# Benign Oscillation of Stochastic Gradient Descent with Large Learning Rate

**Miao Lu**[1*] **Beining Wu**[2*] **Xiaodong Yang**[3] **Difan Zou**[4]
[1] Department of Management Science and Engineering, Stanford University
[2] Department of Statistics, University of Chicago
[3] Department of Statistics, Harvard University
[4] Department of Computer Science & Institute of Data Science, The University of Hong Kong
miaolu@stanford.edu, beiningw@uchicago.edu, xyang@g.harvard.edu, dzou@cs.hku.hk

## Abstract

In this work, we theoretically investigate the generalization properties of neural networks (NN) trained by stochastic gradient descent (SGD) with *large learning rates*. Under such a training regime, our finding is that, the *oscillation* of the NN weights caused by SGD with large learning rates turns out to be beneficial to the generalization of the NN, potentially improving over the same NN trained by SGD with small learning rates that converges more smoothly. In view of this finding, we call such a phenomenon "*benign oscillation*".

## 1 Introduction

While deep neural networks (NNs) have achieved tremendous empirical success in various domains including images, language processing, decision-making, etc, the theoretical understanding of deep learning is still far behind satisfactory, especially the relationships between optimization of the NN and its generalization. From the viewpoint of optimization, using *large learning rates* in NN training has been empirically shown to be of vital importance for generalization (He et al., 2016; Xing et al., 2018; Damian et al., 2022; Kaur et al., 2023). Nevertheless, a principled theoretical understanding of the mechanism behind the benefits of large learning rate training still remains limited.

To better capture the key ingredients in the training dynamics of stochastic gradient descent (SGD) with large learning rates, we train a ResNet (He et al., 2016) using SGD with small and large learning rates and present the training and testing results in Figure 1. When using a large learning rate SGD, we can observe an "oscillating" training curve, i.e., the training loss fluctuates at different iterations (generally this happens only when the learning rate exceeds the inverse of the objective smoothness), while for small learning rate SGD, the training curve is smooth and converges rapidly. On the other hand, the smooth convergence in training loss cannot bring any benefit for the test accuracy – SGD with large learning rates achieves a significantly higher test accuracy than SGD with small learning rates. These empirical observations suggest that the *oscillation* during training can be closely tied to the better generalization performance achieved by SGD with large learning rates.

In this paper, we study the learning dynamics of SGD with large learning rates by investigating the *oscillation* happening during the optimization process, and explain its benefit to the generalization performance. The key message is that compared to the smooth convergence achieved by SGD with small learning rates, *the oscillation prevents the over-greedy convergence and serves as the engine that drives the learning of less-prominent data patterns*. These data patterns would be beneficial for the NN to generalize well on unseen testing data. Thus we explain from the theoretical side why large learning rate training can help NN to generalize better in practice.

[*] Equal Contributions.

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

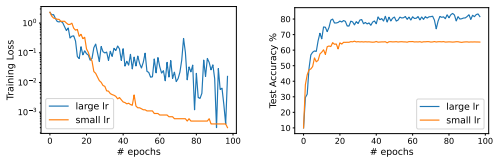


Figure 1: Training and test performance of ResNet-18 on CIFAR-10 dataset, when trained via SGD with small and large learning rates ($\eta = 0.01$ vs. $\eta = 0.75$). We adopt the same configuration as in Andriushchenko et al. (2023): using weight decay but no momentum and no data augmentation. A clear difference between the large learning rate training and small learning rate training can be observed: SGD with a large learning rate leads to an "oscillating" training curve with higher testing accuracy; SGD with a small learning rate has a rapid and smooth convergence but gives lower testing accuracy.

Our investigation of SGD with large learning rates for NN training builds on the feature learning perspective of deep learning theory (Allen-Zhu and Li, 2022), which explicitly considers data models consisting of different types of features and noise. For the sake of our goal, we devise a feature-noise data model consisting of two types of features that have different strengths and different distributions. Based upon the new data model, by carefully tracking the process of feature learning of a NN trained by SGD with large or small learning rates, we prove that only when trained with large learning rates can the NN effectively learn the key features for generalizing to *each* new data point. The NN trained by small learning rate SGD fails to generalize to certain testing data because of the limited learning of the features which are crucial to the generalization to those new data points. Our theory identifies the core incentives for the superior performance of learning the key features with large learning rate SGD as the *oscillation* during NN training. Intuitively, the oscillation can prevent over-greedy convergence which could only leverage the most prominent components of the data, thus allowing for all the useful components to be discovered and learned by gradient descent. In view of our findings, we refer to such a phenomenon as "*benign oscillation*".

## 2 Problem Setting

**Data generation model.** We let $\mathbf{v} \perp \mathbf{u} \in \mathbb{R}^d$ be two fixed vectors, denoting the signal (or feature) part shared by each data point. Then each data point, denoted by $(\mathbf{x}, y)$ where $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)})$ contains 3 patches, is generated as following: let $y \in \{1, -1\}$ be independently generated according to $\mathbb{P}(y = 1) = \mathbb{P}(y = -1) = 1/2$, and **(i)** *weak signal patch:* one patch of $\mathbf{x}$ is taken by the weak signal $y \cdot \mathbf{v}$; **(ii)** *strong signal patch:* with probability $1 - \rho$, one patch of $\mathbf{x}$ that is different from $y \cdot \mathbf{v}$, is taken by the strong signal $y \cdot \mathbf{u}$. **(iii)** *noise patch:* all remaining patches are taken by independent Gaussian noise $\boldsymbol{\xi} \sim N(0, \sigma_p^2(\mathbf{I}_d - \mathbf{v}\mathbf{v}^\top/\|\mathbf{v}\|_2^2 - \mathbf{u}\mathbf{u}^\top/\|\mathbf{u}\|_2^2))$ for some variance $\sigma_p > 0$.

**Two-layer CNN.** We consider a two-layer convolutional neural network (CNN) with filters applied to the three patches separately. We assign the parameters of the second layer of the CNN to a fixed $+1$ and $-1$, respectively. Formally, the CNN $f(\cdot; \mathbf{W}) : \mathbb{R}^{3d} \mapsto \mathbb{R}$ is defined as

$$f(\mathbf{x}; \mathbf{W}) = \sum_{j \in \{\pm 1\}} j F_j(\mathbf{x}; \mathbf{W}_j), \quad F_j(\mathbf{x}; \mathbf{W}_j) = \frac{1}{m} \sum_{r \in [m]} \sum_{p=1}^{3} \sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}^{(p)} \rangle), \tag{1}$$

where $m \in \mathbb{N}_+$ is the number of filters (i.e., neurons), $\sigma(z) = (\max\{z, 0\})^2$ is the $\mathrm{ReLU}^2$ activation function, and $\mathbf{w}_{j,r} \in \mathbb{R}^d$ denotes the weights of the $r$-th neuron of $F_j$. We use $\mathbf{W} = \{\mathbf{W}_j\}_{j \in \{\pm 1\}}$ and $\mathbf{W}_j = \{\mathbf{w}_{j,r}\}_{r \in [m]}$ to denote the collection of the weights.

**Loss function and SGD.** Having access to $n$ i.i.d. samples from the data generation model, $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i \in [n]}$, we solve a *binary classification* task by minimizing the following *mean squared loss*,

$$L(\mathbf{W}) = \frac{1}{n} \sum_{i \in [n]} \ell(f(\mathbf{x}_i; \mathbf{W}), y_i) = \frac{1}{2n} \sum_{i \in [n]} \big(f(\mathbf{x}_i; \mathbf{W}) - y_i\big)^2, \tag{2}$$

where $\ell(f(\mathbf{x}_i; \mathbf{W}), y_i) = (f(\mathbf{x}_i; \mathbf{W}) - y_i)^2/2$ is the loss on a single data point. Inspired by "edge of stability" (Cohen et al., 2020), adopting mean squared error is believed to make it easier to identify the effects of large learning rates. Besides, mean squared loss has also been demonstrated to be comparable or even better than cross-entropy loss in many classification tasks (Hui, 2020).

We optimize the loss function (2) via *multi-pass stochastic gradient descent* (SGD), initializing from some Gaussian weights, where each entry of $\mathbf{W}_{+1}^{(0)}$ and $\mathbf{W}_{-1}^{(0)}$ is sampled from $N(0, \sigma_0^2)$. The SGD goes for several epochs. In each epoch, we use each data $(\mathbf{x}_i, y_i)$ for exactly once, in the exact order of $(\mathbf{x}_1, y_1) \to (\mathbf{x}_2, y_2) \to \cdots \to (\mathbf{x}_n, y_n)$.Thus, the weights of the CNN are updated by

$$\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} - \frac{j\eta}{m} \cdot \big(f(\mathbf{W}^{(t)}, \mathbf{x}_{i_t}) - y_{i_t}\big) \cdot \sum_{p=1}^{3} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{x}_{i_t}^{(p)} \rangle) \cdot \mathbf{x}_{i_t}^{(p)}, \tag{3}$$

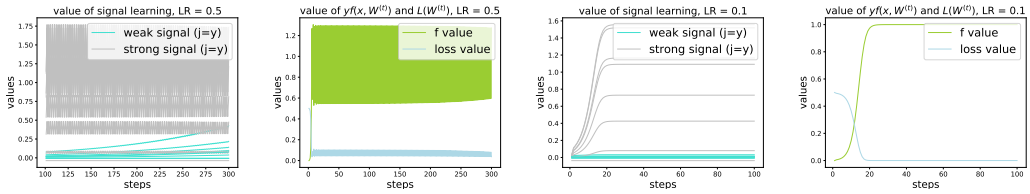


Figure 2: The progress of signal learning and the values of $yf(\mathbf{x}; \mathbf{W}^{(t)})$ and $L(\mathbf{W}^{(t)})$ under different learning rate $\eta$. The CNN in the first two figures is trained by SGD with $\eta = 0.5$ (large LR), while the CNN in the last two figures is trained by SGD with $\eta = 0.1$ (small LR). For signal learning (first and third figures), the gray lines depict the strong signal learning $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u} \rangle$ by all neurons $r \in [m]$, and the light blue lines depict the weak signal learning $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v} \rangle$ by all neurons $r \in [m]$. As we can see, with large LR, the value of CNN oscillates around $y$, and $\sum_t (1 - yf(\mathbf{x}; \mathbf{W}^{(t)}))$ is going to increase, which, as our theory indicates, incentivizes $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v} \rangle$ to increase. In contrast, with small LR, $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v} \rangle$ would stay at the same scale as its initialization.

for each $j \in \{\pm 1\}$ and $r \in [m]$, where $i_t = (t+1) \mod n$ and $\eta > 0$ is the learning rate.

**Generalization via signal (feature) learning.** Our goal is to study the generalization property of the CNN trained by SGD (3). Given a new testing data point $(\mathbf{x}^\diamond, y^\diamond)$ sampled from the data generation model, we measure the generalization of the CNN by the correctness of the classification,

$$\mathbb{E}[\mathbf{1}\{y^\diamond \cdot f(\mathbf{x}^\diamond; \mathbf{W}_{\mathtt{sgd}}) > 0\}] = \mathbb{P}(y^\diamond \cdot f(\mathbf{x}^\diamond; \mathbf{W}_{\mathtt{sgd}}) > 0),$$

where $\mathbf{W}_{\mathtt{sgd}}$ denotes the weights trained by SGD (Zhang et al., 2021).

We investigate the generalization property via looking through the process of signal (feature) learning. Specifically, according to the SGD updates (3), the weights of the CNN is a linear combination of the initialization, the strong signal, the weak signal, and the noise vectors. The relative scales of the combination coefficients actually imply how the weights learn the strong signal $\mathbf{u}$, the weak signal $\mathbf{v}$, or memorizing the noise which determines how the CNN can generalize. Thus, our main focus in the sequel would be studying the dynamics of the inner products $\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u} \rangle$, $\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v} \rangle$, and $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle$. We will show that under large learning rate SGD training $\langle \mathbf{w}_{j,r}, j\mathbf{v} \rangle$ can be effectively learned to a relatively large scale compared to its initialization.

**We refer to Appendix A for more details and remarks on the problem setup and contributions.**

## 3 Understand the Oscillation: Single Training Data Case

Before giving our main theory on large learning rate SGD training, let's first study a simplified setup where we consider only a *single* training data point consisting only of a weak signal patch $y \cdot \mathbf{v}$ and a strong signal patch $y \cdot \mathbf{u}$, without any noise patch. Such a setting helps to illustrate the key insights behind our main theory regarding the understanding of oscillation. Without loss of generality, we denote the single training data as $(\mathbf{x}, y)$ with $\mathbf{x} = (y \cdot \mathbf{v}, y \cdot \mathbf{u})$, and we can also simplify the CNN expression (1) and the SGD updates (3) to

$$f(\mathbf{x}; \mathbf{W}) = \sum_{j \in \{\pm 1\}} jF_j(\mathbf{x}; \mathbf{W}_j), \quad F_j(\mathbf{x}; \mathbf{W}_j) = \frac{1}{m} \sum_{r \in [m]} \sigma(\langle \mathbf{w}_{j,r}, y\mathbf{u} \rangle) + \sigma(\langle \mathbf{w}_{j,r}, y\mathbf{v} \rangle), \quad (4)$$

$$\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} - \frac{\eta j}{m} \cdot \big(f(\mathbf{x}; \mathbf{W}^{(t)}) - y\big) \cdot \Big(\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{u} \rangle) \cdot y\mathbf{u} + \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{v} \rangle) \cdot y\mathbf{v}\Big). \quad (5)$$

**Review: small learning rate training regime.** Firstly, we make a review of what may happen when using SGD with a small learning rate $\eta$. The following proposition proves that in this case the CNN can **not** make much progress in learning the weak signal $y \cdot \mathbf{v}$.

**Proposition 1** (Small learning rate training: single training data (informal))**.** *Under mild conditions on* $(d, m, \sigma_0, \|\mathbf{u}\|_2, \|\mathbf{v}\|_2)$*, if we choose learning rate* $\eta \leq m/(6\|\mathbf{u}\|_2^2)$ *small enough, then with high probability, the training loss can smoothly converge with* $\max_{j,r} |\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v} \rangle| \leq \sigma_0 \|\mathbf{v}\|_2$.

Please refer to Appendix G for more details about the proposition. It shows that in the small learning rate training regime, the CNN only learns the weak signal $y \cdot \mathbf{v}$ to the same scale as its initialization. CNN trained in this manner may fail to generalize to testing data without strong features (substituted by a noise patch $\boldsymbol{\xi}$), because it would make predictions relying mainly on the random noise. On the contrary, in the following we intuitively explain that under certain large learning rate training regime, the CNN can learn the weak signal up $y \cdot \mathbf{v}$ to a constant level higher than its initialization. Such a phenomenon is depicted in Figure 2 on an 8-neuron CNN trained by SGD with $\eta = 0.1$.

**Theoretical motivations: large learning rate regime and oscillation.** When using a large enough learning rate $\eta$ that exceeds the twice inversed smoothness, the weights of the CNN would keep

oscillating, which makes the value of $f(\mathbf{x}; \mathbf{W}^{(t)})$ fluctuate around $y$. The key finding towards our theory is that the fluctuations of $f(\mathbf{x}; \mathbf{W}^{(t)})$ around $y$ would *not* cancel with each other. Instead, the oscillation accumulates over time, which serves as the engine driving the learning of the weak signal $y \cdot \mathbf{v}$. In the sequel, we explain why the cancellation does not happen.

The core idea is that, with a reasonably large learning rate, the CNN weights will be quickly enlarged from the learning of strong feature $\mathbf{u}$ and then keep oscillating, but still stay well bounded. As a result of the SGD updates (5), the summation of the gradient terms is also well bounded. More specifically, let's look carefully into the dynamics of learning the strong signal $y \cdot \mathbf{u}$. For some time steps $t_0, t_1$ and certain neuron $r \in [m]$, it holds from (5) that

$$\mathcal{O}(1) = \Big|\langle \mathbf{w}_{y,r}^{(t_1+1)}, y\mathbf{u}\rangle - \langle \mathbf{w}_{y,r}^{(t_0)}, y\mathbf{u}\rangle\Big| \approx \Theta\bigg(\sum_{s=t_0}^{t_1} \big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big) \cdot \langle \mathbf{w}_{y,r}^{(s)}, y\mathbf{u}\rangle\bigg). \tag{6}$$

Now we split the summation on the right hand side of (6) into two parts: one part is $\mathcal{S}^+$ containing $s$ such that $yf(\mathbf{x}; \mathbf{W}^{(s)}) > 1$ and the other part is $\mathcal{S}^-$ containing $s$ such that $yf(\mathbf{x}; \mathbf{W}^{(s)}) < 1$. It turns out that if the weak signal component of the CNN is relatively small compared with the strong signal component, the whole behavior will be dominated by the dynamics of the strong signal component. In other words, when $yf(\mathbf{x}; \mathbf{W}) > 1$, the inner products $\langle \mathbf{w}_{y,r}, y\mathbf{u}\rangle$ would also take a relatively large value. Conversely, when $yf(\mathbf{x}; \mathbf{W}) < 1$, the inner products $\langle \mathbf{w}_{y,r}, y\mathbf{u}\rangle$ would also take a relatively small value. Consequently, in view of (6), we can see that the total increase of $\langle \mathbf{w}_{y,r}^{(s)}, y\mathbf{u}\rangle$ and decrease of $\langle \mathbf{w}_{y,r}^{(s)}, y\mathbf{u}\rangle$ during the oscillation period are approximately balanced, i.e.,

$$\sum_{s\in\mathcal{S}^+} \underbrace{\big(yf(\mathbf{x}; \mathbf{W}^{(s)}) - 1\big)}_{yf(\mathbf{x};\mathbf{W}^{(s)})>1} \cdot \underbrace{\langle \mathbf{w}_{y,r}^{(s)}, y\mathbf{u}\rangle}_{\text{relatively large}} \approx \sum_{s\in\mathcal{S}^-} \underbrace{\big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big)}_{yf(\mathbf{x};\mathbf{W}^{(s)})<1} \cdot \underbrace{\langle \mathbf{w}_{y,r}^{(s)}, y\mathbf{u}\rangle}_{\text{relatively small}}.$$

Consequently, the summation of $1 - yf(\mathbf{x}; \mathbf{W}^{(s)})$ over $s \in \mathcal{S}^-$ would take a larger value than the summation of $yf(\mathbf{x}; \mathbf{W}^{(s)}) - 1$ over $s \in \mathcal{S}^+$. This means that the whole summation

$$\sum_{s\in\mathcal{S}^+\cup\mathcal{S}^-} \big(yf(\mathbf{x}; \mathbf{W}^{(s)}) - 1\big) = \sum_{s\in\mathcal{S}^-} \big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big) - \sum_{s\in\mathcal{S}^+} \big(yf(\mathbf{x}; \mathbf{W}^{(s)} - 1)\big) \gtrsim 0.$$

That is, the oscillation of $f(\mathbf{x}; \mathbf{W}^{(s)})$ around the label $y$ over time does *not* tend to cancel with each other. Instead, the summation of the fluctuations would have a determined sign. Furthermore, if the CNN values are bounded away from the label by a uniform constant $\delta$ (i.e., the magnitude of the oscillation), we can further prove that

$$\sum_{s\in\mathcal{S}^+\cup\mathcal{S}^-} \big(yf(\mathbf{x}; \mathbf{W}^{(s)}) - 1\big) \gtrsim \Omega\big(\delta \cdot \max\{|\mathcal{S}^+|, |\mathcal{S}^-|\}\big) \gtrsim \Omega\big(\delta \cdot (t_1 - t_0)\big). \tag{7}$$

This is the key motivation behind our theory for studying the oscillating SGD. With (7) in hand, we can further show that once the weak signal component of the CNN is still small, which means that the weak signal hasn't been well learned, the linear accumulation of the oscillation would incentivize the learning of the weak signal by a careful analysis of the updates (5).

**Outcome of oscillation.** Based on previous discussions, we can arrive at our result for the simplified setup of this explanatory section: the oscillating SGD can indeed make progress in learning the weak signal $y \cdot \mathbf{v}$, which helps the CNN to generalize to new data points which possibly lacks strong signal $y \cdot \mathbf{u}$. This is summarized in the following (informal) theorem and corollary.

**Theorem 2** (Large LR training, single data case (informal))**.** *Under mild conditions on dimension $d$, width $m$, initialization $\sigma_0$, and learning rate $\eta > m/(4\|\mathbf{u}\|_2^2)$, if the SGD training* (5) *oscillates in the sense that $|yf(\mathbf{x}; \mathbf{W}^{(t)}) - 1| \geq \delta$ for some constant $\delta > 0$ and each $t \geq 0$, then with high probability there exists a $t^\star \leq T_{\max}$ with $T_{\max} \in \mathrm{poly}(d, m, \eta^{-1}, \delta^{-1}, \|\mathbf{u}\|_2^{-1}, \|\mathbf{v}\|_2^{-1})$ such that*

$$\frac{1}{m}\sum_{r\in[m]} \sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle) \gtrsim \delta, \quad \forall t \geq t^\star.$$

Please refer to Appendix F for formal and detailed statement of Theorem 2 and its proofs. Theorem 2 shows that via oscillating SGD training, the CNN learns the weak signal $y \cdot \mathbf{v}$ up to a constant scale of $\delta$, which is typically much larger than the scale of its initialization, since

$$\frac{1}{m}\sum_{r\in[m]} \sigma(\langle \mathbf{w}_{y,r}^{(0)}, y\mathbf{v}\rangle) \lesssim \widetilde{\mathcal{O}}(\sigma_0^2 \cdot \|\mathbf{v}\|_2^2) \ll \delta \lesssim \frac{1}{m}\sum_{r\in[m]} \sigma(\langle \mathbf{w}_{y,r}^{(t^\star)}, y\mathbf{v}\rangle),$$

whenever the initializations of the CNN is small. We remark that here we mainly consider neurons with $j = y$ and testing data with label $y$ since the CNN is trained only on a single data with label $y$.

**We refer to Appendix B for our main theory on the multiple data setup introduced in Section 2.**

## References


ALLEN-ZHU, Z. and LI, Y. (2022). Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*.

ANDRIUSHCHENKO, M., VARRE, A. V., PILLAUD-VIVIEN, L. and FLAMMARION, N. (2023). Sgd with large step sizes learns sparse features. In *International Conference on Machine Learning*. PMLR.

ARORA, S., LI, Z. and PANIGRAHI, A. (2022). Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*. PMLR.

CAO, Y., CHEN, Z., BELKIN, M. and GU, Q. (2022). Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems* **35** 25237–25250.

CHEN, L. and BRUNA, J. (2022). On gradient descent convergence beyond the edge of stability. *arXiv preprint arXiv:2206.04172* .

CHEN, Z., DENG, Y., WU, Y., GU, Q. and LI, Y. (2022). Towards understanding the mixture-of-experts layer in deep learning. *Advances in neural information processing systems* **35** 23049–23062.

COHEN, J., KAUR, S., LI, Y., KOLTER, J. Z. and TALWALKAR, A. (2020). Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*.

COHEN, J. M., GHORBANI, B., KRISHNAN, S., AGARWAL, N., MEDAPATI, S., BADURA, M., SUO, D., CARDOZE, D., NADO, Z., DAHL, G. E. ET AL. (2022). Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484* .

DAMIAN, A., NICHANI, E. and LEE, J. D. (2022). Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations*.

HE, K., ZHANG, X., REN, S. and SUN, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

HUANG, W., CAO, Y., WANG, H., CAO, X. and SUZUKI, T. (2023). Graph neural networks provably benefit from structural information: A feature learning perspective. *arXiv preprint arXiv:2306.13926* .

HUI, L. (2020). Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *The Ninth International Conference on Learning Representations (ICLR 2021)*.

JASTRZEBSKI, S., ARPIT, D., ASTRAND, O., KERG, G. B., WANG, H., XIONG, C., SOCHER, R., CHO, K. and GERAS, K. J. (2021). Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *International Conference on Machine Learning*. PMLR.

KAUR, S., COHEN, J. and LIPTON, Z. C. (2023). On the maximum hessian eigenvalue and generalization. In *Proceedings on*. PMLR.

LI, Y., WEI, C. and MA, T. (2019). Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems* **32**.

WANG, Z., LI, Z. and LI, J. (2022). Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability. *Advances in Neural Information Processing Systems* **35** 9983–9994.

WEN, Z. and LI, Y. (2021). Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*. PMLR.

WU, J., BRAVERMAN, V. and LEE, J. D. (2023). Implicit bias of gradient descent for logistic regression at the edge of stability. *arXiv preprint arXiv:2305.11788* .

WU, J., ZOU, D., BRAVERMAN, V. and GU, Q. (2021). Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate. In *International Conference on Learning Representation (ICLR)*.

XING, C., ARPIT, D., TSIRIGOTIS, C. and BENGIO, Y. (2018). A walk with sgd. *arXiv preprint arXiv:1802.08770* .

YANG, N., TANG, C. and TU, Y. (2023). Stochastic gradient descent introduces an effective landscape-dependent regularization favoring flat solutions. *Physical Review Letters* **130** 237101.

ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM* **64** 107–115.

ZHU, X., WANG, Z., WANG, X., ZHOU, M. and GE, R. (2022). Understanding edge-of-stability training dynamics with a minimalist example. In *The Eleventh International Conference on Learning Representations*.

ZOU, D., CAO, Y., LI, Y. and GU, Q. (2022). Understanding the generalization of adam in learning neural networks with proper regularization. In *The Eleventh International Conference on Learning Representations*.

ZOU, D., CAO, Y., LI, Y. and GU, Q. (2023). The benefits of mixup for feature learning. In *Proceedings of the 40th International Conference on Machine Learning*.

## Contents

# A More Remarks on Our Contributions and the Problem Setup

## A.1 Our Contributions

Our contributions to explaining large learning rate NN training are in the theoretical side, three folds.

**Dynamic analysis framework for SGD with large learning rates.** We provide a theoretical framework to understand and explain the oscillation in NN training via SGD with a large learning rate. Specifically, we consider a feature-noise data generation model consisting of two types of features – the *strong features* and the *weak features* – that have different strengths and distributions to capture our core ideas towards explaining the relationships between large learning rate SGD training and generalization. Then, our theoretical framework establishes a sharp characterization of the training dynamics of these features and noises, based on which we can precisely analyze the generalization of NN trained by SGD with small or large learning rates. We remark that in general studying the NN optimization dynamics when the learning rate is greater than twice inversed smoothness is quite challenging, and our theoretical analysis framework based upon the feature-noise model potentially provides useful guidance which can be leveraged to study other nonconvex optimization problems.

**A new theoretical argument for feature learning driven by oscillation.** The key to explaining the large learning rate training regime is a new theory on learning the weak features driven by oscillation. As we illustrate in Section 3, the oscillation of the NN values (predictions) around the target (label) does *not* cancel with each other. Instead, the fluctuations accumulate linearly over time. This further serves as the engine driving the learning of the weak features, resulting in better generalization. This characterizes the distinctive training dynamics of SGD under the large learning rate training regime, revealing the benefits of the oscillation in learning useful data patterns.

**Division for generalization by different learning rates.** In contrast to effectively learning the weak features by large learning rate oscillating training, we also show that the smooth and rapid convergence achieved by SGD with small learning rates would *not* help NN learn the weak features, thus being unable to generalize to the new data without strong features. This gives a division of the generalization property of NNs trained by large and small learning rates.

## A.2 More Remarks on the Problem Setup (Section 2)

**Data generation model.** For simplicity, we refer to the data with strong signal as the *strong data*, denoted by $((y\mathbf{u}, y\mathbf{v}, \boldsymbol{\xi}), y)$, and we refer to the data with only weak signal as the *weak data*, denoted by $((\widetilde{\boldsymbol{\xi}}, y\mathbf{v}, \boldsymbol{\xi}), y)$. Here by "strong", we mean a vector with a larger $\ell_2$-norm, as we specify in the main theory part. Intuitively, the weak signal $y \cdot \mathbf{v}$ can be interpreted as the invariant and common signals across data like the shape of key objects in an image. The strong signal $y \cdot \mathbf{u}$ can be understood as the background or the domain information which is stronger but only appears in a certain fraction of all data points. This indicates that in order for a classifier to generalize to all new data, it must effectively learn the weak signal.

Our proposed data generation model is adapted from the feature-learning-based line of research on deep learning (Allen-Zhu and Li, 2022; Cao et al., 2022; Zou et al., 2023), and it can serve as a good theoretical platform to explain the relationships between oscillating NN training with large learning rates and NN generalization. Finally, we remark that this data model can be extended for generality, e.g., multiple features, more patches, multi-class data. In fact, as long as the signal and noise patches have properly different strength and fractions, our theoretical analysis can be directly applied.

**Generalization via signal (feature) learning.** We investigate the generalization property via looking through the process of signal (feature) learning. Specifically, by the SGD updates (3), the weights $\mathbf{w}_{j,r}^{(t)}$ of the CNN is a linear combination of the initialization $\mathbf{w}_{j,r}^{(0)}$, the strong signal $j \cdot \mathbf{u}$, the weak signal $y \cdot \mathbf{v}$, and the noise vectors $\boldsymbol{\xi}_i$, $\widetilde{\boldsymbol{\xi}}_i$. This motivates us to consider the following representation

of the weights, for $j \in \{\pm 1\}$, $r \in [m]$,

$$\mathbf{w}_{j,r}^{(t)} \approx \mathbf{w}_{j,r}^{(0)} + \frac{\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u} \rangle}{\|\mathbf{u}\|_2^2} \cdot j\mathbf{u} + \frac{\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v} \rangle}{\|\mathbf{v}\|_2^2} \cdot j\mathbf{v} + \text{noise parts}. \tag{8}$$

The relative scales of these combination coefficients actually imply how the weights learn the strong signal $\mathbf{u}$, the weak signal $\mathbf{v}$, or memorizing the noise which determines how the CNN can generalize. To be more explicit, consider that by (8), when a new testing data point $(\mathbf{x}^\diamond, y^\diamond)$ comes, the CNN predicts $y^\diamond$ as a function of $\langle \mathbf{w}_{j,r}^{(t)}, y^\diamond \mathbf{u} \rangle$, $\langle \mathbf{w}_{j,r}^{(t)}, y^\diamond \mathbf{v} \rangle$, and $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}^\diamond \rangle$ (possibly without $\langle \mathbf{w}_{j,r}^{(t)}, y^\diamond \mathbf{u} \rangle$, recall (1)), which can be writen as

$$\langle \mathbf{w}_{j,r}^{(t)}, y^\diamond \mathbf{u} \rangle = \langle \mathbf{w}_{j,r}^{(0)}, y^\diamond \mathbf{u} \rangle + y^\diamond j \cdot \rho_{j,r}(\mathbf{u}), \quad \langle \mathbf{w}_{j,r}^{(t)}, y^\diamond \mathbf{v} \rangle = \langle \mathbf{w}_{j,r}^{(0)}, y^\diamond \mathbf{v} \rangle + y^\diamond j \cdot \rho_{j,r}(\mathbf{v}),$$
$$\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}^\diamond \rangle = \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}^\diamond \rangle + \sum_{i=1}^{n} \frac{\rho_{j,r}(\boldsymbol{\xi}_i)}{\|\boldsymbol{\xi}_i\|_2^2} \cdot \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}^\diamond \rangle.$$

Thus the key for the CNN to generalize well to a new data point is to make enough process in signal learning: $\rho_{j,r}(\mathbf{u})$ and $\rho_{j,r}(\mathbf{v})$. More importantly, since under our data model the strong signal $y \cdot \mathbf{u}$ disappears with a constant probability $\rho$, it is vital for the CNN to learn the weak signal well as it is the only patch carrying the information of label when lacking the strong signal patch. Naively fitting the training data through the strong signal component would result in poor generalization when the testing data point loses the strong signal patch.

As is shown by Cao et al. (2022), the CNN tends to fit the training dataset using patches with higher strength when trained by small learning rate gradient descents. Therefore, in such a training regime, the CNN tends to fit the training data using the strong signal $y \cdot \mathbf{u}$, making less progress in learning the weak signal $y \cdot \mathbf{v}$, thus resulting in misclassification when the testing data lacks the strong signal component. On the contrary, our paper investigates the large learning rate regime, and suggests that the *oscillation* of SGD is beneficial for learning the weak signal, giving better generalization results.

In the following and before diving into our main theory, we explain more on the dynamics of the representation coefficients. Note that by the decomposition (8),

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

By a standard concentration argument, we can see that, when the scale $\sigma_0$ of the weight initialization is small and the dimension $d$ is sufficiently large, approximately the coefficients are

$$\rho_{j,r}(\mathbf{u}) \approx \langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u} \rangle, \quad \rho_{j,r}(\mathbf{v}) \approx \langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v} \rangle, \quad \rho_{j,r}(\boldsymbol{\xi}_i) \approx \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle.$$

Thus, our main focus in the sequel would be studying the dynamics of the inner products $\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u} \rangle$, $\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v} \rangle$, and $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle$. We will show that under large learning rate SGD training $\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v} \rangle$ can be effectively learned to a relatively large scale compared to its initialization. This is further provably useful for generalizing to all new data points.

**Implications to the simple signal-noise model.** Our findings could also be adapted to explain the setting with data model considered by Cao et al. (2022). In that case, one data point $\mathbf{x} = (y \cdot \boldsymbol{\mu}, \boldsymbol{\xi})$ consists of a single signal patch $y \cdot \boldsymbol{\mu}$ and a single Gaussian noise patch $\boldsymbol{\xi}$. Therefore a trained neural network can generalize to new data points only when it learns the common signal vector $\boldsymbol{\mu}$.

As is shown by Cao et al. (2022), under small learning rate training regime, if the data model has a low *signal-to-noise ratio* (SNR), that is, the strength of the noise patch is relatively stronger than the the strength of the signal patch, then overfitting the training data would result in poor generalization (harmful overfitting). That is because the neural network would memorize the noise patch quickly so as to fit the data, and consequently the signal patch is not well learned. In contrast, we can show that under the oscillating SGD training regime, the signal can also be well learned even with a low SNR. The mechanism behind this is still that the oscillation during training would accumulate and incentivize the neural network towards signal learning.

## B Main Theory

In this section, we present our main theory on benign oscillation of SGD training with large learning rates for the setup introduced in Section 2. We will first introduce the key conditions and assumptions required by our theory in Section B.1. Then we present our theoretical results in Section B.2. Finally in Section B.2, we also compare large learning rate oscillating training to small learning rate training.

### B.1 Key Conditions and Assumptions

Before presenting our theoretical results, we first outline the key conditions and assumptions needed on the model and the training dynamics. Firstly, our results are based upon the following conditions on the initialization scale $\sigma_0$, dimension $d$, number of data $n$, and neural network width $m$.

**Assumption 3** (Conditions on hyperparameters)**.** *Suppose that the following conditions hold: (i) the CNN weight initialization scale $\sigma_0 = \widetilde{\Theta}(\max\{\|\mathbf{u}\|_2, \|\mathbf{v}\|_2, \sigma_p\sqrt{d}\}^{-1} \cdot d^{-1/2})$; (ii) the dimension $d = \Omega(n^2, \mathrm{polylog}(m))$; (iii) the signal strength: $\|\mathbf{v}\|_2 \leq 0.1\|\mathbf{u}\|_2$, $\|\mathbf{u}\|_2^{-2} + \|\mathbf{v}\|_2^{-2} \leq n(\sigma_p^2 d)^{-1}$. (iv) the learning rate $m/(4\|\mathbf{u}\|_2^2) \leq \eta \leq 2m/(5\|\mathbf{u}\|_2^2)$. (v) the weak data fraction $\rho \leq c$ for some small constant $c$.*

We explain the conditions in Assumption 3 one by one. The conditions on the initialization scale $\sigma_0$ and the learning rate $\eta$ are to ensure that the whole training process is well bounded while oscillates (rather than converging smoothly). The condition on the dimension $d$ puts us in the regime of high dimensions for which independent Gaussian noise has small correlations. Finally, the conditions on the signal strength separate the strong signal from the weak signal by $\ell_2$-norm. Also, we ensure that the data are not too noisy by restricting the variance of the Gaussian noise.

The next assumption is on the training process, which requires that the SGD oscillates. For simplicity, we denote the index of weak training data points lacking the strong feature patch as $\mathcal{W}$.

**Assumption 4** (Oscillating SGD)**.** *We assume that there exists a constant $\delta \in (0.2, 0.8)$, such that $|y_{i_t} f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}) - 1| \geq \delta$ holds for any $t \geq 0$ such that $i_t \notin \mathcal{W}$.*

Through Assumption 4, we require that the value of $yf(\mathbf{x}; \mathbf{W}^{(t)})$ on data points with strong features oscillates around the desired value, 1, by a scale of $\delta \in (0.2, 0.8)$, i.e., the magnitude of the oscillation is at least $\delta$. Here the range for $\delta$ is only for technical considerations to simplify the theoretical analyses.

It is notable that Assumption 4 implicitly requires that the learning rate $\eta$ should be scaled properly. A large $\eta$ forces the training trajectories to escape from the regular region, while a small $\eta$ shall result in smooth convergence. In both cases the phenomenon described in Assumption 4 does not happen. We also remind readers that the $\eta$ condition in Assumption 3 is only sufficient for the regularities such as boundedness and sign stability. Readers can refer to Appendix F.7 for a discussion of the necessary conditions of Assumption 4.

We remark that in general the dynamics of the training process could be quite subtle when oscillation happens, and there exist other more complicated patterns of oscillations if one deliberately chooses a specific learning rate $\eta$. Our work focuses on a relatively simple but common pattern of oscillation. It turns out that under the oscillation pattern in Assumption 4, we can show the benefits of oscillation on the generalization properties of the CNN. Actually, we can also extend our theoretical analysis to a weakened version of Assumption 4 that the time average of $|y_{i_t} f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}) - 1|$ is larger than $\delta$.

Finally, we remark that we only assume the oscillation on strong data, since intuitively on weak data the CNN fits the label via the weak features and the noise (both have smaller strength than strong features) and may converge slower and more smoothly. *See Section B.3 for experimental evidence.*

### B.2 Main Theoretical Results

Our main results are that, under previous conditions and assumptions on the hyperparameters and the training dynamics, the CNN can make enough progress in learning the weak signal $\mathbf{v}$ thanks to the oscillation happening during training. We refer to Appendix H for a detailed proof of the results.

**Theorem 5** (Weak signal learning: oscillating training with large learning rate)**.** *Under Assumptions 3 and 4, w.p. at least $1 - 1/\mathrm{poly}(d)$, there exists $t^\star \leq \mathrm{poly}(d, m, n, \delta^{-1}, \eta^{-1}\sigma_p^{-1}, \|\mathbf{u}\|_2^{-1}, \|\mathbf{v}\|_2^{-1})$*

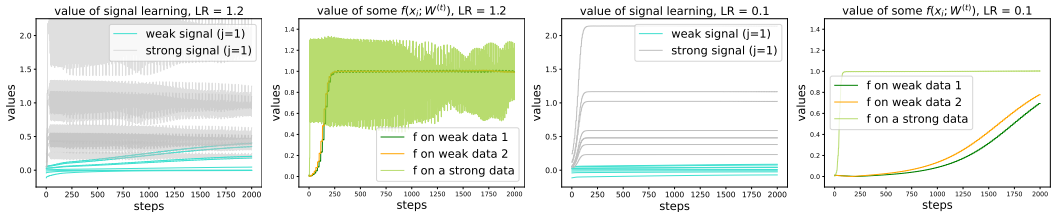


Figure 3: The dynamics of signal learning under a large learning rate $\eta_{\text{large}} = 1.2$ and a small learning rate $\eta_{\text{small}} = 0.1$. The values of signal learning are obtained by characterizing the inner products $\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u} \rangle$ and $\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v} \rangle$. It can be seen that when using the large LR, strong signal learning as well as the NN outputs will oscillate, during which weak signal will be gradually learned. When using the small LR, strong signal learning will converge quickly, and the weak signal learning will stay at the same scale as its initialization.

*such that*

$$\max_{j\in\{\pm 1\}} \frac{1}{m} \sum_{r\in[m]} \sigma(\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v} \rangle) - \frac{1}{m} \sum_{r\in[m]} \sigma(\langle \mathbf{w}_{-j,r}^{(t)}, j\mathbf{v} \rangle) \geq \frac{\delta}{4}.$$

In contrast, under the small learning rate regime, the CNN would not learn the weak features, which is the following proposition with proofs in Appendix I.

**Proposition 6** (Small learning rate training)**.** *Under Assumption 35 on* $(d, m, \sigma_0, \|\mathbf{u}\|_2, \|\mathbf{v}\|_2)$*, if we choose learning rate* $\eta \leq m/(6\|\mathbf{u}\|_2^2)$ *small enough, then with high probability, the training loss can smoothly converge with*

$$\max_{j,r} |\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v} \rangle| \leq \sigma_0 \|\mathbf{v}\|_2.$$

**Division of generalization.** Suppose we are given a new testing data point $(\mathbf{x}^\diamond, y^\diamond)$ with an input $\mathbf{x}^\diamond = (y^\diamond \mathbf{v}, \boldsymbol{\xi}^\diamond, \widetilde{\boldsymbol{\xi}}^\diamond)$ only consisting of the weak signal $\mathbf{v}$. Then a reliable prediction can only count on utilizing the weak signal $\mathbf{v}$. In the regime specified by Theorem 5, there holds $y^\diamond \cdot f(\mathbf{x}^\diamond; \mathbf{W}^{(t)}) \geq \delta/4 - o(1) > 0$ corresponding to correct prediction almost certainly. In contrast, when applying the small learning rate, as specified by Proposition 6, the trained NN fails to take advantage of the weak signal $\mathbf{v}$ from the data $\mathbf{x}^\diamond$. Therefore, it will be likely to make the prediction based on a random guess (the randomness stems from the random initialization and the noise patches $\boldsymbol{\xi}^\diamond, \widetilde{\boldsymbol{\xi}}^\diamond$ ). Consequently, note that the weak data takes up $\rho$ fraction of the dataset (see our data model in Section 2), SGD with large LR will achieve a $\Theta(\rho)$ higher test accuracy than SGD with small LR, demonstrating the benefit of oscillation and large learning training in terms of the generalization ability.

### B.3 Numerical Experiments

In this part, we conduct numerical experiments to demonstrate our findings on "*benign oscillation*". We follow the same data generation model and optimization algorithm as we described as Section 2. Specifically, we consider a dataset with $n = 16$ and $|\mathcal{W}| = 2$, that is, $\rho \approx 0.125$. The dimension is $d = 64$, and the number of neurons for each direction $j$ is $m = 8$. We generate the data with strong signal $\|\mathbf{u}\|_2 = 2$, weak signal $\|\mathbf{v}\|_2 = 0.4$, and noise $\|\boldsymbol{\xi}\|_2 \approx \sigma_p d^{1/2} = 0.8$.

**Weak signal learning.** We run the SGD to train the CNN with two different scale of learning rates: a large learning rate $\eta_{\text{large}} = 1.2$, a small learning rate $\eta_{\text{small}} = 0.1$. We plot the dynamics of signal learning for each neuron $r \in [m]$ from these two training regimes in Figure 3. The first two figures plot the large learning rate training, and the last two figures plot the small learning rate training.

As we can see from Figure 3, with large learning rate SGD training, the CNN can effectively learn the weak signal to a scale much larger than the initialization. On the contrary, by small learning rate SGD training, the CNN does not learn the weak signal since it just remains at the same level as the initialization. This demonstrates our main theory in Section B.

Furthermore, in Figure 3, we plot the values of $y_i f(\mathbf{x}_i; \mathbf{W}^{(t)})$ on certain data points $i \in [n]$ and the value of $L(\mathbf{W}^{(t)})$ for the two training regimes. In specific, we plot the values of $y_i f(\mathbf{x}_i; \mathbf{W}^{(t)})$ on a strong data $i \notin \mathcal{W}$ (randomly sampled) and the value of $y_i f(\mathbf{x}_i; \mathbf{W}^{(t)})$ on the weak data $i \in \mathcal{W}$. As we can see from the large learning rate training case, the value of $f$ on the strong data oscillates while the values of $f$ on the weak data do not. This matches our theoretical assumption in Assumption 4 that the oscillation in $f$ value only happens for strong data. Also, for the small learning rate training case, the values of $f$ on the weak data converge slower than those on the strong data. This is because

on the weak data the CNN mainly uses the noise to fit the target which is of lower strength than the strong signals on strong data. But still, the noise out weights the weak signals and consequently the CNN makes no progress in learning the weak signal if trained smoothly.

**Generalization properties.** Finally, we test the CNN trained by two different learning rates on new testing data generated in the same way as the training data, The testing data size $32$ with $4$ weak data points. We repeat the testing evaluation over $5$ random seeds and take the average. The result is that for the CNN trained by $\eta_{\rm large}$ the test accuracy is $99.38\%$, and for the CNN trained by $\eta_{\rm small}$ the test accuracy is $93.75\%$, matching our theoretical insights that large learning rate training benefits NN generalization. For the CNN trained by $\eta_{\rm small}$, it misclassifies certain weak data points. As we previously discussed, on data without strong signal, the CNN approximately uses a random guess.

# C Related Works

In this section, we discuss the related works.

**Large learning rate NN training.** Gradient descent training coped with large learning rates for deep learning is receiving an ever increasing attention for recent years (Cohen et al., 2020; Jastrzebski et al., 2021; Andriushchenko et al., 2023). For GD training, the phenomenon of "*edge of stability*" (Cohen et al., 2020, 2022) showed that the sharpness of the loss Hessian would finally hover just above $2/\eta$ and thus a larger learning rate would prefer a flatter minimum and possibly better generalization, and have received great attention in recent years (Arora et al., 2022; Chen and Bruna, 2022; Damian et al., 2022; Wang et al., 2022; Zhu et al., 2022). Besides, Li et al. (2019) studied the regularization effect of large learning rates of SGD at initialization which results in better generalization than using a small initial learning rate training. Wu et al. (2021) studied the implicit bias of SGD with a moderate large learning rate for overparametrized linear regression. Wu et al. (2023) then studied the implicit bias of large learning rate GD training in logistic regression. In addition, Andriushchenko et al. (2023) showed that SGD with a large learning rate can help NNs to learn sparse features from data, but did not provide rigorous theoretical justifications. We highlight that our theoretical work on large learning rate SGD builds upon a multi-pass fashion of SGD and a feature-noise data generation model (see Section 2), which is different from previous works (Li et al., 2019; Wu et al., 2021; Andriushchenko et al., 2023) where noise-approximated-SGD is adopted for analysis. Also, we study the behavior of large LR SGD by focusing on the role of *oscillation*, which is largely different from the prior works.

**Feature learning in deep learning theory.** There has been a long line of research in deep learning theory from the perspective of *feature learning* during training of neural network (Allen-Zhu and Li, 2022; Wen and Li, 2021; Zou et al., 2022; Cao et al., 2022; Chen et al., 2022; Zou et al., 2023; Huang et al., 2023; Yang et al., 2023). The idea is that, by explicitly characterizing the dynamics of feature learning during training, one can figure out how different algorithms and data structures can influence the learning of features by the neural network, further uncovering the properties of interest in deep learning, e.g., ensemble (Allen-Zhu and Li, 2022), adaptive gradients (Zou et al., 2022), the phenomenon of benign overfitting (Cao et al., 2022), data augmentation via mixup (Zou et al., 2023), etc. Specifically, the work of Cao et al. (2022) showed that under small learning rate regimes, training on data with low *signal-to-noise ratio* (SNR) would result in *harmful overfitting*, leading to poor generalization abilities of the neural network. Our work extends this line of research to the less theoretically understood regime of large learning rates by characterizing the feature learning process when oscillation happens during gradient descent and explaining its benefits to generalization.

# D Conclusions

This work theoretically investigated NN training with large learning rates and established a theoretical framework to understand the oscillation phenomenon. We revealed the benefit of oscillation to the NN generalization, which we summarize as the phenomenon of "*benign oscillation*". Our theory demystified the phenomenon based on a feature learning perspective and showed that the oscillation can drive the learning of weak but important patterns from data that are crucial to generalization. Our theory shed light on the understanding of large learning rate NN training and provided useful guidance towards the optimization analysis when smooth convergence is guaranteed.

## E Preliminary Lemmas on Concentration

In this section, we give finite-sample concentration results to characterize the high-probability deterministic properties of the random elements in the problem. Now we fix a small constant $p > 0$.

**Lemma 7.** *Suppose that $n \geq 8\log(4/p)$, then with probability at least $1 - p$, we have that*

$$|\{i \in [n] : y_i = 1\}| \wedge |\{i \in [n] : y_i = -1\}| \geq \frac{n}{4}.$$

**Lemma 8.** *Suppose that $n \geq 8\log(4/p)$, then with probability at least $1 - p$, we have that*

$$|\mathcal{W}| \leq \frac{2\rho}{n}.$$

**Lemma 9.** *Suppose that $d = \Omega(\log(4n/p))$, then with probability $1 - p$, we have that*

$$\sigma_p^2 d/2 \leq \|\boldsymbol{\xi}\|_2^2 \leq 3\sigma_p^2 d/2,$$
$$|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle| \leq 2\sigma_p^2 \cdot \sqrt{d\log(2n/p)}$$

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

**Lemma 10.** *Suppose that $d \geq \Omega(\log(mn/p))$. Then with probability at least $1 - p$, we have that*

$$\sigma_0\|\mathbf{u}\|/2 \leq \max_{j,r} \langle \mathbf{w}_{j,r}^{(0)}, j\mathbf{u} \rangle \leq \sqrt{2\log(16m/p)} \cdot \sigma_0\|\mathbf{u}\|,$$
$$\min_{j,r} \langle \mathbf{w}_{j,r}^{(0)}, j\mathbf{u} \rangle \geq -\sqrt{2\log(16m/p)} \cdot \sigma_0\|\mathbf{u}\|,$$
$$\sigma_0\|\mathbf{v}\|/2 \leq \max_{j,r} \langle \mathbf{w}_{j,r}^{(0)}, j\mathbf{v} \rangle \leq \sqrt{2\log(16m/p)} \cdot \sigma_0\|\mathbf{v}\|,$$
$$\min_{j,r} \langle \mathbf{w}_{j,r}^{(0)}, j\mathbf{v} \rangle \geq -\sqrt{2\log(16m/p)} \cdot \sigma_0\|\mathbf{v}\|.$$
$$\sigma_0\sigma_p\sqrt{d}/4 \leq \max_{r \in [m]} j \cdot \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle \leq 2\sqrt{\log(8mn/p)} \cdot \sigma_0\sigma_p\sqrt{d}, \quad \forall i \in [n],$$

Please refer to Appendix A in Cao et al. (2022) for proofs for the above lemmas.

## F Proof for One-data Case

In this section, we give a formal statement of our main theory on the single noiseless training data setup, Theorem 2. We also provide the detailed proofs for this theorem. We begin with the formal statement on the conditions and assumptions required for Theorem 2.

Firstly, we put requirements on the data model, initialization, and learning rates.

**Assumption 11** (Conditions on hyperparameters)**.** *Suppose that the following holds:*

1. *The learning rate $m/(4\|\mathbf{u}\|_2^2) \leq \eta \leq 2m/(5\|\mathbf{u}\|_2^2)$*
2. *The weight initialization scale $\sigma_0 = \widetilde{\Theta}(\max\{\|\mathbf{u}\|_2, \|\mathbf{v}\|_2\}^{-1} \cdot d^{-1/2})$;*
3. *The signal strength $\|\mathbf{u}\|_2 > 0.1 \cdot \|\mathbf{v}\|_2$*
4. *The dimension $d$ satisfies $d = \Omega(\mathrm{polylog}(m))$.*

The first condition on the learning rate guarantees the regularity of the training trajectories, including the boundedness and the sign stability. The second condition on the weight initialization scale makes sure that the CNN is not initialized too large, which is common in practice and also helps regularize the training trajectory. The third condition on the signal strength separates the strong signal from the weak signal by their $\ell_2$-norm. Finally, the last condition on the dimension $d$ puts us in the regime of high dimensions for which independent Gaussian noise has small correlations.

The next assumption is on the training process, which requires that the SGD oscillates.

**Assumption 12** (Oscillations during training: single training data point case)**.** *We assume that there exists some constant $\delta \in (0.2, 0.8)$, such that $|yf(\mathbf{x}; \mathbf{W}^{(t)}) - 1| \geq \delta$ for any $t \geq 0$, where $(\mathbf{x}, y)$ denotes the single data point, $\mathbf{W}^{(t)}$ denotes the weights found by SGD* (4)*.*

We have discussed this assumption on oscillation in Sections 3 and B. Please refer to the main part of this paper for more illustration on the oscillation assumption.

With Assumptions 11 and 12, our formal statement of Theorem 2 is the following theorem.

**Theorem 13** (Restatement of Theorem 2)**.** *Under Assumptions 11 and 12, with probability at least* $1 - 1/\mathrm{poly}(d)$*, there exists a step* $T_{(\mathbf{v})}$ *such that for* $j = y$ *and any* $t \geq T_{(\mathbf{v})}$*, it holds that*

$$\frac{1}{m}\sum_{r\in[m]}\sigma(\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle) - \frac{1}{m}\sum_{r\in[m]}\sigma(\langle \mathbf{w}_{-j,r}^{(t)}, j\mathbf{v}\rangle) \geq \frac{\delta}{4},$$

*where* $\delta > 0$ *is specified in Assumption 12 and* $T_{(\mathbf{v})} \leq \widetilde{\Theta}\big(m\cdot\eta\cdot\|\mathbf{v}\|_2^{-2}\cdot\delta^{-1}\cdot\log(m\delta\sigma_0^{-1}\|\mathbf{v}\|_2^{-2})\big)$.

*Proof of Theorem 13.* Please refer to Appendix F.3 for a detailed proof. □

The following of this section is organized as following. Appendix F.1 presents important properties of the whole training dynamics and the CNN, which serve as the basis for all the following proofs. Appendix F.2 presents the fundamental step that allows for proving weak signal learning. Based on that, we prove the main theorem in Appendix F.3. Finally, all the remaining subsections prove other lemmas involved in Appendix F.

### F.1 Basic Properties of Oscillating Dynamics and the Two-Layer CNN

**Properties of training dynamics.** We first define some neuron subsets. For $j \in \{\pm 1\}$, we define

$$\mathcal{U}_{j,+}^{(t)} = \{r\in[m] : \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle > 0\}, \quad \mathcal{U}_{j,-}^{(t)} = \{r\in[m] : \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle \leq 0\}.$$

and

$$\mathcal{V}_{j,+}^{(t)} = \{r\in[m] : \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle > 0\}, \quad \mathcal{V}_{j,-}^{(t)} = \{r\in[m] : \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle \leq 0\}.$$

According to the update formula (4), we know that the (stochastic) gradient descent iterates the inner products $\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle,\ j = \pm 1$ as follows:

$$\langle \mathbf{w}_{y,r}^{(t+1)}, y\mathbf{u}\rangle = \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle + \frac{\eta\|\mathbf{u}\|_2^2}{m}\cdot\big(1 - yf(\mathbf{x};\mathbf{W}^{(t)})\big)\cdot\sigma'\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle\big), \tag{9}$$

$$\langle \mathbf{w}_{-y,r}^{(t+1)}, -y\mathbf{u}\rangle = \langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle + \frac{\eta\|\mathbf{u}\|_2^2}{m}\cdot\big(1 - yf(\mathbf{x};\mathbf{W}^{(t)})\big)\cdot\sigma'\big(-\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle\big). \tag{10}$$

Analogously, we have that

$$\langle \mathbf{w}_{y,r}^{(t+1)}, y\mathbf{v}\rangle = \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle + \frac{\eta\|\mathbf{v}\|_2^2}{m}\cdot\big(1 - yf(\mathbf{x};\mathbf{W}^{(t)})\big)\cdot\sigma'\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle\big),$$

$$\langle \mathbf{w}_{-y,r}^{(t+1)}, -y\mathbf{v}\rangle = \langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{v}\rangle + \frac{\eta\|\mathbf{v}\|_2^2}{m}\cdot\big(1 - yf(\mathbf{x};\mathbf{W}^{(t)})\big)\cdot\sigma'\big(-\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{v}\rangle\big).$$

We invite readers to some facts that will help understand the behavior of the inner products during the training processes. First, we note that for any $r \in \mathcal{U}_{-y,+}^{(0)}$, $\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle > 0$ stays fixed at its initialization, thus automatically keep fixed signs. Same phenomenon can be verified on $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle$ with $r \in \mathcal{U}_{y,-}^{(0)}$. The intuition behind this phenomenon is that, prediction $f(\cdot;\mathbf{W}^{(t)})$ on the single data with label $y$, i.e.,

$$\begin{aligned}
&f(\mathbf{x};\mathbf{W}^{(t)})\\
&\quad= \frac{y}{m}\sum_{r\in[m]}\Big\{\sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle\big) + \sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle\big) - \sigma\big(\langle \mathbf{w}_{-y,r}^{(t)}, y\mathbf{u}\rangle\big) - \sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, -y\mathbf{v}\rangle\big)\Big\}\\
&\quad= \frac{y}{m}\Big\{\sum_{r\in\mathcal{U}_{y,+}^{(t)}}\sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle\big) + \sum_{r\in\mathcal{V}_{y,+}^{(t)}}\sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle\big)\\
&\qquad - \sum_{r\in\mathcal{U}_{-y,-}^{(t)}}\sigma\big(-\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle\big) - \sum_{r\in\mathcal{V}_{-y,-}^{(t)}}\sigma\big(-\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{v}\rangle\big)\Big\},
\end{aligned}$$

only involves $\mathbf{w}_{y,r}^{(t)}$ with $r \in \mathcal{U}_{y,+}^{(t)} \cup \mathcal{V}_{y,+}^{(t)}$ and $\mathbf{w}_{-y,r}^{(t)}$ with $r \in \mathcal{U}_{-y,-}^{(t)} \cup \mathcal{V}_{-y,-}^{(t)}$. Moreover, the orthogonality assumption $\langle \mathbf{u}, \mathbf{v} \rangle = 0$ ensures that, the dynamics between $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u} \rangle$ and $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v} \rangle$ are independent throughout the gradient descent process.

Thanks to these facts and the signal strength regime in Assumption 12, we can then retreat to tracking the movements of $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u} \rangle, r \in \mathcal{U}_{y,+}^{(t)}$ and $\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u} \rangle, r \in \mathcal{U}_{-y,-}^{(t)}$ whenever the weak signal is not learned and the strong signal dominates. Ideally, only the inner products $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u} \rangle, r \in \mathcal{U}_{y,+}^{(t)}$ will take the lead.

A natural and crucial question is the boundedness of these inner products, which turned out to be the cornerstone for the subsequent analysis. A straightforward but helpful lemma indicates that the inner products that are initialized to be the maximal (resp. minimal) among all inner products continue to be the maximal (resp. minimal) throughout the training process. To put formally, we have

**Lemma 14** (Maximum and minimum neurons)**.** *Suppose that the signs of all the related inner products do not change throughout $[t_1, t_2]$. Then we have*

$$\begin{aligned} \operatorname*{argmax}_{r \in [m]} \langle \mathbf{w}_{y,r}^{(t_1)}, y\mathbf{u} \rangle &= \operatorname*{argmax}_{r \in [m]} \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u} \rangle, \\ \operatorname*{argmin}_{r \in [m]} \langle \mathbf{w}_{-y,r}^{(t_1)}, -y\mathbf{u} \rangle &= \operatorname*{argmin}_{r \in [m]} \langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u} \rangle \end{aligned}$$

*hold for all $t \in [t_1, t_2]$.*

*Proof of Lemma 14.* See the first part in Section F.4. □

A direct profit of this lemma is that it suffices to track two specific indices, the maximum and the minimum, to analyze upper and lower bounds for all $r \in [m]$.

**Single neuron behaves similarly to CNN.** The proof of boundedness utilizes another property of two-layer CNN defined in Equation (1) that exhibits the connections between the behavior of inner product $\langle \mathbf{w}_{y,r}, y\mathbf{u} \rangle$ and the outcome of the model $f(\cdot)$. We state it as follows.

**Lemma 15** (Single neuron imitates entire CNN)**.** *Define the major part of $yf(\mathbf{x}; \mathbf{W})$ as*

$$g(\mathbf{x}, y; \mathbf{W}) = \frac{1}{m} \sum_{r \in [m]} \sigma\big(\langle \mathbf{w}_{y,r}, y\mathbf{u} \rangle\big).$$

*Suppose that there exists $t_1 < t_2$ such that $\mathcal{U}_{y,+}^{(t)} \equiv \mathcal{U}_{y,+}^{(t_1)}$ for all $t \in [t_1, t_2]$. Then, for any $c > 0$, $g(\mathbf{x}, y; \mathbf{W}^{(t)}) \geq c$ implies that*

$$\max_{r \in [m]} \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u} \rangle \geq (\beta_{\mathbf{u}}^{*(t_1)} mc)^{1/2}.$$

*On the other hand, $g(\mathbf{x}; \mathbf{W}^{(t)}) \leq c$ implies that*

$$\max_{r \in [m]} \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u} \rangle \leq (\beta_{\mathbf{u}}^{*(t_1)} mc)^{1/2}.$$

*Here $\beta_{\mathbf{u}}^{*(t)} = \max_{r \in [m]} \sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u} \rangle) / \sum_{r \in [m]} \sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u} \rangle)$.*

*Proof of Lemma 15.* See the second part in Section F.4. For multiple data setting, the proof can be found in Section H.6. □

Being a subtle condition required in the previous lemmas, whether the signs of these inner products are changing throughout the process remains unknown, making the behaviors of these inner products more complicated. The answer to this question is affirmative, as we're able to prove that, under proper conditions, the signs of these inner products are fixed throughout the process. Due to some technical restrictions in the proof, the boundedness and the sign stability are proved simultaneously through a sophisticated inductive argument. The formal statement of the lemma is as follows.

We first define a stopping time. Let

$$T_{(\mathbf{v})} = \min_{t\geq 0}\left\{t : \frac{1}{m}\sum_{r\in[m]}\sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle\big) > \delta\right\}.$$

Such a stopping time helps to control the non-dominating terms in the CNN. Moreover, once the process reaches $T_{(\mathbf{v})}$, the conclusion in Theorem 13 is nearly achieved. With the help of the controls over the non-dominating terms, we have the following results.

**Lemma 16** (Boundedness and sign stability: single training data case)**.** *Suppose Assumptions 11 and 12 hold. With probability at least* $1 - 1/\mathrm{poly}(d)$*, we have the following bounds:*

$$\max_r \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle \leq 1.5\cdot\big(1.05\beta_{\mathbf{u}}^* m\big)^{1/2}, \quad \forall t \leq T_{(\mathbf{v})}, \tag{11}$$

$$\min_r \langle -\mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle \geq -\sqrt{2\log(16m/p)}\cdot\sigma_0\|\mathbf{u}\|_2 \quad \forall t \leq T_{(\mathbf{v})} \tag{12}$$

$$\min_r \langle -\mathbf{w}_{-y,r}^{(t)}, -y\mathbf{v}\rangle \geq -\sqrt{2\log(16m/p)}\cdot\sigma_0\|\mathbf{v}\|_2 \quad \forall t \leq T_{(\mathbf{v})} \tag{13}$$

$$0 \leq yf(\mathbf{x}; \mathbf{W}^{(t)}) \leq 3, \quad \forall t \leq T_{(\mathbf{v})}.$$

*Here* $\beta_{\mathbf{u}}^* = \beta_{\mathbf{u}}^{*(0)}$ *is defined in Lemma 15. Besides, the sign stability for* $\mathbf{u}$ *and* $\mathbf{v}$ *is true on* $[0, T_{(\mathbf{v})}]$*, i.e.* $\mathcal{U}_{\pm y,\pm}^{(t)}$ *and* $\mathcal{V}_{\pm y,\pm}^{(t)}$ *remains invariant in* $t$*, and the superscript* $(t)$ *can be dropped.*

*Proof of Lemma 16.* See the third part in Section F.4. □

Thanks to the stopping times defined previously which put controls on the scale of $\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle$, the major part function $g$ defined in the previous lemma dominates the entire CNN, as the negative parts $\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle$, $\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{v}\rangle$ can be lower bounded in Lemma 16 through a delicate analysis of the whole dynamics.

These three lemmas reveal the key properties of the training dynamic. Several remarks are put here again. Lemmas 14 and 15 are temporarily local, with the condition on the local sign stability. They are not informative until we are able to extend the sign stability to a wider sense, which is achieved by Lemma 16. Nevertheless, these two local lemmas are used frequently throughout the subsequent analysis, so we single them out here to make the proof more readable.

### F.2 Fundamental Reasons towards the Weak Signal Learning

Previous section present several basic properties of the training dynamics as well as the the prediction model-CNN itself. However, they are insufficient in interpreting the driving force of the weak signal learning with oscillation. In the lemma below, we discover a quantitative interpretation towards the increasing on $\langle \mathbf{w}_{y,r}^{(T)}, y\mathbf{v}\rangle$.

**Lemma 17.** *Under Assumption 11 and 12, suppose that there exists* $t_0 \leq t_1$*, such that:*

1. *Sign stability holds for* $\mathbf{u}$ *and* $\mathbf{v}$ *on* $[t_0, t_1]$*;*
2. $\max_{r\in[m]}\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle < B\cdot(\beta_{\mathbf{u}}^* m)^{1/2},\ \forall t \in [t_0, t_1]$*;*
3. $\min_{r\in[m]}\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle > -0.1$ *and* $\min_{r\in[m]}\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{v}\rangle > -0.1,\ \forall t \in [t_0, t_1]$*;*
4. $m^{-1}\sum_{r\in[m]}\sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle\big) < \delta,\ \forall t \in [t_0, t_1]$.
5. $-2 \leq 1 - yf(\mathbf{x}; \mathbf{W}^{(t)}) \leq 1,\ \forall t \in [t_0, t_1]$.

*Then we have*

$$\sum_{s=t_0}^{t_1-1}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big) \geq 2\epsilon\cdot(t_1 - t_0) - \frac{mB}{\eta\|\mathbf{v}\|_2^2\sqrt{1.05-\delta}}.$$

*And consequently for $r \in \mathcal{V}_{y,+} = \{r \in [m] : \langle \mathbf{w}_{y,r}^{(0)}, y\mathbf{v}\rangle > 0\}$, it holds that*

$$\langle \mathbf{w}_{y,r}^{(t_1)}, y\mathbf{v}\rangle \geq \langle \mathbf{w}_{y,r}^{(t_0)}, y\mathbf{v}\rangle \cdot \exp\Big\{\frac{\eta\|\mathbf{v}\|_2^2}{m}\cdot \epsilon t - \frac{\|\mathbf{v}\|_2^2}{\|\mathbf{u}\|_2^2}\cdot \frac{B}{(1.05-\delta)^{1/2}}\Big\},$$

*where $\epsilon = (\delta - \delta(1.05-\delta)^{1/2})/4$.*

*Proof of Lemma 17.* See Section F.5. □

This lemma asserts that, stable oscillation in an bounded area could leads to a linear increasing lower bound for $\sum_t (1 - yf_t)$. This is the first part of Lemma 17. The second part of this lemma relates the increasing speed of $\langle \mathbf{w}_{j,r}, j\mathbf{v}\rangle$ to the summation of $1 - yf_t$. The derivation of this part relies on the fact that $\alpha = \|\mathbf{v}\|_2^2/\|\mathbf{u}\|_2^2 < 1$ and is close to 0. Then for $r \in \mathcal{V}_{y,+}$ we can approximate the ratio intuitively with Taylor expansion

$$\begin{aligned}\frac{\mathbf{w}_{y,r}^{(t)} y\mathbf{v}}{\mathbf{w}_{y,r}^{(t0)} y\mathbf{v}} &= \prod_{t'=0}^{t-1}\big\{1 + \alpha\tilde{\eta}(1 - yf(\mathbf{x}; \mathbf{W}^{(t')}))\big\} \\ &\approx 1 + \alpha\tilde{\eta}\sum_{t'=0}^{t-1}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t')})\big).\end{aligned}$$

The proof of this lemma justifies this intuition formally with more delicate analysis.

### F.3 Proof of Theorem 13

*Proof of Theorem 13.* We prove Theorem 13 by contradiction. Recall that in the previous section we have defined:

$$T_{(\mathbf{v})} = \min\{t : m^{-1}\cdot \sum_r \sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle\big) > \delta\}.$$

With this definition, our goal boils down to proving that $T_{(\mathbf{v})}$ is bounded by a finite time with explicit expression. To put precisely, we prove that $T_{(\mathbf{v})} < T_0 = \frac{m}{\eta\|\mathbf{v}\|_2^2\epsilon}\cdot\Big\{\log\frac{2\sqrt{m\delta}}{\sigma_0\|\mathbf{v}\|_2^2} + 1.5\frac{\|\mathbf{v}\|_2^2}{\|\mathbf{u}\|_2^2}\cdot\sqrt{\frac{1.05}{1-\delta}}\Big\}$, by contradiction. Here $\epsilon > 0$ is specified in Lemma 17 and $\delta$ is specified in Assumption 4. Suppose otherwise that $T_{(\mathbf{v})} \geq T_0$, then Lemma 17 implies that,

$$\begin{aligned}\langle \mathbf{w}_{y,r^*}^{(T_0)}, y\mathbf{v}\rangle &= \langle \mathbf{w}_{y,r^*}^{(0)}, y\mathbf{v}\rangle \cdot \exp\Big\{\frac{\eta\|\mathbf{v}\|_2^2}{m}\epsilon\cdot T_0 - \frac{1.5\|\mathbf{v}\|_2^2}{\|\mathbf{u}\|_2^2}\cdot\sqrt{\frac{1.0}{1.05-\delta}}\Big\} \\ &\geq \frac{1}{2}\sigma_0\|\mathbf{v}\|\cdot\frac{2\sqrt{m\delta}}{\sigma_0\|\mathbf{v}\|} \\ &\geq \sqrt{m\delta}.\end{aligned}$$

This leads to

$$\frac{1}{m}\sum_{r\in[m]}\sigma\big(\langle \mathbf{w}_{y,r}^{(T_0)}, y\mathbf{v}\rangle\big) \geq \frac{1}{m}\sigma\big(\langle \mathbf{w}_{y,r^*}, y\mathbf{v}\rangle\big) \geq \delta,$$

which contradicts to the definition of $T_{(\mathbf{v})}$, and therefore $T_{(\mathbf{v})} \leq T_0$.

Now we prove that the sequence $\{m^{-1}\sum_r \sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle\big)\}_{t\geq T_{(\mathbf{v})}}$ does not fall below $\delta/2$. Intuitively, as long as $\sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle\big) < \delta$, the sequence would continue to increase as the results in Lemma 17 are revived based on the boundedness. The analysis resembles the proof of Lemma 16 with slight differences. We provide the following proposition and with the analysis delayed to Section F.6.

**Proposition 18.** *It holds that*

$$\frac{1}{m}\sum_{r\in[m]}\sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle\big) > \frac{\delta}{2}, \quad \forall t \geq T_{(\mathbf{v})}.$$

This proposition finalizes the proof of Theorem 13, and we're done. □

### F.4 Proof of Lemmas in Section F.1

*Proof of Lemma 14.* By our assumption that the signs of the inner products does not change throughout $[t_1, t_2]$, it is straightforward that $\mathcal{U}_{y,+}^{(t)} = \mathcal{U}_{y,+}^{(t_1)}$ for every $t \in [t_1, t_2]$. Same is true for $\mathcal{U}_{-y,-}^{(t)}$. Therefore, we're able to drop to superscript $(t)$ temporarily, as the attention is restricted to a local interval $[t_1, t_2]$.

Regarding the maximal index, introducing $r' \neq r \in \mathcal{U}_{y,+}$, we have following relation

$$\begin{aligned}
\operatorname*{argmax}_{r\in[m]} \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle &= \operatorname*{argmax}_{r\in\mathcal{U}_{y,+}} \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle, \\
&= \operatorname*{argmax}_{r\in\mathcal{U}_{y,+}} \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle / \langle \mathbf{w}_{y,r'}^{(t)}, y\mathbf{u}\rangle \\
&= \operatorname*{argmax}_{r\in\mathcal{U}_{y,+}} \frac{\langle \mathbf{w}_{y,r}^{(t_1)}, y\mathbf{u}\rangle \prod_{t'=t_1}^{t-1} \Big\{1 + \frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big)\Big\}}{\langle \mathbf{w}_{y,r'}^{(t_1)}, y\mathbf{u}\rangle \prod_{t'=t_1}^{t-1} \Big\{1 + \frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big)\Big\}} \\
&= \operatorname*{argmax}_{r\in\mathcal{U}_{y,+}} \langle \mathbf{w}_{y,r}^{(t_1)}, y\mathbf{u}\rangle / \langle \mathbf{w}_{y,r'}^{(t_1)}, y\mathbf{u}\rangle \\
&= \operatorname*{argmax}_{r\in\mathcal{U}_{y,+}} \langle \mathbf{w}_{y,r}^{(t_1)}, y\mathbf{u}\rangle.
\end{aligned}$$

The same relation can be verified for $\langle \mathbf{w}_{-y,r}, y\mathbf{u}\rangle$ with $r \in \mathcal{U}_{-y,-}$, finishing the proof. □

*Proof of Lemma 15.* Again, the local sign stability assumption ensures that each inner product grows proportionally and the superscript $(t)$ in the neuron index sets can be dropped, with

$$\begin{aligned}
g(\mathbf{x}, y; \mathbf{W}^{(t)}) &= \frac{1}{m} \sum_{r\in[m]} \sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle\big) \\
&= \frac{1}{m} \sum_{r\in\mathcal{U}_{y,+}} \sigma\big(\langle \mathbf{w}_{y,r}^{(t_1)}, y\mathbf{u}\rangle\big) \cdot \prod_{t'=t_1}^{t-1} \Big\{1 + \frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t')})\big)\Big\}^2 \\
&= \frac{\max_{r\in[m]} \sigma\big(\langle \mathbf{w}_{y,r}^{(t_1)}, y\mathbf{u}\rangle\big)}{m\beta_{\mathbf{u}}^{*,(t_1)}} \prod_{t'=t_1}^{t-1} \Big\{1 + \frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t')})\big)\Big\}^2 \\
&= \frac{\sigma\big(\max_{r\in[m]} \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle\big)}{m\beta_{\mathbf{u}}^{*,(t_1)}}
\end{aligned}$$

Here the second line and the last equality is true because Equation (9) implies that all the positive $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle$ iterates by sequentially multiplying the same factor $\Big\{1 + \frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t')})\big)\Big\}$. The third equality comes from the definition of $\beta_{\mathbf{u}}^{*,(t_1)}$ in Lemma 15. Therefore, $g(\mathbf{x}, y; \mathbf{W}^{(t)}) \geq c$ implies that $\sigma(\max_{r\in[m]} \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle) > \beta_{\mathbf{u}}^{*,(t_1)} mc$ and the desired lower bound follows. The upper bound can be proved analogously and is omitted here. □

*Proof of Lemma 16.* A roadmap is provided to help understanding how every single step is achieved so that the readers are encouraged to skip the details without leaving the key ideas behind. Therefore, we only suggest the readers with special interests to check through the full and detailed proof.

**Recap on Notations.** Recall that, $\mathcal{U}_{j,+}^{(t)}$ is the set of indices $r \in [m]$ such that $\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle > 0$ and $\mathcal{U}_{j,-}^{(t)}$ is the set of indices $r \in [m]$ such that $\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle \leq 0$. Specially, let $\mathcal{U}_{y,+} = \mathcal{U}_{y,+}^{(0)}$ and $\mathcal{U}_{-y,-} = \mathcal{U}_{-y,-}^{(0)}$. With probability one, it holds that $\mathcal{U}_{j,-} \cup \mathcal{U}_{j,+} = [m]$. Let $r^* \coloneqq \operatorname{argmax}_{r\in[m]} \langle \mathbf{w}_{y,r}^{(0)}, y\mathbf{u}\rangle$ and $r_* \coloneqq \operatorname{argmin}_{r\in[m]} \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{u}\rangle$. The index $r^*$ (resp. $r_*$) will be the maximum (resp. minimum) throughout the process as we are able to extend the results in Lemma 14 globally.

We also introduce several supplemental notations here to facilitate the proof. Recursively, we define

$$\bar{T}_k \coloneqq \min\big\{t : t > \bar{T}_{k-1},\ yf(\mathbf{x}; \mathbf{W}^{(t)}) \geq 1 \text{ and } yf(\mathbf{x}; \mathbf{W}^{(t-1)}) < 1\big\},$$

with $\bar{T}_0 = 0$. Similarly, we define that

$$\underline{T}_k \coloneqq \min\left\{t : t > \underline{T}_{k-1},\ yf(\mathbf{x}; \mathbf{W}^{(t)}) < 1 \text{ and } yf(\mathbf{x}; \mathbf{W}^{(t-1)}) \geq 1\right\},$$

and $\underline{T}_0 = 0$.

**Roadmap.** From a high level, three steps are required to establish the full proof:

1. Verify that the lower bound in Inequality (12) in Lemma 16 holds for $t \in [0, \bar{T}_1]$ with a direct monotonicity argument. Additionally, we can prove that the upper bound in Inequality (11) hold for $t \in [0, \bar{T}_1)$ by using Lemma 15 and the definition of $\bar{T}_1$. The signs do not change in this stage, as shown in the details below.
2. We extend the results in Lemma 16 to $t \in [\bar{T}_1, \bar{T}_2)$ with repeated use of Lemma 15. The sign stability is guaranteed from an intermediate upper bound on $|1 - yf(\mathbf{x}; \mathbf{W}^{(t)})|$.
3. Note that the condition on $[0, \bar{T}_1)$ (which is proved in the first step) required for the proof of the second step is again true for $t \in [0, \bar{T}_2)$, which is a consequence of the second step. Thus we can repeat the second step to extend the results in Lemma 16 to $t \in [\bar{T}_2, \bar{T}_3)$, and so on. So the results are true for all $t \leq T_{(\mathbf{v})}$.

The first and the last step above are relatively straightforward. However, the second step requires a delicate break-down analysis. Here we provide a detailed roadmap for the second step. The goal is to prove the results in Lemma 16, restricted to $t \in [\bar{T}_1, \bar{T}_2)$. This would be achieved in 4 split steps:

2.1 Firstly, we prove the upper bound in Inequality (11) for $t = \bar{T}_1$ by using one-step gradient descent and the upper bound for $\langle \mathbf{w}_{y,r^*}^{(\bar{T}_1 - 1)}, y\mathbf{u}\rangle$. With a monotonicity argument, $\langle \mathbf{w}_{y,r^*}^{(t)}, y\mathbf{u}\rangle \leq \langle \mathbf{w}_{y,r^*}^{(\bar{T}_1)}, y\mathbf{u}\rangle,\ t \in [\bar{T}_1, \underline{T}_1]$. Besides, for $t \in [\bar{T}_1, \underline{T}_1)$ we can give a lower bound on $\langle \mathbf{w}_{y,r^*}^{(t)}, y\mathbf{u}\rangle$ with the help of Lemma 15.

2.2 Based on previous lower and tight upper bounds on $\langle \mathbf{w}_{y,r^*}^{(\bar{T}_1)}, y\mathbf{u}\rangle$, we can derive a worse-case lower bound on $\langle \mathbf{w}_{y,r^*}^{(\underline{T}_1)}, y\mathbf{u}\rangle$ with one-step gradient descent. We use this worst-case tight lower bound to conclude the sign stability. This step is free of the lower bound on $\langle \mathbf{w}_{-y,r_*}^{(t)}, -y\mathbf{u}\rangle$ for $t \in (\bar{T}_1, \underline{T}_2]$, which we have not yet proved to be true.

2.3 Now we lower bound $\langle \mathbf{w}_{-y,r_*}^{(\underline{T}_1)}, -y\mathbf{u}\rangle$ and $\langle \mathbf{w}_{-y,r_*}^{(\underline{T}_1)}, -y\mathbf{v}\rangle$. This can be achieved by a delicate usage of the lower bound on $\langle \mathbf{w}_{y,r^*}^{(\underline{T}_1)}, y\mathbf{u}\rangle$, which we have proved in Step 2.2, plus an inequality that connects the relative increment of $\langle \mathbf{w}_{y,r^*}^{(t)}, y\mathbf{u}\rangle$ and $\langle \mathbf{w}_{-y,r_*}^{(t)}, -y\mathbf{u}\rangle$ or $\langle \mathbf{w}_{-y,r_*}^{(t)}, -y\mathbf{v}\rangle$. Thus we have proved Inequalities (12) and (13) for $t = \underline{T}_1$, and this can be further extended to the entire $[\bar{T}_1, \bar{T}_2]$ by another monotonicity argument since $\underline{T}_1$ si the local minima.

2.4 The remaining to is upper bound $\langle \mathbf{w}_{y,r^*}^{(t)}, y\mathbf{u}\rangle$ for $t \in (\underline{T}_1, \bar{T}_2)$. This is again a consequence of Lemma 15, as exactly what has been done to upper bound $\langle \mathbf{w}_{y,r^*}^{(t)}, y\mathbf{u}\rangle, t \leq \bar{T}_1$ in Step 1.

Now with the roadmap in mind, we are ready to dive into the details of every step.

**Step 1: Pre-$\bar{T}_1$ Analysis.** Lemma 10 indicates that the lower bound in Inequality (12) holds at initialization $t = 0$ under Assumption 11. Moreover, upper bound on maximal initial inner products in Lemma 10 and Assumption 11 indicates that

$$\begin{aligned}
|yf(\mathbf{x}; \mathbf{W}^{(0)})| &\leq 2 \max_{r\in[m]} \langle \mathbf{w}_{y,r}^{(0)}, y\mathbf{u}\rangle^2 \vee \max_{r\in[m]} \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{u}\rangle^2 \\
&= O\Big(\sigma_0^2 \|\mathbf{u}\|_2^2 \cdot \mathrm{polylog}(m, d)\Big) \\
&\ll 1,
\end{aligned}$$

From this we know that $\bar{T}_1 \geq 1$ and the upper bound in Inequality (11) is true at $t = 0$.

The first step to do is to extend the lower on $\langle \mathbf{w}_{-y,r}, -y\mathbf{u}\rangle$ to $[1, \bar{T}_1]$. Definition of $\bar{T}_1$ implies that $yf(\mathbf{x}; \mathbf{W}^{(t)}) \leq 1$ for $t \in [0, \bar{T}_1)$. Therefore for $r \in \mathcal{U}_{-y,-}^{(0)}$, $1 - yf(\mathbf{x}; \mathbf{W}^{(t)}) > 0,\ t \in [0, \bar{T}_1)$ and Equation (10) gives that

$$\begin{aligned}\langle \mathbf{w}_{-y,r}^{(t+1)}, -y\mathbf{u}\rangle &= \langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle + \frac{\eta\|\mathbf{u}\|^2}{m}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big) \cdot \sigma'\big(-\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle\big) \\ &= \langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle - \frac{2\eta\|\mathbf{u}\|^2}{m}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big) \cdot \langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle \qquad (14) \\ &\geq \langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle.\end{aligned}$$

And furthermore $\langle \mathbf{w}_{-y,r}^{(\bar{T}_1)}, -y\mathbf{u}\rangle \geq \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{u}\rangle \geq -\sqrt{2\log(16m/p)} \cdot \sigma_0\|\mathbf{u}\|_2$. Taking the minimum with respect to $r \in [m]$ gives the result and analogous results can be derived for $\langle \mathbf{w}_{-y,r}^{(\bar{T}_1)}, -y\mathbf{v}\rangle$. So the lower bounds in Inequality (12) and (13)] hold for $t \in [1, \bar{T}_1]$. Equation (9) implies that $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle, r \in \mathcal{U}_{y,+}^{(0)}$ increasing for all $t < \bar{T}_1$. A natural consequence is that $yf(\mathbf{x}; \mathbf{W}^{(t)})$ is non-decreasing in $t$ in this stage, as every summand (with the negative sign before) in the summation is non-decreasing.

Now we prove the sign stability. For $r \in \mathcal{U}_{y,+}^{(0)}$, we already know that $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle > \langle \mathbf{w}_{y,r}^{(0)}, y\mathbf{u}\rangle > 0$, hence $r \in \mathcal{U}_{y,+}^{(t)}$ is non-decreasing in $t$ by induction. Additionally, for $r \in \mathcal{U}_{y,-}^{(0)}$, $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle$ stays fixed in $t$, as mentioned in Section F.1. Two points together ensure that $\mathcal{U}_{y,+}^{(t)} \equiv \mathcal{U}_{y,+}^{(0)}$ for $t \in [1, \bar{T}_1]$.

For $r \in \mathcal{U}_{-y,-}^{(0)}$, let's take a closer look at Equation (14) with $t = 0$:

$$\begin{aligned}\langle \mathbf{w}_{-y,r}^{(1)}, -y\mathbf{u}\rangle &= \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{u}\rangle - \frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(0)})\big) \cdot \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{u}\rangle \\ &= \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{u}\rangle \cdot \underbrace{\Big\{1 - \frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(0)})\big)\Big\}}_{>0 \text{ if } \eta < \left(1 - o(1)\right)/2 \cdot m\|\mathbf{u}\|_2^{-2}}.\end{aligned}$$

Assumption 11 ensures that $\eta < 0.4m\|\mathbf{u}\|_2^{-2} < (1 - o(1))/2 \cdot m\|\mathbf{u}\|_2^{-2}$ so the sign change does not happen at $t = 0$. As mentioned before, $yf(\mathbf{x}; \mathbf{W}^{(t)}) \leq 1$ is non-decreasing in $t$ at this stage, and putting them together we know

$$\Big\{1 - \frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big)\Big\} \geq 0 \text{ for } t \in [1, \bar{T}_1).$$

The same can be verified for $\langle \mathbf{w}_{-y,r}, -y\mathbf{v}\rangle$, and therefore, the sign stability for $\mathcal{U}_{y,-1}^{(t)}$ and $\mathcal{V}_{y,-1}^{(t)}$ is ensured for $t \in [0, \bar{T}_1]$ and the lower bound in Inequality (12) holds for $t \in [1, \bar{T}_1]$.

Now we turn to prove the upper bound (11). Note that, $yf(\mathbf{x}; \mathbf{W}^{(t)}) \leq 1, t < \bar{T}_1$ and the definition of $T_{(\mathbf{v})}$ implies that $g(\mathbf{x}, y; \mathbf{W}^{(t)}) \leq 1 + 2(\sqrt{2\log(16m/p)} \cdot \sigma_0\|\mathbf{u}\|_2)^2 \leq 1.05$. Now that the local sign stability holds for $t \in [0, \bar{T}_1)$, Lemma 15 gives that

$$\max_{r\in[m]} \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle \leq \big(1.05\beta_{\mathbf{u}}^* m\big)^{1/2}. \qquad (15)$$

Here $\beta_{\mathbf{u}}^*$ is defined in 15 with the superscript $(0)$ dropped.

**Step 2.1: Bounding** $\max_r \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle,\ t \in [\bar{T}_1, \underline{T}_1)$**.** In order for the upper bound to be tight, we need a lower bound on $yf(\mathbf{x}; \mathbf{W}^{(\bar{T}-1)})$. Let $\tilde{\eta} = 2\eta\|\mathbf{u}\|_2^2/m > 1/2$, we have the following result.

**Proposition 19.** *For every $k \geq 1$, suppose that*

$$E_{\bar{T}_k - 1} \coloneqq \frac{1}{m}\sum_{r\in[m]} \Big\{\sigma\big(-\langle \mathbf{w}_{-y,r}^{(\bar{T}_k - 1)}, -y\mathbf{u}\rangle\big) + \sigma\big(-\langle \mathbf{w}_{-y,r}^{(\bar{T}_k - 1)}, -y\mathbf{v}\rangle\big)\Big\} < \delta/2.$$

*Then we have*

$$yf(\mathbf{x}; \mathbf{W}^{(\bar{T}_k - 1)}) \geq \frac{1}{2\tilde{\eta}}\Big(2 + \tilde{\eta} - \sqrt{\tilde{\eta}^2 + 4\tilde{\eta}}\Big). \qquad (16)$$

*Moreover, it holds that*

$$yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k)}) \le yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)}) \cdot \Big\{1+\tilde{\eta}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2 + 2E_{\bar{T}_k-1}.$$

*Also, it is notable that the result here still holds for $\bar{T}_k \ge T_{(\mathbf{v})}$.*

*Proof of Proposition 19.* See Section F.6. □

Clearly, the condition required for Proposition 19 is true for before $\bar{T}_1$. Consider one-step gradient descent at $t = \bar{T}_1 - 1$ with Equation 9:

$$\begin{aligned}
\langle \mathbf{w}_{y,r^*}^{(\bar{T}_1)}, y\mathbf{u}\rangle &= \langle \mathbf{w}_{y,r^*}^{(\bar{T}_1-1)}, y\mathbf{u}\rangle + \frac{\eta\|\mathbf{u}\|_2^2}{m}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_1-1)})\big)\cdot\sigma'\big(\langle \mathbf{w}_{y,r^*}^{(\bar{T}_1-1)}, y\mathbf{u}\rangle\big)\\
&\le \langle \mathbf{w}_{y,r^*}^{(\bar{T}_1-1)}, y\mathbf{u}\rangle\Big\{1+(1+\tilde{\eta}/2-\sqrt{\tilde{\eta}^2/4+\tilde{\eta}})\Big\}\\
&\le 1.5\cdot\big(1.05\beta_{\mathbf{u}}^* m\big)^{1/2}. \qquad (17)
\end{aligned}$$

Here the first inequality above comes from Inequality (16), and the second inequality is from taking the suprema $\tilde{\eta} = 1/2$ and Inequality (15). Additionally, we can further deliver an upper bound on $yf(\mathbf{x};\mathbf{W}^{(\bar{T}_1)})$. From Inequality (33) in the proof of Proposition 19 and Inequality (16), we have

$$\begin{aligned}
yf(\mathbf{x};\mathbf{W}^{(\bar{T}_1)}) &\le yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)}) \cdot \Big\{1+\tilde{\eta}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2 + o(1)\\
&\le \Big\{1+\tilde{\eta}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2 + o(1)\\
&\le \Big\{1+\big(\tilde{\eta}-1-\tilde{\eta}/2+\sqrt{\tilde{\eta}^2/4+\tilde{\eta}}\big)\Big\}^2 + o(1)\\
&= \big(\tilde{\eta}/2+\sqrt{\tilde{\eta}^2/4+\tilde{\eta}}\big)^2 + o(1). \qquad (18)
\end{aligned}$$

Since $\tilde{\eta} < 1$, we know that $yf(\mathbf{x};\mathbf{W}^{(\bar{T}_1)}) \le 3$ and $|yf(\mathbf{x};\mathbf{W}^{(\bar{T}_1)}) - 1| \le 2$. We keep $\tilde{\eta}$ in the upper bound above for deriving a sufficient condition on $\tilde{\eta}$ for the sign stability.

Now we look to the lower bound. Definition of $\bar{T}_1, T_{(\mathbf{v})}$ and Assumption 12 implies that $yf(\mathbf{x};\mathbf{W}^t) > 1+\delta$, hence $g(\mathbf{x}, y; \mathbf{W}^{(t)}) > 1+\delta-\delta = 1$ for $t \in [\bar{T}_1, \underline{T}_1)$. Lemma 15 implies that

$$\langle \mathbf{w}_{y,r^*}^{(t)}, y\mathbf{u}\rangle > (\beta_{\mathbf{u}}^* m)^{1/2}, \quad t \in [\bar{T}_1, \underline{T}_1).$$

**Step 2.2: Lower Bounding $\langle \mathbf{w}_{y,r}^{(\underline{T}_1)}, y\mathbf{u}\rangle$** Note that $yf(\mathbf{x};\mathbf{W}^{(t)}) \le yf(\mathbf{x};\mathbf{W}^{(\bar{T}_1)})$ from the local monotonicity. Consider one-step gradient descent with Equation (9):

$$\begin{aligned}
\langle \mathbf{w}_{y,r^*}^{(\underline{T}_1)}, y\mathbf{u}\rangle &= \langle \mathbf{w}_{y,r^*}^{(\underline{T}_1-1)}, y\mathbf{u}\rangle \cdot \Big\{1-\tilde{\eta}\big(yf(\mathbf{x};\mathbf{W}^{(\underline{T}_1-1)})-1\big)\Big\}\\
&\ge \langle \mathbf{w}_{y,r^*}^{(\underline{T}_1-1)}, y\mathbf{u}\rangle \cdot \Big\{1-\tilde{\eta}\big(yf(\mathbf{x};\mathbf{W}^{(\bar{T}_1-1)})-1\big)\Big\}\\
&\ge (\beta_{\mathbf{u}}^* m)^{1/2} \cdot \underbrace{\Big\{1-\tilde{\eta}\big((\tilde{\eta}/2+\sqrt{\tilde{\eta}^2/4+\tilde{\eta}})^2-1\big)-o(1)\Big\}}_{>0 \text{ with } \tilde{\eta}<4/5}. \qquad (19)
\end{aligned}$$

Here the last inequality is a consequence of Inequality (18)] and the deterministic estimation (verified with software) that

$$\min_{\tilde{\eta}\in[1/2,4/5]} \Big\{1-\tilde{\eta}\big((\tilde{\eta}/2+\sqrt{\tilde{\eta}^2/4+\tilde{\eta}})^2-1\big)\Big\} > 0.2. \qquad (20)$$

Every step above is free of the lower bound in Inequality (12). Therefore, the sign stability is true on $[\bar{T}_1, \underline{T}_1]$, because all the inner products are non-decreasing in $t \in [\underline{T}_1, \bar{T}_2]$

**Step 2.3: Lower Bounding $\langle \mathbf{w}_{-y,r}^{(\underline{T}_1)}, -y\mathbf{u}\rangle$ and $\langle \mathbf{w}_{-y,r}^{(\underline{T}_1)}, -y\mathbf{v}\rangle$** The sign stability on $[0, \underline{T}_1]$ guarantees that for every $t \leq \underline{T}_1 - 1$, $1 \pm \tilde{\eta}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)}\big) > 0$ consequently $1 + \tilde{\eta}\frac{\|\mathbf{v}\|_2^2}{\|\mathbf{u}\|_2^2}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)}\big) > 0$. From Equation (10), we have

$$\begin{aligned}\langle \mathbf{w}_{-y,r}^{(\underline{T}_1)}, -y\mathbf{u}\rangle &= \langle \mathbf{w}_{-y,}^{(0)}, -y\mathbf{u}\rangle \prod_{t=0}^{\underline{T}_1 - 1} \Big\{1 - \tilde{\eta}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)}\big)\Big\} \\ &\geq \langle \mathbf{w}_{-y,}^{(0)}, -y\mathbf{u}\rangle \exp\Big\{ - \tilde{\eta} \sum_{t=0}^{\underline{T}_1 - 1} \big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)}\big)\Big\} \end{aligned} \tag{21}$$

On the other aspect

$$\begin{aligned}\frac{\langle \mathbf{w}_{y,r^*}^{(\underline{T}_1)}, y\mathbf{u}\rangle}{\langle \mathbf{w}_{y,r^*}^{(0)}, y\mathbf{u}\rangle} &= \prod_{t=0}^{\underline{T}_1 - 1} \Big\{1 + \tilde{\eta}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)}\big)\Big\} \\ &\leq \exp\Big\{ \sum_{t=0}^{\underline{T}_1 - 1} \tilde{\eta}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big)\Big\}\end{aligned}$$

However, $\langle \mathbf{w}_{y,r^*}^{(0)}, y\mathbf{u}\rangle \leq (\beta_{\mathbf{u}}^* m)^{1/2} \cdot \sqrt{2\log(16m/p)} \cdot \sigma_0 \|\mathbf{u}\|_2$ and Inequality (19) imply that

$$\langle \mathbf{w}_{y,r^*}^{(\underline{T}_1)}, y\mathbf{u}\rangle / \langle \mathbf{w}_{y,r^*}^{(0)}, y\mathbf{u}\rangle \geq \frac{(\beta_{\mathbf{u}}^* m)^{1/2} \cdot (0.2 - o(1))}{(\beta_{\mathbf{u}}^* m)^{1/2} \cdot \sqrt{2\log(16m/p)} \cdot \sigma_0 \|\mathbf{u}\|_2} \geq 1. \tag{22}$$

Combining Inequality (22) and (19) we can obtain that $\sum_{t=0}^{\underline{T}_1 - 1} \big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big) \geq 0$. Together with Inequality (21), this in turns leads to

$$\begin{aligned}0 > \min_r \langle \mathbf{w}_{-y,r}^{(\underline{T}_1)}, -y\mathbf{u}\rangle &= \min_r \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{u}\rangle \prod_{t=0}^{\underline{T}_1 - 1} \Big\{1 - \tilde{\eta}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big)\Big\} \\ &\geq \min_r \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{u}\rangle \\ &\geq -\sqrt{2\log(16m/p)} \cdot \sigma_0 \|\mathbf{u}\|_2.\end{aligned}$$

Analogously we have that

$$\begin{aligned}0 > \min_r \langle \mathbf{w}_{-y,r}^{(\underline{T}_1)}, -y\mathbf{v}\rangle &= \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{v}\rangle \prod_{t=0}^{\underline{T}_1 - 1} \Big\{1 - \tilde{\eta}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big)\Big\} \\ &\geq \min_r \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{v}\rangle \\ &\geq -\sqrt{2\log(16m/p)} \cdot \sigma_0 \|\mathbf{v}\|_2.\end{aligned}$$

In conclusion, we have that

$$\begin{aligned}&\max_r \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle \geq 0.2(\beta_{\mathbf{u}}^* m)^{1/2},\ t \in [\bar{T}_1, \bar{T}_2], \\ &\min_r \langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle \geq -\sqrt{2\log(16m/p)} \cdot \sigma_0 \|\mathbf{u}\|_2,\ t \in [\bar{T}_1, \bar{T}_2], \\ &\min_r \langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{v}\rangle \geq -\sqrt{2\log(16m/p)} \cdot \sigma_0 \|\mathbf{v}\|_2,\ t \in [\bar{T}_1, \bar{T}_2].\end{aligned}$$

Additionally, the lower bound on $\max_r \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle$ indicates that the sign stability holds on $[\bar{T}_1, \bar{T}_2)$. Since the last drop step does not change the sign of $\langle \mathbf{w}_{y,r}, y\mathbf{u}\rangle$. Once the $\mathbf{u}$-sign stability holds, $\mathbf{v}$-sign stability can be easily derived. To see this, note that the $\mathbf{u}$-sign stability implies that for any $t \in [0, T_{(\mathbf{v})}]$, we have $1 \pm \frac{\eta\|\mathbf{u}\|_2^2}{m} \cdot \big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)}\big) > 0$. Now $\|\mathbf{u}\| > \|\mathbf{v}\|$, one clearly sees that $1 \pm \frac{\eta\|\mathbf{v}\|_2^2}{m} \cdot \big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)}\big) > 0$ and the $\mathbf{v}$-sign stability holds.

**Step 2.4: Upper Bounding $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle,\ t \in (\underline{T}_1, \bar{T}_2)$.** This is exactly the same to the proof of Inequality 15, and is thus omitted.

**Step 3. Finalizing Proof.** At this point, all the results in Lemma 16 have been proved to be true on $t \in [\bar{T}_1, \bar{T}_2)$. It is important to note that the only inductive hypothesis used for the local extension on $[\bar{T}_1, \bar{T}_2)$ is the lower bound on $\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle,\ t \le \bar{T}_1$. The rest part merely comes from the definition of $\bar{T}_k$ and $\underline{T}_k$ and Assumption 12. □

**Remark 20.** *In the proof, we implicitly utilized the fact that $\bar{T}_k, \underline{T}_k < \infty,\ \forall k \ge 0$. One may conjecture that there could be cases that $yf(\mathbf{x}; \mathbf{W}^{(t)}) > 1$(resp. $< 1$) once $t \ge \bar{T}_k$ (resp. $\underline{T}_k$. However, Assumption 12 (guaranteed by tuning a proper $\eta$) indicates that this cannot happen. One can argue that once $yf(\mathbf{x}; \mathbf{W}^{(t)}) > 1$, the Assumption 12 enables the dynamic to bounce back towards 1 with at least exponential rate. Therefore, $yf(\mathbf{x}; \mathbf{W}^{(t)})$ falls below $1 + \delta$ within a few steps and Assumption 12 forces $yf(\mathbf{x}; \mathbf{W}^{(t)}) < 1 - \delta$, and $\underline{T}_k < \infty$. Same argument can be used to prove that $\bar{T}_k < \infty$.*

### F.5 Proof of Lemma 17

*Proof of Lemma 17.* Let $r^* = \mathrm{argmax}_r \langle \mathbf{w}_{y,r}^{(t_0)}, y\mathbf{u}\rangle$. For any $0 \le t_0 < t_1$, Equation 9 implies that

$$\begin{aligned}\langle \mathbf{w}_{y,r^*}^{(t_1)}, y\mathbf{u}\rangle =& \langle \mathbf{w}_{y,r^*}^{(t_0)}, y\mathbf{u}\rangle + \frac{2\eta\|\mathbf{u}\|_2^2}{m} \cdot \sum_{\substack{s\in[t_0,t_1-1]:\\ yf(\mathbf{x};\mathbf{W}^{(s)})\ge 1}} \big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big) \cdot \langle \mathbf{w}_{y,r^*}^{(s)}, y\mathbf{u}\rangle \\ &+ \frac{2\eta\|\mathbf{u}\|_2^2}{m} \cdot \sum_{\substack{s\in[t_0,t_1-1]:\\ yf(\mathbf{x};\mathbf{W}^{(s)})< 1}} \big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big) \cdot \langle \mathbf{w}_{y,r^*}^{(s)}, y\mathbf{u}\rangle. \end{aligned} \tag{23}$$

For $s \in [t_0, t]$, $m^{-1}\sum_r \sigma\big(\langle \mathbf{w}_{y,r}^{(s)}, y\mathbf{v}\rangle\big) < \delta$. Therefore, Assumption 12 and $yf(\mathbf{x}; \mathbf{W}^{(s)}) \ge 1$ (and $> 1 + \delta$) imply that $g(\mathbf{x}, y; \mathbf{W}^{(t)}) \ge 1 + \delta - \delta$, hence with Lemma 15 we have that

$$\langle \mathbf{w}_{y,r^*}^{(s)}, y\mathbf{u}\rangle \ge (\beta_{\mathbf{u}}^* m)^{1/2}, \tag{24}$$

On the other aspect, $\min_r \langle \mathbf{w}_{-y,r}^{(s)}, -y\mathbf{u}\rangle > -0.1$ and $\min_r \langle \mathbf{w}_{-y,r}^{(s)}, -y\mathbf{v}\rangle > -0.1$ lead to

$$\frac{1}{m}\sum_{r\in[m]} \Big\{\sigma\big(-\langle \mathbf{w}_{-y,r}^{(s)}, -y\mathbf{u}\rangle\big) + \sigma\big(-\langle \mathbf{w}_{-y,r}^{(s)}, -y\mathbf{v}\rangle\big)\Big\} \le 2 \times 0.1^2 < 0.05.$$

Therefore $yf(\mathbf{x}; \mathbf{W}^{(s)}) \le 1$ along with Assumption 12 indicate $g(\mathbf{x}, y; \mathbf{W}^{(s)}) \le 1 - \delta + 0.05$, hence

$$\langle \mathbf{w}_{y,r^*}^{(s)}, y\mathbf{u}\rangle \le (1.05 - \delta)^{1/2} \cdot (\beta_{\mathbf{u}}^* m)^{1/2}. \tag{25}$$

Meanwhile, the boundedness and $\mathbf{u}$-sign stability imply that

$$\Big|\langle \mathbf{w}_{y,r^*}^{(t_1)}, y\mathbf{u}\rangle - \langle \mathbf{w}_{y,r^*}^{(t_0)}, y\mathbf{u}\rangle\Big| \le |\langle \mathbf{w}_{y,r^*}^{(t_1)}, y\mathbf{u}\rangle| \vee |\langle \mathbf{w}_{y,r^*}^{(t_0)}, y\mathbf{u}\rangle| \le B(\beta_{\mathbf{u}}^* m)^{1/2}. \tag{26}$$

Putting Inequality (24), (25), (26) and Equation (23) together, we have

$$
\begin{aligned}
B \cdot (\beta_{\mathbf{u}}^* m)^{1/2} \geq \Bigg| & \frac{2\eta\|\mathbf{u}\|_2^2}{m} \cdot \sum_{\substack{s\in[t_0,t_1-1]:\\ yf(\mathbf{x};\mathbf{W}^{(s)})\geq 1}} \big(1 - yf(\mathbf{x};\mathbf{W}^{(s)})\big) \cdot \langle \mathbf{w}_{y,r^*}^{(s)}, y\mathbf{u}\rangle \\
& + \frac{2\eta\|\mathbf{u}\|_2^2}{m} \cdot \sum_{\substack{s\in[t_0,t_1-1]:\\ yf(\mathbf{x};\mathbf{W}^{(s)})< 1}} \big(1 - yf(\mathbf{x};\mathbf{W}^{(s)})\big) \cdot \langle \mathbf{w}_{y,r^*}^{(s)}, y\mathbf{u}\rangle \Bigg| \\
\geq & \frac{2\eta\|\mathbf{u}\|_2^2}{m} \cdot \sum_{\substack{s\in[t_0,t_1-1]:\\ yf(\mathbf{x};\mathbf{W}^{(s)})\geq 1}} \big(yf(\mathbf{x};\mathbf{W}^{(s)}) - 1\big) \cdot \langle \mathbf{w}_{y,r^*}^{(s)}, y\mathbf{u}\rangle \\
& - \frac{2\eta\|\mathbf{u}\|_2^2}{m} \cdot \sum_{\substack{s\in[t_0,t_1-1]:\\ yf(\mathbf{x};\mathbf{W}^{(s)})< 1}} \big(1 - yf(\mathbf{x};\mathbf{W}^{(s)})\big) \cdot \langle \mathbf{w}_{y,r^*}^{(s)}, y\mathbf{u}\rangle \\
\geq & \frac{2\eta\|\mathbf{u}\|_2^2}{m} \cdot (\beta_{\mathbf{u}}^* m)^{1/2} \cdot \Bigg\{ \sum_{\substack{s\in[t_0,t_1-1]:\\ yf(\mathbf{x};\mathbf{W}^{(s)})\geq 1}} \big(yf(\mathbf{x};\mathbf{W}^{(s)}) - 1\big) \\
& - (1.05-\delta)^{1/2} \sum_{\substack{s\in[t_0,t_1-1]:\\ yf(\mathbf{x};\mathbf{W}^{(s)})< 1}} \big(1 - yf(\mathbf{x};\mathbf{W}^{(s)})\big) \cdot \langle \mathbf{w}_{y,r^*}^{(s)}, y\mathbf{u}\rangle \Bigg\}
\end{aligned}
$$

This is equivalent to:

$$
\begin{aligned}
\sum_{\substack{s\in[t_0,t_1-1]:\\ yf(\mathbf{x};\mathbf{W}^{(s)})< 1}} \big(1 - yf(\mathbf{x};\mathbf{W}^{(s)})\big) \cdot \langle \mathbf{w}_{y,r^*}^{(s)}, y\mathbf{u}\rangle \geq & \sum_{\substack{s\in[t_0,t_1-1]:\\ yf(\mathbf{x};\mathbf{W}^{(s)})\geq 1}} (1.05-\delta)^{-1/2} \cdot \big(yf(\mathbf{x};\mathbf{W}^{(s)}) - 1\big) \\
& - \underbrace{\frac{mB}{2\eta\|\mathbf{u}\|_2^2\sqrt{1.05-\delta}}}_{:=\Delta(B)}.
\end{aligned}
$$

Now we are ready to lower bound the summation that we are interested in. Since

$$|\{s \in [t_0, t_1 - 1] : yf(\mathbf{x}; \mathbf{W}^{(s)}) < 1\}| + |\{s \in [t_0, t_1 - 1] : yf(\mathbf{x}; \mathbf{W}^{(s)}) > 1\}| = t_1 - t_0,$$

we have either: $|\{s \in [t_0, t_1 - 1] : yf(\mathbf{x}; \mathbf{W}^{(s)}) > 1\}| > (t_1 - t_0)/2$, which implies that

$$
\begin{aligned}
\sum_{s=t_0}^{t} \big(1 - yf(\mathbf{x};\mathbf{W}^{(s)})\big) = & \sum_{t_0\leq s\leq t_1-1: yf(\mathbf{x};\mathbf{W}^{(s)})<1} \big(1 - yf(\mathbf{x};\mathbf{W}^{(s)})\big) \\
& - \sum_{t_0\leq s\leq t_1-1: yf(\mathbf{x};\mathbf{W}^{(s)})>1} \big(yf(\mathbf{x};\mathbf{W}^{(s)}) - 1\big) \\
\geq & \big((1.05-\delta)^{-1/2} - 1\big) \cdot \sum_{t_0\leq s\leq t_1-1: yf(\mathbf{x};\mathbf{W}^{(s)})>1} \big(yf(\mathbf{x};\mathbf{W}^{(s)}) - 1\big) - \Delta(B) \\
\geq & \frac{1}{2}\big(\delta(1.05-\delta)^{-1/2} - \delta\big) \cdot (t_1 - t_0) - \Delta(B), \qquad (27)
\end{aligned}
$$

or $|\{s \in [t_0, t_1 - 1] : yf(\mathbf{x}; \mathbf{W}^{(s)}) < 1\}| > (t - t_0)/2$, which implies that

$$
\sum_{s=t_0}^{t_1-1} \big(1 - yf(\mathbf{x};\mathbf{W}^{(s)})\big) = \sum_{t_0\leq s\leq t_1-1: yf(\mathbf{x};\mathbf{W}^{(s)})<1} \big(yf(\mathbf{x};\mathbf{W}^{(s)}) - 1\big)
$$

$$
\begin{aligned}
&- \sum_{t_0 \le s \le t_1 - 1 : yf(\mathbf{x}; \mathbf{W}^{(s)}) > 1} \big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big) \\
&\ge \big(1 - (1.05 - \delta)^{1/2}\big) \cdot \sum_{t_0 \le s \le t_1 - 1_1 - 1 : yf(\mathbf{x}; \mathbf{W}^{(s)}) < 1} \big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big) \\
&\quad - (1.05 - \delta)^{1/2} \cdot \Delta \\
&\ge \frac{1}{2}\big(\delta - \delta(1.05 - \delta)^{1/2}\big) \cdot (t_1 - t_0) - (1.05 - \delta)^{1/2} \cdot \Delta(B). \qquad (28)
\end{aligned}
$$

In both cases we have used Assumption 12 to bound $yf(\mathbf{x}; \mathbf{W}^{(s)})$ from 1. Combining (27) and (28), we have that

$$
\sum_{s=t_0}^{t_1 - 1} \big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big) \ge \frac{1}{2}\big(\delta - \delta(1.05 - \delta)^{1/2}\big) \cdot (t_1 - t_0) - \Delta(B).
$$

For simplicity, we denote $\alpha = \|\mathbf{v}\|^2 / \|\mathbf{u}\|^2$ and $\epsilon = \big(\delta - \delta(1.05 - \delta)^{1/2}\big)/4$. Note that from the $\mathbf{v}$-sign stability condition, we have

$$
1 + \frac{2m\|\mathbf{v}\|_2^2}{m}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big) > 0, \quad \forall t \in [t_0, t_1].
$$

For $r \in \mathcal{V}_{y,+}$, note that $-2 \le 1 - yf(\mathbf{x}; \mathbf{W}^{(t)}) \le 1$. Therefore, we can lower bound the logarithmic ratio $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle / \langle \mathbf{w}_{y,r}^{(0)}, y\mathbf{v}\rangle$ as:

$$
\begin{aligned}
&\sum_{t=t_0}^{t_1 - 1} \log\Big\{1 + \alpha\tilde{\eta}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big)\Big\} = \sum_{t=t_0}^{t_1 - 1} \int_0^{\alpha} \frac{\tilde{\eta}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big)}{1 + \tilde{\eta}z\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big)} dz \\
&= \sum_{t=t_0}^{t_1 - 1} \int_0^{\alpha} \frac{\tilde{\eta}\Big(\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big) + 2\Big)}{1 + \tilde{\eta}z\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big)} dz - \sum_{t=t_0}^{t_1 - 1} \int_0^{\alpha} \frac{2\tilde{\eta}}{1 + \tilde{\eta}z\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big)} dz \\
&\ge \sum_{t=t_0}^{t_1 - 1} \int_0^{\alpha} \frac{\tilde{\eta}\Big(\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big) + 2\Big)}{1 + \tilde{\eta}z} dz - \sum_{t=t_0}^{t_1 - 1} \int_0^{\alpha} \frac{2\tilde{\eta}}{1 - 2\tilde{\eta}z} dz \\
&\ge \int_0^{\alpha} \frac{\tilde{\eta}\Big(\sum_{t=t_0}^{t_1 - 1}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big) + 2(t_1 - t_0)\Big)}{1 + \tilde{\eta}z} dz - \int_0^{\alpha} \frac{2\tilde{\eta}(t_1 - t_0)}{1 - 2\tilde{\eta}z} dz \\
&\overset{\text{(Prop. 19)}}{\ge} \int_0^{\alpha} \frac{\tilde{\eta}\Big(2\epsilon(t_1 - t_0) - \Delta(B) + 2(t_1 - t_0)\Big)}{1 + \tilde{\eta}z} dz - \int_0^{\alpha} \frac{2\tilde{\eta}(t_1 - t_0)}{1 - 2\tilde{\eta}z} dz \\
&\ge \Big(2\epsilon(t_1 - t_0) - \Delta(B) + 2(t_1 - t_0)\Big)\log(1 + \alpha\tilde{\eta}) + (t_1 - t_0)\log(1 - 2\alpha\tilde{\eta}) \\
&\ge \big(2\epsilon(t_1 - t_0) - \Delta(B)\big)\log(1 + \alpha\tilde{\eta}) + (t_1 - t_0)\log\big\{(1 + \alpha\tilde{\eta})^2 \cdot (1 - 2\alpha\tilde{\eta})\big\}.
\end{aligned}
$$

Moreover $\alpha\tilde{\eta} > \log(1 + \alpha\tilde{\eta}) \ge \frac{1}{2}\alpha\tilde{\eta}$ since $0 < \alpha\tilde{\eta} < 1$. Condition 11 guarantees that $\alpha \ll \epsilon = \big(\delta - \delta \cdot (1.05 - \delta)^{1/2}\big)/4$ and $2\alpha^3\tilde{\eta}^3 + 3\alpha^2\tilde{\eta}^2 \le 4\alpha^2 < 1/5 \wedge 2\epsilon/5$, so

$$
\begin{aligned}
\log\big\{(1 + \alpha\tilde{\eta})^2 \cdot (1 - 2\alpha\tilde{\eta})\big\} &= \log\big\{1 - 3\alpha^2\tilde{\eta}^2 - 2\alpha^3\tilde{\eta}^3\big\} \\
&= -\log(1 + \frac{3\alpha^2\tilde{\eta}^2 + 2\alpha^3\tilde{\eta}^3}{1 - 3\alpha^2\tilde{\eta}^2 - 2\alpha^3\tilde{\eta}^3}) \\
&\ge \frac{-3\alpha^2\tilde{\eta}^2 - 2\alpha^3\tilde{\eta}^3}{1 - 3\alpha^2\tilde{\eta}^2 - 2\alpha^3\tilde{\eta}^3} \\
&\ge -5\alpha^2
\end{aligned}
$$

Putting them all together, we have

$$
\sum_{t=t_0}^{t_1 - 1} \log\Big\{1 + \alpha\tilde{\eta}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t)})\big)\Big\} \ge \Big(\epsilon - 5\alpha^2\Big)(t_1 - t_0) - \Delta(B) \cdot \log(1 + \alpha\tilde{\eta})
$$

$$\geq \frac{1}{2}\alpha\tilde{\eta}\epsilon \cdot (t_1 - t_0) - \Delta(B)\alpha\tilde{\eta}$$

And consequently we have

$$\begin{aligned}\langle \mathbf{w}_{y,r}^{(t_1)}, y\mathbf{v}\rangle &= \langle \mathbf{w}_{y,r}^{(t_0)}, y\mathbf{v}\rangle \cdot \prod_{t=t_0}^{t_1-1}\{1 + \alpha\tilde{\eta}(1 - yf(\mathbf{x}; \mathbf{W}^{(t)}))\} \\ &\geq \langle \mathbf{w}_{y,r}^{(t_0)}, y\mathbf{v}\rangle \cdot \exp\Big\{\frac{\eta\|\mathbf{v}\|_2^2}{m} \cdot \epsilon t - \Delta(B)\alpha\tilde{\eta}\Big\},\end{aligned}$$

which concludes the proof. □

### F.6 Proof of Technical Results

*Proof of Proposition 18.* We want to track the sequence after it falls below $\delta$. To this end, we define two stopping times:

$$\begin{aligned}T_{(\mathbf{v}),\delta,L} &= \min\{t \geq T_{(\mathbf{v})} : m^{-1}\sum_r \sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle\big) < \delta\}; \\ T_{(\mathbf{v})}^{+,2} &= \min\{t \geq T_{(\mathbf{v}),\delta,L} : m^{-1}\sum_r \sigma\big(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle\big) \geq \delta\}.\end{aligned}$$

If $T_{(\mathbf{v}),\delta,L} = \infty$ then the proof is over. Otherwise we prove that, before $T_{(\mathbf{v})}^{+,2} \leq \infty$ (possibly equal), the sequence will never fall below $\delta/2$.

Let's take a closer look at controls over the negative parts while the weak signal remain learned. Note that for $r \in \mathcal{V}_{y,+}$, we have that

$$\begin{aligned}\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle &= \langle \mathbf{w}_{y,r}^{(0)}, y\mathbf{v}\rangle \cdot \prod_{s=0}^{t-1}\big\{1 + \frac{2\eta\|\mathbf{v}\|}{m} \cdot \big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big)\big\} \\ &\leq \langle \mathbf{w}_{y,r}^{(0)}, y\mathbf{v}\rangle \cdot \exp\Big\{\frac{2\eta\|\mathbf{v}\|_2^2}{m}\sum_{s=0}^{t-1}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big)\Big\}.\end{aligned}$$

Meanwhile, for $t \in [T_{(\mathbf{v})}, T_{(\mathbf{v}),\delta,L}]$, we have that $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle > (\beta_{\mathbf{v}}^* m\delta)^{1/2} \gg \langle \mathbf{w}_{y,r}^{(0)}, y\mathbf{v}\rangle$. Hence

$$\exp\Big\{\frac{2\eta\|\mathbf{v}\|_2^2}{m}\sum_{s=0}^{t-1}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big)\Big\} > 1, \quad t \in [T_{(\mathbf{v})}, T_{(\mathbf{v}),\delta,L} - 1]$$

and in consequence we have that $\sum_{s=0}^{t-1}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big) > 0$. On the other hand, for $r \in \mathcal{V}_{-y,-}$, $\langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{v}\rangle < 0$ and we have that

$$\begin{aligned}\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{v}\rangle &= \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{v}\rangle \cdot \prod_{s=1}\big\{1 - \frac{2\eta\|\mathbf{v}\|_2^2}{m} \cdot \big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big)\big\} \\ &\geq \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{v}\rangle \cdot \exp\Big\{-\frac{2\eta\|\mathbf{v}\|_2^2}{m} \cdot \sum_{s=0}^{t-1}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big)\Big\} \\ &\geq \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{v}\rangle \big/ \exp\Big\{\underbrace{\frac{2\eta\|\mathbf{v}\|_2^2}{m} \cdot \sum_{s=0}^{t-1}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(s)})\big)}_{>0}\Big\} \\ &\geq \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{v}\rangle \\ &\geq -\sqrt{2\log(16m/p)} \cdot \sigma_0\|\mathbf{v}\|_2, \qquad (29)\end{aligned}$$

for all $t \in [T_{(\mathbf{v})}, T_{(\mathbf{v}),\delta,L} - 1]$. Analogously, we have that $\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle \geq \langle \mathbf{w}_{-y,r}^{(0)}, -y\mathbf{u}\rangle \geq -\sqrt{2\log(16m/p)} \cdot \sigma_0 \|\mathbf{u}\|_2$, for all $t \in [T_{(\mathbf{v})}, T_{(\mathbf{v}),\delta,L} - 1]$. Therefore, we have that

$$\begin{aligned} E_t &\coloneqq \frac{1}{m} \sum_{r\in[m]} \Big\{ \sigma\big(-\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle\big) + \sigma\big(-\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{v}\rangle\big) \Big\} \\ &\leq 2 \cdot \max_{r\in[m]} \sigma\big(-\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle\big) \vee \sigma\big(-\langle \mathbf{w}_{-y,r}^{(t)}, -y\mathbf{v}\rangle\big) \\ &\leq 2(\sqrt{2\log(16m/p)} \cdot \sigma_0 \|\mathbf{u}\|_2)^2 \\ &\ll \delta/2. \end{aligned} \tag{30}$$

This allows us to leverage Proposition 19 to upper bound $yf(\mathbf{x}; \mathbf{W}^{(t)})$ for $t \in [T_{(\mathbf{v})}, T_{(\mathbf{v}),\delta,L} - 1]$. We locate the last time before $T_{(\mathbf{v}),\delta,L}$ that $yf$ just bounces up over 1, which is, $\bar{T}_{k^*} \coloneqq \max\{\bar{T}_k : \bar{T}_k \leq T_{(\mathbf{v}),\delta,L}\}$. Then Proposition 19 with Inequality (30) implies that

$$\begin{aligned} yf(\mathbf{x}; \mathbf{W}^{(\bar{T}_{k^*}-1)}) &\leq yf(\mathbf{x}; \mathbf{W}^{(\bar{T}_{k^*}-1)}) \cdot \Big\{1 + \tilde{\eta}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2 + 2E_{\bar{T}_k - 1} \\ &\leq \big\{1 + \tilde{\eta}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(\bar{T}_{k^*}-1)})\big)\big\}^2 + o(1) \\ &\leq \big\{1 + \tilde{\eta}\big(1 - \frac{1}{2\tilde{\eta}}(\tilde{\eta} + 2 - \sqrt{\tilde{\eta}^2 + 4\tilde{\eta}})\big)\big\}^2 + o(1) \\ &\leq 3. \end{aligned}$$

Here we have applied the fact that $\tilde{\eta} < 1$.

On the other hand, we have that

$$\frac{1}{m} \sum_{r\in[m]} \sigma\big(\langle \mathbf{w}_{y,r}^{(\bar{T}_{k^*}-1)}, y\mathbf{u}\rangle\big) \leq 1.05,$$

Lemma 15 indicates that $\max_{r\in[m]} \langle \mathbf{w}_{y,r}^{(\bar{T}_{k^*}-1)}, y\mathbf{u}\rangle \leq (1.05\beta_{\mathbf{u}}^* m)^{1/2}$. As Inequality (17). One step gradient descent gives that

$$\langle \mathbf{w}_{y,r}^{(\bar{T}_{k^*})}, y\mathbf{u}\rangle \leq \langle \mathbf{w}_{y,r}^{(\bar{T}_{k^*}-1)}, y\mathbf{u}\rangle \cdot \big\{1 + \frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(\bar{T}_{k^*}-1)})\big)\big\} \leq 1.5 \cdot \big(1.05\beta_{\mathbf{u}}^* m\big)^{1/2}.$$

Now we consider the scale of these inner products right at $T_{(\mathbf{v}),\delta,L}$. From the definition of $T_{(\mathbf{v}),\delta,L}$ we know that $yf > 1$ before this step, which means $1 < yf(\mathbf{x}; \mathbf{W}^{(t)}) < yf(\mathbf{x}; \mathbf{W}^{(\bar{T}_{k^*})}) \leq 3,\ t \in [\bar{T}_{k^*}, T_{(\mathbf{v}),\delta,L}]$. We first state that $m^{-1} \sum_{r\in[m]} \sigma\big(\langle \mathbf{w}_{y,r}^{(T_{(\mathbf{v}),\delta,L})}, y\mathbf{v}\rangle\big)$ is not far away from $\delta$. Note that

$$\begin{aligned} \frac{1}{m} \sum_{r\in[m]} \sigma\big(\langle \mathbf{w}_{y,r}^{(T_{(\mathbf{v}),\delta,L})}, y\mathbf{v}\rangle\big) &= \frac{1}{m} \sum_{r\in[m]} \sigma\big(\langle \mathbf{w}_{y,r}^{(T_{(\mathbf{v}),\delta,L}-1)}, y\mathbf{v}\rangle\big) \cdot \big\{1 + \\ &\qquad \frac{2\eta\|\mathbf{v}\|_2^2}{m} \cdot \big(1 - yf(\mathbf{x}; \mathbf{W}^{(T_{(\mathbf{v}),\delta,L}-1)})\big)\big\} \\ &\geq \delta \cdot \big\{1 + \frac{2\eta\|\mathbf{v}\|_2^2}{m} \cdot \big(1 - 3\big)\big\} \\ &\geq \delta \cdot (1 - 2\|\mathbf{v}\|_2^2/\|\mathbf{u}\|_2^2) \\ &\geq \frac{3}{4}\delta. \end{aligned}$$

Last inequality uses the fact that $\|\mathbf{v}\|_2^2/\|\mathbf{u}\|_2^2 \leq 1/4$. Again we can leverage Inequality (29) with $t = T_{(\mathbf{v}),\delta,L}$ to obtain that

$$\min_{r\in[m]} \langle \mathbf{w}_{-y,r}^{(T_{(\mathbf{v}),\delta,L})}, -y\mathbf{u}\rangle \geq -\sqrt{2\log(16m/p)} \cdot \sigma_0 \|\mathbf{u}\|_2; \tag{31}$$

$$\min_{r\in[m]} \langle \mathbf{w}_{-y,r}^{(T_{(\mathbf{v}),\delta,L})}, -y\mathbf{v}\rangle \geq -\sqrt{2\log(16m/p)} \cdot \sigma_0 \|\mathbf{v}\|_2. \tag{32}$$

Same to Inequality (19), we can also obtain the lower bound for $\max_{r\in[m]}\langle \mathbf{w}_{y,r}^{(T_{(\mathbf{v}),\delta,L})}, y\mathbf{u}\rangle$, which is

$$\begin{aligned}\max_{r\in[m]}\langle \mathbf{w}_{y,r}^{(T_{(\mathbf{v}),\delta,L})}, y\mathbf{u}\rangle &\geq \max_{r\in[m]}\langle \mathbf{w}_{y,r}^{(T_{(\mathbf{v}),\delta,L}-1)}, y\mathbf{u}\rangle \cdot \big\{1+\frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1-yf(\mathbf{x};\mathbf{W}^{(T_{(\mathbf{v}),\delta,L}-1)})\big)\big\}\\ &\geq 0.2\cdot\big(\beta_{\mathbf{u}}^* m\big)^{1/2}\cdot .\end{aligned}$$

For $t\in[T_{(\mathbf{v}),\delta,L}, T_{(\mathbf{v})}^{+,2}]$, we have that $m^{-1}\sum_r \sigma\big(\langle \mathbf{w}_{y,r}, y\mathbf{v}\rangle\big) < \delta$. Moreover, Inequalities (31) and (32) imply that $E_{T_{(\mathbf{v}),\delta,L}} \leq 2(\sqrt{2\log(16m/p)}\cdot\sigma_0\|\mathbf{u}\|_2)^2 \ll 1$. Additionally, we have that $\max_{r\in[m]}\langle \mathbf{w}_{y,r}^{(T_{(\mathbf{v}),\delta,L})}, y\mathbf{u}\rangle < \max_{r\in[m]}\langle \mathbf{w}_{y,r}^{(\bar{T}_{k^*})}, y\mathbf{u}\rangle < 1.5\cdot(1.05\beta_{\mathbf{u}}^* m)^{1/2}$. One can fine the next $\underline{T}_{k^*}$ and $\bar{T}_{k^*+1} > T_{(\mathbf{v}),\delta,L}$ and then apply an inductive argument exactly the same to the proof of Lemma 16 to conclude that: for $t\in t\in[T_{(\mathbf{v}),\delta,L}, T_{(\mathbf{v})}^{+,2}]$ it holds that

$$\begin{aligned}\max_{j,r}\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle &\leq 1.5\cdot(1.05\beta_{\mathbf{u}}^* m)^{1/2};\\ \max_{r}\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle &\geq 0.2\cdot(\beta_{\mathbf{u}}^* m)^{1/2};\\ \min_{r}\langle -\mathbf{w}_{-y,r}^{(t)}, -y\mathbf{u}\rangle &> -\sqrt{2\log(16m/p)}\cdot\sigma_0\|\mathbf{u}\|_2;\\ \min_{r}\langle -\mathbf{w}_{-y,r}^{(t)}, -y\mathbf{v}\rangle &> -\sqrt{2\log(16m/p)}\cdot\sigma_0\|\mathbf{u}\|_2;\\ 0\leq yf(\mathbf{x};\mathbf{W}^{(t)}) &\leq 3.\end{aligned}$$

And Lemma 17 implies that for any $t_1 \geq\geq T_{(\mathbf{v}),\delta,L}$, it holds that

$$\begin{aligned}\frac{1}{m}\sum_r \sigma\big(\langle \mathbf{w}_{y,v}^{(t_1)}, y\mathbf{v}\rangle\big) \geq& \frac{1}{m}\sum_r \sigma\big(\langle \mathbf{w}_{y,v}^{(T_{(\mathbf{v}),\delta,L})}, y\mathbf{v}\rangle\big)\cdot\exp\{-\frac{2\|\mathbf{v}\|_2^2}{\|\mathbf{u}\|_2^2}\cdot\frac{1.5\sqrt{1.05}}{\sqrt{1.05-\delta}}\}\\ &\geq \frac{1}{2}\delta.\end{aligned}$$

In conclusion, we obtained that $m^{-1}\sum_{r\in[m]}\sigma\big(\langle \mathbf{w}_{y,r}, y\mathbf{v}\rangle\big)\geq\delta/2,\ \forall t\geq T_{(\mathbf{v})}$. This finishes the proof of Proposition 18. □

*Proof of Proposition 19.* We continue with the notation $\tilde{\eta}=2\eta\|\mathbf{u}\|_2^2/m$. Define the function

$$h_{\tilde{\eta}}(z) \coloneqq \big(1+\tilde{\eta}(1-z)\big)^2\cdot z.$$

Note that $\|\mathbf{v}\|\leq\|\mathbf{u}\|$, and $2\eta\|\mathbf{u}\|_2^2/m\cdot(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)}))<2\eta\|\mathbf{u}\|_2^2/m<1$ from the definition of $\bar{T}_k-1$, we have

$$\begin{aligned}yf(\mathbf{x},y;\mathbf{W}^{(\bar{T}_k)}) =& \frac{1}{m}\sum_{r\in[m]}\sigma\big(\langle \mathbf{w}_{y,r}^{(\bar{T}_k-1)}, y\mathbf{u}\rangle\big)\cdot\Big\{1+\frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2\\ &+\frac{1}{m}\sum_{r\in[m]}\sigma\big(\langle \mathbf{w}_{y,r}^{(\bar{T}_k-1)}, y\mathbf{v}\rangle\big)\cdot\Big\{1+\frac{2\eta\|\mathbf{v}\|_2^2}{m}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2\\ &-\frac{1}{m}\sum_{r\in[m]}\sigma\big(-\langle \mathbf{w}_{-y,r}^{(\bar{T}_k-1)}, -y\mathbf{u}\rangle\big)\cdot\Big\{1-\frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2\\ &-\frac{1}{m}\sum_{r\in[m]}\sigma\big(-\langle \mathbf{w}_{-y,r}^{(\bar{T}_k-1)}, -y\mathbf{v}\rangle\big)\cdot\Big\{1-\frac{2\eta\|\mathbf{v}\|_2^2}{m}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2\\ \leq& \frac{1}{m}\sum_{r\in[m]}\Big(\sigma\big(\langle \mathbf{w}_{y,r}^{(\bar{T}_k-1)}, y\mathbf{u}\rangle\big)+\sigma\big(\langle \mathbf{w}_{y,r}^{(\bar{T}_k-1)}, y\mathbf{v}\rangle\big)\Big)\cdot\Big\{1+\frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{m}\sum_{r\in[m]}\Big(\sigma\big(-\langle \mathbf{w}_{-y,r}^{(\bar{T}_k-1)}, -y\mathbf{u}\rangle\big)+\sigma\big(-\langle \mathbf{w}_{-y,r}^{(\bar{T}_k-1)}, -y\mathbf{v}\rangle\big)\Big)\\
&\qquad\cdot\Big\{1-\frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2\\
&= yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\cdot\Big\{1+\tilde{\eta}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2+\frac{1}{m}\sum_{r\in[m]}\Big(\sigma\big(-\langle \mathbf{w}_{-y,r}^{(\bar{T}_k-1)}, -y\mathbf{u}\rangle\big)+\\
&\qquad\sigma\big(-\langle \mathbf{w}_{-y,r}^{(\bar{T}_k-1)}, -y\mathbf{v}\rangle\big)\Big)\cdot\underbrace{\Big\{2\tilde{\eta}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2}_{<2\tilde{\eta}<2}\\
&\le yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\cdot\Big\{1+\tilde{\eta}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2+\delta
\end{aligned}
\tag{33}
$$

By Assumption 12 we know that $yf(\mathbf{x};\mathbf{W}^{(\bar{T}+1)}) > 1+\delta$. Definition of $T_{(\mathbf{v})}$ along with with Inequality (33) imply that

$$
h_{\tilde{\eta}}\big(f(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big) = yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\cdot\Big\{1+\tilde{\eta}\big(1-yf(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)})\big)\Big\}^2 \ge 1+\delta-\delta = 1 \tag{34}
$$

Now we consider the equation $h_{\tilde{\eta}}-1=0$, one can easily verify that it has three roots

$$
\begin{aligned}
z_1 =& 1,\\
z_2 =& \frac{\tilde{\eta}+2-\sqrt{\tilde{\eta}^2+4\tilde{\eta}}}{2\tilde{\eta}},\\
z_3 =& \frac{\tilde{\eta}+2+\sqrt{\tilde{\eta}^2+4\tilde{\eta}}}{2\tilde{\eta}} > \frac{2\tilde{\eta}+2}{2\tilde{\eta}} > 1.
\end{aligned}
$$

And the second root $z_2 < 1$ if and only if $\tilde{\eta} > 1/2$.

Now if $0 < \tilde{\eta} \le 1/2$, then $z_1 \le z_2 < z_3$, and then $h(z) < 1$ for $z < z_1 = 1$. Therefore Inequality (34) implies that $f(\mathbf{x};\mathbf{W}^{(\bar{T}_k-1)}) \ge z_1 = 1$, and recursively implies that $f(\mathbf{x};\mathbf{W}^{(0)}) \ge 1$, which contradicts with the fact that $f(\mathbf{x},y;\mathbf{W}^{(0)}) \le 2\sqrt{2\log(16m/p)}\cdot\sigma_0\|\mathbf{u}\|_2$. So in order for $\bar{T}_1 < \infty$, we must have $\tilde{\eta} > 1/2$. In conclusion, one necessary condition for the stable oscillation Assumption 12 is that $\tilde{\eta} > 1/2$, and

$$
yf(\mathbf{x};\mathbf{W}^{(\bar{T}-1)}) \ge z_2 = \frac{\tilde{\eta}+2-\sqrt{\tilde{\eta}^2+4\tilde{\eta}}}{2\tilde{\eta}}.
$$

This finishes the proof of Proposition 19. □

### F.7 Discussion: Necessary Condition for $\delta$-Oscillation

Inequality (18) provides an upper bound involving $\tilde{\eta}$, which should be compatible with Assumption 12, hence

$$
(\tilde{\eta}/2+\sqrt{\tilde{\eta}^2/4+\tilde{\eta}})^2 > 1+\delta \Leftrightarrow \tilde{\eta} > (1+\delta^{-1})(\sqrt{1+\delta}-1).
$$

One can verify with software that RHS is monotonically increasing in $\delta\in[0,1]$ with minimal value $0.5$, when $\tilde{\eta}=0$, which is in line with the weakest oscillation condition discovered in the last part. And the maximal value taken at $\delta=1$ is less than $0.83$.

On the other hand, Inequality (16) should also be compatible with the Assumption 4, which indicates

$$
1-\delta > yf(\mathbf{x};\mathbf{W}^{(\bar{T}_{k-1})}) > \frac{2+\tilde{\eta}-\sqrt{\tilde{\eta}^2+4\tilde{\eta}}}{2\tilde{\eta}}.
$$

The readers can see that it is equivalent to $\tilde{\eta} > \delta^{-1}\big((1-\delta)^{-1/2}-1\big)$. Furthermore, we have that $\delta^{-1}\big((1-\delta)^{-1/2}-1\big) > (1+\delta^{-1})(\sqrt{1+\delta}-1)$ thus it is a stronger requirement on $\eta$.

## G One-data Case: Small Learning Rate Regime

This section focuses on training our model with one single data point $(\mathbf{x}, y)$ where $\mathbf{x} = (y\mathbf{u}, y\mathbf{v})$ contains two signal patch with $\mathbf{u}$ much stronger than $\mathbf{v}$. Therefore, the whole objective can be rearranged by

$$L(\mathbf{W}) = \frac{1}{2}\big(f(\mathbf{x}; \mathbf{W}) - y\big)^2 = \frac{1}{2}\left(\sum_{j\in\{\pm 1\}} \frac{j}{m} \sum_{r\in[m]} [\sigma(\langle \mathbf{w}_{j,r}, y\mathbf{u}\rangle) + \sigma(\langle \mathbf{w}_{j,r}, y\mathbf{v}\rangle)] - y\right)^2.$$

In this simplified setting, each weight vector is updated by

$$\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} - \frac{jy\eta}{m} \cdot \big(f(\mathbf{x}; \mathbf{W}^{(t)}) - y\big) \cdot \Big(\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{u}\rangle)\mathbf{u} + \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{v}\rangle)\mathbf{v}\Big).$$

Then we can directly obtain updating rules of the following inner products

$$\langle \mathbf{w}_{j,r}^{(t+1)}, \mathbf{u}\rangle = \langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}\rangle - \frac{jy\eta}{m} \cdot \big(f(\mathbf{x}; \mathbf{W}^{(t)}) - y\big) \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{u}\rangle)\|\mathbf{u}\|^2,$$
$$\langle \mathbf{w}_{j,r}^{(t+1)}, \mathbf{v}\rangle = \langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}\rangle - \frac{jy\eta}{m} \cdot \big(f(\mathbf{x}; \mathbf{W}^{(t)}) - y\big) \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{v}\rangle)\|\mathbf{v}\|^2,$$

since $\mathbf{u} \perp \mathbf{v}$ are assumed to be orthogonal. In this section, we would also denote the fitting residual at iteration $t$ as $\ell^{(t)} = f(\mathbf{x}; \mathbf{W}^{(t)}) - y$ for convenience.

To prepare for the following analysis, the following lemma provides a high-probability bound on the initialization $\mathbf{W}^{(0)}$.

**Lemma 21.** *Suppose that $d \geq \Omega(\log(m/p))$, $m = \Omega(\log(1/p))$. Then with probability at least $1-p$, there holds*

$$\sigma_0\|\mathbf{u}\|/2 \leq \max_{r\in[m]} \left\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\right\rangle \leq \sqrt{2\log(8m/p)}\sigma_0\|\mathbf{u}\|,$$
$$\sigma_0\|\mathbf{v}\|/2 \leq \max_{r\in[m]} \left\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\right\rangle \leq \sqrt{2\log(8m/p)}\sigma_0\|\mathbf{v}\|,$$

*for all $j \in \{\pm 1\}$.*

**Assumption 22.** *Suppose that the following holds:*

1. *The weight initialization scale $\sigma_0 = \widetilde{\Theta}(\|\mathbf{u}\|^{-1})$;*
2. *The signal strength $\|\mathbf{u}\|_2 > \widetilde{\Omega}(m^2) \cdot \|\mathbf{v}\|_2$;*
3. *The dimension $d$ satisfies $d = \Omega(\mathrm{polylog}(m))$.*

**Theorem 23** (Restatement of Proposition 1)**.** *Under Assumption 22 on $(d, m, \sigma_0, \|\mathbf{u}\|, \|\mathbf{v}\|)$, if we choose the learning rate $\eta \leq \frac{m}{6\|\mathbf{u}\|^2}$ small enough and $\epsilon \in (0, 1)$, then with high probability $1 - 1/\mathrm{poly}(d, m)$, there exist*

$$T^\dagger = \frac{m}{\eta(1-\tau)\|\mathbf{u}\|^2}\log\left(\frac{2\iota}{\sigma_0\|\mathbf{u}\|}\right), \quad T = T^\dagger + \left\lfloor \frac{Cm^3}{2\eta\epsilon\|\mathbf{u}\|^2}\right\rfloor$$

*such that: (i) average loss over iterations $[T^\dagger, T]$ decreased to $2\epsilon$, i.e. $\frac{1}{T-T^\dagger+1}\sum_{s=T^\dagger}^{T} L(\mathbf{W}^{(s)}) \leq 2\epsilon$; (ii) the model does not learn weak signal $\mathbf{v}$ well enough, compared to initialization, i.e. $\max_{j,r} |\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle| \leq \sigma_0\|\mathbf{v}\|$.*

In the small learning rate regime, the dynamics will go through two stages, in which the strong signal $\mathbf{u}$ will be firstly learned exponentially fast, and subsequently fully fit the given data point $(\mathbf{x}, y)$ therefore stabilizing the training process.

**Lemma 24.** *For any $0 < \tau < 1$ to be tuned later, suppose that at some time $t$ there holds*

$$\max_r \left\{|\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}\rangle|, \quad |\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle|\right\} \leq \sqrt{\frac{\tau}{2}},$$

*then we can lower bound the fitting residual $-y\ell^{(t)} \geq 1 - \tau$.*

*Proof.* Plug into the CNN model definition

$$-y\ell^{(t)} = 1 - F_y(\mathbf{x}; \mathbf{W}^{(t)}) + F_{-y}(\mathbf{x}; \mathbf{W}^{(t)}) \geq 1 - F_y(\mathbf{x}; \mathbf{W}^{(t)}).$$

We can upper bound $F_y(\mathbf{x}; \mathbf{W}^{(t)})$ further by

$$\begin{aligned} F_y(\mathbf{x}; \mathbf{W}^{(t)}) =& \frac{1}{m} \sum_{r\in[m]} \left[\sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle) + \sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle)\right] \\ \leq& \max_r \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle^2 + \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v}\rangle^2 \leq \tau. \end{aligned}$$

Then it follows that $-y\ell^{(t)} \geq 1 - \tau$. □

### G.1 Stage 1. Exponential Growth.

We will mainly track the maximal inner-product between $\mathbf{w}$ and signal vectors $\mathbf{v}, \mathbf{u}$,

$$\Psi^{(t)} = \max_{j,r} |\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle|, \quad \Phi^{(t)} = \max_{j,r} |\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}\rangle|.$$

In the following, we would take

$$\begin{aligned} \tau =& \max\left\{2\sigma_0\|\mathbf{u}\| \left(2\log(8m/p)\right)^{1/2-\|\mathbf{v}\|^2/4\|\mathbf{u}\|^2}, 1 - \frac{6\sqrt{2}\|\mathbf{v}\|^2}{\|\mathbf{u}\|^2 \log(1/\sqrt{2\log(8m/p)})}\right\}, \\ \iota =& \sigma_0\|\mathbf{u}\| \exp\left[\frac{1-\tau}{6}\log\left(\frac{\sqrt{\tau/2}}{\sigma_0\|\mathbf{u}\|\sqrt{2\log(8m/p)}}\right)\right]. \end{aligned}$$

By the conditions in Proposition 1 upon $\|\mathbf{v}\|^2/\|\mathbf{u}\|^2$, we find $\tau, \iota$ both constant in $(0, 1)$.

**Lemma 25.** *Under the same condition as Proposition 1, there exists time*

$$T^\dagger = \frac{m}{\eta(1-\tau)\|\mathbf{u}\|^2}\log\left(\frac{2\iota}{\sigma_0\|\mathbf{u}\|}\right),$$

*such that: (i) The model learns strong signal to a constant level* $\max_r \langle \mathbf{w}_{y,r}^{(T^\dagger)}, y\mathbf{u}\rangle \geq \iota$. *(ii) Compared to random initialization, the model does not learn weak signal that much* $\max_{j,r} |\langle \mathbf{w}_{j,r}^{(T^\dagger)}, \mathbf{v}\rangle| \leq \sigma_0\|\mathbf{v}\|$.

*Proof.* Firstly, we would find $\Psi^{(t)}, \Phi^{(t)}$ having an exponentially growing upper bound. Recursively, we would have

$$\begin{aligned} \Psi^{(t+1)} \leq& \Psi^{(t)} + \max_{j,r}\left|\frac{jy\eta}{m}\cdot\left(f(\mathbf{x}; \mathbf{W}^{(t)}) - y\right)\cdot\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{v}\rangle)\|\mathbf{v}\|^2\right| \\ =& \Psi^{(t)} + \frac{\eta}{m}\left|\ell^{(t)}\right|\|\mathbf{v}\|^2 \cdot \max_{j,r}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{v}\rangle) \\ \leq& \Psi^{(t)} + \frac{2\eta}{m}\left|\ell^{(t)}\right|\|\mathbf{v}\|^2\Psi^{(t)} \leq \exp\left(\frac{6\eta\|\mathbf{v}\|^2}{m}\right)\Psi^{(t)}. \end{aligned}$$

Therefore, $\Psi^{(t)} \leq \exp\left(\frac{6\eta\|\mathbf{v}\|^2 t}{m}\right)\Psi^{(0)} \leq \exp\left(\frac{6\eta\|\mathbf{v}\|^2 t}{m}\right)\sqrt{2\log(8m/p)}\sigma_0\|\mathbf{v}\|$. It follows similarly that

$$\Phi^{(t)} \leq \exp\left(\frac{6\eta\|\mathbf{u}\|^2 t}{m}\right)\Phi^{(0)} \leq \exp\left(\frac{6\eta\|\mathbf{u}\|^2 t}{m}\right)\sqrt{2\log(8m/p)}\sigma_0\|\mathbf{u}\|.$$

Note that growing rates of these two bounds differ a lot due to the different magnitudes of $\|\mathbf{u}\|$ and $\|\mathbf{v}\|$. Our subsequent analysis illustrates that $\Phi^{(t)}$ can grow into a constant-level magnitude since strong signal $\mathbf{u}$ is significant enough. We can track how well our model learns $\mathbf{u}$ by

$$A^{(t)} = \max_r \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle.$$

By definition, $A^{(t)} \le \Phi^{(t)}$ also admits an exponentially upper bound. For a certain $\tau \in (0,1)$, derived from the previous exponential upper bound, $\max\{\Phi^{(t)}, \Psi^{(t)}\} \le \sqrt{\tau/2}$ remains true at least until

$$T_1 = \frac{m}{6\eta\|\mathbf{u}\|^2} \log\left(\frac{\sqrt{\tau/2}}{\sigma_0\|\mathbf{u}\|\sqrt{2\log(8m/p)}}\right).$$

Consequently, until at least $T_1$, we are able to use Lemma 24 to conclude $-y\ell^{(t)} \ge 1-\tau$, which enables lower bounding $A^{(t)}$. Start with updating rule

$$\begin{aligned}\langle \mathbf{w}_{y,r}^{(t+1)}, y\mathbf{u}\rangle =& \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle + \frac{\eta}{m}\cdot\big(-y\ell^{(t)}\big)\cdot\sigma'(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle)\|\mathbf{u}\|^2 \\ \ge& \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle + \frac{2\eta(1-\tau)\|\mathbf{u}\|^2}{m}\max\{\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle, 0\},\end{aligned}$$

and take maximum over $r \in [m]$ to see

$$A^{(t+1)} \ge A^{(t)} + \frac{2\eta(1-\tau)\|\mathbf{u}\|^2}{m}A^{(t)} \ge \exp\left(\frac{\eta(1-\tau)\|\mathbf{u}\|^2}{m}\right)A^{(t)},$$

where the last equality is by $1+z \ge \exp(z/2)$ for any $0 \le z \le 2$. Consequently, we would have

$$A^{(t)} \ge \exp\left(\frac{\eta(1-\tau)\|\mathbf{u}\|^2 t}{m}\right)A^{(0)} \ge \exp\left(\frac{\eta(1-\tau)\|\mathbf{u}\|^2 t}{m}\right)\sigma_0\|\mathbf{u}\|/2,$$

at least until $t \le T_1$. Define

$$T_2 = \frac{m}{\eta(1-\tau)\|\mathbf{u}\|^2}\log\left(\frac{2\iota}{\sigma_0\|\mathbf{u}\|}\right) \le T_1$$

where the inequality is due to the scaling of $\iota$ upon $\tau$. Plugging $T_2$ into the exponential lower bound, we can conclude that $\Phi^{(T_2)} \ge A^{(T_2)} \ge \iota$ already grows up to a constant level magnitude by the time $T_2$. Lastly, plug the definition of $T_2$ to upper bound

$$\Psi^{(T_2)} \le \exp\left(\frac{6\|\mathbf{v}\|^2}{(1-\tau)\|\mathbf{u}\|^2}\log\left(\frac{2\iota}{\sigma_0\|\mathbf{u}\|}\right)\right)\sqrt{2\log(8m/p)}\sigma_0\|\mathbf{v}\| \le \sigma_0\|\mathbf{v}\|.$$

In conclusion, by taking $T^\dagger = T_2$, this lemma is completely proved. □

### G.2 Stage 2. Stabilized Convergence.

In the second stage, our lemmas would suggest that before the model really learns the weak signal $\mathbf{v}$ (i.e. before $\max_{j,r}|\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle|$ breaks the $O(\sigma_0\|\mathbf{v}\|)$ upper bound), the model already fits the given data point by exploiting strong signal $\mathbf{u}$ and decreasing the loss to $\epsilon$.

**Lemma 26.** *For any $\epsilon \in (0,1)$, there exists time*

$$T = T^\dagger + \left\lfloor \frac{Cm^3}{2\eta\epsilon\|\mathbf{u}\|^2}\right\rfloor$$

*such that: (i) Average loss over iterations within this stage has decreased to $2\epsilon$ $\frac{1}{T-T^\dagger+1}\sum_{s=T^\dagger}^{T} L(\mathbf{W}^{(s)}) \le \epsilon + \epsilon^2 \le 2\epsilon$. (ii) All through the training dynamics $0 \le t \le T$, there holds $\max_{j,r}|\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle| \le \sigma_0\|\mathbf{v}\|$.*

Firstly, we identify when the upper bound on $\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle$ breaks and find that the conclusions of Lemma 25 still holds before that time.

**Lemma 27.** *Take $\eta \le \frac{m}{6\|\mathbf{u}\|^2}$. There exists a time*

$$T^\ddagger = \frac{m}{6\eta\|\mathbf{v}\|^2}\log\left(\frac{\sqrt{\tau/2}}{\sigma_0\|\mathbf{v}\|\sqrt{2\log(8m/p)}}\right) \ge T^\dagger$$

*such that*

$$\max_r \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle \ge \iota/2, \quad \max_{j,r}|\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle| \le \sigma_0\|\mathbf{v}\|$$

*hold for any $T^\dagger \le t \le T^\ddagger$.*

*Proof.* Firstly, we need to adopt the exponential upper bound derived in proving Lemma 25,

$$\Psi^{(t)} \leq \exp\left(\frac{6\eta\|\mathbf{v}\|^2 t}{m}\right)\Psi^{(0)} \leq \exp\left(\frac{6\eta\|\mathbf{v}\|^2 t}{m}\right)\sqrt{2\log(8m/p)}\sigma_0\|\mathbf{v}\|.$$

Then we naturally find that before $T^\ddagger$, it would always hold that $\max_{j,r}|\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle| \leq \sigma_0\|\mathbf{v}\|$. Due to the conditions on $\|\mathbf{u}\|/\|\mathbf{v}\|$, $T^\ddagger$ is found to be much larger than $T^\dagger$. Then proceed by induction to prove the other assertion. At time $t = T^\dagger$, the lower bound $\max_r\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle \geq \iota/2$ holds as a consequence of the previous lemma. Suppose it holds until time $t$. Restate the updating rule by

$$\langle \mathbf{w}_{y,r}^{(t+1)}, y\mathbf{u}\rangle = \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle + \frac{\eta}{m}\cdot\left(1 - yf(\mathbf{x};\mathbf{W}^{(t)})\right)\cdot\sigma'(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle)\|\mathbf{u}\|^2,$$

from which we find $\max_r\langle \mathbf{w}_{y,r}^{(t+1)}, y\mathbf{u}\rangle \geq \max_r\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle$ must hold if $yf(\mathbf{x};\mathbf{W}^{(t)}) \leq 1$. Otherwise, once $yf(\mathbf{x};\mathbf{W}^{(t)}) > 1$, it immediately follows that

$$\begin{aligned}
1 <& yf(\mathbf{x};\mathbf{W}^{(t)}) = F_y(\mathbf{x};\mathbf{W}^{(t)}) - F_{-y}(\mathbf{x};\mathbf{W}^{(t)}) \\
\leq& F_y(\mathbf{x};\mathbf{W}^{(t)}) = \frac{1}{m}\sum_{r\in[m]}\left[\sigma(\langle \mathbf{w}_{y,r}, y\mathbf{u}\rangle) + \sigma(\langle \mathbf{w}_{y,r}, y\mathbf{v}\rangle)\right] \\
\leq& \max_r\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle^2 + \sigma_0^2\|\mathbf{v}\|^2.
\end{aligned}$$

Consequently, for the specific filter $r^* = \operatorname{argmax}_r\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle$, there holds

$$\begin{aligned}
\langle \mathbf{w}_{y,r^*}^{(t+1)}, y\mathbf{u}\rangle \geq& \langle \mathbf{w}_{y,r^*}^{(t)}, y\mathbf{u}\rangle - \frac{3\eta}{m}\langle \mathbf{w}_{y,r^*}^{(t)}, y\mathbf{u}\rangle\|\mathbf{u}\|^2 \\
\geq& \left(1 - \sigma_0^2\|\mathbf{v}\|^2\right)\left(1 - \frac{3\eta}{m}\|\mathbf{u}\|^2\right) \geq \frac{\iota}{2},
\end{aligned}$$

where the last inequality is enabled by taking $\eta \leq \frac{m}{6\|\mathbf{u}\|^2}$ and $\sigma_0 \leq \sqrt{1-\iota}/\|\mathbf{v}\|$. □

Our subsequently analysis confirms that even before $T^\ddagger$, the model can already fit the given data point by exploiting $\mathbf{u}$. For the given $0 < \epsilon < 1$, define a reference point $\mathbf{W}^*$ as

$$\mathbf{w}_{j,r}^* = \frac{4m(1+\epsilon)}{\iota}\cdot\frac{\mathbf{u}}{\|\mathbf{u}\|^2}, \quad j \in \{\pm 1\}, r \in [m].$$

**Lemma 28.** *Under the same condition as the previous lemma, for all $T^\dagger \leq t \leq T^\ddagger$, there holds $y\left\langle \nabla f(\mathbf{x};\mathbf{W}^{(t)}), \mathbf{W}^*\right\rangle \geq 2(1+\epsilon)$.*

*Proof.* Recall that $f(\mathbf{x};\mathbf{W}) = \sum_{j,r}\frac{j}{m}\left[\sigma(\langle \mathbf{w}_{j,r}, y\mathbf{u}\rangle) + \sigma(\langle \mathbf{w}_{j,r}, y\mathbf{v}\rangle)\right]$ and $\mathbf{u} \perp \mathbf{v}$, so we have

$$\begin{aligned}
y\left\langle \nabla f(\mathbf{x};\mathbf{W}^{(t)}), \mathbf{W}^*\right\rangle =& \frac{1}{m}\sum_{j,r}\sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{u}\rangle\right)\langle \mathbf{w}_{j,r}^*, y\mathbf{u}\rangle \\
=& \sum_{j,r}\sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{u}\rangle\right)\frac{4(1+\epsilon)}{\iota} \\
\geq& \max_r\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle\frac{4(1+\epsilon)}{\iota} \geq 2(1+\epsilon),
\end{aligned}$$

where the last inequality is by $\max_r\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle \geq \iota/2$ as shown by the previous lemma. □

**Lemma 29.** *Continued from the previous setting, we know for $T^\dagger \leq t \leq T^\ddagger$,*

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

*Proof.* Firstly expand the difference by

$$\begin{aligned}
&\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \\
=& 2\eta\left\langle \nabla L(\mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^*\right\rangle - \eta^2\|\nabla L(\mathbf{W}^{(t)})\|_F^2.
\end{aligned} \tag{35}$$

With only one data point, $\nabla L(\mathbf{W}^{(t)}) = \ell^{(t)} \nabla f(\mathbf{x}; \mathbf{W}^{(t)})$ admits a simplified expression, where $\ell^{(t)} = f(\mathbf{W}^{(t)}, \mathbf{x}) - y$ denotes the fitting residual. Since the neural network $f(\mathbf{W}, \mathbf{x})$ is 2-homogeneous in $\mathbf{W}$ due to the activation function $\sigma(z) = \max\{z, 0\}^2$, we can have

$$\left\langle \nabla f(\mathbf{x}; \mathbf{W}^{(t)}), \mathbf{W}^{(t)} \right\rangle = 2f(\mathbf{x}; \mathbf{W}^{(t)}).$$

Stack these observations into the first term of previous difference expansion to obtain

$$\begin{aligned}
&\left\langle \nabla L(\mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \right\rangle \\
=&\ell^{(t)} \left\langle \nabla f(\mathbf{x}; \mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \right\rangle \\
=&\ell^{(t)} \left( 2f(\mathbf{x}; \mathbf{W}^{(t)}) - \left\langle \nabla f(\mathbf{x}; \mathbf{W}^{(t)}), \mathbf{W}^* \right\rangle \right) \\
=&2\ell^{(t)} \left( f(\mathbf{x}; \mathbf{W}^{(t)}) - y \right) + \ell^{(t)} y \left( 2 - y \left\langle \nabla f(\mathbf{x}; \mathbf{W}^{(t)}), \mathbf{W}^* \right\rangle \right).
\end{aligned}$$

Note that the first term is exactly $4L(\mathbf{W}^{(t)})$. As for the second term, we need to plug in Lemma 28 to see $2 - y \left\langle \nabla f(\mathbf{x}; \mathbf{W}^{(t)}), \mathbf{W}^* \right\rangle \leq -2\epsilon < 0$, so that

$$\left| \ell^{(t)} y \left( 2 - y \left\langle \nabla f(\mathbf{x}; \mathbf{W}^{(t)}), \mathbf{W}^* \right\rangle \right) \right| \leq \frac{1}{2} \ell^{(t)2} + 2\epsilon^2 = L(\mathbf{W}^{(t)}) + 2\epsilon^2.$$

As a result, we would know $\left\langle \nabla L(\mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \right\rangle \geq 3L(\mathbf{W}^{(t)}) - 2\epsilon^2$. Next, an upper bound on the second order term $\eta^2 \|\nabla L(\mathbf{W}^{(t)})\|_F^2$ is given by

$$\begin{aligned}
\eta^2 \|\nabla L(\mathbf{W}^{(t)})\|_F^2 =& \eta^2 \ell^{(t)2} \left[ \|\mathbf{u}\|^2 \sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}, y\mathbf{u} \rangle) + \|\mathbf{v}\|^2 \sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}, y\mathbf{v} \rangle) \right] \\
\leq& O(\max\{\|\mathbf{u}\|^2, \|\mathbf{v}\|^2\}) \cdot \eta^2 L(\mathbf{W}^{(t)}),
\end{aligned}$$

since the dynamics of inner products $\langle \mathbf{w}_{j,r}, y\mathbf{u} \rangle, \langle \mathbf{w}_{j,r}, y\mathbf{v} \rangle$ are well bounded by $O(1)$. Via scaling $\eta \cdot O(\max\{\|\mathbf{u}\|^2, \|\mathbf{v}\|^2\}) \leq 1$, we would know $\eta^2 \|\nabla L(\mathbf{W}^{(t)})\|_F^2 \leq \eta L(\mathbf{W}^{(t)})$. Eventually, continued from (35), we can completely prove this lemma. □

*Proof of Lemma 26.* Continued from Lemma 29, for any $t \geq T^\dagger$,

$$\frac{1}{t - T^\dagger + 1} \sum_{s=T^\dagger}^{t} L(\mathbf{W}^{(s)}) \leq \frac{\|\mathbf{W}^{(T^\dagger)} - \mathbf{W}^*\|_F^2}{2\eta(t - T^\dagger + 1)} + \epsilon^2.$$

Before proceeding to scale time $t$, it would be helpful to decompose $\|\mathbf{W}^{(T^\dagger)} - \mathbf{W}^*\|_F^2$ and have an upper bound,

$$\begin{aligned}
&\|\mathbf{W}^{(T^\dagger)} - \mathbf{W}^*\|_F^2 \\
=& \sum_{j,r} \frac{\left\langle \mathbf{w}_{j,r}^{(T^\dagger)} - \mathbf{w}_{j,r}^*, \mathbf{u} \right\rangle^2}{\|\mathbf{u}\|^2} + \frac{\left\langle \mathbf{w}_{j,r}^{(T^\dagger)} - \mathbf{w}_{j,r}^*, \mathbf{v} \right\rangle^2}{\|\mathbf{v}\|^2} + \left\| \left( \mathbf{I}_d - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2} - \frac{\mathbf{u}\mathbf{u}^\top}{\|\mathbf{u}\|_2^2} \right) (\mathbf{w}_{j,r}^{(T^\dagger)} - \mathbf{w}_{j,r}^*) \right\|^2 \\
\leq& \sum_{j,r} \frac{2\left\langle \mathbf{w}_{j,r}^{(T^\dagger)}, \mathbf{u} \right\rangle^2 + 2\left\langle \mathbf{w}_{j,r}^*, \mathbf{u} \right\rangle^2}{\|\mathbf{u}\|^2} + \frac{\left\langle \mathbf{w}_{j,r}^{(T^\dagger)}, \mathbf{v} \right\rangle^2}{\|\mathbf{v}\|^2} + \left\| \left( \mathbf{I}_d - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2} - \frac{\mathbf{u}\mathbf{u}^\top}{\|\mathbf{u}\|_2^2} \right) \mathbf{w}_{j,r}^{(0)} \right\|^2,
\end{aligned}$$

where we exploit the fact that $\mathbf{w}^*$ is parallel to $\mathbf{u}$, and the gradient steps only updates $\mathbf{w}$ along the directions of $\mathbf{u}, \mathbf{v}$. Recall that $\langle \mathbf{w}^{(T^\dagger)}, \mathbf{u} \rangle = \Omega(1)$, $\langle \mathbf{w}^{(T^\dagger)}, \mathbf{v} \rangle = O(\sigma_0 \|\mathbf{v}\|)$, $\|\mathbf{w}_{j,r}^{(0)}\| = O(\sigma_0 \sqrt{d})$, the leading term would be $\frac{\left\langle \mathbf{w}_{j,r}^*, \mathbf{u} \right\rangle^2}{\|\mathbf{u}\|^2}$. Therefore, we would conclude that $\|\mathbf{W}^{(T^\dagger)} - \mathbf{W}^*\|_F^2 \leq Cm^3 / \|\mathbf{u}\|^2$. As a result, average loss after iterations $T^\dagger$ can be bounded by

$$\frac{1}{t - T^\dagger + 1} \sum_{s=T^\dagger}^{t} L(\mathbf{W}^{(s)}) \leq \frac{Cm^3}{2\eta \|\mathbf{u}\|^2 (t - T^\dagger + 1)} + \epsilon^2.$$

Then choose $T = T^\dagger + \left\lfloor \frac{Cm^3}{2\eta\epsilon\|\mathbf{u}\|^2} \right\rfloor$. Since $\frac{\|\mathbf{u}\|^2}{\|\mathbf{v}\|^2} \geq \widetilde{\Omega}(m^2)$, we can verify that $T \leq T^\ddagger$ where $T^\ddagger$ is given in Lemma 27 until when the weak signal cannot be fully learned. In conclusion, the final output would be $\frac{1}{T-T^\dagger+1}\sum_{s=T^\dagger}^{T} L(\mathbf{W}^{(s)}) \leq \epsilon + \epsilon^2 \leq 2\epsilon$. □

Combine Lemmas 25 and 26 to obtain the full version of Theorem 23.

# H Proofs for Main Theoretical Results

## H.1 Preliminary Analysis

Recall that $\mathcal{W}$ is the index set of training data points which lack the strong feature patch. By (3), the CNN weights are updated according to

$$\begin{aligned}\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} - \frac{j\eta}{m}\cdot\big(f(\mathbf{x}_{i_t};\mathbf{W}^{(t)}) - y_{i_t}\big)\cdot\Big(&\sigma'(\langle\mathbf{w}_{j,r}^{(t)}, y_{i_t}\mathbf{u}\rangle)\cdot y_{i_t}\mathbf{u}\cdot\mathbf{1}\{i_t\notin\mathcal{W}\}\\ &+\sigma'(\langle\mathbf{w}_{j,r}^{(t)}, y_{i_t}\mathbf{v}\rangle)\cdot y_{i_t}\mathbf{v} + \sigma'(\langle\mathbf{w}_{j,r}^{(t)}\cdot\boldsymbol{\xi}_{i_t}\rangle)\cdot\boldsymbol{\xi}_{i_t}\\ &+\sigma'(\langle\mathbf{w}_{j,r}^{(t)}\cdot\widetilde{\boldsymbol{\xi}}_{i_t}\rangle)\cdot\widetilde{\boldsymbol{\xi}}_{i_t},\cdot\mathbf{1}\{i_t\in\mathcal{W}\}.\Big)\end{aligned} \tag{36}$$

Also, recall that the correct index sets for strong and weak signal patch are defined as

$$\mathcal{U}_{j,+}^{(t)} := \Big\{r\in[m] : \langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle \geq 0\Big\}, \quad \mathcal{V}_{j,+}^{(t)} := \Big\{r\in[m] : \langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle \geq 0\Big\},$$

By the CNN expression (1), for each $j\in\{\pm1\}$, the inner products that matter are

*(i) positive neurons:* $\langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle$ for $r\in\mathcal{U}_{j,+}^{(t)}$, $\langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle$ for $r\in\mathcal{V}_{j,+}^{(t)}$, and $\langle\mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle$.

*(ii) negative neurons:* $\langle\mathbf{w}_{-j,r}^{(t)}, j\mathbf{u}\rangle$ for $r\notin\mathcal{U}_{-j,+}^{(t)}$, $\langle\mathbf{w}_{-j,r}^{(t)}, j\mathbf{v}\rangle$ for $r\notin\mathcal{V}_{-j,+}^{(t)}$, and $\langle\mathbf{w}_{-j,r}^{(t)}, \boldsymbol{\xi}_i\rangle$.

By (36), the update formula of these inner products of interests are given by

$$\begin{aligned}\langle\mathbf{w}_{j,r}^{(t+1)}, j\mathbf{u}\rangle &= \langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle\\ &\quad+\frac{\eta\|\mathbf{u}\|_2^2}{m}\cdot\big(1-y_{i_t}f(\mathbf{x}_{i_t};\mathbf{W}^{(t)})\big)\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle jy_{i_t})\cdot\mathbf{1}\{i_t\notin\mathcal{W}\},\end{aligned} \tag{37}$$

$$\langle\mathbf{w}_{j,r}^{(t+1)}, j\mathbf{v}\rangle = \langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle + \frac{\eta\|\mathbf{v}\|_2^2}{m}\cdot\big(1-y_{i_t}f(\mathbf{x}_{i_t};\mathbf{W}^{(t)})\big)\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(t)}\cdot j\mathbf{v}\rangle jy_{i_t}), \tag{38}$$

$$\begin{aligned}\langle\mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i\rangle = \langle\mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle + \frac{\eta\cdot jy_{i_t}}{m}\cdot\big(1-y_{i_t}f(\mathbf{x}_{i_t};\mathbf{W}^{(t)})\big)\cdot\Big(&\sigma'(\langle\mathbf{w}_{j,r}^{(t)}\cdot\boldsymbol{\xi}_{i_t}\rangle)\cdot\langle\boldsymbol{\xi}_{i_t}, \boldsymbol{\xi}_i\rangle\\ &+\sigma'(\langle\mathbf{w}_{j,r}^{(t)}\cdot\widetilde{\boldsymbol{\xi}}_{i_t}\rangle)\cdot\langle\widetilde{\boldsymbol{\xi}}_{i_t}, \boldsymbol{\xi}_i\rangle\cdot\mathbf{1}\{i_t\in\mathcal{W}\}\Big).\end{aligned} \tag{39}$$

## H.2 Overview of Analysis

The road map towards proving our main theorem follows the same logic for dealing with the single data setup (Appendix F). Basically, as long as the weak signal component and the noise component are not learned to a large enough scale, we can prove that the strong signal part would dominate and the oscillation would accumulate at a linear rate. This further gives Lemma 31 which shows the CNN would learn effectively learn the weak signal $\langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle$. Meanwhile, we can prove that the influences from the negative part of weak signal learning $\langle\mathbf{w}_{-j,r}^{(t)}, j\mathbf{v}\rangle$ and the noise patch $\langle\mathbf{w}_{\pm1,r}^{(t)}, \boldsymbol{\xi}_i\rangle$ can be well controlled (Propositions 32). Putting all together, we can prove the main result Theorem 5.

To be formal, we define two important stopping times as follows:

$$T_{(\mathbf{v})}^{j} = \inf\left\{t : \frac{1}{m}\sum_{r\in[m]}\sigma(\langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle) \geq \delta/2\right\},$$

$$T_{(\boldsymbol{\xi})} = \inf\left\{t : \max_{r\in[m],j\in\{\pm1\}}\left\{\max_{i\in[n]}\left|\langle\mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle\right|, \max_{i\in\mathcal{W}}\left|\langle\mathbf{w}_{j,r}^{(t)}, \widetilde{\boldsymbol{\xi}}_i\rangle\right|\right\} \geq \delta/4\right\}.$$

We recap that $\mathcal{W}$ denotes the index set of weak data, and $\widetilde{\boldsymbol{\xi}}_i$ denotes the Gaussian noise appearing on the lacking strong signal patch for those weak data. Also we note that $T^j_{\mathbf{v}}, T_{\boldsymbol{\xi}} \leq +\infty$, where the equal sign is attainable. We then define $T^j_{\max} = \min\{T^j_{(\mathbf{v})}, T_{(\boldsymbol{\xi})}\}$.

On the first place, following the same arguments as in the single data setup (Appendix F), we have the following boundedness and sign stability results.

**Lemma 30** (Boundedness and sign stability)**.** *Under Assumption 3 and 4, for $j \in \{\pm 1\}$ and $0 \leq t \leq T^j_{\max}$, the followings hold with probability at least $1 - 1/\mathrm{poly}(d)$:*

1. *it holds that $\mathcal{U}^{(t)}_{j,+} = \mathcal{U}^{(0)}_{j,+} \neq \emptyset$ and $\mathcal{V}^{(t)}_{j,+} = \mathcal{V}^{(0)}_{j,+} \neq \emptyset$ for any $t \in [0, T^j_{\max}]$. Hence the superscript $(t)$ and $(0)$ can be dropped;*
2. *for any $t \in [0, T^j_{\max}]$, we have that*

$$\max_{r\in[m]} \langle \mathbf{w}^{(t)}_{j,r}, j\mathbf{u}\rangle \leq 1.5 \cdot (1.05\beta^*_{\mathbf{u},j} m)^{1/2}, \tag{40}$$

   *where $\beta^*_{\mathbf{u},j}$ is defined in Lemma 15;*

3. *for any $t \in [0, T^j_{\max}]$ it holds that*

$$\min_{r\in[m]} \langle \mathbf{w}^{(t)}_{-j,r}, -j\mathbf{u}\rangle \geq -\sqrt{2\log(16m/p)} \cdot \sigma_0 \|\mathbf{u}\|_2, \tag{41}$$

$$\min_{r\in[m]} \langle \mathbf{w}^{(t)}_{-j,r}, -j\mathbf{v}\rangle \geq -\sqrt{2\log(16m/p)} \cdot \sigma_0 \|\mathbf{v}\|_2; \tag{42}$$

4. *for any $t \in [0, T^j_{\max}]$ such that $y_{i_t} = j$, it holds that*

$$\big|1 - y_{i_t} f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)})\big| \leq 2.$$

*Proof of Lemma 30.* See Section H.4. □

**Lemma 31** (Weak signal learning)**.** *Under Assumptions 3 and 4, with probability at least $1 - 1/\mathrm{poly}(d)$, it holds that for any $j \in \{\pm 1\}$ and $0 \leq t_1 \leq t_2 \leq T^j_{\max}$ that*

$$\sum_{\substack{s=t_1\\ y_{i_s}=j}}^{t_2} 1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)}) \geq \frac{\delta}{16} \cdot \big(1 - (1.05 - \delta/4)^{\frac{1}{2}}\big) \cdot (t_2 - t_1 + 1) - \frac{m(1.05)^{\frac{1}{2}}}{2\eta\|\mathbf{u}\|_2^2 (1.05 - \delta/4)^{\frac{1}{2}}},$$

*where $\delta$ is specified in Assumption 4. Moreover, for $r \in \mathcal{V}^{(0)}_{j,+}$, we have that*

$$\langle \mathbf{w}^{(t+1)}_{j,r}, j\mathbf{v}\rangle \geq \langle \mathbf{w}^{(t_0)}_{j,r}, j\mathbf{v}\rangle \cdot \exp\left\{ \frac{\eta\|\mathbf{v}\|_2^2}{32m} \cdot \big(\delta - \delta(1.05 - \delta/4)^{\frac{1}{2}}\big) \cdot (t - t_0 + 1) - \frac{\|\mathbf{v}\|_2^2 (1.05)^{\frac{1}{2}}}{\|\mathbf{u}\|_2^2 (1.05 - \delta/4)^{\frac{1}{2}}} \right\}. \tag{43}$$

*Proof of Lemma 31.* See Appendix H.5 for a detailed proof. □

For simplicity, we define the maximum absolute value of the noise inner products over data as

$$\Upsilon^{(t)} = \max_{r\in[m], j\in\{\pm 1\}} \left\{ \max_{i\in[n]} \Big|\langle \mathbf{w}^{(t)}_{j,r}, \boldsymbol{\xi}_i\rangle\Big|, \max_{i\in\mathcal{W}} \Big|\langle \mathbf{w}^{(t)}_{j,r}, \widetilde{\boldsymbol{\xi}}_i\rangle\Big| \right\}$$

**Proposition 32** (Noise memorization)**.** *Under Assumptions 3 and 4, then with probability at least $1 - 1/\mathrm{poly}(d)$, it holds for any $t_0 \leq t_1 \leq T^j_{\max} - \widetilde{T}_{(\boldsymbol{\xi})}$, and $j \in \{\pm 1\}$ that*

$$\Upsilon^{(t)} \leq \Upsilon^{(t_1)} \cdot (1 + \epsilon), \quad \forall r \in [m], \quad \forall t_1 \leq t \leq t_1 + \widetilde{T}_{(\boldsymbol{\xi})}$$

*where $\widetilde{T}_{(\boldsymbol{\xi})} = \widetilde{\Theta}(mn \cdot \eta^{-1} \cdot \epsilon \cdot (1+\epsilon)^{-1} \cdot (\sigma_p^2 d)^{-1})$.*

*Proof of Proposition 32.* See Appendix H.6 for a detailed proof. □

### H.3 Proof of Theorem 5

With Lemma 31 and Proposition 32, we are ready to prove Theorem 5.

*Proof of Theorem 5.* For fixed $j$, we prove that $T_{\max}^j$ is bounded by a finite number by contradiction. Specifically, we prove that

$$T_{\max}^j \le T_{j,0} \coloneqq \frac{32m}{\eta\|\mathbf{v}\|_2^2}\cdot(\delta-\delta(1.05-\delta/4)^{1/2})^{-1}\cdot\left\{\log\frac{2\sqrt{m\delta}}{\sigma_0\|\mathbf{v}\|_2}+1.5\frac{\|\mathbf{v}\|_2^2}{\|\mathbf{u}\|_2^2}\cdot\sqrt{\frac{1.05}{1.05-\delta/4}}\right\}$$

Suppose that the result fails, then $T_{\max}^j = T_{(\mathbf{v})}^j \wedge T_{(\boldsymbol{\xi})} \ge T_{j,0}$. Then Lemma 30 and Lemma 31 hold on $[0, T_{(\mathbf{v})}^j]$. By applying the lower bound in Inequality (43) as well as Lemma (10), we have that

$$\begin{aligned}
&\max_{r\in[m]}\langle\mathbf{w}_{j,r}^{(T_{j,0})}, j\mathbf{v}\rangle \\
&\quad\ge \langle\mathbf{w}_{j,r}^{(0)}, j\mathbf{v}\rangle\cdot\exp\left\{\frac{\eta\|\mathbf{v}\|_2^2}{32m}\cdot\big(\delta-\delta(1.05-\delta/4)^{\frac{1}{2}}\big)\cdot T_{j,0}-1.5\frac{\|\mathbf{v}\|_2^2}{\|\mathbf{u}\|_2^2}\cdot\sqrt{\frac{1.05}{1.05-\delta/4}}\right\} \\
&\quad\ge \frac{1}{2}\sigma_0\|\mathbf{v}\|\cdot\frac{\sqrt{m\delta}}{\sigma_0\|\mathbf{v}\|_2/2}=\sqrt{m\delta}.
\end{aligned}$$

This leads to $m^{-1}\sum_r\sigma(\langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle)\ge\delta$, which clearly contradicts with the definition of $T_{\max}^j = T_{(\boldsymbol{\xi})}\wedge T_{(\mathbf{v})}^j$, hence it must be that $T_{\max}^j \le T_{j,0}$. In the sequel, we prove that $T_{\max}^j = T_{(\mathbf{v})}^j$, for which our conclusion directly follows. Again we prove by contradiction. If $T_{\max}^j = T_{(\boldsymbol{\xi})}$, then $\Upsilon^{(t)}$ would reach $\delta/4$ for time less than $T_{j,0}$. But by Proposition 32 with $\epsilon=1$, we know that it takes at least

$$K\widetilde{T}_{(\boldsymbol{\xi})} \coloneqq \widetilde{\Theta}\left(\frac{mn}{\eta}\sigma_p^2 d\cdot\log\frac{\delta}{\sigma_0\sigma_p\sqrt{d}}\right)$$

steps to reach $\delta/4$. By Assumption 3, we can also find that $K\widetilde{T}_{(\boldsymbol{\xi})}\ge T_{j,0}$, which is a contradiction. Therefore, we have that $T_{\max}^j = T_{(\mathbf{v})}^j$, which means that at $t^\star = T_{\max}^j$,

$$\frac{1}{m}\sum_{r\in[m]}\sigma(\langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle)\ge\frac{\delta}{2}.$$

In the meanwhile, Proposition 30 guarantees that at time $t^\star = T_{\max}^j$,

$$\frac{1}{m}\sum_{r\in[m]}\sigma(\langle\mathbf{w}_{-j,r}^{(t)}, j\mathbf{v}\rangle)\le\frac{\delta}{4}.$$

Thus, we conclude that at time $t^\star$,

$$\frac{1}{m}\sum_{r\in[m]}\sigma(\langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle)-\frac{1}{m}\sum_{r\in[m]}\sigma(\langle\mathbf{w}_{-j,r}^{(t)}, j\mathbf{v}\rangle)\ge\frac{\delta}{4}.$$

This finishes the proof of Theorem 5. □

### H.4 Proof of Lemma 30

*Proof of Lemma 30.* Recall that we defined two times:

$$\begin{aligned}
T_{(\boldsymbol{\xi})} &= \min\left\{t\ge 0: \max_{j,r,i\in[n]}|\langle\mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle| \vee \max_{j,r,i\in\mathcal{W}}|\langle\mathbf{w}_{j,r}^{(t)}, \widetilde{\boldsymbol{\xi}}_i\rangle|\ge\delta/4\right\}; \\
T_{(\mathbf{v})}^j &= \min\left\{t\ge 0: \frac{1}{m}\sum_{r\in[m]}\sigma\big(\langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle\big)\ge\delta/2\right\},
\end{aligned}$$

We define some notations to simplify our presentation. Let $\tilde{f}_t = y_{i_t} f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)})$. For $s \geq 0$ define time

$$t_j(s) = \min\{t \in \mathbb{N} : t > t_j(s-1),\ y_{i_t} = j\}.$$

Where $t_j(0) = \min\{t \in \mathbb{N} : y_{i_t} = j\}$. Let $\tilde{\eta} \coloneqq 2\eta\|\mathbf{u}\|_2^2/m$. $\alpha = \|\mathbf{v}\|_2^2/\|\mathbf{u}\|_2^2$. Then, for fixed $j$, we define

$$\bar{S}_{j,k} \coloneqq \min\{s \in \mathbb{N} : s > \bar{S}_{j,k-1} \text{ such that } \tilde{f}_{t_j(s)} \geq 1 \text{ and } \tilde{f}_{t_j(\max\{s' < s : i_{t_j(s')} \notin \mathcal{W}\})} < 1\}.$$

and

$$\underline{S}_{j,k} \coloneqq \min\{s \in \mathbb{N} : s > \underline{S}_{j,k-1} \text{ such that } i_{t_j(s)} \notin \mathcal{W},\ \tilde{f}_{t_j(s)} < 1 \text{ and } \tilde{f}_{t_j(\max\{s' < s : i_{t_j(s')} \notin \mathcal{W}\})} \geq 1\}.$$

Moreover we define $\mathcal{U}_{j,+} \coloneqq \{r \in [m] : \langle \mathbf{w}_{j,r}^{(t_j(0))}, j\mathbf{v}\rangle > 0\}$ and $\mathcal{U}_{j,-} \coloneqq \{r \in [m] : \langle \mathbf{w}_{j,r}^{(t_j(0))}, j\mathbf{v}\rangle \leq 0\}$. $\mathcal{V}_{j,\pm}$ are defined analogously.

The following analysis is nearly the same to the proof of one data case in the sense that the steps are organized exactly the same to the proof of one data case. Therefore we suggest readers to refer to the roadmap provided in Section F.1 frequently for better understanding.

The following proposition indicates that we can separate neurons into two parts, with each individual part learning one kinds of sample independently.

**Proposition 33.** *If the sign stability holds locally for all the inner products, it holds that*

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t_j(s+1))}, j\mathbf{u}\rangle &= \langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u}\rangle\{1 + \tilde{\eta}(1 - \tilde{f}_{t_j(s)}) \cdot \mathbf{1}_{i_{t_j(s)} \in \mathcal{W}}\}, \quad r \in \mathcal{U}_{j,+}; \\
\langle \mathbf{w}_{-j,r}^{(t_j(s+1))}, -j\mathbf{u}\rangle &= \langle \mathbf{w}_{-j,r}^{(t_j(s))}, -j\mathbf{u}\rangle\{1 - \tilde{\eta}(1 - \tilde{f}_{t_j(s)}) \cdot \mathbf{1}_{i_{t_j(s)} \in \mathcal{W}}\}, \quad r \in \mathcal{U}_{-j,-};
\end{aligned} \tag{44}$$

*Proof of Proposition 33.* See Section H.6. □

We note that for $t \leq T_{(\mathbf{v})}^j \wedge T_{(\boldsymbol{\xi})}$, which is the time scale we're mainly focusing on, it holds that

$$\begin{aligned}
&\max_{j,r,i} |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle| \vee |\langle \mathbf{w}_{j,r}^{(t)}, \widetilde{\boldsymbol{\xi}}_i\rangle| \leq \frac{\delta}{4}, \\
&\frac{1}{m}\sum_{r \in [m]} \sigma\big(\langle \mathbf{w}_{j,r}, j\mathbf{v}\rangle\big) < \delta/2.
\end{aligned}$$

Now we begin the proof of boundedness and sign stability with step-by-step analysis. The $j$ is arbitrary but fixed throughout analysis.

**Step 1: Pre-$\bar{S}_{j,1}$ Analysis.** Clearly at $s = 0$, the lower bounds in Inequality (41) and (41) are guaranteed with Lemma 10. And we know that $\tilde{f}_{t_j(0)} \ll 1$ so $\bar{S}_{j,1} \geq 1$. The initialization also guarantees that the upper bound in Inequality (40) holds at $s = 0$.

For $s \in [0, \bar{S}_{j,k-1})$, the definition of $\bar{S}_{j,k-1}$ indicates that $\tilde{f}_{t_j(s)} \leq 1$. Therefore, from Proposition 33 we can see that the $\langle \mathbf{w}_{j,r}^{t_j(s)}, j\mathbf{u}\rangle,\ r \in \mathcal{U}_{j,+}$ and $\langle \mathbf{w}_{-j,r}^{t_j(s)}, -j\mathbf{u}\rangle,\ r \in \mathcal{U}_{-j,-}$ are non-decreasing in $s$ during this stage. One naturally infers that

$$\begin{aligned}
\langle \mathbf{w}_{-j,r}^{(t_j(\bar{S}_{j,1}))}, -j\mathbf{u}\rangle &\geq \langle \mathbf{w}_{-j,r}^{(t_j(s))}, -j\mathbf{u}\rangle \\
&\geq \langle \mathbf{w}_{-j,r}^{(t_j(0))}, -j\mathbf{u}\rangle \\
&\geq -\sqrt{2\log(16m/p)} \cdot \sigma_0\|\mathbf{u}\|_2, \quad \forall r \in \mathcal{U}_{-j,-}.
\end{aligned}$$

Same can be verified for $\langle \mathbf{w}_{-j,r}^{(t_j(s))}, -j\mathbf{v}\rangle$ with $r \in \mathcal{V}_{-j,-}$. Hence the lower bounds in Inequality (41) and (42) are successfully extended to $s \in [0, \bar{S}_{j,k}]$. Also, for these inner products, the sign stability

holds on $[0, \bar{S}_{j,1}]$, since $\tilde{\eta} < 4/5$ and we have

$$1 \pm \frac{2\eta\|\mathbf{u}\|_2^2}{m} \cdot (1 - \tilde{f}_{t_j(s)}) \geq 1 - \tilde{\eta}\big(1 - o(1)\big) \geq 0,$$
$$1 \pm \frac{2\eta\|\mathbf{v}\|_2^2}{m} \cdot (1 - \tilde{f}_{t_j(s)}) \geq 1 - \alpha\tilde{\eta}\big(1 - o(1)\big) \geq 0.$$

The we turn to upper bound $\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u}\rangle$ for $s \in [0, \bar{S}_{j,1})$. Note that, for $s$ such that $i_{t_j(s)} \notin \mathcal{W}$, the definition of $T_{\max}^j$ implies that

$$\begin{aligned}
1 - \delta \geq \tilde{f}_{t_j(s)} \geq \frac{1}{m}\sum_{r\in[m]} \sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u}\rangle\big) + \frac{1}{m}\sum_{r\in[m]} \sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{v}\rangle\big) \\
- \frac{1}{m}\sum_{r\in[m]} \sigma\big(-\langle \mathbf{w}_{-j,r}^{(t_j(s))}, -j\mathbf{u}\rangle\big) - \frac{1}{m}\sum_{r\in[m]} \sigma\big(-\langle \mathbf{w}_{-j,r}^{(t_j(s))}, -j\mathbf{v}\rangle\big) \\
- \Big|\frac{1}{m}\sum_{r\in[m]} \sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(s))}, \boldsymbol{\xi}_i\rangle\big) + \frac{1}{m}\sum_{r\in[m]} \sigma\big(\langle \mathbf{w}_{-j,r}^{(t_j(s))}, \boldsymbol{\xi}_i\rangle\big)\Big| \\
\geq \frac{1}{m}\sum_{r\in[m]} \sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u}\rangle\big) - 2 \times o(1) - 2 \times \delta/4.
\end{aligned}$$

Therefore, $m^{-1}\sum_{r\in[m]} \sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u}\rangle\big) \leq 1 - \delta/2 + o(1) \leq 1.05$. Thanks to the local sign stability, Lemma 15 implies that $\max_{r\in[m]}\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u}\rangle \leq (1.05\beta_{\mathbf{u},j}^* m)^{1/2}$. If otherwise $s \in [0, \bar{S}_{j,k-1})$ follows $i_{t_j(s)} \in \mathcal{W}$, then we choose $\tilde{s} = \max\{s' \leq s : i_{t_j(s)} \notin \mathcal{W}\}$

$$\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u}\rangle = \langle \mathbf{w}_{j,r}^{(t_j(s'))}, j\mathbf{u}\rangle \leq (1.05\beta_{\mathbf{u},j}^* m)^{1/2}.$$

Thus the upper bound is done for $s \in [0, \bar{S}_{j,1}]$.

**Step 2.1: Bounding $\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u}\rangle$ with $s \in [\bar{S}_{j,1}, \underline{S}_{j,1})$.** It's easy to verify that $i_{t_j(\bar{S}_{j,k}-1)} \notin \mathcal{W}$ and $i_{t_j(\bar{S}_{j,k})} \notin \mathcal{W}$ for $t_j(\bar{S}_{j,k} - 1) \leq T_{(\boldsymbol{v})}$, because otherwise the increment on $\langle \mathbf{w}_{j,r}, j\mathbf{v}\rangle$ at this step cannot force $\tilde{f}$ to bounce above $1 + \delta$, in case that $m^{-1}\sum\sigma(\langle \mathbf{w}_{j,r}, j\mathbf{v}\rangle) < \delta/2$. We begin with a proposition that is parallel to Proposition 19.

**Proposition 34.** *Without loss of generality, we can assume that $i_{t_j(\bar{S}_{j,k}-1)} \notin \mathcal{W}$. Otherwise we can find the last step before $\bar{S}_{j,k}$ such that $i_{t_j(s)} \notin \mathcal{W}$. Suppose that*

$$E_{t_j(\bar{S}_{j,k}-1)} \coloneqq \frac{1}{m}\sum_{r\in[m]} \Big\{\sigma\big(-\langle \mathbf{w}_{-j,r}^{(t_j(\bar{S}_{j,k}-1))}, -j\mathbf{u}\rangle\big) + \sigma\big(-\langle \mathbf{w}_{j,r}^{(t_j(\bar{S}_{j,k}-1))}, -j\mathbf{v}\rangle\big)\Big\} \leq \delta/4,$$
$$\Upsilon_{t_j(\bar{S}_{j,k}-1)} \coloneqq \max_{j,r,i\in[n]} |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle| \vee \max_{j,r,i\in\mathcal{W}} |\langle \mathbf{w}_{j,r}^{(t)}, \widetilde{\boldsymbol{\xi}}_i\rangle| \leq \delta/4$$

*then we have that*

$$\tilde{f}_{t_j(\bar{S}_{j,k}-1)} \geq \frac{\tilde{\eta} + 2 - \sqrt{\tilde{\eta}^2 + 4\tilde{\eta}}}{2\tilde{\eta}}. \tag{45}$$

*and that*

$$\begin{aligned}
\tilde{f}_{t_j(\bar{S}_{j,k})} &\leq \Big\{1 - \tilde{\eta}(1 - \tilde{f}_{t_j(\bar{S}_{j,k}-1)})\Big\}^2 + 2E_{t_j(\bar{S}_{j,k}-1)} + 2\Upsilon_{t_j(\bar{S}_{j,k}-1)} \\
&\leq \big(\tilde{\eta}/2 + \sqrt{\tilde{\eta}^2/4 + \tilde{\eta}}\big)^2 + 2E_{t_j(\bar{S}_{j,k}-1)} + 2\Upsilon_{t_j(\bar{S}_{j,k}-1)}.
\end{aligned}$$

*Proof.* See Section H.6. □

With this proposition, we can derive an upper bound on $\langle \mathbf{w}_{j,r}^{(t_j(\bar{S}_{j,1}))}, j\mathbf{u}\rangle$. One step gradient with Equation (44) implies that for $r \in \mathcal{U}_{j,+}$ we have

$$\begin{aligned}\langle \mathbf{w}_{j,r}^{\left(t_j(\bar{S}_{j,1})\right)}, j\mathbf{u}\rangle &= \langle \mathbf{w}_{j,r}^{\left(t_j(\bar{S}_{j,1})-1\right)}, j\mathbf{u}\rangle \cdot \{1+\tilde{\eta}(1-\tilde{f}_{t_j(\bar{S}_{j,1})-1})\} \\ &\leq \langle \mathbf{w}_{j,r}^{\left(t_j(\bar{S}_{j,1})-1\right)}, j\mathbf{u}\rangle \cdot \big\{1-\tilde{\eta}\big(1-\frac{\tilde{\eta}+2-\sqrt{\tilde{\eta}^2+4\tilde{\eta}}}{2\tilde{\eta}}\big)\big\} \\ &\leq 1.5\cdot(1.05\beta_{\mathbf{u},j}^* m)^{1/2}.\end{aligned}$$

Here the first inequality comes from Inequality (45) and the second inequality comes from monotonicity and taking $\tilde{\eta}=1/2$.

This upper bound continues to hold for $\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u}\rangle$ with $s \in [\bar{S}_{j,1}, \underline{S}_{j,1})$ because of monotonicity. We consider the lower bound for $\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u}\rangle$ with $s \in [\bar{S}_{j,1}, \underline{S}_{j,1})$ and $i_{t_j(s)} \notin \mathcal{W}$. For these $s$, we have that

$$\begin{aligned}1+\delta < \tilde{f}_{t_j(s)} &\leq \frac{1}{m}\sum_{r\in[m]}\Big\{\sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u}\rangle\big)\Big\} + \frac{1}{m}\sum_{r\in[m]}\Big\{\sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{v}\rangle\big)\Big\} \\ &\quad + (\text{negative part}) + (\text{noise part}) \\ &\leq \frac{1}{m}\sum_{r\in[m]}\Big\{\sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u}\rangle\big)\Big\} + \delta/2 + 2\times\delta/4.\end{aligned}$$

Combining with Lemma 15, we obtain $\max_{r\in[m]}\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u}\rangle \geq (\beta_{\mathbf{u},j}^* m)^{1/2}$, for $s \in [\bar{S}_{j,1}, \underline{S}_{j,1})$ and $i_{t_j(s)} \notin \mathcal{W}$.

**Step 2.2: Lower Bounding $\max_r \langle \mathbf{w}_{y,r}^{(t_j(\underline{S}_{j,k}))}, y\mathbf{u}\rangle$.** From one-step gradient descent with Equation (44) we know that for $r \in \mathcal{U}_{j,+}$, it holds that

$$\begin{aligned}\langle \mathbf{w}_{j,r}^{\left(t_j(\underline{S}_{j,1})\right)}, j\mathbf{u}\rangle &= \langle \mathbf{w}_{j,r}^{\left(t_j(\underline{S}_{j,1}-1)\right)}, j\mathbf{u}\rangle \cdot \{1+\tilde{\eta}(1-\tilde{f}_{t_j(\underline{S}_{j,1}-1)})\} \\ &\geq \langle \mathbf{w}_{j,r}^{\left(t_j(\underline{S}_{j,1}-1)\right)}, j\mathbf{u}\rangle \cdot \big\{1-\tilde{\eta}\big(1-\tilde{f}_{t_j(\bar{S}_{j,1})}\big)\big\} \\ &\geq \langle \mathbf{w}_{j,r}^{\left(t_j(\underline{S}_{j,1}-1)\right)}, j\mathbf{u}\rangle \cdot \big\{1-\tilde{\eta}\big(1-\frac{\tilde{\eta}+2-\sqrt{\tilde{\eta}^2+4\tilde{\eta}}}{2\tilde{\eta}}\big)-o(1)\big\} \\ &\geq 0.2\cdot(\beta_{\mathbf{u},j}^* m)^{1/2}. \qquad (46)\end{aligned}$$

This is again a consequence from Inequality (20), which we have used in the proof of Lemma 16. This positive non-vanishing lower bound guarantees that our $\tilde{\eta}$ choice is sufficient for the sign stability to hold for $s \in [\bar{S}_{j,1}, \underline{S}_{j,2}]$.

**Step 2.3: Lower Bounding $\langle \mathbf{w}_{-j,r}^{(t_j(\underline{S}_{-j,1}))}, -j\mathbf{u}\rangle$ and $\langle \mathbf{w}_{-j,r}^{(t_j(\underline{S}_{j,1}))}, -j\mathbf{v}\rangle$.** It suffices to consider $r \in \mathcal{V}_{-j,-}$ and $r \in \mathcal{U}_{-j,-}$. Inequality (46) indicates that

$$1 \ll \frac{\langle \mathbf{w}_{j,r}^{\left(t_j(\underline{S}_{j,1})\right)}, j\mathbf{u}\rangle}{\langle \mathbf{w}_{j,r}^{\left(t_j(0)\right)}, j\mathbf{u}\rangle} = \prod_{s=0, i_{t_j(s)}\notin\mathcal{W}}^{\underline{S}_{j,1}-1}\{1+\tilde{\eta}(1-\tilde{f}_{t_j(s)})\} \leq \exp\Big\{\tilde{\eta}\sum_{s=0, i_{t_j(s)}\notin\mathcal{W}}^{\underline{S}_{j,1}-1}(1-\tilde{f}_{t_j(s)})\Big\}.$$

Therefore, $\sum_{s=0,i_{t_j(s)}\in\mathcal{W}}^{\underline{S}_{j,1}-1}(1-\tilde{f}_{t_j(s)})>0$. Now for $r\in\mathcal{U}_{-j,-}$, we have that

$$\begin{aligned}
\langle \mathbf{w}_{-j,r}^{\left(t_j(\underline{S}_{j,1})\right)}, -j\mathbf{u}\rangle &= \langle \mathbf{w}_{-j,r}^{\left(t_j(0)\right)}, -j\mathbf{u}\rangle \cdot \prod_{s=0}^{\underline{S}_{j,1}-1}\{1-\tilde{\eta}(1-\tilde{f}_{t_j(\underline{S}_{j,1}-1)})\} \\
&\geq \langle \mathbf{w}_{-j,r}^{\left(t_j(0)\right)}, -j\mathbf{u}\rangle \cdot \exp\Big\{-\tilde{\eta}\sum_{s=0,i_{t_j(s)}\notin\mathcal{W}}^{\underline{S}_{j,1}-1}(1-\tilde{f}_{t_j(\underline{S}_{j,1}-1)})\Big\} \\
&\geq \langle \mathbf{w}_{-j,r}^{\left(t_j(0)\right)}, -j\mathbf{u}\rangle \\
&\geq \sqrt{2\log(16m/p)}\cdot\sigma_0\|\mathbf{u}\|_2. \qquad (47)
\end{aligned}$$

On the other hand, for $s\in[0,\underline{S}_{j,1}-1]$ such that $i_{t_j(s)}\in\mathcal{W}$, condition $t\leq T_{\max}^j=T_{(\mathbf{v})}^j\wedge T_{(\boldsymbol{\xi})}$ guarantees that

$$\begin{aligned}
\tilde{f}_{t_j(s)} &= \frac{1}{m}\sum_{r\in[m]}\Big\{\sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{v}\rangle\big)-\sigma\big(-\langle \mathbf{w}_{-j,r}^{(t_j(s))}, -j\mathbf{v}\rangle\big)\Big\} \\
&\quad+\frac{1}{m}\sum_{r\in[m]}\Big\{\sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(s))}, \boldsymbol{\xi}_{i_{t_j(s)}}\rangle\big)-\sigma\big(\langle \mathbf{w}_{-j,r}^{(t_j(s))}, \boldsymbol{\xi}_{i_{t_j(s)}}\rangle\big)\Big\} \\
&\quad+\frac{1}{m}\sum_{r\in[m]}\Big\{\sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(s))}, \widetilde{\boldsymbol{\xi}}_{i_{t_j(s)}}\rangle\big)-\sigma\big(\langle \mathbf{w}_{-j,r}^{(t_j(s))}, \widetilde{\boldsymbol{\xi}}_{i_{t_j(s)}}\rangle\big)\Big\} \\
&\leq \delta/2+2\times\delta/4 \\
&\leq 1.
\end{aligned}$$

Hence we derive that

$$\sum_{s\in[0,\underline{S}_{j,1}-1]}\Big(1-\tilde{f}_{t_j(s)}\Big)=\sum_{\substack{s\in[0,\underline{S}_{j,1}-1]\\ i_{t_j(s)}\in\mathcal{W}}}\Big(1-\tilde{f}_{t_j(s)}\Big)+\sum_{\substack{s\in[0,\underline{S}_{j,1}-1]\\ i_{t_j(s)}\in\mathcal{W}}}\Big(1-\tilde{f}_{t_j(s)}\Big)\geq 0.$$

And in consequence, for $r\in\mathcal{V}_{j,-}$ it holds that

$$\begin{aligned}
\langle \mathbf{w}_{-j,r}^{(t_j(\underline{S}_{j,1}))}, -j\mathbf{v}\rangle &= \langle \mathbf{w}_{-j,r}^{(t_j(0))}, -j\mathbf{v}\rangle \cdot \prod_{s=0}^{\underline{S}_{j,1}-1}\{1-\alpha\tilde{\eta}(1-\tilde{f}_{t_j(s)})\} \\
&\geq \langle \mathbf{w}_{-j,r}^{(t_j(0))}, -j\mathbf{v}\rangle \cdot \exp\Big\{-\alpha\tilde{\eta}\sum_{s=0}^{\underline{S}_{j,1}-1}(1-\tilde{f}_{t_j(s)})\Big\} \\
&\geq \langle \mathbf{w}_{-j,r}^{(t_j(0))}, -j\mathbf{v}\rangle \\
&\geq -\sqrt{2\log(16m/p)}\cdot\sigma_0\|\mathbf{v}\|_2. \qquad (48)
\end{aligned}$$

The monotonicity again extends lower bounds in Inequality (48) and (47) to $s\in[\bar{S}_{j,1},\bar{S}_{j,2}]$. And the sign stability naturally holds on this interval.

**Step 2.4 & Finalizing.** Thanks to the sign stability proved in the last step, we can now apply Lemma 15 to derive that the upper bound in Inequality (40) continues to hold for $s\in[\underline{S}_{j,1},\bar{S}_{j,2}]$, with exactly the same argument to the Step 1. With an inductive argument that exactly repeats the argument above, we can infer that all the results in Lemma 30 holds for $t_j(s)$ with $\in[\bar{S}_{j,1},\bar{S}_{j,2}]$. Proposition 33 implies that for any $t\in\mathbb{N}$ we can find $s_t\in\mathbb{N}$ such that $t_j(s_t)\leq t<t_j(s_t+1)$ and $\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle=\langle \mathbf{w}_{j,r}^{(t_j(s_t))}, j\mathbf{u}\rangle$ (so are all the other inner products), so all the results in Lemma 30 hold for $t\in\mathbb{N}$. □

### H.5 Proof of Lemma 31

*Proof of Lemma 31.* For any $t_0 \le t_1 \le t_2 \le T^j_{\max}$, $j \in \{\pm 1\}$, and $r \in \mathcal{U}^{(t)}_{j,+}$, by (37),

$$\begin{aligned}\langle \mathbf{w}^{(t_2+1)}_{j,r}, j\mathbf{u}\rangle = \langle \mathbf{w}^{(t_1)}_{j,r}, j\mathbf{u}\rangle + \frac{2\eta\|\mathbf{u}\|_2^2}{m} \cdot \sum_{\substack{t_1\le s\le t_2\\ y_{i_s}=j, i_s\notin\mathcal{W}\\ y_{i_s}f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})>1}} \big(1 - y_{i_s} f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})\big)\cdot\langle \mathbf{w}^{(s)}_{j,r}, j\mathbf{u}\rangle \\ + \frac{2\eta\|\mathbf{u}\|_2^2}{m} \cdot \sum_{\substack{t_1\le s\le t_2\\ y_{i_s}=j, i_s\notin\mathcal{W}\\ y_{i_s}f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})<1}} \big(1 - y_{i_s} f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})\big)\cdot\langle \mathbf{w}^{(s)}_{j,r}, j\mathbf{u}\rangle,\end{aligned}$$

Note that for $t_0 \le t_1 \le t_2 \le T^j_{\max}(t_0)$, we can apply the conclusions of Lemma 30. Specifically, we consider the maximal neuron $r^* = \mathrm{argmax}\langle \mathbf{w}^{(t)}_{j,r}, j\mathbf{u}\rangle$, and

$$\begin{aligned}(1.05)^{\frac{1}{2}}\cdot(\beta^*_{\mathbf{u},j}m)^{\frac{1}{2}} &\ge \Big|\langle \mathbf{w}^{(t_2+1)}_{j,r^*}, j\mathbf{u}\rangle - \langle \mathbf{w}^{(t_1)}_{j,r^*}, j\mathbf{u}\rangle\Big| \qquad (49)\\ &= \frac{2\eta\|\mathbf{u}\|_2^2}{m}\cdot\Bigg|\sum_{\substack{t_1\le s\le t_2\\ y_{i_s}=j, i_s\notin\mathcal{W}\\ y_{i_s}f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})>1}} \big(1 - y_{i_s} f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})\big)\cdot \underbrace{\langle \mathbf{w}^{(s)}_{j,r^*}, j\mathbf{u}\rangle}_{>(\beta^*_{\mathbf{u},j}m)^{\frac{1}{2}}}\\ &\quad - \sum_{\substack{t_1\le s\le t_2\\ y_{i_s}=j, i_s\notin\mathcal{W}\\ y_{i_s}f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})<1}} \big(1 - y_{i_s} f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})\big)\cdot \underbrace{\langle \mathbf{w}^{(s)}_{j,r^*}, j\mathbf{u}\rangle}_{<(1.05-\delta/4)^{\frac{1}{2}}\cdot(\beta^*_{\mathbf{u},j}m)^{\frac{1}{2}}}\Bigg|,\end{aligned}$$

where the red remarks follow from Lemma 15. Rearranging terms, we conclude from (49) that

$$\begin{aligned}&\sum_{\substack{t_1\le s\le t_2\\ y_{i_s}=j, i_s\notin\mathcal{W}\\ y_{i_s}f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})<1}} \big(1 - y_{i_s} f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})\big) \qquad (50)\\ &\ge (1.05-\delta/4)^{-\frac{1}{2}}\cdot \sum_{\substack{t_1\le s\le t_2\\ y_{i_s}=j, i_s\notin\mathcal{W}\\ y_{i_s}f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})>1}} \big(y_{i_s} f(\mathbf{x}_{i_s};\mathbf{W}^{(s)}) - 1\big) - \frac{m(1.05)^{\frac{1}{2}}}{2\eta\|\mathbf{u}\|_2^2(1.05-\delta/4)^{\frac{1}{2}}},\end{aligned}$$

Under Assumption 3, with probability at least $1 - 1/\mathrm{poly}(d)$, it holds that

$$\sharp\{t_1 \le s \le t_2 : y_{i_s} = j, i_s \notin \mathcal{W}\} \ge \frac{1}{4}\cdot(t_2 - t_1 + 1), \qquad (51)$$

Combining (50) and (51), we can finally prove that

$$\sum_{\substack{s=t_1\\ y_{i_s}=j, i_s\notin\mathcal{W}}}^{t_2} 1 - y_{i_s} f(\mathbf{x}_{i_s};\mathbf{W}^{(s)}) \ge \frac{\delta}{8}\cdot\big(1-(1.05-\delta/4)^{\frac{1}{2}}\big)\cdot(t_2-t_1+1) - \frac{m(1.05)^{\frac{1}{2}}}{2\eta\|\mathbf{u}\|_2^2(1.05-\delta/4)^{\frac{1}{2}}}. \qquad (52)$$

Finally, for the weak data $i_s \in \mathcal{W}$, under Assumption 3, with probability at least $1 - 1/\mathrm{poly}(d)$,

$$\sharp\{t_1 \le s \le t_2 : y_{i_s} = j, i_s \in \mathcal{W}\} \le 2\rho\cdot(t_2 - t_1 + 1) \le \frac{\delta}{32}\cdot\big(1-(1.05-\delta/4)^{\frac{1}{2}}\big)\cdot(t_2-t_1+1),$$

where the second inequality follows from the condition on $\rho$ by Assumption 3. By Lemma 30, we have that $|1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})| \leq 2$ for $t_0 \leq s \leq T^j_{\max}(t_0)$. Therefore, we have that

$$\sum_{\substack{s=t_1 \\ y_{i_s}=j, i_s \in \mathcal{W}}}^{t_2} 1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)}) \geq -\frac{\delta}{16} \cdot \big(1 - (1.05 - \delta/4)^{\frac{1}{2}}\big) \cdot (t_2 - t_1 + 1). \tag{53}$$

Combining (52) and (53), we can conclude that

$$\sum_{\substack{s=t_1 \\ y_{i_s}=j}}^{t_2} 1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)}) \geq \frac{\delta}{16} \cdot \big(1 - (1.05 - \delta/4)^{\frac{1}{2}}\big) \cdot (t_2 - t_1 + 1) - \frac{m(1.05)^{\frac{1}{2}}}{2\eta \|\mathbf{u}\|_2^2 (1.05 - \delta/4)^{\frac{1}{2}}}.$$

This finishes the proof of the first part in Lemma 31. Now we consider the second part. For simplicity, we denote by $\alpha = \|\mathbf{v}\|_2^2 / \|\mathbf{u}\|_2^2$ and $\tilde{\eta} = 2\|\mathbf{u}\|_2^2 / m$. Consider that for any $0 \leq t \leq T^j_{\max}$, $j \in \{\pm 1\}$, and $r \in \mathcal{V}^{(t)}_{j,+}$, due to Lemma 30, the $\mathbf{v}$-sign stability condition is true on $[0, T^j_{\max}(t_0)]$. In view of (38), this means that

$$1 + \alpha\tilde{\eta} \cdot \big(1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})\big) > 0, \quad \forall t_0 \leq s \leq T^j_{\max} \text{ s.t. } y_{i_s} = j.$$

Then for any $t_0 \leq t \leq T^j_{\max}(t_0)$, $t_0 \leq s \leq t$, and $r \in \mathcal{C}^{(t)}_{j,+}$, since $-2 \leq 1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(t)}) \leq 2$ due to Lemma 30, we can lower bound the relative increment as

$$\begin{aligned}
&\sum_{\substack{s=t_0 \\ y_{i_s}=j}}^{t} \log\Big\{1 + \alpha\tilde{\eta}\big(1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})\big)\Big\} = \sum_{\substack{s=t_0 \\ y_{i_s}=j}}^{t} \int_0^{\alpha} \frac{\tilde{\eta}\big(1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})\big)}{1 + \tilde{\eta} z \big(1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})\big)} dz \\
&= \sum_{\substack{s=t_0 \\ y_{i_s}=j}}^{t} \int_0^{\alpha} \frac{\tilde{\eta}\Big(\big(1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})\big) + 2\Big)}{1 + \tilde{\eta} z \big(1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})\big)} dz - \sum_{\substack{s=t_0 \\ y_{i_s}=j}}^{t} \int_0^{\alpha} \frac{2\tilde{\eta}}{1 + \tilde{\eta} z \big(1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})\big)} dz \\
&\geq \sum_{\substack{s=t_0 \\ y_{i_s}=j}}^{t} \int_0^{\alpha} \frac{\tilde{\eta}\Big(\big(1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})\big) + 2\Big)}{1 + 2\tilde{\eta} z} dz - \sum_{\substack{s=t_0 \\ y_{i_s}=j}}^{t} \int_0^{\alpha} \frac{2\tilde{\eta}}{1 - 2\tilde{\eta} z} dz \\
&\geq \int_0^{\alpha} \frac{\tilde{\eta}\Big(\sum_{s=t_0, y_{i_s}=j}^{t} \big(1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(t)})\big) + 2N_j(t_0, t)\Big)}{1 + 2\tilde{\eta} z} dz - \int_0^{\alpha} \frac{2\tilde{\eta} N_j(t_0, t)}{1 - 2\tilde{\eta} z} dz,
\end{aligned} \tag{54}$$

where for simplicity we denote $N_j(t_0, t) = \sharp\{t_0 \leq s \leq t : y_{i_s} = j\}$. Now we can use Proposition 31 to lower bound the right hand side of (54), Denoting $\epsilon = \delta \cdot (1 - (1.05 - \delta/4)^{\frac{1}{2}})/16$, we have that

$$\begin{aligned}
&\sum_{\substack{s=t_0 \\ y_{i_s}=j}}^{t} \log\Big\{1 + \alpha\tilde{\eta}\big(1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})\big)\Big\} \\
&\geq \int_0^{\alpha} \frac{\tilde{\eta}\Big(\epsilon(t - t_0 + 1) - \Delta + 2N_j(t_0, t)\Big)}{1 + 2\tilde{\eta} z} dz - \int_0^{\alpha} \frac{2\tilde{\eta} N_j(t_0, t)}{1 - 2\tilde{\eta}} dz \\
&\geq \left(\frac{\epsilon}{2}(t - t_0 + 1) - \frac{\Delta}{2} + N_j(t_0, t)\right) \cdot \log(1 + 2\alpha\tilde{\eta}) + N_j(t_0, t) \cdot \log(1 - 2\alpha\tilde{\eta}) \\
&\geq \left(\frac{\epsilon}{2}(t - t_0 + 1) - \frac{\Delta}{2}\right) \cdot \log(1 + 2\alpha\tilde{\eta}) + N_j(t_0, t) \cdot \log\big\{(1 + 2\alpha\tilde{\eta}) \cdot (1 - 2\alpha\tilde{\eta})\big\},
\end{aligned}$$

where $\Delta$ is defined in Lemma 31. Moreover since for our choice of $\alpha\tilde{\eta} \ll 1$ in Assumption 3,

$$\log(1 + 2\alpha\tilde{\eta}) \geq \alpha\tilde{\eta}, \quad \log\big\{(1 + 2\alpha\tilde{\eta}) \cdot (1 - 2\alpha\tilde{\eta})\big\} = \log\big\{1 - 4\alpha^2\tilde{\eta}^2\big\} \geq -2\alpha^2\tilde{\eta}^2,$$

and using the fact that with probability at least $1-1/\mathrm{poly}(d)$ it holds that $N_j(t_0,t) \le (t-t_0+1)/2$, we finally have that

$$\sum_{\substack{s=t_0\\ y_{i_s}=j}}^{t} \log\Big\{1+\alpha\tilde{\eta}\big(1-y_{i_s}f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})\big)\Big\} \ge \frac{\alpha\tilde{\eta}}{2}\cdot(\epsilon-2\alpha\tilde{\eta})\cdot(t_1-t_0+1)-\Delta\cdot\log(1+2\alpha\tilde{\eta})$$
$$\ge \frac{1}{4}\alpha\tilde{\eta}\cdot\epsilon\cdot(t_1-t_0+1)-2\alpha\tilde{\eta}\cdot\Delta.$$

Finally, using (38) again, we can lower bound our target as

$$\begin{aligned}\langle \mathbf{w}_{j,r}^{(t+1)}, j\mathbf{v}\rangle &= \langle \mathbf{w}_{j,r}^{(t_0)}, j\mathbf{v}\rangle \cdot \prod_{\substack{s=t_0\\ y_{i_s}=j}}^{t}\Big(1+\alpha\tilde{\eta}\big(1-y_{i_s}f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})\big)\Big)\\ &= \langle \mathbf{w}_{j,r}^{(t_0)}, j\mathbf{v}\rangle \cdot \exp\left\{\sum_{\substack{s=t_0\\ y_{i_s}=j}}^{t}\log\Big\{1+\alpha\tilde{\eta}\big(1-y_{i_s}f(\mathbf{x}_{i_s};\mathbf{W}^{(s)})\big)\Big\}\right\}\\ &\ge \langle \mathbf{w}_{j,r}^{(t_0)}, j\mathbf{v}\rangle \cdot \exp\left\{\frac{1}{4}\alpha\tilde{\eta}\cdot\epsilon\cdot(t_1-t_0+1)-2\alpha\tilde{\eta}\cdot\Delta\right\}.\end{aligned}$$

Plugging in the definition of $\epsilon$, $\Delta$, $\alpha$, and $\tilde{\eta}$, we can arrive at

$$\begin{aligned}&\langle \mathbf{w}_{j,r}^{(t+1)}, j\mathbf{v}\rangle\\ &\quad\ge \langle \mathbf{w}_{j,r}^{(t_0)}, j\mathbf{v}\rangle \cdot \exp\left\{\frac{\eta\|\mathbf{v}\|_2^2}{32m}\cdot\big(\delta-\delta(1.05-\delta/4)^{\frac{1}{2}}\big)\cdot(t-t_0+1)-\frac{\|\mathbf{v}\|_2^2(1.05)^{\frac{1}{2}}}{\|\mathbf{u}\|_2^2(1.05-\delta/4)^{\frac{1}{2}}}\right\}.\end{aligned}$$

This finishes the proof of Lemma 31. □

### H.6 Technical Results and Proof

*Proof of Lemma 15 in Multiple Data Setting.* In the multiple data setting, the (positive) inner products only changes at the steps where the corresponding data label aligns with the directions of the neurons (i.e., $j=\pm 1$). Define

$$\beta_{\mathbf{u},j}^{*,(t_1)} = \frac{\max_r \sigma(\langle \mathbf{w}_{j,r}^{(t_1)}, j\mathbf{u}\rangle)}{\sum_r \sigma(\langle \mathbf{w}_{j,r}^{(t_1)}, j\mathbf{u}\rangle)}.$$

Again, the local sign stability assumption ensures that each inner product grows proportionally and the superscript $(t)$ in the neuron index sets can be dropped, with

$$\begin{aligned}&\frac{1}{m}\sum_{r\in[m]}\sigma\big(\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle\big)\\ &\quad= \frac{1}{m}\sum_{r\in\mathcal{U}_{j,+}}\sigma\big(\langle \mathbf{w}_{j,r}^{(t_1)}, j\mathbf{u}\rangle\big)\cdot\prod_{\substack{t'\in[t_1,t-1]:\\ y_{i_t}=j,\ i_t\notin\mathcal{W}}}\Big\{1+\frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1-y_{i_{t'}}f(\mathbf{x}_{i_{t'}};\mathbf{W}^{(t')})\big)\Big\}^2\\ &\quad= \frac{\max_{r\in[m]}\sigma\big(\langle \mathbf{w}_{j,r}^{(t_1)}, j\mathbf{u}\rangle\big)}{m\beta_{\mathbf{u},j}^{*,(t_1)}}\prod_{\substack{t'\in[t_1,t-1]:\\ y_{i_{t'}}=j,\ i_t\notin\mathcal{W}}}\Big\{1+\frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1-y_{i_{t'}}f(\mathbf{x}_{i_{t'}};\mathbf{W}^{(t')})\big)\Big\}^2\\ &\quad= \frac{\sigma\big(\max_{r\in[m]}\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle\big)}{m\beta_{\mathbf{u},j}^{*,(t_1)}}\end{aligned}$$

Here the second line and the last equality is true because Equation (9) implies that all the positive $\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle$ iterates by sequentially multiplying the same factor $\Big\{1 + \frac{2\eta\|\mathbf{u}\|_2^2}{m}\big(1 - yf(\mathbf{x}; \mathbf{W}^{(t')})\big)\Big\}$. The third equality comes from the definition of $\beta_{\mathbf{u}}^{*,(t_1)}$ in Lemma 15.

Therefore, $m^{-1} \cdot \sum_{r\in[m]} \sigma(\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle) > c$ implies that $\sigma(\max_{r\in[m]}\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u}\rangle) > \beta_{\mathbf{u},j}^{*,(t_1)} mc$ and the desired lower bound follows. The upper bound can be proved analogously. □

*Proof of Proposition 32.* We prove the result by induction. For step $t = t_1$, the result holds trivially. Suppose that this result holds for each step $t_1, \cdots, t$, then by (39), we have that

$$
\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i\rangle &= \langle \mathbf{w}_{j,r}^{(t_1)}, \boldsymbol{\xi}_i\rangle + \sum_{s=t_1}^{t} \frac{\eta \cdot jy_{i_s}}{m} \cdot \big(1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})\big) \cdot \Big(\sigma'(\langle \mathbf{w}_{j,r}^{(s)} \cdot \boldsymbol{\xi}_{i_s}\rangle) \cdot \langle \boldsymbol{\xi}_{i_s}, \boldsymbol{\xi}_i\rangle \\
&\qquad + \sigma'(\langle \mathbf{w}_{j,r}^{(s)} \cdot \widetilde{\boldsymbol{\xi}}_{i_s}\rangle) \cdot \langle \widetilde{\boldsymbol{\xi}}_{i_s}, \boldsymbol{\xi}_i\rangle \cdot \mathbf{1}\{i_s \in \mathcal{W}\}\Big) \\
&= \langle \mathbf{w}_{j,r}^{(t_1)}, \boldsymbol{\xi}_i\rangle + \sum_{\substack{s=t_1 \\ i_s = i}}^{t} \frac{\eta\|\boldsymbol{\xi}_i\|_2^2 \cdot jy_{i_s}}{m} \cdot \big(1 - y_i f(\mathbf{x}_i; \mathbf{W}^{(s)})\big) \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(s)} \cdot \boldsymbol{\xi}_i\rangle) \\
&\qquad + \sum_{\substack{s=t_1 \\ i_s \neq i}}^{t} \frac{\eta\|\boldsymbol{\xi}_i\|_2^2 \cdot jy_{i_s}}{m} \cdot \big(1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})\big) \cdot \Big(\sigma'(\langle \mathbf{w}_{j,r}^{(s)} \cdot \boldsymbol{\xi}_{i_s}\rangle) \cdot \frac{\langle \boldsymbol{\xi}_{i_s}, \boldsymbol{\xi}_i\rangle}{\|\boldsymbol{\xi}_i\|_2^2} \\
&\qquad\qquad + \sigma'(\langle \mathbf{w}_{j,r}^{(s)} \cdot \widetilde{\boldsymbol{\xi}}_{i_s}\rangle) \cdot \frac{\langle \widetilde{\boldsymbol{\xi}}_{i_s}, \boldsymbol{\xi}_i\rangle}{\|\boldsymbol{\xi}_i\|_2^2} \cdot \mathbf{1}\{i_s \in \mathcal{W}\}\Big)
\end{aligned}
$$

Taking absolute value, applying the definition of $\Upsilon^{(s)}$ we obtain that

$$
\begin{aligned}
&\Big|\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i\rangle\Big| \qquad (55)\\
&\quad \le \Big|\langle \mathbf{w}_{j,r}^{(t_1)}, \boldsymbol{\xi}_i\rangle\Big| \\
&\qquad + \bigg|\sum_{\substack{s=t_1 \\ i_s = i}}^{t} \frac{\eta\|\boldsymbol{\xi}_i\|_2^2 \cdot jy_{i_s}}{m} \cdot \big(1 - y_i f(\mathbf{x}_i; \mathbf{W}^{(s)})\big) \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(s)} \cdot \boldsymbol{\xi}_i\rangle)\bigg| \\
&\qquad + \bigg|\sum_{\substack{s=t_1 \\ i_s \neq i}}^{t} \frac{\eta\|\boldsymbol{\xi}_i\|_2^2 \cdot jy_{i_s}}{m} \cdot \big(1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})\big) \\
&\qquad\qquad \cdot \bigg(\sigma'(\langle \mathbf{w}_{j,r}^{(s)} \cdot \boldsymbol{\xi}_{i_s}\rangle) \cdot \frac{\langle \boldsymbol{\xi}_{i_s}, \boldsymbol{\xi}_i\rangle}{\|\boldsymbol{\xi}_i\|_2^2} + \sigma'(\langle \mathbf{w}_{j,r}^{(s)} \cdot \widetilde{\boldsymbol{\xi}}_{i_s}\rangle) \cdot \frac{\langle \widetilde{\boldsymbol{\xi}}_{i_s}, \boldsymbol{\xi}_i\rangle}{\|\boldsymbol{\xi}_i\|_2^2} \cdot \mathbf{1}\{i_s \in \mathcal{W}\}\bigg)\bigg| \\
&\quad \le \Upsilon^{(t_1)} + \sum_{\substack{s=t_1 \\ i_s = i}}^{t} \frac{2\eta\|\boldsymbol{\xi}_i\|_2^2}{m} \cdot \big|1 - y_i f(\mathbf{x}_i; \mathbf{W}^{(s)})\big| \cdot \Upsilon^{(s)} \\
&\qquad + \sum_{\substack{s=t_1 \\ i_s \neq i}}^{t} \frac{2\eta\|\boldsymbol{\xi}_i\|_2^2}{m} \cdot \big|1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})\big| \cdot \Upsilon^{(s)} \cdot \bigg(\frac{|\langle \boldsymbol{\xi}_{i_s}, \boldsymbol{\xi}_i\rangle|}{\|\boldsymbol{\xi}_i\|_2^2} + \frac{|\langle \widetilde{\boldsymbol{\xi}}_{i_s}, \boldsymbol{\xi}_i\rangle|}{\|\boldsymbol{\xi}_i\|_2^2}\bigg).
\end{aligned}
$$

Now using the fact that with probability at least $1 - 1/\mathrm{poly}(d)$, $|1 - y_{i_s} f(\mathbf{x}_{i_s}; \mathbf{W}^{(s)})| \leq 2$, $\|\boldsymbol{\xi}_i\|_2^2 \leq 3\sigma_p^2 d/2$, and that $|\langle \boldsymbol{\xi}, \boldsymbol{\xi}' \rangle| / \|\boldsymbol{\xi}\|_2^2 \leq \widetilde{\mathcal{O}}(d^{-1/2})$ for i.i.d. $\boldsymbol{\xi}$, $\boldsymbol{\xi}'$, we can further upper bound (55) as

$$\begin{aligned}
\max_{i\in[n]} \Big| \langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle \Big| &\leq \Upsilon^{(t_1)} + \frac{6\eta\sigma_p^2 d}{m} \cdot \left( \sum_{\substack{s=t_1 \\ i_s = i}}^{t} \Upsilon^{(s)} + \widetilde{\mathcal{O}}(d^{-1/2}) \cdot \sum_{\substack{s=t_1 \\ i_s \neq i}}^{t} \Upsilon^{(s)} \right) \qquad (56) \\
&= \Upsilon^{(t_1)} + \frac{6\eta\sigma_p^2 d}{m} \cdot \left( \sum_{\substack{s=t_1 \\ i_s = i}}^{t} \Upsilon^{(t_1)} + \widetilde{\mathcal{O}}(d^{-1/2}) \cdot \sum_{\substack{s=t_1 \\ i_s \neq i}}^{t} \Upsilon^{(t_1)} \right) \\
&\quad + \frac{6\eta\sigma_p^2 d}{m} \cdot \left( \sum_{\substack{s=t_1 \\ i_s = i}}^{t} \big( \Upsilon^{(t_1)} - \Upsilon^{(s)} \big) + \widetilde{\mathcal{O}}(d^{-1/2}) \cdot \sum_{\substack{s=t_1 \\ i_s \neq i}}^{t} \big( \Upsilon^{(t_1)} - \Upsilon^{(s)} \big) \right).
\end{aligned}$$

By our induction, we have that $\Upsilon^{(s)} - \Upsilon^{(t_1)} \leq \Upsilon^{(t_1)} \cdot \epsilon$, for which we can further bound (56) as

$$\begin{aligned}
\max_{i\in[n]} \Big| \langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle \Big| &\leq \Upsilon^{(t_1)} \cdot \left[ 1 + \frac{6\eta\sigma_p^2 d}{m} \cdot (1+\epsilon) \cdot \left( \sum_{\substack{s=t_1 \\ i_s = i}}^{t} 1 + \widetilde{\mathcal{O}}(d^{-1/2}) \cdot \sum_{\substack{s=t_1 \\ i_s \neq i}}^{t} 1 \right) \right] \\
&\leq \Upsilon^{(t_1)} \cdot \left[ 1 + \frac{6\eta\sigma_p^2 d}{m} \cdot (1+\epsilon) \cdot \left( \frac{2(t - t_1 + 1)}{n} + \widetilde{\mathcal{O}}(d^{-1/2}) \cdot (t - t_1 + 1) \right) \right] \\
&\leq \Upsilon^{(t_1)} \cdot \left[ 1 + \frac{18\eta\sigma_p^2 d}{mn} \cdot (1+\epsilon) \cdot (t - t_1 + 1) \right],
\end{aligned}$$

where in the first inequality we utilize the fact that $\sharp\{t_1 \leq s \leq t : i_s = i\} \leq 2(t - t_1 + 1)/n$, and in the last inequality we use the condition that $d = \widetilde{\Omega}(n^2)$. Therefore, when $t \leq t_1 + \widetilde{T}_{(\boldsymbol{\xi})} - 1$ with $\widetilde{T}_{(\boldsymbol{\xi})} = \widetilde{\Theta}(mn \cdot \eta^{-1} \cdot \epsilon \cdot (1+\epsilon)^{-1} \cdot (\sigma_p^2 d)^{-1})$, it holds that

$$\max_{i\in[n]} \Big| \langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle \Big| \leq \Upsilon^{(t_1)} \cdot (1+\epsilon). \qquad (57)$$

By using the same argument as proving (57), we can also show that for $t \leq t_1 + \widetilde{T}_{(\boldsymbol{\xi})} - 1$

$$\max_{i\in\mathcal{W}} \Big| \langle \mathbf{w}_{j,r}^{(t+1)}, \widetilde{\boldsymbol{\xi}}_i \rangle \Big| \leq \Upsilon^{(t_1)} \cdot (1+\epsilon). \qquad (58)$$

By combining (57) and (58), we can arrive at

$$\Upsilon^{(t+1)} = \max_{j\in\{pm\}} \left\{ \max_{i\in[n]} \Big| \langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle \Big|, \max_{i\in\mathcal{W}} \Big| \langle \mathbf{w}_{j,r}^{(t+1)}, \widetilde{\boldsymbol{\xi}}_i \rangle \Big| \right\} \leq \Upsilon^{(t_1)} \cdot (1+\epsilon).$$

Thus we have proved our induction statement for step $t+1$. Repeating the induction completes the proof of Proposition 32. □

*Proof of Proposition 33.* From Equation (37), for $r \in \mathcal{U}_{j,+}$ we can infer that

$$\langle \mathbf{w}_{j,r}^{(t_j(s+1))}, j\mathbf{u} \rangle = \langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u} \rangle + \sum_{t=t_j(s)}^{t_j(s+1)-1} \frac{\eta \|\mathbf{u}\|_2^2}{m} \cdot \big(1 - \tilde{f}_t\big) \cdot \sigma'\big( \langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u} \rangle \cdot j y_{i_t} \big) \cdot \mathbf{1}_{i_t \in \mathcal{W}}.$$

From the definition of $t_j(s)$ we know that for $t \in (t_j(s), t_j(s+1))$, $y_{i_t} j = -1$. On the other hand $y_{i_{t_j(s)}} j = 1$ and $r \in \mathcal{U}_{j,+}$, hence we obtain that

$$\langle \mathbf{w}_{j,r}^{(t_j(s)+1)}, j\mathbf{u} \rangle = \langle \mathbf{w}_{j,r}^{(t_j(s))}, j\mathbf{u} \rangle + \frac{2\eta \|\mathbf{u}\|_2^2}{m} \cdot \big(1 - \tilde{f}_{t_j(s)}\big) \cdot \langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u} \rangle \cdot \mathbf{1}_{i_{t_j(s)} \in \mathcal{W}}.$$

Since the sign stability holds, this multiplication by a non-negative factor does not change the sign of the inner products. Therefore we have that

$$\langle \mathbf{w}_{j,r}^{\left(t_j(s+1)\right)}, j\mathbf{u}\rangle = \langle \mathbf{w}_{j,r}^{\left(t_j(s+1)-1\right)}, j\mathbf{u}\rangle = \cdots = \langle \mathbf{w}_{j,r}^{\left(t_j(s)+1\right)}, j\mathbf{u}\rangle,$$

which concludes our results. Others can be proved analogously and are omitted. □

*Proof of Proposition 34.* We expand $\tilde{f}_{t_j(\bar{S}_{j,k})}$ as follows.

$$\begin{aligned}
\tilde{f}_{t_j(\bar{S}_{j,k})} \leq{}& \frac{1}{m}\sum_{r\in[m]} \sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(\bar{S}_{j,k}-1))}, j\mathbf{u}\rangle\big)\cdot\Big\{1+\tilde{\eta}\big(1-\tilde{f}_{t_j(\bar{S}_{j,k}-1)}\big)\Big\}^2 \\
&+\frac{1}{m}\sum_{r\in[m]} \sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(\bar{S}_{j,k}-1))}, j\mathbf{v}\rangle\big)\cdot\Big\{1+\alpha\tilde{\eta}\big(1-\tilde{f}_{t_j(\bar{S}_{j,k}-1)}\big)\Big\}^2 \\
&-\frac{1}{m}\sum_{r\in[m]} \sigma\big(-\langle \mathbf{w}_{-j,r}^{(t_j(\bar{S}_{j,k}-1))}, -j\mathbf{u}\rangle\big)\cdot\Big\{1+\tilde{\eta}\big(1+\tilde{f}_{t_j(\bar{S}_{j,k}-1)}\big)\Big\}^2 \\
&-\frac{1}{m}\sum_{r\in[m]} \sigma\big(-\langle \mathbf{w}_{j,r}^{(t_j(\bar{S}_{j,k}-1))}, -j\mathbf{v}\rangle\big)\cdot\Big\{1-\alpha\tilde{\eta}\big(1-\tilde{f}_{t_j(\bar{S}_{j,k}-1)}\big)\Big\}^2 \\
&+\Big|\frac{1}{m}\sum_{r\in[m]}\langle \mathbf{w}_{j,r}^{(t_j(\bar{S}_{j,k}))}, \boldsymbol{\xi}_{i_{t_j(\bar{S}_{j,k})}}\rangle+\frac{1}{m}\sum_{r\in[m]}\langle \mathbf{w}_{-j,r}^{(t_j(\bar{S}_{j,k}))}, \boldsymbol{\xi}_{i_{t_j(\bar{S}_{j,k})}}\rangle\Big| \\
\leq{}& \frac{1}{m}\sum_{r\in[m]}\Big\{\sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(\bar{S}_{j,k}-1))}, j\mathbf{u}\rangle\big)+\sigma\big(\langle \mathbf{w}_{j,r}^{(t_j(\bar{S}_{j,k}-1))}, j\mathbf{v}\rangle\big) \\
&-\sigma\big(-\langle \mathbf{w}_{-j,r}^{(t_j(\bar{S}_{j,k}-1))}, -j\mathbf{u}\rangle\big)-\sigma\big(-\langle \mathbf{w}_{j,r}^{(t_j(\bar{S}_{j,k}-1))}, -j\mathbf{v}\rangle\big)\Big\}\cdot\Big\{1+\tilde{\eta}\big(1-\tilde{f}_{t_j(\bar{S}_{j,k}-1)}\big)\Big\}^2 \\
&+\frac{1}{m}\sum_{r\in[m]}\Big\{\sigma\big(-\langle \mathbf{w}_{-j,r}^{(t_j(\bar{S}_{j,k}-1))}, -j\mathbf{u}\rangle\big)+\sigma\big(-\langle \mathbf{w}_{j,r}^{(t_j(\bar{S}_{j,k}-1))}, -j\mathbf{v}\rangle\big)\Big\}\cdot\Big\{2\tilde{\eta}\big(1-\tilde{f}_{t_j(\bar{S}_{j,k}-1)}\big)\Big\}+\delta/2 \\
\leq{}& \tilde{f}_{t_j(\bar{S}_{j,k}-1)}\cdot\Big\{1+\tilde{\eta}\big(1-\tilde{f}_{t_j(\bar{S}_{j,k}-1)}\big)\Big\}^2+E_{t_j(\bar{S}_{j,k}-1)}+\delta/2.
\end{aligned}$$

By Assumption 12 we know that $\tilde{f}_{t_j(\bar{S}_{j,k})} > 1+\delta$. From the discussion in the proof of Lemma 19, we know that once

$$E_{t_j(\bar{S}_{j,k}-1)} \coloneqq \frac{1}{m}\sum_{r\in[m]}\Big\{\sigma\big(-\langle \mathbf{w}_{-j,r}^{(t_j(\bar{S}_{j,k}-1))}, -j\mathbf{u}\rangle\big)+\sigma\big(-\langle \mathbf{w}_{j,r}^{(t_j(\bar{S}_{j,k}-1))}, -j\mathbf{v}\rangle\big)\Big\} \leq \delta/4,$$

we have that

$$\tilde{f}_{t_j(\bar{S}_{j,k}-1)} \geq \frac{\tilde{\eta}+2-\sqrt{\tilde{\eta}^2+4\tilde{\eta}}}{2\tilde{\eta}}.$$

and that

$$\begin{aligned}
\tilde{f}_{t_j(\bar{S}_{j,k})} &\leq \Big\{1-\tilde{\eta}(1-\tilde{f}_{t_j(\bar{S}_{j,k}-1)})\Big\}^2+2E_{t_j(\bar{S}_{j,k}-1)}+\delta/2 \\
&\leq \big(\tilde{\eta}/2+\sqrt{\tilde{\eta}^2/4+\tilde{\eta}}\big)^2+2E_{t_j(\bar{S}_{j,k}-1)}+\delta/2.
\end{aligned}$$

This finishes the proof of Proposition 34 □

# I Multiple-data Case: Small Learning Rate

Recall that $\mathcal{W}$ is the index set of training data points which lack the strong feature patch. By (3), the CNN weights are updated according to

$$\begin{aligned}\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} - \frac{j\eta}{m}\cdot\big(f(\mathbf{x}_{i_t};\mathbf{W}^{(t)}) - y_{i_t}\big)\cdot\Big(&\sigma'(\langle\mathbf{w}_{j,r}^{(t)}, y_{i_t}\mathbf{u}\rangle)\cdot y_{i_t}\mathbf{u}\cdot\mathbf{1}\{i_t\notin\mathcal{W}\}\\ &+\sigma'(\langle\mathbf{w}_{j,r}^{(t)}, y_{i_t}\mathbf{v}\rangle)\cdot y_{i_t}\mathbf{v} + \sigma'(\langle\mathbf{w}_{j,r}^{(t)}\cdot\boldsymbol{\xi}_{i_t}\rangle)\cdot\boldsymbol{\xi}_{i_t}\\ &+\sigma'(\langle\mathbf{w}_{j,r}^{(t)}\cdot\widetilde{\boldsymbol{\xi}}_{i_t}\rangle)\cdot\widetilde{\boldsymbol{\xi}}_{i_t},\cdot\mathbf{1}\{i_t\in\mathcal{W}\}.\Big)\end{aligned}$$

Also, by the CNN expression (1), for each $j\in\{\pm1\}$, the inner products that matter are

*(i) positive neurons:* $\langle\mathbf{w}_{j,r}^{(t+1)}, j\mathbf{u}\rangle$ for $r\in\mathcal{B}_{j,+}^{(t)}$, $\langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle$ for $r\in\mathcal{C}_{j,+}^{(t)}$, and $\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}_i\rangle$.

*(ii) negative neurons:* $\langle\mathbf{w}_{-j,r}^{(t+1)}, j\mathbf{u}\rangle$ for $r\notin\mathcal{B}_{-j,+}^{(t)}$, $\langle\mathbf{w}_{-j,r}^{(t)}, j\mathbf{v}\rangle$ for $r\notin\mathcal{C}_{-j,+}^{(t)}$, and $\langle\mathbf{w}_{-j,r}^{(t)},\boldsymbol{\xi}_i\rangle$.

Subsequently, update formulas of those inner products of interests are given by

$$\begin{aligned}\langle\mathbf{w}_{j,r}^{(t+1)}, j\mathbf{u}\rangle &= \langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle\\ &\quad+\frac{\eta\|\mathbf{u}\|_2^2}{m}\cdot\big(1-y_{i_t}f(\mathbf{x}_{i_t};\mathbf{W}^{(t)})\big)\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle jy_{i_t})\cdot\mathbf{1}\{i_t\notin\mathcal{W}\},\\ \langle\mathbf{w}_{j,r}^{(t+1)}, j\mathbf{v}\rangle &= \langle\mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle + \frac{\eta\|\mathbf{v}\|_2^2}{m}\cdot\big(1-y_{i_t}f(\mathbf{x}_{i_t};\mathbf{W}^{(t)})\big)\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(t)}\cdot j\mathbf{v}\rangle jy_{i_t}),\\ \langle\mathbf{w}_{j,r}^{(t+1)},\boldsymbol{\xi}_i\rangle &= \langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}_i\rangle + \frac{\eta\cdot jy_{i_t}}{m}\cdot\big(1-y_{i_t}f(\mathbf{x}_{i_t};\mathbf{W}^{(t)})\big)\cdot\Big(\sigma'(\langle\mathbf{w}_{j,r}^{(t)}\cdot\boldsymbol{\xi}_{i_t}\rangle)\cdot\langle\boldsymbol{\xi}_{i_t},\boldsymbol{\xi}_i\rangle\\ &\quad+\sigma'(\langle\mathbf{w}_{j,r}^{(t)}\cdot\widetilde{\boldsymbol{\xi}}_{i_t}\rangle)\cdot\langle\widetilde{\boldsymbol{\xi}}_{i_t},\boldsymbol{\xi}_i\rangle\cdot\mathbf{1}\{i_t\in\mathcal{W}\}\Big).\end{aligned}$$

**Assumption 35.** *Suppose that the following holds: For some $\epsilon\in(0,1)$,*

1. *The weight initialization scale $\sigma_0=\widetilde{\Theta}(\|\mathbf{u}\|^{-1})$;*
2. *Strong signal strength $\|\mathbf{u}\|_2 > \widetilde{\Omega}(m/\sqrt{n})\cdot\sigma_p\sqrt{d}$ and weak signal strength $\sigma_p\sqrt{d}\geq\widetilde{\Omega}\left(m/\sqrt{n}\right)\cdot\|\mathbf{v}\|$;*
3. *The dimension $d$ satisfies $d=\Omega(\mathrm{polylog}(m,n))$.*

**Theorem 36** (Restatement of Proposition 6)**.** *Under Assumption 35 on $(d,m,\sigma_0,\|\mathbf{u}\|,\|\mathbf{v}\|,\epsilon)$, if we choose the learning rate $\eta\leq\frac{m}{6\|\mathbf{u}\|^2}$ small enough and $\epsilon'\in(0,1)$, then with high probability $1-1/\mathrm{poly}(d,m)$, there exist*

$$T^\dagger=\frac{4m}{\eta(1-\tau)(1-\rho)\|\mathbf{u}\|^2}\log\left(\frac{2\iota}{\sigma_0\|\mathbf{u}\|}\right),\quad T=T^\dagger+\left\lfloor\frac{Cm^3}{2\eta\epsilon\|\mathbf{u}\|^2}\right\rfloor,$$
$$T'=T+\left\lfloor\frac{\left\|\mathbf{W}^{(T)}-\mathbf{W}^\star\right\|_F^2}{2\eta\epsilon}\right\rfloor,$$

*such that: (i) average loss on samples $\mathcal{W}^c$ decreased to $3\epsilon$ over iterations $[T^\dagger,T]$, i.e.*

$$\frac{1}{2n}\sum_{i\in\mathcal{W}^c}\min_{T^\dagger\leq t\leq T}\Big(y_i-f(\mathbf{x}_i;\mathbf{W}^{(t)})\Big)^2\leq3\epsilon;$$

*(ii) average loss on samples $\mathcal{W}$ decreased to $3\epsilon'$ over iterations $[T,T']$, i.e.*

$$\frac{1}{2n}\sum_{i\in\mathcal{W}}\min_{T\leq t\leq T'}\Big(y_i-f(\mathbf{x}_i;\mathbf{W}^{(t)})\Big)^2\leq3\epsilon';$$

*(iii) the model does not learn weak signal $\mathbf{v}$ well enough even until $T'$, compared to initialization, i.e. $\max_{j,r}|\langle\mathbf{w}_{j,r}^{(t)},\mathbf{v}\rangle|\leq\sigma_0\|\mathbf{v}\|, t\leq T'$.*

For convenience, we also write $\ell_i^{(t)} = f(\mathbf{x}_i; \mathbf{W}^{(t)}) - y_i$ as the fitting residual.

**Lemma 37.** *For any $0 < \tau < 1$ to be tuned later, suppose that at some time $t$ there holds*

$$\max_{j,r} \left\{ |\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u} \rangle|, \quad |\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v} \rangle|, \quad \max_{i\in[n]} |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle|, \quad \max_{i\in\mathcal{W}} |\langle \mathbf{w}_{j,r}^{(t)}, \widetilde{\boldsymbol{\xi}}_i \rangle| \right\} \le \sqrt{\frac{\tau}{3}},$$

*then we can lower bound the fitting residual $-y\ell_i^{(t)} \ge 1 - \tau$ for every $i \in [n]$.*

*Proof.* Plug into the CNN model definition

$$-y\ell_i^{(t)} = 1 - F_y(\mathbf{x}_i; \mathbf{W}^{(t)}) + F_{-y}(\mathbf{x}_i; \mathbf{W}^{(t)}) \ge 1 - F_y(\mathbf{x}_i; \mathbf{W}^{(t)}).$$

If $i \in \mathcal{W}^c$, we can upper bound $F_y(\mathbf{x}_i; \mathbf{W}^{(t)})$ further by

$$\begin{aligned} F_y(\mathbf{x}_i; \mathbf{W}^{(t)}) =& \frac{1}{m} \sum_{r\in[m]} \Big[ \sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u} \rangle) + \sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v} \rangle) + \sigma(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \Big] \\ \le& \max_r \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{u} \rangle^2 + \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v} \rangle^2 + \langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i \rangle^2 \le \tau. \end{aligned}$$

Otherwise, if $i \in \mathcal{W}$, we also have

$$\begin{aligned} F_y(\mathbf{x}_i; \mathbf{W}^{(t)}) =& \frac{1}{m} \sum_{r\in[m]} \Big[ \sigma(\langle \mathbf{w}_{y,r}^{(t)}, \widetilde{\boldsymbol{\xi}}_i \rangle) + \sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v} \rangle) + \sigma(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \Big] \\ \le& \max_r \langle \mathbf{w}_{y,r}^{(t)}, \widetilde{\boldsymbol{\xi}}_i \rangle^2 + \langle \mathbf{w}_{y,r}^{(t)}, y\mathbf{v} \rangle^2 + \langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i \rangle^2 \le \tau. \end{aligned}$$

Then it follows that $-y\ell^{(t)} \ge 1 - \tau$. □

### I.1 Stage 1. Learn Strong Signal Exponentially Fast

We will mainly track the maximal inner-product between $\mathbf{w}$ and signal vectors $\mathbf{v}, \mathbf{u}$,

$$\begin{aligned} \Psi^{(t)} =& \max_{j,r} |\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v} \rangle|, \quad \Phi^{(t)} = \max_{j,r} |\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u} \rangle|, \\ \Gamma_i^{(t)} =& \max_{j,r} |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle|, \quad i \in [n]; \\ \widetilde{\Gamma}_i^{(t)} =& \max_{j,r} |\langle \mathbf{w}_{j,r}^{(t)}, \widetilde{\boldsymbol{\xi}}_i \rangle|, \quad i \in \mathcal{W}. \end{aligned}$$

**Lemma 38.** *Ever since initialization, at least until time $T_+ := \frac{nm}{3\eta(4+\rho)\sigma_p^2 d}$, there still holds that*

$$\max_{i\in[n]} \Gamma_i^{(t)} \le \sigma_0 \sigma_p \sqrt{d}, \quad \max_{i\in\mathcal{W}} \widetilde{\Gamma}_i^{(t)} \le \sigma_0 \sigma_p \sqrt{d}. \tag{59}$$

*Proof.* For those inner products with noise vectors, $\forall i \in [n]$, the updating rules become

$$\begin{aligned} |\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle| \le& |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle| + \frac{\eta}{m} \cdot |\ell_{i_t}^{(t)}| \cdot \Big( \sigma'(\langle \mathbf{w}_{j,r}^{(t)} \cdot \boldsymbol{\xi}_{i_t} \rangle) \cdot |\langle \boldsymbol{\xi}_{i_t}, \boldsymbol{\xi}_i \rangle| \\ &+ \sigma'(\langle \mathbf{w}_{j,r}^{(t)} \cdot \widetilde{\boldsymbol{\xi}}_{i_t} \rangle) \cdot |\langle \widetilde{\boldsymbol{\xi}}_{i_t}, \boldsymbol{\xi}_i \rangle| \cdot \mathbf{1}\{i_t \in \mathcal{W}\} \Big) \\ \le& |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle| + \frac{6\eta}{m} \Big( |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_{i_t} \rangle| \cdot |\langle \boldsymbol{\xi}_{i_t}, \boldsymbol{\xi}_i \rangle| + |\langle \mathbf{w}_{j,r}^{(t)}, \widetilde{\boldsymbol{\xi}}_{i_t} \rangle| \cdot |\langle \widetilde{\boldsymbol{\xi}}_{i_t}, \boldsymbol{\xi}_i \rangle| \cdot \mathbf{1}\{i_t \in \mathcal{W}\} \Big). \end{aligned}$$

By Taking maximum over $r \in [m]$, we conclude that

$$\Gamma_i^{(t+1)} \le \Gamma_i^{(t)} + \frac{6\eta}{m} \Big( \Gamma_{i_t}^{(t)} |\langle \boldsymbol{\xi}_{i_t}, \boldsymbol{\xi}_i \rangle| + \widetilde{\Gamma}_{i_t}^{(t)} |\langle \widetilde{\boldsymbol{\xi}}_{i_t}, \boldsymbol{\xi}_i \rangle| \cdot \mathbf{1}\{i_t \in \mathcal{W}\} \Big), \quad \forall i \in [n].$$

Similarly, we also have,

$$\widetilde{\Gamma}_i^{(t+1)} \le \widetilde{\Gamma}_i^{(t)} + \frac{6\eta}{m} \Big( \Gamma_{i_t}^{(t)} |\langle \boldsymbol{\xi}_{i_t}, \widetilde{\boldsymbol{\xi}}_i \rangle| + \widetilde{\Gamma}_{i_t}^{(t)} |\langle \widetilde{\boldsymbol{\xi}}_{i_t}, \widetilde{\boldsymbol{\xi}}_i \rangle| \cdot \mathbf{1}\{i_t \in \mathcal{W}\} \Big), \quad \forall i \in \mathcal{W}.$$

We then use induction to rigorously prove our conclusion. Firstly, (59) holds at time $t = 0$. Now suppose that (59) holds until some $\widetilde{T} < T_+$. Fixing some $i \in [n]$,

$$\begin{aligned}\Gamma_i^{(\widetilde{T}+1)} \leq & \frac{6\eta\sigma_0\sigma_p\sqrt{d}}{m}\sum_{t\leq\widetilde{T}}\Big(|\langle\boldsymbol{\xi}_{i_t},\boldsymbol{\xi}_i\rangle| + |\langle\widetilde{\boldsymbol{\xi}}_{i_t},\boldsymbol{\xi}_i\rangle|\cdot\mathbf{1}\{i_t\in\mathcal{W}\}\Big)\\ \leq & \frac{6\eta\sigma_0\sigma_p\sqrt{d}}{m}\left(\frac{3\widetilde{T}\sigma_p^2 d}{2n} + 2\widetilde{T}(1+\rho)\sigma_p^2\sqrt{d\log(4n^2/p)}\right)\\ \leq & \frac{3\eta\sigma_0\sigma_p\sqrt{d}(4+\rho)\widetilde{T}\sigma_p^2 d}{nm} \leq \sigma_0\sigma_p\sqrt{d}.\end{aligned}$$

The first inequality is by induction hypothesis. The second inequality is because that there are at most $\widetilde{T}/n$ many $i_t$'s would equal $i$ and at most $\rho\widetilde{T}$ many $i_t$'s would be in $\mathcal{W}$, and we also use Lemma 9 to control the norm and correlations between noise vectors. The third inequality is by scaling $d \geq 16n^2\log(4n^2/p)$, while the last inequality is due to $\widetilde{T} < T_+$. Similarly, we can also control $\widetilde{\Gamma}_i^{(t+1)}$ for some fixed $i \in \mathcal{W}$,

$$\begin{aligned}\widetilde{\Gamma}_i^{(\widetilde{T}+1)} \leq & \frac{6\eta\sigma_0\sigma_p\sqrt{d}}{m}\sum_{t\leq\widetilde{T}}\Big(|\langle\boldsymbol{\xi}_{i_t},\widetilde{\boldsymbol{\xi}}_i\rangle| + |\langle\widetilde{\boldsymbol{\xi}}_{i_t},\widetilde{\boldsymbol{\xi}}_i\rangle|\cdot\mathbf{1}\{i_t\in\mathcal{W}\}\Big)\\ \leq & \frac{6\eta\sigma_0\sigma_p\sqrt{d}}{m}\left(\frac{3\widetilde{T}\sigma_p^2 d}{2n} + 2\widetilde{T}(1+\rho)\sigma_p^2\sqrt{d\log(4n^2/p)}\right) \leq \sigma_0\sigma_p\sqrt{d},\end{aligned}$$

where the second inequality is because there are at most $\widetilde{T}/n$ many $i_t$'s would equal $i \in \mathcal{W}$. In conclusion, (59) holds at least until $T_+$. □

In the following, we would take

$$\begin{aligned}\tau =& \max\left\{2\sigma_0\|\mathbf{u}\|\left(2\log(8m/p)\right)^{1/2-\|\mathbf{v}\|^2/4(\sigma_p^2 d)}, 1 - \frac{6\sqrt{2}\sigma_p^2 d}{\|\mathbf{u}\|^2\log(1/\sqrt{2\log(8m/p)})}\right\},\\ \iota =& 2\sigma_0\|\mathbf{u}\|\exp\left[\frac{(1-\tau)(1-\rho)\|\mathbf{u}\|^2}{3(4+\rho)\sigma_p^2 d}\right].\end{aligned}$$

By the conditions in Proposition 1 upon $\|\mathbf{v}\|^2/\|\mathbf{u}\|^2$, we find $\tau, \iota$ both constant in $(0, 1)$.

**Lemma 39.** *Under the same condition as Proposition 1, there exists time*

$$T^\dagger = \frac{4m}{\eta(1-\tau)(1-\rho)\|\mathbf{u}\|^2}\log\left(\frac{2\iota}{\sigma_0\|\mathbf{u}\|}\right),$$

*such that: (i) The model learns strong signal to a constant level,*

$$\max_r j\langle\mathbf{w}_{j,r}^{(T^\dagger)},\mathbf{u}\rangle \geq \iota, \quad \forall j\in\{\pm 1\}.$$

*(ii) Compared to random initialization, the model does not learn weak signal that much,* $\max_{j,r}|\langle\mathbf{w}_{j,r}^{(T^\dagger)},\mathbf{v}\rangle| \leq \sigma_0\|\mathbf{v}\|$.

*Proof.* Firstly, we would find $\Psi^{(t)}, \Phi^{(t)}$ having an exponentially growing upper bound. Recursively, we would have

$$\begin{aligned}\Psi^{(t+1)} \leq & \Psi^{(t)} + \max_{j,r}\left|\frac{jy\eta}{m}\cdot\big(f(\mathbf{x}_i;\mathbf{W}^{(t)}) - y\big)\cdot\sigma'(\langle\mathbf{w}_{j,r}^{(t)},y\mathbf{v}\rangle)\|\mathbf{v}\|^2\right|\\ = & \Psi^{(t)} + \frac{\eta}{m}\left|\ell_i^{(t)}\right|\|\mathbf{v}\|^2\cdot\max_{j,r}\sigma'(\langle\mathbf{w}_{j,r}^{(t)},y\mathbf{v}\rangle)\\ \leq & \Psi^{(t)} + \frac{2\eta}{m}\left|\ell_i^{(t)}\right|\|\mathbf{v}\|^2\Psi^{(t)} \leq \exp\left(\frac{6\eta\|\mathbf{v}\|^2}{m}\right)\Psi^{(t)}.\end{aligned}$$

Therefore, $\Psi^{(t)} \le \exp\left(\frac{6\eta\|\mathbf{v}\|^2 t}{m}\right)\Psi^{(0)} \le \exp\left(\frac{6\eta\|\mathbf{v}\|^2 t}{m}\right)\sqrt{2\log(8m/p)}\sigma_0\|\mathbf{v}\|$. It follows similarly that

$$\Phi^{(t)} \le \exp\left(\frac{6\eta(1-\rho)\|\mathbf{u}\|^2 t}{m}\right)\Phi^{(0)} \le \exp\left(\frac{6\eta(1-\rho)\|\mathbf{u}\|^2 t}{m}\right)\sqrt{2\log(8m/p)}\sigma_0\|\mathbf{u}\|.$$

The extra factor $1-\rho$ appears because only a $1-\rho$ proportion of data points would contain $\mathbf{u}$, and therefore contribute to evolution of $\Phi^{(t)}$. Note that growing rates of these two bounds differ a lot due to the different magnitudes of $(1-\rho)\|\mathbf{u}\|$ and $\|\mathbf{v}\|$.

Our subsequent analysis illustrates that $\Phi^{(t)}$ can grow into a constant-level magnitude since strong signal $\mathbf{u}$ is significant enough. We can track how well our model learns $\mathbf{u}$ by

$$A_1^{(t)} = \max_r \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle, \quad A_{-1}^{(t)} = \max_{r, i\notin\mathcal{W}} \langle \mathbf{w}_{-1,r}^{(t)}, -\mathbf{u}\rangle.$$

By definition, $A_1^{(t)}, A_{-1}^{(t)} \le \Phi^{(t)}$ also admits an exponentially upper bound. For a certain $\tau \in (0,1)$, derived from the previous exponential upper bound, $\max\{\Phi^{(t)}, \Psi^{(t)}\} \le \sqrt{\tau/3}$ remains true at least until

$$T_1 = \frac{m}{6\eta(1-\rho)\|\mathbf{u}\|^2}\log\left(\frac{\sqrt{\tau/2}}{\sigma_0\|\mathbf{u}\|\sqrt{2\log(8m/p)}}\right).$$

Moreover, since $(1-\rho)\|\mathbf{u}\|^2 \gg \sigma_p^2 d/n$, we also know $T_1 \le T_+$ where $T_+$ comes from Lemma 38 and therefore $\Gamma_i^{(t)} \le \sigma_0\sigma_p\sqrt{d} \le \sqrt{\tau/3}, \widetilde{\Gamma}_i^{(t)} \le \sigma_0\sigma_p\sqrt{d} \le \sqrt{\tau/3}$ until $t \le T_1$. Consequently, until at least $T_1$, we are able to use Lemma 37 to conclude $-y_{i_t}\ell_{i_t}^{(t)} \ge 1-\tau$, which enables lower bounding $A^{(t)}$.

The $i_t$-th sample would be used to update parameters, according to our multi-pass SGD setting. If $i_t \in \mathcal{W}$, then $\langle \mathbf{w}_{j,r}^{(t+1)}, \mathbf{u}\rangle = \langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u}\rangle$ holds for any $j, r$. If $i_t \notin \mathcal{W}$ but $y_{i_t} = -1$, then $\max_r \langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle = \max_r \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle \ge 0$ since that neuron will not be activated. Otherwise, only if $i_t \notin \mathcal{W}$ and $y_{i_t} = 1$, the updating rule becomes

$$\begin{aligned}\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle =& \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle + \frac{\eta}{m}\cdot\big(-y_{i_t}\ell_{i_t}^{(t)}\big)\cdot\sigma'(\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle)\|\mathbf{u}\|^2 \\ \ge& \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle + \frac{2\eta(1-\tau)\|\mathbf{u}\|^2}{m}\max\{\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle, 0\}.\end{aligned}$$

Take maximum over $r \in [m]$ to see

$$A_1^{(t+1)} \ge A_1^{(t)} + \frac{2\eta(1-\tau)\|\mathbf{u}\|^2}{m}A_1^{(t)} \ge \exp\left(\frac{\eta(1-\tau)\|\mathbf{u}\|^2}{m}\right)A_1^{(t)},$$

where the last equality is by $1+z \ge \exp(z/2)$ for any $0 \le z \le 2$. Consequently, when $t$ is large, we would have

$$\begin{aligned}A_1^{(t)} \ge& \exp\left(\frac{\eta(1-\tau)\|\mathbf{u}\|^2}{m}\sum_{t'\le t}\mathbf{1}\{i_{t'} \notin \mathcal{W}, y_{i_t} = 1\}\right)A_1^{(0)} \\ \ge& \exp\left(\frac{\eta(1-\tau)\|\mathbf{u}\|^2(1-\rho)t}{4m}\right)\sigma_0\|\mathbf{u}\|/2,\end{aligned}$$

at least until $t \le T_1$. We use the fact that $\sum_{t'\le t}\mathbf{1}\{i_{t'} \notin \mathcal{W}, y_{i_t} = 1\} \ge (1-\rho)t/4$ because the sample labels are balanced (Lemma 7) and $1-\rho$ proportion of samples come with the strong signal. In the same manner, we would have

$$A_{-1}^{(t)} \ge \exp\left(\frac{\eta(1-\tau)\|\mathbf{u}\|^2(1-\rho)t}{4m}\right)\sigma_0\|\mathbf{u}\|/2.$$

Define the time when $A_{\pm 1}^{(t)}$ both break $\iota$,

$$T_2 = \frac{4m}{\eta(1-\tau)(1-\rho)\|\mathbf{u}\|^2}\log\left(\frac{2\iota}{\sigma_0\|\mathbf{u}\|}\right) \le T_1$$

where the inequality is due to the scaling of $\iota$ upon $\tau$. Moreover, we also need $T_2 \leq T_+$, where $T_+$ is the time that $\langle \mathbf{w}_{j,r}^{(t)}, \widetilde{\boldsymbol{\xi}}_i \rangle, \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$ remains in $O(\sigma_0 \sigma_p \sqrt{d})$. And this requirement is also achieved by the selection of $\tau$ and $\iota$. Plugging $T_2$ into the exponential lower bound, we can conclude that $\Phi^{(T_2)} \geq A_{\pm 1}^{(T_2)} \geq \iota$ already grows up to a constant level magnitude by the time $T_2$. Lastly, plug the definition of $T_2$ to upper bound

$$\Psi^{(T_2)} \leq \exp\left( \frac{24\|\mathbf{v}\|^2}{(1-\tau)(1-\rho)\|\mathbf{u}\|^2} \log\left( \frac{2\iota}{\sigma_0 \|\mathbf{u}\|} \right) \right) \sqrt{2\log(8m/p)} \sigma_0 \|\mathbf{v}\| \leq \sigma_0 \|\mathbf{v}\|.$$

In conclusion, by taking $T^\dagger = T_2$, this lemma is completely proved. □

### I.2 Stage 2. Exploit Strong Signal

In the second stage, our lemmas would suggest that before the model really learns the weak signal $\mathbf{v}$ or memorizes any noise vector, the model already fits a proportion $1-\rho$ of data points by exploiting strong signal $\mathbf{u}$.

**Lemma 40.** *There exists time*

$$T = T^\dagger + \left\lfloor \frac{Cm^3}{2\eta\epsilon\|\mathbf{u}\|^2} \right\rfloor$$

*such that: (i) Average loss over iterations within this stage has decreased to* $2\epsilon$*,*

$$\frac{1}{2n} \sum_{i \in \mathcal{W}^c} \min_{T^\dagger \leq t \leq T} \left( y_i - f(\boldsymbol{x}_i; \mathbf{W}^{(t)}) \right)^2 \leq 3\epsilon.$$

*(ii) All through the training dynamics* $0 \leq t \leq T$*, there holds* $\max_{j,r} |\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v} \rangle| \leq \sigma_0 \|\mathbf{v}\|$*. (iii) All through the training dynamics* $0 \leq t \leq T$*, there holds*

$$\max_{j,r,i \in [n]} |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle| \leq \sigma_0 \sigma_p \sqrt{d}, \quad \max_{j,r,i \in \mathcal{W}} |\langle \mathbf{w}_{j,r}^{(t)}, \widetilde{\boldsymbol{\xi}}_i \rangle| \leq \sigma_0 \sigma_p \sqrt{d}.$$

In studying the second stage, we firstly identify when the upper bound on $\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v} \rangle$ breaks and find that the conclusions of Lemma 39 still holds before that time.

**Lemma 41.** *Take* $\eta \leq \frac{m}{6\|\mathbf{u}\|^2}$*. There exists a time*

$$T^\ddagger = \frac{m}{6\eta\|\mathbf{v}\|^2} \log\left( \frac{\sqrt{\tau/2}}{\sigma_0 \|\mathbf{v}\| \sqrt{2\log(8m/p)}} \right) \geq T^\dagger$$

*such that* (59)*,*

$$\max_{j,r} |\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v} \rangle| \leq \sigma_0 \|\mathbf{v}\|, \quad \max_r \langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u} \rangle \geq \iota/2, \quad \forall j \in \{\pm 1\},$$

*hold for any* $T^\dagger \leq t \leq T^\ddagger$*.*

*Proof.* Firstly, we need to adopt the exponential upper bound derived in proving Lemma 39,

$$\Psi^{(t)} \leq \exp\left( \frac{6\eta\|\mathbf{v}\|^2 t}{m} \right) \Psi^{(0)} \leq \exp\left( \frac{6\eta\|\mathbf{v}\|^2 t}{m} \right) \sqrt{2\log(8m/p)} \sigma_0 \|\mathbf{v}\|.$$

Then we naturally find that before $T^\ddagger$, it would always hold that $\max_{j,r} |\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v} \rangle| \leq \sigma_0 \|\mathbf{v}\|$. Due to the conditions on $\|\mathbf{u}\|/\|\mathbf{v}\|$, $T^\ddagger$ is found to be much larger than $T^\dagger$. Then we proceed to prove the other assertion by induction. At time $t = T^\dagger$, the lower bounds $\max_{j,r} \langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u} \rangle \geq \iota/2, j = \pm 1$ hold as a consequence of the previous lemma.

Suppose it holds until time $t$. If $i_t \in \mathcal{W}$, then $\langle \mathbf{w}_{j,r}^{(t+1)}, \mathbf{u} \rangle = \langle \mathbf{w}_{j,r}^{(t)}, \mathbf{u} \rangle$ holds for any $j, r$. If $i_t \notin \mathcal{W}$ but $y_{i_t} = -1$, then $\max_r \langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u} \rangle = \max_r \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u} \rangle \geq 0$ since that neuron will not be activated. Otherwise, if $i_t \notin \mathcal{W}$ and $y_{i_t} = 1$, restate the updating rule by

$$\langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u} \rangle = \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u} \rangle + \frac{\eta}{m} \cdot \left( 1 - y_{i_t} f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}) \right) \cdot \sigma'(\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u} \rangle) \|\mathbf{u}\|^2,$$

from which we find $\max_{r,i\notin\mathcal{W}}\langle \mathbf{w}_{y_i,r}^{(t+1)}, y_i\mathbf{u}\rangle \geq \max_{r,i\notin\mathcal{W}}\langle \mathbf{w}_{y_i,r}^{(t)}, y_i\mathbf{u}\rangle$ must hold if $y_{i_t}f(\mathbf{x}_{i_t};\mathbf{W}^{(t)})\leq 1$. Otherwise, once $y_{i_t}f(\mathbf{x}_{i_t};\mathbf{W}^{(t)})>1$, it immediately follows that

$$\begin{aligned}1<&y_{i_t}f(\mathbf{x}_{i_t};\mathbf{W}^{(t)}) = F_{y_{i_t}}(\mathbf{x}_{i_t};\mathbf{W}^{(t)}) - F_{-y_{i_t}}(\mathbf{x}_{i_t};\mathbf{W}^{(t)})\\ \leq& F_{y_{i_t}}(\mathbf{x}_{i_t};\mathbf{W}^{(t)}) = \frac{1}{m}\sum_{r\in[m]}\left[\sigma(\langle \mathbf{w}_{1,r}, y_{i_t}\mathbf{u}\rangle) + \sigma(\langle \mathbf{w}_{1,r}, y_{i_t}\mathbf{v}\rangle) + \sigma(\langle \mathbf{w}_{1,r}, \xi_{i_t}\rangle)\right]\\ \leq& \max_r \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle^2 + \sigma_0^2\|\mathbf{v}\|^2 + \sigma_0^2\sigma_p^2 d.\end{aligned}$$

Consequently, for the specific neuron $r^* = \mathrm{argmax}_r \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{u}\rangle^2$, there holds

$$\begin{aligned}\langle \mathbf{w}_{1,r^*}^{(t+1)}, \mathbf{u}\rangle \geq& \langle \mathbf{w}_{1,r^*}^{(t)}, \mathbf{u}\rangle - \frac{3\eta}{m}\langle \mathbf{w}_{1,r^*}^{(t)}, \mathbf{u}\rangle\|\mathbf{u}\|^2\\ \geq& \left(1-\sigma_0^2\|\mathbf{v}\|^2 - \sigma_0^2\sigma_p^2 d\right)\left(1-\frac{3\eta}{m}\|\mathbf{u}\|^2\right) \geq \frac{\iota}{2},\end{aligned}$$

where the last inequality is enabled by taking $\eta \leq \frac{m}{6\|\mathbf{u}\|^2}$ and $\sigma_0 \leq \sqrt{1-\iota}/\sqrt{\|\mathbf{v}\|^2+\sigma_p^2 d}$. Therefore, we find that $\max_r \langle \mathbf{w}_{1,r}^{(t+1)}, \mathbf{u}\rangle \geq \iota/2$ must hold no matter what $i_t$ is. In the same way, one can also obtain $\max_r \langle \mathbf{w}_{-1,r}^{(t+1)}, -\mathbf{u}\rangle \geq \iota/2$. In conclusion, the induction proof is complete. □

Our subsequently analysis confirms that even before $T^\ddagger$, the model can already fit those data points with strong signal by exploiting $\mathbf{u}$. For the given $0<\epsilon<1$, define a reference point $\mathbf{W}^*$ as

$$\mathbf{w}_{j,r}^* = \frac{4m(1+\epsilon)}{\iota}\cdot\frac{\mathbf{u}}{\|\mathbf{u}\|^2}, \quad j\in\{\pm 1\}, r\in[m]. \tag{60}$$

**Lemma 42.** *Under the same condition as the previous lemma, for all $T^\dagger \leq t \leq T^\ddagger$, there holds $y_i\left\langle \nabla f(\mathbf{x}_i;\mathbf{W}^{(t)}), \mathbf{W}^*\right\rangle \geq 2(1+\epsilon)$ for any $i\notin\mathcal{W}$.*

*Proof.* Recall that $f(\mathbf{x}_i;\mathbf{W}) = \sum_{j,r}\frac{j}{m}\left[\sigma(\langle \mathbf{w}_{j,r}, y_i\mathbf{u}\rangle) + \sigma(\langle \mathbf{w}_{j,r}, y_i\mathbf{v}\rangle) + \sigma(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}_i\rangle)\right]$ and $\mathbf{u}\perp \mathrm{span}(\mathbf{v}, \boldsymbol{\xi}_i)$, so we have

$$\begin{aligned}y_i\left\langle \nabla f(\mathbf{x}_i;\mathbf{W}^{(t)}), \mathbf{W}^*\right\rangle =& \frac{1}{m}\sum_{j,r}\sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, y_i\mathbf{u}\rangle\right)\langle \mathbf{w}_{j,r}^*, y_i\mathbf{u}\rangle\\ =& \sum_{j,r}\sigma'\left(\langle \mathbf{w}_{j,r}^{(t)}, y_i\mathbf{u}\rangle\right)\frac{4(1+\epsilon)}{\iota}\\ \geq& \max_r\langle \mathbf{w}_{y_i,r}^{(t)}, y_i\mathbf{u}\rangle\frac{4(1+\epsilon)}{\iota} \geq 2(1+\epsilon),\end{aligned}$$

where the last inequality is by $\max_r\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{u}\rangle \geq \iota/2$ for any $j\in\{\pm 1\}$ as shown by the previous lemma. □

**Lemma 43.** *Continued from the previous setting, for $T^\dagger \leq t \leq T^\ddagger$, if $i_t\notin\mathcal{W}$, there holds*

$$\|\mathbf{W}^{(t)}-\mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)}-\mathbf{W}^*\|_F^2 \geq 2\eta\left(f(\mathbf{x}_{i_t};\mathbf{W}^{(t)}) - y_{i_t}\right)^2 - 2\eta\epsilon^2.$$

*Proof.* Firstly expand the difference by

$$\begin{aligned}&\|\mathbf{W}^{(t)}-\mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)}-\mathbf{W}^*\|_F^2\\ =&2\eta\left\langle \ell_{i_t}^{(t)}\nabla f(\mathbf{x}_{i_t};\mathbf{W}^{(t)}), \mathbf{W}^{(t)}-\mathbf{W}^*\right\rangle - \eta^2\left|\ell_{i_t}^{(t)}\right|^2\|\nabla f(\mathbf{x}_{i_t};\mathbf{W}^{(t)})\|_F^2.\end{aligned} \tag{61}$$

Since the neural network $f(\mathbf{x};\mathbf{W})$ is 2-homogeneous in $\mathbf{W}$ due to the activation function $\sigma(z) = \max\{z,0\}^2$, we can have

$$\left\langle \nabla f(\mathbf{x};\mathbf{W}^{(t)}), \mathbf{W}^{(t)}\right\rangle = 2f(\mathbf{x};\mathbf{W}^{(t)}).$$

Stack these observations into the first term of previous difference expansion to obtain

$$\begin{aligned}
&\left\langle \ell_{i_t}^{(t)} \nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \right\rangle \\
=&\ell_{i_t}^{(t)} \left( 2f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}) - \left\langle \nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}), \mathbf{W}^* \right\rangle \right) \\
=&2\ell_{i_t}^{(t)} \left( f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}) - y_{i_t} \right) + \ell_{i_t}^{(t)} y_{i_t} \left( 2 - y_{i_t} \left\langle \nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}), \mathbf{W}^* \right\rangle \right).
\end{aligned}$$

Note that the first term is exactly $2\left(f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}) - y_{i_t}\right)^2$. As for the second term, since $i_t \notin \mathcal{W}$, we need to plug in Lemma 42 to see $2 - y_{i_t} \left\langle \nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}), \mathbf{W}^* \right\rangle \leq -2\epsilon < 0$, so that

$$\left| \ell_{i_t}^{(t)} y_{i_t} \left( 2 - y_{i_t} \left\langle \nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}), \mathbf{W}^* \right\rangle \right) \right| \leq \frac{1}{2} \ell_{i_t}^{(t)2} + 2\epsilon^2.$$

As a result, we would know $\left\langle \ell_{i_t}^{(t)} \nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \right\rangle \geq \frac{3}{2} \left(f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}) - y_{i_t}\right)^2 - 2\epsilon^2$. Next, an upper bound on the second order term $\eta^2 \|\nabla L(\mathbf{W}^{(t)})\|_F^2$ is given by

$$\begin{aligned}
&\eta^2 \left| \ell_{i_t}^{(t)} \right|^2 \|\nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)})\|_F^2 \\
=&\eta^2 \ell_{i_t}^{(t)2} \left[ \|\mathbf{u}\|^2 \sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{u} \rangle)^2 + \|\mathbf{v}\|^2 \sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{v} \rangle)^2 + \|\boldsymbol{\xi}_{i_t}\|^2 \sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_{i_t} \rangle)^2 \right] \\
\leq& O(\max\{\|\mathbf{u}\|^2, \|\mathbf{v}\|^2, \|\boldsymbol{\xi}_{i_t}\|^2\}) \cdot \eta^2 \ell_{i_t}^{(t)2},
\end{aligned}$$

since the dynamics of inner products $\langle \mathbf{w}_{j,r}, y\mathbf{u} \rangle, \langle \mathbf{w}_{j,r}, y\mathbf{v} \rangle, \langle \mathbf{w}_{j,r}, \boldsymbol{\xi}_i \rangle$ are well bounded by $O(1)$. Via scaling $\eta \cdot O(\max\{\|\mathbf{u}\|^2, \|\mathbf{v}\|^2, \|\boldsymbol{\xi}_{i_t}\|^2\}) \leq 1$, we would know $\eta^2 \left| \ell_{i_t}^{(t)} \right|^2 \|\nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)})\|_F^2 \leq \eta \ell_{i_t}^{(t)2}$. Eventually, continued from (61), we can completely prove this lemma. □

**Lemma 44.** *Continued from the previous setting, for $T^\dagger \leq t \leq T^\ddagger$, if $i_t \in \mathcal{W}$, there holds*

$$\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \geq -C\eta\sigma_0^2(\|\mathbf{v}\|^2 + \sigma_p^2 d). \tag{62}$$

*Proof.* Same as the last lemma, from the SGD setting, we have

$$\begin{aligned}
&\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \\
=&2\eta \left\langle \ell_{i_t}^{(t)} \nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \right\rangle - \eta^2 \left| \ell_{i_t}^{(t)} \right|^2 \|\nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)})\|_F^2;
\end{aligned} \tag{63}$$

from the 2-homogeneity, it follows that

$$\left\langle \nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}), \mathbf{W}^{(t)} \right\rangle = 2f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}).$$

Since $i_t \in \mathcal{W}$, every $\nabla_{\mathbf{w}_{j,r}} f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)})$ is in $\operatorname{span}\left(\mathbf{v}, \boldsymbol{\xi}_{i_t}, \widetilde{\boldsymbol{\xi}}_{i_t}\right) \perp \mathbf{u}$, so

$$\left\langle \nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}), \mathbf{W}^* \right\rangle = 0.$$

As a result, the first-order term in (63) can be bounded by

$$\begin{aligned}
&\left| 2\eta \left\langle \ell_{i_t}^{(t)} \nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \right\rangle \right| \\
=&4\eta \left| \left( f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}) - y_{i_t} \right) f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}) \right| \\
\leq&\frac{12\eta}{m} \left| \sum_{j,r} \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \widetilde{\boldsymbol{\xi}} \rangle) + \sum_{j,r} \sigma(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{v} \rangle) + \sum_{j,r} \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_{i_t} \rangle) \right| \\
\leq&O\left( \eta\sigma_0^2 \|\boldsymbol{v}\|^2 + \eta\sigma_0^2 \sigma_p^2 d \right).
\end{aligned}$$

We can also deal with the second-order term in (61) by

$$\begin{aligned}
&\eta^2 \left|\ell_{i_t}^{(t)}\right|^2 \|\nabla f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)})\|_F^2 \\
\leq &\eta^2 \ell_{i_t}^{(t)2} \left[ \|\widetilde{\boldsymbol{\xi}}_{i_t}\|^2 \sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \widetilde{\boldsymbol{\xi}}_{i_t} \rangle)^2 + \|\mathbf{v}\|^2 \sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y\mathbf{v} \rangle)^2 + \|\boldsymbol{\xi}_{i_t}\|^2 \sum_{j,r} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_{i_t} \rangle)^2 \right] \\
\leq &O(\eta^2(\sigma_0^4 \|\mathbf{v}\|^4 + \sigma_0^4 \sigma_p^4 d^2)),
\end{aligned}$$

where the last inequality is due to $\ell_{i_t}^{(t)}$ being $O(1)$. Since we already take $\eta \leq \frac{m}{6\|\mathbf{u}\|^2}$, the second-order term is ignorable compared to the first-order term. Therefore, we can conclude (62). □

*Proof of Lemma 40.* Continued from Lemmas 43 and 44, for any $t \geq T^\dagger$,

$$\begin{aligned}
&\frac{1}{t - T^\dagger + 1} \sum_{s=T^\dagger}^{t} \mathbf{1}\{i_s \notin \mathcal{W}\} \cdot \frac{1}{2} \left( f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}) - y_{i_t} \right)^2 \\
\leq &\frac{\|\mathbf{W}^{(T^\dagger)} - \mathbf{W}^*\|_F^2}{2\eta(t - T^\dagger + 1)} + \epsilon^2(1 - \rho) + C\sigma_0^2(\|\mathbf{v}\|^2 + \sigma_p^2 d)\rho.
\end{aligned}$$

Before proceeding to scale time $t$, it would be helpful to decompose $\|\mathbf{W}^{(T^\dagger)} - \mathbf{W}^*\|_F^2$ and have an upper bound,

$$\begin{aligned}
&\|\mathbf{W}^{(T^\dagger)} - \mathbf{W}^*\|_F^2 \\
= &\sum_{j,r} \frac{\left\langle \mathbf{w}_{j,r}^{(T^\dagger)} - \mathbf{w}_{j,r}^*, \mathbf{u} \right\rangle^2}{\|\mathbf{u}\|^2} + \frac{\left\langle \mathbf{w}_{j,r}^{(T^\dagger)} - \mathbf{w}_{j,r}^*, \mathbf{v} \right\rangle^2}{\|\mathbf{v}\|^2} + \left\| \mathbf{P}_{\boldsymbol{\xi}, \widetilde{\boldsymbol{\xi}}} (\mathbf{w}_{j,r}^{(T^\dagger)} - \mathbf{w}_{j,r}^*) \right\|^2 \\
&\quad + \left\| \left( \mathbf{I}_d - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2} - \frac{\mathbf{u}\mathbf{u}^\top}{\|\mathbf{u}\|_2^2} - \mathbf{P}_{\boldsymbol{\xi}, \widetilde{\boldsymbol{\xi}}} \right) (\mathbf{w}_{j,r}^{(T^\dagger)} - \mathbf{w}_{j,r}^*) \right\|^2 \\
\leq &\sum_{j,r} \frac{2\left\langle \mathbf{w}_{j,r}^{(T^\dagger)}, \mathbf{u} \right\rangle^2 + 2\left\langle \mathbf{w}_{j,r}^*, \mathbf{u} \right\rangle^2}{\|\mathbf{u}\|^2} + \frac{\left\langle \mathbf{w}_{j,r}^{(T^\dagger)}, \mathbf{v} \right\rangle^2}{\|\mathbf{v}\|^2} + \left\| \mathbf{P}_{\boldsymbol{\xi}, \widetilde{\boldsymbol{\xi}}} \mathbf{w}_{j,r}^{(T^\dagger)} \right\|^2 \\
&\quad + \left\| \left( \mathbf{I}_d - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2} - \frac{\mathbf{u}\mathbf{u}^\top}{\|\mathbf{u}\|_2^2} - \mathbf{P}_{\boldsymbol{\xi}, \widetilde{\boldsymbol{\xi}}} \right) (\mathbf{w}_{j,r}^{(T^\dagger)} - \mathbf{w}_{j,r}^*) \right\|^2,
\end{aligned}$$

where $\mathbf{P}_{\boldsymbol{\xi}, \widetilde{\boldsymbol{\xi}}}$ denotes the projection matrix onto linear space $\mathrm{span}\left(\boldsymbol{\xi}, \widetilde{\boldsymbol{\xi}}\right)$. In these derivations, we exploit the fact that $\mathbf{w}^*$ is parallel to $\mathbf{u}$, and the gradient steps only updates $\mathbf{w}$ along the directions of $\mathbf{u}, \mathbf{v}$. Recall that $\langle \mathbf{w}^{(T^\dagger)}, \mathbf{u} \rangle = \Omega(1)$, $\langle \mathbf{w}^{(T^\dagger)}, \mathbf{v} \rangle = O(\sigma_0 \|\mathbf{v}\|)$, $\|\mathbf{w}_{j,r}^{(0)}\| = O(\sigma_0 \sqrt{d})$ and $\langle \mathbf{w}^{(T^\dagger)}, \boldsymbol{\xi}_i \rangle = O(\sigma_0 \sigma_p \sqrt{d})$, the leading term would be $\frac{\left\langle \mathbf{w}_{j,r}^*, \mathbf{u} \right\rangle^2}{\|\mathbf{u}\|^2}$. Therefore, we would conclude that $\|\mathbf{W}^{(T^\dagger)} - \mathbf{W}^*\|_F^2 \leq Cm^3 / \|\mathbf{u}\|^2$. As a result, average loss after iterations $T^\dagger$ can be bounded by

$$\begin{aligned}
&\frac{1}{t - T^\dagger + 1} \sum_{s=T^\dagger}^{t} \mathbf{1}\{i_s \notin \mathcal{W}\} \cdot \frac{1}{2} \left( f(\mathbf{x}_{i_t}; \mathbf{W}^{(t)}) - y_{i_t} \right)^2 \\
\leq &\frac{Cm^3}{2\eta \|\mathbf{u}\|^2 (t - T^\dagger + 1)} + \epsilon^2(1 - \rho) + C\sigma_0^2(\|\mathbf{v}\|^2 + \sigma_p^2 d)\rho.
\end{aligned}$$

Then choose $T = T^\dagger + \left\lfloor \frac{Cm^3}{2\eta\epsilon \|\mathbf{u}\|^2} \right\rfloor$. Since $\frac{\|\mathbf{u}\|^2}{\|\mathbf{v}\|^2} \geq \widetilde{\Omega}(m^2)$, we can verify that $T \leq T^\ddagger$ where $T^\ddagger$ is given in Lemma 41 until when the weak signal cannot be fully learned. Moreover, we also have $T \leq T_+$ where $T_+$ is given in Lemma 38 when the noise is not memorized.

In conclusion, via scaling $\sigma_0^2 \leq \epsilon/C\rho(\|\mathbf{v}\|^2 + \sigma_p^2 d)$, the final output would be

$$\begin{aligned}
&\frac{1}{2n}\sum_{i\in\mathcal{W}^c}\min_{T^\dagger\leq t\leq T}\left(y_i - f(\boldsymbol{x}_i;\mathbf{W}^{(t)})\right)^2\\
\leq&\frac{1}{t-T^\dagger+1}\sum_{s=T^\dagger}^{t}\mathbf{1}\{i_s\notin\mathcal{W}\}\cdot\frac{1}{2}\left(f(\mathbf{x}_{i_t};\mathbf{W}^{(t)}) - y_{i_t}\right)^2\\
\leq&\epsilon + \epsilon^2(1-\rho) + C\sigma_0^2(\|\mathbf{v}\|^2+\sigma_p^2 d)\rho \leq 3\epsilon,
\end{aligned}$$

ending the proof. □

### I.3 Stage 3. Memorize Noise

After the second stage, the model already fits those data points with strong signal by exploiting $\mathbf{u}$. Subsequently, in the following third stage, residual $\ell_i^{(t)}, i\in\mathcal{W}^c$ would remain quite small, preventing the model from keep learning $\mathbf{u}$.

On the contrary, since $f(\mathbf{x}_i;\mathbf{W}^{(t)}) = O(\sigma_0^2)$ is still far from its label $y_i$ for each sample $i\in\mathcal{W}$ without the strong signal. Therefore, the weight vectors will still evolve in the directions perpendicular to $\mathbf{u}$. In Assumption 35, the ratio between weak signal $\mathbf{v}$ and pure noise $\sigma_p\sqrt{d}$ is scaled by

$$\frac{\|\mathbf{v}\|}{\sigma_p\sqrt{d}} \leq \widetilde{O}\left(\frac{1}{\sqrt{n}}\right).$$

Therefore, the model will eventually interpolates the whole dataset by memorizing noise vectors $(\widetilde{\boldsymbol{\xi}}_i, \boldsymbol{\xi}_i), i\in\mathcal{W}$, Define a reference point $\mathbf{W}^\star$ by

$$\mathbf{w}_{j,r}^\star = \mathbf{w}_{j,r}^* + \frac{4m(1+\epsilon')}{\iota}\left[\sum_{i\in\mathcal{W}}\mathbf{1}\{y_i=j\}\cdot\frac{\boldsymbol{\xi}_i}{\|\boldsymbol{\xi}_i\|} + \mathbf{1}\{y_i=j\}\cdot\frac{\widetilde{\boldsymbol{\xi}}_i}{\|\widetilde{\boldsymbol{\xi}}\|_i}\right],\quad j\in\{\pm1\}, r\in[m],$$

where $\mathbf{w}_{j,r}^*$ defined in (60) is the reference point used in the second stage. The following lemma is an adaptation of Theorem 4.4 of Cao et al. (2022) onto SGD with square loss.

**Lemma 45.** *Under the same setup as before, for some $\epsilon'\in(0,1)$, let*

$$T' = T + \left\lfloor\frac{\left\|\mathbf{W}^{(T)} - \mathbf{W}^\star\right\|_F^2}{2\eta\epsilon}\right\rfloor,$$

*where $T$ is the end of the second stage, Lemma 40. Then we would have $\max_{j,r}|\langle\mathbf{w}^{(t)},\mathbf{v}\rangle| \leq 2\sigma_0\|\mathbf{v}\|$ even until $t\leq T'$. But the whole dataset has already been interpolated during this interval,*

$$\frac{1}{2n}\sum_{i\in\mathcal{W}}\min_{T\leq t\leq T'}\left(y_i - f(\mathbf{x}_i;\mathbf{W}^{(t)})\right)^2 \leq 3\epsilon'.$$

*Proof Sketch.* As the closing stage of the training dynamics, the evolution during this interval is trivial based on all techniques developed in Sections G and I. Inner products $\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}_i\rangle, \langle\mathbf{w}_{j,r}^{(t)},\widetilde{\boldsymbol{\xi}}_i\rangle, i\in\mathcal{W}$ would firstly go through a substage in which they exponentially increase to a constant level. And then the model will fit all samples indexed by $\mathcal{W}$ by memorizing these noise vectors in polynomial time. All through this interval, $\max_{j,r}|\langle\mathbf{w}^{(t)},\mathbf{v}\rangle| \leq 2\sigma_0\|\mathbf{v}\|$ would stay $\Omega(\sigma_0\|\mathbf{v}\|)$ due to the scale of $\|\mathbf{v}\|/(\sigma_p\sqrt{d})$. A detailed proof is omitted here for readability. □

Combine Lemmas 39, 40 and 45 to obtain the full version of Theorem 36.