Benign Overfitting and Grokking in ReLU Networks for XOR Cluster Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Neural networks trained by gradient descent (GD) have exhibited a number of surprising generalization behaviors. First, they can achieve a perfect fit to noisy training data and still generalize near-optimally, showing that overfitting can sometimes be benign. Second, they can undergo a period of classical, harmful overfitting—achieving a perfect fit to training data with near-random performance on test data—before transitioning ("grokking") to near-optimal generalization later in training. In this work, we show that both of these phenomena provably occur in two-layer ReLU networks trained by GD on XOR cluster data where a constant fraction of the training labels are flipped. In this setting, we show that after the first step of GD, the network achieves 100% training accuracy, perfectly fitting the noisy labels in the training data, but achieves near-random test accuracy. At a later training step, the network achieves near-optimal test accuracy while still fitting the random labels in the training data, exhibiting a "grokking" phenomenon. This provides the first theoretical result of benign overfitting in neural network classification when the data distribution is not linearly separable. Our proofs rely on analyzing the feature learning process under GD, which reveals that the network implements a non-generalizable linear classifier after one step and gradually learns generalizable features in later steps.

1 Introduction

Classical wisdom in machine learning regards overfitting to noisy training data as harmful for generalization, and regularization techniques such as early stopping have been developed to prevent overfitting. However, modern neural networks can exhibit a number of counterintuitive phenomena that contravene this classical wisdom. Two intriguing phenomena that have attracted significant attention in recent years are *benign overfitting* (Bartlett et al., 2020) and *grokking* (Power et al., 2022):

- **Benign overfitting:** A model perfectly fits noisily labeled training data, but still achieves near-optimal test error.
- **Grokking:** A model initially achieves perfect training accuracy but no generalization (i.e. no better than a random predictor), and upon further training, transitions to almost perfect generalization.

Recent theoretical work has established benign overfitting in a variety of settings, including linear regression (Hastie et al., 2019; Bartlett et al., 2020), linear classification (Chatterji & Long, 2021a; Wang & Thrampoulidis, 2021), kernel methods (Belkin et al., 2019; Liang & Rakhlin, 2020), and neural network classification (Frei et al., 2022b; Kou et al., 2023). However, existing results of benign overfitting in neural network classification settings are restricted to linearly separable data distributions, leaving open the question of how benign overfitting can occur in fully non-linear

Submitted to the Mathematics of Modern Machine Learning Workshop at NeurIPS 2023. Do not distribute.

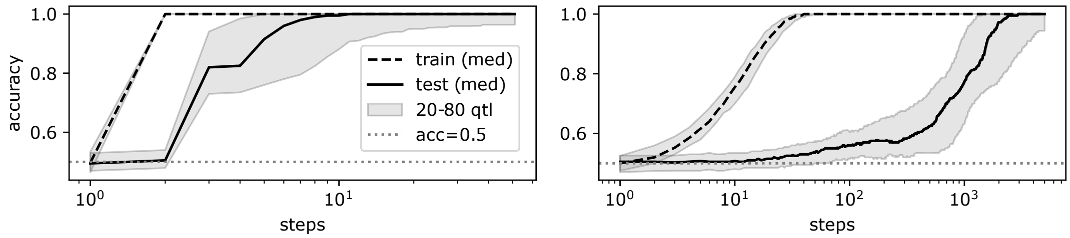

Figure 1: Comparing train and test accuracies of a two-layer neural network (2.1) trained on noisily labeled XOR data over 100 independent runs. *Left/right panel* shows benign overfitting and grokking when the step size is larger/smaller compared to the weight initialization scale. For plotting the x-axis, we add 1 to time so that the initialization $t = 0$ can be shown in log scale. See Appendix A.7 for details of the experimental setup.

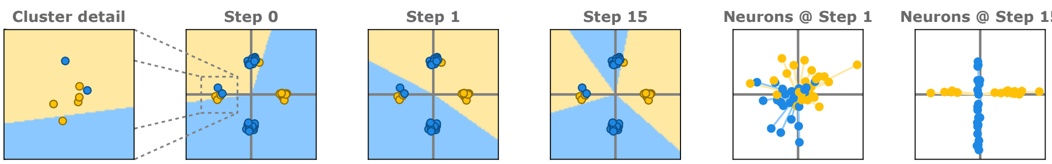

Figure 2: *Left four panels*: 2-dimensional projection of the noisily labeled XOR cluster data (Definition 2.1) and the decision boundary of the neural network (2.1) classifier restricted to the subspace spanned by the cluster means at times $t = 0, 1$ and 15. *Right two panels*: 2-dimensional projection of the neuron weights plotted at times $t = 1$ and 15.

settings. For grokking, several recent papers (Nanda et al., 2023; Gromov, 2023; Varma et al., 2023) have proposed explanations, but to the best of our knowledge, no prior work has established a rigorous proof of grokking in a neural network setting.

In this work, we characterize a setting in which both benign overfitting and grokking provably occur. We consider a two-layer ReLU network trained by gradient descent on a binary classification task defined by an XOR cluster data distribution (Figure 2). Specifically, datapoints from the positive class are drawn from a mixture of two high-dimensional Gaussian distributions $\frac{1}{2}N(\mu_1, I) + \frac{1}{2}N(-\mu_1, I)$, and datapoints from the negative class are drawn from $\frac{1}{2}N(\mu_2, I) + \frac{1}{2}N(-\mu_2, I)$, where μ_1 and μ_2 are orthogonal vectors. We then allow a constant fraction of the labels to be flipped. In this setting, we rigorously prove the following results: (i) **One-step catastrophic overfitting:** After one gradient descent step, the network perfectly fits every single training datapoint (no matter if it has a clean or flipped label), but has test accuracy close to 50%, performing no better than random guessing. (ii) **Grokking and benign overfitting:** After training for more steps, the network undergoes a "grokking" period from catastrophic to benign overfitting—it eventually reaches near 100% test accuracy, while maintaining 100% training accuracy the whole time. This behavior can be seen in Figure 1, where we also see that with a smaller step size the same grokking phenomenon occurs but with a delayed time for both overfitting and generalization.

Our results provide the first theoretical characterization of benign overfitting in a truly non-linear setting involving training a neural network on a non-linearly separable distribution. Interestingly, prior work on benign overfitting in neural networks for linearly separable distributions (Frei et al., 2022b; Cao et al., 2022; Xu & Gu, 2023; Kou et al., 2023) have not shown a time separation between catastrophic overfitting and generalization, which suggests that the XOR cluster data setting is fundamentally different.

2 Preliminaries

2.1 Notation

For a vector x, denote its Euclidean norm by $\|x\|$. Denote the sign of a scalar x by $\text{sgn}(x)$. Denote by $\sum_j q_j N(\mu_j, \Sigma_j)$ a mixture of Gaussian distributions, namely, with probability q_j, the sample is generated from $N(\mu_j, \Sigma_j)$. For a finite set $\mathcal{A} = \{a_i\}_{i=1}^n$, denote the uniform distribution on $\mathcal{A}$

by Unif$\mathcal{A}$. For an integer $d \geq 1$, denote the set $\{1, \cdots, d\}$ by $[d]$. For a finite set $\mathcal{A}$, let $|\mathcal{A}|$ be its cardinality. We use $\{\pm\mu\}$ to represent the set $\{+\mu, -\mu\}$. For two positive sequences $\{x_n\}, \{y_n\}$, we say $x_n = O(y_n)$ (respectively $x_n = \Omega(y_n)$), if there exists a universal constant $C > 0$ such that $x_n \leq Cy_n$ (respectively $x_n \geq Cy_n$) for all n. We say $x_n = \Theta(y_n)$ if $x_n = O(y_n)$ and $y_n = O(x_n)$.

2.2 Data Generation Setting

Let $\mu_1, \mu_2 \in \mathbb{R}^p$ be two orthogonal vectors, i.e. $\mu_1^\top \mu_2 = 0$.[1] Let $\eta \in [0, 1/2)$ be the label flipping probability.

Definition 2.1 (XOR cluster data)**.** Define P_{clean} as the distribution over the space $\mathbb{R}^p \times \{\pm 1\}$ of labelled data such that a datapoint $(x, \widetilde{y}) \sim P_{\text{clean}}$ is generated according to the following procedure: First, sample the label $\widetilde{y} \sim \text{Unif}\{\pm 1\}$. Second, generate x as follows: if $\widetilde{y} = 1$, then $x \sim \frac{1}{2}N(+\mu_1, I_p) + \frac{1}{2}N(-\mu_1, I_p)$; if $\widetilde{y} = -1$, then $x \sim \frac{1}{2}N(+\mu_2, I_p) + \frac{1}{2}N(-\mu_2, I_p)$. Define P to be the distribution over $\mathbb{R}^p \times \{\pm 1\}$ which is the η-noise-corrupted version of P_{clean}, namely: to generate a sample $(x, y) \sim P$, first generate $(x, \widetilde{y}) \sim P_{\text{clean}}$, and then let $y = \widetilde{y}$ with probability $1 - \eta$, and $y = -\widetilde{y}$ with probability η.

We consider n training datapoints $\{(x_i, y_i)\}_{i=1}^n$ generated i.i.d from the distribution P. We assume the sample size n to be sufficiently large (i.e., larger than any universal constant appearing in this paper). For simplicity, we assume $\|\mu_1\| = \|\mu_2\|$, omit the subscripts and denote them by $\|\mu\|$.

2.3 Neural Network, Loss Function, and Training Procedure

We consider a two-layer neural network of width m of the form

$$f(x; W) := \sum_{j=1}^{m} a_j \phi(\langle w_j, x \rangle), \tag{2.1}$$

where $w_1, \ldots, w_m \in \mathbb{R}^p$ are the first-layer weights, $a_1, \ldots, a_m \in \mathbb{R}$ are the second-layer weights, and the activation $\phi(z) := \max\{0, z\}$ is the ReLU function. We denote $W = [w_1, \ldots, w_m] \in \mathbb{R}^{p \times m}$ and $a = [a_1, \ldots, a_m]^\top \in \mathbb{R}^m$. We assume the second-layer weights are sampled according to $a_j \overset{\text{i.i.d.}}{\sim} \text{Unif}\{\pm \frac{1}{\sqrt{m}}\}$ and are fixed during the training process.

We define the empirical risk using the logistic loss function $\ell(z) = \log(1 + \exp(-z))$:

$$\widehat{L}(W) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i f(x_i; W)).$$

We use gradient descent (GD) $W^{(t+1)} = W^{(t)} - \alpha \nabla \widehat{L}\left(W^{(t)}\right)$ to update the first-layer weight matrix W, where α is the step size. Specifically, at time $t = 0$ we randomly initialize the weights by $w_j^{(0)} \overset{\text{i.i.d.}}{\sim} N\big(0, \omega_{\text{init}}^2 I_p\big), j \in [m]$, where ω_{init}^2 is the initialization variance; at each time step $t = 0, 1, 2, \ldots$, the GD update can be calculated as

$$w_j^{(t+1)} - w_j^{(t)} = -\alpha \frac{\partial \widehat{L}(W^{(t)})}{\partial w_j} = \frac{\alpha a_j}{n} \sum_{i=1}^{n} g_i^{(t)} \phi'(\langle w_j^{(t)}, x_i \rangle) y_i x_i, \quad j \in [m], \tag{2.2}$$

where $g_i^{(t)} := -\ell'(y_i f(x_i; W^{(t)}))$.

3 Main Results

Given a large enough universal constant C, we make the following assumptions:

(A1) The norm of the mean satisfies $\|\mu\|^2 \geq C n^{0.51} \sqrt{p}$.

(A2) The dimension of the feature space satisfies $p \geq C n^2 \|\mu\|^2$.

[1]Our results hold when μ_1 and μ_2 are near-orthogonal. We assume exact orthogonality for ease of presentation.

(A3) The noise rate satisfies $\eta \leq 1/C$.

(A4) The step size satisfies $\alpha \leq 1/(Cnp)$.

(A5) The initialization variance satisfies $\omega_{\text{init}}\, nm^{3/2}p \leq \alpha\|\mu\|^2$.

(A6) The number of neurons satisfies $m \geq Cn^{0.02}$.

Assumption (A1) concerns the signal-to-noise ratio (SNR) in the distribution, where the order 0.51 can be extended to any constant strictly larger than $\frac{1}{2}$. The assumption of high-dimensionality (A2) is important for enabling benign overfitting, and implies that the training datapoints are near-orthogonal. For a given n, these two assumptions are simultaneously satisfied if $\|\mu\| = \Theta(p^\beta)$ where $\beta \in (\frac{1}{4}, \frac{1}{2})$ and p is a sufficiently large polynomial in n. Assumption (A3) ensures that the label noise rate is at most a constant. While Assumption (A4) ensures the step size is small enough to allow for a variant of smoothness between different steps, Assumption (A5) ensures that the step size is large relative to the initialization scale so that the behavior of the network after a single step of GD is significantly different from that at random initialization. Assumption (A6) ensures the number of neurons is large enough to allow for concentration arguments at random initialization.

With these assumptions in place, we can state our main theorem which characterizes the training error and test error of the neural network at different times during the training trajectory.

Theorem 3.1. *Suppose that Assumptions (A1)-(A6) hold. With probability at least* $1 - n^{-\Omega(1)} - O(1/\sqrt{m})$ *over the random data generation and initialization of the weights, we have:*

- *The classifier* $\operatorname{sgn}(f(x; W^{(t)}))$ *can correctly classify all training datapoints for* $1 \leq t \leq \sqrt{n}$*:*
$$y_i = \operatorname{sgn}(f(x_i; W^{(t)})), \quad \forall i \in [n].$$
- *The classifier* $\operatorname{sgn}(f(x; W^{(t)}))$ *has near-random test error at* $t = 1$*:*
$$\tfrac{1}{2}(1 - n^{-\Omega(1)}) \leq \mathbb{P}_{(x,y)\sim P_{\mathit{clean}}}(y \neq \operatorname{sgn}(f(x; W^{(1)}))) \leq \tfrac{1}{2}(1 + n^{-\Omega(1)}).$$
- *The classifier* $\operatorname{sgn}(f(x; W^{(t)}))$ *generalizes when* $Cn^{0.01} \leq t \leq \sqrt{n}$*:*
$$\mathbb{P}_{(x,y)\sim P_{\mathit{clean}}}(y \neq \operatorname{sgn}(f(x; W^{(t)}))) \leq \exp(-\Omega(n^{0.99}\|\mu\|^4/p)) = \exp(-\Omega(n^{2.01})).$$

Theorem 3.1 shows that at time $t = 1$, the network achieves 100% training accuracy despite the constant fraction of flipped labels in the training data. The second part of the theorem shows that this overfitting is catastrophic as the test error is close to that of a random guess. On the other hand, by the first and third parts of the theorem, as long as the time step t satisfies $Cn^{0.01} \leq t \leq \sqrt{n}$, the network continues to overfit to the training data while simultaneously achieving test error $\exp(-\Omega(n^{2.01}))$, which guarantees a near-zero test error for large n. In particular, the network exhibits benign overfitting, and it achieves this by grokking. Notably, Theorem 3.1 is the first guarantee for benign overfitting in neural network classification for a nonlinear data distribution, in contrast to prior works which required linearly separable distributions (Frei et al., 2022b, 2023a; Cao et al., 2022; Xu & Gu, 2023; Kou et al., 2023; Kornowski et al., 2023). In Appendix A.1, we provide an overview of the key ingredients to the proof of Theorem 3.1.

4 Discussion

We have shown that two-layer neural networks trained on XOR cluster data with random label noise by GD reveal both benign overfitting and grokking. There are a few natural questions for future research. First, our analysis requires an upper bound on the number of training steps due to technical reasons; it is intriguing to understand the generalization behavior as time grows to infinity. Second, our proof crucially relies upon the assumption that the training data are nearly-orthogonal which requires that the ambient dimension is large relative to the number of samples. Prior work has shown with experiments that overfitting is less benign in this setting when the dimension is small relative to the number of samples (Frei et al., 2022a, Fig. 2); a precise characterization of the effect of high-dimensional data on generalization remains open.

References

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Yuan Cao, Zixiang Chen, Mikhail Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *arXiv preprint arXiv:2202.06526*, 2022.

Niladri S. Chatterji and Philip M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021a.

Niladri S. Chatterji and Philip M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021b. URL http://jmlr.org/papers/v22/20-974.html.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Spencer Frei, Niladri S Chatterji, and Peter L Bartlett. Random feature amplification: Feature learning and generalization in neural networks. *Preprint, arXiv:2202.07626*, 2022a.

Spencer Frei, Niladri S. Chatterji, and Peter L. Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory (COLT)*, 2022b.

Spencer Frei, Gal Vardi, Peter L. Bartlett, and Nathan Srebro. Benign overfitting in linear classifiers and leaky relu networks from kkt conditions for margin maximization. In *Conference on Learning Theory (COLT)*, 2023a.

Spencer Frei, Gal Vardi, Peter L. Bartlett, Nathan Srebro, and Wei Hu. Implicit bias in leaky relu networks trained on high-dimensional data. In *International Conference on Learning Representations*, 2023b.

Andrey Gromov. Grokking modular arithmetic. *Preprint, arXiv:2301.02679*, 2023.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Guy Kornowski, Gilad Yehudai, and Ohad Shamir. From tempered to benign overfitting in relu neural networks. *Preprint, arXiv:2305.15141*, 2023.

Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting for two-layer relu convolutional networks. In *International Conference on Machine Learning (ICML)*, 2023.

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

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *Preprint, arXiv:2301.05217*, 2023.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *Preprint, arXiv:2201.02177*, 2022.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *Preprint, arXiv:2309.02390*, 2023.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Ke Wang and Christos Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting and regularization. *Preprint, arXiv:2011.09148*, 2021.

Xingyu Xu and Yuantao Gu. Benign overfitting of non-smooth neural networks beyond lazy training. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11094–11117. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/xu23k.html.

A Appendix organization

A.1 Proof Sketch 7
A.1.1 Additional Notations 7
A.1.2 Properties of the Training Data and Random Initialization 8
A.1.3 Proof Sketch for Theorem 3.1 9
A.2 Properties of the training data 11
A.2.1 Proof of Lemma A.1 11
A.2.2 Proof of Corollary A.2 14
A.3 Properties of the initial weights and activation patterns 14
A.3.1 Proof of Lemma A.3 14
A.3.2 Proof of Lemma A.4 15
A.3.3 Proof of the Probability bound of the “Good run" event 20
A.4 Trajectory Analysis of the Neurons 20
A.4.1 Proof of Lemma A.11 21
A.4.2 Proof of Lemma A.12 22
A.4.3 Proof of Corollary A.13 23
A.4.4 Proof of Lemma A.14 24
A.4.5 Proof of Lemma A.15 26
A.5 Proof of the Main Theorem 29
A.5.1 Proof of Theorem A.16: 1-step Overfitting 29
A.5.2 Proof of Theorem A.8: Generalization 30
A.5.3 Proof of Theorem A.21: 1-step Test Accuracy 34
A.6 Probability Lemmas 40
A.7 Experimental details 42

A.1 Proof Sketch

A.1.1 Additional Notations

We first introduce some additional notation. For a matrix X, denote its Frobenius norm by $\|X\|_F$ and its spectral norm by $\|X\|$. Denote the indicator function by $\mathbb{I}(\cdot)$. Denote the cosine similarity of two vectors u, v by $\text{cossim}(u, v) := \frac{\langle u,v \rangle}{\|u\|\|v\|}$. For a random variable X, denote its expectation by $\mathbb{E}[X]$. Denote the c.d.f of standard normal distribution by $\Phi(\cdot)$ and the p.d.f. of standard normal distribution by $\Phi'(\cdot)$. Denote $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$. Denote the Bernoulli distribution which takes 1 with probability $p \in (0, 1)$ by $\text{Bern}(p)$. Denote the Binomial distribution with size n and probability p by $\text{B}(n, p)$. For a random variable X, denote its variance by $\text{Var}(X)$; and its absolute third central moment by $\rho(X)$. For $i \in [n]$, let $\bar{x}_i \in \mathsf{centers} = \{\pm\mu_1, \pm\mu_2\}$ be the mean of the Gaussian from which the sample (x_i, y_i) is drawn from. For each $\nu \in \mathsf{centers}$, define $\mathcal{I}_\nu = \{i \in [n] : \bar{x}_i = \nu\}$, i.e., the set of indices i such that x_i belongs to the cluster centered at ν. Thus, $\{\mathcal{I}_\nu\}_{\nu \in \mathsf{centers}}$ is a partition of $[n]$. Moreover, define $\mathcal{C} = \{i \in [n] : y_i = \widetilde{y}_i\}$ and $\mathcal{N} = \{i \in [n] : y_i \neq \widetilde{y}_i\}$ to be the set of clean and noisy samples, respectively. Further we define for each $\nu \in \mathsf{centers}$ the following sets:

$$\mathcal{C}_\nu := \mathcal{C} \cap \mathcal{I}_\nu \quad \text{and} \quad \mathcal{N}_\nu := \mathcal{N} \cap \mathcal{I}_\nu.$$

Let $c_\nu = |\mathcal{C}_\nu|$ and $n_\nu = |\mathcal{N}_\nu|$. Define the training input data matrix $X = [x_1, \ldots, x_n]^\top$. Let $\varepsilon \in (0, 10^{-3}/4)$ be a universal constant.

In Appendix A.1.2, we present several properties satisfied with high probability by the training data and random initialization, which are crucial in our proof. In Appendix A.1.3, we outline the major steps in the proof of Theorem 3.1.

A.1.2 Properties of the Training Data and Random Initialization

Lemma A.1 (Properties of training data)**.** *Suppose Assumptions (A1) and (A2) hold. Let the training data $\{(x_i, y_i)\}_{i=1}^n$ be sampled i.i.d from P as in Definition 2.1. With probability at least $1 - O(n^{-\varepsilon})$ the training data satisfy properties (E1)-(E4) defined below.*

(E1) *For all $k \in [n]$, $\max_{\nu \in \text{centers}} \langle x_k - \bar{x}_k, \nu \rangle \leq 10\sqrt{\log n}\|\mu\|$ and $|\|x_k\|^2 - p - \|\mu\|^2| \leq 10\sqrt{p \log n}$,*

(E2) *For each $i, k \in [n]$ such that $i \neq k$, we have $|\langle x_i, x_k \rangle - \langle \bar{x}_i, \bar{x}_k \rangle| \leq 10\sqrt{p \log n}$,*

(E3) *For $\nu \in$ centers, we have $|c_\nu + n_\nu - n/4| \leq \sqrt{\varepsilon n \log n}$ and $|n_\nu - \eta n| \leq \eta\sqrt{\varepsilon n \log n}$.*

(E4) *For $\nu \in$ centers, we have $|c_\nu + n_\nu - c_{-\nu} - n_{-\nu}| \geq n^{1/2-\varepsilon}$ and $|n_\nu - n_{-\nu}| \geq \eta n^{1/2-\varepsilon}$.*

Denote by $\mathcal{G}_{data}$ the set of training data satisfying conditions (E1)-(E4)*. Thus, the result can be stated succinctly as $\mathbb{P}(X \in \mathcal{G}_{data}) \geq 1 - O(n^{-\varepsilon})$.*

The proof of Lemma A.1 can be found in Appendix A.2.1. Conditions (E1) and (E2) are essentially the same as Frei et al. (2022b, Lemma 4.3) or Chatterji & Long (2021b, Lemma 10). Conditions (E3) and (E4) concern the number of clean and noisy examples in each cluster, and can be proved by concentration and anti-concentration arguments, respectively.

Lemma A.1 has an important corollary.

Corollary A.2 (Near-orthogonality of training data)**.** *Suppose Assumptions (A1), (A2), and Conditions (E1), (E2) from Lemma A.1 all hold. Then*

$$|\text{cossim}(x_i, x_k)| \leq \frac{2}{Cn^2}$$

for all $1 \leq i \neq k \leq n$.

This near-orthogonality comes from the high dimensionality of the feature space (i.e., Assumption (A2)) and will be crucially used throughout the proofs on optimization and generalization of the network. The proof of Corollary A.2 can be found in Appendix A.2.1.

Next, we divide the neuron indices into two sets according to the sign of the corresponding second-layer weight:

$$\mathcal{J}_{\texttt{Pos}} := \{j \in [m] : a_j > 0\}; \quad \mathcal{J}_{\texttt{Neg}} := \{j \in [m] : a_j < 0\}.$$

We will conveniently call them positive and negative neurons. Our next lemma shows that some properties of the random initialization hold with a large probability. The proof details can be found in Appendix A.3.1.

Lemma A.3 (Properties of the random weight initialization)**.** *Suppose Assumptions (A2) and (A6) hold. The followings hold with probability at least $1 - O(n^{-\varepsilon})$ over the random initialization:*

(D1) $\left\|W^{(0)}\right\|_F^2 \leq \frac{3}{2}\omega_{init}^2 mp$, *and* *(D2)* $|\mathcal{J}_{\texttt{Pos}}| \geq m/3$ *and* $|\mathcal{J}_{\texttt{Neg}}| \geq m/3$.

Denote the set of $W^{(0)}$ satisfying condition (D1) by $\mathcal{G}_W$. Denote the set of $a = (a_j)_{j=1}^m$ satisfying condition (D2) by $\mathcal{G}_A$. Then $\mathbb{P}(a \in \mathcal{G}_A, W^{(0)} \in \mathcal{G}_W) \geq 1 - O(n^{-\varepsilon})$.

We say that the sample i *activates* neuron j at time t if $\langle w_j^{(t)}, x_i \rangle > 0$. Now, for each neuron $j \in [m]$, time $t \geq 0$ and $\nu \in$ centers, define the set of indices i of samples x_i with clean (resp. noisy) labels from the cluster centered at ν that activates neuron j at time t:

$$\mathcal{C}_{\nu,j}^{(t)} := \{i \in \mathcal{C}_\nu : \langle w_j^{(t)}, x_i \rangle > 0\} \quad \text{(resp. } \mathcal{N}_{\nu,j}^{(t)} := \{i \in \mathcal{N}_\nu : \langle w_j^{(t)}, x_i \rangle > 0\}). \tag{A.1}$$

Moreover, we define

$$d_{\nu,j}^{(t)} := |\mathcal{C}_{\nu,j}^{(t)}| - |\mathcal{N}_{\nu,j}^{(t)}|, \quad \text{and} \quad D_{\nu,j}^{(t)} := d_{\nu,j}^{(t)} - d_{-\nu,j}^{(t)}.$$

For $\kappa \in [0, 1/2)$ and $\nu \in \texttt{centers}$, a neuron j is said to be (ν, κ)-*aligned* if

$$D_{\nu,j}^{(0)} > n^{1/2-\kappa}, \quad \text{and} \quad \max\{d_{-\nu,j}^{(0)}, d_{\nu,j}^{(0)}\} < \min\{c_\nu, c_{-\nu}\} - 2(n_{+\nu} + n_{-\nu}) - \sqrt{n} \tag{A.2}$$

The first condition ensures that at initialization, there are at least $n^{1/2-\kappa}$ many more samples from cluster ν activating the j-th neuron than from cluster $-\nu$ after accounting for cancellations from the noisy labels. The second is a technical condition necessary for trajectory analysis. A neuron j is said to be $(\pm\nu, \kappa)$-*aligned* if it is either (ν, κ)-aligned or $(-\nu, \kappa)$-aligned.

Lemma A.4 (Properties of the interaction between training data and initial weights)**.** *Suppose Assumptions (A1)-(A3) and (A6) hold. Given $a \in \mathcal{G}_A, X \in \mathcal{G}_{data}$, the followings hold with probability at least $1 - O(n^{-\varepsilon})$ over the random initialization $W^{(0)}$:*

(B1) For all $i \in [n]$, the sample x_i activates a large proportion of positive and negative neurons, i.e., $|\{j \in \mathcal{J}_{\texttt{Pos}} : \langle w_j^{(0)}, x_i\rangle > 0\}| \geq m/7$ and $|\{j \in \mathcal{J}_{\texttt{Neg}} : \langle w_j^{(0)}, x_i\rangle > 0\}| \geq m/7$ both hold.

(B2) For all $\nu \in \texttt{centers}$ and $\kappa \in [0, \frac{1}{2})$, both $|\{j \in \mathcal{J}_{\texttt{Pos}} : j$ is (ν, κ)-aligned$\}| \geq mn^{-10\varepsilon}$, and $|\{j \in \mathcal{J}_{\texttt{Neg}} : j$ is (ν, κ)-aligned$\}| \geq mn^{-10\varepsilon}$.

(B3) For all $\nu \in \texttt{centers}$, we have $\big|\{j \in \mathcal{J}_{\texttt{Pos}} : j$ is $(\pm\nu, 20\varepsilon)$-aligned$\}\big| \geq (1 - 10n^{-20\varepsilon})|\mathcal{J}_{\texttt{Pos}}|$. Moreover, the same statement holds if "$\mathcal{J}_{\texttt{Pos}}$" is replaced with "$\mathcal{J}_{\texttt{Neg}}$" everywhere.

(B4) For all $\nu \in \texttt{centers}$ and $\kappa \in [0, \frac{1}{2})$, let $\mathcal{J}_{\nu,\texttt{Pos}}^{\kappa} := \{j \in \mathcal{J}_{\texttt{Pos}} : j$ is (ν, κ)-aligned$\}$. Then $\sum_{j \in \mathcal{J}_{\nu,\texttt{Pos}}^{\kappa}} (c_\nu - n_\nu - d_{-\nu,j}^{(0)}) \geq \frac{n}{10}|\mathcal{J}_{\nu,\texttt{Pos}}^{\kappa}|$. Moreover, the same statement holds if "$\mathcal{J}_{\texttt{Pos}}$" is replaced with "$\mathcal{J}_{\texttt{Neg}}$" everywhere.

Condition (B1) makes sure that the neurons spread uniformly at initialization so that each datapoint activates at least a constant fraction of positive and negative neurons. Condition (B2) guarantees that for each $\nu \in \texttt{centers}$, there are a fraction of neurons aligning with ν more than $-\nu$. Condition (B3) shows that most neurons will somewhat align with either ν or $-\nu$. Condition (B4) is a technical concentration result. For proof details, see Appendix A.3.2.

Define the set $\mathcal{G}_{\text{good}}$ as

$$\mathcal{G}_{\text{good}} := \{(a, W^{(0)}, X) : a \in \mathcal{G}_A, X \in \mathcal{G}_{\text{data}}, W^{(0)} \in \mathcal{G}_W \text{ and conditions (B1)-(B4) hold}\},$$

whose probability is lower bounded by $\mathbb{P}((a, W^{(0)}, X) \in \mathcal{G}_{\text{good}}) \geq 1 - O(n^{-\varepsilon})$. This is a consequence of Lemmas A.1, A.3 and A.4 (see Appendix A.3.3).

Definition A.5. If the training data X and the initialization $a, W^{(0)}$ belong to $\mathcal{G}_{\text{good}}$, we define this circumstance as a "good run."

A.1.3 Proof Sketch for Theorem 3.1

In order for the network to learn a generalizable solution for the XOR cluster distribution, we would like positive neurons' (i.e., those with $a_j > 0$) weights w_j to align with $\pm\mu_1$, and negative neurons' weights to align with $\pm\mu_2$; we prove that this is satisfied for $t \in [Cn^{0.01}, \sqrt{n}]$. However, for $t = 1$, we show that the network only approximates a linear classifier, which can fit the training data in high dimension but has trivial test error. Figure 3 plots the evolution of the distribution of positive neurons' projections onto both μ_1 and μ_2, confirming that these neurons are much more aligned with $\pm\mu_1$ at a later training time, while they cannot distinguish $\pm\mu_1$ and $\pm\mu_2$ at $t = 1$.

Below we give a sketch of the proofs, and details are in Appendix A.5.

One-Step Catastrophic Overfitting: Under a good run, we have the following approximation for each neuron after the first iteration:

$$w_j^{(1)} \approx \frac{\alpha a_j}{2n} \sum_{i=1}^{n} \mathbb{I}(\langle w_j^{(0)}, x_i\rangle > 0) y_i x_i, \quad j \in [m].$$

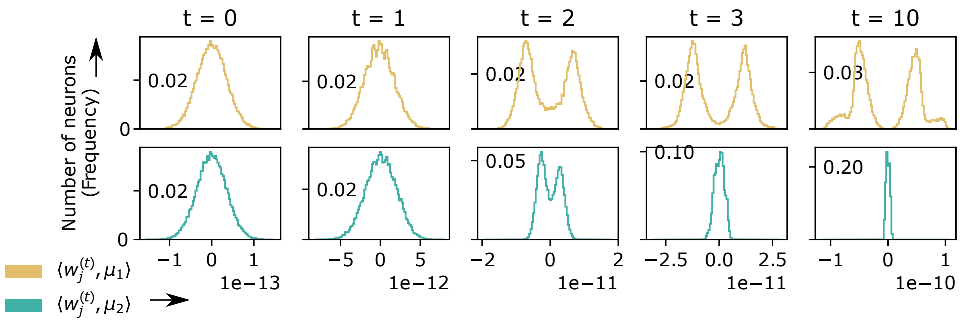

Figure 3: Histograms of inner products between positive neurons and μ_1 or μ_2 pooled over 100 independent runs under the same setting as in Figure 1. *Top (resp. bottom) row:* Inner products between positive neurons and μ_1 (resp. μ_2). While the distributions of the projections of positive neurons $w_j^{(t)}$ onto the μ_1 and μ_2 directions are nearly the same at times $t = 0, 1$, they become significantly more aligned with $\pm\mu_1$ over time. See Appendix A.7 for details of the experimental setup.

For details of this approximation, see Appendix A.4.

Let $s_{ij} := \mathbb{I}(\langle w_j^{(0)}, x_i \rangle > 0)$. Then, for sufficiently large m, we can approximate the neural network output at $t = 1$ as

$$\begin{aligned}\sum_{j=1}^{m} a_j \phi(\langle w_j^{(1)}, x \rangle) &\approx \frac{\alpha}{2n} \sum_{j=1}^{m} a_j \phi(a_j \langle \sum_{i=1}^{n} s_{ij} y_i x_i, x \rangle) \\ &\overset{a.s.}{\to} \frac{\alpha}{4n} \langle \sum_{i=1}^{n} \mathbb{E}[s_{ij}] y_i x_i, x \rangle = \frac{\alpha}{8n} \langle \sum_{i=1}^{n} y_i x_i, x \rangle.\end{aligned} \tag{A.3}$$

The convergence above follows from Lemma A.6 below and that the first-layer weights and second-layer weights are independent at initialization. This implies that the neural network classifier $\mathrm{sgn}(f(\cdot; W^{(1)}))$ behaves similarly to the linear classifier $\mathrm{sgn}(\langle \sum_{i=1}^{n} y_i x_i, \cdot \rangle)$. It can be shown that this linear classifier achieves 100% training accuracy whenever the training data are near orthogonal (Frei et al., 2023b, Appendix D), but because each class has two clusters with opposing means, linear classifiers only achieve 50% test error for the XOR cluster distribution. Thus at time $t = 1$, the network is able to fit the training data but is not capable of generalizing.

Lemma A.6. *Let $\{a_j\}$ and $\{b_j\}$ be two independent sequences of random variables with $a_j \overset{i.i.d.}{\sim} \mathrm{Unif}\{\pm \frac{1}{\sqrt{m}}\}$, and $\mathbb{E}[b_j] = b, \mathbb{E}[|b_j|] < \infty$. Then $\sum_{j=1}^{m} a_j \phi(a_j b_j) \to b/2$ almost surely as $m \to \infty$.*

Proof. Note that the ReLU function satisfies $x = \phi(x) - \phi(-x)$, and $\mathbb{E}[a_j \phi(a_j b_j)] = \mathbb{E}[\phi(b_j) - \phi(-b_j)]/2m = \mathbb{E}[b_j]/2m$. Then the result follows from the strong law of large number. $\square$

Multi-Step Generalization: Next, we show that positive (resp. negative) neurons gradually align with one of $\pm\mu_1$ (resp. $\pm\mu_2$), and forget both of $\pm\mu_2$ (resp. $\pm\mu_1$), making the network generalizable. Taking the direction $+\mu_1$ as an example, we define sets of neurons

$$\mathcal{J}_1 = \{j \in \mathcal{J}_{\texttt{Pos}} : j \text{ is } (+\mu_1, 20\varepsilon)\text{-aligned}\}; \quad \mathcal{J}_2 = \{j \in \mathcal{J}_{\texttt{Neg}} : j \text{ is } (\pm\mu_1, 20\varepsilon)\text{-aligned}\}.$$

We have by conditions (B2)-(B3) of Lemma A.4 that under a good run,

$$|\mathcal{J}_1| \geq m n^{-10\varepsilon}, \quad |\mathcal{J}_2| \geq (1 - 10 n^{-20\varepsilon}) |\mathcal{J}_{\texttt{Neg}}|,$$

which implies that $\mathcal{J}_1$ contains a certain proportion of $\mathcal{J}_{\texttt{Pos}}$ and $\mathcal{J}_2$ covers most of $\mathcal{J}_{\texttt{Neg}}$. The next lemma shows that neurons in $\mathcal{J}_1$ will keep aligning with $+\mu_1$, but neurons in $\mathcal{J}_2$ will gradually forget $+\mu_1$.

Lemma A.7. *Suppose that Assumptions (A1)-(A6) hold. Under a good run, we have that for $1 \le t \le \sqrt{n}$,*

$$\mathrm{cossim}(\sum_{j\in\mathcal{J}_1} w_j^{(t)}, +\mu_1) = \Omega(\frac{\sqrt{n}\|\mu\|}{\sqrt{p}});$$

$$\mathrm{cossim}(\sum_{j\in\mathcal{J}_2} w_j^{(t)}, +\mu_1) = O(\frac{\sqrt{n}\|\mu\|}{\sqrt{p}}(\frac{1}{t} + \sqrt{\frac{\log n}{n}})).$$

We can see that when t is large, $\mathrm{cossim}(\sum_{j\in\mathcal{J}_2} w_j^{(t)}, +\mu_1) = o(\mathrm{cossim}(\sum_{j\in\mathcal{J}_1} w_j^{(t)}, +\mu_1))$, thus for $x \sim N(+\mu_1, I_p)$, neurons with $j \in \mathcal{J}_1$ will dominate the output of $f(x; W^{(t)})$. For the other three clusters centered at $-\mu_1, +\mu_2, -\mu_2$ we have similar results, which then lead the model to generalization. Formally, we have the following theorem on generalization.

Theorem A.8. *Suppose that Assumptions (A1)-(A6) hold. Under a good run, for $Cn^{10\varepsilon} \le t \le \sqrt{n}$, the generalization error of classifier $\mathrm{sgn}(f(x, W^{(t)}))$ has an upper bound*

$$\mathbb{P}_{(x,y)\sim P_{clean}}(y \ne \mathrm{sgn}(f(x; W^{(t)}))) \le \exp\big(-\Omega(\frac{n^{1-20\varepsilon}\|\mu\|^4}{p})\big).$$

A.2 Properties of the training data

A.2.1 Proof of Lemma A.1

Lemma A.1 (Properties of training data)**.** *Suppose Assumptions (A1) and (A2) hold. Let the training data $\{(x_i, y_i)\}_{i=1}^n$ be sampled i.i.d from P as in Definition 2.1. With probability at least $1 - O(n^{-\varepsilon})$ the training data satisfy properties (E1)-(E4) defined below.*

(E1) *For all $k \in [n]$, $\max_{\nu\in\mathsf{centers}} \langle x_k - \bar{x}_k, \nu\rangle \le 10\sqrt{\log n}\|\mu\|$ and $|\|x_k\|^2 - p - \|\mu\|^2| \le 10\sqrt{p\log n}$,*

(E2) *For each $i, k \in [n]$ such that $i \ne k$, we have $|\langle x_i, x_k\rangle - \langle \bar{x}_i, \bar{x}_k\rangle| \le 10\sqrt{p\log n}$,*

(E3) *For $\nu \in$ centers, we have $|c_\nu + n_\nu - n/4| \le \sqrt{\varepsilon n\log n}$ and $|n_\nu - \eta n| \le \eta\sqrt{\varepsilon n\log n}$.*

(E4) *For $\nu \in$ centers, we have $|c_\nu + n_\nu - c_{-\nu} - n_{-\nu}| \ge n^{1/2-\varepsilon}$ and $|n_\nu - n_{-\nu}| \ge \eta n^{1/2-\varepsilon}$.*

Denote by $\mathcal{G}_{data}$ the set of training data satisfying conditions (E1)-(E4)*. Thus, the result can be stated succinctly as $\mathbb{P}(X \in \mathcal{G}_{data}) \ge 1 - O(n^{-\varepsilon})$.*

Proof. Before proceeding with the proof, we recall that centers $= \{\pm\mu_1, \pm\mu_2\}$. We first show that (E1) holds with large probability. To this end, fix $k \in [n]$. We have by the construction of x_k in Section 2.2 that $x_k \sim N(\bar{x}_k, I_p)$ for some $\bar{x}_k \in \{\pm\mu_1, \pm\mu_2\}$. Let $\xi_k = x_k - \bar{x}_k$. By Lemma A.26, we have

$$\mathbb{P}\big(\|\xi_k\| > \sqrt{p(t+1)}\big) \le \mathbb{P}\big(\big|\|\xi_k\|^2 - p\big| > pt\big) \le 2\exp(-pt^2/8), \quad \forall t \in (0, 1). \tag{A.4}$$

Note that for any fixed non-zero vector $\nu \in \mathbb{R}^p$, we have $\langle \nu, \xi_k\rangle \sim N(0, \|\nu\|^2)$. Therefore, again by Lemma A.26, we have

$$\mathbb{P}(|\langle \nu, \xi_k\rangle| > t\|\nu\|) \le \exp(-t^2/2), \quad \forall t \ge 1 \tag{A.5}$$

where the parameter t in both inequality will be chosen later. To show that the first inequality of (E1) holds w.h.p, we show the complement event $\mathcal{F}_k := \{\max_{\nu\in\mathsf{centers}} \langle \xi_k, \nu\rangle > t\|\mu\|\}$ has low probability. Applying the union bound,

$$\begin{aligned}\mathbb{P}(\mathcal{F}_k) &\le \sum_{\nu\in\{\pm\mu_1,\pm\mu_2\}} \mathbb{P}(|\langle \xi_k, \nu\rangle| > t\|\mu\|) \quad \because \text{Union bound}\\ &\le 4\exp(-t^2/2) \quad \because \text{Inequality (A.5).}\end{aligned}$$

Let $\delta := n^{-\varepsilon}$. Picking $t = \sqrt{2\log(16n/\delta)}$ in inequality (A.5) and applying the union bound again, we have

$$\mathbb{P}(\textstyle\bigcup_{k=1}^n \mathcal{F}_k) \le 4n\exp(-t^2/2) \le \delta/4. \tag{A.6}$$

Next, fix $t_1 \in (0,1)$ and $t_2 \geq 1$ arbitrary. To show that the second inequality of (E1) holds w.h.p, we first prove an intermediate step: the complement event $\mathcal{E}_k := \{|\|x_k\|^2 - p - \|\mu\|^2| > pt_1 + 2\|\mu\|t_2\}$ has low probability. Towards this, first note that since

$$\|x_k\|^2 = \|\bar{x}_k\|^2 + \|\xi_k\|^2 + 2\langle \bar{x}_k, \xi_k\rangle = \|\mu\|^2 + \|\xi_k\|^2 + 2\langle \bar{x}_k, \xi_k\rangle$$

we have the alternative characterization of $\mathcal{E}_k$ as

$$\mathcal{E}_k = \{|\|\xi_k\|^2 - p + 2\langle \bar{x}_k, \xi_k\rangle| > pt_1 + 2\|\mu\|t_2\}.$$

Next, recall the fact: if $X, Y \in \mathbb{R}$ are random variables and $a, b \in \mathbb{R}$ are constants, then

$$\mathbb{P}(|X+Y| > a+b) \leq \mathbb{P}(|X| > a) + \mathbb{P}(|Y| > b). \tag{A.7}$$

To see this, first note that $|X+Y| \leq |X| + |Y|$ by the triangle inequality. From this we deduce that $\mathbb{P}(|X+Y| > a+b) \leq \mathbb{P}(|X| + |Y| > a+b)$. Now, by the union bound, we have

$$\mathbb{P}(|X| + |Y| > a+b) \leq \mathbb{P}(\{|X| > a\} \cup \{|Y| > b\}) \leq \mathbb{P}(|X| > a) + \mathbb{P}(|Y| > b)$$

which proves (A.7). Now, to upper bound $\mathbb{P}(\mathcal{E}_k)$, note that

$$\begin{aligned}\mathbb{P}(\mathcal{E}_k) &= \mathbb{P}(|\|\xi_k\|^2 - p + 2\langle \bar{x}_k, \xi_k\rangle| > pt_1 + 2\|\mu\|t_2) \\ &\leq \mathbb{P}\big(\big|\|\xi_k\|^2 - p\big| > pt_1\big) + \mathbb{P}(|\langle \bar{x}_k, \xi_k\rangle| > t_2\|\mu\|) \quad \because \text{Inequality (A.7)} \\ &\leq 2\exp(-pt_1^2/8) + \exp(-t_2^2/2). \quad \because \text{Inequalities (A.4) and (A.5)}\end{aligned} \tag{A.8}$$

Inequality (A.8) is the crucial intermediate step to proving the second inequality of (E1). It will be convenient to complete the proof of the second inequality of (E1) simultaneously with that of (E2). To this end, we next prove an analogous intermediate step to (E2).

Fix $s_1, s_2 \geq 1$ to be chosen later. Define the event $\mathcal{E}_{ij} := \{|\langle x_i, x_j\rangle - \langle \bar{x}_i, \bar{x}_j\rangle| > s_1\sqrt{p} + 2t_2\|\mu\|\}$ for each pair $i, j \in [n]$ such that $1 \leq i \neq j \leq n$. We upper bound $\mathbb{P}(\mathcal{E}_{ij})$ in similar fashion as in (A.8). To this end, fix $i, j \in [n]$ such that $i \neq j$. Note that the identity $\langle x_i, x_j\rangle = \xi_i^\top \xi_j + \bar{x}_i^\top \bar{x}_j + \xi_i^\top \bar{x}_j + \xi_j^\top \bar{x}_i$ implies that $|\langle x_i, x_j\rangle - \langle \bar{x}_i, \bar{x}_j\rangle| = |\xi_i^\top \xi_j + \xi_i^\top \bar{x}_j + \xi_j^\top \bar{x}_i|$. Now, we claim that

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{ij}) &= \mathbb{P}(|\xi_i^\top \xi_j + \xi_i^\top \bar{x}_j + \xi_j^\top \bar{x}_i| \geq s_1\sqrt{p} + 2t_2\|\mu\|) \\ &\leq \mathbb{P}(|\xi_i^\top \xi_j| > s_1\sqrt{p}) + \mathbb{P}(|\xi_i^\top \bar{x}_j| > t_2\|\mu\|) + \mathbb{P}(|\xi_j^\top \bar{x}_i| > t_2\|\mu\|) \\ &\leq \exp(-s_1^2/2s_2) + 2\exp(-p(s_2-1)^2/8) + 2\exp(-t_2^2/2),\end{aligned} \tag{A.9}$$

The first inequality simply follows from applying (A.7) twice. Moreover, $\mathbb{P}(|\xi_i^\top \bar{x}_j| > t_2\|\mu\|)$ and $\mathbb{P}(|\xi_j^\top \bar{x}_i| > t_2\|\mu\|) \leq \exp(-t_2^2/2)$ follows from (A.5). To prove the claim, it remains to prove

$$\begin{aligned}&\mathbb{P}(|\langle \xi_i, \xi_j\rangle| > s_1\sqrt{p}) \\ &\leq \mathbb{P}\big(|\langle \xi_i, \xi_j\rangle| > s_1\sqrt{p} \;\big|\; \|\xi_j\| \leq \sqrt{s_2 p}\big) + \mathbb{P}(\|\xi_j\| > \sqrt{s_2 p}) \quad \because \text{law of total expectation} \\ &\leq \exp(-s_1^2/2s_2) + 2\exp(-p(s_2-1)^2/8).\end{aligned} \tag{A.10}$$

To prove the inequality at (A.10), first we get $\mathbb{P}(\|\xi_j\| > \sqrt{s_2 p}) \leq 2\exp(-p(s_2-1)^2/8)$ by applying (A.4) to upper bounds the second summand of the left-hand side of (A.10). For upper bounding the first summand, first let $\mathbb{P}\big(|\langle \xi_i, \xi_j\rangle| > s_1\sqrt{p} \;\big|\; \xi_j\big)$ be the conditional probability conditioned on a realization of ξ_j (while ξ_i remains random). Then by definition

$$\mathbb{P}\big(|\langle \xi_i, \xi_j\rangle| > s_1\sqrt{p} \;\big|\; \|\xi_j\| \leq \sqrt{s_2 p}\big) = \mathbb{E}_{\xi_j}[\mathbb{P}\big(|\langle \xi_i, \xi_j\rangle| > s_1\sqrt{p} \;\big|\; \xi_j\big) \;\big|\; \|\xi_j\| \leq \sqrt{s_2 p}]. \tag{A.11}$$

For fixed ξ_j such that $\|\xi_j\| \leq \sqrt{s_2 p}$, we have by (A.5) that

$$\mathbb{P}\big(|\langle \xi_i, \xi_j\rangle| > s_1\sqrt{p} \;\big|\; \xi_j\big) = \mathbb{P}\big(|\langle \xi_i, \xi_j\rangle| > \|\xi_j\|(s_1\sqrt{p}/\|\xi_j\|) \;\big|\; \xi_j\big) \leq \exp(-(s_1\sqrt{p}/\|\xi_j\|)^2/2).$$

Continue to assume fixed ξ_j such that $\|\xi_j\| \leq \sqrt{s_2 p}$, note that $s_1\sqrt{p}/\|\xi_j\| \geq s_1\sqrt{p}/\sqrt{s_2 p} = s_1/\sqrt{s_2}$ implies

$$\exp(-(s_1\sqrt{p}/\|\xi_j\|)^2/2) \leq \exp(-(s_1/\sqrt{s_2})^2/2).$$

Hence, $\mathbb{P}\big(|\langle \xi_i, \xi_j\rangle| > s_1\sqrt{p} \;\big|\; \xi_j\big) \leq \exp(-s_1^2/2s_2)$. Applying $\mathbb{E}_{\xi_j}[\,\cdot \;\big|\; \|\xi_j\| \leq \sqrt{s_2 p}\,]$ to both side of the preceding inequality, we get $\mathbb{P}\big(|\langle \xi_i, \xi_j\rangle| > s_1\sqrt{p} \;\big|\; \|\xi_j\| \leq \sqrt{s_2 p}\big) \leq \exp(-s_1^2/2s_2)$ which upper bounds the first summand of the left-hand side of (A.10). We now choose the values for $t_1 = \sqrt{8\log(16n/\delta)/p}$, $t_2 = \sqrt{2\log(16n^2/\delta)}$, $s_1 = 2\sqrt{\log(8n^2/\delta)}$, and $s_2 = 1 + \sqrt{8\log(16n^2/\delta)/p}$. Recall that $\delta = n^{-\varepsilon}$ and n is sufficiently large, then we have

$$\sqrt{\log(16n^2/\delta)/p} = \sqrt{\log(16n^{2+\varepsilon})/p} \le \sqrt{3\log(16n)/p} \le 1$$

by Assumptions (A1) and (A2). Combining (A.8) and (A.9) then applying the union bound, we have

$$\begin{aligned}\mathbb{P}((\cup_{k=1}^n \mathcal{E}_k) \cup (\cup_{i,j\in[n]:i\neq j}\mathcal{E}_{ij})) &\le \textstyle\sum_{k=1}^n \mathbb{P}(\mathcal{E}_k) + \sum_{i,j\in[n]:i\neq j}\mathbb{P}(\mathcal{E}_{ij}) \\ &\le 2n\exp(-\tfrac{pt_1^2}{8}) + n^2[2\exp(-\tfrac{t_2^2}{2}) + \exp(-\tfrac{s_1^2}{2s_2}) + 2\exp(-\tfrac{p(s_2-1)^2}{8})] \le \delta.\end{aligned} \tag{A.12}$$

Moreover, plugging the above values of t_1, t_2 and s_1 into the definition of $\mathcal{E}_k$ and $\mathcal{E}_{ij}$, we see that (E1) and (E2) are satisfied since they contain the complement of the event in (A.12).

Next, show that (E3) holds with large probability. We prove the inequality involving $|c_\nu + n_\nu - n/4|$ portion of (E3). Proofs for the rest of the inequalities in (E3) follow analogously using the same technique below. Recall from the data generation model, for each $k \in [n]$, $\bar{x}_k$ is sampled i.i.d $\sim$ Unif$\{\pm\mu_1, \pm\mu_2\}$. Define the following indicator random variable:

$$\mathbb{I}_\nu(k) = \begin{cases} 1 & \text{if } \bar{x}_k = \nu \\ 0 & \text{otherwise,}\end{cases} \quad \text{for each } k \in [n], \text{ and } \nu \in \{\pm\mu_1, \pm\mu_2\}$$

Then we have $\sum_\nu \mathbb{I}_\mu(k) = 1$ for each k, and $\mathbb{E}[\mathbb{I}_\nu(k)] = n/4$ for each ν. Applying Hoeffding's inequality, we obtain

$$\mathbb{P}(|\textstyle\sum_{k=1}^n \mathbb{I}_\nu(k) - n/4| > t\sqrt{n}) \le 2\exp(-2t^2).$$

Applying the union bound, we have

$$\mathbb{P}(\max_\nu |\textstyle\sum_{k=1}^n \mathbb{I}_\nu(k) - n/4| > t\sqrt{n}) \le 8\exp(-2t^2). \tag{A.13}$$

Thus we can bound the above tail probability by $O(\delta)$ by letting $t = \sqrt{\log(1/\delta)/2}$, and the upper bound $t\sqrt{n} \le \sqrt{n\log(1/\delta)} = \sqrt{n\varepsilon\log(n)}$.

Next, show that (E4) holds with large probability. We prove the inequality involving $|c_\nu + n_\nu - c_{-\nu} - n_{-\nu}|$ portion of (E4). Proofs for the rest of the inequalities in (E4) follow analogously using the same technique below. Note that for each k,

$$\mathbb{E}[\mathbb{I}_\nu(k) - \mathbb{I}_{-\nu}(k)] = 0; \quad \mathbb{E}[|\mathbb{I}_\nu(k) - \mathbb{I}_{-\nu}(k)|^l] = \frac{1}{4} \text{ for any } l \ge 1.$$

It yields that

$$\rho(\mathbb{I}_\nu(k) - \mathbb{I}_{-\nu}(k))/\text{Var}(\mathbb{I}_\nu(k) - \mathbb{I}_{-\nu}(k))^{3/2} = 2.$$

Applying the Berry-Esseen theorem (Lemma A.28), we have

$$\mathbb{P}(|c_\nu + n_\nu - c_{-\nu} - n_{-\nu}| > t\sqrt{n}) = \mathbb{P}(|\sum_{k=1}^n (\mathbb{I}_\nu(k) - \mathbb{I}_{-\nu}(k))| > t\sqrt{n}) \ge 2\bar{\Phi}(2t) - \frac{12}{\sqrt{n}}.$$

Let $t = n^{-\varepsilon}$. By $\Phi(t) \le 1/2 + \Phi'(0)t$, we have

$$\mathbb{P}(|\sum_{k=1}^n (\mathbb{I}_\nu(k) - \mathbb{I}_{-\nu}(k))| > t\sqrt{n}) \ge 1 - \frac{4}{\sqrt{2\pi}n^\varepsilon} - \frac{12}{\sqrt{n}} = 1 - O(\delta). \tag{A.14}$$

Combining (A.6), (A.12)-(A.14), we prove that conditions (E1)-(E4) hold with probability at least $1 - O(\delta)$ over the randomness of the training data. □

A.2.2 Proof of Corollary A.2

Corollary A.2 (Near-orthogonality of training data)**.** *Suppose Assumptions (A1), (A2), and Conditions (E1), (E2) from Lemma A.1 all hold. Then*

$$|\mathrm{cossim}(x_i, x_k)| \le \frac{2}{Cn^2}$$

for all $1 \le i \ne k \le n$*.*

Proof. By Lemma A.1, we have that under (E1) and (E2), when $i \ne j$,

$$\frac{|\langle x_i, x_j\rangle|}{\|x_i\| \cdot \|x_j\|} \le \frac{\|\mu\|^2 + C_n\sqrt{p}}{p + \|\mu\|^2 - C_n\sqrt{p}} \le \frac{2\|\mu\|^2}{p} \le \frac{2}{Cn^2},$$

for sufficiently large p. Here the second inequality comes from Assumption (A1); and the last inequality comes from Assumption (A2). ☐

A.3 Properties of the initial weights and activation patterns

We begin with additional notations that is used for the proofs of Lemmas A.3 and A.4. Following the notations in Xu & Gu (2023), we simplify the notation of $\mathcal{J}_{\texttt{Pos}}$ and $\mathcal{J}_{\texttt{Neg}}$ defined in Section A.1 as

$$\mathcal{J}_{\texttt{P}} := \mathcal{J}_{\texttt{Pos}} = \{j \in [m] : a_j > 0\}; \quad \mathcal{J}_{\texttt{N}} := \mathcal{J}_{\texttt{Neg}} = \{j \in [m] : a_j < 0\}.$$

We denote the set of pairs (i, j) such that the neuron j is active with respect to the sample x_i at time t by $\mathcal{A}^{(t)}$, i.e., define

$$\mathcal{A}^{(t)} := \{(i, j) \in [n] \times [m] : \langle w_j^{(t)}, x_i\rangle > 0\}.$$

Define subsets $\mathcal{A}^{i,(t)}$ and $\mathcal{A}_j^{(t)}$ of $\mathcal{A}^{(t)}$ where i (resp. j) is a sample (resp. neuron) index:

$$\mathcal{A}^{i,(t)} := \{j \in [m] : \langle w_j^{(t)}, x_i\rangle > 0\},$$
$$\mathcal{A}_j^{(t)} := \{i \in [n] : \langle w_j^{(t)}, x_i\rangle > 0\}.$$

Define

$$\mathcal{C}_{\nu,j}^{(t)} = \mathcal{C}_\nu \cap \mathcal{A}_j^{(t)}; \quad \mathcal{N}_{\nu,j}^{(t)} = \mathcal{N}_\nu \cap \mathcal{A}_j^{(t)}, \text{ for } j \in [m], \nu \in \mathsf{centers}.$$

Note that the above definition is equivalent to (A.1) from the main text.

Let $n_{\pm\nu} := n_\nu + n_{-\nu}$. For $\nu \in \mathsf{centers}$, we denote the sets of indices j of (ν, κ)-*aligned* neurons (see (A.2) in the main text for the definition of (ν, κ)-aligned-ness) with parameter $\kappa \in [0, \frac{1}{2})$:

$$\mathcal{J}_\nu^\kappa := \{j \in [m] : D_{\nu,j}^{(0)} > n^{1/2-\kappa}, \text{ and } d_{-\nu,j}^{(0)} < \min\{c_\nu, c_{-\nu}\} - 2n_{\pm\nu} - \sqrt{n}\}.$$

Thus, we have by definition that

$$\mathcal{J}_\nu^\kappa = \{j \in \mathcal{J}_{\texttt{P}} : \text{neuron } j \text{ is } (\nu, \kappa)\text{-aligned}\}$$

Further we denote

$$\mathcal{J}_{\texttt{P}}^{i,(t)} = \mathcal{J}_{\texttt{P}} \cap \mathcal{A}^{i,(t)}; \quad \mathcal{J}_{\texttt{N}}^{i,(t)} = \mathcal{J}_{\texttt{N}} \cap \mathcal{A}^{i,(t)}. \tag{A.15}$$

Finally, we denote

$$\mathcal{J}_{\nu,\texttt{P}}^\kappa = \mathcal{J}_{\texttt{P}} \cap \mathcal{J}_\nu^\kappa; \quad \mathcal{J}_{\nu,\texttt{N}}^\kappa = \mathcal{J}_{\texttt{N}} \cap \mathcal{J}_\nu^\kappa. \tag{A.16}$$

A.3.1 Proof of Lemma A.3

Lemma A.3 (Properties of the random weight initialization)**.** *Suppose Assumptions (A2) and (A6) hold. The followings hold with probability at least* $1 - O(n^{-\varepsilon})$ *over the random initialization:*

(D1) $\|W^{(0)}\|_F^2 \le \frac{3}{2}\omega_{\mathit{init}}^2 mp$, *and* *(D2)* $|\mathcal{J}_{\texttt{Pos}}| \ge m/3$ *and* $|\mathcal{J}_{\texttt{Neg}}| \ge m/3$.

Denote the set of $W^{(0)}$ satisfying condition (D1) by $\mathcal{G}_W$. Denote the set of $a = (a_j)_{j=1}^m$ satisfying condition (D2) by $\mathcal{G}_A$. Then $\mathbb{P}(a \in \mathcal{G}_A, W^{(0)} \in \mathcal{G}_W) \geq 1 - O(n^{-\varepsilon})$.

Proof. Recall earlier for simplicity, we defined for simplicity $\mathcal{J}_{\texttt{P}} = \mathcal{J}_{\texttt{Pos}}$ and $\mathcal{J}_{\texttt{N}} = \mathcal{J}_{\texttt{Neg}}$. Let $\delta = n^{-\varepsilon}$. Then (D1) is proved to hold with probability $1 - O(\delta)$ in the Lemma 4.2 of Frei et al. (2022b). For (D2), since $|\mathcal{J}_{\texttt{P}}|$ and $|\mathcal{J}_{\texttt{N}}|$ both follow distribution $\mathrm{B}(m, 1/2)$, it suffices to show that $\mathbb{P}(|\mathcal{J}_{\texttt{P}}| \geq m/3) \geq 1 - \delta$. Applying Hoeffding's inequality, we have

$$\mathbb{P}(|\mathcal{J}_{\texttt{P}}| \leq m/3) = \mathbb{P}(|\mathcal{J}_{\texttt{P}}| - m/2 \leq -m/6) \leq \exp(-m/18) \leq \delta,$$

where the last inequality comes from Assumption (A6). □

A.3.2 Proof of Lemma A.4

Lemma A.4 (Properties of the interaction between training data and initial weights)**.** *Suppose Assumptions (A1)-(A3) and (A6) hold. Given $a \in \mathcal{G}_A, X \in \mathcal{G}_{data}$, the followings hold with probability at least $1 - O(n^{-\varepsilon})$ over the random initialization $W^{(0)}$:*

(B1) For all $i \in [n]$, the sample x_i activates a large proportion of positive and negative neurons, i.e., $|\{j \in \mathcal{J}_{\texttt{Pos}} : \langle w_j^{(0)}, x_i \rangle > 0\}| \geq m/7$ and $|\{j \in \mathcal{J}_{\texttt{Neg}} : \langle w_j^{(0)}, x_i \rangle > 0\}| \geq m/7$ both hold.

(B2) For all $\nu \in \mathsf{centers}$ and $\kappa \in [0, \frac{1}{2})$, both $|\{j \in \mathcal{J}_{\texttt{Pos}} : j \text{ is } (\nu, \kappa)\text{-aligned}\}| \geq mn^{-10\varepsilon}$, and $|\{j \in \mathcal{J}_{\texttt{Neg}} : j \text{ is } (\nu, \kappa)\text{-aligned}\}| \geq mn^{-10\varepsilon}$.

(B3) For all $\nu \in \mathsf{centers}$, we have $\big|\{j \in \mathcal{J}_{\texttt{Pos}} : j \text{ is } (\pm\nu, 20\varepsilon)\text{-aligned}\}\big| \geq (1 - 10n^{-20\varepsilon})|\mathcal{J}_{\texttt{Pos}}|$. Moreover, the same statement holds if "$\mathcal{J}_{\texttt{Pos}}$" is replaced with "$\mathcal{J}_{\texttt{Neg}}$" everywhere.

(B4) For all $\nu \in \mathsf{centers}$ and $\kappa \in [0, \frac{1}{2})$, let $\mathcal{J}^{\kappa}_{\nu,\texttt{Pos}} := \{j \in \mathcal{J}_{\texttt{Pos}} : j \text{ is } (\nu, \kappa)\text{-aligned}\}$. Then $\sum_{j \in \mathcal{J}^{\kappa}_{\nu,\texttt{Pos}}} (c_\nu - n_\nu - d^{(0)}_{-\nu,j}) \geq \frac{n}{10}|\mathcal{J}^{\kappa}_{\nu,\texttt{Pos}}|$. Moreover, the same statement holds if "$\mathcal{J}_{\texttt{Pos}}$" is replaced with "$\mathcal{J}_{\texttt{Neg}}$" everywhere.

Before we proceed with the proof of Lemma A.4, we consider the following restatements of (B1) through (B4):

(B'1) For each $i \in [n]$, x_i activates a constant fraction of neurons initially, i.e. for each $i \in [n]$ the sets $\mathcal{J}_{\texttt{P}}^{i,(0)}$ and $\mathcal{J}_{\texttt{N}}^{i,(0)}$ defined at (A.15) satisfy

$$|\mathcal{J}_{\texttt{P}}^{i,(0)}| \geq m/7 \quad \text{and} \quad |\mathcal{J}_{\texttt{N}}^{i,(0)}| \geq m/7.$$

(B'2) For $\nu \in \mathsf{centers}$ and $\kappa \in [0, 1/2)$, we have $\min\{|\mathcal{J}^{\kappa}_{\nu,\texttt{P}}|, |\mathcal{J}^{\kappa}_{\nu,\texttt{N}}|\} \geq mn^{-10\varepsilon}$.

(B'3) For $\nu \in \mathsf{centers}$, we have $\big|\mathcal{J}^{20\varepsilon}_{\nu,\texttt{P}} \cup \mathcal{J}^{20\varepsilon}_{-\nu,\texttt{P}}\big| \geq (1 - 10n^{-20\varepsilon})|\mathcal{J}_{\texttt{P}}|$ and $\big|\mathcal{J}^{20\varepsilon}_{\nu,\texttt{N}} \cup \mathcal{J}^{20\varepsilon}_{-\nu,\texttt{N}}\big| \geq (1 - 10n^{-20\varepsilon})|\mathcal{J}_{\texttt{N}}|$.

(B'4) For $\nu \in \mathsf{centers}$ and $\kappa \in [0, \frac{1}{2})$, we have $\sum_{j \in \mathcal{J}} (c_\nu - d^{(0)}_{-\nu,j}) \geq \frac{n}{10}|\mathcal{J}|$, where $\mathcal{J} \in \{\mathcal{J}^{\kappa}_{\nu,\texttt{P}}, \mathcal{J}^{\kappa}_{\nu,\texttt{N}}\}$.

Unwinding the definitions, we note that the (B'1) through (B'4) are equivalent to the (B1) through (B4) of Lemma A.4

Proof. Let $\delta = n^{-\varepsilon}$. Throughout this proof, we implicitly condition on the fixed $\{a_j\} \in \mathcal{G}_A$ and $\{x_i\} \in \mathcal{G}_{\text{data}}$, i.e., when writing a probability and expectation we write $\mathbb{P}(\,\cdot\,|\{a_j\}, \{x_i\})$ and $\mathbb{E}[\,\cdot\,|\{a_j\}, \{x_i\}]$ to denote $\mathbb{P}(\,\cdot\,)$ and $\mathbb{E}[\,\cdot\,]$ respectively.

Proof of condition (B1): Define the following events for each $i \in [n]$:

$$\mathcal{P}_i := \{|\mathcal{J}_{\texttt{P}}^{i,(0)}| \geq m/7\}; \quad \mathcal{N}_i := \{|\mathcal{J}_{\texttt{N}}^{i,(0)}| \geq m/7\}.$$

We first show that $\cap_{i=1}^n (\mathcal{P}_i \cap \mathcal{N}_i)$ occurs with large probability. To this end, applying the union bound, we have

$$\mathbb{P}\big(\cap_{i=1}^n (\mathcal{P}_i \cap \mathcal{N}_i)\big) = 1 - \mathbb{P}\big(\cup_{i=1}^n (\mathcal{P}_i^c \cup \mathcal{N}_i^c)\big) \geq 1 - \sum_{i=1}^n \big(\mathbb{P}(\mathcal{P}_i^c) + \mathbb{P}(\mathcal{N}_i^c)\big).$$

Note that $\mathcal{P}_i$ and $\mathcal{N}_i$ are defined completely analogously corresponding to when $a_j > 0$ and $a_j < 0$, respectively. Thus, to prove (B1), it suffices to show that $\mathbb{P}(\mathcal{P}_i^c) \leq \delta/(4n)$ for each i, or equivalently,

$$\mathbb{P}\big(\sum_{j \in \mathcal{J}_{\mathtt{P}}} U_j \leq \frac{m}{7}\big) \leq \frac{\delta}{4n}$$

holds for each $i \in [n]$, where $U_j := \mathbb{I}(\langle w_j^{(0)}, x_i \rangle > 0)$. Note that given x_i and $\mathcal{J}_{\mathtt{P}}$, $\{U_j\}_{j \in \mathcal{J}_{\mathtt{P}}}$ are i.i.d Bernoulli random variables with mean $1/2$, thus we have

$$\mathbb{P}\big(\sum_{j \in \mathcal{J}_{\mathtt{P}}} U_j \leq \frac{m}{7}\big) \leq \mathbb{P}\big(\sum_{j \in \mathcal{J}_{\mathtt{P}}} (U_j - \frac{1}{2}) \leq (\frac{1}{7} - \frac{1}{6})m\big) \leq \exp(-2m(\frac{1}{6} - \frac{1}{7})^2) \leq \frac{\delta}{4n},$$

where the first inequality uses $|\mathcal{J}_{\mathtt{P}}| \geq m/3$; the second inequality comes from Hoeffding's inequality; and the third inequality uses Assumption (A6). Now we have proved that (B1) holds with probability at least $1 - \delta/2$.

Proof of condition (B2): Without loss of generality, we only prove the results for $\mathcal{J}_{\nu,\mathtt{P}}^{\kappa}$. Note that $\mathcal{J}_{\nu,\mathtt{P}}^{\kappa_1} \subseteq \mathcal{J}_{\nu,\mathtt{P}}^{\kappa_2}$ for $\kappa_1 < \kappa_2$. Thus we only consider the case $\kappa = 0$. It suffices to show that for each $j \in [m]$,

$$\mathbb{P}(D_{\nu,j}^{(0)} > \sqrt{n}) \geq 8n^{-10\varepsilon} \quad \text{and} \quad \mathbb{P}(d_{\mu,j}^{(0)} \geq \min\{c_\nu, c_{-\nu}\} - 2n_{\pm\nu} - \sqrt{n}) \leq n^{-10\varepsilon}, \mu \in \{\pm\nu\}. \tag{A.17}$$

Suppose (A.17) holds for any $\nu \in \{\pm\mu_1, \pm\mu_2\}$. Applying the inequality $P(A \cap B) \geq 1 - P(A^c) - P(B^c)$, we have

$$\mathbb{P}(D_{\nu,j}^{(0)} > \sqrt{n}, d_{\mu,j}^{(0)} < \min\{c_\nu, c_{-\nu}\} - 2n_{\pm\nu} - \sqrt{n}, \mu \in \{\pm\nu\}) \geq 8n^{-10\varepsilon} - 2n^{-10\varepsilon} = 6n^{-10\varepsilon}.$$

Then we have

$$\mathbb{E}[|\mathcal{J}_{\nu,\mathtt{P}}|] \geq 6n^{-10\varepsilon}|\mathcal{J}_{\mathtt{P}}| \geq \frac{2m}{n^{10\varepsilon}},$$

where the last inequality uses $\min\{|\mathcal{J}_{\mathtt{P}}|, |\mathcal{J}_{\mathtt{N}}|\} \geq m/3$, which comes from the definition of $\mathcal{G}_A$. Note that given $\{a_j\}$ and $\{x_i\}$, $|\mathcal{J}_{\nu,\mathtt{P}}|$ is the summation of i.i.d Bernoulli random variables. Applying Hoeffding's inequality, we obtain

$$\mathbb{P}(|\mathcal{J}_{\nu,\mathtt{P}}| \leq \frac{m}{n^{10\varepsilon}}) \leq \mathbb{P}(|\mathcal{J}_{\nu,\mathtt{P}}| - \mathbb{E}[|\mathcal{J}_{\nu,\mathtt{P}}|] \leq -\frac{m}{n^{10\varepsilon}}) \leq \exp(-\frac{2m^2}{n^{20\varepsilon}|\mathcal{J}_{\mathtt{P}}|}) \leq n^{-\varepsilon},$$

where the last inequality uses $|\mathcal{J}_{\mathtt{P}}| = m - |\mathcal{J}_{\mathtt{N}}| \leq 2m/3$, $20\varepsilon \leq 0.01$, and Assumption (A6). Applying the union bound, we have

$$\mathbb{P}(\cap_{\nu \in \{\pm\mu_1, \pm\mu_2\}}\{|\mathcal{J}_{\nu,\mathtt{P}}| > m/n^{10\varepsilon}\}) \geq 1 - 4n^{-\varepsilon}.$$

Thus it remains to show (A.17). Without loss of generality, we will only prove (A.17) for $\nu = +\mu_1$, which can be easily extended to other ν's. Recall that $X = [x_1, \ldots, x_n]^\top$ is the given training data. Let $V = Xw_j^{(0)}$, then $V \sim N(0, XX^\top)$. Let $Z = [z_1, \cdots, z_n]^\top, z_i = v_i/\|x_i\|, i \in [n]$. Denote $\Sigma = \mathrm{Cov}(Z)$. Then $Z \sim N(0, \Sigma)$. By Corollary A.2, we have

$$\Sigma_{ii} = 1; \quad |\Sigma_{ij}| \leq \frac{2}{Cn^2}$$

for $1 \leq i \neq j \leq n$. Denote

$$\mathcal{A}_1 = \mathcal{C}_{+\mu_1} \cup \mathcal{N}_{-\mu_1}; \quad \mathcal{A}_2 = \mathcal{C}_{-\mu_1} \cup \mathcal{N}_{+\mu_1}.$$

By the definition of $\mathcal{G}_{\text{data}}$ and (E3) in Lemma A.1, we have

$$||\mathcal{A}_1| - |\mathcal{A}_2|| \leq |c_{+\mu_1} - c_{-\mu_1}| + |n_{+\mu_1} - n_{-\mu_1}| \leq (1+\eta)\sqrt{n\varepsilon\log(n)}; \tag{A.18}$$

$$|\mathcal{A}_1| + |\mathcal{A}_2| = c_{+\mu_1} + n_{+\mu_1} + c_{-\mu_1} + n_{-\mu_1} \geq \frac{n}{2} - 2\sqrt{n\varepsilon\log(n)} = \frac{n}{2} - o(n) \tag{A.19}$$

for sufficiently large n. Note that equivalently, we can rewrite $D^{(0)}_{+\mu_1,j}$ as

$$\sum_{i\in\mathcal{A}_1}\mathbb{I}(z_i>0)-\sum_{i\in\mathcal{A}_2}\mathbb{I}(z_i>0). \tag{A.20}$$

Since we want to give a lower bound for $D^{(0)}_{+\mu_1,j}$, below we only consider the case when $|\mathcal{A}_1|<|\mathcal{A}_2|$. With the new expression of $D^{(0)}_{+\mu_1,j}$, we have

$$\mathbb{P}(D^{(0)}_{+\mu_1,j}>\sqrt{n})=\sum_{k=0}^{\lfloor|\mathcal{A}_1|-\sqrt{n}\rfloor}\sum_{\substack{\mathcal{B}_2\subseteq\mathcal{A}_2\\|\mathcal{B}_2|=k}}\sum_{\substack{\mathcal{B}_1\subseteq\mathcal{A}_1\\|\mathcal{B}_1|>k+\sqrt{n}}}\mathbb{E}\Big[\prod_{i\in\mathcal{B}_1\cup\mathcal{B}_2}\mathbb{I}(z_i>0)\cdot\prod_{i\in(\mathcal{A}_1\backslash\mathcal{B}_1)\cup(\mathcal{A}_2\backslash\mathcal{B}_2)}\mathbb{I}(z_i\le 0)\Big]. \tag{A.21}$$

By Lemma A.25, we have

$$\mathbb{E}\Big[\prod_{i\in\mathcal{B}_1\cup\mathcal{B}_2}\mathbb{I}(z_i>0)\cdot\prod_{i\in(\mathcal{A}_1\backslash\mathcal{B}_1)\cup(\mathcal{A}_2\backslash\mathcal{B}_2)}\mathbb{I}(z_i\le 0)\Big]\ge\gamma^{|\mathcal{A}_1|+|\mathcal{A}_2|}, \tag{A.22}$$

where $\gamma=1/2-4/(Cn)$. Let $Z'=[z'_1,\cdots,z'_n]^\top\sim N(0,I_n)$. Denote $\Delta_j:=\sum_{i\in\mathcal{A}_1}\mathbb{I}(z'_i>0)-\sum_{i\in\mathcal{A}_2}\mathbb{I}(z'_i>0)$, and $n_\Delta=|\mathcal{A}_1|+|\mathcal{A}_2|$. Then we have $\Delta_j\sim B(|\mathcal{A}_1|,1/2)-B(|\mathcal{A}_2|,1/2)$, $\mathbb{E}[\Delta_j]=(|\mathcal{A}_1|-|\mathcal{A}_2|)/2$, and

$$\frac{\mathbb{E}[\Delta_j]}{\sqrt{n_\Delta}}\ge\frac{-(1+\eta)\sqrt{n\varepsilon\log(n)}}{2\sqrt{n/2-o(n)}}\ge-\sqrt{n\varepsilon\log(n)} \tag{A.23}$$

by (A.18) and (A.19). Here the last inequality comes from Assumption (A3). Combining (A.21) and (A.22), we have

$$\begin{aligned}\mathbb{P}(D^{(0)}_{+\mu_1,j}>\sqrt{n})&\ge\sum_{k=0}^{\lfloor|\mathcal{A}_1|-\sqrt{n}\rfloor}\sum_{\substack{\mathcal{B}_2\subseteq\mathcal{A}_2\\|\mathcal{B}_2|=k}}\sum_{\substack{\mathcal{B}_1\subseteq\mathcal{A}_1\\|\mathcal{B}_1|>k+\sqrt{n}}}\gamma^{|\mathcal{A}_1|+|\mathcal{A}_2|}\\&=(2\gamma)^{|\mathcal{A}_1|+|\mathcal{A}_2|}\sum_{k=0}^{\lfloor|\mathcal{A}_1|-\sqrt{n}\rfloor}\sum_{\substack{\mathcal{B}_2\subseteq\mathcal{A}_2\\|\mathcal{B}_2|=k}}\sum_{\substack{\mathcal{B}_1\subseteq\mathcal{A}_1\\|\mathcal{B}_1|>k+\sqrt{n}}}(\frac{1}{2})^{|\mathcal{A}_1|+|\mathcal{A}_2|}\\&=(2\gamma)^{|\mathcal{A}_1|+|\mathcal{A}_2|}\mathbb{P}(\Delta_j>\sqrt{n})\\&\ge(1-\frac{8}{Cn})^n\mathbb{P}(\Delta_j>\sqrt{n})\ge(1-\frac{8}{C})\mathbb{P}(\Delta_j>\sqrt{n}),\end{aligned} \tag{A.24}$$

where the second equation uses the decomposition of $\mathbb{P}(\Delta_j>\sqrt{n})$; the second inequality uses $|\mathcal{A}_1|+|\mathcal{A}_2|\le n$; and the last inequality uses $f(n)=(1-8/(Cn))^n$ is a monotonically increasing function for $n\ge 1$. Note that

$$\begin{aligned}\mathbb{P}(\Delta_j>\sqrt{n})&=\mathbb{P}\Big(\frac{\Delta_j-\mathbb{E}[\Delta_j]}{\sqrt{n_\Delta}/2}>\frac{\sqrt{n}-\mathbb{E}[\Delta_j]}{\sqrt{n_\Delta}/2}\Big)\\&\ge\bar{\Phi}\Big(\frac{\sqrt{n}-\mathbb{E}[\Delta_j]}{\sqrt{n_\Delta}/2}\Big)-O(\frac{1}{\sqrt{n}})\ge\bar{\Phi}(2(\sqrt{3}+\sqrt{\varepsilon\log(n)}))-O(\frac{1}{\sqrt{n}}),\end{aligned}$$

where the first inequality uses Berry-Esseen theorem (Lemma A.28), and the second inequality is from (A.19) and (A.23). If $\sqrt{\varepsilon\log(n)}\le\sqrt{3}$, then $\bar{\Phi}(2(\sqrt{3}+\sqrt{\varepsilon\log(n)}))-O(1/\sqrt{n})=\Omega(1)$, which gives a constant lower bound for $\mathbb{P}(\Delta_j>\sqrt{n})$. If $\sqrt{\varepsilon\log(n)}>\sqrt{3}$, we have

$$\begin{aligned}\bar{\Phi}(2(\sqrt{3}+\sqrt{\varepsilon\log(n)}))\ge\bar{\Phi}(4\sqrt{\varepsilon\log(n)})&\ge\frac{1}{8\sqrt{2\pi\varepsilon\log(n)}}\exp(-8\varepsilon\log(n))\\&=\frac{1}{8\sqrt{2\pi\varepsilon\log(n)}n^{8\varepsilon}}\ge\frac{17}{n^{10\varepsilon}},\end{aligned}$$

for sufficiently large n. Here the second inequality uses $\bar{\Phi}(x) \geq \Phi'(x)/(2x)$ for $x \geq 1$. Combining both situations, we have

$$\mathbb{P}(\Delta_j > \sqrt{n}) \geq \frac{17}{n^{10\varepsilon}} - \frac{C_{\text{BE}}}{\sqrt{n/3}} \geq \frac{16}{n^{10\varepsilon}} \tag{A.25}$$

for sufficiently large n. Combining (A.24) and (A.25), we have

$$\mathbb{P}(D^{(0)}_{+\mu_1,j} > \sqrt{n}) \geq (1 - \frac{8}{C})\frac{16}{n^{10\varepsilon}} \geq \frac{8}{n^{10\varepsilon}}$$

for $C \geq 16$. It remains to prove

$$\mathbb{P}(d^{(0)}_{\mu,j} \geq \min\{c_{+\mu_1}, c_{-\mu_1}\} - 2n_{\pm\mu_1} - \sqrt{n}) \leq \frac{1}{n^{10\varepsilon}}, \mu \in \{\pm\mu_1\}.$$

Without loss of generality, below we prove it for $\mu = +\mu_1$. According to condition (E3) in Lemma A.1, we have

$$\min\{c_{+\mu_1}, c_{-\mu_1}\} - 2n_{\pm\mu_1} - \sqrt{n} \geq (\frac{1}{4} - 5\eta)n - 6\sqrt{n\varepsilon\log(n)} - \sqrt{n} \geq (\frac{1}{5} - \frac{5}{C})n \geq \frac{n}{6} \tag{A.26}$$

for $C \geq 150$ and sufficiently large n. Here the second inequality is from Assumption (A3). Thus it suffices to prove $\mathbb{P}(d^{(0)}_{+\mu_1,j} \geq n/6) \leq n^{-10\varepsilon}$. Note that

$$d^{(0)}_{+\mu_1,j} = \sum_{i \in \mathcal{C}_{+\mu_1}} \mathbb{I}(z_i > 0) - \sum_{i \in \mathcal{N}_{+\mu_1}} \mathbb{I}(z_i > 0).$$

Denote

$$\Delta'_j := \sum_{i \in \mathcal{C}_{+\mu_1}} \mathbb{I}(z'_i > 0) - \sum_{i \in \mathcal{N}_{+\mu_1}} \mathbb{I}(z'_i > 0).$$

Following the same proof procedure for the anti-concentration result of $D^{(0)}_{+\mu_1,j}$, we have

$$\mathbb{P}(d^{(0)}_{+\mu_1,j} \geq \frac{n}{6}) \leq (2\gamma_2)^{c_{+\mu_1}+n_{+\mu_1}} \mathbb{P}(\Delta'_j \geq \frac{n}{6}),$$

where $\gamma_2 = 1/2 + 4/(Cn)$. According to condition (E3) in Lemma A.1, we have $c_{+\mu_1} - n_{+\mu_1} \leq (1/4 - 2\eta)n + 2\sqrt{n\varepsilon\log(n)}$. It yields that

$$\mathbb{E}[\Delta'_j] = \frac{c_{+\mu_1} - n_{+\mu_1}}{2} \leq (1/8 - \eta)n + \sqrt{n\varepsilon\log(n)} \leq n/7.$$

Applying Hoeffding's inequality, we have

$$\mathbb{P}(\Delta'_j \geq n/6) \leq \mathbb{P}(\Delta'_j - \mathbb{E}[\Delta'_j] \geq n/42) \leq \exp(-\Omega(n)).$$

Combining the inequalities above, we have

$$\mathbb{P}(d^{(0)}_{+\mu_1,j} \geq n/6) \leq (1 + \frac{8}{Cn})^{c_{+\mu_1}+n_{+\mu_1}} \mathbb{P}(\Delta'_j \geq n/6) = \exp(-\Omega(n)) \leq \frac{1}{n^{10\varepsilon}}, \tag{A.27}$$

where the equation uses $(1 + 8/(Cn))^{c_{+\mu_1}+n_{+\mu_1}} \leq (1 + 8/(Cn))^n \leq \exp(8/C)$. Now we have completed the proof for (B2).

Proof of condition (B3): Without loss of generality, we only prove the results for $\mathcal{J}^{20\varepsilon}_{+\mu_1,\mathtt{P}} \cup \mathcal{J}^{20\varepsilon}_{-\mu_1,\mathtt{P}}$. By Berry-Essen theorem, we have

$$\begin{aligned}\mathbb{P}(|\Delta_j| \leq n^{1/2-20\varepsilon}) &= \mathbb{P}\big(\frac{\Delta_j - \mathbb{E}[\Delta_j]}{\sqrt{n_\Delta}/2} \in [-\frac{\mathbb{E}[\Delta_j]}{\sqrt{n_\Delta}/2} - \frac{2}{n^{20\varepsilon}}, -\frac{\mathbb{E}[\Delta_j]}{\sqrt{n_\Delta}/2} + \frac{2}{n^{20\varepsilon}}]\big) \\ &\leq 2[\Phi(\frac{2}{n^{20\varepsilon}}) - \Phi(0)] + O(\frac{1}{\sqrt{n}}) \leq 4n^{-20\varepsilon},\end{aligned}$$

where the first inequality uses $\Phi(b) - \Phi(a) \leq 2(\Phi((b-a)/2) - \Phi(0)), b \geq a$; the second inequality uses $\Phi(x) - \Phi(0) \leq \Phi'(0)x, x \geq 0$ and $20\varepsilon < 1/2$. It yields that

$$\mathbb{P}(|D^{(0)}_{+\mu_1,j}| \leq n^{1/2-20\varepsilon}) \leq 2\mathbb{P}(|\Delta_j| \leq n^{1/2-20\varepsilon}) \leq 8n^{-20\varepsilon},$$

where the first inequality is from Lemma A.24. Combined with (A.26) and (A.27), we have

$$\begin{aligned}
&\mathbb{P}(|D_{\nu,j}^{(0)}| > n^{1/2-20\varepsilon}, d_{\nu,j}^{(0)} < \min\{c_\nu, c_{-\nu}\} - 2n_{\pm\nu} - \sqrt{n}, \nu \in \{\pm\mu_1\}) \\
&\geq \mathbb{P}(|D_{\nu,j}^{(0)}| > n^{1/2-20\varepsilon}, d_{\nu,j}^{(0)} < n/6, \nu \in \{\pm\mu_1\}) \\
&\geq 1 - 8n^{-20\varepsilon} - 2\exp(-\Omega(n)) \geq 1 - 9n^{-20\varepsilon},
\end{aligned}$$

where the second inequality uses $D_{\nu,j}^{(0)} = -D_{-\nu,j}^{(0)}$ and $\mathbb{P}(\cap_{i=1}^n A_i) = 1 - \mathbb{P}(\cup_{i=1}^n A_i^c) \geq 1 - \sum_{i=1}^n \mathbb{P}(A_i^c)$. Note that given $\{a_j\}$ and $\{x_i\}$, $|\mathcal{J}_{\nu,\mathrm{P}} \cup \mathcal{J}_{-\nu,\mathrm{P}}|$ is the summation of i.i.d Bernoulli random variables with expectation larger than $1 - 9n^{-20\varepsilon}$. Applying Hoeffding's inequality, we obtain

$$\begin{aligned}
&\mathbb{P}(|\mathcal{J}_{+\mu_1,\mathrm{P}}^{20\varepsilon} \cup \mathcal{J}_{-\mu_1,\mathrm{P}}^{20\varepsilon}| < |\mathcal{J}_\mathrm{P}|(1 - 10n^{-20\varepsilon})) \\
&\leq \mathbb{P}(|\mathcal{J}_{+\mu_1,\mathrm{P}}^{20\varepsilon} \cup \mathcal{J}_{-\mu_1,\mathrm{P}}^{20\varepsilon}| - \mathbb{E}[|\mathcal{J}_{+\mu_1,\mathrm{P}}^{20\varepsilon} \cup \mathcal{J}_{-\mu_1,\mathrm{P}}^{20\varepsilon}|] < -|\mathcal{J}_\mathrm{P}|n^{-20\varepsilon}) \\
&\leq \exp(-2|\mathcal{J}_\mathrm{P}|n^{-40\varepsilon}) \leq n^{-\varepsilon},
\end{aligned}$$

where the first inequality uses $\mathbb{E}[|\mathcal{J}_{+\mu_1,\mathrm{P}}^{20\varepsilon} \cup \mathcal{J}_{-\mu_1,\mathrm{P}}|] \geq |\mathcal{J}_\mathrm{P}^{20\varepsilon}|(1 - 9n^{-20\varepsilon})$ and the last inequality is from Assumption (A6) and $40\varepsilon < 0.01$.

Proof of condition (B4): Lastly we show that (B4) also holds with probability at least $1 - O(n^{-\varepsilon})$. Without loss of generality, we only prove it for $\mathcal{J}_{+\mu_1,\mathrm{P}}^\kappa$. Referring back to the definition of $\mathcal{J}_{+\mu_1,\mathrm{P}}^\kappa$ in equation (A.16), it is crucial to note that it solely imposes upper bounds on $d_{-\mu_1,j}^{(0)}$. Consequently, the average of $d_{-\mu_1,j}^{(0)}$ in $\mathcal{J}_{+\mu_1,\mathrm{P}}^\kappa$ is no more than the average of $d_{-\mu_1,j}^{(0)}$ in $\mathcal{J}_\mathrm{P}$, which imposes no constraints on $d_{-\mu_1,j}^{(0)}$. Armed with this understanding, when $|\mathcal{J}_{+\mu_1,\mathrm{P}}^\kappa| > 0$, we have that with probability 1,

$$\frac{1}{|\mathcal{J}_{+\mu_1,\mathrm{P}}^\kappa|} \sum_{j \in \mathcal{J}_{+\mu_1,\mathrm{P}}^\kappa} (c_{+\mu_1} - n_{+\mu_1} - d_{-\mu_1,j}^{(0)}) \geq \frac{1}{|\mathcal{J}_\mathrm{P}|} \sum_{j \in \mathcal{J}_\mathrm{P}} (c_{+\mu_1} - n_{+\mu_1} - d_{-\mu_1,j}^{(0)}).$$

Thus it suffices to show that

$$\frac{1}{|\mathcal{J}_\mathrm{P}|} \sum_{j \in \mathcal{J}_\mathrm{P}} (c_{+\mu_1} - n_{+\mu_1} - d_{-\mu_1,j}^{(0)}) \geq \frac{n}{10} \tag{A.28}$$

with probability at least $1 - O(\delta)$. Note that given the training data X, $\{d_{-\mu_1,j}^{(0)}\}_{j=1}^m$ are i.i.d random variables with $\mathbb{E}[d_{-\mu_1,j}^{(0)}] = (c_{-\mu_1} - n_{-\mu_1})/2$, which comes from the symmetry of the distribution of $w_j^{(0)}$. Then we have

$$\mathbb{E}[c_{+\mu_1} - n_{+\mu_1} - d_{-\mu_1,j}^{(0)}] = c_{+\mu_1} - n_{+\mu_1}(c_{-\mu_1} - n_{-\mu_1})/2 \geq (\frac{1}{8} - 5\eta)n - 5\sqrt{n\varepsilon\log(n)} \geq \frac{n}{9}. \tag{A.29}$$

Here the first inequality uses (E3) in Lemma A.1 and the second inequality uses Assumption (A3). Applying Hoeffding's inequality, we obtain

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{|\mathcal{J}_\mathrm{P}|} \sum_{j \in \mathcal{J}_\mathrm{P}} (c_{+\mu_1} - n_{+\mu_1} - d_{-\mu_1,j}^{(0)}) < \frac{n}{10}\Big) \\
&= \mathbb{P}\Big(\sum_{j \in \mathcal{J}_\mathrm{P}} (d_{-\mu_1,j}^{(0)} - \mathbb{E}[d_{-\mu_1,j}^{(0)}]) > \big(c_{+\mu_1} - n_{+\mu_1} - \frac{n}{10} - \mathbb{E}[d_{-\mu_1,j}^{(0)}]\big)|\mathcal{J}_\mathrm{P}|\Big) \\
&\leq \mathbb{P}\Big(\sum_{j \in \mathcal{J}_\mathrm{P}} (d_{-\mu_1,j}^{(0)} - \mathbb{E}[d_{-\mu_1,j}^{(0)}]) > \frac{n}{90}|\mathcal{J}_\mathrm{P}|\Big) \leq \exp\Big(-\frac{n^2|\mathcal{J}_\mathrm{P}|}{4050(c_{-\mu_1} + n_{-\mu_1})^2}\Big) \leq \delta,
\end{aligned}$$

where the first inequality uses (A.29), the second inequality uses Hoeffding's inequality and the bounds of $d_{-\mu_1,j}^{(0)}$, i.e. $-n_{-\mu_1} \leq d_{-\mu_1,j}^{(0)} \leq c_{-\mu_1}$, and the last inequality uses Assumption (A6). It proves (A.28). □

Remark A.9. In the proof of (B2), note that when $\Sigma = I_n$, z_i are independent with each other. Then (A.17) can be proved by applying Hoeffding's inequality. In our setting, Σ is close to the identity matrix, which means that $\{z_i\}$ are weakly dependent and inspires us to prove similar results.

A.3.3 Proof of the Probability bound of the "Good run" event

Combining the probability lower bound parts of Lemma A.1,A.3 and A.4, we have

$$\begin{aligned}
&\mathbb{P}((a, W^{(0)}, X) \in \mathcal{G}_{\text{good}}) \\
&\quad \geq \mathbb{P}(a \in \mathcal{G}_A, X \in \mathcal{G}_{\text{data}}, \text{(B1)-(B4) are satisfied}) - \mathbb{P}(W^{(0)} \notin \mathcal{G}_W) \\
&\quad \geq \mathbb{P}(\text{(B1)-(B4) are satisfied} \,|\, a \in \mathcal{G}_A, X \in \mathcal{G}_{\text{data}})\mathbb{P}(a \in \mathcal{G}_A, X \in \mathcal{G}_{\text{data}}) - O(n^{-\varepsilon}) \\
&\quad \geq (1 - O(n^{-\varepsilon}))(1 - O(n^{-\varepsilon})) - O(n^{-\varepsilon}) = 1 - O(n^{-\varepsilon}),
\end{aligned}$$

as desired.

A.4 Trajectory Analysis of the Neurons

Let $t \geq 0$ be an arbitrary step. Denote $z_i^{(t)} := y_i f(x_i; W^{(t)})$, and $h_i^{(t)} := g_i^{(t)} - 1/2$. Then we can decompose (2.2) as

$$w_j^{(t+1)} - w_j^{(t)} = \tfrac{\alpha a_j}{2n} \textstyle\sum_{i=1}^n \phi'(\langle w_j^{(t)}, x_i \rangle) y_i x_i + \tfrac{\alpha a_j}{n} \sum_{i=1}^n h_i^{(t)} \phi'(\langle w_j^{(t)}, x_i \rangle) y_i x_i. \tag{A.30}$$

Remark A.10. When $|z_i^{(t)}|$ is sufficiently small, we can use $1/2$ as an approximation for the negative derivative of the logistic loss by first-order Taylor's expansion and we will show that the training dynamics is nearly the same in the first $O(p)$ steps.

Lemma A.11. *Suppose that Assumptions (A1)-(A6) hold. Under a good run, for* $0 \leq t \leq 1/(\sqrt{n}p\alpha) - 2$*, we have* $\max_{i \in [n]} |h_i^{(t)}| \leq 2/n^{3/2}$.

Lemma A.12. *Suppose that Assumptions (A1)-(A6) hold. Under a good run, for* $0 \leq t \leq 1/(\sqrt{n}p\alpha) - 2$*, we have that for each* $k \in [n]$*,*

$$\begin{aligned}
&\left| \langle w_j^{(t+1)} - w_j^{(t)}, x_k \rangle - \frac{\alpha a_j}{2n} \big[y_k \phi'(\langle w_j^{(t)}, x_k \rangle) p + y_{\bar{x}_k} D_{\bar{x}_k, j}^{(t)} \|\mu\|^2 \big] \right| \\
&\qquad\qquad \leq \frac{4\alpha}{n^{5/2}\sqrt{m}} \big[\phi'(\langle w_j^{(t)}, x_k \rangle) p + \frac{C_n n^{1.99} \|\mu\|^2}{3C} \big], \text{ and}
\end{aligned} \tag{A.31}$$

$$\left| \langle w_j^{(t+1)} - w_j^{(t)}, \nu \rangle - \frac{\alpha a_j}{2n} y_\nu D_{\nu, j}^{(t)} \|\mu\|^2 \right| \leq \frac{5\alpha}{n^{3/2}\sqrt{m}} \|\mu\|^2. \tag{A.32}$$

where $C_n := 10\sqrt{\log(n)}$*,* $\bar{x}_k \in \mathsf{centers}$ *is defined as the cluster mean for sample* (x_k, y_k)*, and* y_ν *is defined as the clean label for cluster centered at* ν *(i.e.* $y_\nu = 1$ *for* $\nu \in \{\pm\mu_1\}$*,* $y_\nu = -1$ *for* $\nu \in \{\pm\mu_2\}$*).*

Taking a closer look at (A.31), we see that if $a_j y_k > 0$, and x_k activates neuron w_j at time s, then x_k will activate neuron $w_j^{(t)}$ for any $t \in [s, 1/(\sqrt{n}p\alpha) - 2]$. Moreover, if $a_j y_k < 0$, and x_k activates neuron w_j at time s, then x_k will not activate neuron w_j at time $s+1$, which implies that there is an upper bound for the inner product $\langle w_j^{(t)}, x_k \rangle$. These observations are stated as the corollary below:

Corollary A.13. *Suppose that Assumptions (A1)-(A6) hold. Under a good run, for any pair* $(j, k) \in [m] \times [n]$*, the following is true:*

(F1) When $a_j y_k > 0$*, if there exists some* $0 \leq s < 1/(\sqrt{n}p\alpha) - 2$ *such that* $\langle w_j^{(s)}, x_k \rangle > 0$*, then for any* $s \leq t \leq 1/(\sqrt{n}p\alpha) - 2$*, we have* $\langle w_j^{(t)}, x_k \rangle > 0$*.*

(F2) When $a_j y_k < 0$*, for any* $0 \leq t \leq 1/(\sqrt{n}p\alpha) - 2$ *we have that* $\langle w_j^{(t)}, x_k \rangle \leq \frac{\alpha}{\sqrt{m}} \|\mu\|^2$*.*

(F3) When $a_j y_k < 0$*, for any* $0 \leq t \leq 1/(\sqrt{n}p\alpha) - 3$ *we have that* $\langle w_j^{(t)}, x_k \rangle > 0$ *implies* $\langle w_j^{(t+1)}, x_k \rangle < 0$*.*

A.4.1 Proof of Lemma A.11

Lemma A.11. *Suppose that Assumptions (A1)-(A6) hold. Under a good run, for* $0 \le t \le 1/(\sqrt{n}p\alpha) - 2$*, we have* $\max_{i\in[n]} |h_i^{(t)}| \le 2/n^{3/2}$.

Proof. It suffices to show that for $0 \le t \le 1/(\sqrt{n}p\alpha) - 2$,

$$\max_{i\in[n]} |h_i^{(t)}| \le \frac{2\alpha p}{n}(t+2).$$

We prove the result by an induction on t. Denote

$$P(t): \quad \max_{i\in[n]} |h_i^{(\tau)}| \le \frac{2\alpha p}{n}(t+2), \quad \forall \tau \le t.$$

When $t = 0$, we have

$$|h_i^{(0)}| \le \frac{p\omega_{\text{init}}\sqrt{3m}}{2} \le \frac{\sqrt{3}\alpha\|\mu\|^2}{4nm} \le \frac{4\alpha p}{n}$$

by Lemma A.18, Assumption (A2) and (A5). Thus $P(0)$ holds. Suppose $P(t)$ holds and $t \le 1/(\sqrt{n}p\alpha) - 3$, then we have

$$|h_i^{(\tau)}| \le \frac{2\alpha p}{\sqrt{n}}(\tau+2) \le \frac{2}{\sqrt{n}}; \quad \frac{1}{2} - \frac{2}{\sqrt{n}} \le g_i^{(\tau)} \le \frac{1}{2} + \frac{2}{\sqrt{n}}, \quad \forall \tau \le t,$$

which yields that $\max_{i\in[n]} g_i^{(\tau)} \le 1$. Further we have that for each pair $(j,k) \in [m] \times [n]$,

$$\begin{aligned} |\langle w_j^{(\tau+1)} - w_j^{(\tau)}, x_k\rangle| &= \Big|\frac{\alpha a_j}{n}\sum_{i=1}^n g_i^{(\tau)}\phi'(\langle w_j^{(\tau)}, x_i\rangle)y_i\langle x_i, x_k\rangle\Big| \\ &\le \frac{\alpha}{n\sqrt{m}}\max_{i\in[n]} g_i^{(\tau)}(2p + 2n\|\mu\|^2) \le \frac{4\alpha p}{n\sqrt{m}}, \end{aligned}$$

where the first inequality uses $\|x_i\|^2 \le 2p$, $|\langle x_i, x_j\rangle| \le 2\mu^2$, which comes from Lemma A.1, and the second inequality uses Assumption (A2). It yields that for each pair $(j,k) \in [m] \times [n]$,

$$|\langle w_j^{(t+1)}, x_k\rangle| \le \sum_{\tau=0}^{t} |\langle w_j^{(\tau+1)} - w_j^{(\tau)}, x_k\rangle| + |\langle w_j^{(0)}, x_k\rangle| \le \frac{4\alpha p}{n\sqrt{m}}(t+1) + \sqrt{2p}\|w_j^{(0)}\| \le \frac{4\alpha p}{n\sqrt{m}}(t+2),$$

where the last inequality uses Lemma A.3 and Assumption (A5). Then we have that for each $k \in [n]$,

$$|f(x_k; W^{(t+1)})| \le \sum_{j=1}^m |a_j\langle w_j^{(t+1)}, x_k\rangle| \le \sqrt{m}\max_{j\in[m]} |\langle w_j^{(t+1)}, x_k\rangle| \le \frac{4\alpha p}{n}(t+2).$$

By $|1/(1+\exp(z)) - 1/2| \le |z|/2, \forall z$, we have for each $i \in [n]$,

$$|h_i^{(t+1)}| \le \frac{1}{2}|z_i^{(t+1)}| = \frac{1}{2}|f(x_i; W^{(t+1)})| \le \frac{2\alpha p}{n}(t+2).$$

Thus $P(t+1)$ is proved. □

A.4.2 Proof of Lemma A.12

Lemma A.12. *Suppose that Assumptions (A1)-(A6) hold. Under a good run, for* $0 \leq t \leq 1/(\sqrt{n}p\alpha) - 2$*, we have that for each* $k \in [n]$*,*

$$\begin{aligned} &\Big|\langle w_j^{(t+1)} - w_j^{(t)}, x_k\rangle - \frac{\alpha a_j}{2n}\big[y_k \phi'(\langle w_j^{(t)}, x_k\rangle)p + y_{\bar{x}_k} D_{\bar{x}_k,j}^{(t)}\|\mu\|^2\big]\Big| \\ &\qquad\qquad \leq \frac{4\alpha}{n^{5/2}\sqrt{m}}\big[\phi'(\langle w_j^{(t)}, x_k\rangle)p + \frac{C_n n^{1.99}\|\mu\|^2}{3C}\big], \text{ and} \end{aligned} \tag{A.31}$$

$$\Big|\langle w_j^{(t+1)} - w_j^{(t)}, \nu\rangle - \frac{\alpha a_j}{2n} y_\nu D_{\nu,j}^{(t)}\|\mu\|^2\Big| \leq \frac{5\alpha}{n^{3/2}\sqrt{m}}\|\mu\|^2. \tag{A.32}$$

where $C_n := 10\sqrt{\log(n)}$*,* $\bar{x}_k \in \mathsf{centers}$ *is defined as the cluster mean for sample* (x_k, y_k)*, and* y_ν *is defined as the clean label for cluster centered at* ν *(i.e.* $y_\nu = 1$ *for* $\nu \in \{\pm\mu_1\}$*,* $y_\nu = -1$ *for* $\nu \in \{\pm\mu_2\}$*).*

Proof. First we have

$$\begin{aligned} \Big|\frac{\alpha a_j}{n}\sum_{i=1}^n h_i^{(t)}\phi'(\langle w_j^{(t)}, x_i\rangle)y_i\langle x_i, x_k\rangle\Big| &\leq \frac{2\alpha}{n^{5/2}\sqrt{m}}\sum_{i=1}^n \phi'(\langle w_j^{(t)}, x_i\rangle)|\langle x_i, x_k\rangle| \\ &\leq \frac{2\alpha}{n^{5/2}\sqrt{m}}\big[\phi'(\langle w_j^{(t)}, x_k\rangle)\|x_k\|^2 + \sum_{i\neq k}|\langle x_i, x_k\rangle|\big] \\ &\leq \frac{4\alpha}{n^{5/2}\sqrt{m}}\big[\phi'(\langle w_j^{(t)}, x_k\rangle)p + n\|\mu\|^2\big], \end{aligned} \tag{A.33}$$

where the first inequality uses $\max_i h_i^{(t)} \leq 2n^{-3/2}$, which is from Lemma A.11; the third inequality uses $\|x_k\|^2 \leq 2p, |\langle x_i, x_k\rangle| \leq 2\|\mu\|^2$, which is induced by Lemma A.1. Next we have the following decomposition:

$$\begin{aligned} &\sum_{i=1}^n \phi'(\langle w_j^{(t)}, x_i\rangle)\langle y_i x_i, x_k\rangle \\ =& y_k\phi'(\langle w_j^{(t)}, x_k\rangle)(\|x_k\|^2 - p - \|\mu\|^2) + \sum_{i\neq k}\phi'(\langle w_j^{(t)}, x_i\rangle)y_i(\langle x_i, x_k\rangle - \langle\bar{x}_i, \bar{x}_k\rangle) \\ &+ y_k\phi'(\langle w_j^{(t)}, x_k\rangle)(p + \|\mu\|^2) + \sum_{i\neq k}\phi'(\langle w_j^{(t)}, x_i\rangle)y_i\langle\bar{x}_i, \bar{x}_k\rangle \\ =& y_k\phi'(\langle w_j^{(t)}, x_k\rangle)(\|x_k\|^2 - p - \|\mu\|^2) + \sum_{i\neq k}\phi'(\langle w_j^{(t)}, x_i\rangle)y_i(\langle x_i, x_k\rangle - \langle\bar{x}_i, \bar{x}_k\rangle) \\ &+ y_k\phi'(\langle w_j^{(t)}, x_k\rangle)p + y_{\bar{x}_k}D_{\bar{x}_k,j}^{(t)}\|\mu\|^2 + \sum_{i:\bar{x}_i\notin\{\pm\bar{x}_k\}}\phi'(\langle w_j^{(t)}, x_i\rangle)y_i\langle\bar{x}_i, \bar{x}_k\rangle, \end{aligned} \tag{A.34}$$

where the second equation uses the definition of $D_{\nu,j}^{(t)}$. Recall that $C_n = 10\sqrt{\log(n)}$. Combining with results in Lemma A.1, (A.34) yields that

$$\Big|\sum_{i=1}^n \phi'(\langle w_j^{(t)}, x_i\rangle)\langle y_i x_i, x_k\rangle - \big[y_k\phi'(\langle w_j^{(t)}, x_k\rangle)p + y_{\bar{x}_k}D_{\bar{x}_k,j}^{(t)}\|\mu\|^2\big]\Big| \leq nC_n\sqrt{p} + 2n\|\mu\| \leq 2nC_n\sqrt{p}, \tag{A.35}$$

where the first inequality uses (E1) and (E2) in Lemma A.1 and the second inequality uses Assumption (A2). Recall the decomposition (A.30) of the gradient descent update, we have

$$\langle w_j^{(t+1)} - w_j^{(t)}, x_k\rangle = \frac{\alpha a_j}{2n}\sum_{i=1}^n \phi'(\langle w_j^{(t)}, x_i\rangle)\langle y_i x_i, x_k\rangle + \frac{\alpha a_j}{n}\sum_{i=1}^n h_i^{(t)}\phi'(\langle w_j^{(t)}, x_i\rangle)\langle y_i x_i, x_k\rangle \tag{A.36}$$

Then combining (A.33), (A.35), and (A.36), we have

$$\begin{aligned}
&\Big|\langle w_j^{(t+1)} - w_j^{(t)}, x_k\rangle - \frac{\alpha a_j}{2n}\big[y_k\phi'(\langle w_j^{(t)}, x_k\rangle)p + y_{\bar{x}_k}D_{\bar{x}_k,j}^{(t)}\|\mu\|^2\big]\Big| \\
&\le \frac{4\alpha}{n^{5/2}\sqrt{m}}\big[\phi'(\langle w_j^{(t)}, x_k\rangle)p + n\|\mu\|^2\big] + \frac{\alpha C_n\sqrt{p}}{\sqrt{m}} \\
&\le \frac{4\alpha}{n^{5/2}\sqrt{m}}\big[\phi'(\langle w_j^{(t)}, x_k\rangle)p + n\|\mu\|^2 + \frac{C_n n^{2-0.01}\|\mu\|^2}{4C}\big] \\
&\le \frac{4\alpha}{n^{5/2}\sqrt{m}}\big[\phi'(\langle w_j^{(t)}, x_k\rangle)p + \frac{C_n n^{2-0.01}\|\mu\|^2}{3C}\big],
\end{aligned}$$

where the second inequality uses Assumption (A1) and the last inequality holds for large enough n.

Now we turn to prove (A.32). Similar to (A.36), we have a decomposition for $\langle w_j^{(t+1)} - w_j^{(t)}, \nu\rangle$:

$$\langle w_j^{(t+1)} - w_j^{(t)}, \nu\rangle = \frac{\alpha a_j}{2n}\sum_{i=1}^n \phi'(\langle w_j^{(t)}, x_i\rangle)\langle y_i x_i, \nu\rangle + \frac{\alpha a_j}{n}\sum_{i=1}^n h_i^{(t)}\phi'(\langle w_j^{(t)}, x_i\rangle)\langle y_i x_i, \nu\rangle.$$

Similar to (A.33), we have

$$\Big|\frac{\alpha a_j}{n}\sum_{i=1}^n h_i^{(t)}\phi'(\langle w_j^{(t)}, x_i\rangle)y_i\langle x_i, \nu\rangle\Big| \le \frac{4\alpha}{n^{3/2}\sqrt{m}}\|\mu\|^2$$

by Lemma A.11 and $|\langle x_i, \nu\rangle| \le 2\|\mu\|^2$, which induced by (E1) in Lemma A.1. Similar to (A.35), we have

$$\Big|\sum_{i=1}^n \phi'(\langle w_j^{(t)}, x_i\rangle)\langle y_i x_i, \nu\rangle - y_\nu D_{\nu,j}^{(t)}\|\mu\|^2\Big| = \Big|\sum_{i=1}^n \phi'(\langle w_j^{(t)}, x_i\rangle)y_i\langle x_i - \bar{x}_i, \nu\rangle\Big| \le nC_n\|\mu\| \tag{A.37}$$

by (E1) in Lemma A.1. Combining the inequalities above, we have

$$\Big|\langle w_j^{(t+1)} - w_j^{(t)}, \nu\rangle - \frac{\alpha a_j}{2n}y_\nu D_{\nu,j}^{(t)}\|\mu\|^2\Big| \le \frac{4\alpha}{n^{3/2}\sqrt{m}}\|\mu\|^2 + \frac{\alpha C_n}{2\sqrt{m}}\|\mu\| \le \frac{5\alpha}{n^{3/2}\sqrt{m}}\|\mu\|^2$$

for large enough n. Here the last inequality uses

$$\|\mu\|^2 \ge Cn^{0.51}\sqrt{p} \ge C^{3/2}n^{1.51}\|\mu\|,$$

which comes from Assumptions (A1)-(A2). □

A.4.3 Proof of Corollary A.13

Corollary A.13. *Suppose that Assumptions (A1)-(A6) hold. Under a good run, for any pair $(j,k) \in [m]\times[n]$, the following is true:*

(F1) When $a_j y_k > 0$, if there exists some $0 \le s < 1/(\sqrt{n}p\alpha) - 2$ such that $\langle w_j^{(s)}, x_k\rangle > 0$, then for any $s \le t \le 1/(\sqrt{n}p\alpha) - 2$, we have $\langle w_j^{(t)}, x_k\rangle > 0$.

(F2) When $a_j y_k < 0$, for any $0 \le t \le 1/(\sqrt{n}p\alpha) - 2$ we have that $\langle w_j^{(t)}, x_k\rangle \le \frac{\alpha}{\sqrt{m}}\|\mu\|^2$.

(F3) When $a_j y_k < 0$, for any $0 \le t \le 1/(\sqrt{n}p\alpha) - 3$ we have that $\langle w_j^{(t)}, x_k\rangle > 0$ implies $\langle w_j^{(t+1)}, x_k\rangle < 0$.

Proof. **(F1):** It suffices to show the result holds for $t = s+1$, then by induction we can prove it for all $s \le t \le 1/(\sqrt{n}p\alpha) - 2$. Note that $a_j y_k = 1/\sqrt{m}$ and $\langle w_j^{(s)}, x_k\rangle > 0$, by (A.31), we have

$$\langle w_j^{(s+1)} - w_j^{(s)}, x_k\rangle \ge \frac{\alpha}{2n\sqrt{m}}(p - n\|\mu\|^2) - \frac{4\alpha}{n^{5/2}\sqrt{m}}\big[p + \frac{C_n n^{1.99}\|\mu\|^2}{3C}\big] \ge \frac{\alpha p}{4n\sqrt{m}} > 0, \tag{A.38}$$

where the second inequality uses Assumption (A2).

(F2): We prove (F2) by induction. Denote

$$Q(t): \quad \langle w_j^{(t)}, x_k \rangle \le \frac{\alpha}{\sqrt{m}} \|\mu\|^2.$$

When $t = 0$, by the definition of a good run, we have

$$|\langle w_j^{(0)}, x_k \rangle| \le \|w_j^{(0)}\| \cdot \|x_k\| \le \|W^{(0)}\|_F \cdot \sqrt{2p} \le \omega_{\text{init}} p\sqrt{3m} \le \frac{\alpha}{Cn\sqrt{m}} \|\mu\|^2, \tag{A.39}$$

where the second inequality uses Lemma A.1; the third inequality uses Lemma A.3; and the last inequality is from Assumption (A5). Thus $Q(0)$ holds. Suppose $Q(t)$ holds and $t \le 1/(\sqrt{n}p\alpha) - 3$. If $\langle w_j^{(t)}, x_k \rangle < 0$, we have

$$\langle w_j^{(t+1)}, x_k \rangle \le \langle w_j^{(t+1)} - w_j^{(t)}, x_k \rangle \le \frac{\alpha a_j y_{\bar{x}_k}}{2n} D_{\bar{x}_k, j}^{(t)} \|\mu\|^2 + \frac{4\alpha C_n}{3Cn^{0.51}\sqrt{m}} \|\mu\|^2 \le \frac{\alpha}{\sqrt{m}} \|\mu\|^2,$$

where the second inequality uses (A.31) and $\phi'(\langle w_j^{(t)}, x_k \rangle) = 0$; and the third inequality uses $D_{\nu,j}^{(t)} \le n$ and n is large enough. If $\langle w_j^{(t)}, x_k \rangle > 0$, we have

$$\begin{aligned} \langle w_j^{(t+1)} - w_j^{(t)}, x_k \rangle &\le -\frac{\alpha}{2n\sqrt{m}} (p - n\|\mu\|^2) + \frac{4\alpha}{n^{5/2}\sqrt{m}} [p + \frac{C_n n^{1.99} \|\mu\|^2}{3C}] \\ &\le -\frac{\alpha}{2n\sqrt{m}} (p - n\|\mu\|^2) + \frac{8\alpha p}{n^{5/2}\sqrt{m}}, \end{aligned}$$

where the first inequality uses (A.31) and $\phi'(\langle w_j^{(t)}, x_k \rangle) = 1$; and the second inequality uses Assumption (A2). Combined with the inductive hypothesis, we have

$$\langle w_j^{(t+1)}, x_k \rangle = \langle w_j^{(t)}, x_k \rangle + \langle w_j^{(t+1)} - w_j^{(t)}, x_k \rangle \le \frac{\alpha}{\sqrt{m}} \|\mu\|^2 - \frac{\alpha}{2n\sqrt{m}} (p - n\|\mu\|^2) + \frac{8\alpha p}{n^{5/2}\sqrt{m}} < 0$$

by Assumption (A2). Thus $Q(t+1)$ holds. And (F3) is also proved by the last inequality. □

A.4.4 Proof of Lemma A.14

Since the analysis on one cluster can be similarly replicated on other clusters, below we will focus on analyzing the cluster centered at $+\mu_1$. Given the training set, $D_{+\mu_1, j}^{(0)}$ is a function of the random initialization $w_j^{(0)}$. $D_{+\mu_1, j}^{(0)}$ plays an important role in determining the direction that $w_j^{(t)}, t \ge 1$ aligns with and the sign of the inner product $\langle w_j^{(t)}, x_k \rangle$. For $\bar{x}_k \in \{\pm\mu_1\}$, $y_{\bar{x}_k} = 1$. Then for each $t \le 1/(\sqrt{n}p\alpha) - 2$, (A.31) is simplified to

$$\left| \langle w_j^{(t+1)} - w_j^{(t)}, x_k \rangle - \frac{\alpha a_j y_k p}{2n} \right| \le \frac{4\alpha p}{n^{5/2}\sqrt{m}} + \frac{\alpha}{2\sqrt{m}} \|\mu\|^2, \quad \text{when } \langle w_j^{(t)}, x_k \rangle > 0; \tag{A.40}$$

$$\left| \langle w_j^{(t+1)} - w_j^{(t)}, x_k \rangle - \frac{\alpha a_j}{2n} D_{\bar{x}_k, j}^{(t)} \|\mu\|^2 \right| \le \frac{4\alpha C_n}{3Cn^{0.01}\sqrt{mn}} \|\mu\|^2, \quad \text{when } \langle w_j^{(t)}, x_k \rangle \le 0. \tag{A.41}$$

Here $C_n = 10\sqrt{\log(n)}$ is defined in Lemma A.12. We will elaborate on the outcomes for neurons with $a_j > 0$ and $a_j < 0$ separately in the following lemmas.

Lemma A.14. *Suppose that Assumptions (A1)-(A6) hold. Under a good run, we have that for any $j \in \mathcal{J}_{+\mu_1, \mathtt{P}}^{20\varepsilon}$ (or equivalently, for any neuron $j \in \mathcal{J}_{\mathtt{Pos}}$ that is $(\mu_1, 20\varepsilon)$-aligned)), the followings hold for $1 \le t \le 1/(\sqrt{n}p\alpha) - 2$:*

(G1)

$$\mathcal{C}_{+\mu_1, j}^{(t)} = \mathcal{C}_{+\mu_1}; \quad \mathcal{C}_{-\mu_1, j}^{(t)} = \mathcal{C}_{-\mu_1, j}^{(0)}; \quad \mathcal{N}_{-\mu_1, j}^{(t)} = \varnothing; \quad D_{+\mu_1, j}^{(t)} > c_{+\mu_1} - n_{+\mu_1} - d_{-\mu_1, j}^{(0)}.$$

(G2)

$$\langle w_j^{(t)} - w_j^{(t-1)}, \mu_1 \rangle \ge \frac{\alpha}{4n\sqrt{m}} D_{+\mu_1, j}^{(t-1)} \|\mu\|^2.$$

Proof. Given $j \in \mathcal{J}^{20\varepsilon}_{+\mu_1,\mathrm{P}}$, when $t=0$, for $x_k \in \mathcal{C}^{(0)}_{+\mu_1,j}$, we have $a_j y_k > 0$. Thus by Corollary A.13, we have

$$x_k \in \mathcal{C}^{(t)}_{+\mu_1,j}, \quad 0 \le t \le 1/(\sqrt{n}p\alpha) - 2. \tag{A.42}$$

Similarly we have that for $x_k \in \mathcal{C}^{(0)}_{-\mu_1,j}$,

$$x_k \in \mathcal{C}^{(t)}_{-\mu_1,j}, \quad 0 \le t \le 1/(\sqrt{n}p\alpha) - 2; \tag{A.43}$$

and for $x_k \in \mathcal{N}^{(0)}_{-\mu_1,j}$, $x_k \notin \mathcal{N}^{(1)}_{-\mu_1,j}$ since $a_j y_k < 0$.

Next for $x_k \in \mathcal{C}_{+\mu_1} \backslash \mathcal{C}^{(0)}_{+\mu_1,j}$, we have

$$\begin{aligned}\langle w_j^{(1)} - w_j^{(0)}, x_k\rangle &\ge \frac{\alpha a_j}{2n} D^{(0)}_{+\mu_1,j}\|\mu\|^2 - \frac{4\alpha C_n}{3Cn^{0.01}\sqrt{mn}}\|\mu\|^2 \\ &\ge \frac{\alpha}{2n^{20\varepsilon}\sqrt{mn}}\|\mu\|^2 - \frac{4\alpha C_n}{3Cn^{0.01}\sqrt{mn}}\|\mu\|^2 \ge \frac{\alpha}{4n^{20\varepsilon}\sqrt{mn}}\|\mu\|^2,\end{aligned} \tag{A.44}$$

where the first inequality is from (A.41); the second inequality uses $D^{(0)}_{+\mu_1,j} > n^{1/2-20\varepsilon}$, which is from $j \in \mathcal{J}^{20\varepsilon}_{+\mu_1,\mathrm{P}}$; and the last inequality uses $40\varepsilon < 0.01$. It yields that

$$\langle w_j^{(1)}, x_k\rangle \ge \langle w_j^{(1)} - w_j^{(0)}, x_k\rangle - \|w_j^{(0)}\| \cdot \|x_k\| \ge \frac{\alpha}{4n^{20\varepsilon}\sqrt{mn}}\|\mu\|^2 - \frac{\alpha}{Cn\sqrt{m}}\|\mu\|^2 > 0, \tag{A.45}$$

where the second inequality uses (A.39). Thus we have

$$\mathcal{C}_{+\mu_1} \backslash \mathcal{C}^{(0)}_{+\mu_1,j} \subseteq \mathcal{C}^{(1)}_{+\mu_1,j}.$$

Combined with (A.42), we obtain $\mathcal{C}^{(1)}_{+\mu_1,j} = \mathcal{C}_{+\mu_1}$. Then by Corollary A.13, we have

$$\mathcal{C}^{(t)}_{+\mu_1,j} = \mathcal{C}_{+\mu_1}, \quad 0 \le t \le 1/(\sqrt{n}p\alpha) - 2.$$

For $x_k \in \big(\mathcal{C}_{-\mu_1} \backslash \mathcal{C}^{(0)}_{-\mu_1,j}\big) \cup \big(\mathcal{N}_{-\mu_1} \backslash \mathcal{N}^{(0)}_{-\mu_1,j}\big)$, Following similar analysis of (A.45), we have

$$\langle w_j^{(1)}, x_k\rangle \le \langle w_j^{(1)} - w_j^{(0)}, x_k\rangle + \|w_j^{(0)}\| \cdot \|x_k\| \le -\Big(\frac{\alpha}{4n^{20\varepsilon}\sqrt{mn}}\|\mu\|^2 - \frac{\alpha}{Cn\sqrt{m}}\|\mu\|^2\Big) < 0. \tag{A.46}$$

Thus we have $\mathcal{C}_{-\mu_1} \backslash \mathcal{C}^{(0)}_{-\mu_1,j} \notin \mathcal{C}^{(1)}_{-\mu_1,j}$, and $\mathcal{N}_{-\mu_1} \backslash \mathcal{N}^{(0)}_{-\mu_1,j} \notin \mathcal{N}^{(1)}_{-\mu_1,j}$. Combined with (A.43) and $\mathcal{N}^{(0)}_{-\mu_1,j} \notin \mathcal{N}^{(1)}_{-\mu_1,j}$, we obtain

$$\mathcal{C}^{(1)}_{-\mu_1,j} = \mathcal{C}^{(0)}_{-\mu_1,j}; \quad \mathcal{N}^{(1)}_{-\mu_1,j} = \varnothing.$$

It yields that

$$D^{(1)}_{+\mu_1,j} = c_{+\mu_1} - |\mathcal{N}^{(1)}_{+\mu_1,j}| - |\mathcal{C}^{(0)}_{-\mu_1,j}| > c_{+\mu_1} - n_{+\mu_1} - d^{(0)}_{-\mu_1,j} > \sqrt{n},$$

where the last inequality uses $d^{(0)}_{+\mu_1,j} < \min\{c_{+\mu_1}, c_{-\mu_1}\} - 2n_{\pm\mu_1} - \sqrt{n}$ and

$$c_{+\mu_1} - n_{+\mu_1} - d^{(0)}_{-\mu_1,j} > \sqrt{n} + d^{(0)}_{+\mu_1,j} - d^{(0)}_{-\mu_1,j} > \sqrt{n}.$$

Thus (G1) holds for $t=1$. Then (G1) is proved by replicating the same analysis and employing induction.

For the inner product with the cluster mean $+\mu_1$, by (A.32) we have

$$\langle w_j^{(t+1)} - w_j^{(t)}, \mu_1\rangle \ge \frac{\alpha}{2n\sqrt{m}} D^{(t)}_{+\mu_1,j}\|\mu\|^2 - \frac{5C_n\alpha}{n^{3/2}\sqrt{m}}\|\mu\|^2 \ge \frac{\alpha}{4n\sqrt{m}} D^{(t)}_{+\mu_1,j}\|\mu\|^2,$$

where the last inequality uses $D^{(t)}_{+\mu_1,j} > 0$. □

A.4.5 Proof of Lemma A.15

Lemma A.15. *Suppose that Assumptions (A1)-(A6) hold. Under a good run, for any* $j \in \mathcal{J}^{20\varepsilon}_{+\mu_1,\mathtt{N}} \cup \mathcal{J}^{20\varepsilon}_{-\mu_1,\mathtt{N}}$ *(or equivalently, for any neuron* $j \in \mathcal{J}_{\mathtt{Neg}}$ *that is* $(\pm\mu_1, 20\varepsilon)$*-aligned), the followings hold for* $2 \leq t \leq 1/(\sqrt{n}p\alpha) - 2$.

$$\mathcal{N}^{(t)}_{+\mu_1,j} = \mathcal{N}_{+\mu_1}, \mathcal{N}^{(t)}_{-\mu_1,j} = \mathcal{N}_{-\mu_1}; \tag{A.47}$$

$$-n - \Delta_{\mu_1}(t-2) \leq \sum_{s=0}^{t} D^{(s)}_{\nu,j} \leq n + \Delta_{\mu_1}(t-2), \quad \nu \in \{\pm\mu_1\}, \tag{A.48}$$

where $\Delta_{\mu_1} := |n_{+\mu_1} - n_{-\mu_1}| + \sqrt{n}$.

Proof. For a given $\nu \in \{\pm\mu_1\}$, suppose $j \in \mathcal{J}^{20\varepsilon}_{\nu,\mathtt{N}}$. Then we have

$$a_j < 0; \quad D^{(0)}_{\nu,j} > n^{1/2-20\varepsilon}; \quad d^{(0)}_{\nu,j} \leq \min\{c_\nu, c_{-\nu} - 2n_{\pm\nu} - \sqrt{n}\} \tag{A.49}$$

according to the definition (A.16). Note that we study the same data as in Lemma A.14 and only $\mathrm{sgn}(a_j)$ is flipped in the trajectory analysis compared to the setting in Lemma A.14, our analysis in the first two iterations follows similar procedures in Lemma A.14. For $x_k \in \mathcal{C}^{(0)}_{\nu,j} \cup \mathcal{C}^{(0)}_{-\nu,j}$, $a_j y_k < 0$, by Corollary A.13, we have

$$\langle w^{(1)}_j, x_k \rangle < 0. \tag{A.50}$$

For $x_k \in \mathcal{N}^{(0)}_{\nu,j} \cup \mathcal{N}^{(0)}_{-\nu,j}$, $a_j y_k > 0$, by Corollary A.13, we have

$$\langle w^{(t)}_j, x_k \rangle > 0 \tag{A.51}$$

for any $t \leq 1/(\sqrt{n}p\alpha) - 2$. For $x_k \in \big(\mathcal{C}_\nu \backslash \mathcal{C}^{(0)}_{\nu,j}\big) \cup \big(\mathcal{N}_\nu \backslash \mathcal{N}^{(0)}_{\nu,j}\big)$, similar to (A.44), we have

$$\langle w^{(1)}_j - w^{(0)}_j, x_k \rangle \leq -\big(\frac{\alpha a_j}{2n} D^{(0)}_{+\mu_1,j}\|\mu\|^2 - \frac{4\alpha C_n}{3Cn^{0.01}\sqrt{mn}}\|\mu\|^2\big) \leq -\frac{\alpha}{4n^{20\varepsilon}\sqrt{mn}}\|\mu\|^2 < 0,$$

then similar to (A.45), we have

$$\langle w^{(1)}_j, x_k \rangle \leq -\langle w^{(1)}_j - w^{(0)}_j, x_k \rangle + \|w^{(0)}_j\| \cdot \|x_k\| \leq -\frac{\alpha}{4n^{20\varepsilon}\sqrt{mn}}\|\mu\|^2 + \frac{\alpha}{Cn\sqrt{m}}\|\mu\|^2 < 0. \tag{A.52}$$

For $x_k \in \big(\mathcal{C}_{-\nu} \backslash \mathcal{C}^{(0)}_{-\nu,j}\big) \cup \big(\mathcal{N}_{-\nu} \backslash \mathcal{N}^{(0)}_{-\nu,j}\big)$, similar to (A.46), we have

$$\langle w^{(1)}_j, x_k \rangle \geq \langle w^{(1)}_j - w^{(0)}_j, x_k \rangle - \|w^{(0)}_j\| \cdot \|x_k\| \geq \frac{\alpha}{4n^{20\varepsilon}\sqrt{mn}}\|\mu\|^2 - \frac{\alpha}{Cn\sqrt{m}}\|\mu\|^2 > 0. \tag{A.53}$$

Combining (A.50)-(A.53), we have

$$\mathcal{C}^{(1)}_{\nu,j} = \varnothing; \quad \mathcal{C}^{(1)}_{-\nu,j} = \mathcal{C}_{-\nu} \backslash \mathcal{C}^{(0)}_{-\nu,j}; \quad \mathcal{N}^{(1)}_{\nu,j} = \mathcal{N}^{(0)}_{\nu,j}; \quad \mathcal{N}^{(1)}_{-\nu,j} = \mathcal{N}_{-\nu}. \tag{A.54}$$

Thus by the definition of $D^{(1)}_{\nu,j}$, we have

$$D^{(1)}_{\nu,j} = -|\mathcal{N}^{(0)}_{\nu,j}| - c_{-\nu} + |\mathcal{C}^{(0)}_{-\nu,j}| + n_{-\nu} \leq -|\mathcal{N}^{(0)}_{\nu,j}| - c_{-\nu} + d^{(0)}_{-\nu,j} + 2n_{-\nu}. \tag{A.55}$$

It further yields that

$$D^{(1)}_{\nu,j} + D^{(0)}_{\nu,j} \leq -|\mathcal{N}^{(0)}_{\nu,j}| - c_{-\nu} + 2n_{-\nu} + d^{(0)}_{\nu,j} \leq -c_{-\nu} + 2n_{-\nu} + d^{(0)}_{\nu,j} < -\sqrt{n},$$

where the first inequality uses (A.55) and the definition of $D^{(0)}_{\nu,j}$, and the third inequality uses (A.49).

After the second iteration, for $x_k \in \mathcal{N}_\nu \backslash \mathcal{N}_{\nu,j}^{(1)}$, $\langle w_j^{(0)}, x_k\rangle < 0, \langle w_j^{(1)}, x_k\rangle < 0$. Then we have

$$\langle w_j^{(2)} - w_j^{(0)}, x_k\rangle \geq -\frac{\alpha}{2n\sqrt{m}}(D_{\nu,j}^{(0)} + D_{\nu,j}^{(1)})\|\mu\|^2 - \frac{4\alpha C_n}{3Cn^{0.01}\sqrt{mn}}\|\mu\|^2$$
$$> \frac{\alpha}{2\sqrt{mn}}\|\mu\|^2 - \frac{4\alpha C_n}{3Cn^{0.01}\sqrt{mn}}\|\mu\|^2,$$

where the first inequality uses (A.41), and the second inequality uses $D_{\nu,j}^{(1)} + D_{\nu,j}^{(0)} < -\sqrt{n}$. It further yields that

$$\langle w_j^{(2)}, x_k\rangle \geq \langle w_j^{(2)} - w_j^{(0)}, x_k\rangle - \|w_j^{(0)}\|\cdot\|x_k\| \geq \frac{\alpha}{2\sqrt{mn}}\|\mu\|^2 - \frac{4\alpha C_n}{3Cn^{0.01}\sqrt{mn}}\|\mu\|^2 - \frac{\alpha}{Cn\sqrt{m}}\|\mu\|^2 > 0. \tag{A.56}$$

For $x_k \in \mathcal{N}_{\nu,j}^{(1)} \cup \mathcal{N}_{-\nu}$, note that $a_j y_k > 0$. Then by Corollary A.13, we have $\langle w_j^{(2)}, x_k\rangle > 0$. Combined with (A.56), we obtain $\mathcal{N}_{\nu,j}^{(2)} = \mathcal{N}_\nu, \mathcal{N}_{-\nu,j}^{(2)} = \mathcal{N}_{-\nu}$. Again by Corollary A.13, we have that for $2 \leq t \leq 1/(\sqrt{n}p\alpha) - 2$,

$$\mathcal{N}_{\nu,j}^{(t)} = \mathcal{N}_\nu, \quad \mathcal{N}_{-\nu,j}^{(t)} = \mathcal{N}_{-\nu}, \tag{A.57}$$

i.e. for $t \geq 2$, neurons with $j \in \mathcal{J}_{\nu,\mathtt{N}}^{20\varepsilon} \cup \mathcal{J}_{-\nu,\mathtt{N}}^{20\varepsilon}$ are active for all noisy points in $\mathcal{N}_{\pm\mu_1}$, which proves (A.47).

For $x_k \in \mathcal{C}_{-\nu,j}^{(1)}$, note that $a_j y_k < 0$ and $\langle w_j^{(1)}, x_k\rangle > 0$. Then by Corollary A.13, we have $\langle w_j^{(2)}, x_k\rangle < 0$. For $x_k \in \mathcal{C}_{-\nu} \backslash \mathcal{C}_{-\nu,j}^{(1)}$, by (A.54) we have $\langle w_j^{(0)}, x_k\rangle > 0, \langle w_j^{(1)}, x_k\rangle < 0$. It yields that

$$\langle w_j^{(2)} - w_j^{(0)}, x_k\rangle \leq -\frac{\alpha}{2n\sqrt{m}}(p + D_{\nu,j}^{(1)}\|\mu\|^2) + \frac{4\alpha p}{n^{5/2}\sqrt{m}} + \frac{\alpha}{2\sqrt{m}}\|\mu\|^2 + \frac{4\alpha C_n}{3Cn^{0.01}\sqrt{mn}}\|\mu\|^2 \leq -\frac{\alpha p}{4n\sqrt{m}},$$

where the first inequality uses (A.40) and (A.41), and the second inequality uses Assumption (A2). It further yields that

$$\langle w_j^{(2)}, x_k\rangle < \langle w_j^{(2)} - w_j^{(0)}, x_k\rangle + \|w_j^{(0)}\| \cdot \|x_k\| \leq -\frac{\alpha p}{4n\sqrt{m}} + \frac{\alpha}{Cn\sqrt{m}}\|\mu\|^2 < 0 \tag{A.58}$$

by Assumption (A2). Thus we have $\mathcal{C}_{-\nu,j}^{(2)} = \varnothing$.

For $x_k \in \mathcal{C}_{\nu,j}^{(0)}$, $\langle w_j^{(0)}, x_k\rangle > 0, \langle w_j^{(1)}, x_k\rangle < 0$, which is similar to the setting of $\mathcal{C}_{-\nu} \backslash \mathcal{C}_{-\nu,j}^{(1)}$. Repeating the analysis above, we have

$$\langle w_j^{(2)}, x_k\rangle < 0.$$

For $x_k \in \mathcal{C}_\nu \backslash \mathcal{C}_{\nu,j}^{(0)}$, note that $\langle w_j^{(0)}, x_k\rangle < 0, \langle w_j^{(1)}, x_k\rangle < 0$, then we have

$$\langle w_j^{(2)} - w_j^{(0)}, x_k\rangle \geq -\frac{\alpha}{2n\sqrt{m}}(D_{\nu,j}^{(0)} + D_{\nu,j}^{(1)})\|\mu\|^2 - \frac{4\alpha C_n}{3Cn^{0.01}\sqrt{mn}}\|\mu\|^2$$
$$> \frac{\alpha}{2\sqrt{mn}}\|\mu\|^2 - \frac{4\alpha C_n}{3Cn^{0.01}\sqrt{mn}}\|\mu\|^2 > 0,$$

where the first inequality uses (A.41) and the second inequality uses (A.55). Combining the inequalities above, we obtain

$$\mathcal{C}_{\nu,j}^{(2)} = \mathcal{C}_\nu \backslash \mathcal{C}_{\nu,j}^{(0)}; \quad \mathcal{C}_{-\nu,j}^{(2)} = \varnothing; \quad \mathcal{N}_{\nu,j}^{(2)} = \mathcal{N}_\nu; \quad \mathcal{N}_{-\nu,j}^{(2)} = \mathcal{N}_{-\nu}. \tag{A.59}$$

Combining (A.54) and (A.59), we have

$$\sum_{s=0}^{2} D_{\nu,j}^{(s)} = c_\nu - c_{-\nu} - n_\nu + 3n_{-\nu} - 2|\mathcal{N}_\nu^{(0)}|,$$

and it yields that

$$c_\nu - c_{-\nu} - 3n_\nu + 3n_{-\nu} \le \sum_{s=0}^{2} D_{\nu,j}^{(s)} \le c_\nu - c_{-\nu} + 3n_{-\nu} - n_\nu.$$

It remains to prove (A.48). It suffices to prove

$$c_\nu - 2c_{-\nu} - 4n_\nu + 3n_{-\nu} - \Delta_{\mu_1}(t-2) \le \sum_{s=0}^{t} D_{\nu,j}^{(s)} \le (2c_\nu - c_{-\nu} + 4n_{-\nu} - n_\nu) + \Delta_{\mu_1}(t-2), \nu \in \{\pm\mu_1\},$$

since $2c_\nu - c_{-\nu} + 4n_{-\nu} - n_\nu \le n$ and $c_\nu - 2c_{-\nu} - 4n_\nu + 3n_{-\nu} \ge -n$ by Lemma A.1. Without loss of generality, below we only show the proof of the right-hand side. Denote $\mathcal{T} = \{t \in [T], t \ge 3, D_{\nu,j}^{(t)} > \Delta_{\mu_1}\} = \{t_i\}_{i=1}^{K}, t_1 < t_2 < \cdots < t_K$. To prove the right-hand side of (A.48), it suffices to show that the followings hold

$$\sum_{t=t_i}^{s} D_{\nu,j}^{(t)} \le c_\nu + n_{-\nu} + \Delta_{\mu_1}(s - t_i); \tag{A.60}$$

$$\sum_{t=t_i}^{t_{i+1}-1} D_{\nu,j}^{(t)} \le \Delta_{\mu_1}(t_{i+1} - t_i) \tag{A.61}$$

for any $i \in [K]$ and all $s \in [t_i, t_{i+1} - 2]$. (A.60) directly follows from the definition of the set $\mathcal{T}$ and the fact that $D_{\nu,j}^{(t)} \le c_\nu + n_{-\nu}$ for any j, t. For a given $t_i, t_i \in \mathcal{T}$, we have $D_{\nu,j}^{(t_i)} > \Delta_{\mu_1} \ge \sqrt{n}$. By (A.41), we have that for any $x_k \in \mathcal{C}_\nu \backslash \mathcal{C}_\nu^{(t_i)}(j)$,

$$\begin{aligned}\langle w_j^{(t_i+1)}, x_k\rangle &\le \langle w_j^{(t_i+1)} - w_j^{(t_i)}, x_k\rangle \le -\frac{\alpha}{2n\sqrt{m}} D_{\nu,j}^{(t_i)} \|\mu\|^2 + \frac{4\alpha C_n}{3Cn^{0.01}\sqrt{mn}}\|\mu\|^2 \\ &\le -\frac{\alpha}{4n\sqrt{m}} D_{\nu,j}^{(t_i)} \|\mu\|^2 < 0,\end{aligned} \tag{A.62}$$

which implies that $w_j^{(t_i+1)}$ is still inactive for those x_k that didn't activate $w_j^{(t_i)}$. For any $x_k \in \mathcal{C}_{\nu,j}^{(t_i)}$, since $a_j y_k < 0$, by Corollary A.13, we have

$$\langle w_j^{(t_i)}, x_k\rangle \le \frac{\alpha\|\mu\|^2}{\sqrt{m}}.$$

Combined with (A.40), we have

$$\begin{aligned}\langle w_j^{(t_i+1)}, x_k\rangle &= \langle w_j^{(t_i+1)} - w_j^{(t_i)}, x_k\rangle + \langle w_j^{(t_i)}, x_k\rangle \\ &\le -\frac{\alpha p}{2n\sqrt{m}} + \frac{4\alpha p}{n^{5/2}\sqrt{m}} + \frac{3\alpha}{2\sqrt{m}}\|\mu\|^2 \le -\frac{\alpha p}{4n\sqrt{m}} < 0\end{aligned} \tag{A.63}$$

where the second inequality uses Assumption (A2). Combining (A.62) and (A.63), we have $\mathcal{C}_{\nu,j}^{(t_i+1)} = \varnothing$, and

$$\langle w_j^{(t_i+1)}, x_k\rangle \le -\frac{\alpha}{2n\sqrt{m}} D_{\nu,j}^{(t_i)} \|\mu\|^2 + \frac{4\alpha C_n}{3Cn^{0.01}\sqrt{mn}}\|\mu\|^2 \tag{A.64}$$

for all $x_k \in \mathcal{C}_\nu$. It yields that

$$D_{\nu,j}^{(t_i+1)} = |\mathcal{C}_{\nu,j}^{(t_i+1)}| - |\mathcal{C}_{-\nu,j}^{(t_i+1)}| + n_{-\nu} - n_\nu = -|\mathcal{C}_{-\nu,j}^{(t_i+1)}| + n_{-\nu} - n_\nu \le |n_{+\mu_1} - n_{-\mu_1}|,$$

where the first equation uses (A.47). It implies that $t_{i+1} - t_i > 1$. Let $t_i^\star = \min\{t \in \mathbb{N} : t_i + 1 < t \le t_{i+1}, \mathcal{C}_\nu^{(t)}(j) \ne \varnothing\}$. We claim that $t_i^\star$ is well-defined for each i, because $\mathcal{C}_\nu^{(t_{i+1})}(j) \ne \varnothing$. Otherwise we have $D_{\nu,j}^{(t_{i+1})} \le |n_{+\mu_1} - n_{-\mu_1}| < \Delta_{\mu_1}$, which contradicts to the definition of the set $\mathcal{T}$. Thus $t_i^\star$ always exists. Choose one point from the set $\mathcal{C}_{\nu,j}^{(t_i^\star)}$ and denote it as $x_k^\star$. Note that for any

$t \in [t_i+1, t_i^\star - 1]$, we have $\mathcal{C}_\nu^{(t)}(j) = \varnothing$, $D_{\nu,j}^{(t)} \le |n_{+\mu_1} - n_{-\mu_1}|$, and by (A.41),

$$\langle w_j^{(t+1)} - w_j^{(t)}, x_k^\star \rangle \le -\frac{\alpha}{2n\sqrt{m}} D_{\nu,j}^{(t)} \|\mu\|^2 + \frac{4\alpha C_n}{3Cn^{0.01}\sqrt{mn}} \|\mu\|^2.$$

Combined with (A.64), it yields that

$$\begin{aligned} 0 \le \langle w_j^{(t_i^\star)}, x_k^\star \rangle &= \sum_{t=t_i+1}^{t_i^\star - 1} \langle w_j^{(t+1)} - w_j^{(t)}, x_k^\star \rangle + \langle w_j^{(t_i+1)}, x_k^\star \rangle \\ &\le -\frac{\alpha \|\mu\|^2}{2n\sqrt{m}} \Big(D_{\nu,j}^{(t_i)} + \sum_{t=t_i+1}^{t_i^\star-1} D_{\nu,j}^{(t)} - \frac{4\sqrt{n}C_n}{3Cn^{0.01}} (t_i^\star - t_i) \Big). \end{aligned}$$

It further yields that

$$\sum_{t=t_i}^{t_i^\star - 1} D_{\nu,j}^{(t)} \le \frac{4\sqrt{n}C_n}{3Cn^{0.01}} (t_i^\star - t_i) \le \sqrt{n}(t_i^\star - t_i).$$

If $t_i^\star = t_{i+1}$, then we've proved (A.61). If $t_i^\star < t_{i+1}$, then we have

$$\sum_{t=t_i}^{t_{i+1}-1} D_{\nu,j}^{(t)} = \sum_{t=t_i}^{t_i^\star-1} D_{\nu,j}^{(t)} + \sum_{t=t^\star}^{t_{i+1}-1} D_{\nu,j}^{(t)} \le \sqrt{n}(t^\star - t_i) + \Delta_{\mu_1}(t_{i+1} - t^\star) \le \Delta_{\mu_1}(t_{i+1} - t_i),$$

which proves the right side. For the left side, similarly we denote $\mathcal{T}_- = \{t \in [T], t \ge 3, D_{\nu,j}^{(t)} < -\Delta_{\mu_1}\} = \{t_i\}_{i=1}^K, t_1 < t_2 < \cdots < t_K$. Following the same analysis, we can prove that the followings hold

$$\sum_{t=t_i}^{s} D_{\nu,j}^{(t)} \ge -c_{-\nu} - n_\nu - \Delta_{\mu_1}(s - t_i); \quad \sum_{t=t_i}^{t_{i+1}-1} D_{\nu,j}^{(t)} \ge -\Delta_{\mu_1}(t_{i+1} - t_i)$$

for any $i \in [K]$ and all $s \in [t_i, t_{i+1} - 2]$. It proves the left-hand side of (A.48). $\square$

A.5 Proof of the Main Theorem

We rigorously prove Theorem 3.1 in this section. The upper bound of t in the theorems below is $1/(\sqrt{n}p\alpha) - 2$, which by Assumption (A4), is larger than $\sqrt{n}$, the upper bound of t in Theorem 3.1.

A.5.1 Proof of Theorem A.16: 1-step Overfitting

Theorem A.16. *Suppose that Assumptions (A1)-(A6) hold. Under a good run, the classifier* $\mathrm{sgn}(f(x, W^{(t)}))$ *can correctly classify all training datapoints for* $1 \le t \le 1/(\sqrt{n}p\alpha) - 2$.

Proof. Without loss of generality, we only consider datapoints in the cluster $\mathcal{C}_{+\mu_1} \cup \mathcal{N}_{+\mu_1}$. According to (B1) in Lemma A.4, we have that under a good run, $|\mathcal{J}_{\mathtt{P}}^{i,(0)}| \ge m/7, |\mathcal{J}_{\mathtt{N}}^{i,(0)}| \ge m/7$ for each $i \in [n]$. For $x_k \in \mathcal{C}_{+\mu_1}$, by Corollary A.13, we have

$$\langle w_j^{(s)}, x_k \rangle > 0$$

for all $j \in \mathcal{J}_{\mathtt{P}}^{k,(0)}$ and $0 \le s \le 1/(\sqrt{n}p\alpha) - 2$; and

$$\langle w_j^{(s)}, x_k \rangle \le \frac{\alpha}{\sqrt{m}} \|\mu\|^2$$

for all $j \in \mathcal{J}_{\texttt{N}}$ and $0 \le s \le 1/(\sqrt{n}p\alpha) - 2$. Then for $1 \le t \le 1/(\sqrt{n}p\alpha) - 2$, we have

$$\begin{aligned}
\sum_{j=1}^{m} a_j \phi(\langle w_j^{(t)}, x_k \rangle) &\ge \sum_{j \in \mathcal{J}_{\texttt{P}}^{k,(0)}} \frac{1}{\sqrt{m}} \phi(\langle w_j^{(t)}, x_k \rangle) - \sum_{j: a_j < 0} \frac{1}{\sqrt{m}} \phi(\langle w_j^{(t)}, x_k \rangle) \\
&\ge \sum_{j \in \mathcal{J}_{\texttt{P}}^{k,(0)}} \sum_{s=0}^{t-1} \frac{1}{\sqrt{m}} \langle w_j^{(s+1)} - w_j^{(s)}, x_k \rangle - \sum_{j: a_j < 0} \frac{\alpha}{m} \|\mu\|^2 \\
&\ge \frac{\alpha p t}{4nm} |\mathcal{J}_{\texttt{P}}^{k,(0)}| - \frac{\alpha |\mathcal{J}_{\texttt{N}}|}{m} \|\mu\|^2 \\
&\ge \frac{\alpha p t}{28n} - \alpha \|\mu\|^2 > 0,
\end{aligned}$$

where the first inequality uses $\phi(x) \ge 0, \forall x$; the second inequality uses the definition of $\mathcal{J}_{\texttt{P}}^{k,(0)}$ and (F2) in Corollary A.13; the third inequality uses (A.38) in Corollary A.13; and the last inequality is from Assumption (A2). For $x_k \in \mathcal{N}_{+\mu_1}$, similarly we have

$$\begin{aligned}
\sum_{j=1}^{m} a_j \phi(\langle w_j^{(t)}, x_k \rangle) &\le - \sum_{j \in \mathcal{J}_{\texttt{N}}^{k,(0)}} \frac{1}{\sqrt{m}} \phi(\langle w_j^{(t)}, x_k \rangle) + \sum_{j: a_j > 0} \frac{1}{\sqrt{m}} \phi(\langle w_j^{(t)}, x_k \rangle) \\
&\le - \sum_{j \in \mathcal{J}_{\texttt{N}}^{k,(0)}} \sum_{s=1}^{t} \frac{1}{\sqrt{m}} \langle w_j^{(s)} - w_j^{(s-1)}, x_k \rangle + \sum_{j: a_j > 0} \frac{\alpha}{\sqrt{m}} \|\mu\|^2 \\
&\le -\left(\frac{\alpha p t}{28n} - \alpha \|\mu\|^2\right) < 0.
\end{aligned}$$

Thus our classifier can correctly classify all training datapoints for $1 \le t \le 1/(\sqrt{n}p\alpha) - 2$. □

A.5.2 Proof of Theorem A.8: Generalization

Before proceeding with the proof of Theorem A.8, we first state a technical lemma:

Lemma A.17. *Suppose that* $\|W\| > 0$*. Then there exists a constant* $c > 0$ *such that*

$$\mathbb{P}_{(x,\widetilde{y}) \sim P_{clean}}(\widetilde{y} \neq \operatorname{sgn}(f(x;W))) \le \max_{\nu \in \mathsf{centers}} 2 \exp\left(-c \left(\frac{\mathbb{E}_{x \sim N(\nu, I_p)}[f(x;W)]}{\|W\|_F}\right)^2\right).$$

Proof. It suffices to prove that for each $\nu \in \mathsf{centers}$,

$$\mathbb{P}_{x \sim N(\nu, I_p)}(y_\nu f(x;W) < 0) \le 2 \exp\left(-c \left(\frac{\mathbb{E}_{x \sim N(\nu, I_p)}[f(x;W)]}{\|W\|_F}\right)^2\right). \tag{A.65}$$

Then applying the law of total expectation, we have

$$\begin{aligned}
\mathbb{P}_{(x,\widetilde{y}) \sim P_{\text{clean}}}(\widetilde{y} \neq \operatorname{sgn}(f(x;W))) &= \frac{1}{4} \sum_{\nu \in \mathsf{centers}} \mathbb{P}_{x \sim N(\nu, I_p)}(y_\nu \neq \operatorname{sgn}(f(x;W))) \\
&\le \frac{1}{2} \sum_{\nu \in \mathsf{centers}} \exp\left(-c \left(\frac{\mathbb{E}_{x \sim N(\nu, I_p)}[f(x;W)]}{\|W\|_F}\right)^2\right) \\
&\le \max_{\nu \in \mathsf{centers}} 2 \exp\left(-c \left(\frac{\mathbb{E}_{x \sim N(\nu, I_p)}[f(x;W)]}{\|W\|_F}\right)^2\right).
\end{aligned}$$

Since for each ν, $N(\nu, I_p)$ is 1-strongly log-concave, we plug in $\lambda = 1$ in the proof of Lemma 4.1 in Frei et al. (2022b). Then (A.65) is obtained.

□

Our next theorem shows that the generalization risk is small for large t. Recall the definition of $\mathcal{J}_1$ and $\mathcal{J}_2$, we equivalently write them as

$$\mathcal{J}_1 = \mathcal{J}^{20\varepsilon}_{+\mu_1,\mathtt{P}} = \{j \in [m] : a_j > 0, D^{(0)}_{+\mu_1,j} > n^{1/2-20\varepsilon}, d^{(0)}_{+\mu_1,j} < \min\{c_{+\mu_1}, c_{-\mu_1}\} - 2n_{\pm\mu_1} - \sqrt{n}\};$$
$$\mathcal{J}_2 = \mathcal{J}^{20\varepsilon}_{+\mu_1,\mathtt{N}} \cup \mathcal{J}^{20\varepsilon}_{-\mu_1,\mathtt{N}} = \{j \in [m] : a_j < 0, D^{(0)}_{\nu,j} > n^{1/2-20\varepsilon},$$
$$d^{(0)}_{\nu,j} < \min\{c_\nu, c_{-\nu}\} - 2n_{\pm\mu_1} - \sqrt{n}, \nu \in \{\pm\mu_1\}\}.$$

Here $\mathcal{J}^{20\varepsilon}_{+\mu_1,\mathtt{P}}$, $\mathcal{J}^{20\varepsilon}_{+\mu_1,\mathtt{N}}$, and $\mathcal{J}^{20\varepsilon}_{-\mu_1,\mathtt{N}}$ are defined in (A.16). By Lemma A.4, we know that under a good run,

$$|\mathcal{J}_1| \geq \frac{m}{n^{10\varepsilon}}, \quad |\mathcal{J}_2| \geq (1 - \frac{10}{n^{20\varepsilon}})|\mathcal{J}_{\mathtt{N}}|. \tag{A.66}$$

Theorem A.8. *Suppose that Assumptions (A1)-(A6) hold. Under a good run, for $Cn^{10\varepsilon} \leq t \leq \sqrt{n}$, the generalization error of classifier $\mathrm{sgn}(f(x, W^{(t)}))$ has an upper bound*

$$\mathbb{P}_{(x,y)\sim P_{clean}}(y \neq \mathrm{sgn}(f(x; W^{(t)}))) \leq \exp\big(-\Omega(\frac{n^{1-20\varepsilon}\|\mu\|^4}{p})\big).$$

Proof. Without loss of generality, we consider x follows $N(+\mu_1, I_p)$. Then we have

$$\begin{aligned}\mathbb{E}_x[yf(x, W^{(t)})] &= \sum_{j=1}^m a_j \mathbb{E}_x[\phi(\langle w^{(t)}_j, x\rangle)] \\ &\geq \frac{1}{\sqrt{m}}\Big[\sum_{j:a_j>0}\phi\big(\langle w^{(t)}_j, \mathbb{E}[x]\rangle\big) - \sum_{j:a_j<0}\mathbb{E}_x[\phi(\langle w^{(t)}_j, x\rangle)\Big] \\ &\geq \frac{1}{\sqrt{m}}\sum_{j:j\in\mathcal{J}_1}\phi\big(\langle w^{(t)}_j, \mu_1\rangle\big) - \frac{1}{\sqrt{m}}\sum_{j:a_j<0}\mathbb{E}_x[\phi(\langle w^{(t)}_j, x\rangle)],\end{aligned} \tag{A.67}$$

where the first inequality uses Jensen's inequality. By Lemma A.14, we have that for $j \in \mathcal{J}_1$,

$$\begin{aligned}\langle w^{(t)}_j, \mu_1\rangle &= \sum_{s=0}^{t-1}\langle w^{(s+1)}_j - w^{(s)}_j, \mu_1\rangle + \langle w^{(0)}_j, \mu_1\rangle \\ &\geq \frac{\alpha}{4n\sqrt{m}}\sum_{s=0}^{t-1} D^{(s)}_{+\mu_1,j}\|\mu\|^2 - \omega_{\text{init}}\sqrt{3mp/2}\|\mu\| \\ &\geq \frac{\alpha\|\mu\|^2}{4n\sqrt{m}}\big[n^{1/2-20\varepsilon} + (c_{+\mu_1} - n_{+\mu_1} - d^{(0)}_{-\mu_1,j})(t-1)\big] - \omega_{\text{init}}\sqrt{3mp/2}\|\mu\| \\ &\geq \frac{\alpha\|\mu\|^2}{4n\sqrt{m}}(c_{+\mu_1} - n_{+\mu_1} - d^{(0)}_{-\mu_1,j})(t-1),\end{aligned}$$

where the first inequality is from Lemma A.14 and Lemma A.3; the second inequality uses the property that for $j \in \mathcal{J}_1$, $D^{(s)}_{+\mu_1,j} \geq c_{+\mu_1} - n_{+\mu_1} - d^{(0)}_{-\mu_1}(j), s \geq 1$, which is also from Lemma A.14; and the third inequality uses Assumption (A5). It yields that

$$\sum_{j:j\in\mathcal{J}_1}\phi\big(\langle w^{(t)}_j, \mu_1\rangle\big) \geq \frac{\alpha\|\mu\|^2(t-1)}{4n\sqrt{m}}\sum_{j\in\mathcal{J}_1}\big(c_{+\mu_1} - d^{(0)}_{-\mu_1}(j) - n_{+\mu_1}\big) \geq \frac{\alpha\|\mu\|^2(t-1)}{40\sqrt{m}}|\mathcal{J}_1|, \tag{A.68}$$

where the last inequality uses (B4) in Lemma A.4. For the second term in (A.67), note that we have $\phi(\lambda x) = \lambda\phi(x), \forall\lambda > 0$, and by Jensen's inequality, $\phi(x_1 + x_2) \leq \phi(x_1) + \phi(x_2), \forall x_1, x_2 \in \mathbb{R}$. Then we have

$$\mathbb{E}_x[\phi(\langle w, x\rangle)] \leq \phi(\langle w, \mu_1\rangle) + \mathbb{E}_x[\phi(\langle w, x - \mu_1\rangle)] = \phi(\langle w, \mu_1\rangle) + \sqrt{\frac{1}{2\pi}}\|w\|, \tag{A.69}$$

where the last equation uses the expectation of half-normal distribution. By Lemma A.11, we have $g_i^{(t)} \le 1$, and

$$\|w_j^{(t+1)} - w_j^{(t)}\| = \Big\|\frac{\alpha a_j}{n}\sum_{i=1}^n g_i^{(t)}\phi'(\langle w_j^{(\tau)}, x_i\rangle)y_i x_i\Big\|$$
$$\le \frac{\alpha}{n\sqrt{m}}\max_{i\in[n]} g_i^{(t)}\sqrt{\sum_{i=1}^n \|x_i\|^2 + \sum_{i\neq j}|\langle x_i, x_j\rangle|} \le \frac{2\alpha\sqrt{p}}{\sqrt{mn}}, \quad 0 \le t \le 1/(\sqrt{n}p\alpha) - 2,$$

where the last inequality uses $\|x_i\|^2 \le 2p$, $|\langle x_i, x_j\rangle| \le 2\mu^2$, which comes from Lemma A.1, and Assumption (A2). It yields that for each $j \in [m]$,

$$\|w_j^{(t)}\| \le \sum_{\tau=0}^{t-1}\|w_j^{(\tau+1)} - w_j^{(\tau)}\| + \|w_j^{(0)}\| \le \frac{2\alpha\sqrt{p}t}{\sqrt{nm}} + \|w_j^{(0)}\| \le \frac{3\alpha\sqrt{p}t}{\sqrt{mn}}, \tag{A.70}$$

where the last inequality uses Lemma A.3. Then we consider the decomposition of $\sum_{j:a_j<0}\phi(\langle w_j^{(t)}, \mu_1\rangle)$:

$$\sum_{j:a_j<0}\phi(\langle w_j^{(t)}, \mu_1\rangle) = \sum_{j\in\mathcal{J}_2}\phi(\langle w_j^{(t)}, \mu_1\rangle) + \sum_{j\in\mathcal{J}_{\mathtt{N}}, j\notin\mathcal{J}_2}\phi(\langle w_j^{(t)}, \mu_1\rangle).$$

For the first term, we have

$$\begin{aligned}
&\sum_{j\in\mathcal{J}_2}\phi(\langle w_j^{(t)}, \mu_1\rangle) \\
&\le \sum_{j\in\mathcal{J}_2}\Big[\sum_{s=0}^{t-1}\phi(\langle w_j^{(s+1)} - w_j^{(s)}, \mu_1\rangle) + \phi(\langle w_j^{(0)}, \mu_1\rangle)\Big] \\
&\le \sum_{j\in\mathcal{J}_2}\Big[\sum_{s=0}^{t-1}\big(\frac{\alpha\|\mu\|^2}{2n\sqrt{m}}D_{+\mu_1,j}^{(s)} + \frac{5\alpha\|\mu\|^2}{n\sqrt{mn}}\big) + \omega_{\text{init}}\sqrt{3mp/2}\|\mu\|\Big] \\
&\le \sum_{j\in\mathcal{J}_2}\Big[\frac{\alpha\|\mu\|^2}{2n\sqrt{m}}(n + \Delta_{\mu_1}(t-2)) + \frac{5\alpha\|\mu\|^2 t}{n\sqrt{mn}} + \omega_{\text{init}}\sqrt{3mp/2}\|\mu\|\Big] \\
&\le \sum_{j\in\mathcal{J}_2}\frac{\alpha\|\mu\|^2}{2n\sqrt{m}}[n + 1 + (\Delta_{\mu_1}+1)(t-2)] \le \frac{\alpha\|\mu\|^2}{2n\sqrt{m}}[n + 1 + (\Delta_{\mu_1}+1)(t-2)]|\mathcal{J}_{\mathtt{N}}|,
\end{aligned} \tag{A.71}$$

where the second inequality uses (A.32) in Lemma A.12; the third inequality uses Lemma A.15; and the fourth inequality uses Assumptions (A1) and (A5). For the second term, we have

$$\begin{aligned}
&\sum_{j\in\mathcal{J}_{\mathtt{N}}, j\notin\mathcal{J}_2}\phi(\langle w_j^{(t)}, \mu_1\rangle) \\
&\le \sum_{j\in\mathcal{J}_{\mathtt{N}}, j\notin\mathcal{J}_2}\Big[\sum_{s=0}^{t-1}\phi(\langle w_j^{(s+1)} - w_j^{(s)}, \mu_1\rangle) + \phi(\langle w_j^{(0)}, \mu_1\rangle)\Big] \\
&\le \sum_{j\in\mathcal{J}_{\mathtt{N}}, j\notin\mathcal{J}_2}\Big[\sum_{s=0}^{t-1}\big(\frac{\alpha\|\mu\|^2}{2n\sqrt{m}}D_{+\mu_1,j}^{(s)} + \frac{5\alpha\|\mu\|^2}{n\sqrt{mn}}\big) + \omega_{\text{init}}\sqrt{3mp/2}\|\mu\|\Big] \\
&\le \sum_{j\in\mathcal{J}_{\mathtt{N}}, j\notin\mathcal{J}_2}\frac{\alpha t(c_{+\mu_1} + n_{-\mu_1} + 1)\|\mu\|^2}{n\sqrt{m}} \\
&= \frac{\alpha t(c_{-\mu_1} + n_{+\mu_1} + 1)\|\mu\|^2}{n\sqrt{m}}(|\mathcal{J}_{\mathtt{N}}| - |\mathcal{J}_2 \cup \mathcal{J}_3|) \\
&\le \frac{10\alpha t\|\mu\|^2}{n^{20\varepsilon}\sqrt{m}}|\mathcal{J}_{\mathtt{N}}|,
\end{aligned} \tag{A.72}$$

where the second inequality uses (A.32) in Lemma A.12; the third inequality uses $D_{\nu,j}^{(t)} \le c_\nu + n_{-\nu}$ and Assumption (A5); and the last inequality uses (A.66) and $c_{-\mu_1} + n_{+\mu_1} + 1 \le n$. Combining (A.69), (A.70), (A.71), and (A.72), we have

$$\begin{aligned}
\sum_{j:a_j<0} \mathbb{E}_x[\phi(\langle w_j^{(t)}, x\rangle)] &\le \sum_{j:a_j<0} \phi(\langle w_j^{(t)}, \mu_1\rangle) + \sqrt{\frac{1}{2\pi}} \sum_{j:a_j<0} \|w_j^{(t)}\| \\
&= \sum_{j\in\mathcal{J}_2} \phi(\langle w_j^{(t)}, \mu_1\rangle) + \sum_{j\in\mathcal{J}_{\mathtt{N}}, j\notin\mathcal{J}_2} \phi(\langle w_j^{(t)}, \mu_1\rangle) + \sqrt{\frac{1}{2\pi}} \sum_{j:a_j<0} \|w_j^{(t)}\| \\
&\le \frac{\alpha\|\mu\|^2 t\sqrt{m}}{2n}\big[\frac{n+1}{t} + (\Delta_{\mu_1}+1) + \frac{20n}{n^{20\varepsilon}} + \frac{3\sqrt{2np}}{\sqrt{\pi}\|\mu\|^2}\big].
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathbb{E}_{x\sim N(+\mu_1, I_p)}[yf(x, W^{(t)})] \\
&\ge \frac{\alpha\|\mu\|^2(t-1)}{40m}|\mathcal{J}_1| - \frac{\alpha\|\mu\|^2 t}{2n}\big[\frac{n+1}{t} + (\Delta_{\mu_1}+1) + \frac{20n}{n^{20\varepsilon}} + \frac{3\sqrt{2np}}{\sqrt{\pi}\|\mu\|^2}\big] \\
&\ge \frac{\alpha\|\mu\|^2 t}{2}\Big[\frac{1}{20n^{10\varepsilon}}(1-\frac{1}{t}) - \frac{2}{t} - \frac{\Delta_{\mu_1}+1}{n} - \frac{20}{n^{20\varepsilon}} - \frac{6\sqrt{p}}{\sqrt{2\pi n}\|\mu\|^2}\Big] \qquad \text{(A.73)} \\
&\ge \frac{\alpha\|\mu\|^2 t}{2}\Big[\frac{1}{20n^{10\varepsilon}}(1-\frac{1}{t}) - \frac{2}{t} - \frac{2\eta\sqrt{n\varepsilon\log(n)}+1}{n} - \frac{20}{n^{20\varepsilon}} - \frac{6}{\sqrt{2\pi}Cn}\Big] \ge \frac{\alpha\|\mu\|^2 t}{80n^{10\varepsilon}}
\end{aligned}$$

for $t \ge Cn^{10\varepsilon}$ when C is large enough. Here the second inequality uses $|\mathcal{J}_1| \ge mn^{-10\varepsilon}$; the third inequality uses (E3) in Lemma A.1 and Assumption (A1); and the last inequality uses $\varepsilon < 0.01$. By (A.70), it follows that $\|W^{(t)}\|_F \le 3\alpha t\sqrt{p/n}$. Thus we have

$$\frac{\mathbb{E}_{x\sim N(+\mu_1, I_p)}[yf(x, W^{(t)})]}{\|W^{(t)}\|_F} \ge \frac{\sqrt{n}\|\mu\|^2}{240\sqrt{p}n^{10\varepsilon}}.$$

This lower bound for the normalized margin can be easily extended to the other ν's. Applying Lemma A.17, we have

$$\mathbb{P}_{(x,y)\sim P_{\text{clean}}}(y \ne \text{sgn}(f(x; W^{(t)}))) \le 2\exp\left(-\frac{cn^{1-20\varepsilon}\|\mu\|^4}{240^2 p}\right) = \exp\big(-\Omega(\frac{n^{1-20\varepsilon}\|\mu\|^4}{p})\big).$$

□

Lemma A.7. *Suppose that Assumptions (A1)-(A6) hold. Under a good run, we have that for* $1 \le t \le \sqrt{n}$,

$$\text{cossim}(\sum_{j\in\mathcal{J}_1} w_j^{(t)}, +\mu_1) = \Omega(\frac{\sqrt{n}\|\mu\|}{\sqrt{p}});$$

$$\text{cossim}(\sum_{j\in\mathcal{J}_2} w_j^{(t)}, +\mu_1) = O(\frac{\sqrt{n}\|\mu\|}{\sqrt{p}}(\frac{1}{t} + \sqrt{\frac{\log n}{n}})).$$

Proof. This lemma is essentially implied by the proof of Lemma A.8. By (A.70), we have

$$\|\sum_{j:j\in\mathcal{J}_1} w_j^{(t)}\| \le \sum_{j:j\in\mathcal{J}_1} \|w_j^{(t)}\| \le |\mathcal{J}_1|\frac{3\alpha\sqrt{p}t}{\sqrt{mn}}.$$

By (A.68), we have

$$\langle \sum_{j:j\in\mathcal{J}_1} w_j^{(t)}, +\mu_1\rangle \ge \frac{\alpha\|\mu\|^2(t-1)}{40\sqrt{m}}|\mathcal{J}_1|.$$

Combining the inequalities above, we obtain

$$\text{cossim}(\sum_{j:j\in\mathcal{J}_1} w_j^{(t)}, +\mu_1) \geq \frac{\sqrt{n}\|\mu\|(t-1)}{120\sqrt{p}t} = \Omega(\frac{\sqrt{n}\|\mu\|}{\sqrt{p}}).$$

Again by (A.70), we have

$$\|\sum_{j:j\in\mathcal{J}_2} w_j^{(t)}\| \leq \sum_{j:j\in\mathcal{J}_2} \|w_j^{(t)}\| \leq |\mathcal{J}_2|\frac{3\alpha\sqrt{p}t}{\sqrt{mn}}.$$

By (A.71), we have

$$\langle\sum_{j\in\mathcal{J}_2} w_j^{(t)}, \mu_1\rangle \leq \frac{\alpha\|\mu\|^2}{2n\sqrt{m}}[n+1+(\Delta_{\mu_1}+1)(t-2)]|\mathcal{J}_2|$$

Combining the inequalities above, we obtain

$$\text{cossim}(\sum_{j\in\mathcal{J}_2} w_j^{(t)}, +\mu_1) \leq \frac{\|\mu\|}{6\sqrt{np}}[\frac{n}{t}+(\Delta_{\mu_1}+1)] = O(\frac{\sqrt{n}\|\mu\|}{\sqrt{p}}(\frac{1}{t}+\frac{\sqrt{\log(n)}}{\sqrt{n}})),$$

where the last inequality uses $\Delta_{\mu_1} = o(\sqrt{n\log(n)})$, which comes from Lemma A.1. □

A.5.3 Proof of Theorem A.21: 1-step Test Accuracy

Before stating the proof, we begin with the necessary definitions and a preliminary result. Recall that $h_i^{(t)} = g_i^{(t)} - 1/2$ and the decomposition (A.30). When $t = 0$, we denote

$$w_{j,\mathtt{T}}^{(1)} := w_j^{(0)} + \frac{\alpha a_j}{2n}\sum_{i=1}^{n}\phi'(\langle w_j^{(0)}, x_i\rangle)y_i x_i, \quad j\in[m] \tag{A.74}$$

and $W_{\mathtt{T}}^{(1)} := [w_{1,\mathtt{T}}^{(1)}, \cdots, w_{m,\mathtt{T}}^{(1)}]^\top$. Next lemma shows that $W_{\mathtt{T}}^{(1)}$ is a good approximation of $W^{(1)}$ with a large probability.

Lemma A.18. *Suppose Assumptions (A1) and (A2) hold. Given $\{x_i\}\in\mathcal{G}_{data}$ and $W^{(0)}\in\mathcal{G}_W$, we have*

$$|h_i^{(0)}| \leq p\omega_{init}\sqrt{3m}/2;$$
$$\|W_{\mathtt{T}}^{(1)} - W^{(1)}\|_F = \sqrt{\sum_{j=1}^{m}\|w_{j,\mathtt{T}}^{(1)} - w_j^{(1)}\|^2} \leq \frac{\alpha\omega_{init}p^{3/2}\sqrt{3m}}{\sqrt{n}}.$$

Proof. Let $z_i^{(t)} = y_i f(x_i; W^{(t)})$. Note that $\ell'(z) = -1/(1+\exp(z))$, we have $|-\ell'(z) - 1/2| \leq |z|/2$. It yields that

$$\begin{aligned}|h_i^{(0)}| &\leq \frac{1}{2}|z_i^{(0)}| \leq \frac{1}{2}\sum_{j=1}|a_j\langle w_j^{(0)}, x_i\rangle| \leq \frac{1}{2}\sqrt{\sum_{j=1}^{m} a_j^2\sum_{j=1}^{m}\|w_j^{(0)}\|^2\cdot\|x\|^2} \\ &= \frac{1}{2}\|W^{(0)}\|_F\cdot\|x_i\| \leq \frac{1}{2}p\omega_{\text{init}}\sqrt{3m},\end{aligned} \tag{A.75}$$

where the first inequality uses $h_i^{(t)} = g_i^{(t)} - 1/2$ and $g_i^{(t)} := -\ell'(z_i^{(t)})$; the second inequality uses triangle inequality; the third inequality uses Cauchy-Schwarz inequality; and the last inequality uses

(E1) in Lemma A.1 and (D1) in Lemma A.3. Denote $h_{\max} = \max_{i\in[n]} |h_i^{(0)}|$. Then we have

$$\begin{aligned}\|w_{j,\mathtt{T}}^{(1)} - w_j^{(1)}\| &= \frac{\alpha}{n\sqrt{m}} \|\sum_{i=1}^n h_i^{(0)} \phi'(\langle w_j^{(0)}, x_i\rangle) y_i x_i\| \\ &\le \frac{\alpha h_{\max}}{n\sqrt{m}} \sqrt{\sum_{i=1}^n \|x_i\|^2 + n(n-1)\max_{i\neq j} |x_i^\top x_j|} \\ &\le \frac{\alpha h_{\max}}{n\sqrt{m}} \sqrt{4np} \le \frac{\sqrt{3}\alpha\omega_{\text{init}} p^{3/2}}{\sqrt{n}},\end{aligned}$$

where the second inequality uses $\|x_i\|^2 \le 2p$ and $p \ge Cn^2\|\mu\|^2$, which come from (E1) and (E2) in Lemma A.1 and Assumption (A2) respectively, and the third inequality uses (A.75). Further we have

$$\|W_{\mathtt{T}}^{(1)} - W^{(1)}\|_F = \sqrt{\sum_{j=1}^m \|w_{j,\mathtt{T}}^{(1)} - w_j^{(1)}\|^2} \le \frac{\alpha\omega_{\text{init}} p^{3/2}\sqrt{3m}}{\sqrt{n}}.$$

□

Lemma A.19. *Suppose that Assumptions (A1)-(A6) hold. Given $X \in \mathcal{G}_{data}$, for each $j \in [m]$, we have*

$$\begin{aligned}n/24 \le \mathit{Var}(D_{+\mu_1,j}^{(0)}) \le n/2;\\ \mathbb{E}\big[\big|D_{+\mu_1,j}^{(0)}) - \mathbb{E}[D_{+\mu_1,j}^{(0)})]\big|^3\big] \le n^{3/2}.\end{aligned}$$

Proof. Recall that $\mathcal{A}_1 = \mathcal{C}_{+\mu_1} \cup \mathcal{N}_{-\mu_1}$, $\mathcal{A}_2 = \mathcal{C}_{-\mu_1} \cup \mathcal{N}_{+\mu_1}$. According to equation (A.20), we have

$$D_{+\mu_1,j}^{(0)} = \sum_{i\in\mathcal{A}_1} \mathbb{I}(z_i > 0) - \sum_{i\in\mathcal{A}_2} \mathbb{I}(z_i > 0). \tag{A.76}$$

According to Lemma A.24, we have

$$\begin{aligned}\text{Var}(D_{+\mu_1,j}^{(0)}) &= \mathbb{E}_B[f_1(b_1,\cdots,b_n)] \ge \frac{1}{2}\mathbb{E}_{B'}[f_1(b_1',\cdots,b_n')] \\ &= \frac{1}{2}\text{Var}_{B'}(\sum_{i\in\mathcal{A}_1} b_i' - \sum_{i\in\mathcal{A}_2} b_i') = \frac{|\mathcal{A}_1| + |\mathcal{A}_2|}{8} \ge \frac{n}{24},\end{aligned}$$

where $f_1(b_1,\cdots,b_n) := (\sum_{i\in\mathcal{A}_1} b_i - \sum_{i\in\mathcal{A}_2} b_i - (|\mathcal{A}_1| - |\mathcal{A}_2|)/2)^2 \ge 0$, and b_i' are i.i.d Bernoulli random variables defined in Lemma A.24, and the last inequality is from (A.19). On the other side, similarly we have

$$\text{Var}(D_{+\mu_1,j}^{(0)}) \le 2\mathbb{E}_{B'}[f_1(b_1',\cdots,b_n')] = (|\mathcal{A}_1| + |\mathcal{A}_2|)/2 \le n/2, \tag{A.77}$$

where the last inequality is from (E3) in Lemma A.1. Denote $f_2(b_1,\cdots,b_n) := (\sum_{i\in\mathcal{A}_1} b_i - \sum_{i\in\mathcal{A}_2} b_i - (|\mathcal{A}_1| - |\mathcal{A}_2|)/2)^4 \ge 0$, then we have

$$\begin{aligned}\mathbb{E}[|D_{+\mu_1,j}^{(0)} - \mathbb{E}[D_{+\mu_1,j}^{(0)}]|^4] &= \mathbb{E}_B[f_2(b_1,\cdots,b_n)] \le 2\mathbb{E}_{B'}[f_2(b_1',\cdots,b_n')] \\ &= 2\mathbb{E}_{B'}\big[\big[\sum_{i\in\mathcal{A}_1}(b_i' - \frac{1}{2}) - \sum_{i\in\mathcal{A}_2}(b_i' - \frac{1}{2})\big]^4\big] \\ &\le 16\mathbb{E}_{B'}\big[\big[\sum_{i\in\mathcal{A}_1}(b_i' - \frac{1}{2})\big]^4 + \big[\sum_{i\in\mathcal{A}_2}(b_i' - \frac{1}{2})\big]^4\big] \\ &\le 4(|\mathcal{A}_1|^2 + |\mathcal{A}_2|^2) \le n^2,\end{aligned} \tag{A.78}$$

where the first inequality uses Lemma A.24; the second inequality uses $(a+b)^4 \le 8(a^4 + b^4)$; the third inequality uses the formula of the fourth central moment of a binomial distribution with parameter equal to $1/2$, i.e. $\mu_4(\mathrm{B}(n, 1/2)) = n(1 + (3n-6)/4)/4 \le n^2/4$; and the last inequality is from (E3) in Lemma A.1. Combining (A.77) and (A.78), we have

$$\mathbb{E}\big[\big|D^{(0)}_{+\mu_1,j}) - \mathbb{E}[D^{(0)}_{+\mu_1,j})]\big|^3\big] \le \sqrt{\mathrm{Var}(D^{(0)}_{+\mu_1,j})\mathbb{E}[|D^{(0)}_{+\mu_1,j} - \mathbb{E}[D^{(0)}_{+\mu_1,j}]|^4]} \le n^{3/2}$$

by applying the Cauchy-Schwarz inequality.

□

Lemma A.20. *Suppose that Assumptions (A1)-(A6) hold. Given $X = [x_1, \cdots, x_n]^\top \in \mathcal{G}_{data}$, we have*

$$\mathbb{P}\Big(\Big|\sum_{j=1}^m a_j\phi(a_j D^{(0)}_{+\mu_1,j}) - \frac{1}{2}\mathbb{E}[D^{(0)}_{+\mu_1,j}]\Big| > t\Big) \le 2\bar{\Phi}\Big(\frac{t\sqrt{m}}{3C_n\sqrt{n\varepsilon}}\Big) + \frac{C}{\sqrt{m}};$$

$$\mathbb{P}\Big(\Big|\sum_{j=1}^m a_j|a_j D^{(0)}_{+\mu_1,j}|\Big| > t\Big) \le 2\bar{\Phi}\Big(\frac{t\sqrt{m}}{3C_n\sqrt{n\varepsilon}}\Big) + \frac{C}{\sqrt{m}}.$$

Proof. In this proof, by convention all $\mathbb{P}(\cdot), \mathbb{E}[\cdot], \mathrm{Var}(\cdot), \rho(\cdot)$ are implicitly conditioned on a fixed X. Denote the expectation of $D^{(0)}_{+\mu_1,j}$ by $e_{+\mu_1}$. Note that conditioning on X, $\{a_j\phi(a_j D^{(0)}_{+\mu_1,j})\}_{j\ge 1}$ are i.i.d, and the expectation of $D^{(0)}_{+\mu_1,j}$ is

$$e_{+\mu_1} = (c_{+\mu_1} - n_{+\mu_1} - c_{-\mu_1} + n_{-\mu_1})/2 \le 2C_n\sqrt{n\varepsilon}, \tag{A.79}$$

where the inequality uses (E3) in Lemma A.1. By Lemma A.19, we have

$$\frac{n}{24} \le \mathrm{Var}\big(D^{(0)}_{+\mu_1,j}\big) \le \frac{n}{2}; \quad \rho(D^{(0)}_{+\mu_1,j}) \le n^{3/2}. \tag{A.80}$$

Denote

$$\sigma^2_{+\mu_1} = \mathrm{Var}\big(ma_j\phi(a_j D^{(0)}_{+\mu_1,j})\big); \quad \rho_{+\mu_1} = \rho(ma_j\phi(a_j D^{(0)}_{+\mu_1,j})).$$

Combining (A.80) and results in Lemma A.23, we have

$$\mathbb{E}[ma_j\phi(a_j D^{(0)}_{+\mu_1,j})] = \frac{e_{+\mu_1}}{2}; \quad \max\{\frac{n}{48}, \frac{e^2_{+\mu_1}}{4}\} \le \sigma^2_{+\mu_1} \le \max\{\frac{n}{2}, \frac{e^2_{+\mu_1}}{2}\}; \quad \rho_{+\mu_1} \le 32\max\{n^{3/2}, |e_{+\mu_1}|^3\}. \tag{A.81}$$

Applying Berry-Esseen theorem, we have

$$\mathbb{P}\Big(\Big|\sum_{j=1}^m a_j\phi(a_j D^{(0)}_{+\mu_1,j}) - \frac{1}{2}e_{+\mu_1}\Big| > t\Big) \le 2\bar{\Phi}\Big(\frac{t\sqrt{m}}{\sigma_{+\mu_1}}\Big) + \frac{C_{\mathrm{BE}}\rho_{+\mu_1}}{\sigma^3_{+\mu_1}\sqrt{m}} \le 2\bar{\Phi}\Big(\frac{t\sqrt{m}}{\sqrt{n} + 2C_n\sqrt{n\varepsilon}}\Big) + \frac{C}{\sqrt{m}}$$

for some universal constant $C > 0$. Here the second inequality uses $\sigma^2_{+\mu_1} \le (\sqrt{n} + |e_{+\mu_1}|)^2$, which comes from (A.81), and the last inequality uses (A.79). By the symmetry of a_j, we have

$$\mathbb{E}[ma_j|a_j D^{(0)}_{+\mu_1,j}|] = 0; \quad \mathrm{Var}(ma_j|a_j D^{(0)}_{+\mu_1,j}|) = \mathbb{E}[(D^{(0)}_{+\mu_1,j})^2]; \quad \rho(ma_j|a_j D^{(0)}_{+\mu_1,j}|) = \mathbb{E}[|D^{(0)}_{+\mu_1,j}|^3].$$

By (A.80), we have

$$\frac{n}{24} + e^2_{+\mu_1} \le \mathbb{E}[(D^{(0)}_{+\mu_1,j})^2] \le \frac{n}{2} + e^2_{+\mu_1}; \quad \mathbb{E}[|D^{(0)}_{+\mu_1,j}|^3] \le 8(\rho(D^{(0)}_{+\mu_1,j}) + |e_{+\mu_1}|^3) \le 8(n^{3/2} + |e_{+\mu_1}|^3). \tag{A.82}$$

Similarly, applying Berry-Esseen theorem, we have

$$\mathbb{P}\Big(\Big|\sum_{j=1}^m a_j|a_j D^{(0)}_{+\mu_1,j}|\Big| > t\Big) \le 2\bar{\Phi}\Big(\frac{t\sqrt{m}}{\sqrt{n} + 2C_n\sqrt{n\varepsilon}}\Big) + \frac{C}{\sqrt{m}},$$

where the inequality uses $\mathrm{Var}(ma_j|a_j D^{(0)}_{+\mu_1,j}|) \le (\sqrt{n} + |e_{+\mu_1}|)^2$ and (A.79). Then the results of this lemma are proved by noting that $C_n\sqrt{\varepsilon} \ge 1$ for large enough n. □

Theorem A.21. *Suppose that Assumptions (A1)-(A6) hold. With probability at least $1 - 3C/\sqrt{m} - 2n^{-\varepsilon}$ over the initialization of the weights and the generation of training data, after one iteration, the* *classifier* $\mathrm{sgn}(f(x, W^{(1)}))$ *exhibits a generalization risk with the following bounds:*

$$\tfrac{1}{2}(1 - n^{-\varepsilon}) \le \mathbb{P}_{(x,y)\sim P_{clean}}(y \neq \mathrm{sgn}(f(x; W^{(1)}))) \le \tfrac{1}{2}(1 + n^{-\varepsilon}).$$

Proof. For any given training data $X \in \mathcal{G}_{\text{data}}$, denote the expectation of $D^{(0)}_{\nu,j}$ by e_ν, i.e.

$$e_\nu := \mathbb{E}[D^{(0)}_{\nu,j}|X] = (c_\nu - n_\nu - c_{-\nu} + n_{-\nu})/2, \quad \nu \in \{\pm\mu_1, \pm\mu_2\}, \tag{A.83}$$

and a set of parameters $\mathcal{G}_X$:

$$\begin{aligned}\mathcal{G}_X := \big\{(a, W^{(0)}) :& |\sum_{j=1}^m a_j \phi(a_j D^{(0)}_{\nu,j}) - e_\nu/2| \le 3C_n\sqrt{n\varepsilon/m}\log(m),\\ & |\sum_{j=1}^m a_j |a_j D^{(0)}_{\nu,j}|| \le 3C_n\sqrt{n\varepsilon/m}\log(m), a \in \mathcal{G}_A, W^{(0)} \in \mathcal{G}_W\big\}.\end{aligned}$$

Applying the union bound, we have

$$\mathbb{P}(\mathcal{G}_X|X \in \mathcal{G}_{\text{data}}) \ge 1 - \exp(-\Omega(\log^2(m))) - \frac{2C}{\sqrt{m}} - n^{-\varepsilon}$$

by Lemma A.20 and A.3. Further we have

$$\begin{aligned}\mathbb{P}((a, W^{(0)}) \in \mathcal{G}_X, X \in \mathcal{G}_{\text{data}}) &\ge \mathbb{P}(\mathcal{G}_X|X \in \mathcal{G}_{\text{data}})\mathbb{P}(X \in \mathcal{G}_{\text{data}})\\ &\ge 1 - \exp(-\log^2(m)/2) - \frac{2C}{\sqrt{m}} - 2n^{-\varepsilon}\\ &\ge 1 - \frac{3C}{\sqrt{m}} - 2n^{-\varepsilon}.\end{aligned}$$

Define events $\mathcal{F}_{\text{test},\nu}$ for test data:

$$\begin{aligned}\mathcal{F}_{\text{test},\nu} = \{x \in \mathbb{R}^p :& |\|x\|^2 - p - \|\mu\|^2| \le C_n\sqrt{p};\\ & |\langle x, x_i\rangle - \langle \nu, \bar{x}_i\rangle| \le C_n\sqrt{p} \text{ for all } i \in [n]\}, \quad \nu \in \{\pm\mu_1, \pm\mu_2\}.\end{aligned}$$

Treat $\{x\} \cup \{x_i\}_{i=1}^n$ as a new 'training' set with $n+1$ datapoints. Following the proof procedure in Lemma A.1, we can show that $\mathbb{P}_{x\sim N(\nu,I_p)}(x \in \mathcal{F}_{\text{test}}|X \in \mathcal{G}_{\text{data}}) \ge 1 - n^{-\varepsilon}$, where $\mathcal{F}_{\text{test}} := \cup_{\nu\in\{\pm\mu_1,\pm\mu_2\}}\mathcal{F}_{\text{test},\nu}$. And $\mathcal{F}_{\text{test}}$ is a symmetric set for x, i.e., if $x \in \mathcal{F}$, then $-x$ also belongs to $\mathcal{F}_{\text{test}}$. In the remaining proof, by convention all probabilities and expectations are implicitly conditioned on fixed $X \in \mathcal{G}_{\text{data}}$ and $a, W^{(0)} \in \mathcal{G}_X$. Therefore, to simplify notation, we write $\mathbb{P}(\,\cdot\,)$ and $\mathbb{E}[\,\cdot\,]$ to denote $\mathbb{P}(\,\cdot\,|a, W^{(0)}, \{x_i\})$ and $\mathbb{E}[\,\cdot\,|a, W^{(0)}, \{x_i\}]$, respectively. In other words, the randomness is over the test data (x, y), conditioned on a fixed initialization and training data. We first look at the clusters centered at $\pm\mu_1$, i.e. $x \sim N(\pm\mu_1, I_p), y = 1$. Then we have

$$\begin{aligned}&\mathbb{P}_{x\sim N(\pm\mu_1,I_p)}(y \neq \mathrm{sgn}(f(x, W^{(1)}))) = \mathbb{P}_{x\sim N(\pm\mu_1,I_p)}(f(x, W^{(1)}) \le 0)\\ &\quad= \frac{1}{2}\mathbb{P}_{x\sim N(\mu_1,I_p)}(f(x, W^{(1)}) \le 0) + \frac{1}{2}\mathbb{P}_{x\sim N(\mu_1,I_p)}(f(-x, W^{(1)}) \le 0).\end{aligned} \tag{A.84}$$

Note that given $W^{(0)}$ and X, we have with probability 1 that

$$\begin{aligned}|f(x; W^{(1)}) - f(x; W^{(1)} - W^{(0)})| &= \Big|\sum_{j=1}^m a_j[\phi(\langle w^{(1)}_j, x\rangle) - \phi(\langle w^{(1)}_j - w^{(0)}_j, x\rangle)]\Big|\\ &\le \sum_{j=1}^m |a_j\langle w^{(0)}_j, x\rangle| \le \sqrt{\sum_{j=1}^m a_j^2 \sum_{j=1}^m \|w^{(0)}_j\|^2 \cdot \|x\|^2}\\ &= \|W^{(0)}\|_F \cdot \|x\| \le \omega_{\text{init}}\sqrt{3mp/2}\|x\|,\end{aligned} \tag{A.85}$$

where the first inequality comes from the 1-Lipschitz continuity of $\phi(\cdot)$; the second inequality uses Cauchy-Schwarz inequality; and the last inequality uses Lemma A.3. Next, recall that W_{T} is defined as in (A.74). By the same argument above, we have

$$\begin{aligned}
&|f(x;W^{(1)}-W^{(0)})-f(x;W^{(1)}_{\mathtt{T}}-W^{(0)})| \\
&=\Big|\sum_{j=1}^m a_j[\phi(\langle w_j^{(1)}-w_j^{(0)},x\rangle)-\phi(\langle w_{j,\mathtt{T}}^{(1)}-w_j^{(0)},x\rangle)]\Big| \\
&\le \sum_{j=1}^m |a_j\langle w_j^{(1)}-w_{j,\mathtt{T}}^{(1)},x\rangle| \le \sqrt{\sum_{j=1}^m a_j^2\sum_{j=1}^m \|w_j^{(1)}-w_{j,\mathtt{T}}^{(1)}\|^2\cdot\|x\|^2} = \|W^{(1)}-W^{(1)}_{\mathtt{T}}\|_F\cdot\|x\| \\
&\le \alpha\omega_{\text{init}}p\sqrt{3mp/n}\|x\| \le \omega_{\text{init}}\sqrt{3mp/n}\|x\|, \qquad \text{(A.86)}
\end{aligned}$$

where the first inequality comes from the 1-Lipschitz continuity of $\phi(\cdot)$; the second inequality uses Cauchy-Schwarz inequality; the third inequality uses Lemma A.18; and the last inequality uses Assumption (A3). Using (A.85) and (A.86), we have by the triangle inequality that

$$|f(x;W^{(1)})-f(x;W^{(1)}_{\mathtt{T}}-W^{(0)})| \le 2\omega_{\text{init}}\sqrt{mp}\|x\| =: \epsilon_x, \quad \text{that for any } x\in\mathbb{R}^p. \qquad \text{(A.87)}$$

Recall that

$$\langle w_{j,\mathtt{T}}^{(1)}-w_j^{(0)},x\rangle = \frac{\alpha a_j}{2n}\sum_{i=1}^n \phi'(\langle w_j^{(0)},x_i\rangle)\langle y_ix_i,x\rangle.$$

Then under a good run, for $x\in\mathcal{F}_{\text{test}}$, we have that with probability 1,

$$\Big|\langle w_{j,\mathtt{T}}^{(1)}-w_j^{(0)},x\rangle-\frac{\alpha a_j}{2n}D^{(0)}_{+\mu_1,j}\|\mu\|^2\Big| \le \frac{\alpha}{\sqrt{m}}C_n\sqrt{p},$$

where the inequality uses the definition of $\mathcal{F}_{\text{test}}$. It yields that

$$\Big|f(x;W^{(1)}_{\mathtt{T}}-W^{(0)})-\sum_{j=1}^m\frac{\alpha a_j}{2n}\phi(a_jD^{(0)}_{+\mu_1,j})\|\mu\|^2\Big| \le \alpha C_n\sqrt{p}. \qquad \text{(A.88)}$$

According to the definition of $\mathcal{G}_X$, we have

$$\Big|\sum_{j=1}^m\frac{\alpha a_j}{2n}\phi(a_jD^{(0)}_{+\mu_1,j})\|\mu\|^2-\frac{\alpha\|\mu\|^2}{4n}e_{+\mu_1}\Big| \le \frac{3\alpha C_n\sqrt{\varepsilon}\log(m)}{2\sqrt{mn}}\|\mu\|^2. \qquad \text{(A.89)}$$

Combining (A.87)-(A.89), we have

$$\Big|f(x;W^{(1)})-\frac{\alpha\|\mu\|^2}{4n}e_{+\mu_1}\Big| \le \epsilon_x+\alpha C_n\sqrt{p}+\frac{3\alpha C_n\sqrt{\varepsilon}\log(m)}{2\sqrt{mn}}\|\mu\|^2. \qquad \text{(A.90)}$$

The above inequality immediately implies that

$$\mathbb{P}(f(x;W^{(1)})\le 0|\mathcal{F}_{\text{test}}) \ge \mathbb{P}(\frac{\alpha\|\mu\|^2}{2n}e_{+\mu_1}\le -\epsilon_x-\alpha C_n\sqrt{p}-\frac{3\alpha C_n\sqrt{\varepsilon}\log(m)}{2\sqrt{mn}}\|\mu\|^2|\mathcal{F}_{\text{test}}). \qquad \text{(A.91)}$$

Similar to (A.90), for $-x\sim N(-\mu_1,I_p)$, we have

$$\Big|f(-x;W^{(1)})-\frac{\alpha\|\mu\|^2}{2n}e_{-\mu_1}\Big| \le \epsilon_x+\alpha C_n\sqrt{p}+\frac{3\alpha C_n\sqrt{\varepsilon}\log(m)}{2\sqrt{mn}}\|\mu\|^2.$$

Note that by definition, $e_{-\mu_1}=-e_{+\mu_1}$, the above inequality immediately implies that

$$\mathbb{P}(f(-x;W^{(1)})\le 0|\mathcal{F}_{\text{test}}) \ge \mathbb{P}(\frac{\alpha\|\mu\|^2}{2n}e_{+\mu_1}\ge \epsilon_x+\alpha C_n\sqrt{p}+\frac{3\alpha C_n\sqrt{\varepsilon}\log(m)}{2\sqrt{mn}}\|\mu\|^2|\mathcal{F}_{\text{test}}). \qquad \text{(A.92)}$$

According to the definition of $\mathcal{G}_{\text{test}}$, we have $\epsilon_x\le 4\omega_{\text{init}}\sqrt{m}p^{3/2}$. According to the definition of $\mathcal{G}_{\text{data}}$, we have

$$\begin{aligned}
|c_\nu-n_\nu-c_{-\nu}+n_{-\nu}| &\ge |c_\nu-c_{-\nu}|-|n_\nu-n_{-\nu}| \ge |c_\nu+n_\nu-c_{-\nu}-n_{-\nu}|-2|n_\nu-n_{-\nu}| \\
&\ge (1-2\eta)n^{1/2-\varepsilon}\ge n^{1/2-\varepsilon}/2.
\end{aligned}$$

Thus we have $|e_{+\mu_1}| \geq n^{1/2-\varepsilon}/4$. It yields that

$$
\begin{aligned}
&\frac{\alpha\|\mu\|^2}{2n}|e_{+\mu_1}| - \epsilon_x - \alpha C_n\sqrt{p} - \frac{3\alpha C_n\sqrt{\varepsilon}\log(m)}{2\sqrt{mn}}\|\mu\|^2 \\
\geq& \frac{\alpha\|\mu\|^2}{\sqrt{n}}\Big(\frac{1}{8n^{\varepsilon}} - 4\sqrt{mn}p^{3/2}\frac{\omega_{\text{init}}}{\alpha\|\mu\|^2} - C_n\sqrt{\frac{np}{\|\mu\|^4}} - \frac{3C_n\sqrt{\varepsilon}\log(m)}{2\sqrt{m}}\Big) \\
\geq& \frac{\alpha\|\mu\|^2}{\sqrt{n}}\Big(\frac{1}{8n^{\varepsilon}} - \frac{2}{m\sqrt{n}} - \frac{C_n}{3Cn^{0.01}} - \frac{3C_n}{2\sqrt{C}n^{0.01}}\Big) > 0,
\end{aligned} \tag{A.93}
$$

where the first inequality uses $|e_{+\mu_1}| \geq n^{1/2-\varepsilon}/4$ and $\epsilon_x \leq 4\omega_{\text{init}}\sqrt{m}p^{3/2}$; the second inequality uses Assumption (A5), (A1) and (A6); and the last inequality uses n is large enough. Combining (A.91)-(A.93), we have

$$
\begin{aligned}
&\mathbb{P}(f(x;W^{(1)}) \leq 0|\mathcal{F}_{\text{test}}) + \mathbb{P}(f(-x;W^{(1)}) \leq 0|\mathcal{F}_{\text{test}}) \\
\geq&\mathbb{P}(\frac{\alpha\|\mu\|^2}{2n}|e_{+\mu_1}| \geq \epsilon_x + \alpha C_n\sqrt{p} + \frac{3\alpha C_n\sqrt{\varepsilon}\log(m)}{2\sqrt{mn}}\|\mu\|^2|\mathcal{F}_{\text{test}}) = 1,
\end{aligned} \tag{A.94}
$$

where the inequality uses $\epsilon_x \geq 0$. Following a similar procedure, for the other side, we have

$$
\begin{aligned}
&\mathbb{P}(f(x;W^{(1)}) \leq 0|\mathcal{F}_{\text{test}}) + \mathbb{P}(f(-x;W^{(1)}) \leq 0|\mathcal{F}_{\text{test}}) \\
\leq&\mathbb{P}(\frac{\alpha\|\mu\|^2}{2n}|e_{+\mu_1}| \geq -\epsilon_x - \alpha C_n\sqrt{p} - \frac{3\alpha C_n\sqrt{\varepsilon}\log(m)}{2\sqrt{mn}}\|\mu\|^2|\mathcal{F}_{\text{test}}) = 1.
\end{aligned} \tag{A.95}
$$

Combining (A.94) and (A.95), we have

$$
\mathbb{P}(f(x;W^{(1)}) \leq 0|\mathcal{F}_{\text{test}}) + \mathbb{P}(f(-x;W^{(1)}) \leq 0|\mathcal{F}_{\text{test}}) = 1.
$$

Following the same procedure, we have that for any $\nu \in \{\pm\mu_1, \pm\mu_2\}$,

$$
\mathbb{P}_{x\sim N(\nu,I_p)}(yf(x;W^{(1)}) \leq 0|\mathcal{F}_{\text{test}}) + \mathbb{P}_{x\sim N(\nu,I_p)}(yf(-x;W^{(1)}) \leq 0|\mathcal{F}_{\text{test}}) = 1.
$$

Then for $(x,y) \sim P_{\text{clean}}$, we have

$$
\mathbb{P}_{(x,y)\sim P_{\text{clean}}}(yf(x;W^{(1)}) \leq 0) \geq \mathbb{P}(yf(x;W^{(1)}) \leq 0|\mathcal{F}_{\text{test}})\mathbb{P}(\mathcal{F}_{\text{test}}) \geq \frac{1}{2}(1-n^{-\varepsilon});
$$

$$
\mathbb{P}_{(x,y)\sim P_{\text{clean}}}(yf(x;W^{(1)}) \leq 0) \leq \mathbb{P}(yf(x;W^{(1)}) \leq 0|\mathcal{F}_{\text{test}})\mathbb{P}(\mathcal{F}_{\text{test}}) + \mathbb{P}(\mathcal{F}_{\text{test}}^c) \leq \frac{1}{2}(1+n^{-\varepsilon}).
$$

□

Lemma A.22. *Suppose that Assumptions (A1)-(A6) hold. With probability at least* $1 - 3C/\sqrt{m} - 2n^{-\varepsilon}$ *over the initialization of the weights and the generation of training data, we have*

$$
\mathbb{P}_{x\sim N(+\mu_1,I_p)}\Big(\Big|f(x;W^{(1)}) - \sum_{j=1}^{m}\frac{\alpha a_j}{2n}\phi(a_j D_{+\mu_1,j}^{(0)})\|\mu\|^2\Big| \leq 2\alpha C_n\sqrt{p}\Big) \geq 1 - O(n^{-\varepsilon}).
$$

Proof. We have

$$
|f(x;W^{(1)}) - \sum_{i=1}^{m}\frac{\alpha a_j}{2n}\phi(a_j D_{+\mu_1,j}^{(0)})\|\mu\|^2| \leq 4\omega_{\text{init}}p\sqrt{mp} + \alpha C_n\sqrt{p} \leq 2\alpha C_n\sqrt{p}.
$$

Here the first inequality uses (A.87), (A.88) and $\|x\| \leq \sqrt{2p}$, and the second inequality is from Assumption (A5).

□

A.6 Probability Lemmas

Lemma A.23. *Suppose we have a random variable g that has finite L_3 norm and a Rademacher variable a that is independent with g. Then we have*

$$\max\{\frac{1}{2}Var(g), \frac{1}{4}(\mathbb{E}[g])^2\} \le Var(a\phi(ag)) \le \max\{Var(g), \frac{1}{2}(\mathbb{E}[g])^2\}; \tag{A.96}$$

$$\mathbb{E}\big[\big|a\phi(ag) - \mathbb{E}[a\phi(ag))]\big|^3\big] \le 32\max\{\mathbb{E}[|g-\mathbb{E}[g]|^3], |\mathbb{E}[g]|^3\}. \tag{A.97}$$

Proof. The expectation of the random variable $a\phi(ag)$ is

$$\mathbb{E}[a\phi(ag)] = \frac{1}{2}\mathbb{E}[\phi(g) - \phi(-g)] = \frac{1}{2}\mathbb{E}[g], \tag{A.98}$$

where the first equation uses the law of expectation, and the second equation uses $\phi(x) - \phi(-x) = x$. The second moment of $a\phi(ag)$ is

$$\mathbb{E}[(a\phi(ag))^2] = \mathbb{E}[\phi(ag)^2] = \frac{1}{2}\mathbb{E}[\phi(g)^2 + \phi(-g)^2] = \frac{1}{2}\mathbb{E}[g^2], \tag{A.99}$$

where the last equation uses $\phi(x)^2 + \phi(-x)^2 = x^2$. Combining (A.98) and (A.99), we have

$$\mathrm{Var}(a\phi(ag)) = \frac{1}{2}\mathbb{E}[g^2] - \frac{1}{4}(\mathbb{E}[g])^2 = \frac{1}{2}\mathrm{Var}(g) + \frac{1}{4}(\mathbb{E}[g])^2,$$

which implies (A.96). Moreover, for a random variable X that has finite L_3 norm, we have

$$\|X - \mathbb{E}[X]\|_3 \le \|X\|_3 + \|\mathbb{E}[X]\|_3 \le \|X\|_3 + \mathbb{E}[|X|] \le 2\|X\|_3,$$

where the second inequality is due to $\|\mathbb{E}[X]\|_3 = |\mathbb{E}[X]|$ and the last inequality is due to $\|X\|_1 \le \|X\|_3$. Thus we have

$$\mathbb{E}\big[\big|a\phi(ag) - \frac{1}{2}\mathbb{E}[g]\big|^3\big] \le 8\mathbb{E}[|a\phi(ag)|^3] = 4\mathbb{E}[\phi(g)^3 + \phi(-g)^3] = 4\mathbb{E}[|g|^3],$$

where the last equation is due to $\phi(x)^3 + \phi(-x)^3 = |x|^3$. Then by $\|g\|_3 \le \|g - \mathbb{E}[g]\|_3 + |\mathbb{E}[g]|$, we have

$$\mathbb{E}\big[\big|a\phi(ag) - \frac{1}{2}\mathbb{E}[g]\big|^3\big] \le 4\big(\|g-\mathbb{E}[g]\|_3 + |\mathbb{E}[g]|\big)^3 \le 32\max\{\mathbb{E}[|g-\mathbb{E}[g]|^3], |\mathbb{E}[g]|^3\}.$$

□

Lemma A.24. *Suppose $Z = [z_1, \cdots, z_n]^\top \sim N(0, \Sigma)$, where $\Sigma_{ii} = 1$, and $|\Sigma_{ij}| \le 1/(Cn^2), 1 \le i \ne j \le n$. And $Z' = [z'_1, \cdots, z'_n]^\top \sim N(0, \mathbb{I}_n)$. Let $b_i = \mathbb{I}(z_i > 0)$ and $b'_i = \mathbb{I}(z'_i > 0), i \in [n]$ be Bernoulli random variables. Let $B = [b_1, \cdots, b_n]^\top$ and $B' = [b'_1, \cdots, b'_n]^\top$. Then we have that for any non-negative function $f : \mathbb{R}^n \to \mathbb{R}^+ \cup \{0\}$,*

$$\frac{1}{2}\mathbb{E}_{B'}[f(b'_1, \cdots, b'_n)] \le \mathbb{E}_B[f(b_1, \cdots, b_n)] \le 2\mathbb{E}_{B'}[f(b'_1, \cdots, b'_n)].$$

Proof. Note that for any fixed value $(b_1, \cdots, b_n) \in \{0,1\}^n$, $\mathbb{P}_{B'}(b'_1, \cdots, b'_n) = (1/2)^n$. Then we have

$$\begin{aligned}\mathbb{E}_B[f(b_1, \cdots, b_n)] &= \sum_{b_1,\cdots,b_n} f(b_1, \cdots, b_n)\mathbb{P}_B(b_1, \cdots, b_n) \\ &\ge (2\gamma_1)^n \sum_{b_1,\cdots,b_n} f(b_1, \cdots, b_n)\mathbb{P}_{B'}(b_1, \cdots, b_n) \\ &= (2\gamma_1)^n \mathbb{E}_{B'}[f(b_1, \cdots, b_n)],\end{aligned} \tag{A.100}$$

where the inequality comes from Lemma A.25. On the other side, similarly we have

$$\mathbb{E}_B[f(b_1, \cdots, b_n)] \le (2\gamma_2)^n \mathbb{E}_{B'}[f(b_1, \cdots, b_n)]. \tag{A.101}$$

By $C > 8$, we have $(2\gamma_1)^n = (1-4/(Cn))^n \geq 1-4/(Cn) \geq 1/2$ and $(2\gamma_2)^n = (1+4/(Cn))^n \leq \exp(4/C) \leq \exp(1/2) \leq 2$. Combining these results with (A.100) and (A.101), we have

$$\frac{1}{2}\mathbb{E}_{B'}[f(b_1', \cdots, b_n')] \leq \mathbb{E}_B[f(b_1, \cdots, b_n)] \leq 2\mathbb{E}_{B'}[f(b_1', \cdots, b_n')].$$

□

Lemma A.25. *Suppose $Z = [z_1, \cdots, z_n]^\top \sim N(0, \Sigma)$, where $\Sigma_{ii} = 1$, and $|\Sigma_{ij}| \leq 1/(Cn^2), 1 \leq i \neq j \leq n$. Then we have that for any subset $\mathcal{A} \subseteq [n]$,*

$$\gamma_1^n \leq \mathbb{E}[\prod_{i\in\mathcal{A}} \mathbb{I}(z_i > 0) \cdot \prod_{i\in[n]\setminus\mathcal{A}} \mathbb{I}(z_i < 0)] \leq \gamma_2^n$$

for $\gamma_1 = 1/2 - 2/(Cn)$ and $\gamma_2 = 1/2 + 2/(Cn)$.

Proof. We first prove the result for $\mathcal{A} = [n]$. Note that

$$\mathbb{P}(z_1 > 0, \cdots, z_n > 0) = \mathbb{P}(z_1 > 0)\prod_{k=2}^n \mathbb{P}(z_k > 0|z_{k-1} > 0, \cdots, z_1 > 0). \tag{A.102}$$

Let $Z_{k-1} = [z_1, \cdots, z_{k-1}]^\top$ and denote the covariance matrix of $[z_1, \cdots, z_k]$ as

$$\begin{bmatrix} \Sigma_{k-1} & \epsilon_k \\ \epsilon_k^\top & 1 \end{bmatrix},$$

where $\Sigma_{k-1} = \text{Cov}(Z_{k-1})$ and $\epsilon_k = \text{Cov}(Z_{k-1}, z_k)$. Then $|\epsilon_{kj}| \leq 1/(Cn^2)$ for $j \in [k-1]$, and the conditional distribution of $z_k|Z_{k-1}$ is $N(\epsilon_k^\top\Sigma_{k-1}^{-1}Z_{k-1}, 1 - \epsilon_k^\top\Sigma_{k-1}^{-1}\epsilon_k)$. By Gershgorin circle theorem, we have

$$1 - \frac{1}{Cn} \leq \lambda_{\min}(\Sigma_{k-1}) \leq \lambda_{\max}(\Sigma_{k-1}) \leq 1 + \frac{1}{Cn}.$$

Denote $f_{k-1}(\cdot)$ as the density function of Z_{k-1}. Then we have

$$\begin{aligned}
\mathbb{P}(z_k > 0|z_{k-1} > 0, \cdots, z_1 > 0) &= \int_0^\infty \cdots \int_0^\infty f_{k-1}(Z_{k-1})\bar{\Phi}\Big(\frac{-\epsilon_k^\top\Sigma_{k-1}^{-1}Z_{k-1}}{\sqrt{1-\epsilon_k^\top\Sigma_{k-1}^{-1}\epsilon_k}}\Big)dz_1\cdots dz_{k-1} \\
&\geq \int_{\|\Sigma_{k-1}^{-1/2}Z_{k-1}\|\leq 2\sqrt{n}} f_{k-1}(Z_{k-1})\bar{\Phi}\Big(\frac{-\epsilon_k\Sigma_{k-1}^{-1}Z_{k-1}}{\sqrt{1-\epsilon_k^\top\Sigma_{k-1}^{-1}\epsilon_k}}\Big)dz_1\cdots dz_{k-1} \\
&\geq \Big(\frac{1}{2} - \frac{\|\Sigma_{k-1}^{-1/2}\epsilon_k\|\cdot 2\sqrt{n}}{\sqrt{2\pi(1-\epsilon_k^\top\Sigma_{k-1}^{-1}\epsilon_k)}}\Big)\mathbb{P}(\|\Sigma_{k-1}^{-1/2}Z_{k-1}\| \leq 2\sqrt{n}) \\
&\geq (\frac{1}{2} - \frac{2\sqrt{2}}{nC\sqrt{\pi}})\mathbb{P}(\|\Sigma_{k-1}^{-1/2}Z_{k-1}\| \leq 2\sqrt{n}) \\
&\geq (\frac{1}{2} - \frac{2\sqrt{2}}{nC\sqrt{\pi}})(1-\exp(-n)) \geq \frac{1}{2} - \frac{2}{Cn}
\end{aligned} \tag{A.103}$$

for sufficiently large n. Here the second inequality uses $|\Phi(x) - \Phi(0)| \leq \Phi'(0)|x|$ and Cauchy-Schwarz inequality; the third inequality uses $\sigma_{\min}(\Sigma_{k-1}) = \lambda_{\min}(\Sigma_{k-1}) \geq 1/2$ and $\|\Sigma_{k-1}^{-1/2}\epsilon_k\| \leq \sqrt{2}\|\epsilon_k\| \leq \sqrt{2}n^{-3/2}/C$; and the fourth inequality uses the concentration inequality for chi-square random variables in Lemma A.26. Then the result is proved by combining (A.102) and (A.103). On the other side, we have

$$\mathbb{P}(z_k > 0|z_{k-1} > 0, \cdots, z_1 > 0) \leq \int_{\|\Sigma_{k-1}^{-1/2} Z_{k-1}\| \leq 2\sqrt{n}} f_{k-1}(Z_{k-1}) \bar{\Phi}\Big(\frac{-\epsilon_k \Sigma_{k-1}^{-1} Z_{k-1}}{\sqrt{1-\epsilon_k^\top \Sigma_{k-1}^{-1}\epsilon_k}}\Big) dz_1 \cdots dz_{k-1}$$

$$+ \mathbb{P}(\|\Sigma_{k-1}^{-1/2} Z_{k-1}\| > 2\sqrt{n})$$

$$\leq \Big(\frac{1}{2} + \frac{\|\Sigma_{k-1}^{-1/2}\epsilon_k\| \cdot 2\sqrt{n}}{\sqrt{2\pi(1-\epsilon_k^\top \Sigma_{k-1}^{-1}\epsilon_k)}}\Big) + \mathbb{P}(\|\Sigma_{k-1}^{-1/2} Z_{k-1}\| > 2\sqrt{n})$$

$$\leq \frac{1}{2} + \frac{2\sqrt{2}}{nC\sqrt{\pi}} + \exp(-n) \leq \frac{1}{2} + \frac{2}{Cn}.$$

Note that our proof does not use any information related to $\mathcal{A}$, thus we can extend the result for any subset $\mathcal{A} \subseteq [n]$. $\square$

Lemma A.26. *For X_k i.i.d $\sim N(0, \sigma^2), 1 \leq k \leq n$, we have*

$$\Phi'(t)/t \leq \mathbb{P}(|X_1| \geq t\sigma) \leq \exp(-t^2/2), \quad \forall t \geq 1;$$

$$\mathbb{P}(\Big|\frac{1}{n\sigma^2}\sum_{k=1}^n X_k^2 - 1\Big| \geq t) \leq 2\exp(-nt^2/8), \quad \forall t \in (0,1).$$

Proof. For the first inequality, we note that

$$\bar{\Phi}(t) = \int_t^{+\infty} \frac{x}{\sqrt{2\pi}x}\exp(-\frac{1}{2}x^2)dx \leq \int_t^{+\infty} \frac{1}{2\sqrt{2\pi}t}\exp(-\frac{1}{2}x^2)dx^2 = \frac{\Phi'(t)}{t}.$$

It yields that for any $t \geq 1$,

$$\mathbb{P}(|X_1| \geq t\sigma) = 2\bar{\Phi}(t) \leq 2\Phi'(t)/t \leq \exp(-t^2/2).$$

On the other side, we have

$$\bar{\Phi}(t) \geq \int_t^{+\infty} \frac{\frac{1+x^2}{x^2}}{\sqrt{2\pi}\frac{1+t^2}{t^2}}\exp(-\frac{1}{2}x^2)dx = \frac{1}{\sqrt{2\pi}}\frac{t^2}{1+t^2}\Big(-\frac{\exp(-\frac{x^2}{2})}{x}\Big)\Big|_{x=t}^{+\infty} = \frac{t}{1+t^2}\Phi'(t).$$

When $t \geq 1$, it further yields that $\bar{\Phi}(t) \geq \Phi'(t)/(2t)$. Thus we have

$$\mathbb{P}(|X_1| \geq t\sigma) = 2\bar{\Phi}(t) \geq \Phi'(t)/t.$$

The second inequality is Example 2.11 in Wainwright (2019) $\square$

Lemma A.27 (Hoeffding's inequality, Equation (2.11) in Wainwright (2019))**.** *Let $X_k, 1 \leq k \leq n$ be a series of independent random variables with $X_k \in [a, b]$. Then*

$$\mathbb{P}(\sum_{k=1}^n (X_k - \mathbb{E}[X_k]) \geq t) \leq \exp\Big(-\frac{2t^2}{n(b-a)^2}\Big), \quad \forall t \geq 0.$$

Lemma A.28. *[Berry-Esseen Theorem, Theorem 3.4.17 in Durrett (2019)] Let $X_1, \cdots, X_n$ are i.i.d. random variables with $\mathbb{E}[X_i] = 0, Var(X_i) = \sigma^2$, and $\mathbb{E}[|X_i|^3] = \rho < \infty$. If $F_n(x)$ is the distribution of $\sum_{i=1}^n X_i/(\sigma\sqrt{n})$, then*

$$|F_n(x) - \Phi(x)| \leq \frac{3\rho}{\sigma^3\sqrt{n}}.$$

A.7 Experimental details

In our experiments, dimension $p = 40000$, number of train/test samples $n = 200$ $\mu = 2.5\sqrt{p/n}$, number of neurons $m = 1000$, label noise rate $\eta = 0.05$, and initial weight scale $\omega_{\text{init}} = 10^{-15}$. For Figure 3, 2, and 1-left, the step size $\alpha = 10^{-12}$. For Figure 4 and 1-right, $\alpha = 10^{-16}$.

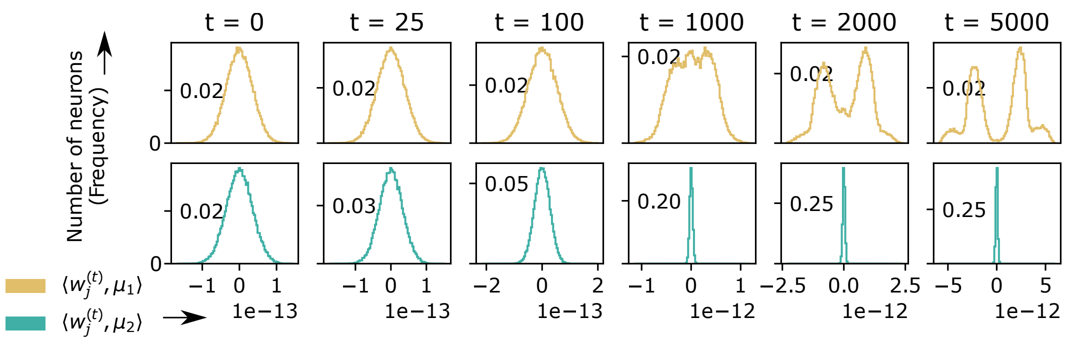

Figure 4: Histograms of inner products between positive neurons and μ's pooled over 100 independent runs under the same setting as in Figure 1 but with a smaller step size. *Top (resp. bottom) row:* Inner products between positive neurons and μ_1 (resp. μ_2). While the projections of positive neurons $w_j^{(t)}$ onto the μ_1 and μ_2 directions have nearly the same distribution when the network cannot generalize, they become much more aligned with $\pm\mu_1$ when the network can generalize.