$. To achieve this, we write $\mathbf{x} = (\widehat{y}\boldsymbol{\mu}, \boldsymbol{\xi})$, and get

$$\begin{aligned}
\widehat{y}f(\mathbf{W}^{(t)}, \mathbf{x}) &= \frac{1}{m}\sum_{j,r} \widehat{y}j[\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle)] \\
&= \frac{1}{m}\sum_{r}[\sigma(\langle \mathbf{w}_{\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{\widehat{y},r}^{(t)}, \boldsymbol{\xi}\rangle)] - \frac{1}{m}\sum_{r}[\sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}\rangle)]
\end{aligned} \tag{E.2}$$

Now consider first the expressions $\langle \mathbf{w}_{j,r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle$ for $j = \pm\widehat{y}$. Recall from (5.1) the signal-noise decomposition of $\mathbf{w}_{j,r}^{(t)}$:

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

hence the inner product with $j = \widehat{y}$ can be bounded as

$$\begin{aligned}
\langle \mathbf{w}_{\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle &= \langle \mathbf{w}_{\widehat{y},r}^{(0)}, \widehat{y}\boldsymbol{\mu}\rangle + \gamma_{\widehat{y},r}^{(t)} + \sum_{i=1}^{n} \overline{\rho}_{\widehat{y},r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \langle \boldsymbol{\xi}_i, \widehat{y}\boldsymbol{\mu}\rangle + \sum_{i=1}^{n} \underline{\rho}_{\widehat{y},r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \langle \boldsymbol{\xi}_i, \widehat{y}\boldsymbol{\mu}\rangle \\
&\geq \gamma_{\widehat{y},r}^{(t)} - \sqrt{2\log(12m/\delta)} \cdot \sigma_0\|\boldsymbol{\mu}\|_2 \\
&\qquad - \sqrt{2\log(6n/\delta)} \cdot \sigma_p\|\boldsymbol{\mu}\|_2 \cdot (\sigma_p^2 d/2)^{-1}\bigg[\sum_{i=1}^{n} \overline{\rho}_{\widehat{y},r,i}^{(t)} + \sum_{i=1}^{n} |\underline{\rho}_{\widehat{y},r,i}^{(t)}|\bigg] \\
&= \gamma_{\widehat{y},r}^{(t)} - \Theta(\sqrt{\log(m/\delta)}\sigma_0\|\boldsymbol{\mu}\|_2) - \Theta\big(\sqrt{\log(n/\delta)} \cdot (\sigma_p d)^{-1}\|\boldsymbol{\mu}\|_2\big) \cdot \Theta(\mathrm{SNR}^{-2}) \cdot \gamma_{\widehat{y},r}^{(t)} \\
&= \big[1 - \Theta\big(\sqrt{\log(n/\delta)} \cdot \sigma_p/\|\boldsymbol{\mu}\|_2\big)\big]\gamma_{\widehat{y},r}^{(t)} - \Theta\big(\sqrt{\log(m/\delta)}(\sigma_p d)^{-1}\sqrt{n}\|\boldsymbol{\mu}\|_2\big) \\
&= \Theta(\gamma_{\widehat{y},r}^{(t)}),
\end{aligned} \tag{E.3}$$

where the inequality is by Lemma B.4 and Lemma B.5; the second equality is obtained by plugging in the coefficient orders we summarized at the start of the section; the third equality is by the condition $\sigma_0 \leq C^{-1}(\sigma_p d)^{-1}\sqrt{n}$ in Condition 4.1 and $\mathrm{SNR} = \|\boldsymbol{\mu}\|_2/\sigma_p\sqrt{d}$; for the fourth equality, notice that $\gamma_{j,r}^{(t)} = \Omega(\widehat{\gamma})$, also $\sqrt{\log(n/\delta)} \cdot \sigma_p/\|\boldsymbol{\mu}\|_2 \leq 1/\sqrt{C}$ and $\sqrt{\log(m/\delta)}(\sigma_p d)^{-1}\sqrt{n}\|\boldsymbol{\mu}\|_2/\widehat{\gamma} = \sqrt{\log(m/\delta)}\sigma_p/(\sqrt{n}\|\boldsymbol{\mu}\|_2) \leq \sqrt{\log(m/\delta)/n} \cdot 1/\big(\sqrt{C\log(n/\delta)}\big) \leq 1/(C\sqrt{\log(n/\delta)})$ holds by $\|\boldsymbol{\mu}\|_2^2 \geq C \cdot \sigma_p^2\log(n/\delta)$ and $n \geq C\log(m/\delta)$ in Condition 4.1, so for sufficiently large constant C the equality

holds. Moreover, we can deduce in a similar manner that

$$\begin{aligned}\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle &= \langle \mathbf{w}_{-\widehat{y},r}^{(0)}, \widehat{y}\boldsymbol{\mu}\rangle - \gamma_{-\widehat{y},r}^{(t)} + \sum_{i=1}^{n} \overline{\rho}_{-\widehat{y},r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \langle \boldsymbol{\xi}_i, -\widehat{y}\boldsymbol{\mu}\rangle + \sum_{i=1}^{n} \underline{\rho}_{-\widehat{y},r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \langle \boldsymbol{\xi}_i, \widehat{y}\boldsymbol{\mu}\rangle \\ &\leq -\gamma_{-\widehat{y},r}^{(t)} + \sqrt{2\log(8m/\delta)} \cdot \sigma_0 \|\boldsymbol{\mu}\|_2 \\ &\quad + \sqrt{2\log(6n/\delta)} \cdot \sigma_p \|\boldsymbol{\mu}\|_2 \cdot (\sigma_p^2 d/2)^{-1} \Big[\sum_{i=1}^{n} \overline{\rho}_{-\widehat{y},r,i}^{(t)} + \sum_{i=1}^{n} |\underline{\rho}_{-\widehat{y},r,i}^{(t)}| \Big] \\ &= -\Theta(\gamma_{-\widehat{y},r}^{(t)}) < 0,\end{aligned} \tag{E.4}$$

where the second equality holds based on similar analyses as in (E.3).

Denote $g(\boldsymbol{\xi})$ as $\sum_r \sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}\rangle)$. According to Theorem 5.2.2 in Vershynin (2018), we know that for any $x \geq 0$ it holds that

$$\mathbb{P}(g(\boldsymbol{\xi}) - \mathbb{E}g(\boldsymbol{\xi}) \geq x) \leq \exp\Big(- \frac{cx^2}{\sigma_p^2 \|g\|_{\mathrm{Lip}}^2} \Big), \tag{E.5}$$

where c is a constant. To calculate the Lipschitz norm, we have

$$\begin{aligned}|g(\boldsymbol{\xi}) - g(\boldsymbol{\xi}')| &= \Big| \sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}\rangle) - \sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}'\rangle) \Big| \\ &\leq \sum_{r=1}^{m} \big| \sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}\rangle) - \sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}'\rangle) \big| \\ &\leq \sum_{r=1}^{m} |\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi} - \boldsymbol{\xi}'\rangle| \\ &\leq \sum_{r=1}^{m} \big\| \mathbf{w}_{-\widehat{y},r}^{(t)} \big\|_2 \cdot \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2,\end{aligned}$$

where the first inequality is by triangle inequality; the second inequality is by the property of ReLU; the last inequality is by Cauchy-Schwartz inequality. Therefore, we have

$$\|g\|_{\mathrm{Lip}} \leq \sum_{r=1}^{m} \big\| \mathbf{w}_{-\widehat{y},r}^{(t)} \big\|_2, \tag{E.6}$$

and since $\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}\rangle \sim \mathcal{N}\big(0, \|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2^2 \sigma_p^2\big)$, we can get

$$\mathbb{E}g(\boldsymbol{\xi}) = \sum_{r=1}^{m} \mathbb{E}\sigma(\langle \mathbf{w}_{-\widehat{y},r}^{(t)}, \boldsymbol{\xi}\rangle) = \sum_{r=1}^{m} \frac{\|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2 \sigma_p}{\sqrt{2\pi}} = \frac{\sigma_p}{\sqrt{2\pi}} \sum_{r=1}^{m} \|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2.$$

Next we seek to upper bound the 2-norm of $\mathbf{w}_{j,r}^{(t)}$. First, we tackle the noise section in the decomposition, namely:

$$\begin{aligned}&\Big\| \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i \Big\|_2^2 \\ &= \sum_{i=1}^{n} {\rho_{j,r,i}^{(t)}}^2 \cdot \|\boldsymbol{\xi}_i\|_2^{-2} + 2 \sum_{1 \leq i_1 < i_2 \leq n} \rho_{j,r,i_1}^{(t)} \rho_{j,r,i_2}^{(t)} \cdot \|\boldsymbol{\xi}_{i_1}\|_2^{-2} \cdot \|\boldsymbol{\xi}_{i_2}\|_2^{-2} \cdot \langle \boldsymbol{\xi}_{i_1}, \boldsymbol{\xi}_{i_2}\rangle \\ &\leq 4\sigma_p^{-2} d^{-1} \sum_{i=1}^{n} {\rho_{j,r,i}^{(t)}}^2 + 2 \sum_{1 \leq i_1 < i_2 \leq n} |\rho_{j,r,i_1}^{(t)} \rho_{j,r,i_2}^{(t)}| \cdot (16\sigma_p^{-4} d^{-2}) \cdot (2\sigma_p^2 \sqrt{d\log(6n^2/\delta)})\end{aligned}$$

$$
\begin{aligned}
&= 4\sigma_p^{-2}d^{-1}\sum_{i=1}^{n}{\rho_{j,r,i}^{(t)}}^2 + 32\sigma_p^{-2}d^{-3/2}\sqrt{\log(6n^2/\delta)}\bigg[\bigg(\sum_{i=1}^{n}|\rho_{j,r,i}^{(t)}|\bigg)^2 - \sum_{i=1}^{n}{\rho_{j,r,i}^{(t)}}^2\bigg] \\
&= \Theta(\sigma_p^{-2}d^{-1})\sum_{i=1}^{n}{\rho_{j,r,i}^{(t)}}^2 + \widetilde{\Theta}(\sigma_p^{-2}d^{-3/2})\bigg(\sum_{i=1}^{n}|\rho_{j,r,i}^{(t)}|\bigg)^2 \\
&\leq \big[\Theta(\sigma_p^{-2}d^{-1}n^{-1}) + \widetilde{\Theta}(\sigma_p^{-2}d^{-3/2})\big]\bigg(\sum_{i=1}^{n}|\overline{\rho}_{j,r,i}^{(t)}| + \sum_{i=1}^{n}|\underline{\rho}_{j,r,i}^{(t)}|\bigg)^2 \\
&\leq \Theta(\sigma_p^{-2}d^{-1}n^{-1})\bigg(\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}\bigg)^2
\end{aligned}
$$

where for the first inequality we used Lemma B.4; for the second inequality we used the definition of $\overline{\rho}, \underline{\rho}$; for the second to last equation we plugged in coefficient orders. We can thus upper bound the norm of $\mathbf{w}_{j,r}^{(t)}$ as:

$$
\begin{aligned}
\|\mathbf{w}_{j,r}^{(t)}\|_2 &\leq \|\mathbf{w}_{j,r}^{(0)}\|_2 + \gamma_{j,r}^{(t)}\cdot\|\boldsymbol{\mu}\|_2^{-1} + \bigg\|\sum_{i=1}^{n}\rho_{j,r,i}^{(t)}\cdot\|\boldsymbol{\xi}_i\|_2^{-2}\cdot\boldsymbol{\xi}_i\bigg\|_2 \\
&\leq \|\mathbf{w}_{j,r}^{(0)}\|_2 + \gamma_{j,r}^{(t)}\cdot\|\boldsymbol{\mu}\|_2^{-1} + \Theta(\sigma_p^{-1}d^{-1/2}n^{-1/2})\cdot\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)} \\
&= \Theta(\sigma_p^{-1}d^{-1/2}n^{-1/2})\cdot\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)} \qquad \text{(E.7)}
\end{aligned}
$$

where the first inequality is due to the triangle inequality, and the equality is due to the following comparisons:

$$
\frac{\gamma_{j,r}^{(t)}\cdot\|\boldsymbol{\mu}\|_2^{-1}}{\Theta(\sigma_p^{-1}d^{-1/2}n^{-1/2})\cdot\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}} = \Theta(\sigma_p d^{1/2}n^{1/2}\|\boldsymbol{\mu}\|_2^{-1}\mathrm{SNR}^2) = \Theta(\sigma_p^{-1}d^{-1/2}n^{1/2}\|\boldsymbol{\mu}\|_2) = O(1)
$$

based on the coefficient order $\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}/\gamma_{j,r}^{(t)} = \Theta(\mathrm{SNR}^{-2})$, the definition $\mathrm{SNR} = \|\boldsymbol{\mu}\|_2/(\sigma_p\sqrt{d})$, and the condition for d in Condition 4.1; and also

$$
\frac{\|\mathbf{w}_{j,r}^{(0)}\|_2}{\Theta(\sigma_p^{-1}d^{-1/2}n^{-1/2})\cdot\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}} = \frac{\Theta(\sigma_0\sqrt{d})}{\Theta(\sigma_p^{-1}d^{-1/2}n^{-1/2})\cdot\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}} = O(\sigma_0\sigma_p d n^{-1/2}) = O(1)
$$

based on Lemma B.5, the coefficient order $\sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)} = \Omega(n)$, and the condition for σ_0 in Condition 4.1. With this and (E.3), we give an analysis of the following the key component,

$$
\frac{\sum_r\sigma(\langle\mathbf{w}_{\widehat{y},r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle)}{\sigma_p\sum_{r=1}^{m}\big\|\mathbf{w}_{-\widehat{y},r}^{(t)}\big\|_2} \geq \frac{\Theta\big(\sum_r\gamma_{\widehat{y},r}^{(t)}\big)}{\Theta(d^{-1/2}n^{-1/2})\cdot\sum_{r,i}\overline{\rho}_{-\widehat{y},r,i}^{(t)}} = \Theta(d^{1/2}n^{1/2}\mathrm{SNR}^2) = \Theta(n^{1/2}\|\boldsymbol{\mu}\|_2^2/\sigma_p^2 d^{1/2}) \qquad \text{(E.8)}
$$

By (E.8) and $n\|\boldsymbol{\mu}\|_2^4 \geq C_1\sigma_p^4 d$ where C_1 is a sufficiently large constant, it directly follows that

$$
\sum_r\sigma(\langle\mathbf{w}_{\widehat{y},r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle) - \frac{\sigma_p}{\sqrt{2\pi}}\sum_{r=1}^{m}\|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2 > 0. \qquad \text{(E.9)}
$$

Now using the method in (E.5) with the results above, we plug (E.4) into (E.2) and then (E.1), to obtain

$$
\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\boldsymbol{W}^{(t)},\mathbf{x})\leq 0\big) \leq \mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\bigg(\sum_r\sigma(\langle\mathbf{w}_{-\widehat{y},r}^{(t)},\boldsymbol{\xi}\rangle) \geq \sum_r\sigma(\langle\mathbf{w}_{\widehat{y},r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle)\bigg)
$$

$$
\begin{aligned}
&= \mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\bigg(g(\boldsymbol{\xi}) - \mathbb{E}g(\boldsymbol{\xi}) \geq \sum_r \sigma(\langle \mathbf{w}_{\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle) - \frac{\sigma_p}{\sqrt{2\pi}}\sum_{r=1}^m \|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2\bigg) \\
&\leq \exp\bigg[-\frac{c\Big(\sum_r \sigma(\langle \mathbf{w}_{\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle) - (\sigma_p/\sqrt{2\pi})\sum_{r=1}^m \|\mathbf{w}_{-\widehat{y},r}^{(t)}\|_2\Big)^2}{\sigma_p^2\Big(\sum_{r=1}^m \big\|\mathbf{w}_{-\widehat{y},r}^{(t)}\big\|_2\Big)^2}\bigg] \\
&= \exp\bigg[-c\Big(\frac{\sum_r \sigma(\langle \mathbf{w}_{\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle)}{\sigma_p\sum_{r=1}^m \big\|\mathbf{w}_{-\widehat{y},r}^{(t)}\big\|_2} - 1/\sqrt{2\pi}\Big)^2\bigg] \\
&\leq \exp(c/2\pi)\exp\bigg(-0.5c\Big(\frac{\sum_r \sigma(\langle \mathbf{w}_{\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle)}{\sigma_p\sum_{r=1}^m \big\|\mathbf{w}_{-\widehat{y},r}^{(t)}\big\|_2}\Big)^2\bigg)
\end{aligned}
\tag{E.10}
$$

where the second inequality is by (E.9) and plugging (E.6) into (E.5), the third inequality is due to the fact that $(s-t)^2 \geq s^2/2 - t^2, \forall s,t \geq 0$.

And we can get from (E.8) and (E.10) that

$$
\begin{aligned}
\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\boldsymbol{W}^{(t)},\mathbf{x}) \leq 0\big) &\leq \exp(c/2\pi)\exp\bigg(-0.5c\Big(\frac{\sum_r \sigma(\langle \mathbf{w}_{\widehat{y},r}^{(t)}, \widehat{y}\boldsymbol{\mu}\rangle)}{\sigma_p\sum_{r=1}^m \big\|\mathbf{w}_{-\widehat{y},r}^{(t)}\big\|_2}\Big)^2\bigg) \\
&= \exp\Big(\frac{c}{2\pi} - \frac{n\|\boldsymbol{\mu}\|_2^4}{C\sigma_p^4 d}\Big) \\
&\leq \exp\Big(-\frac{n\|\boldsymbol{\mu}\|_2^4}{2C\sigma_p^4 d}\Big) \\
&= \exp\Big(-\frac{n\|\boldsymbol{\mu}\|_2^4}{C_2\sigma_p^4 d}\Big),
\end{aligned}
$$

where $C = O(1)$; the last inequality holds if we choose $C_1 \geq cC/\pi$; the last equality holds if we choose C_2 as $2C$.

□

E.2. Test Error Lower Bound

In this section, we will give the lower bound of the test error at iteration t defined in Theorem 4.2 when the training loss converges to ϵ, which, together with Theorem 4.2, shows a sharp phase transition. First, we give the proof of key Lemma 5.8.

Proof of Lemma 5.8. Without loss of generality, let $\max\Big\{\sum_r \gamma_{1,r}^{(t)}, \sum_r \gamma_{-1,r}^{(t)}\Big\} = \sum_r \gamma_{1,r}^{(t)}$. Denote $\mathbf{v} = \lambda \cdot \sum_i \mathbb{1}(y_i = 1)\boldsymbol{\xi}_i$, where $\lambda = C_7\|\boldsymbol{\mu}\|_2^2/(d\sigma_p^2)$ and C_7 is a sufficiently large constant. Then we only need to prove that

$$
\underbrace{g(\boldsymbol{\xi}+\mathbf{v}) - g(\boldsymbol{\xi}) + g(-\boldsymbol{\xi}+\mathbf{v}) - g(-\boldsymbol{\xi})}_{I} \geq 4C_6\sum_r \gamma_{1,r}^{(t)}. \tag{E.11}
$$

Since ReLU is a convex activation function, we have that

$$
\sigma(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}+\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}\rangle) \geq \sigma'(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}\rangle)\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle \tag{E.12}
$$

$$
\sigma(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}+\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}\rangle) \geq \sigma'(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}\rangle)\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle. \tag{E.13}
$$

Adding (E.12) and (E.13) we have that almost surely for all $\boldsymbol{\xi}$

$$\begin{aligned}
&\sigma(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}+\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}\rangle) + \sigma(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}+\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}\rangle) \\
&\geq \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v}\rangle \\
&\geq \lambda\bigg[\sum_{y_i=1} \overline{\rho}_{1,r,i}^{(t)} - 2n\sqrt{\log(12mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d} - 5n^2\alpha\sqrt{\log(6n^2/\delta)/d}\bigg],
\end{aligned} \tag{E.14}$$

where the last inequality is by (C.14) and Lemma B.5. Since ReLU is a Liptchitz, we also have that

$$\begin{aligned}
&\sigma(\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}+\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}\rangle) + \sigma(\langle \mathbf{w}_{-1,r}^{(t)}, -\boldsymbol{\xi}+\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{-1,r}^{(t)}, -\boldsymbol{\xi}\rangle) \\
&\leq 2|\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{v}\rangle| \\
&\leq 2\lambda\bigg[\sum_{y_i=1} \underline{\rho}_{-1,r,i}^{(t)} + 2n\sqrt{\log(12mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d} + 5n^2\alpha\sqrt{\log(6n^2/\delta)/d}\bigg],
\end{aligned} \tag{E.15}$$

where the last inequality is by (C.13) and Lemma B.5. Therefore, by plugging (E.14) and (E.15) into left hand side I in (E.11), we have that

$$\begin{aligned}
&g(\boldsymbol{\xi}+\mathbf{v}) - g(\boldsymbol{\xi}) + g(-\boldsymbol{\xi}+\mathbf{v}) - g(-\boldsymbol{\xi}) \\
&\geq \lambda\bigg[\sum_r\sum_{y_i=1} \overline{\rho}_{1,r,i}^{(t)} - 6nm\sqrt{\log(12mn/\delta)}\cdot\sigma_0\sigma_p\sqrt{d} - 15mn^2\alpha\sqrt{\log(6n^2/\delta)/d}\bigg] \\
&\geq (\lambda/2)\cdot\sum_r\sum_{y_i=1} \overline{\rho}_{1,r,i}^{(t)} \\
&\geq \lambda/2\cdot\Theta(\mathrm{SNR}^{-2})\sum_r \gamma_{1,r}^{(t)} \\
&\geq 4C_6\sum_r \gamma_{1,r}^{(t)},
\end{aligned}$$

where the second inequality is by Lemma D.1 and Condition 4.1; the third inequality is by Lemma D.7. Finally, it is worth noting that the norm

$$\|\mathbf{v}\|_2 = \|\lambda\cdot\sum_i \mathbb{1}(y_i=1)\boldsymbol{\xi}_i\|_2 = \Theta\bigg(\sqrt{\frac{n\|\boldsymbol{\mu}\|_2^4}{\sigma_p^4 d}}\bigg) \leq 0.06\sigma_p,$$

where the last inequality is by condition $n\|\boldsymbol{\mu}\|_2^4 \leq C_3\sigma_p^4 d$ with sufficiently large C_3 in Theorem 4.2, which completes the proof. □

Then we present an important Lemma, which bounds the Total Variation (TV) distance between two Gaussian with the same covariance matrix.

Lemma E.2 (Proposition 2.1 in Devroye et al. (2018))**.** *The TV distance between $\mathcal{N}(0, \sigma_p^2\mathbf{I}_d)$ and $\mathcal{N}(\mathbf{v}, \sigma_p^2\mathbf{I}_d)$ is smaller than $\|\mathbf{v}\|_2/2\sigma_p$.*

Finally, we can prove the third part of Theorem 4.2: given Lemma E.2 and Lemma 5.8.

Theorem E.3 (Third part of Theorem 4.2)**.** *Suppose that $n\|\boldsymbol{\mu}\|_2^4 \leq C_3 d\sigma_p^4$, then we have that $L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) \geq p + 0.1$, where C_3 is an sufficiently large absolute constant.*

Proof. For the sake of convenience, we use $(\mathbf{x}, \widehat{y}, y) \sim \mathcal{D}$ to denote the following: data point $(\mathbf{x}, y)$ follows distribution $\mathcal{D}$

defined in Definition 1.1, and $\widehat{y}$ is its true label. By (E.1), we have

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(y \neq \text{sign}(f(\mathbf{W}^{(t)},\mathbf{x}))\big)\\
&= p\cdot\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x})\geq 0\big) + (1-p)\cdot\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x})\leq 0\big)\\
&= p + (1-2p)\cdot\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x})\leq 0\big).
\end{aligned} \tag{E.16}$$

Therefore, it suffices to provide a lower bound for $\mathbb{P}_{(\mathbf{x},\widehat{y})\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x})\leq 0\big)$. To achieve this, we have

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\boldsymbol{W}^{(t)},\mathbf{x})\leq 0\big)\\
&= \mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\bigg(\sum_r \sigma(\langle\mathbf{w}_{-\widehat{y},r}^{(t)},\boldsymbol{\xi}\rangle) - \sum_r \sigma(\langle\mathbf{w}_{\widehat{y},r}^{(t)},\boldsymbol{\xi}\rangle) \geq \sum_r \sigma(\langle\mathbf{w}_{\widehat{y},r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle) - \sum_r \sigma(\langle\mathbf{w}_{-\widehat{y},r}^{(t)},\widehat{y}\boldsymbol{\mu}\rangle)\bigg)\\
&\geq 0.5\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\bigg(\bigg|\sum_r \sigma(\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}\rangle) - \sum_r \sigma(\langle\mathbf{w}_{-1,r}^{(t)},\boldsymbol{\xi}\rangle)\bigg| \geq C_6\max\bigg\{\sum_r \gamma_{1,r}^{(t)}, \sum_r \gamma_{-1,r}^{(t)}\bigg\}\bigg)
\end{aligned} \tag{E.17}$$

where C_6 is a constant, the inequality holds since if $\left|\sum_r \sigma(\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}\rangle) - \sum_r \sigma(\langle\mathbf{w}_{-1,r}^{(t)},\boldsymbol{\xi}\rangle)\right|$ is too large we can always pick a corresponding $\widehat{y}$ given $\boldsymbol{\xi}$ to make a wrong prediction. Let $g(\boldsymbol{\xi}) = \sum_r \sigma(\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}\rangle) - \sum_r \sigma(\langle\mathbf{w}_{-1,r}^{(t)},\boldsymbol{\xi}\rangle)$. Denote the set

$$\Omega := \bigg\{\boldsymbol{\xi}\bigg||g(\boldsymbol{\xi})| \geq C_6\max\bigg\{\sum_r \gamma_{1,r}^{(t)}, \sum_r \gamma_{-1,r}^{(t)}\bigg\}\bigg\}.$$

By plugging the definition of Ω into (E.17), we have

$$\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\boldsymbol{W}^{(t)},\mathbf{x})\leq 0\big) \geq 0.5\mathbb{P}(\Omega) \tag{E.18}$$

Next, we will give a lower bound of $\mathbb{P}(\Omega)$. By Lemma 5.8, we have that $\sum_j[g(j\boldsymbol{\xi}+\mathbf{v}) - g(j\boldsymbol{\xi})] \geq 4C_6\max_j\left\{\sum_r \gamma_{j,r}^{(t)}\right\}$

Therefore, by pigeon's hole principle, there must exist one of the $\boldsymbol{\xi}$, $\boldsymbol{\xi}+\mathbf{v}$, $-\boldsymbol{\xi}$, $-\boldsymbol{\xi}+\mathbf{v}$ belongs Ω. So we have proved that $\Omega\cup-\Omega\cup\Omega-\{\mathbf{v}\}\cup-\Omega-\{\mathbf{v}\} = \mathbb{R}^d$. Therefore at least one of $\mathbb{P}(\Omega), \mathbb{P}(-\Omega), \mathbb{P}(\Omega-\{\mathbf{v}\}), \mathbb{P}(\Omega-\{\mathbf{v}\}), \mathbb{P}(-\Omega-\{\mathbf{v}\})$ is greater than 0.25. Notice that $\mathbb{P}(-\Omega) = \mathbb{P}(\Omega)$ and

$$\begin{aligned}
|\mathbb{P}(\Omega) - \mathbb{P}(\Omega-\mathbf{v})| &= |\mathbb{P}_{\boldsymbol{\xi}\sim\mathcal{N}(0,\sigma_p^2\mathbf{I}_d)}(\boldsymbol{\xi}\in\Omega) - \mathbb{P}_{\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{v},\sigma_p^2\mathbf{I}_d)}(\boldsymbol{\xi}\in\Omega)|\\
&\leq \text{TV}(\mathcal{N}(0,\sigma_p^2\mathbf{I}_d),\mathcal{N}(\mathbf{v},\sigma_p^2\mathbf{I}_d))\\
&\leq \frac{\|\mathbf{v}\|_2}{2\sigma_p}\\
&\leq 0.03,
\end{aligned}$$

where the first inequality is by the definition of Total variation (TV) distance, the second inequality is by Lemma E.2.

Therefore we have proved that $\mathbb{P}(\Omega)\geq 0.22$, and plugging this into (E.16) and (E.18), we get

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(y \neq \text{sign}(f(\mathbf{W}^{(t)},\mathbf{x}))\big)\\
&= p + (1-2p)\cdot\mathbb{P}_{(\mathbf{x},\widehat{y},y)\sim\mathcal{D}}\big(\widehat{y}f(\mathbf{W}^{(t)},\mathbf{x})\leq 0\big)\\
&\geq p + (0.5-p)\cdot\mathbb{P}(\Omega)\\
&\geq 0.78p + 0.11\\
&\geq p + 0.1,
\end{aligned}$$

where the last inequality is by $p < 1/C$ from Condition 4.1 and by choosing $C > 22$ a sufficiently large constant, which completes the proof. □