Journal of Machine Learning Research 24 (2023) 1-49 Submitted 10/22; Revised 7/23; Published 8/23

Random Feature Amplification: Feature Learning and Generalization in Neural Networks

Spencer Frei FREI@BERKELEY.EDU
Simons Institute for the Theory of Computing, University of California, Berkeley, Calvin Lab #230, Berkeley, CA 94720

Niladri S. Chatterji NILADRI@CS.STANFORD.EDU
Computer Science Department, Stanford University, 353 Jane Stanford Way, Stanford, CA 94305

Peter L. Bartlett PETER@BERKELEY.EDU
University of California, Berkeley & Google DeepMind, 367 Evans Hall #3860 Berkeley, CA 94720

Editor: Joan Bruna

Abstract

In this work, we provide a characterization of the feature-learning process in two-layer ReLU networks trained by gradient descent on the logistic loss following random initialization. We consider data with binary labels that are generated by an XOR-like function of the input features. We permit a constant fraction of the training labels to be corrupted by an adversary. We show that, although linear classifiers are no better than random guessing for the distribution we consider, two-layer ReLU networks trained by gradient descent achieve generalization error close to the label noise rate. We develop a novel proof technique that shows that at initialization, the vast majority of neurons function as random features that are only weakly correlated with useful features, and the gradient descent dynamics 'amplify' these weak, random features to strong, useful features.

Keywords: feature learning, generalization, neural networks, classification, label noise

1. Introduction

A number of recent works have developed optimization and generalization guarantees for neural networks in the 'neural tangent kernel regime', namely, where the behavior of the neural network can be well-approximated by the linearization of the network around its random initialization (Jacot et al., 2018; Allen-Zhu et al., 2019; Zou et al., 2019; Du et al., 2019; Arora et al., 2019; Soltanolkotabi et al., 2019). Although these works provide a deep understanding of the behavior of neural networks in the early stages of training—where the network parameters are close to their initial values—they fail to capture a number of meaningful characteristics of practical neural networks such as the ability to learn features that differ significantly from those found at random initialization (Fort et al., 2020; Long, 2021). This points to the need for analyses of neural network training that can characterize how gradient descent is able to learn meaningful features.

A remarkable feature of neural networks is that despite their capacity to overfit, when trained by gradient descent they are capable of feature-learning even when there is significant label noise in the training data. Label noise is a common feature in modern machine learning datasets like ImageNet (Shankar et al., 2020), and moreover, some of the most interesting

©2023 Spencer Frei, Niladri S. Chatterji, Peter L. Bartlett.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v24/22-1132.html.

behaviors of neural networks have been observed when they are trained on datasets with artificially introduced random label noise (Zhang et al., 2017). This points to the importance of theoretically understanding the effect of noisy labels on the neural network training process. A handful of recent works have sought to understand the training dynamics of neural networks in the presence of noisy labels, but were either restricted to neural networks in the neural tangent kernel (NTK) regime, where feature learning is impossible (Hu et al., 2020; Ji et al., 2021); failed to provide generalization guarantees for the resulting network (Li et al., 2019); or only applied in settings where linear classifiers perform well (Frei et al., 2021).

In this work, we characterize the feature learning process of, and provide generalization guarantees for, two-layer ReLU networks trained by gradient descent on a data distribution where no linear classifier (that use input features) can perform better than random guessing. In particular, we consider two-layer ReLU networks where the first layer is trained while the second layer is fixed at its initial values, and we assume the data comes from a uniform mixture of four clusters of data, with means at $+\mu_1, -\mu_1, +\mu_2, -\mu_2$, where $\mu_1, \mu_2 \in \mathbb{R}^d$ are orthogonal. Clean labels are initially generated by an XOR function of the clusters: data from the $+\mu_1$ and $-\mu_1$ clusters have the clean label $+1$, and data from the $+\mu_2$ and $-\mu_2$ clusters have the label -1. We then allow for a constant fraction of these labels to be corrupted arbitrarily. Our results show that, provided gradient descent is initialized randomly with a sufficiently small initialization variance and provided the learning rate is sufficiently large, then with high probability gradient descent produces a network that correctly classifies every 'clean' test example and incorrectly classifies every 'noisy' test example. We point the reader to Figure 1 to see an example of the data distribution and the decision boundary learned in this setting. Our results hold for networks of essentially constant width and for arbitrarily small initialization variance. This is in contrast to the neural tangent kernel approaches where the initialization scale is relatively much larger that prevents features to change substantially during training.

Our proof follows by characterizing the types of features that individual neurons learn throughout the training process. We show that at random initialization, provided the width of the network is a sufficiently large constant, most neurons are 'weak' random features: they have a normalized correlation of order $O(1/\sqrt{d})$, where d is the input dimension, with at least one of the cluster means $\{\pm\mu_1, \pm\mu_2\}$. After initialization, provided the learning rate is sufficiently large, a single step of gradient descent amplifies these neurons from 'weak' random features to 'strong', learned features: the normalized correlations with the cluster means improve from order $O(1/\sqrt{d})$ to order $O(1)$. In the later part of the training process, we show that the gradient descent dynamics ensure that if a neuron is highly correlated with a given cluster center μ_s after the first step, then (1) its norm increases throughout training, so that the network relies more upon this neuron to determine the network output, and (2) the neuron becomes orthogonal to the opposing cluster center $\mu_{s'}$, $s' \neq s$, so that the neuron is useful only for samples from the cluster center μ_s. We show that having properties (1) and (2) is sufficient for producing a network that classifies all of the clean samples correctly and noisy samples incorrectly. A key difficulty in showing each of these facts is the presence of noisy training labels, which could in principle prevent the network from learning useful features; a careful analysis shows that this barrier is surmountable provided the fraction of noisy labels is smaller than an absolute constant.

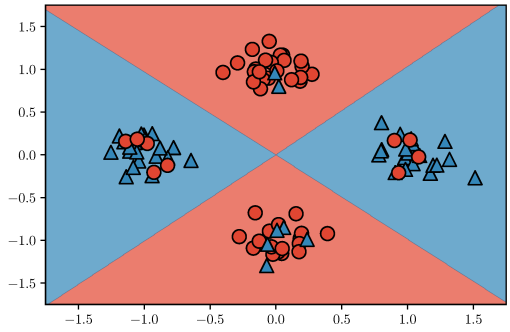

Figure 1: We consider a noisy 2-XOR cluster distribution where opposing cluster means share the same initial 'clean' label but a constant fraction of the labels are corrupted by an adversary. The figure is for the special case of Gaussian cluster distributions in $d = 2$ dimensions with in-cluster variance $\sigma^2 = 1/50$ when labels are flipped with probability 15%. We plot the decision boundary resulting from training a two-layer ReLU network given $n = 5000$ samples (we plot only a subset of the training samples to more clearly illustrate the labels of the samples). The network was trained for $T = 3000$ iterations, with network width $m = 500$, step-size $\alpha = 0.05$, and initialization variance $\omega_{\text{init}}^2 = 1/(32m)$.

1.1 Related work

As mentioned in the previous subsection, a number of works have highlighted the need to develop analyses of neural network training that go 'beyond' the NTK, or equivalently, neural networks that lie in the 'feature learning regime'. One collection of works has focused on developing separations between what hypothesis classes can be learned efficiently using neural networks in the feature learning regime versus what can be learned using approaches based on kernels or random features (Yehudai and Shamir, 2019; Allen-Zhu and Li, 2019; Ghorbani et al., 2019; Wei et al., 2019; Daniely and Malach, 2020; Allen-Zhu and Li, 2021; Malach et al., 2021; Abbe et al., 2021). One example of such a hypothesis class includes single neurons $x \mapsto \phi(\langle w, x \rangle)$, which can be efficiently learned using gradient descent on neural networks beyond the kernel regime (Frei et al., 2020; Yehudai and Shamir, 2020) but cannot be efficiently learned using random features or kernel-based methods (Yehudai and Shamir, 2019; Kamath et al., 2020). For a more detailed comparison of recent work on separations between what is learnable using kernel methods versus what is learnable using neural networks in the feature learning regime, we refer the reader to Table 2 and Appendix A of Malach et al. (2021). We note that two concurrent works have shown that a single step of gradient descent suffices for feature-learning behavior in neural networks (Ba et al., 2022; Damian et al., 2022). We also show that a single step of gradient descent suffices for learning data-dependent features, but our analysis also requires training for more than one step so that the learned features become more 'refined' (see Conditions 9 and 10 as well as Lemma 11 below).

Another line of work utilizes the mean field approximation to connect the training dynamics of infinitely wide neural networks to that of the solution to a partial differential

equation (Mei et al., 2018; Chizat and Bach, 2018; Wei et al., 2019; Chen et al., 2020; Fang et al., 2021). This approach allows for the network weights to traverse far from the initialization and learn features. These works provide a useful characterization of the limiting behavior of neural networks as they become infinitely wide. By contrast, in this work we provide a guarantee for neural network optimization and generalization for networks of constant width (for a constant level of failure probability).

A handful of other works have explored the behavior of neural networks trained by gradient descent for variants of the XOR distribution we consider in this work. Wei et al. (2019) used the mean field approximation to show that infinite-width two-layer networks trained by gradient flow will generalize well. Bai and Lee (2020) considered two-layer neural networks with smooth activations trained with additional 'random sign' and $\|W\|_{2,4}^8$ penalty regularization. They showed that when training with a large random initialization and a very large network, the second-order term of the Taylor expansion of the network around its initialization dominates the training dynamics and has a good optimization landscape provided the weights are close enough to initialization. They used this to derive a generalization guarantee for the resulting network. Although the work Bai and Lee (2020) is a strict improvement over standard NTK-based approaches, their analysis is more similar to the kernel-based analysis than the feature-learning approach we take here. Finally, Daniely and Malach (2020) provided a characterization of learning a noiseless parity over the binary cube when performing gradient descent on the population risk (i.e., assuming infinite samples). Their analysis relies upon a neuron-by-neuron characterization of the learning process, similar to ours, but it is unclear how their analysis would proceed without access to infinite samples or if there are noisy labels. Indeed, much of the difficulty in characterizing feature-learning in neural networks comes from the possibility that neural networks could simply memorize the sampled training data rather than learn useful representations that enable generalization to unseen test data. In contrast to all of the above works, our work provides a novel characterization of how feature-learning occurs in finite-width neural networks that are trained in the finite-sample setting and when a substantial portion of the training labels are adversarially corrupted.

Finally, since our analysis shows that early-stopped gradient descent with a small initialization variance produces neural networks with rather simple decision boundaries which essentially ignore the noisy labels (see Fig. 1), our work is related to a series of works on the simplicity bias of gradient descent (Phuong and Lampert, 2021; Lyu et al., 2021; Boursier et al., 2022; Frei et al., 2023). The aforementioned works all rely upon data that is either nearly-orthogonal or exactly orthogonal, while we make no such assumption. On the other hand, these other works characterize the behavior of gradient descent throughout the entire training trajectory, while we require early-stopping.

2. Preliminaries

We begin with describing our notational conventions. We denote $\|x\|$ as the Euclidean norm of a vector x. We will use uppercase letters to refer to matrices, with $\|W\|_F$ denoting the Frobenius norm of a matrix, and $\|W\|_2$ denoting the spectral norm. Given a matrix $W \in \mathbb{R}^{m\times d}$ we let $w_1, \ldots, w_m$ denote the rows of this matrix. Given any positive integer k, let $[k] = \{1, 2, \ldots, k\}$.

We next describe the distributional setting. We consider a joint distribution P over $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ constructed as follows.

1. First, define a cluster distribution $\mathsf{P}_{\text{clust}}$ over $\mathbb{R}^d$, which we assume to be log-concave[1] and satisfies $\mathbb{E}_{z\sim\mathsf{P}_{\text{clust}}}[z] = 0$ and $\mathbb{E}_{z\sim\mathsf{P}_{\text{clust}}}[zz^\top] = \sigma^2 I_d$ where $\sigma > 0$ is a fixed parameter.

2. Let $\mu_1, \mu_2 \in \mathbb{R}^d$ be unit norm orthogonal vectors, so $\langle \mu_1, \mu_2 \rangle = 0$ and $\|\mu_i\| = 1$ for $i = 1, 2$. The positive clusters are centered at μ_1 and $-\mu_1$, while the negative clusters are centered at μ_2 and $-\mu_2$.

3. The distribution of 'clean' samples $\widetilde{\mathsf{P}}$ is an XOR-like mixture distribution consisting of four independent cluster distributions $\{\mathsf{P}^{(i)}_{\text{clust}}\}_{i=1}^4$ centered at $\mu_1, -\mu_1, \mu_2, -\mu_2$ with labels $+1, +1, -1, -1$ respectively. That is, for example, for $(x, \tilde{y}) \sim \mathsf{P}^{(1)}_{\text{clust}}$, $x = \mu_1 + z$ where $z \sim \mathsf{P}_{\text{clust}}$ and $\tilde{y} = 1$. The distribution of clean samples is the uniform mixture $\widetilde{\mathsf{P}} := \frac{1}{4}\left[\mathsf{P}^{(1)}_{\text{clust}} + \mathsf{P}^{(2)}_{\text{clust}} + \mathsf{P}^{(3)}_{\text{clust}} + \mathsf{P}^{(4)}_{\text{clust}}\right]$.

4. Finally, the data distribution P is constructed by introducing label noise to $\widetilde{\mathsf{P}}$. The distribution P has the same marginal distribution over x as $\widetilde{\mathsf{P}}$, but for a given $(x, y) \sim \mathsf{P}$, the label y is equal to $\tilde{y}$ with probability $1 - \eta(x)$ and is equal to $-\tilde{y}$ with probability $\eta(x)$ for some $\eta(x) \in [0, 1]$. We call $\eta := \mathbb{E}_{x\sim\mathsf{P}}[\eta(x)]$ the *noise rate*.

We assume the training data S is generated as i.i.d. samples from P,

$$S := \{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathsf{P}^n.$$

The samples $\{(x_i, y_i)\}_{i=1}^n$ can be partitioned into *clean* and *noisy* samples, where we use the notation $\mathcal{C}, \mathcal{N} \subset [n]$ to denote the indices corresponding to the clean and noisy samples. In particular, using the notation $\tilde{y}_i$ to denote the clean label for the i-th sample, we have

$$y_i = \begin{cases} \tilde{y}_i, & i \in \mathcal{C}, \\ -\tilde{y}_i, & i \in \mathcal{N}. \end{cases}$$

We will consider the regime where the noise rate $\eta \approx |\mathcal{N}|/n$ is smaller than a constant. In Figure 1, we illustrate what samples from this distribution look like.

We analyze the classification error attained by neural networks trained by gradient descent with the logistic loss given the dataset S. In particular, we consider the class of one-hidden-layer ReLU networks consisting of m neurons with first layer weights $W \in \mathbb{R}^{m\times d}$,

$$x \mapsto f(x; W) := \sum_{j=1}^m a_j \phi(\langle w_j, x \rangle), \quad \text{where} \quad \phi(t) := \max\{0, t\}. \tag{1}$$

We will use the convention that ϕ is applied entry-wise, so that $\phi(Wx)$ has j-th component $\phi(\langle w_j, x\rangle)$. For simplicity, we assume that m is an even number and that half of the second layer weights a_j are initialized at the value of $+1/\sqrt{m}$, and the other half are initialized at

1. That is, $z \sim \mathsf{P}_{\text{clust}}$ has a probability density function p_z satisfying $p_z(x) = \exp(-U(x))$ for some convex function $U : \mathbb{R}^d \to \mathbb{R}$.

the value $-1/\sqrt{m}$. (Our results hold for odd m by setting $a_m = 0$.) We assume the second layer weights are fixed at their initialized values throughout training. This assumption allows for a more simplified analysis as it allows for a static partition of the neurons into 'positive' neurons (those for which $a_j > 0$) and 'negative' neurons ($a_j < 0$) throughout training. We believe it is possible to extend our analysis to the setting where both layers are trained but we do not pursue this question in this work.

Let $\ell(z) := \log(1 + \exp(-z))$ be the logistic loss. We consider the gradient descent algorithm on the empirical risk $\widehat{L}(W)$ corresponding to weights W the n samples $\{(x_i, y_i)\}_{i=1}^n$, where

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

The population risk under the logistic loss is defined as

$$L(W) := \mathbb{E}_{(x,y)\sim \mathsf{P}} \left[\ell\big(yf(x; W)\big)\right].$$

We consider ReLU networks trained by gradient descent on the first layer weights with fixed learning rate $\alpha > 0$ and with random initialization $[W^{(0)}]_{i,j} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \omega_{\text{init}}^2)$. In particular,

$$W^{(t+1)} = W^{(t)} - \alpha \nabla \widehat{L}(W^{(t)}) = W^{(t)} - \frac{\alpha}{n} \sum_{i=1}^{n} \ell'\big(y_i f(x_i; W^{(t)})\big) y_i \nabla f(x_i; W^{(t)}).$$

Note that since the ReLU activation $\phi(q) = \max(0, q)$ is not differentiable at 0, we use any subgradient value $\phi'(0) \in [0, 1]$ when performing gradient descent. (Our results do not depend on the value chosen for the subgradient.)

We let $C > 1$ denote a positive absolute constant that is large enough. Given a failure probability $\delta \in (0, 1/2)$ we make the following assumptions going forward:

(A1) The dimension $d \geq C \max\left\{\log^2(n/\delta), \log(m/\delta)\right\}$;

(A2) The in-cluster variance $\sigma^2 \leq 1/(C^2 d)$;

(A3) The sample size $n \geq C \log(m/\delta)$;

(A4) The noise rate $\eta \leq 1/C$;

(A5) The number of hidden nodes satisfies $m \geq C \log(1/\delta)$;

(A6) The variance at initialization satisfies $0 < \omega_{\text{init}}^2 \leq \frac{1}{C^4 md}$;

(A7) The step-size α satisfies $1/(2\sqrt{C}) \leq \alpha \leq 1/\sqrt{C}$.

The first four assumptions above concern the distribution and the relationship between the number of samples, dimension, and number of neurons in the network. These assumptions are relatively mild as they only require that the dimension and number of samples are logarithmically large. These assumptions ensure that the signal-to-noise ratio in the model is quite high, and that in the setting with no label noise $\eta = 0$, the optimal test error achievable is $o_n(1)$ (see Appendix D for more details). The final three assumptions concern the hyperparameters for the model and the optimization algorithm. Assumption (A5) ensures

that the network is wide enough to ensure there are enough random features at initialization for gradient descent to "amplify". It is noteworthy that assumption (A6) permits arbitrarily small (but nonzero) initialization variance. The assumption (A7) ensures that the step-size is large enough so that significant features can be learned after a single step of gradient descent but small enough so that optimization is stable.

3. Main results

Our main contribution is summarized in the following theorem.

Theorem 1 *Let $\delta \in (0, 1/2)$. For all $C > 1$ sufficiently large, under the assumptions (A1) through (A7), by running gradient descent with step-size α for $T = 1 + 1/(4\alpha)$ iterations, with probability at least $1 - 4\delta$ over the random initialization and the draws of the samples we have,*

1. *For the training points:*

$$\begin{aligned} \text{for all } i \in \mathcal{C}, \quad & y_i = \operatorname{sgn}\left(f(x_i; W^{(T)})\right), \\ \text{while for all } i \in \mathcal{N}, \quad & y_i \neq \operatorname{sgn}\left(f(x_i; W^{(T)})\right). \end{aligned}$$

2. *Further, the test error satisfies*

$$\mathbb{P}_{(x,y)\sim \mathsf{P}}\big(y \neq \operatorname{sgn}(f(x; W^{(T)}))\big) \leq \eta + C\sqrt{\frac{\log(1/\delta)}{n}}.$$

Theorem 1 shows that at time T, gradient descent learns a network that accurately classifies every clean sample, and incorrectly classifies every noisy sample, and achieves population risk close to the noise rate η. In Figure 1, we plot the decision boundary for a neural network trained by gradient descent when 15% of the training labels are flipped and we observe that indeed every noisy sample is incorrectly classified and every clean sample is correctly classified.

It is worth noting that the decision boundary displayed in Figure 1 is rather simple. Our proof below will show that this simplicity is due to the fact that nearly every neuron in the neural network will become highly correlated to one of the four cluster means $\{\pm\mu_1, \pm\mu_2\}$ so that the neural network essentially acts as the low-complexity classifier $x \mapsto \operatorname{sgn}(|\langle \mu_1, x\rangle| - |\langle \mu_2, x\rangle|)$. The main technical contribution of our work is the characterization of this feature-learning process and an examination of how it proceeds in the presence of noisy labels.

Let us remark that previous works on the generalization of neural networks in the feature-learning regime for variants of the XOR problem we study (without label noise) have sample complexities of order $O(\sqrt{d/n})$, which is an improvement over kernel-based methods which have sample complexity $\Omega(\sqrt{d^2/n})$ (Wei et al., 2019; Bai and Lee, 2020). By contrast, Theorem 1 provides a dimension-independent rate of $O(\sqrt{1/n})$. This difference is due to the fact that they consider an XOR problem with a lower signal-to-noise ratio than the one we consider. In particular, they assume the features are uniform on the hypercube

$\{\pm 1\}^d$ with labels given by $y = \mathrm{sgn}(x_i x_j)$ for distinct coordinates $i \neq j$. Since the variance in every direction is the same, the signal-to-noise ratio is thus of order $\Theta(1/d)$. In our setting, the variance in the signal directions is larger: the variance in the direction of μ_1 and μ_2 is equal to $1 + \sigma^2$ while the variance in the direction of any vector orthogonal to μ_1 and μ_2 is σ^2. Thus, the signal-to-noise ratio in our setting is of order $\Theta\left(\frac{1+\sigma^2}{d\sigma^2}\right) = \Omega(1)$ by Assumption (A2).

We note that our analysis does not rely upon the neural tangent kernel approximation. One way to see this is to observe that the assumption on the width of the network given in Assumption (A5) only requires the width to be larger than a fixed constant for a constant level of failure probability. Moreover, we show explicitly in the following proposition that for each sample, the feature maps given by the hidden layer activations change significantly from their values at random initialization, an essential characteristic of neural networks in the feature-learning regime (Yang and Hu, 2021).

Proposition 2 *Under the settings of Theorem 1, with probability at least* $1 - 4\delta$ *over the random initialization and draws of the samples, the feature maps of the neural network at time* $T = 1 + 1/(4\alpha)$ *satisfy, for all* $i \in [n]$,

$$\frac{\|\phi(W^{(T)}x_i) - \phi(W^{(0)}x_i)\|}{\|\phi(W^{(0)}x_i)\|} \geq \frac{1}{C\omega_{\mathrm{init}}\sqrt{md}} \geq \frac{1}{C}.$$

In particular, as $\omega_{\mathrm{init}}\sqrt{md} \to 0$*, the relative change in each sample's feature map is unbounded.*

The proof of Proposition 2 is given in Appendix C.

In the next section, we provide the proof of Theorem 1. The proof follows by concretely characterizing the type of features that different neurons learn throughout the training process.

4. Proofs

In this section, we provide an overview of the proof of Theorem 1. The detailed proofs are collected below in Appendix A.

We begin by introducing some additional notation that will be needed throughout the proofs. As stated above, the set of samples $\{(x_i, y_i)\}_{i=1}^n$ can be partitioned into clean samples and noisy samples, which are identified by the index sets $\mathcal{C}, \mathcal{N} \subset [n]$, respectively, and $\mathcal{C} \cup \mathcal{N} = [n]$. Each sample comes from one of four clusters, with possible means $\{\pm\mu_1, \pm\mu_2\}$, and we will identify these samples with $I_{+\mu_1}$, $I_{-\mu_1}$, $I_{+\mu_2}$, $I_{-\mu_2} \subset [n]$. We further decompose each of these cluster identification sets into the clean and noisy parts, that is, $I_{+\mu_1} = I_{+\mu_1}^{\mathcal{C}} \cup I_{+\mu_1}^{\mathcal{N}}$, and similarly for $I_{-\mu_1}$, $I_{+\mu_2}$, and $I_{-\mu_2}$. This notation allows for us to write $i \in I_{-\mu_1}$ when we mean $(x_i, y_i) = (-\mu_1 + z, -1)$, where $z \sim \mathsf{P}_{\mathrm{clust}}$. We use the short-hand notation $I_{\pm\mu_1}$ to denote $I_{+\mu_1} \cup I_{-\mu_1}$ and likewise for $I_{\pm\mu_2}$.

We note that there exists a natural neural network consisting of four ReLU neurons that can classify the (clean) data with high accuracy:

$$f^\star(x; W) := |\langle \mu_1, x\rangle| - |\langle \mu_2, x\rangle| = \phi(\langle \mu_1, x\rangle) + \phi(\langle -\mu_1, x\rangle) - \phi(\langle \mu_2, x\rangle) - \phi(\langle -\mu_2, x\rangle). \quad (2)$$

This ideal low-complexity classifier is suggestive of the following possibility: for positive neurons, corresponding to second-layer weights satisfying $a_j > 0$, the neurons become adapted to either the $+\mu_1$ cluster or the $-\mu_1$ cluster, depending upon the sign of $\langle w_j^{(0)}, \mu_1 \rangle$ at initialization. For negative neurons, corresponding to neurons with $a_j < 0$, the neurons become adapted to either the $+\mu_2$ cluster or the $-\mu_2$ cluster depending on the sign of $\langle w_j^{(0)}, \mu_2 \rangle$ at initialization. This is at a high-level the argument that we show below.

In the remainder of this section assume that Assumptions (A1) through (A7) are in force.

4.1 Random Initialization and Sample Properties

We begin with an analysis of the properties of the random initialization. In the lemma below, we derive concentration results on the norm of the random weights, as well as a count for the number of neurons that are correlated with a fixed vector at a given threshold level. The correlation part of the lemma will be the basis of a 'random feature amplification' phenomenon, whereby the relatively small (random) correlations of the neurons with different cluster means at initialization will be amplified into strong correlations by gradient descent.

Lemma 3 *Let $\delta \in (0, 1/2)$ and let $C_0 > 1$ be any absolute constant. Let $\mu \in \mathbb{R}^d$ satisfy $\|\mu\| = 1$. With probability at least $1 - \delta$, we have for all $j \in [m]$,*

$$\frac{1}{2}\omega_{\text{init}}\sqrt{d} \leq \|w_j^{(0)}\| \leq \frac{3}{2}\omega_{\text{init}}\sqrt{d},$$

and

$$\sum_{j=1}^{m} \mathbb{1}\left(|\langle w_j^{(0)}, \mu \rangle| \geq \frac{\omega_{\text{init}}}{2C_0}\right) \geq m \cdot \left(1 - \frac{1}{2C_0} - \sqrt{\frac{2\log(4/\delta)}{m}}\right).$$

Recall from (2) that there exists a neural network with four ReLU neurons that achieves high accuracy on the clean distribution $\tilde{\mathsf{P}}$, with the neuron weights corresponding to the four cluster means $\{\pm\mu_1, \pm\mu_2\}$. As we noted previously, a potential mechanism for neural network learning would be that most of the positive neurons (with second layer weights $a_j > 0$) become highly correlated with one of the $\pm\mu_1$ clusters while most of the negative neurons become highly correlated with one of the $\pm\mu_2$ clusters. If j-th neuron's weight w_j is highly correlated with a cluster mean $\mu \in \{\pm\mu_1, \pm\mu_2\}$, then for all samples x coming from the cluster μ, the sign of the activation for a neuron on the sample $\text{sgn}(\langle w_j, x \rangle)$ would be the same as the activation if the weight were exactly the cluster mean, $\text{sgn}(\langle \mu, x \rangle)$, so that the j-th neuron behaves similarly to the cluster mean μ. If this occurs we say that the j-th neuron *captures* the cluster with mean μ.

We show below that this 'capturing' phenomenon can be shown through a two-step process: first, at initialization, most of the positive neurons will have a normalized correlation with μ_1 of order $\Theta(1/\sqrt{d})$, and similarly most of the negative neurons will have a normalized correlation with μ_2 of order $\Theta(1/\sqrt{d})$. This is Lemma 4 below. Next, we show that by taking a single gradient step with a sufficiently large step-size, the normalized correlations for these neurons will improve from order $\Theta(1/\sqrt{d})$ to order $\Theta(1)$. This result, shown later in Lemma 12, is what we refer to as the 'random feature amplification' phenomenon, whereby

the random features at initialization are amplified into useful features by gradient descent. Towards this end, we characterize the correlations of the neurons with the cluster means at initialization in the following lemma.

Lemma 4 *Let $\delta \in (0, 1/2)$. For any absolute constant $C_0 > 1$, if C is sufficiently large, with probability at least $1 - \delta$ over the random initialization, there exist sets of neurons $J_{+\mu_1}, J_{-\mu_1}, J_{+\mu_2}, J_{-\mu_2} \subset [m]$ satisfying the following:*

$$\text{for } \mu \in \{\pm\mu_1\}, \quad |J_\mu| := \left|\left\{j : a_j > 0,\ \left\langle \frac{w_j^{(0)}}{\|w_j^{(0)}\|}, \mu \right\rangle \geq \frac{1}{3C_0\sqrt{d}}\right\}\right| \geq \frac{m}{4}\left(1 - \frac{1}{C_0}\right)^2,$$

$$\text{for } \mu \in \{\pm\mu_2\}, \quad |J_\mu| := \left|\left\{j : a_j < 0,\ \left\langle \frac{w_j^{(0)}}{\|w_j^{(0)}\|}, \mu \right\rangle \geq \frac{1}{3C_0\sqrt{d}}\right\}\right| \geq \frac{m}{4}\left(1 - \frac{1}{C_0}\right)^2.$$

In particular, $J := J_{\pm\mu_1} \cup J_{\pm\mu_2}$ satisfies $|J| \geq m(1 - 1/C_0)^2$.

Lemma 4 identifies a set of candidate neurons that are partially correlated with the cluster *means* $\{\pm\mu_1, \pm\mu_2\}$. We would like to translate this result into a statement about the data, and to do so, we first need to provide some basic facts about samples from the distribution. The reader may find it helpful to refer back to the beginning of Section 4 where we introduce the $I_{\pm\mu_i}$ notation.

Lemma 5 *There is a universal constant $C_1 \geq 2$ such that the following holds. For any $\delta \in (0, 1/2)$, for all $C > 1$ large enough, with probability at least $1 - \delta$ over $S \sim \mathsf{P}^n$, the following holds.*

(a) For each $\mu \in \{\pm\mu_1, \pm\mu_2\}$ and $\mu^\perp$ orthogonal to μ,

$$\text{for all } i \in I_\mu, \quad \langle x_i, \mu\rangle \geq 1 - C_1\sigma\sqrt{d} \geq 1 - 1/C_1, \quad \text{and} \quad |\langle x_i, \mu^\perp\rangle| \leq C_1\sigma\sqrt{d} \leq 1/C_1.$$

(b) For all $\mu \in \{\pm\mu_1, \pm\mu_2\}$, for any $i \in I_\mu$, $\|x_i - \mu\|^2 \leq C_1\sigma^2 d \leq 1/C_1$.

(c) The fraction of noisy points $\frac{|\mathcal{N}|}{n} \leq \eta + C_1\sqrt{\log(1/\delta)/n} \leq \eta + 1/C_1$.

(d) For any cluster $\mu \in \{\pm\mu_1, \pm\mu_2\}$ and any $0 \leq t \leq T - 1$, we have

$$\frac{1}{4} - C_1\sqrt{\frac{\log(1/\delta)}{n}} \leq \frac{1}{n}|I_\mu| \leq \frac{1}{4} + C_1\sqrt{\frac{\log(1/\delta)}{n}}.$$

Now, recall that Lemma 4 shows that a large fraction of the neurons will 'capture' at least one of the four cluster centers with a normalized correlation of $\langle w_j^{(0)}/\|w_j^{(0)}\|, \mu_s\rangle \geq \Omega(1/\sqrt{d})$. Since the within-cluster variance is of order $\sigma = O(1/\sqrt{d})$, there is not enough signal for these neurons to capture all *samples* within each cluster. However, the following lemma demonstrates that capturing the cluster mean with a normalized correlation threshold of order $1/\sqrt{d}$ suffices to guarantee that a strictly larger portion of the samples from that cluster will be captured than not. This technical lemma will be key to our subsequent analysis.

Lemma 6 *There exists a universal constant $C_2 > 1$ such that for any $\delta \in (0, 1/2)$, for all $C > 1$ large enough, with probability at least $1 - 2\delta$, both Lemma 5 and the following event holds. For any $j \in [m]$ satisfying $\langle w_j^{(0)}/\|w_j^{(0)}\|, \mu\rangle \geq 1/(3C_0\sqrt{d})$ for some $\mu \in \{\pm\mu_1, \pm\mu_2\}$, it holds that*

$$\sum_{i \in I_{+\mu}^{\mathcal{C}}} \phi'(\langle w_j^{(0)}, x_i\rangle) - \sum_{i \in I_{-\mu}^{\mathcal{C}}} \phi'(\langle w_j^{(0)}, x_i\rangle) \geq \frac{n}{C_2}.$$

In light of the above, we introduce the following definition.

Definition 7 *We define the event where all parts of Lemma 3, Lemma 4 (with $C_0 = 4^5 \cdot 1024^2 \exp(4)$), Lemma 5, and Lemma 6 hold a* good run.

By the above lemmas, we know that for any $\delta \in (0, 1/2)$, for all $C > 1$ large enough, a good run occurs with probability at least $1 - 4\delta$. In the remainder of this section, we will assume that a good run occurs.

4.2 Sufficient Conditions for a Large Margin Classifier via a Good Subnetwork

Our proof will rely upon the notion of a good *subnetwork* of the neural network. For index set $\tilde{J} \subset [m]$ and matrix $W \in \mathbb{R}^{m \times d}$, denote by $W_{\tilde{J}} \in \mathbb{R}^{|\tilde{J}| \times d}$ as the sub-matrix of W consisting of rows with indices from $\tilde{J}$. Denote by $f^{\tilde{J}}(x; \cdot)$ the subnetwork consisting of rows from $\tilde{J}$,

$$f^{\tilde{J}}(x; W) := \sum_{j \in \tilde{J}} a_j \sigma(\langle w_j, x\rangle).$$

The below lemma demonstrates that in order to show that the neural network produces a good margin, it suffices to show that there exists a large subnetwork that produces a good margin provided that the weights of the network are bounded.

Lemma 8 *Let $J \subset [m]$, and denote $J^c = [m] \setminus J$. If $W \in \mathbb{R}^{m \times d}$ is such that $\|W\|_F \leq 1$ and there is a constant $C_f > 1$ such that $yf^J(x; W) \geq 1/C_f$ for some $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$, then provided $\|x\| \leq 2$ and $|J^c|/m \leq 1/(16C_f^2)$, we have $yf(x; W) \geq 1/(2C_f)$.*

Lemma 8 demonstrates that in order to show the neural network classifies an example correctly, it suffices to identify a large subnetwork that does so. The rest of our proof is dedicated to showing that this happens. The subnetwork that performs well is defined in terms of the neurons $j \in J_{\pm\mu_1} \cup J_{\pm\mu_2}$, where the index sets $J_{\pm\mu_1} \cup J_{\pm\mu_2}$ are defined in Lemma 4 and are shown to constitute a large fraction of all of the neurons: for each $\mu \in \{\pm\mu_1, \pm\mu_2\}$, the set J_μ has cardinality at least $|J_\mu| \geq \frac{m}{4}(1 - 1/C_0)^2$, where $C_0 > 1$ is a large constant. We next define two conditions that we will show suffice for showing this subnetwork classifies examples correctly, which we refer to as the *neuron alignment condition* and the *almost-orthogonality condition*. We describe the first of these below.

Condition 9 (Neuron alignment condition) *We say that the* neuron alignment condition *holds at time t if the subsets of neurons $J_{\pm\mu_1}$ and $J_{\pm\mu_2}$ defined in Lemma 4 satisfy the following: for every $\mu \in \{\pm\mu_1, \pm\mu_2\}$, and for all $j \in J_\mu$,*

$$\phi'(\langle w_j^{(t)}, x_k\rangle) = 1 \text{ for all } k \in I_\mu, \quad \text{and} \quad \phi'(\langle w_j^{(t)}, x_k\rangle) = 0 \text{ for all } k \in I_{-\mu}.$$

The neuron alignment condition loosely states that there is a substantial number of neurons (the neurons in the sets $J_{\pm\mu_1} \cup J_{\pm\mu_2}$) that completely capture each of the clusters in the sense that all samples within each cluster have the same ReLU activation, which is "on" on one of the clusters and "off" on the opposing cluster. By Lemma 4, we know that there is a large fraction of neurons that catch the cluster *means* $\{\pm\mu_1, \pm\mu_2\}$ at initialization. However, as we argued prior to Lemma 6, because the normalized correlation between the neurons at initialization and the cluster means is of order $1/\sqrt{d}$ while the variance within each cluster is also of order $1/\sqrt{d}$, a substantial portion of the examples within each cluster will *not* be captured by a neuron at initialization. We briefly note here that in the next section, we will show that a single step of gradient descent suffices to address this problem.

We next introduce the notion of *almost-orthogonality*, which will be key to showing that the subnetwork is able to classify examples correctly with a positive margin. This condition ensures that the $J_{\pm\mu_1}$ neurons capture the $\pm\mu_1$ clusters, and are *almost orthogonal* to data from the $\pm\mu_2$ clusters, and vice versa for the $J_{\pm\mu_2}$ neurons. In particular, this will allow for us to say that the subnetwork satisfies $f^J(x; W) \approx |\langle \mu_1, x\rangle| - |\langle \mu_2, x\rangle|$, which one can verify produces a good margin for clean data $(x, \tilde{y}) \sim \tilde{\mathsf{P}}$.

Condition 10 (Almost-orthogonality) *We say* almost-orthogonality holds up to time τ *if for all* $t \le \tau$,

$$\text{for all } j \in J_{\pm\mu_1}, \quad |\langle w_j^{(t)}, \mu_2\rangle| \le 3\alpha|a_j|, \quad \text{and} \quad \text{for all } j \in J_{\pm\mu_2}, \quad |\langle w_j^{(t)}, \mu_1\rangle| \le 3\alpha|a_j|.$$

The almost-orthogonality condition ensures that the projection of the $J_{\pm\mu_1}$ (resp. $J_{\pm\mu_2}$) neurons onto the space spanned by μ_2 (resp. μ_1) remains small for all iterates of gradient descent up to time τ.

In the next lemma, we show that the combination of neuron alignment and almost-orthogonality suffices to produce a good subnetwork margin. Note that we consider times $t \ge 1$ with foresight, as we shall eventually show that neuron alignment and almost-orthogonality hold for all $t \ge 1$.

Lemma 11 *Let $J = J_{\pm\mu_1} \cup J_{\pm\mu_2}$, where the sets $J_{\pm\mu_1}$ and $J_{\pm\mu_2}$ are defined in Lemma 4. Suppose that neuron alignment (Condition 9) and almost-orthogonality (Condition 10) hold at times $\tau = 1, \dots, T-1 = 1/(4\alpha)$. Then, on a good run, for all $C > 1$ large enough, at time $T = 1 + 1/(4\alpha)$, we have $\|W^{(T)}\|_F \le 1$, and that*

$$\text{for all } i \in \mathcal{C}, \quad y_i f^J(x_i; W^{(T)}) \ge \frac{1}{C_3} > 0, \quad \text{and}$$
$$\text{for all } i \in \mathcal{N}, \quad y_i f^J(x_i; W^{(T)}) \le -\frac{1}{C_3} < 0,$$

where $C_3 = 4096\exp(2)/(1 - 1/C_0)^2$ and $C_0 > 1$ is the constant from Lemma 4.

Lemma 11 demonstrates that in order to show that a given subnetwork $f^J(x; W)$ accurately classifies all of the clean data, it suffices to show that neuron alignment and almost-orthogonality hold for a sufficiently large (but constant) number of steps. By Lemma 8, this translates to a guarantee for the entire network $f(x; W)$ if we can show that the subnetwork is sufficiently large, which we can ensure by taking J as the union of the sets

$J_{\pm\mu_1}, J_{\pm\mu_2} \subset [m]$ as in Lemma 4 and by taking the constant $C_0 > 1$ from that lemma to be sufficiently large. Thus, to complete the proof, we need only verify that neuron alignment and almost-orthogonality hold for a sufficiently large but constant number of steps. This is what we show in the next subsection.

We note that both neuron alignment and almost-orthogonality are needed in order to ensure that the subnetwork $f^J(x; W)$ behaves like the simple classifier $|\langle \mu_1, x\rangle| - |\langle \mu_2, x\rangle|$. For instance, consider what happens if half of the positive neurons (corresponding to $j \in [m]$ with $a_j > 0$) are proportional to $\mu_1 + 100\mu_2$ and the other half are proportional to $-\mu_1 - 100\mu_2$, and likewise half of the negative neurons are proportional to μ_2 and the other half are proportional to $-\mu_2$. Then the neuron alignment condition would hold, but almost-orthogonality would not hold, and the network would behave like the predictor $|\langle \mu_1 + 100\mu_2, x\rangle| - |\langle \mu_2, x\rangle|$ and not generalize well. Thus, in addition to showing that the neurons are highly correlated with the cluster means from a given class, we must also show that they are nearly orthogonal to the cluster means from the opposite class.

4.3 Gradient Descent Produces a Large Margin Classifier

As mentioned previously, we cannot expect neuron alignment to hold at initialization, as the random features that define the subnetwork f^J have per-neuron normalized correlations $\langle w_j^{(0)}/\|w_j^{(0)}\|, \mu_1\rangle$ of order $O(1/\sqrt{d})$, while the fluctuations within each cluster $\langle w_j^{(0)}/\|w_j^{(0)}\|, \mu_s - x_i\rangle$ are also of order $\sigma = O(1/\sqrt{d})$. This means that many samples x_i belonging to a cluster μ_s will satisfy $\mathrm{sgn}(\langle w_j^{(0)}, x_i\rangle) \neq \mathrm{sgn}(\langle w_j^{(0)}, \mu_s\rangle)$, preventing the satisfaction of the neuron alignment condition. This is where Lemma 6 will play a role: although the random features have normalized correlations of order $\Theta(1/\sqrt{d})$ with the cluster means, this signal provides an 'edge' in terms of the ReLU activations of samples within the cluster. That is, having $\langle w_j^{(0)}/\|w_j^{(0)}\|, \mu_s\rangle \geq c/\sqrt{d}$ is sufficient to guarantee that the fraction of samples within the μ_s cluster sharing the same sign as $\langle w_j^{(0)}, \mu_s\rangle$ is at least $1/2 + \Delta$ for some absolute constant $\Delta > 0$. This provides enough signal for gradient descent to latch onto and 'amplify' the normalized per-neuron correlations from $\langle w_j^{(0)}/\|w_j^{(0)}\|, \mu_s\rangle \geq c/\sqrt{d}$ to $\langle w_j^{(1)}/\|w_j^{(1)}\|, \mu_2\rangle \geq c'$ after one sufficiently large step. Since now the normalized correlations are of order 1 while the within-cluster fluctuations are of order $1/\sqrt{d}$, this allows for neuron alignment to hold after a single step of gradient descent.

Lemma 12 *For $C > 1$ sufficiently large, on a good run Condition 9 holds at time $t = 1$. Moreover, letting $C_2 > 1$ denote the constant from Lemma 6, the per-neuron normalized correlations satisfy*

$$\text{for every } \mu \in \{\pm\mu_1, \pm\mu_2\} \text{ and every } j \in J_\mu, \quad \left\langle w_j^{(1)}/\|w_j^{(1)}\|, \mu \right\rangle \geq \frac{1}{16C_2}.$$

We now know that neuron alignment holds at time $t = 1$, and that the number of neurons that are characterized by the alignment condition is quite large (precisely, $m(1 - 1/C_0)^2$ for a large constant C_0). By Lemma 11, if we can show that (i) neuron alignment continues to hold for a certain number of steps, (ii) almost-orthogonality holds throughout these steps, and (iii) we early-stop so that the hidden layer weights are not too large, then there will be

a large subnetwork that classifies clean examples with a positive margin. In the next lemma, we inductively argue that this is the case.

Lemma 13 *For $C > 1$ sufficiently large, on a good run, for every time $t = 1, \ldots, 1/(4\alpha)$, neuron alignment (Condition 9) holds at time t and almost-orthogonality (Condition 10) holds up to time t.*

We emphasize that although Lemma 12 shows that neuron alignment holds at time $t = 1$, this is not sufficient to guarantee generalization since we must ensure that the positive (respectively negative) neurons are not highly correlated to $\pm\mu_2$ (respectively $\pm\mu_1$) since this could result in inaccurate predictions as outlined at the end of Section 4.2. This potential problem is precisely what almost-orthogonality (Condition 10) prevents, and Lemma 13 shows that by running gradient descent for a large (but constant) number of steps, we can guarantee that both neuron alignment and almost-orthogonality hold up to time $T - 1 = 1/(4\alpha)$. By Lemma 11, this implies that at time T the subnetwork $f^J(x; W^{(T)})$ classifies all of the clean examples correctly, and by Lemma 8 this implies that the full network $f(x; W^{(T)})$ classifies all of the clean examples correctly with small $\|W^{(T)}\|_F$. From here the proof of Theorem 1 is a straightforward Rademacher-complexity based argument; the details are provided in Appendix A.4.

5. Discussion

We have shown that two-layer neural networks with ReLU activations trained by gradient descent can achieve small test error on a distribution for which linear classifiers perform no better than random guessing. We developed a novel proof technique that detailed how using a random initialization provides a collection of random features that gradient descent is able to amplify into stronger, useful features for prediction. Importantly, our analysis holds when a constant fraction of the training labels are arbitrarily corrupted.

Our analysis requires the usage of early-stopping, so that gradient descent only runs for $T = O(1)$ iterations. We showed that running gradient descent for $O(1)$ iterations is sufficient to achieve classification error close to the noise rate. The reason $T = O(1)$ is helpful is that under this assumption, the weights assigned to each sample in the gradient descent updates (proportional to $-\ell'(y_i f(x_i; W^{(t)}))$) are not too small, so that the useful signals from each sample can be used to push the neural network weights in a good direction. Early-stopping also allows for a uniform convergence-based argument for the generalization error of the trained network. Without early-stopping, there is the potential for the neural network to overfit to noisy labels, and it is a natural question whether the network will still generalize near-optimally when it has overfit (i.e., whether or overfitting is 'benign' as in Bartlett et al. (2020); Frei et al. (2022)).

In Figure 2, we examine the behavior of two-layer ReLU networks trained by gradient descent on the logistic loss for the 2-XOR distribution we consider when 15% of the labels are flipped (for full experimental details, see Appendix E). We consider two distinct settings: a low-dimensional setting where $n \gg d$ and a high-dimensional setting where $d \gg n$. In the low-dimensional setting, the test accuracy decreases after the network overfits to the noisy training data, while in the high-dimensional setting the test accuracy remains at the optimal 85% level even after reaching the point of interpolation. Since our assumptions only

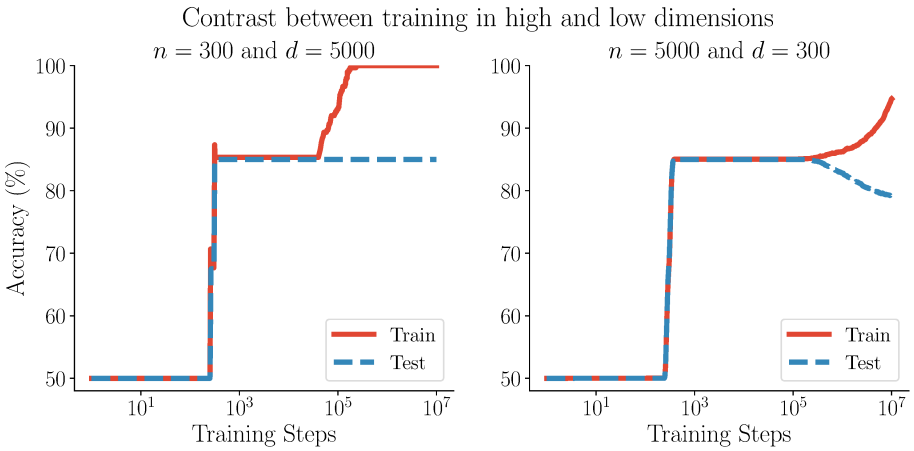

Figure 2: Training and validation accuracy for a two-layer ReLU network with $m = 400$ neurons trained on P (within-cluster variance of $\sigma^2 = 1/d^{1.2}$) when 15% of the labels within each cluster are flipped to the opposing cluster. When the network begins to overfit to the noisy labels, the test accuracy decreases in the $n \gg d$ setting while it remains optimal in the $d \gg n$ setting.

require that the number of samples and dimension are not super-exponential in the other, this suggests that we would need to introduce new techniques, separately tailored to the low-dimensional and high-dimensional settings, in order to characterize the generalization behavior of the network after the point of interpolation.

Another natural direction for future research is to understand whether or not the random feature amplification phenomenon that we identified in two-layer networks has an analogue in deeper networks. Yet another direction is to understand whether this analysis technique can be generalized to settings with more cluster centers.

Acknowledgements

We thank the anonymous reviewers for their numerous suggestions which helped improve the presentation of the paper. We thank Hongren Yan and Yutong Wang for pointing out issues in a previous version of this paper. We gratefully acknowledge the support of the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2023505, DMS-2031883, and #814639.

Appendix A. Omitted Proofs from Section 4

In this appendix, we provide the proofs for all of the lemmas in Section 4. In Section A.1, we provide the proofs for the lemmas that involve concentration inequalities: Lemmas 3, 4, 5, and 6. Next, we prove Lemmas 8 and 11, which show that producing a good subnetwork suffices for the neural network to classify the clean examples correctly and provide a sufficient condition for producing a good subnetwork. In Section A.3, we show that gradient descent produces a good subnetwork. Finally, in Section C, we provide a proof of Proposition 2, which emphasizes that the feature maps produced by gradient descent differ significantly from those found at initialization.

We remind the reader that throughout this section we assume that Assumptions (A1)-(A7) are in effect. We also note that $C > 1$ is always used to denote the constant used in these assumptions.

A.1 Random Initialization and Sample Properties

In this subsection we provide the proofs for Lemmas 3, 4, 5, and 6.

A.1.1 Proof of Lemma 3

We restate the lemma here for the reader's convenience.

Lemma 3 *Let $\delta \in (0, 1/2)$ and let $C_0 > 1$ be any absolute constant. Let $\mu \in \mathbb{R}^d$ satisfy $\|\mu\| = 1$. With probability at least $1 - \delta$, we have for all $j \in [m]$,*

$$\frac{1}{2}\omega_{\text{init}}\sqrt{d} \leq \|w_j^{(0)}\| \leq \frac{3}{2}\omega_{\text{init}}\sqrt{d},$$

and

$$\sum_{j=1}^{m} \mathbb{1}\left(|\langle w_j^{(0)}, \mu\rangle| \geq \frac{\omega_{\text{init}}}{2C_0}\right) \geq m \cdot \left(1 - \frac{1}{2C_0} - \sqrt{\frac{2\log(4/\delta)}{m}}\right).$$

Proof We first prove the first part of the lemma. Note that for fixed $j \in [m]$, there are i.i.d. $z_i \sim \mathsf{N}(0, 1)$ such that

$$\|w_j^{(0)}\|^2 = \sum_{i=1}^{d} (w_j^{(0)})_i^2 = \omega_{\text{init}}^2 \sum_{i=1}^{d} z_i^2 \sim \omega_{\text{init}}^2 \cdot \chi^2(d).$$

By concentration of the χ^2 distribution (see, Wainwright, 2019, Example 2.11), for any $t \in (0, 1)$,

$$\mathbb{P}\left(\left|\frac{1}{d\omega_{\text{init}}^2}\|w_j^{(0)}\|^2 - 1\right| \geq t\right) \leq 2\exp(-dt^2/8).$$

In particular, by taking $t = \sqrt{8\log(4m/\delta)/d}$, we have that for fixed $j \in [m]$, with probability at least $1 - \delta/2m$,

$$\frac{1}{2}d\omega_{\text{init}}^2 \leq (1-t)d\omega_{\text{init}}^2 \leq \|w_j^{(0)}\|^2 \leq (1+t)d\omega_{\text{init}}^2 \leq \frac{3}{2}d\omega_{\text{init}}^2,$$

where we have used Assumption (A1), that is, $d \geq C\log(m/\delta)$for a sufficiently large constant $C > 1$ implies $t \leq 1/2$. Applying a union bound over $j \in [m]$ shows that the bound on the norms at initialization holds over all j with probability at least $1 - \delta/2$.

For the neuron-counting argument, let $z \sim \mathsf{N}(0,1)$ denote a standard normal random variable. Denote by p the probability

$$p := \mathbb{P}(|\langle w_j^{(0)}, \mu\rangle| \geq \omega_{\text{init}}/(2C_0)) = \mathbb{P}_{z\sim\mathsf{N}(0,1)}(|z| \geq 1/(2C_0)).$$

By anti-concentration of the Gaussian, we have

$$1 - p = \mathbb{P}(|z| \leq 1/(2C_0)) \leq 1/(2C_0) \tag{3}$$

Define random variables $U_j := \mathbb{1}(|\langle w_j^{(0)}, \mu\rangle| \geq \omega_{\text{init}}/(2C_0))$. Since U_j are 1-sub-Gaussian, Hoeffding's inequality implies for any $t \geq 0$,

$$\mathbb{P}\left(\left|\sum_{j=1}^{m}(U_j - p)\right| \geq t\right) \leq 2\exp(-t^2/2m).$$

Thus, with probability at least $1 - \delta/2$, we have

$$\begin{aligned}\sum_{j=1}^{m}\mathbb{1}(|\langle w_j^{(0)}, \mu\rangle| \geq \omega_{\text{init}}/(2C_0)) &= \sum_{j=1}^{m} U_j \\ &\geq mp - \sqrt{2m\log(4/\delta)} \\ &\overset{(i)}{\geq} m\cdot\left(1 - \frac{1}{2C_0} - \sqrt{\frac{2\log(4/\delta)}{m}}\right),\end{aligned}$$

where in (i) we use (3).

Taking a union bound over the first and second parts of the proof shows that both claims hold simultaneously with probability at least $1 - \delta$. ■

A.1.2 Proof of Lemma 4

We restate and prove Lemma 4 below.

Lemma 4 *Let $\delta \in (0, 1/2)$. For any absolute constant $C_0 > 1$, if C is sufficiently large, with probability at least $1 - \delta$ over the random initialization, there exist sets of neurons $J_{+\mu_1}, J_{-\mu_1}, J_{+\mu_2}, J_{-\mu_2} \subset [m]$ satisfying the following:*

$$\text{for } \mu \in \{\pm\mu_1\}, \quad |J_\mu| := \left|\left\{j : a_j > 0,\ \left\langle \frac{w_j^{(0)}}{\|w_j^{(0)}\|}, \mu\right\rangle \geq \frac{1}{3C_0\sqrt{d}}\right\}\right| \geq \frac{m}{4}\left(1 - \frac{1}{C_0}\right)^2,$$

$$\text{for } \mu \in \{\pm\mu_2\}, \quad |J_\mu| := \left|\left\{j : a_j < 0,\ \left\langle \frac{w_j^{(0)}}{\|w_j^{(0)}\|}, \mu\right\rangle \geq \frac{1}{3C_0\sqrt{d}}\right\}\right| \geq \frac{m}{4}\left(1 - \frac{1}{C_0}\right)^2.$$

In particular, $J := J_{\pm\mu_1} \cup J_{\pm\mu_2}$ satisfies $|J| \geq m(1 - 1/C_0)^2$.

Proof Fix $C_0 > 1$. Apply Lemma 3 to the positive neurons j satisfying $a_j > 0$ with μ_1. This tells us that with probability at least $1 - \delta/16$,

$$|J_{\pm\mu_1}| := |\{j : a_j > 0,\ |\langle w_j^{(0)}, \mu_1\rangle| \geq \omega_{\text{init}}/2C_0\}| \geq \frac{m}{2}\left(1 - \frac{1}{2C_0} - \sqrt{\frac{4\log(32/\delta)}{m}}\right) \geq \frac{m}{2}(1 - 1/C_0),$$

where we have used Assumption (A5) so that we may take $m \geq 16C_0^2 \log(32/\delta)$. Notice that for $w_j^{(0)} \neq 0$, the event $\{\langle w_j^{(0)}, \mu_1\rangle > 0\}$ depends only on the angle between $w_j^{(0)}$ and μ_1, while the event $\{|\langle w_j^{(0)}, \mu_1\rangle| \geq \xi\omega_{\text{init}}\}$ depends only on the product $\|w_j^{(0)}\|\|\mu_1\|$. Thus the sign of $\langle w_j^{(0)}, \mu_1\rangle$ is independent of whether or not $j \in J_{\pm\mu_1}$. Since $\mathbb{P}(\langle w_j^{(0)}, \mu_1\rangle > 0) = 1/2$, by Hoeffding's inequality we know that with probability at least $1-\delta/16$, at least $\frac{1}{2} - \sqrt{\frac{4\log(32/\delta)}{m}}$ fraction of the indices in $J_{\pm\mu_1}$ satisfy $\langle w_j^{(0)}, \mu_1\rangle > 0$ and likewise at least $\frac{1}{2} - \sqrt{\frac{4\log(32/\delta)}{m}}$ fraction of the indices in $J_{\pm\mu_1}$ satisfy $\langle w_j^{(0)}, \mu_1\rangle < 0$. In particular, taking a union bound we have with probability at least $1 - \delta/8$,

$$\left|\left\{j : a_j > 0,\ \langle w_j^{(0)}, \mu_1\rangle \geq \frac{\omega_{\text{init}}}{2C_0}\right\}\right| \geq \frac{m}{4}(1 - 1/C_0) \cdot \left(1 - \sqrt{\frac{4\log(32/\delta)}{m}}\right) \geq \frac{m}{4}(1 - 1/C_0)^2,$$

where we have again used Assumption (A5). We can argue similarly for neurons satisfying $\langle w_j^{(0)}, -\mu_1\rangle \geq \omega_{\text{init}}/2C_0$ and neurons satisfying $\langle w_j^{(0)}, \pm\mu_2\rangle \geq \omega_{\text{init}}/2C_0$ to get that with probability at least $1 - \delta/2$,

$$\begin{aligned}
|\{j : a_j > 0,\ \langle w_j^{(0)}, \mu_1\rangle \geq \omega_{\text{init}}/2C_0\}| &\geq \frac{m}{4}(1 - 1/C_0)^2,\\
|\{j : a_j > 0,\ \langle w_j^{(0)}, -\mu_1\rangle \geq \omega_{\text{init}}/2C_0\}| &\geq \frac{m}{4}(1 - 1/C_0)^2,\\
|\{j : a_j < 0,\ \langle w_j^{(0)}, \mu_2\rangle \geq \omega_{\text{init}}/2C_0\} &\geq \frac{m}{4}(1 - 1/C_0)^2,\\
|\{j : a_j < 0,\ \langle w_j^{(0)}, -\mu_2\rangle \geq \omega_{\text{init}}/2C_0\}| &\geq \frac{m}{4}(1 - 1/C_0)^2.
\end{aligned}$$

By Lemma 3, we know that with probability at least $1 - \delta/2$, $\|w_j^{(0)}\| \leq \frac{3}{2}\omega_{\text{init}}\sqrt{d}$, and thus whenever $\langle w_j^{(0)}, \mu_1\rangle \geq \omega_{\text{init}}/2C_0$, we have

$$\left\langle \frac{w_j^{(0)}}{\|w_j^{(0)}\|}, \mu_1 \right\rangle \geq \frac{\omega_{\text{init}}}{2C_0\|w_j^{(0)}\|} \geq \frac{1}{3C_0\sqrt{d}}.$$

Taking a union bound, we see that with probability at least $1-\delta$,

$$|J_{+\mu_1}| := |\{j : a_j > 0,\ \langle w_j^{(0)}/\|w_j^{(0)}\|, \mu_1\rangle \geq 1/(3C_0\sqrt{d})\}| \geq \frac{m}{4}(1-1/C_0)^2,$$
$$|J_{-\mu_1}| := |\{j : a_j > 0,\ \langle w_j^{(0)}/\|w_j^{(0)}\|, -\mu_1\rangle \geq 1/(3C_0\sqrt{d})\}| \geq \frac{m}{4}(1-1/C_0)^2,$$
$$|J_{+\mu_2}| := |\{j : a_j < 0,\ \langle w_j^{(0)}/\|w_j^{(0)}\|, \mu_2\rangle \geq 1/(3C_0\sqrt{d})\}| \geq \frac{m}{4}(1-1/C_0)^2,$$
$$|J_{-\mu_2}| := |\{j : a_j < 0,\ \langle w_j^{(0)}/\|w_j^{(0)}\|, -\mu_2\rangle \geq 1/(3C_0\sqrt{d})\}| \geq \frac{m}{4}(1-1/C_0)^2.$$

■

A.1.3 Proof of Lemma 5

We restate and prove Lemma 5 below.

Lemma 5 *There is a universal constant $C_1 \geq 2$ such that the following holds. For any $\delta \in (0, 1/2)$, for all $C > 1$ large enough, with probability at least $1-\delta$ over $S \sim \mathsf{P}^n$, the following holds.*

(a) For each $\mu \in \{\pm\mu_1, \pm\mu_2\}$ and $\mu^\perp$ orthogonal to μ,

$$\text{for all } i \in I_\mu,\quad \langle x_i, \mu\rangle \geq 1 - C_1\sigma\sqrt{d} \geq 1 - 1/C_1, \quad \text{and} \quad |\langle x_i, \mu^\perp\rangle| \leq C_1\sigma\sqrt{d} \leq 1/C_1.$$

(b) For all $\mu \in \{\pm\mu_1, \pm\mu_2\}$, for any $i \in I_\mu$, $\|x_i - \mu\|^2 \leq C_1\sigma^2 d \leq 1/C_1$.

(c) The fraction of noisy points $\frac{|\mathcal{N}|}{n} \leq \eta + C_1\sqrt{\log(1/\delta)/n} \leq \eta + 1/C_1$.

(d) For any cluster $\mu \in \{\pm\mu_1, \pm\mu_2\}$ and any $0 \leq t \leq T-1$, we have

$$\frac{1}{4} - C_1\sqrt{\frac{\log(1/\delta)}{n}} \leq \frac{1}{n}|I_\mu| \leq \frac{1}{4} + C_1\sqrt{\frac{\log(1/\delta)}{n}}.$$

Proof We shall show that each part of the lemma holds with a large enough probability and then take a union bound to establish our claim.

Proof of parts (a) and (b): We consider the case $i \in I_{+\mu_1}$. The cases of $i \in I_\mu$ for $\mu \in \{-\mu_1, \pm\mu_2\}$ follow using identical arguments.

Let $i \in I_{\mu_1}$. We begin by noting that, since $\|\mu\| = 1$, we have by Cauchy–Schwarz,

$$\begin{aligned}\langle x_i, \mu_1\rangle &= \langle x_i - \mu_1, \mu_1\rangle + \langle \mu_1, \mu_1\rangle \\ &\geq 1 - \|x_i - \mu\|. \end{aligned} \tag{4}$$

Therefore, to derive a lower bound on $\langle x_i, \mu_1\rangle$ when $i \in I_{+\mu_1}$, it suffices to derive an upper bound on $\|x_i - \mu\|$ for each i, so that we will first prove part (b).

Since $(x_i - \mu)/\sigma$ is isotropic and log-concave, by concentration of the Euclidean norm of isotropic log-concave random vectors (Adamczak et al., 2014, Theorem 1), there is a universal constant $c > 0$ such that,

$$\mathbb{P}(\|(x_i - \mu)/\sigma\| \geq cu\sqrt{d}) \leq \exp(-cu\sqrt{d}).$$

In particular, using Assumption (A1), we can take $d \geq \log^2(32n/\delta)/c^2$ so that $\exp(-cu\sqrt{d}) \leq \delta/(32n)$ and thus we have with probability at least $1 - \delta/32$, for all $i \in I_{+\mu_1}$,

$$\|x_i - \mu\| \leq c\sigma\sqrt{d}.$$

This, along with Assumption (A2) proves part (b).

Using (4) and Assumption (A2) so that $c\sigma\sqrt{d} < 1$, we have

$$\langle x_i, \mu_1 \rangle \geq 1 - c\sigma\sqrt{d},$$

which proves the first half of part (a) of the lemma when $i \in I_{+\mu_1}$. When $i \in I_{+\mu_1}$, the cluster mean $\mu^\perp$ orthogonal to μ_1 is μ_2, and so we have,

$$|\langle x_i, \mu_2 \rangle| = |\langle x_i - \mu_1, \mu_2 \rangle| \leq \|x_i - \mu_1\| \leq c\sigma\sqrt{d},$$

which completes the proof of the second part of (a) when $\mu = \mu_1$. Taking a union bound over $\mu \in \{\pm\mu_1, \pm\mu_2\}$ shows that parts (a) and (b) hold with probability at least $1 - \delta/8$.

Proof of part (c): We note that $\{\mathbb{1}(y_i \neq \tilde{y}_i)\}_{i=1}^n$ are a collection of n i.i.d. random variables bounded by one with expectation equal to the noise rate η. For some absolute constant $c > 1$, Hoeffding's inequality therefore gives, for any $u \geq 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}(y_i \neq \tilde{y}_i) - \eta \geq u\right) = \mathbb{P}\left(\frac{|\mathcal{N}|}{n} - \eta \geq u\right) \leq \exp\left(-\frac{nu^2}{2c}\right).$$

In particular, for $u = \sqrt{\frac{2c\log(2/\delta)}{n}}$, by Assumption (A3) we have with probability at least $1 - \delta/2$, $|\mathcal{N}|/n \leq \eta + \sqrt{2c\log(2/\delta)/n} \leq \eta + 1/C_1$ by Assumption (A4).

Proof of part (d): We consider the case $\mu = +\mu_1$ with identical arguments holding for $\mu \in \{-\mu_1, \pm\mu_2\}$. Notice that the random variables $\{\mathbb{1}(i \in I_{+\mu_1})\}_{i=1}^n$ are i.i.d. Bernoulli with mean $1/4$, since the samples are drawn uniformly from the four clusters $\{\pm\mu_1, \pm\mu_2\}$. Thus, by Hoeffding's inequality, for some absolute constant $c > 1$, we have with probability at least $1 - \delta/16$,

$$\frac{1}{4} - c\sqrt{\frac{\log(32/\delta)}{n}} \leq \frac{1}{n}|I_{+\mu_1}| \leq \frac{1}{4} + c\sqrt{\frac{\log(32/\delta)}{n}}. \tag{5}$$

Since $\delta \in (0, 1/2)$, there is a larger constant $c' > 0$ such that $c\sqrt{\frac{\log(32/\delta)}{n}} \leq c'\sqrt{\frac{\log(1/\delta)}{n}}$. Taking a union bound over the four clusters shows that part (d) holds with probability at least $1 - \delta/4$.

Thus all four parts (a), (b), (c), (d) hold with probability at least $1 - \delta$. ■

A.1.4 Proof of Lemma 6

We restate and prove Lemma 6 below.

Lemma 6 *There exists a universal constant $C_2 > 1$ such that for any $\delta \in (0, 1/2)$, for all $C > 1$ large enough, with probability at least $1 - 2\delta$, both Lemma 5 and the following event*

holds. For any $j \in [m]$ satisfying $\langle w_j^{(0)}/\|w_j^{(0)}\|, \mu\rangle \geq 1/(3C_0\sqrt{d})$ for some $\mu \in \{\pm\mu_1, \pm\mu_2\}$, it holds that

$$\sum_{i \in I^{\mathcal{C}}_{+\mu}} \phi'(\langle w_j^{(0)}, x_i\rangle) - \sum_{i \in I^{\mathcal{C}}_{-\mu}} \phi'(\langle w_j^{(0)}, x_i\rangle) \geq \frac{n}{C_2}.$$

Proof We shall prove this lemma in two parts. First, we shall define a "good event" $\mathcal{E}$ that occurs with probability at least $1 - 2\delta$. Then via a deterministic argument, we shall show that the lemma holds whenever this good event occurs.

Defining the good event. Fix some $\mu \in \{\pm\mu_1, \pm\mu_2\}$. By definition of P, there are $z_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_{\text{clust}}$ such that

$$N_\mu(j) := \sum_{i \in I^{\mathcal{C}}_{\mu}} \phi'(\langle w_j^{(0)}, x_i\rangle) = \sum_{i \in I^{\mathcal{C}}_{\mu}} \mathbb{1}(\langle w_j^{(0)}, \mu\rangle + \sigma\langle z_i, w_j^{(0)}\rangle > 0).$$

Similarly, there are $u_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_{\text{clust}}$ such that

$$N_{-\mu}(j) := \sum_{i \in I^{\mathcal{C}}_{-\mu}} \phi'(\langle w_j^{(0)}, x_i\rangle) = \sum_{i \in I^{\mathcal{C}}_{-\mu}} \mathbb{1}(-\langle w_j^{(0)}, \mu\rangle + \sigma\langle u_i, w_j^{(0)}\rangle > 0).$$

Thus, if we define,

$$p := \mathbb{P}_{z \sim \mathsf{P}_{\text{clust}}}(\langle w_j^{(0)}, \mu\rangle + \sigma\langle z, w_j^{(0)}\rangle > 0),$$

then we have,

$$N_\mu(j) \sim \mathsf{Binomial}(|I^{\mathcal{C}}_\mu|, p) \quad \text{and} \quad N_{-\mu}(j) \sim \mathsf{Binomial}(|I^{\mathcal{C}}_{-\mu}|, 1-p).$$

This motivates deriving upper and lower bounds for the cardinality of the sets $I^{\mathcal{C}}_\mu$ and $I^{\mathcal{C}}_{-\mu}$. To do so, we first note that with probability at least $1 - \delta$, all of the events in Lemma 5 hold. In particular, by Part (d) of that lemma, we have with probability at least $1 - \delta$, for any $\mu \in \{\pm\mu_1, \pm\mu_2\}$,

$$\frac{1}{4} - C_1\sqrt{\frac{\log(1/\delta)}{n}} \leq \frac{1}{n}|I_\mu| = \frac{1}{n}(|I^{\mathcal{C}}_\mu| + |I^{\mathcal{N}}_\mu|) \leq \frac{1}{4} + C_1\sqrt{\frac{\log(1/\delta)}{n}}. \tag{6}$$

We thus have with probability at least $1 - \delta$, all of the events in Lemma 5 hold, and, for all $\mu \in \{\pm\mu_1, \pm\mu_2\}$,

$$\frac{n}{8} \overset{(i)}{\leq} \frac{n}{4}\left(1 - C_1\sqrt{\frac{\log(1/\delta)}{n}} - \frac{|\mathcal{N}|}{n}\right) \leq |I^{\mathcal{C}}_\mu| \leq \frac{n}{4}\left(1 + C_1\sqrt{\frac{\log(1/\delta)}{n}}\right), \tag{7}$$

where the inequality (i) uses Assumptions (A3) and (A4).

Now, since $N_\mu(j) \sim \mathsf{Binomial}(|I^{\mathcal{C}}_\mu|, p)$, by Hoeffding's inequality and a union bound (over the neurons and over the clusters), there is some $c > 0$ such that with probability at least $1 - \delta$, for all $\mu \in \{\pm\mu_1, \pm\mu_2\}$, for all $j \in [m]$,

$$|I^{\mathcal{C}}_\mu| \cdot \left(p - c\sqrt{\frac{\log(64m/\delta)}{|I^{\mathcal{C}}_\mu|}}\right) \leq N_\mu(j) \leq |I^{\mathcal{C}}_\mu| \cdot \left(p + c\sqrt{\frac{\log(64m/\delta)}{|I^{\mathcal{C}}_\mu|}}\right). \tag{8}$$

Let us define $\mathcal{E}$ to be the event where the events in Lemma 5, and inequalities (7) and (8) all simultaneously hold. By a union bound this happens with probability at least $1 - 2\delta$. This shall determine the success probability of the lemma.

Lemma holds whenever the *good event* $\mathcal{E}$ occurs. In the remainder of the proof let us assume that this event $\mathcal{E}$ occurs; we will show that the lemma holds as a deterministic consequence of these events.

Since the event $\mathcal{E}$ occurs, for all $\mu \in \{\pm\mu_1, \pm\mu_2\}$ and all $j \in [m]$ we have,

$$\begin{aligned} N_\mu(j) &\overset{(i)}{\geq} |I_\mu^{\mathcal{C}}| \left(p - c\sqrt{\frac{\log(64m/\delta)}{|I_\mu^{\mathcal{C}}|}} \right) \\ &\overset{(ii)}{\geq} \frac{n}{4} \left(p - 3c\sqrt{\frac{\log(64m/\delta)}{n}} \right) \left(1 - C_1\sqrt{\frac{\log(1/\delta)}{n}} - \frac{|\mathcal{N}|}{n} \right) \\ &\overset{(iii)}{\geq} \frac{n}{4} \left[\left(1 - \frac{|\mathcal{N}|}{n} \right) p - 4c\sqrt{\frac{\log(64m/\delta)}{n}} \right]. \end{aligned} \tag{9}$$

Above, (i) uses Eq. (8), while (ii) uses Eq. (7). Inequality (iii) uses Assumption (A3) so that $n \geq \log(64m/\delta)$ and by taking c to be a larger constant. Further, for all $\mu \in \{\pm\mu_1, \pm\mu_2\}$ and all $j \in [m]$, we have,

$$\begin{aligned} N_{-\mu}(j) &\leq |I_{-\mu}^{\mathcal{C}}| \left((1-p) + c\sqrt{\frac{\log(32m/\delta)}{|I_{-\mu}|}} \right) \\ &\overset{(i)}{\leq} \frac{n}{4} \left(1 + C_1\sqrt{\frac{\log(1/\delta)}{n}} \right) \left((1-p) + 3c\sqrt{\frac{\log(32m/\delta)}{n}} \right) \\ &\overset{(ii)}{\leq} \frac{n}{4} \left[\left(1 + 4C_1\sqrt{\frac{\log(1/\delta)}{n}} \right) (1-p) + 4c\sqrt{\frac{\log(64m/\delta)}{n}} \right]. \end{aligned} \tag{10}$$

Above, (i) uses (7) and (ii) uses the assumption $n \geq C\log(1/\delta)$ given by (A3). Thus, we have shown that, when the good event $\mathcal{E}$ occurs, then inequalities (9) and (10) hold. In the remainder of the proof, we will show that the lemma follows as a consequence of the inequalities (9) and (10).

In order to show $N_\mu(j) \gg N_{-\mu}(j)$, it suffices to show that p is large enough so that there is sufficient 'edge' for more samples to be captured by w_j than not. To this end, we have for any j such that $\langle w_j^{(0)}, \mu \rangle > 0$,

$$\begin{aligned} p &= \mathbb{P}_{z\sim \mathsf{P}_{\text{clust}}}(\langle w_j^{(0)}, \mu \rangle + \sigma \langle z, w_j^{(0)} \rangle > 0) \\ &= \mathbb{P}_{z\sim \mathsf{P}_{\text{clust}}} \left(\left\langle z, \frac{w_j^{(0)}}{\|w_j^{(0)}\|} \right\rangle > -\frac{\langle w_j^{(0)}, \mu \rangle}{\sigma \|w_j^{(0)}\|} \right) \\ &= \frac{1}{2} + \mathbb{P}_{z\sim \mathsf{P}_{\text{clust}}} \left(\left\langle z, \frac{w_j^{(0)}}{\|w_j^{(0)}\|} \right\rangle \in \left[-\frac{\langle w_j^{(0)}, \mu \rangle}{\sigma \|w_j^{(0)}\|}, 0 \right] \right). \end{aligned} \tag{11}$$

Recall that we are considering neurons $j \in [m]$ such that $\langle w_j^{(0)}/\|w_j^{(0)}\|, \mu \rangle \geq 1/(3C_0\sqrt{d})$. By assumption (A2), for C sufficiently large we have $\sigma\sqrt{d} \leq 1/C \leq 3/C_0$ so that the inclusion

$[-1/9, 0] \subset [-1/(3C_0\sigma\sqrt{d}), 0]$ holds. Thus, we have,

$$p \geq \frac{1}{2} + \mathbb{P}_{z\sim \mathsf{P}_{\text{clust}}}\left(\left\langle z, \frac{w_j^{(0)}}{\|w_j^{(0)}\|}\right\rangle \in \left[-\frac{1}{3C_0\sigma\sqrt{d}}, 0\right]\right)$$
$$\geq \frac{1}{2} + \mathbb{P}_{z\sim \mathsf{P}_{\text{clust}}}\left(\left\langle z, \frac{w_j^{(0)}}{\|w_j^{(0)}\|}\right\rangle \in \left[-\frac{1}{9}, 0\right]\right).$$

Note that $\langle z, w_j^{(0)}/\|w_j^{(0)}\|\rangle$ is the projection of a log-concave isotropic random vector onto the one dimensional subspace spanned by $w_j^{(0)}/\|w_j^{(0)}\|$, and thus by Diakonikolas et al. (2020, Definition 1.2, Fact A.4) there exists an absolute constant $c_1 > 0$ such that

$$\mathbb{P}_{z\sim \mathsf{P}_{\text{clust}}}\left(\left\langle z, \frac{w_j^{(0)}}{\|w_j^{(0)}\|}\right\rangle \in \left[-\frac{1}{9}, 0\right]\right) \geq c_1, \tag{12}$$

and continuing from the previous display we thus have

$$p \geq \frac{1}{2} + c_1. \tag{13}$$

We can thus use the inequalities given in events (9) and (10) to see that for any $\mu \in \{\pm\mu_1, \pm\mu_2\}$ and for any $j \in [m]$ such that $\langle w_j^{(0)}/\|w_j^{(0)}\|, \mu\rangle \geq 1/(3C_0\sqrt{d})$,

$$\begin{aligned}
&N_\mu^{(0)}(j) - N_{-\mu}^{(0)}(j) \\
&\quad\geq \frac{n}{4}\left[\left(1 - \frac{|\mathcal{N}|}{n}\right)p - \left(1 + 4C_1\sqrt{\frac{\log(64m/\delta)}{n}}\right)(1-p) - 8c\sqrt{\frac{\log(64m/\delta)}{n}}\right] \\
&\quad\geq \frac{n}{4}\left[\left(2 - \frac{|\mathcal{N}|}{n}\right)p - 1 - 10c\sqrt{\frac{\log(64m/\delta)}{n}}\right] \\
&\quad\overset{(i)}{\geq} \frac{n}{4}\left[\left(2 - \frac{|\mathcal{N}|}{n}\right)\left(\frac{1}{2} + c_1\right) - 1 - 10c\sqrt{\frac{\log(64m/\delta)}{n}}\right] \\
&\quad\overset{(ii)}{\geq} \frac{n}{4}\left[2c_1 - \frac{|\mathcal{N}|}{n} - 10c\sqrt{\frac{\log(64m/\delta)}{n}}\right] \\
&\quad\overset{(iii)}{\geq} \frac{n}{4}\cdot c_1.
\end{aligned} \tag{14}$$

In (i), we have used (13). Inequality (ii) follows by a direct calculation. Finally, (iii) uses that Assumption (A3) ensures $n \geq 4 \cdot 100c^2c_1^{-2}\log(64m/\delta)$, as well as Lemma 5(c) and Assumption (A4).

This shows that there exists a universal constant $C_2 > 1$ such that whenever event $\mathcal{E}$ occurs, for all $\mu \in \{\pm\mu_1, \pm\mu_2\}$ and

$$\text{for all } j \text{ such that } \langle w_j^{(0)}/\|w_j^{(0)}\|, \mu\rangle \geq 1/(3C_0\sqrt{d}), \quad N_\mu^{(0)}(j) - N_{-\mu}^{(0)}(j) \geq \frac{n}{C_2}. \tag{15}$$

Putting things together. Recall that above we argued that the event $\mathcal{E}$ (which is when the events in Lemma 5 and Equations (7) and (8) all hold simultaneously) occurs with probability at least $1 - 2\delta$, and since this implies the claim in Equation (15) holds, this completes the proof. ■

A.2 Sufficient Conditions for a Large Margin Classifier via a Good Subnetwork

In this subsection, we prove Lemmas 8 and 11, which demonstrate that in order to show the neural network correctly classifies all clean samples, it suffices to show that there exists a large subnetwork that classifies the points correctly. Before we prove this, we introduce the following auxiliary lemma, which bounds the growth of the weights of the network over time. This lemma will be used in a number of places in the remaining proofs.

A.2.1 Auxiliary Lemma on Neuron Weight Growth

Lemma 14 *For $C > 1$ large enough, on a good run we have the following bound on the norms of the weights for times $t \geq 1$:*

1. *For all $j \in [m]$, $\|w_j^{(t)}\| \leq 2|a_j|\alpha t$;*
2. *$\|W^{(t)}\|_F \leq 2\alpha t$.*

Proof First, note that since a good run occurs, Lemma 5 and Assumption (A2) imply that for any sample $i \in [n]$, we have $\|x_i - \mu_s\|^2 \leq C_1\sigma^2 d + \frac{1}{C_1} \leq 1/3$, where μ_s is the cluster mean corresponding to x_i. Therefore, we have for any $i \in [n]$,

$$\|x_i\| \leq (1 + 1/\sqrt{3}) \leq \sqrt{2}. \tag{16}$$

We can thus bound,

$$\begin{aligned}
\|w_j^{(t)} - w_j^{(0)}\| &\leq \alpha \sum_{\tau=0}^{t-1} \|\nabla_j \widehat{L}(W^{(\tau)})\| \\
&= \alpha \sum_{\tau=0}^{t-1} \left\| \frac{1}{n} \sum_{i=1}^{n} -\ell'_{i,\tau} y_i a_j \phi'(\langle w_j^{(\tau)}, x_i \rangle) x_i \right\| \\
&\leq \alpha \sum_{\tau=0}^{t-1} \frac{1}{n} \sum_{i=1}^{n} |\ell'_{i,\tau}||a_j|\phi'(\langle w_j^{(\tau)}, x_i \rangle)\|x_i\| \\
&\leq \alpha t |a_j| \sqrt{2}, \qquad (17)
\end{aligned}$$

where the final inequality uses (16) and $|\ell'_{i,\tau}| \leq 1$. Therefore, by the triangle inequality and Lemma 3,

$$\begin{aligned}
\|w_j^{(t)}\| &\leq \|w_j^{(0)}\| + \sqrt{2}|a_j|\alpha t \\
&\leq \frac{3}{2}\omega_{\text{init}}\sqrt{d} + \sqrt{2}|a_j|\alpha t \\
&\leq 2|a_j|\alpha t,
\end{aligned}$$

where the final inequality uses Assumptions (A6) and (A7) so that $\omega_{\text{init}}\sqrt{md} \le \alpha/3$ for $C > 1$ sufficiently large. The bound on the Frobenius norm follows by noting that $\|W^{(t)}\|_F^2 = \sum_{j=1}^m \|w_j^{(t)}\|^2$ and that $|a_j| = 1/\sqrt{m}$. ■

A.2.2 Proof of Lemma 8

With the above lemma in hand, we now restate and prove Lemma 8.

Lemma 8 *Let $J \subset [m]$, and denote $J^c = [m] \setminus J$. If $W \in \mathbb{R}^{m\times d}$ is such that $\|W\|_F \le 1$ and there is a constant $C_f > 1$ such that $yf^J(x;W) \ge 1/C_f$ for some $(x,y) \in \mathbb{R}^d \times \{\pm 1\}$, then provided $\|x\| \le 2$ and $|J^c|/m \le 1/(16C_f^2)$, we have $yf(x;W) \ge 1/(2C_f)$.*

Proof By definition,

$$f(x;W) = f^J(x;W) + f^{J^c}(x;W) = \sum_{j\in J} a_j\phi(\langle w_j, x\rangle) + \sum_{j\in J^c} a_j\phi(\langle w_j, x\rangle).$$

For the latter term, note that

$$\begin{aligned}
|f^{J^c}(x;W)| &= \left|\sum_{j\in J^c} a_j\phi(\langle w_j, x\rangle)\right| \\
&\overset{(i)}{\le} \sqrt{\sum_{j\in J^c} a_j^2}\sqrt{\sum_{j\in J^c}\langle w_j, x\rangle^2} \\
&= \sqrt{\frac{|J^c|}{m}}\|W_{J^c}x\|_2 \\
&\le \sqrt{\frac{|J^c|}{m}}\|W_{J^c}\|_2\|x\| \\
&\le \sqrt{\frac{|J^c|}{m}}\|W_{J^c}\|_F\|x\|.
\end{aligned}$$

In (i) we use the Cauchy–Schwarz inequality, and that ϕ is 1-Lipschitz with $\phi(0) = 0$. The final claim follows as $\|W_{J^c}\|_F \le \|W\|_F \le 1$, so that

$$f(x;W) \ge f^J(x;W) - \sqrt{\frac{|J^c|}{m}}\|W_{J^c}\|_F\|x\| \ge \frac{1}{C_f} - \frac{1}{4C_f}\cdot 1\cdot 2 = \frac{1}{2C_f}.$$

■

A.2.3 Proof of Lemma 11

In this section we restate and prove Lemma 11.

Lemma 11 *Let $J = J_{\pm\mu_1} \cup J_{\pm\mu_2}$, where the sets $J_{\pm\mu_1}$ and $J_{\pm\mu_2}$ are defined in Lemma 4. Suppose that neuron alignment (Condition 9) and almost-orthogonality (Condition 10) hold*

at times $\tau = 1, \dots, T-1 = 1/(4\alpha)$. Then, on a good run, for all $C > 1$ large enough, at time $T = 1 + 1/(4\alpha)$, we have $\|W^{(T)}\|_F \leq 1$, and that

$$\begin{aligned} &\text{for all } i \in \mathcal{C}, \quad y_i f^J(x_i; W^{(T)}) \geq \frac{1}{C_3} > 0, \quad \text{and} \\ &\text{for all } i \in \mathcal{N}, \quad y_i f^J(x_i; W^{(T)}) \leq -\frac{1}{C_3} < 0, \end{aligned}$$

where $C_3 = 4096 \exp(2)/(1 - 1/C_0)^2$ and $C_0 > 1$ is the constant from Lemma 4.

Proof By Lemma 14, we have that for all $\tau \in \{1, \dots, T\}$

$$\|W^{(\tau)}\|_F \leq 2\alpha\tau \leq 1,$$

since $\tau \leq T = 1/(4\alpha) + 1$. This shows the claimed guarantee for the norm.

We now show the claim for the margin. First, note that we have for any $x \in \mathbb{R}^d$ and $W \in \mathbb{R}^{m \times d}$, since ϕ is 1-Lipschitz, Cauchy–Schwarz gives

$$|f(x; W)| = \left| \sum_{j=1}^{m} a_j \phi(\langle w_j, x \rangle) \right| \leq \sqrt{\sum_{j=1}^{m} a_j^2} \sqrt{\sum_{j=1}^{m} \langle w_j, x \rangle^2} = \|Wx\| \leq \|W\|_2 \|x\|. \tag{18}$$

Using Lemma 5(b), we can therefore bound the neural network output at time τ by

$$|f(x_i; W^{(\tau)})| \leq \|W^{(\tau)}\|_F \|x_i\| \leq 2, \quad \text{for all } i \in [n],\ \tau \leq T - 1.$$

Note that $-\ell'(z)$ is a decreasing function and also that $-\ell'(z) \geq 1/2 \exp(-z)$ on $z \geq 0$. Therefore,

$$-\ell'(y_i f(x_i; W^{(\tau)})) \geq \frac{1}{2} \exp(-2), \quad \text{for all } i \in [n],\ \tau \leq T - 1. \tag{19}$$

We will now show that for sufficiently large t, the network produces a positive margin on the $+\mu_1$. This shall be crucial in showing that the network produces a positive margin on the clean points associated with this cluster, and a negative margin on the noisy points in the cluster.

Recall the notation $-\ell'_{i,t} = -\ell'(y_i f(x_i; W^{(t)}))$. Since neuron alignment holds, we have for $j \in J_{+\mu_1}$ and $\tau \le T-1$,

$$\begin{aligned}
\langle w_j^{(\tau+1)} - w_j^{(\tau)}, \mu_1 \rangle &= \frac{\alpha|a_j|}{n} \sum_{i=1}^{n} -\ell'_{i,\tau} y_i \phi'(\langle w_j^{(\tau)}, x_i \rangle) \langle x_i, \mu_1 \rangle \\
&= \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{C}}_{+\mu_1}} -\ell'_{i,\tau} \langle x_i, \mu_1 \rangle - \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{N}}_{+\mu_1}} -\ell'_{i,\tau} \langle x_i, \mu_1 \rangle \\
&\quad - \frac{\alpha|a_j|}{n} \sum_{i \in I_{+\mu_2}} -\ell'_{i,\tau} y_i \phi'(\langle w_j^{(\tau)}, x_i \rangle) \langle x_i, \mu_1 \rangle \\
&\quad - \frac{\alpha|a_j|}{n} \sum_{i \in I_{-\mu_2}} -\ell'_{i,\tau} y_i \phi'(\langle w_j^{(\tau)}, x_i \rangle) \langle x_i, \mu_1 \rangle \\
&\overset{(i)}{\ge} \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{C}}_{+\mu_1}} -\ell'_{i,\tau} \cdot \frac{1}{2} - \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{N}}_{+\mu_1}} -\ell'_{i,\tau} \cdot \frac{3}{2} \\
&\quad - \frac{\alpha|a_j|}{n} \sum_{i \in I_{+\mu_2}} -\ell'_{i,\tau} \cdot C_1 \sigma \sqrt{d} - \frac{\alpha|a_j|}{n} \sum_{i \in I_{-\mu_2}} -\ell'_{i,\tau} \cdot C_1 \sigma \sqrt{d} \\
&\overset{(ii)}{\ge} \alpha|a_j| \cdot \left[\frac{|I^{\mathcal{C}}_{+\mu_1}|}{n} \cdot \frac{\exp(-2)}{4} - \frac{|I^{\mathcal{N}}_{+\mu_1}|}{n} \cdot \frac{3}{2} - \frac{|I_{\pm\mu_2}|}{n} \cdot C_1 \sigma \sqrt{d} \right] \\
&\overset{(iii)}{\ge} \alpha|a_j| \cdot \left[\frac{\exp(-2)}{32} - \frac{3}{2} \cdot \frac{|\mathcal{N}|}{n} - \frac{C_1}{C} \right] \\
&\overset{(iv)}{\ge} \alpha|a_j| \left[\frac{\exp(-2)}{32} - \frac{3}{2} \left(\frac{1}{C} + C_1 \sqrt{\frac{1}{C}} \right) - \frac{C_1}{C} \right] \\
&\overset{(v)}{\ge} \frac{\exp(-2)\alpha|a_j|}{64}.
\end{aligned} \tag{20}$$

In (i) we use Lemma 5: for the sums over $I^{\mathcal{C}}_{+\mu_1}$ and $I^{\mathcal{N}}_{+\mu_1}$, part (b) of the lemma and the assumption on σ given in Assumption (A2) imply that $\|x_i - \mu_1\| \le 1/2$ for C large enough and hence $\langle x_i, \mu_1 \rangle = \langle x_i - \mu_1, \mu_1 \rangle + 1 \in [1/2, 3/2]$ for $i \in I_{+\mu_1}$. For the sums over $i \in I_{\pm\mu_2}$, part (a) of Lemma 5 implies $|\langle x_i, \mu_1 \rangle| \le C_1 \sigma \sqrt{d}$ and using this with $|y_i \phi'| \le 1$ provides the desired bound. For inequality (ii), we use (19) and that ℓ is 1-Lipschitz. In inequality (iii), we use the lower bound $|I^{\mathcal{C}}_{+\mu_1}| \ge n/8$ given in Eq. (7), as well as $|I_{\pm\mu_2}| \le n$ and the upper bound for σ given in Assumption (A2). For the inequality (iv), we use Lemma 5(c) and the assumptions on the noise rate and number of samples given in Assumptions (A4) and (A3) to bound $|\mathcal{N}|/n \le \eta + C_1 \sqrt{\log(1/\delta)/n} \le 1/C + C_1 \sqrt{1/C}$. Then (v) follows by taking C to be a large enough universal constant.

Summing (20) from $\tau = 1, \dots, T-1$ and using that $j \in J_{+\mu_1}$ implies $\langle w_j^{(1)}, \mu_1 \rangle > 0$, we get that

$$\langle w_j^{(T)}, \mu_1 \rangle \ge \langle w_j^{(T)} - w_j^{(1)}, \mu_1 \rangle \ge \frac{\exp(-2)\alpha|a_j|(T-1)}{64}, \quad \text{for all } j \in J_{+\mu_1}. \tag{21}$$

Thus, we have the following lower bound on the network output at μ_1:

$$\begin{aligned}
f^J(\mu_1; W^{(T)}) &= \sum_{j \in J_{+\mu_1}} a_j \phi(\langle w_j^{(T)}, \mu_1 \rangle) + \sum_{j \in J_{-\mu_1}} a_j \phi(\langle w_j^{(T)}, \mu_1 \rangle) + \sum_{j \in J_{\pm\mu_2}} a_j \phi(\langle w_j^{(T)}, \mu_1 \rangle) \\
&\overset{(i)}{=} \sum_{j \in J_{+\mu_1}} a_j \langle w_j^{(T)}, \mu_1 \rangle + \sum_{j \in J_{\pm\mu_2}} a_j \phi(\langle w_j^{(T)}, \mu_1 \rangle) \\
&\overset{(ii)}{\geq} \sum_{j \in J_{+\mu_1}} a_j \langle w_j^{(T)}, \mu_1 \rangle - \sum_{j \in J_{\pm\mu_2}} |a_j \langle w_j^{(T)}, \mu_1 \rangle| \\
&\overset{(iii)}{\geq} \sum_{j \in J_{+\mu_1}} a_j \langle w_j^{(T)}, \mu_1 \rangle - \frac{3\alpha |J_{\pm\mu_2}|}{m} \\
&\overset{(iv)}{\geq} \alpha \left[\frac{|J_{+\mu_1}|(T-1)\exp(-2)}{64m} - \frac{3|J_{\pm\mu_2}|}{m} \right] \\
&\overset{(v)}{\geq} \alpha \left[\frac{(T-1)\exp(-2)}{256}(1 - 1/C_0)^2 - 3 \right].
\end{aligned}$$

In (i) we use the neuron alignment condition. In (ii) we use that ϕ is 1-Lipschitz. In (iii) we use the almost-orthogonality (Condition 10) and that $|a_j| = 1/\sqrt{m}$. In (iv) we use Eq. (21) and again use the fact that $|a_j| = 1/\sqrt{m}$. Finally, (v) uses Lemma 4, so that we have $|J_{+\mu_1}|/m \geq \frac{1}{4}(1 - 1/C_0)^2$, as well as the fact that $|J_{\pm\mu_2}| \leq m$. In particular, we see that for $T - 1 = 1/(4\alpha)$, we have

$$f^J(\mu_1; W^{(T)}) \geq \frac{\exp(-2)}{1024}(1 - 1/C_0)^2 - 3\alpha \geq \frac{\exp(-2)}{2048}(1 - 1/C_0)^2. \tag{22}$$

In the last inequality, we use the Assumption (A7) and take $C > 1$ large enough so that $\alpha \leq \exp(-2)(1 - 1/C_0)^2/(6 \cdot 1024)$. With a lower bound on the margin for the cluster center μ_1 established, we can translate this result to one for samples using Lemma 5. To do so, note that the sub-network $f^J(\cdot; W)$ is $\|W\|_F$-Lipschitz in the network input, i.e., we have

$$\begin{aligned}
|f^J(x; W) - f^J(x'; W)| &= \left| \sum_{j \in J} a_j [\sigma(\langle w_j, x \rangle) - \sigma(\langle w_j, x' \rangle)] \right| \\
&\leq \|a\| \sqrt{\sum_{j=1}^m \langle w_j, x - x' \rangle^2} \\
&\leq \|W\|_F \|x - x'\|,
\end{aligned}$$

where the first inequality follows by Cauchy–Schwarz inequality and the last inequality follows since $\|a\| = \sum_{j=1}^m a_j^2 = 1$ and $\|W(x - x')\| \leq \|W\|_F \|x - x'\|$. Therefore we can use Lemma 5 (b) to translate (22) into a guarantee for the samples. For any $i \in I_{+\mu_1}^{\mathcal{C}}$, so that

$y_i = +1$,

$$\begin{aligned} y_i f^J(x_i; W^{(T)}) &\geq y_i f^J(\mu_1; W^{(T)}) - \|W^{(T)}\|_F \max_i \|x_i - \mu_1\| \\ &\geq \frac{\exp(-2)}{2048}(1 - 1/C_0)^2 - C_1\sigma\sqrt{d} \\ &\geq \frac{\exp(-2)}{4096}(1 - 1/C_0)^2. \end{aligned}$$

The second inequality uses that $\|W^{(T)}\|_F \leq 1$ and Lemma 5, while the last inequality uses Assumption (A2) so that $C_1\sigma\sqrt{d}$ can be taken smaller than any absolute constant for $C > 1$ sufficiently large.

This completes the proof for samples $i \in I^{\mathcal{C}}_{+\mu_1}$. To see that the network also incorrectly classifies noisy samples, take $i \in I^{\mathcal{N}}_{+\mu_1}$, so that $y_i = -1$. Then, again using Lemma 5(b),

$$\begin{aligned} y_i f^J(x_i W^{(T)}) &= -f^J(x_i W^{(T)}) \\ &\leq -f^J(\mu_1; W^{(T)}) + \|W^{(T)}\|_F \max_i \|x_i - \mu_1\| \\ &\leq -\frac{\exp(-2)}{4096}(1 - 1/C_0)^2, \end{aligned}$$

where the last inequality follows since $\|W_F^{(T)}\| \leq 1$ as we proved above.

For the other clusters, an identical argument to (20) yields

$$\begin{aligned} \langle w_j^{(\tau+1)} - w_j^{(\tau)}, -\mu_1 \rangle &\geq \frac{\alpha|a_j|}{64}\exp(-2), \quad \text{for all } j \in J_{-\mu_1}, \tau \leq T-1, \\ \langle w_j^{(\tau+1)} - w_j^{(\tau)}, \mu_2 \rangle &\geq \frac{\alpha|a_j|}{64}\exp(-2), \quad \text{for all } j \in J_{+\mu_2}, \tau \leq T-1, \\ \langle w_j^{(\tau+1)} - w_j^{(\tau)}, -\mu_2 \rangle &\geq \frac{\alpha|a_j|}{64}\exp(-2), \quad \text{for all } j \in J_{-\mu_2}, \tau \leq T-1. \end{aligned} \tag{23}$$

We can utilize the identities (23) and similar arguments to show that the desired margin condition holds for other clusters $I^{\mathcal{C}}_{-\mu_1}, I^{\mathcal{C}}_{\pm\mu_2}$ so the result holds for all $i \in \mathcal{C}$. ■

A.3 Gradient Descent Produces a Large Margin Classifier

In this section, we show that the sufficient conditions necessary for producing a good subnetwork described in Lemma 11 hold. The first step for this is to show that neuron alignment holds at time $t = 1$.

A.3.1 Proof of Lemma 12

We restate and prove Lemma 12 below.

Lemma 12 *For $C > 1$ sufficiently large, on a good run Condition 9 holds at time $t = 1$. Moreover, letting $C_2 > 1$ denote the constant from Lemma 6, the per-neuron normalized correlations satisfy*

$$\text{for every } \mu \in \{\pm\mu_1, \pm\mu_2\} \text{ and every } j \in J_\mu, \quad \left\langle w_j^{(1)}/\|w_j^{(1)}\|, \mu \right\rangle \geq \frac{1}{16C_2}.$$

Proof Since a good run occurs, all of the events in Lemma 3, Lemma 4, Lemma 5, and Lemma 6 hold. Recall that the sets $J_{\pm\mu_1}$ and $J_{\pm\mu_2}$ were defined in Lemma 4. We will now show that Condition 9 holds for these sets at time $t = 1$. We will demonstrate the first claim in the condition statement (regarding μ_1), that is, for all $j \in J_{+\mu_1}$:

$$\phi'(\langle w_j^{(1)}, x_k\rangle) = 1 \quad \text{for all } k \in I_{+\mu_1},$$
$$\phi'(\langle w_j^{(1)}, x_k\rangle) = 0 \quad \text{for all } k \in I_{-\mu_1}.$$

The remaining parts of the neuron alignment condition concerning $j \in J_{-\mu_1} \cup J_{\pm\mu_2}$ shall follow by using an identical argument.

There are two parts to the neuron alignment condition, let us begin by proving that the first part holds.

Part 1 of NAC: Let us begin by showing that for all $j \in J_{+\mu_1}$:

$$\phi'(\langle w_j^{(1)}, x_k\rangle) = 1 \quad \text{for all } k \in I_{+\mu_1}. \tag{24}$$

Recall that by the definition of the set $J_{+\mu_1}$, we have that for all $j \in J_{+\mu_1}$,

$$\phi'(\langle w_j^{(0)}, \mu_1\rangle) = 1.$$

To show that the first part of NAC holds for the subset $J_{+\mu_1}$, we need to show that a step of gradient descent takes ensures that all of the samples from this cluster are captured by the neurons in $J_{+\mu_1}$. We shall prove this in stages.

1. First, we shall establish a relation between the parameters after one the first step of gradient descent $w_j^{(1)}$ and those at initialization $w_j^{(0)}$.

2. Then, we shall leverage this relation to show that the angle between $w_j^{(1)}$ and μ_1 is small.

3. This, along with the fact that the samples from this cluster are close to its center, shall be sufficient to ensure that (24) is satisfied.

Step 1: First, recall that by the calculation (18), we have $|f(x_i; W^{(0)})| \leq \|W^{(0)}\|_F \|x_i\|$. Thus the bound $\|w_j^{(0)}\| \leq \frac{3}{2}\omega_{\text{init}}\sqrt{d}$ by Lemma 3, the bound $\|x_i\| \leq \sqrt{2}$ from (16) imply that

$$|f(x_i; W^{(0)})| \leq \|W^{(0)}\|_F \|x_i\| \leq 3\omega_{\text{init}}\sqrt{md}.$$

Note that $z \mapsto -\ell'(z)$ is a decreasing function, and thus

$$-\ell'_{i,0} \in \left[-\ell'\left(3\omega_{\text{init}}\sqrt{md}\right), -\ell'\left(-3\omega_{\text{init}}\sqrt{md}\right)\right]. \tag{25}$$

With this in place, let us analyze the gradient update for a neuron in the set $J_{+\mu_1}$. Recall that for such nodes, $a_j = 1/\sqrt{m} > 0$ and therefore,

$$\begin{aligned}
&w_j^{(1)} \\
&= w_j^{(0)} + \frac{\alpha a_j}{n} \sum_{i=1}^{n} -\ell'_{i,0} y_i x_i \phi'(\langle w_j^{(0)}, x_i \rangle) \\
&= w_j^{(0)} + \frac{\alpha a_j}{n} \sum_{i \in I_{+\mu_1}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)\mu_1 + \frac{\alpha a_j}{n} \sum_{i \in I_{-\mu_1}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)(-\mu_1) \\
&\quad + \frac{\alpha a_j}{n} \sum_{i \in I_{+\mu_2}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)\mu_2 + \frac{\alpha a_j}{n} \sum_{i \in I_{-\mu_2}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)(-\mu_2) \\
&\quad + \frac{\alpha a_j}{n} \sum_{i \in I_{+\mu_1}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)(x_i - \mu_1) + \frac{\alpha a_j}{n} \sum_{i \in I_{-\mu_1}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)(x_i - (-\mu_1)) \\
&\quad + \frac{\alpha a_j}{n} \sum_{i \in I_{+\mu_2}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)(x_i - \mu_2) + \frac{\alpha a_j}{n} \sum_{i \in I_{-\mu_2}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)(x_i - (-\mu_2)).
\end{aligned}$$

Define the first "error vector"

$$\begin{aligned}
\zeta_1 := &\frac{\alpha a_j}{n} \sum_{i \in I_{+\mu_1}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)(x_i - \mu_1) + \frac{\alpha a_j}{n} \sum_{i \in I_{-\mu_1}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)(x_i - (-\mu_1)) \\
&+ \frac{\alpha a_j}{n} \sum_{i \in I_{+\mu_2}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)(x_i - \mu_2) + \frac{\alpha a_j}{n} \sum_{i \in I_{-\mu_2}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)(x_i - (-\mu_2)).
\end{aligned}$$

By Lemma 5(b) that provides a bound on the deviation of x_i from its cluster center, and using that $|\ell'(t)|, |\phi'(t)| \leq 1$, we have that

$$\|\zeta_1\| \leq C_1 \alpha a_j \sigma \sqrt{d}. \tag{26}$$

Continuing from above, we get,

$$\begin{aligned}
&w_j^{(1)} - w_j^{(0)} \\
&\quad = \frac{\alpha a_j}{n} \sum_{i \in I_{+\mu_1}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)\mu_1 - \frac{\alpha a_j}{n} \sum_{i \in I_{-\mu_1}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)\mu_1 \\
&\qquad + \frac{\alpha a_j}{n} \sum_{i \in I_{+\mu_2}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)\mu_2 - \frac{\alpha a_j}{n} \sum_{i \in I_{-\mu_2}} -\ell'_{i,0} y_i \phi'(\langle w_j^{(0)}, x_i \rangle)\mu_2 + \zeta_1 \\
&\quad = \frac{\alpha a_j}{n} \sum_{i \in I_{+\mu_1}} -\ell'(0) y_i \phi'(\langle w_j^{(0)}, x_i \rangle)\mu_1 - \frac{\alpha a_j}{n} \sum_{i \in I_{-\mu_1}} -\ell'(0) y_i \phi'(\langle w_j^{(0)}, x_i \rangle)\mu_1 \\
&\qquad + \frac{\alpha a_j}{n} \sum_{i \in I_{+\mu_2}} -\ell'(0) y_i \phi'(\langle w_j^{(0)}, x_i \rangle)\mu_2 - \frac{\alpha a_j}{n} \sum_{i \in I_{-\mu_2}} -\ell'(0) y_i \phi'(\langle w_j^{(0)}, x_i \rangle)\mu_2 + \zeta_1 + \zeta_2,
\end{aligned} \tag{27}$$

where we have defined the second "error vector" ζ_2 as,

$$\begin{aligned}\zeta_2 := \frac{\alpha a_j}{n}\sum_{i\in I_{+\mu_1}}(-\ell'_{i,0}+\ell'(0))\phi'(\langle w_j^{(0)},x_i\rangle)y_i\mu_1 - \frac{\alpha a_j}{n}\sum_{i\in I_{-\mu_1}}(-\ell'_{i,0}+\ell'(0))y_i\phi'(\langle w_j^{(0)},x_i\rangle)\mu_1 \\ - \frac{\alpha a_j}{n}\sum_{i\in I_{+\mu_2}}(-\ell'_{i,0}+\ell'(0))y_i\phi'(\langle w_j^{(0)},x_i\rangle)\mu_2 + \frac{\alpha a_j}{n}\sum_{i\in I_{-\mu_2}}(-\ell'_{i,0}+\ell'(0))y_i\phi'(\langle w_j^{(0)},x_i\rangle)\mu_2.\end{aligned}$$

Applying the triangle inequality and Equation (25),

$$\begin{aligned}\|\zeta_2\| &\le \alpha a_j \max\{\|\mu_1\|,\|\mu_2\|\}\max\left\{\left(-\ell'\left(-3\omega_{\text{init}}\sqrt{md}\right)+\ell'(0)\right),\left(-\ell'\left(3\omega_{\text{init}}\sqrt{md}\right)+\ell'(0)\right)\right\} \\ &\le \alpha a_j\omega_{\text{init}}\sqrt{md}, \end{aligned} \tag{28}$$

where in the last line we have used that $-\ell'$ is $1/4$-Lipschitz and that $\|\mu_1\| = \|\mu_2\| = 1$. Define now, for $j \subset [m]$,

$$\begin{cases} N_{+\mu_1}(j) = \sum_{i\in I_{+\mu_1}} y_i\phi'(\langle w_j^{(0)},x_i\rangle), \\ N_{-\mu_1}(j) = \sum_{i\in I_{-\mu_1}} y_i\phi'(\langle w_j^{(0)},x_i\rangle), \\ N_{+\mu_2}(j) = \sum_{i\in I_{+\mu_2}} y_i\phi'(\langle w_j^{(0)},x_i\rangle), \\ N_{-\mu_2}(j) = \sum_{i\in I_{-\mu_2}} y_i\phi'(\langle w_j^{(0)},x_i\rangle). \end{cases} \tag{29}$$

Substituting the above definition into (27), we then have

$$\begin{aligned} w_j^{(1)} - w_j^{(0)} &= \frac{\alpha a_j}{n}\left[-\ell'(0)\mu_1(N_{+\mu_1}(j)-N_{-\mu_1}(j)) - \ell'(0)\mu_2(N_{+\mu_2}(j)-N_{-\mu_2}(j))\right] + \zeta_1+\zeta_2 \\ &= \frac{\alpha a_j}{2n}\left[\mu_1(N_{+\mu_1}(j)-N_{-\mu_1}(j)) + \mu_2(N_{+\mu_2}(j)-N_{-\mu_2}(j))\right] + \zeta_1+\zeta_2, \end{aligned} \tag{30}$$

where the last equality follows since $-\ell'(0) = 1/2$.

Step 2: Continuing with the plan outlined above, we will now show that $\langle w_j^{(1)}/\|w_j^{(1)}\|, \mu_1\rangle \ge c$ for a universal constant c. We have,

$$\begin{aligned}\langle w_j^{(1)} - w_j^{(0)}, \mu_1\rangle &= \frac{\alpha a_j}{2n}\left[\|\mu_1\|^2(N_{+\mu_1}(j)-N_{-\mu_1}(j)) + \langle\mu_2,\mu_1\rangle(N_{+\mu_2}(j)-N_{-\mu_2}(j))\right] \\ &\quad + \langle\zeta_1,\mu_1\rangle + \langle\zeta_2,\mu_1\rangle \\ &\ge \frac{\alpha a_j}{2n}\left[N_{+\mu_1}(j)-N_{-\mu_1}(j)\right] - \frac{\alpha a_j}{n}\left[C_1 n\sigma\sqrt{d} + n\omega_{\text{init}}\sqrt{md}\right]. \end{aligned} \tag{31}$$

In the last line we have applied the inequalities (26) and (28). Thus, it suffices to derive a lower bound for $N_{+\mu_1}(j) - N_{-\mu_1}(j)$, which is precisely the result that Lemma 6 provides.

We have,

$$
\begin{aligned}
&N_{+\mu_1}(j) - N_{-\mu_1}(j) \\
&\quad = \sum_{i \in I_{+\mu_1}} y_i \phi'(\langle w_j^{(0)}, x_i \rangle) - \sum_{i \in I_{-\mu_1}} \phi'(\langle w_j^{(0)}, x_i \rangle) \qquad (32) \\
&\quad = \sum_{i \in I^{\mathcal{C}}_{+\mu_1}} \phi'(\langle w_j^{(0)}, x_i \rangle) - \sum_{i \in I^{\mathcal{N}}_{+\mu_1}} \phi'(\langle w_j^{(0)}, x_i \rangle) - \sum_{i \in I^{\mathcal{C}}_{-\mu_1}} \phi'(\langle w_j^{(0)}, x_i \rangle) + \sum_{i \in I^{\mathcal{N}}_{-\mu_1}} \phi'(\langle w_j^{(0)}, x_i \rangle) \\
&\quad \geq \sum_{i \in I^{\mathcal{C}}_{+\mu_1}} \phi'(\langle w_j^{(0)}, x_i \rangle) - \sum_{i \in I^{\mathcal{C}}_{-\mu_1}} \phi'(\langle w_j^{(0)}, x_i \rangle) - |\mathcal{N}| \\
&\quad \overset{(i)}{\geq} n \left(\frac{1}{C_2} - \frac{|\mathcal{N}|}{n} \right) \\
&\quad \overset{(ii)}{\geq} \frac{n}{2C_2}. \qquad (33)
\end{aligned}
$$

In (i) we use Lemma 6, while in (ii) we use Assumption (A4) so that $|\mathcal{N}|/n \leq 2\eta \leq 1/2C_2$. Thus, plugging this in to (31) we get that

$$
\langle w_j^{(1)}, \mu_1 \rangle \overset{(i)}{>} \langle w_j^{(1)} - w_j^{(0)}, \mu_1 \rangle \geq \frac{\alpha a_j}{4C_2} - \alpha a_j \left[C_1 \sigma \sqrt{d} + \omega_{\text{init}} \sqrt{md} \right] \overset{(ii)}{\geq} \alpha a_j / 8C_2. \qquad (34)
$$

Inequality (i) uses that $j \in J_{+\mu_1}$ implies $\langle w_j^{(0)}, \mu_1 \rangle > 0$. Inequality (ii) follows by using Assumption (A2), so that $C_1 \sigma \sqrt{d} \leq 1/16C_2$, as well as Assumption (A6) so that for $C > 1$ sufficiently large we have $\omega_{\text{init}} \sqrt{md} \leq 1/16C_2$. Next, we can use Lemma 14 to derive a bound for the normalized margin,

$$
\left\langle \frac{w_j^{(1)}}{\|w_j^{(1)}\|}, \mu_1 \right\rangle \geq \frac{\alpha a_j / 8C_2}{2\alpha a_j} = \frac{1}{16C_2}. \qquad (35)
$$

This completes the proof for the normalized margin claim.

Step 3: To show that the first part of the neuron alignment holds, we want to show that $\langle w_j^{(1)}, x_i \rangle > 0$. We have,

$$
\begin{aligned}
\left\langle \frac{w_j^{(1)}}{\|w_j^{(1)}\|}, x_i \right\rangle &= \left\langle \frac{w_j^{(1)}}{\|w_j^{(1)}\|}, \mu_1 \right\rangle + \left\langle \frac{w_j^{(1)}}{\|w_j^{(1)}\|}, x_i - \mu_1 \right\rangle \\
&\overset{(i)}{\geq} 1/16C_2 - \|x_i - \mu_1\| \\
&\overset{(ii)}{\geq} 1/16C_2 - C_1 \sigma \sqrt{d} \\
&\overset{(iii)}{\geq} 1/32C_2 > 0. \qquad (36)
\end{aligned}
$$

Above, (i) uses (35) and the Cauchy–Schwarz inequality. Inequality (ii) uses Lemma 5. The final inequality (iii) uses Assumption (A2), so that $C_1 \sigma \sqrt{d} \leq 1/64C_2$. This completes the part of neuron alignment concerning neurons $J_{+\mu_1}$ and for samples in cluster $I_{+\mu_1}$.

Part 2 of NAC: To show the part of neuron alignment concerning samples in cluster $I_{-\mu_1}$, note that we still have the identity (35). But for samples $i \in I_{-\mu_1}$, we have

$$\langle w_j^{(1)}, x_i \rangle = \langle w_j^{(1)}, -\mu_1 \rangle + \langle w_j^{(1)}, x_i + \mu_1 \rangle,$$

where $\|x_i + \mu_1\|$ is small, and so the inequality $\langle w_j^{(1)}, x_i \rangle < 0$ follows using the same argument as above. Hence, we have shown that $\phi'(\langle w_j^{(1)}, x_i \rangle)$ for all $i \in I_{-\mu_1}$.

This completes the proof of neuron alignment for the neurons in $J_{+\mu_1}$. An analogous argument can also be used to establish the claim for the neurons in $J_{-\mu_1} \cup J_{\pm\mu_2}$. ■

A.3.2 Proof of Lemma 13

We now show that the neuron alignment condition and almost-orthogonality condition hold for a sufficiently large amount of time.

Lemma 13 *For $C > 1$ sufficiently large, on a good run, for every time $t = 1, \ldots, 1/(4\alpha)$, neuron alignment (Condition 9) holds at time t and almost-orthogonality (Condition 10) holds up to time t.*

Proof The proof is by induction. To see the base case $t = 1$, first, note that neuron alignment holds at time $t = 1$ by Lemma 12. Further, almost-orthogonality holds at time $t = 1$ since by Lemma 14 we have $|\langle w_j^{(1)}, \mu \rangle| \leq \|w_j^{(1)}\| \leq 2|a_j|\alpha t$ for any $\mu \in \{\pm\mu_1, \pm\mu_2\}$. So let us now assume that neuron alignment and almost-orthogonality hold at every time step until time t, and consider the case $t + 1 \leq 1/(4\alpha)$. By Lemma 14, since $t + 1 \leq 1/(4\alpha)$, we have $\|W^{(\tau)}\|_F \leq 1$ for every $\tau \leq t + 1$. Using an identical argument to (19), this implies for all $i \in [n]$ and $\tau \leq \{1, \ldots, 1/(4\alpha)\}$,

$$-\ell'_{i,\tau} := -\ell'(y_i f(x_i; W^{(\tau)})) \geq \frac{1}{2}\exp(-2). \tag{37}$$

This key property will allow us to show that neuron alignment holds at time $t + 1$.

Neuron alignment holds at time $t + 1$. We will first show the result for neurons $j \in J_{+\mu_1}$; the result for neurons in $J_{-\mu_1} \cup J_{\pm\mu_2}$ will follow similarly.

Let $j \in J_{+\mu_1}$, so $a_j = |a_j| = 1/\sqrt{m}$. It suffices to show that for $k \in I_{+\mu_1}$, we have $\langle w_j^{(t+1)}, x_k \rangle > 0$, and for $k \in I_{-\mu_1}$, we have $\langle w_j^{(t+1)}, x_k \rangle < 0$. To show this, we will utilize an argument similar to that we used in the proof of Lemma 12 (see eqs. (35) and (36)), in that we will first show that $\langle w_j^{(t+1)}/\|w_j^{(t+1)}\|, +\mu_1 \rangle \geq c$ for some constant $c > 0$, and then use that the within-cluster variance is of order $\sigma^2 d$ and that $\sigma^2 \ll 1/d$. Towards this end, we first derive a consequence of neuron alignment. Let τ be a time satisfying $1 \leq \tau \leq t$. Then

neuron alignment holds at time τ by the induction hypothesis, so that,

$$\begin{aligned}
&\langle w_j^{(\tau+1)} - w_j^{(\tau)}, +\mu_1\rangle \\
&= \frac{\alpha a_j}{n}\sum_{i=1}^{n} -\ell'_{i,\tau}\phi'(\langle w_j^{(\tau)}, x_i\rangle)\langle y_i x_i, \mu_1\rangle \\
&= \frac{\alpha|a_j|}{n}\sum_{i\in I^{\mathcal{C}}_{+\mu_1}} -\ell'_{i,\tau}\phi'(\langle w_j^{(\tau)}, x_i\rangle)\langle x_i, \mu_1\rangle - \frac{\alpha|a_j|}{n}\sum_{i\in I^{\mathcal{N}}_{+\mu_1}} -\ell'_{i,\tau}\phi'(\langle w_j^{(\tau)}, x_i\rangle)\langle x_i, \mu_1\rangle \\
&\quad + \frac{\alpha|a_j|}{n}\sum_{i\in I^{\mathcal{C}}_{-\mu_1}} -\ell'_{i,\tau}\phi'(\langle w_j^{(\tau)}, x_i\rangle)\langle x_i, \mu_1\rangle - \frac{\alpha|a_j|}{n}\sum_{i\in I^{\mathcal{N}}_{-\mu_1}} -\ell'_{i,\tau}\phi'(\langle w_j^{(\tau)}, x_i\rangle)\langle x_i, \mu_1\rangle \\
&\quad + \frac{\alpha|a_j|}{n}\sum_{i\in I_{+\mu_2}} -\ell'_{i,\tau}y_i\phi'(\langle w_j^{(\tau)}, x_i\rangle)\langle x_i, \mu_1\rangle + \frac{\alpha|a_j|}{n}\sum_{i\in I_{-\mu_2}} -\ell'_{i,\tau}y_i\phi'(\langle w_j^{(\tau)}, x_i\rangle)\langle x_i, \mu_1\rangle \\
&\overset{(i)}{=} \frac{\alpha|a_j|}{n}\sum_{i\in I^{\mathcal{C}}_{+\mu_1}} -\ell'_{i,\tau}\langle x_i, \mu_1\rangle - \frac{\alpha|a_j|}{n}\sum_{i\in I^{\mathcal{N}}_{+\mu_1}} -\ell'_{i,\tau}\langle x_i, \mu_1\rangle \\
&\quad + \frac{\alpha|a_j|}{n}\sum_{i\in I_{+\mu_2}} -\ell'_{i,\tau}y_i\phi'(\langle w_j^{(\tau)}, x_i\rangle)\langle x_i, \mu_1\rangle + \frac{\alpha|a_j|}{n}\sum_{i\in I_{-\mu_2}} -\ell'_{i,\tau}y_i\phi'(\langle w_j^{(\tau)}, x_i\rangle)\langle x_i, \mu_1\rangle.
\end{aligned} \tag{38}$$

In (i) we have used that the neuron alignment condition holds at time τ, and thus $\phi'(\langle w_j^{(\tau)}, x_i\rangle) = 1$ for $i \in I_{+\mu_1}$ and $\phi'(\langle w_j^{(\tau)}, x_i\rangle) = 0$ for $i \in I_{-\mu_1}$. We can bound the terms $\langle x_i, +\mu_1\rangle$ appearing above with Lemma 5, so that

$$\begin{aligned}
\langle w_j^{(\tau+1)} - w_j^{(\tau)}, +\mu_1\rangle &\overset{(i)}{\geq} \frac{\alpha|a_j|}{n}\left[\sum_{i\in I^{\mathcal{C}}_{+\mu_1}} -\ell'_{i,\tau}\left[1 - C_1\sigma\sqrt{d}\right] - 2|\mathcal{N}| - 2\sum_{i\in I_{\pm\mu_2}} -\ell'_{i,\tau}C_1\sigma\sqrt{d}\right] \\
&\overset{(ii)}{\geq} \frac{\alpha|a_j|}{n}\left[\sum_{i\in I^{\mathcal{C}}_{+\mu_1}} \frac{1}{4}\exp(-2) - 2|\mathcal{N}| - 2C_1|I^{\mathcal{C}}_{+\mu_1} \cup I_{\pm\mu_2}|\sigma\sqrt{d}\right] \\
&\overset{(iii)}{\geq} \frac{\alpha|a_j|}{n}\left[\frac{n}{8}\cdot\frac{1}{4}\exp(-2) - 2|\mathcal{N}| - 2C_1 n\sigma\sqrt{d}\right] \\
&\overset{(iv)}{\geq} \frac{\alpha|a_j|\exp(-2)}{64}.
\end{aligned} \tag{39}$$

In (i) we use that $|\ell'| \leq 1$ and Lemma 5, so that $\langle x_i, \mu_1\rangle \geq 1 - C_1\sigma\sqrt{d}$ for $i \in I_{+\mu_1}$, $|\langle x_i, \mu_1\rangle| \leq 2$ for $i \in I_{-\mu_1}$, and $|\langle x_i, \mu_1\rangle| \leq C_1\sigma\sqrt{d}$ for $i \in I_{\pm\mu_2}$. In inequality (ii), we use (37) as well as the fact that Assumption (A2) implies $C_1\sigma\sqrt{d} \leq 1/2$. In (iii) we use parts (c) and (d) of Lemma 5 and Assumption (A3) so that $|I^{\mathcal{C}}_{+\mu_1}| \geq |I_{+\mu_1}| - |\mathcal{N}| \geq n/8$. The final line (iv) follows by using Assumptions (A2) and (A4), so that $2|\mathcal{N}|/n \leq \exp(-2)/128$ and $2C_1\sigma\sqrt{d} \leq \exp(-2)/128$ as well. We have thus shown that if neuron alignment holds at time τ, then for $j \in J_{+\mu}$ we have $\langle w_j^{(\tau+1)} - w_j^{(\tau)}, +\mu_1\rangle \geq \alpha|a_j|\exp(-2)/64$. Telescoping this

inequality from times $\tau = 1, \ldots, t$, we get

$$\langle w_j^{(t+1)}, +\mu_1 \rangle \geq \langle w_j^{(1)}, +\mu_1 \rangle + \frac{\alpha |a_j| t \exp(-2)}{64} \overset{(i)}{\geq} \frac{\alpha |a_j| t \exp(-2)}{64},$$

where inequality (i) uses Lemma 12. By Lemma 14, we have $\|w_j^{(t+1)}\| \leq 2\alpha |a_j|(t+1)$, so that,

$$\left\langle \frac{w_j^{(t+1)}}{\|w_j^{(t+1)}\|}, +\mu_1 \right\rangle \geq \frac{\alpha |a_j| t \exp(-2)}{128 \alpha |a_j| (t+1)} \geq \frac{\exp(-2)}{256}. \tag{40}$$

Using an identical argument to (36), since by Lemma 5(b) and Assumption (A2) we have the inequalities $\|x_k - \mu_1\| \leq C_1 \sigma \sqrt{d} \leq C_1 / C$, by taking $C > 512 C_1 \exp(2)$ we have $\langle w_j^{(t+1)}, x_k \rangle > 0$ for $k \in I_{+\mu_1}$. A symmetric argument shows that $\langle w_j^{(t+1)}, x_k \rangle < 0$ for $k \in I_{-\mu_1}$. This completes the proof that neuron alignment holds for neurons $j \in J_{+\mu_1}$. We can show that neuron alignment holds for neurons in $J_{-\mu_1} \cup J_{\pm \mu_2}$ using an analogous argument.

Almost-orthogonality holds at time $t+1$. We now show that almost-orthogonality continues to hold at time $t+1$ given it holds at time t. We will prove the result for neurons $j \in J_{+\mu_1}$ with an analogous argument holding for the neurons in $J_{-\mu_1} \cup J_{\pm \mu_2}$.

We want to show that, for any neuron $j \in J_{+\mu_1}$ satisfying

$$|\langle w_j^{(t)}, \mu_2 \rangle| \leq 3\alpha |a_j|,$$

we have that $|\langle w_j^{(t+1)}, \mu_2 \rangle| \leq 3\alpha |a_j|$ as well. We will show this by demonstrating that if at time t we have $|\langle w_j^{(t)}, \mu_2 \rangle| \geq \alpha |a_j|$, then $\langle w_j^{(t+1)}, \mu_2 \rangle$ will either change sign or will decrease in magnitude at the next iteration; since the order of norm changes for a single neuron in one step is $O(\alpha |a_j|)$, this will complete the proof.

Consider the case that $\langle w_j^{(t)}, \mu_2 \rangle \geq \alpha |a_j|$; the negative case will follow using a symmetric argument. Since neuron alignment holds, an identical argument used to derive Equations (38)

through (39) implies that

$$
\begin{aligned}
&\langle w_j^{(t+1)} - w_j^{(t)}, \mu_2 \rangle \\
&= \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{C}}_{+\mu_1}} -\ell'_{i,t} \langle x_i, \mu_2 \rangle - \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{N}}_{+\mu_1}} -\ell'_{i,t} \langle x_i, \mu_2 \rangle \\
&\quad + \frac{\alpha|a_j|}{n} \sum_{i \in I_{+\mu_2}} -\ell'_{i,t} y_i \phi'(\langle w_j^{(t)}, x_i \rangle) \langle x_i, \mu_2 \rangle + \frac{\alpha|a_j|}{n} \sum_{i \in I_{-\mu_2}} -\ell'_{i,t} y_i \phi'(\langle w_j^{(t)}, x_i \rangle) \langle x_i, \mu_2 \rangle \\
&\overset{(i)}{\leq} 2C_1 \alpha |a_j| \sigma \sqrt{d} - \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{C}}_{+\mu_2}} -\ell'_{i,t} \phi'(\langle w_j^{(t)}, x_i \rangle) \langle x_i, \mu_2 \rangle + \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{N}}_{+\mu_2}} -\ell'_{i,t} \phi'(\langle w_j^{(t)}, x_i \rangle) \langle x_i, \mu_2 \rangle \\
&\quad - \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{C}}_{-\mu_2}} -\ell'_{i,t} \phi'(\langle w_j^{(t)}, x_i \rangle) \langle x_i, \mu_2 \rangle + \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{N}}_{-\mu_2}} -\ell'_{i,t} \phi'(\langle w_j^{(t)}, x_i \rangle) \langle x_i, \mu_2 \rangle \\
&\overset{(ii)}{\leq} 2C_1 \alpha |a_j| \sigma \sqrt{d} + \frac{2\alpha|a_j||\mathcal{N}|}{n} \\
&\quad - \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{C}}_{+\mu_2}} -\ell'_{i,t} \phi'(\langle w_j^{(t)}, x_i \rangle) \langle x_i, \mu_2 \rangle - \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{C}}_{-\mu_2}} -\ell'_{i,t} \phi'(\langle w_j^{(t)}, x_i \rangle) \langle x_i, \mu_2 \rangle \\
&\overset{(iii)}{\leq} 2C_1 \alpha |a_j| \sigma \sqrt{d} + \frac{2\alpha|a_j||\mathcal{N}|}{n} \\
&\quad - \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{C}}_{+\mu_2}} \frac{1}{2} \exp(-2) \phi'(\langle w_j^{(t)}, x_i \rangle) \cdot \frac{1}{2} + \frac{3}{2} \cdot \frac{\alpha|a_j|}{n} \sum_{i \in I^{\mathcal{C}}_{-\mu_2}} \phi'(\langle w_j^{(t)}, x_i \rangle) \\
&= \alpha |a_j| \Bigg[2C_1 \sigma \sqrt{d} + 2\frac{|\mathcal{N}|}{n} \\
&\quad - \frac{\exp(-2)}{4n} \Bigg(\sum_{i \in I^{\mathcal{C}}_{+\mu_2}} \phi'(\langle w_j^{(t)}, x_i \rangle) - 6\exp(2) \cdot \sum_{i \in I^{\mathcal{C}}_{-\mu_2}} \phi'(\langle w_j^{(t)}, x_i \rangle) \Bigg) \Bigg]. \qquad (41)
\end{aligned}
$$

In (i) we have used Lemma 5, so that $|\langle x_i, \mu_2 \rangle| \leq C_1 \sigma \sqrt{d}$ when $i \in I_{+\mu_1}$. In inequality (ii), we use that Lemma 5 implies $|\langle x_i, \mu_2 \rangle| \leq 2$ for $i \in I_{\pm \mu_2}$, so that by the 1-Lipschitz property of ℓ and ϕ we have,

$$
\left| \sum_{i \in I^{\mathcal{N}}_{\pm \mu_2}} -\ell'_{i,t} y_i \phi'(\langle w_j^{(t)}, x_i \rangle) \langle x_i, \mu_2 \rangle \right| \leq 2|\mathcal{N}|.
$$

In inequality (iii), we have used that ℓ is 1-Lipschitz as well as Lemma 5 so that $|\langle x_i, \mu_2 \rangle| \leq 3/2$ for $i \in I_{-\mu_2}$.

From the above, one can see that if $\sum_{i \in I^{\mathcal{C}}_{+\mu_2}} \phi'(\langle w_j^{(t)}, x_i \rangle) \gg \sum_{i \in I^{\mathcal{C}}_{-\mu_2}} \phi'(\langle w_j^{(t)}, x_i \rangle)$, then we will have that the above quantity is negative, showing that $\langle w_j^{(t)}, \mu_2 \rangle$ will decrease. When $\langle w_j^{(t)}, \mu_2 \rangle$ is large, then this is likely to occur; this is precisely the second part of Lemma 6.

In particular, since $\langle w_j^{(t)}, \mu_2 \rangle \geq \alpha |a_j|$ by assumption, we have

$$\left\langle \frac{w_j^{(t)}}{\|w_j^{(t)}\|}, \mu_2 \right\rangle \overset{(i)}{\geq} \frac{\alpha |a_j|}{3\alpha |a_j| t} \overset{(ii)}{\geq} \frac{4}{3}\alpha \overset{(iii)}{\geq} \frac{1}{2\sqrt{C}}, \tag{42}$$

where (i) follows by Lemma 14 and the fact that we are considering the case when $\langle w_j^{(t)}, \mu_2 \rangle \geq \alpha |a_j|$; inequality (ii) uses that $t \leq 1/(4\alpha)$; and (iii) uses Assumption (A7), so that $\alpha \geq 1/(2\sqrt{C})$. Since the correlation with the cluster mean is of constant order, we can repeat the argument used in (36) to show that the sign of $\langle w_j^{(t)}, x_i \rangle$ is the same as the sign of $\langle w_j^{(t)}, \mu_2 \rangle$ for $i \in I^{\mathcal{C}}_{+\mu_2}$:

$$\begin{aligned}
\langle w_j^{(t)}/\|w_j^{(t)}\|, x_i \rangle &\geq \langle w_j^{(t)}/\|w_j^{(t)}\|, \mu_2 \rangle - \|x_i - \mu_2\| \\
&\overset{(i)}{\geq} \frac{1}{2\sqrt{C}} - C_1 \sigma \sqrt{d} \\
&\overset{(ii)}{>} 0.
\end{aligned}$$

Inequality (i) uses the lower bound in (42) as well as Lemma 5. Inequality (ii) uses assumption (A2), so that $C_1 \sigma \sqrt{d} \leq C_1/C < 1/(2\sqrt{C})$ for C sufficiently large relative to C_1. Using a symmetric argument, we thus have for positive neurons satisfying $\langle w_j^{(t)}, \mu_2 \rangle \geq \alpha |a_j|$,

$$\text{for every } i \in I^{\mathcal{C}}_{+\mu_2}, \quad \phi'(\langle w_j^{(t)}, x_i \rangle) = 1, \quad \text{while for } i \in I^{\mathcal{C}}_{-\mu_2}, \quad \phi'(\langle w_j^{(t)}, x_i \rangle) = 0. \tag{43}$$

Substituting the above into (41), we get,

$$\begin{aligned}
\langle w_j^{(t+1)} - w_j^{(t)}, \mu_2 \rangle &\leq \alpha |a_j| \Bigg[2C_1 \sigma \sqrt{d} + 2\frac{|\mathcal{N}|}{n} \\
&\qquad - \frac{\exp(-2)}{4n} \Bigg(\sum_{i \in I^{\mathcal{C}}_{+\mu_2}} \phi'(\langle w_j^{(t)}, x_i \rangle) - 6\exp(2) \cdot \sum_{i \in I^{\mathcal{C}}_{-\mu_2}} \phi'(\langle w_j^{(t)}, x_i \rangle) \Bigg) \Bigg] \\
&\overset{(i)}{\leq} \alpha |a_j| \left[2C_1 \sigma \sqrt{d} + 2\frac{|\mathcal{N}|}{n} - \frac{\exp(-2)}{4n} |I^{\mathcal{C}}_{+\mu_2}| \right] \\
&\overset{(ii)}{\leq} \alpha |a_j| \left[2C_1 \sigma \sqrt{d} + 3\frac{|\mathcal{N}|}{n} - \frac{\exp(-2)}{4n} |I_{+\mu_2}| \right] \\
&\overset{(iii)}{\leq} \alpha |a_j| \left[2C_1 \sigma \sqrt{d} + 3\frac{|\mathcal{N}|}{n} - \frac{\exp(-2)}{32} \right] \\
&\overset{(iv)}{<} 0.
\end{aligned} \tag{44}$$

The inequality (i) uses eq. (43). Inequality (ii) uses that $|I^{\mathcal{C}}_{+\mu_2}| \geq |I_{+\mu_2}| - |I^{\mathcal{N}}_{+\mu_2}| \geq |I_{+\mu_2}| - |\mathcal{N}|$. Inequality (iii) uses the lower bound on the number of points in cluster μ_2 given in Lemma 5 together with Assumption (A3). The final inequality follows by using Assumption (A2) and

Lemma 5, which allow for us to take $\sigma\sqrt{d}$ and $|\mathcal{N}|/n$ smaller than an absolute constant. This shows that, in the case that $\langle w_j^{(t)}, \mu_2\rangle \geq \alpha|a_j|$, the value of $\langle w_j^{(t+1)}, \mu_2\rangle$ is strictly less than $\langle w_j^{(t)}, \mu_2\rangle$. Since by Lemma 5 we have $\|x_i\| \leq \sqrt{2}$, we have,

$$|\langle w_j^{(t+1)} - w_j^{(t)}, \mu_2\rangle| = \alpha|a_j|\left|\frac{1}{n}\sum_{i=1}^{n} -\ell'_{i,t} y_i \phi'(\langle w_j^{(t)}, x_i\rangle)\langle x_i, \mu_2\rangle\right| \leq 2\alpha|a_j|. \tag{45}$$

As we have shown $\langle w_j^{(t)}, \mu_2\rangle \geq \alpha|a_j|$, this implies $\langle w_j^{(t+1)}, \mu_2\rangle \in [(1-2)\alpha|a_j|, \alpha|a_j|)$, and thus the inequality $|\langle w_j^{(t+1)}, \mu_2\rangle| \leq 3\alpha|a_j|$ holds as desired. This completes the induction in the case that $\langle w_j^{(t)}, \mu_2\rangle \geq \alpha|a_j|$.

For the case $\langle w_j^{(t)}, \mu_2\rangle \leq -\alpha|a_j|$, we can use a nearly identical argument as above to show that
$\langle w_j^{(t+1)} - w_j^{(t)}, \mu_2\rangle > 0$ so that $\langle w_j^{(t+1)}, \mu_2\rangle \in (-\alpha|a_j|, (-1+2)\alpha|a_j|]$. This again gives $|\langle w^{(t+1)}, \mu_2\rangle| \leq 3\alpha|a_j|$.

The only remaining case is when $|\langle w_j^{(t)}, \mu_2\rangle| \leq \alpha|a_j|$. In this case, (45) implies that we have the inequality $|\langle w_j^{(t+1)}, \mu_2\rangle| \leq (1+2)\alpha|a_j| = 3\alpha|a_j|$, completing the induction for the $J_{+\mu_1}$ neurons. The proof that almost-orthogonality holds for neurons $j \in J_{-\mu_1} \cup J_{\pm\mu_2}$ holds using an analogous argument.

■

A.4 Proof of Theorem 1

For the reader's convenience, we restate the theorem below before completing its proof.

Theorem 1 *Let $\delta \in (0, 1/2)$. For all $C > 1$ sufficiently large, under the assumptions (A1) through (A7), by running gradient descent with step-size α for $T = 1 + 1/(4\alpha)$ iterations, with probability at least $1 - 4\delta$ over the random initialization and the draws of the samples we have,*

1. *For the training points:*

$$\begin{aligned} \text{for all } i \in \mathcal{C}, \quad & y_i = \operatorname{sgn}\left(f(x_i; W^{(T)})\right), \\ \text{while for all } i \in \mathcal{N}, \quad & y_i \neq \operatorname{sgn}\left(f(x_i; W^{(T)})\right). \end{aligned}$$

2. *Further, the test error satisfies*

$$\mathbb{P}_{(x,y)\sim \mathsf{P}}\big(y \neq \operatorname{sgn}(f(x; W^{(T)}))\big) \leq \eta + C\sqrt{\frac{\log(1/\delta)}{n}}.$$

Proof First, note that with probability at least $1 - 4\delta$, a good run occurs, so that the results of Lemma 3, Lemma 4, Lemma 5, and Lemma 6 all hold for the absolute constant $C_0 = 4^5 \cdot 1024^2 \exp(4)$. We thus can apply Lemma 13 so that neuron alignment and

almost-orthogonality hold for times $t = 1, \ldots, 1/4\alpha$. Since neuron alignment and almost-orthogonality hold, by Lemma 11, we have,

$$y_i f^J(x_i; W^{(T)}) \geq \frac{\exp(-2)}{4 \cdot 1024}(1 - 1/C_0)^2 \quad \text{for all } i \in \mathcal{C}, \quad \text{and}$$
$$y_i f^J(x_i; W^{(T)}) \leq -\frac{\exp(-2)}{4 \cdot 1024}(1 - 1/C_0)^2 \quad \text{for all } i \in \mathcal{N}. \tag{46}$$

In order to apply Lemma 8, which relates the prediction on the subnetwork to the entire network, we need to ensure that for $C_f = 4 \cdot 1024 \exp(2)/(1-1/C_0)^2$ we have $|J^c|/m \leq 1/16C_f^2$. If we denote by $J^c = [m] \setminus (J_{\pm\mu_1} \cup J_{\pm\mu_2})$, then $|J^c|/m \leq 1 - (1 - 1/C_0)^2 \leq 2/C_0$, so that,

$$\frac{|J^c|}{m} \leq \frac{2}{C_0} \overset{(i)}{=} \frac{\exp(-4)}{2 \cdot 16^2 \cdot 1024^2} \leq \frac{\exp(-4)}{16^2 \cdot 1024^2}\left(1 - \frac{1}{C_0}\right)^2 = \frac{1}{16C_f^2}.$$

The equality (i) follows since $C_0 = 4^5 \cdot 1024^2 \exp(4)$. Thus we may apply Lemma 8. Since $\|W^{(T)}\|_F \leq 1$ by Lemma 11, and since $\|x_i\| \leq 2$ by Lemma 5, the lower bound for clean samples given in (46) can be used in Lemma 8 to get,

$$\text{for all } i \in \mathcal{C}, \quad y_i f(x_i; W^{(T)}) \geq \frac{\exp(-2)}{16 \cdot 1024} =: \gamma > 0. \tag{47}$$

Using a symmetric argument, we have that noisy samples satisfy

$$\text{for all } i \in \mathcal{N}, \quad y_i f(x_i; W^{(T)}) \leq -\frac{\exp(-2)}{16 \cdot 1024} = -\gamma < 0.$$

This shows that the neural network accurately classifiers all of the clean samples correctly at a margin of $\gamma > 0$, and misclassifies all noisy samples incorrectly. Since we have the Frobenius norm bound $\|W^{(T)}\|_F \leq 1$, we can therefore use a simple Rademacher complexity-based argument to derive a generalization bound for the neural network. In particular, let us define the ramp loss

$$r_\gamma(z) := \min(1, \max(0, 1 - z/\gamma)).$$

Then r_γ is γ^{-1}-Lipschitz, and if we denote by

$$\mathcal{F} := \{x \mapsto f(x; W) : \|W\|_F \leq 1\}$$

as the class of two-layer ReLU networks with Frobenius norm at most 1, the expected Rademacher complexity (Shalev-Shwartz and Ben-David, 2014, Lemma 26.9) of the hypothesis class induced by the composition of r_γ with the class of two-layer ReLU networks with Frobenius norm at most 1 satisfies

$$\mathfrak{R}(r_\gamma \circ \mathcal{F}) \leq \gamma^{-1}\mathfrak{R}(\mathcal{F}).$$

Since $\mathbb{E}_{(x,y)\sim\mathsf{P}}[\|x\|^2] \leq 2$, a standard bound on the Rademacher complexity of two-layer ReLU networks (see Proposition 15) therefore implies

$$\mathfrak{R}(r_\gamma \circ \mathcal{F}) \leq \frac{4\gamma^{-1}}{\sqrt{n}}.$$

Finally, note that by (47), we have that the empirical risk under the ramp loss r_γ is at most the risk under the zero-one loss,

$$\frac{1}{n}\sum_{i=1}^{n} r_\gamma(y_i f(x_i; W^{(t)})) \leq \frac{|\mathcal{N}|}{n}. \tag{48}$$

Standard Rademacher complexity generalization bounds (e.g. Shalev-Shwartz and Ben-David (2014, Theorem 26.5)) thus imply

$$\begin{aligned} P\big(y \neq \operatorname{sgn}(f(x; W^{(T)}))\big) &\leq \mathbb{E}[r_\gamma(y f(x; W^{(T)})] \\ &\leq \frac{1}{n}\sum_{i=1}^{n} r_\gamma(y_i f(x_i; W^{(t)})) + \frac{4\gamma^{-1}}{\sqrt{n}} + \sqrt{\frac{2\log(4/\delta)}{n}} \\ &\leq \frac{|\mathcal{N}|}{n} + \frac{4\gamma^{-1}}{\sqrt{n}} + \sqrt{\frac{2\log(4/\delta)}{n}} \\ &\leq \eta + \frac{\sqrt{2C\log(2T/\delta)} + 4\gamma^{-1} + \sqrt{2\log(4/\delta)}}{\sqrt{n}}. \end{aligned}$$

In the last inequality, we have used that $\|W^{(T)}\|_F \leq 1$ and that part (c) of Lemma 5 implies $|\mathcal{N}|/n \leq \eta + \sqrt{2C\log(1/\delta)/n}$. Since $T = 1/(4\alpha) + 1$ and $\alpha \geq 1/(2\sqrt{C})$, this completes the proof. ■

Appendix B. Rademacher Complexity Bound

Below, we provide a characterization of the Rademacher complexity of the class of one-hidden-layer ReLU networks with weights that have a bounded Frobenius norm.

Proposition 15 *Let $R > 0$ be arbitrary, and let $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{Uniform}(\{+1, -1\})$ be independent Rademacher random variables, and let $s \in \mathbb{R}^n$ denote the vector of Rademacher variables. Consider $\widehat{\mathfrak{R}}(\mathcal{F}_R)$, the empirical Rademacher complexity (Bartlett and Mendelson, 2003) of the function class*

$$\mathcal{F}_R := \{x \mapsto f(x; W) : \|W\|_F \leq R\},$$

defined by

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_R) := \mathbb{E}_{s\sim\mathsf{Uniform}(\{+1,-1\})^n}\left[\sup_{\|W\|_F\leq R} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i; W)\right].$$

Then, for $a_j \overset{\text{i.i.d.}}{\sim} \mathsf{Uniform}(\{1/\sqrt{m}, -1/\sqrt{m}\})$, we have

$$\mathfrak{R}(\mathcal{F}_R) := \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}^n}\widehat{\mathfrak{R}}(\mathcal{F}_R) \leq \frac{2R\sqrt{\mathbb{E}_x\left[\|x\|^2\right]}}{\sqrt{n}}.$$

Proof We mimic the proof given in (Ma, 2017, Lecture 8). We have

$$\begin{aligned}
\widehat{\mathfrak{R}}(\mathcal{F}_R) &= \frac{1}{n}\mathbb{E}_{\varepsilon_i}\left[\sup_{\|W\|_F\le R}\sum_{i=1}^n \varepsilon_i f(x_i;W)\right] \\
&= \frac{1}{n}\mathbb{E}_{\varepsilon_i}\left[\sup_{\|W\|_F\le R}\sum_{i=1}^n \varepsilon_i \sum_{j=1}^n a_j\phi(\langle w_j, x_i\rangle)\right] \\
&\overset{(i)}{=} \frac{1}{n}\mathbb{E}_{\varepsilon_i}\left[\sup_{\|W\|_F\le R}\sum_{j=1}^m a_j\|w_j\|_2 \sum_{i=1}^n \varepsilon_i\phi(\langle w_j/\|w_j\|_2, x_i\rangle)\right] \\
&\le \frac{1}{n}\mathbb{E}_{\varepsilon_i}\left[\left(\sup_{\|W\|_F\le R}\sum_{j=1}^m |a_j|\|w_j\|_2\right)\max_{j\in[m]}\left|\sum_{i=1}^n \varepsilon_i\phi(\langle w_j/\|w_j\|_2, x_i\rangle)\right|\right] \\
&\overset{(ii)}{\le} \frac{R}{n}\mathbb{E}_{\varepsilon_i}\left[\max_{j\in[m]}\left|\sum_{i=1}^n \varepsilon_i\phi(\langle w_j/\|w_j\|_2, x_i\rangle)\right|\right] \\
&\le \frac{R}{n}\mathbb{E}_{\varepsilon_i}\left[\sup_{\|\bar w\|\le 1}\left|\sum_{i=1}^n \varepsilon_i\phi(\langle \bar w, x_i\rangle)\right|\right]. \qquad (49)
\end{aligned}$$

In (i) we use the homogeneity of the ReLU activation, and in (ii) we use the Cauchy–Schwarz inequality to get that

$$\sum_{j=1}^m |a_j|\|w_j\|_2 = \frac{1}{\sqrt{m}}\sum_{j=1}^m \|w_j\|_2 \le \frac{1}{\sqrt{m}}\cdot\sqrt{m}\sqrt{\sum_{j=1}^m \|w_j\|^2} = \|W\|_F.$$

From (49), since ϕ is 1-Lipschitz and the zero function is included in the class $\{x\mapsto\phi(\langle\bar w, x\rangle) : \|\bar w\|\le 1\}$, a symmetrization argument yields (Ma, 2017, Lecture 5)

$$\widehat{\mathfrak{R}}(\mathcal{F}_R) \le \frac{2R}{n}\mathbb{E}_{\varepsilon_i}\left[\sup_{\|\bar w\|\le 1}\sum_{i=1}^n \varepsilon_i\phi(\langle\bar w, x_i\rangle)\right] = 2R\cdot\widehat{\mathfrak{R}}\left(\{x\mapsto\phi(\langle\bar w, x\rangle) : \|\bar w\|\le 1\}\right).$$

Finally, as ϕ is 1-Lipschitz, the contraction property of the Rademacher complexity and standard Rademacher complexity bounds for linear hypothesis classes (Shalev-Shwartz and Ben-David, 2014, Lemma 26.10) yields the desired bound. ■

Appendix C. Proof of Proposition 2

We restate and prove Proposition 2 below.

Proposition 2 *Under the settings of Theorem 1, with probability at least $1-4\delta$ over the random initialization and draws of the samples, the feature maps of the neural network at time $T = 1 + 1/(4\alpha)$ satisfy, for all $i\in[n]$,*

$$\frac{\|\phi(W^{(T)}x_i) - \phi(W^{(0)}x_i)\|}{\|\phi(W^{(0)}x_i)\|} \ge \frac{1}{C\omega_{\text{init}}\sqrt{md}} \ge \frac{1}{C}.$$

In particular, as $\omega_{\text{init}}\sqrt{md} \to 0$, the relative change in each sample's feature map is unbounded.

Proof For simplicity, let us denote x_i by the short-hand x. Since the j-th component ($j \in [m]$) of $\phi(Wx)$ is given by $\phi(\langle w_j, x\rangle)$, we have,

$$\|\phi(W^{(T)}x) - \phi(W^{(0)}x)\|^2 = \sum_{j=1}^{m}[\phi(\langle w_j^{(T)}, x\rangle) - \phi(\langle w_j^{(0)}, x\rangle)]^2.$$

To show that the feature map moves significantly, it therefore suffices to derive a lower bound on $|\phi(\langle w_j^{(T)}, x\rangle) - \phi(\langle w_j^{(0)}, x\rangle)|$ for each j. To do so, we will show that for each sample x, a significant number of neurons have large, positive activations, so that $\langle w_j^{(T)}, x\rangle \gg 0$, while the near-zero initialization allows for us to essentially ignore the $\phi(\langle w_j^{(0)}, x\rangle)$ term.

Since neuron alignment holds at times $t = 1, \ldots, T-1$, an identical argument to that of (21) shows that for any $\mu \in \{\pm\mu_1, \pm\mu_2\}$ and $j \in J_\mu$, we have,

$$\langle w_j^{(T)} - w_j^{(1)}, \mu\rangle \geq \frac{\alpha|a_j|(T-1)}{64}\exp(-2) = \frac{|a_j|\exp(-2)}{256}.$$

Moreover, using Equation (34) we also have that $\langle w_j^{(1)} - w_j^{(0)}, \mu\rangle > 0$. Adding this inequality to the preceding display, we get,

$$\begin{aligned}
&\text{for each } j \in J_{+\mu_1}, \quad \langle w_j^{(T)} - w_j^{(0)}, +\mu_1\rangle \geq \frac{|a_j|\exp(-2)}{256},\\
&\text{for each } j \in J_{-\mu_1}, \quad \langle w_j^{(T)} - w_j^{(0)}, -\mu_1\rangle \geq \frac{|a_j|\exp(-2)}{256},\\
&\text{for each } j \in J_{+\mu_2}, \quad \langle w_j^{(T)} - w_j^{(0)}, +\mu_2\rangle \geq \frac{|a_j|\exp(-2)}{256},\\
&\text{for each } j \in J_{-\mu_2}, \quad \langle w_j^{(T)} - w_j^{(0)}, -\mu_2\rangle \geq \frac{|a_j|\exp(-2)}{256}.
\end{aligned} \tag{50}$$

Following an identical calculation used in the proof of Lemma 14 (see Eq. (17)), we know that $\|w_j^{(T)} - w_j^{(0)}\| \leq \sqrt{2}|a_j|\alpha T = \sqrt{2}|a_j|(\alpha+4)$. Since $\alpha \leq 1/10$ we thus have,

$$\text{for every } \mu \in \{\pm\mu_1, \pm\mu_2\} \text{ and each } j \in J_\mu, \quad \|w_j^{(T)} - w_j^{(0)}\| \leq \frac{8}{\sqrt{m}}. \tag{51}$$

Let $\mu(x) \in \{\pm\mu_1, \pm\mu_2\}$ be such that $x \in I_{\mu(x)}$. Then by Lemma 5, we know that $\|x - \mu(x)\| \leq C_1\sigma\sqrt{d}$, so that for any $j \in J_{\mu(x)}$,

$$\begin{aligned}
\langle w_j^{(T)} - w_j^{(0)}, x\rangle &= \left\langle w_j^{(T)} - w_j^{(0)}, \mu\right\rangle + \left\langle w_j^{(T)} - w_j^{(0)}, x - \mu\right\rangle\\
&\overset{(i)}{\geq} \frac{\exp(-2)}{256\sqrt{m}} - C_1\sigma\sqrt{d}\|w_j^{(T)} - w_j^{(0)}\|\\
&\overset{(ii)}{\geq} \frac{\exp(-2)}{256\sqrt{m}} - \frac{8C_1\sigma\sqrt{d}}{\sqrt{m}}\\
&\overset{(iii)}{\geq} \frac{\exp(-2)}{512\sqrt{m}}.
\end{aligned} \tag{52}$$

In inequality (i) we use (50) and $\|x - \mu(x)\| \leq C_1\sigma\sqrt{d}$. In inequality (ii) we use (51), and in inequality (iii) we use Assumption (A2) so that for $C > 1$ sufficiently large, we have $8C_1\sigma\sqrt{d} \leq \exp(-2)/512$. Since $\phi(z_1) - \phi(z_2) = z_1 - z_2$ when both $z_1 > 0$ and $z_2 > 0$, we thus have

$$\begin{aligned} &\text{for all } i \in [n] \text{ and all } j \in J_{\mu(x_i)} \text{ satisfying } \langle w_j^{(0)}, x_i\rangle > 0, \\ &\text{we have} \quad \phi(\langle w_j^{(T)}, x_i\rangle) - \phi(\langle w_j^{(0)}, x_i\rangle) \geq \frac{\exp(-2)}{1024\sqrt{m}}. \end{aligned} \tag{53}$$

Now, note that by Lemma 5, $\|x\| \leq 2$, and by Lemma 5, we have $\|w_j^{(0)}\| \leq 2\omega_{\text{init}}\sqrt{d}$. Continuing from (52), we therefore have for any $j \in J_{\mu(x)}$,

$$\begin{aligned} \langle w_j^{(T)}, x\rangle &\geq \frac{\exp(-2)}{512\sqrt{m}} - \langle w_j^{(0)}, x\rangle \\ &\geq \frac{\exp(-2)}{512\sqrt{m}} - 4\omega_{\text{init}}\sqrt{d} \\ &\overset{(i)}{\geq} \frac{\exp(-2)}{1024\sqrt{m}}, \end{aligned}$$

where inequality (i) uses Assumption (A6) so that for $C > 1$ sufficiently large, we have $\omega_{\text{init}} \leq \exp(-2)/(4096\sqrt{md})$. Since $\phi(\langle w_j^{(0)}, x\rangle) = 0$ for $\langle w_j^{(0)}, x\rangle < 0$, this implies that

$$\begin{aligned} &\text{for all } i \in [n] \text{ and all } j \in J_{\mu(x_i)} \text{ satisfying } \langle w_j^{(0)}, x_i\rangle \leq 0, \\ &\text{we have} \quad \phi(\langle w_j^{(T)}, x_i\rangle) - \phi(\langle w_j^{(T)}, x_i\rangle) \geq \frac{\exp(-2)}{1024\sqrt{m}}. \end{aligned} \tag{54}$$

Putting together (53) and (54), we see that,

$$\begin{aligned} \|\phi(W^{(T)}x_i) - \phi(W^{(0)}x_i)\|^2 &= \sum_{j=1}^{m} |\phi(\langle w_j^{(T)}, x_i\rangle) - \phi(\langle w_j^{(0)}, x_i\rangle)|^2 \\ &\geq |J_{\mu(x_i)}| \cdot \left(\frac{\exp(-2)}{1024\sqrt{m}}\right)^2 \\ &\overset{(i)}{\geq} \frac{\exp(-4)}{1024^2} \cdot \frac{1}{4}\left(1 - \frac{1}{C_0}\right)^2 \\ &\geq \frac{\exp(-4)}{8 \cdot 1024^2}, \end{aligned} \tag{55}$$

where inequality (i) uses the lower bound on $|J_\mu|$ given in Lemma 4.

On the other hand, we have

$$\begin{aligned} \|\phi(W^{(0)}x_i)\| &\overset{(i)}{\leq} \|W^{(0)}\|_F \|x_i\| \\ &\overset{(ii)}{\leq} \frac{3}{2}\omega_{\text{init}}\sqrt{md} \cdot 2, \end{aligned}$$

where (i) uses that ϕ is 1-Lipschitz and (ii) uses Lemma 5 and Lemma 14. Putting this upper bound together with (55), we get,

$$\frac{\|\phi(W^{(T)}x_i) - \phi(W^{(0)}x_i)\|}{\|\phi(W^{(0)}x_i)\|} \geq \frac{\exp(-2)}{16 \cdot 1024\omega_{\text{init}}\sqrt{md}},$$

completing the proof. ■

Appendix D. On the Optimal Error in the Noiseless Setting

In this section we show that in the noiseless setting ($\eta = 0$), under assumptions (A1) through (A3), the optimal error achievable in $O(\sqrt{\log(1/\delta)/n})$ and that this test error is achieved by the classifier $x \mapsto \operatorname{sgn}(|\langle \mu_1, x\rangle| - |\langle \mu_2, x\rangle|)$.

Denote $\nu(x) := |\langle \mu_1, x\rangle| - |\langle \mu_2, x\rangle|$. By definition, the test error for the classifier induced by ν is

$$\begin{aligned}\mathbb{P}(\operatorname{sgn}(\nu(x)) \neq y) &= \mathbb{P}(y\nu(x) < 0) \\ &= \frac{1}{4}\Bigg(\mathbb{P}_{z\sim\mathsf{P}_{\text{clust}}}(\nu(z+\mu_1) < 0) + \mathbb{P}_{z\sim\mathsf{P}_{\text{clust}}}(\nu(z-\mu_1) < 0) \\ &\quad + \mathbb{P}_{z\sim\mathsf{P}_{\text{clust}}}(-\nu(z+\mu_2) < 0) + \mathbb{P}_{z\sim\mathsf{P}_{\text{clust}}}(-\nu(z-\mu_2) < 0)\Bigg).\end{aligned}$$

We shall show that $\mathbb{P}_{z\sim\mathsf{P}_{\text{clust}}}(\nu(z+\mu_1) < 0) = o_n(1)$, and an identical argument will yield the same bound for the remaining three terms. By definition,

$$\begin{aligned}\mathbb{P}_{z\sim\mathsf{P}_{\text{clust}}}(\nu(z+\mu_1) < 0) &= \mathbb{P}_z(|\langle \mu_1, z+\mu_1\rangle| - |\langle \mu_2, z+\mu_1\rangle|) \\ &= \mathbb{P}_z(|1 + \langle \mu_1, z\rangle| - |\langle \mu_2, z\rangle| < 0) \\ &\leq \mathbb{P}_z(|\langle \mu_1, z\rangle| + |\langle \mu_2, z\rangle| > 1) \\ &\leq \mathbb{P}_z(|\langle \mu_1, z\rangle| > 1/3) + \mathbb{P}_z(|\langle \mu_2, z\rangle| > 1/3). \qquad (56)\end{aligned}$$

For $i \in \{1, 2\}$, since $z \sim \mathsf{P}_{\text{clust}}$ is log-concave with $\mathbb{E}[z] = 0$, $\mathbb{E}[zz^\top] = \sigma^2 I$ and $\|\mu_i\| = 1$, $\langle \mu_i, z/\sigma\rangle$ is isotropic and log-concave and hence $\mathbb{P}_{z\sim\mathsf{P}_{\text{clust}}}(|\langle \mu_i, z\rangle| > 1/3) \leq 3\exp(-\sigma^{-1}/3)$ using Lovász and Vempala (2007, Theorem 5.1 and Lemma 5.7). By assumption (A2) this means $\mathbb{P}_{z\sim\mathsf{P}_{\text{clust}}}(|\langle \mu_i, z\rangle| > 1/3) \leq 3\exp(-C\sqrt{d}/3)$. We claim that this quantity is at most $O(\sqrt{\log(1/\delta)/n})$ under assumption (A1). To see this, note that $\exp(-C\sqrt{d}/3) \leq \sqrt{\log(1/\delta)/n}$ if and only if $C\sqrt{d} > \frac{3}{2}\log(n/\log(1/\delta))$. We have,

$$\log^2\left(\frac{n}{\log(1/\delta)}\right) = \left(\log n - \log\frac{1}{\delta}\right)^2 \leq \left(\log n + \frac{1}{\delta}\right)^2 = \log^2(n/\delta).$$

In particular, since by assumption (A1) we have $d \geq C\log^2(n/\delta)$, we also have $d \geq C\log^2(n/\log(1/\delta))$. In particular, $\sqrt{d} > \sqrt{C}\log(n/\log(1/\delta)) > \frac{3}{2}\log(n/\log(1/\delta))$ for $C \geq 2$. This shows that $\exp(-C\sqrt{d}/3) \leq \sqrt{\log(1/\delta)/n}$ and hence $\mathbb{P}_z(|\langle \mu_i, z\rangle| > 1/3) = O(\sqrt{\log(1/\delta)/n})$ for each i. Substituting this into (56) shows that

$$\mathbb{P}_{z\sim\mathsf{P}_{\text{clust}}}(\nu(z+\mu_1) < 0) = O(\sqrt{\log(1/\delta)/n}).$$

Appendix E. Experiment details

We provide here the experimental details for Figure 2. We consider a two-layer ReLU network of the form (1) with $m = 400$ neurons. The within-cluster distribution is Gaussian, $\mathsf{P}_{\text{clust}} \sim \mathsf{N}(0, \sigma^2 I_d)$, where the within-cluster variance is given by $\sigma^2 = 1/d^{1.2}$ and we flip 15% of the labels within each cluster the orthogonal cluster's label. We initialize using centered Gaussians with variance $\omega_{\text{init}}^2 = {}^{0.01}/_{md}$ and run with a step-size of $\alpha = 0.1$. Validation accuracy is measured using $n = 6000$ samples.

References

Emmanuel Abbe, Pritish Kamath, Eran Malach, Colin Sandon, and Nathan Srebro. On the power of differentiable learning versus PAC and SQ learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Radoslaw Adamczak, Rafal Latala, Alexander E. Litvak, Krzysztof Oleszkiewicz, Alain Pajor, and Nicole Tomczak-Jaegermann. A short proof of Paouris' inequality. *Canadian Mathematical Bulletin*, 2014.

Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep learning. *Preprint, arXiv:2001.04413*, 2021.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning (ICML)*, 2019.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yu Bai and Jason Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Peter Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research (JMLR)*, 2003.

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.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Alex Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory (COLT)*, 2022.

Amit Daniely and Eran Malach. Learning parities with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory (COLT)*, 2020.

Simon Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Cong Fang, Jason Lee, Pengkun Yang, and Tong Zhang. Modeling from features: A mean-field framework for over-parameterized deep neural networks. In *Conference on Learning Theory (COLT)*, 2021.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of a single neuron with gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Spencer Frei, Yuan Cao, and Quanquan Gu. Provable generalization of SGD-trained neural networks of any width in the presence of adversarial label noise. In *International Conference on Machine Learning (ICML)*, 2021.

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

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

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Wei Hu, Zhiyuan Li, and Dingli Yu. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. In *International Conference on Learning Representations (ICLR)*, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ziwei Ji, Justin D. Li, and Matus Telgarsky. Early-stopped neural networks are consistent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Pritish Kamath, Omar Montasser, and Nathan Srebro. Approximate is good enough: Probabilistic variants of dimensional and margin complexity. In *Conference on Learning Theory (COLT)*, 2020.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Philip Long. Properties of the after kernel. *Preprint, arXiv:2105.10585*, 2021.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Struct. Algorithms*, 30(3):307–358, 2007. ISSN 1042-9832.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Tengyu Ma. CS229T/STAT231: Statistical Learning Theory lecture notes (Fall 2017). `https://web.archive.org/web/20200901203150/http://web.stanford.edu/class/cs229t/scribe_notes/10_17_final.pdf`, 2017.

Eran Malach, Pritish Kamath, Emmanuel Abbe, and Nathan Srebro. Quantifying the benefit of using differentiable learning over tangent kernels. In *International Conference on Machine Learning (ICML)*, 2021.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 2018.

Mary Phuong and Christoph H Lampert. The inductive bias of re{lu} networks on orthogonally separable data. In *International Conference on Learning Representations*, 2021.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning (ICML)*, 2020.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 2019.

Martin Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

Colin Wei, Jason Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Greg Yang and Edward Hu. Feature learning in infinite-width neural networks. In *International Conference on Machine Learning (ICML)*, 2021.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Gilad Yehudai and Ohad Shamir. Learning a single neuron with gradient methods. In *Conference on Learning Theory (COLT)*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 2019.