Benign Overfitting without Linearity: Neural Network Classifiers Trained by Gradient Descent for Noisy Linear Data

Spencer Frei FREI@BERKELEY.EDU
UC Berkeley

Niladri S. Chatterji NILADRI@CS.STANFORD.EDU
Stanford University

Peter L. Bartlett PETER@BERKELEY.EDU
UC Berkeley

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Benign overfitting, the phenomenon where interpolating models generalize well in the presence of noisy data, was first observed in neural network models trained with gradient descent. To better understand this empirical observation, we consider the generalization error of two-layer neural networks trained to interpolation by gradient descent on the logistic loss following random initialization. We assume the data comes from well-separated class-conditional log-concave distributions and allow for a constant fraction of the training labels to be corrupted by an adversary. We show that in this setting, neural networks exhibit benign overfitting: they can be driven to zero training error, perfectly fitting any noisy training labels, and simultaneously achieve minimax optimal test error. In contrast to previous work on benign overfitting that require linear or kernel-based predictors, our analysis holds in a setting where both the model and learning dynamics are fundamentally nonlinear.

1. Introduction

Trained neural networks have been shown to generalize well to unseen data even when trained to interpolation (that is, vanishingly small training loss) on training data with significant label noise (Zhang et al., 2017; Belkin et al., 2019). This empirical observation is surprising as it appears to violate long standing intuition from statistical learning theory that the greater the capacity of a model to fit randomly labelled data, the worse the model's generalization performance on test data will be. This conflict between theory and practice has led to a surge of theoretical research into the generalization performance of interpolating statistical models to see if this 'benign overfitting' phenomenon can be observed in simpler settings that are more amenable to theoretical investigation. We now understand that benign overfitting can occur in many classical statistical settings, including linear regression (Hastie et al., 2019; Bartlett et al., 2020; Muthukumar et al., 2020; Negrea et al., 2020; Tsigler and Bartlett, 2020; Chinot et al., 2020; Chatterji et al., 2021), sparse linear regression (Koehler et al., 2021; Chatterji and Long, 2021b; Li and Wei, 2021; Wang et al., 2021a), logistic regression (Montanari et al., 2019; Chatterji and Long, 2021a; Liang and Sur, 2020; Muthukumar et al., 2021; Wang et al., 2021b; Minsker et al., 2021), and kernel-based estimators (Belkin et al., 2018; Mei and Montanari, 2019; Liang and Rakhlin, 2020; Liang et al., 2020), among others, and our understanding of when and why this phenomenon occurs in these settings is rapidly increasing. And yet, for the class of models from which the initial motivation for understanding benign overfitting arose—trained neural networks—we understand remarkably little.

In this work, we consider the class of two-layer networks with smoothed leaky ReLU activations trained on data coming from a high-dimensional linearly separable dataset where a constant fraction of the training labels can be adversarially corrupted (Kearns et al., 1994). We demonstrate that networks trained by standard gradient descent on the logistic loss in this setting exhibit benign overfitting: they can be driven to zero loss, and thus interpolate the noisy training data, and simultaneously achieve minimax optimal generalization error.

Our results follow by showing that the training loss can be driven to zero while the expected normalized margin for clean data points is large. The key technical ingredient of the proof for both of these claims is a 'loss ratio bound': we show that the gradient descent dynamics ensure that the loss of each example decreases at roughly the same rate throughout training. This ensures that the noisy points cannot have an outsized influence on the training dynamics, so that we can have control over the normalized margin for clean data points throughout training. At a high-level, this is possible because the data is high-dimensional, which ensures that all data points are roughly mutually orthogonal.

Our results hold for finite width networks, and since the logistic loss is driven to zero, the weights traverse far from their randomly initialized values. As a consequence, this shows benign overfitting behavior in trained neural networks beyond the kernel regime (Jacot et al., 2018).

1.1. Related Work

A number of recent works have characterized the generalization performance of interpolating models. Most related to ours are those in the classification setting. Chatterji and Long (2021a) study the high-dimensional sub-Gaussian mixture model setup we consider here, where labels can be corrupted adversarially, and analyze the performance of the maximum margin linear classifier. They do so by utilizing recent works that show that the weights found by unregularized gradient descent on the logistic loss asymptotically approach the maximum margin classifier for linearly separable data (Soudry et al., 2018; Ji and Telgarsky, 2019). Our proof techniques can be viewed as an extension of some of the techniques developed by Chatterji and Long in the logistic regression setting to two-layer neural networks. Muthukumar et al. (2021) study the behavior of the overparameterized max-margin classifier in a discriminative classification model with label-flipping noise, by connecting the behavior of the max-margin classifier to the ordinary least squares solution. They show that under certain conditions, all training data points become support vectors of the maximum margin classifier (see also, Hsu et al., 2021). Following this, Wang and Thrampoulidis (2021) and Cao et al. (2021) analyze the behavior of the overparameterized max-margin classifier in high dimensional mixture models by exploiting the connection between the max-margin classifier and the OLS solution. In contrast with these works, we consider the generalization performance of an interpolating nonlinear neural network.

A key difficulty in establishing benign overfitting guarantees for trained neural networks lies in demonstrating that the neural network can interpolate the data. Brutzkus et al. (2018) study SGD on two-layer networks with leaky ReLU activations and showed that for linearly separable data, stochastic gradient descent on the hinge loss will converge to zero training loss. They provided guarantees for the test error provided the number of samples is sufficiently large relative to the input dimension and the Bayes error rate is zero, but left open the question of what happens when there is label noise or when the data is high-dimensional. Frei et al. (2021) show that for linear separable data with labels corrupted by adversarial label noise (Kearns et al., 1994), SGD on the logistic loss of two-layer leaky ReLU networks achieves test error that is at most a constant multiple of the square root of the noise rate under mild distributional assumptions. However, their proof technique did not allow for the network to be trained to interpolation. In contrast, we allow for the network to be trained to arbitrarily small loss and hence interpolate noisy data. In principle, this could allow for the noisy samples to adversely influence the classifier, but we show this does not happen.

A series of recent works have exploited the connection between overparameterized neural networks and an infinite width approximation known as the neural tangent kernel (NTK) (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). These works show that for a certain scaling regime of the initialization, learning rate, and width of the network, neural networks trained by gradient descent behave similarly to their linearization around random initialization and can be well-approximated by the NTK. The near-linearity simplifies much of the analysis of neural network optimization and generalization. Indeed, a number of recent works have characterized settings in which neural networks in the kernel regime can exhibit benign overfitting (Liang et al., 2020; Montanari and Zhong, 2021).

Unfortunately, the kernel approximation fails to meaningfully capture a number of aspects of neural networks trained in practical settings, such as the ability to learn features (Yang and Hu, 2021), so that previous kernel-based approaches for understanding neural networks provide a quite restricted viewpoint for understanding neural networks in practice. By contrast, in this work, we develop an analysis of benign overfitting in finite width neural networks trained for many iterations on the logistic loss. We show that gradient descent drives the logistic loss to zero so that the weights grow to infinity, far from the near-initialization region where the kernel approximation holds, while the network simultaneously maintains a positive margin on clean examples. This provides the first guarantee for benign overfitting that does not rely upon an effectively linear evolution of the parameters.

Finally, we note in a concurrent work Cao et al. (2022) characterize the generalization performance of interpolating two-layer convolutional neural networks. They consider a distribution where input features consist of two patches, a 'signal' patch and a 'noise' patch, and binary output labels are a deterministic function of the signal patch. They show that if the signal-to-noise ratio is larger than a threshold value then the interpolating network achieves near-zero test error, while if the signal-to-noise ratio is smaller than the threshold then the interpolating network generalizes poorly. There are a few key differences in our results. First, our setup allows for a constant fraction of the training labels to be random, while in their setting the training labels are a deterministic function of the input features. Achieving near-zero training loss in our setting thus requires overfitting to noisy labels, in contrast to their setting where such overfitting is not possible. Second, they require the input dimension to be at least as large as m^2 (where m is the number of neurons in the network), while our results do not make any assumptions on the relationship between the input dimension and the number of neurons in the network.

2. Preliminaries

In this section we introduce the assumptions on the data generation process, the neural network architecture, and the optimization algorithm we consider.

2.1. Notation

We denote the ℓ^2 norm of a vector $x \in \mathbb{R}^p$ by $\|x\|$. For a matrix $W \in \mathbb{R}^{m\times p}$, we use $\|W\|_F$ to denote its Frobenius norm and $\|W\|_2$ to denote its spectral norm, and we denote its rows by $w_1, \ldots, w_m$. For an integer n, we use the notation $[n]$ to refer to the set $[n] = \{1, 2, \ldots, n\}$.

2.2. Setting

We shall let $C > 1$ denote a positive absolute constant, and our results will hold for all values of C sufficiently large. We consider a mixture model setting similar to one previously considered by Chatterji and Long (2021a), defined in terms of a joint distribution P over $(x, y) \in \mathbb{R}^p \times \{\pm 1\}$. Samples from this distribution can have noisy labels, and so we will find it useful to first describe a 'clean' distribution $\tilde{\mathsf{P}}$ and then define the true distribution P in terms of $\tilde{\mathsf{P}}$. Samples (x, y) from P are constructed as follows:

1. Sample a clean label $\tilde{y} \in \{\pm 1\}$ uniformly at random, $\tilde{y} \sim \mathsf{Uniform}(\{+1, -1\})$.
2. Sample $z \sim \mathsf{P}_{\mathsf{clust}}$ where
   - $\mathsf{P}_{\mathsf{clust}} = \mathsf{P}^{(1)}_{\mathsf{clust}} \times \cdots \times \mathsf{P}^{(p)}_{\mathsf{clust}}$ is a product distribution whose marginals are all mean-zero with sub-Gaussian norm at most one;
   - $\mathsf{P}_{\mathsf{clust}}$ is a λ-strongly log-concave distribution over $\mathbb{R}^p$ for some $\lambda > 0$;[1]
   - for some $\kappa > 0$, it holds that $\mathbb{E}_{z\sim\mathsf{P}_{\mathsf{clust}}}[\|z\|^2] \geq \kappa p$.
3. Generate $\tilde{x} = z + \tilde{y}\mu$.
4. Then, given a noise rate $\eta \in [0, 1/C]$, P is any distribution over $\mathbb{R}^p \times \{\pm 1\}$ such that the marginal distribution of the features for P and $\tilde{\mathsf{P}}$ coincide, and the total variation distance between the two distributions satisfies $d_{\mathsf{TV}}(\tilde{\mathsf{P}}, \mathsf{P}) \leq \eta$. Equivalently, P has the same marginal distribution over x as $\tilde{\mathsf{P}}$, but a sample $(x, y) \sim \mathsf{P}$ has label equal to $\tilde{y}$ with probability $1 - \eta(x)$ and has label equal to $-\tilde{y}$ with probability $\eta(x)$, where $\eta(x) \in [0, 1]$ satisfies $\mathbb{E}_{x\sim\mathsf{P}}[\eta(x)] \leq \eta$.

We note that the above assumptions coincide with those used by Chatterji and Long (2021a) in the linear setting with the exception of the introduction of an assumption of λ strong log-concavity that we introduce. This assumption is needed so that we may employ a concentration inequality for Lipschitz functions for strongly log-concave distributions. We note that variations of this data model have also been studied recently (Wang and Thrampoulidis, 2021; Liang and Recht, 2021; Wang et al., 2021c).

One example of a cluster distribution which satisfies the above assumptions is the (possibly anisotropic) Gaussian.

Example 1 *If $\mathsf{P}_{\mathsf{clust}} = \mathsf{N}(0, \Sigma)$, where $\|\Sigma\|_2 \leq 1$ and $\|\Sigma^{-1}\| \leq 1/\kappa$, and each of the labels are flipped independently with probability η, then all the properties listed above are satisfied.*

[1] That is, $z \sim \mathsf{P}_{\mathsf{clust}}$ has a probability density function p_z satisfying $p_z(x) = \exp(-U(x))$ for some convex function $U : \mathbb{R}^p \to \mathbb{R}$ such that $\nabla^2 U(x) - \lambda I_p$ is positive semidefinite.

Next, we introduce the neural network architecture and the optimization algorithm. We consider one-hidden-layer neural networks of width m that take the form

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

where we denote the input $x \in \mathbb{R}^p$ and emphasize that the network is parameterized by a matrix $W \in \mathbb{R}^{m\times p}$ corresponding to the first layer weights $\{w_j\}_{j=1}^m$. The network's second layer weights $\{a_j\}_{j=1}^m$ are initialized $a_j \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}(\{1/\sqrt{m}, -1/\sqrt{m}\})$ and fixed at their initial values. We assume the activation function ϕ satisfies $\phi(0) = 0$ and is strictly increasing, 1-Lipschitz, and H-smooth, that is, it is twice differentiable almost everywhere and there exist $\gamma, H > 0$ such that

$$0 < \gamma \le \phi'(z) \le 1, \quad \text{and} \quad |\phi''(z)| \le H, \ \forall z \in \mathbb{R}.$$

An example of such a function is a smoothed leaky ReLU activation,

$$\phi_{\text{SLReLU}}(z) = \begin{cases} z - \frac{1-\gamma}{4H}, & z \ge 1/H, \\ \frac{1-\gamma}{4} H z^2 + \frac{1+\gamma}{2} z, & |z| \le 1/H, \\ \gamma z - \frac{1-\gamma}{4H}, & z \le -1/H. \end{cases} \tag{1}$$

As $H \to \infty$, ϕ_{SLReLU} approximates the leaky ReLU activation $z \mapsto \max(\gamma z, z)$. We shall refer to functions ϕ satisfying the above properties as γ-leaky, H-smooth activations.

We assume access to a set of samples $S = \{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathsf{P}^n$. We denote by $\mathcal{C} \subset [n]$ the set of indices corresponding to samples with *clean* labels, and $\mathcal{N}$ as the set of indices corresponding to *noisy* labels, so that $i \in \mathcal{N}$ implies $(x_i, y_i) \sim \mathsf{P}$ is such that $y_i = -\tilde{y}_i$ using the notation above.

Let $\ell(z) = \log(1 + \exp(-z))$ be the logistic loss, and denote the empirical and population risks under ℓ by

$$\widehat{L}(W) := \frac{1}{n}\sum_{i=1}^n \ell(y_i f(x_i; W)) \quad \text{and} \quad L(W) := \mathbb{E}_{(x,y)\sim\mathsf{P}}\left[\ell(y f(x;W))\right].$$

We will also find it useful to treat the function $-\ell'(z) = 1/(1+\exp(z))$ as a loss itself: since ℓ is convex and decreasing, $-\ell'$ is non-negative and decreasing and thus can serve as a surrogate for the 0-1 loss. This trick has been used in a number of recent works on neural network optimization (Cao and Gu, 2020; Frei et al., 2019; Ji and Telgarsky, 2020; Frei et al., 2021). To this end, we introduce the notation,

$$g(z) := -\ell'(z) \quad \text{and} \quad \widehat{G}(W) := \frac{1}{n}\sum_{i=1}^n g(y_i f(x_i; W)).$$

We also introduce notation to refer to the function output and the surrogate loss g evaluated at samples for a given time point,

$$f_i^{(t)} := f(x_i; W^{(t)}) \quad \text{and} \quad g_i^{(t)} := g\big(y_i f_i^{(t)}\big).$$

We initialize the first layer weights independently for each neuron according to standard normals $[W^{(0)}]_{i,j} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \omega_{\text{init}}^2)$, where ω_{init}^2 is the initialization variance. The optimization algorithm we consider is unregularized full-batch gradient descent on $\widehat{L}(W)$ initialized at $W^{(0)}$ with fixed step-size $\alpha > 0$ which has updates

$$W^{(t+1)} = W^{(t)} - \alpha \nabla \widehat{L}(W^{(t)}).$$

Given a failure probability $\delta \in (0, 1/2)$, we make the following assumptions on the parameters in the paper going forward:

(A1) Number of samples $n \geq C \log(1/\delta)$.

(A2) Dimension $p \geq C \max\{n\|\mu\|^2, n^2 \log(n/\delta)\}$.

(A3) Norm of the mean satisfies $\|\mu\|^2 \geq C \log(n/\delta)$.

(A4) Noise rate $\eta \leq 1/C$.

(A5) Step-size $\alpha \leq \left(C \max\left\{1, \frac{H}{\sqrt{m}}\right\} p^2\right)^{-1}$, where ϕ is H-smooth.

(A6) Initialization variance satisfies $\omega_{\text{init}}\sqrt{mp} \leq \alpha$.

Assumptions (A1), (A2), and (A3) above have previously appeared in Chatterji and Long (2021a) and put a constraint on how the number of samples, dimension, and cluster mean separation can relate to one another. One regime captured by these assumptions is when the mean separation satisfies $\|\mu\| = \Theta(p^\beta)$, where $\beta \in (0, 1/2)$ and $p \geq C \max\{n^{\frac{1}{1-2\beta}}, n^2 \log(n/\delta)\}$. Assumption (A6) ensures that the first step of gradient descent dominates the behavior of the neural network relative to that at initialization; this will be key to showing that the network traverses far from initialization after a single step, which we show in Proposition 2. We note that our analysis holds for neural networks of arbitrary width $m \geq 1$.

3. Main Result

Our main result is that when a neural network is trained on samples from the distribution P described in the previous section, it will exhibit benign overfitting: the network achieves arbitrarily small logistic loss, and hence interpolates the noisy training data, and simultaneously achieves test error close to the noise rate.

Theorem 1 *For any γ-leaky, H-smooth activation ϕ, and for all $\kappa \in (0, 1)$, $\lambda > 0$, there is a $C > 1$ such that provided Assumptions (A1) through (A6) are satisfied, the following holds. For any $0 < \varepsilon < 1/2n$, by running gradient descent for $T \geq C\widehat{L}(W^{(0)})/\left(\|\mu\|^2 \alpha \varepsilon^2\right)$ iterations, with probability at least $1 - 2\delta$ over the random initialization and the draws of the samples, the following holds:*

1. *All training points are classified correctly and the training loss satisfies $\widehat{L}(W^{(T)}) \leq \varepsilon$.*

2. *The test error satisfies*

$$\mathbb{P}_{(x,y)\sim \mathsf{P}}\left[y \neq \operatorname{sgn}(f(x; W^{(T)}))\right] \leq \eta + 2\exp\left(-\frac{n\|\mu\|^4}{Cp}\right).$$

Theorem 1 shows that neural networks trained by gradient descent will exhibit benign overfitting: the logistic loss can be driven to zero so that the network interpolates the noisy training data, and the trained network will generalize with classification error close to the noise rate η provided $n\|\mu\|^4 \gg p$. Note that when $\mathsf{P}_{\mathsf{clust}} = \mathsf{N}(0, I)$, Giraud and Verzelen (2019, Appendix B) showed that in the noiseless case ($\eta = 0$), the minimax test error is at least $c\exp(-c' \min(\|\mu\|^2, n\|\mu\|^4/p))$ for some absolute constants $c, c' > 0$. In the setting of random classification noise, where labels are flipped with probability η (i.e., $\eta(x) = \eta$ for all x), this implies that the minimax test error is at least $\eta + c\exp(-c' \min(\|\mu\|^2, n\|\mu\|^4/p))$. By Assumption (A3), $\|\mu\|^2 > n\|\mu\|^4/p$, so that the test error in Theorem 1 is minimax optimal up to constants in the exponent in the setting of random classification noise.

We briefly also compare our results to margin bounds in the literature. Note that even if one could prove that the training data is likely to be separated by a large margin, the bound of Theorem 1 approaches the noise rate faster than the standard margin bounds (Vapnik, 1999; Shawe-Taylor et al., 1998).

We note that our results do not require many of the assumptions typical in theoretical analyses of neural networks: we allow for networks of arbitrary width; we permit arbitrarily small initialization variance; and we allow for the network to be trained for arbitrarily long. In particular, we wish to emphasize that the optimization and generalization analysis used to prove Theorem 1 does not rely upon the neural tangent kernel approximation. One way to see this is that our results cover finite-width networks and require $\|W^{(t)}\| \to \infty$ as $\varepsilon \to 0$ since the logistic loss is never zero. In fact, for the choice of step-size and initialization variance given in Assumptions (A5) and (A6), the weights travel far from their initial values after a single step of gradient descent, as we show in Proposition 2 below.

Proposition 2 *Under the settings of Theorem 1, we have for some absolute constant $C > 1$ with probability at least $1 - 2\delta$ over the random initialization and the draws of the samples,*

$$\frac{\|W^{(1)} - W^{(0)}\|_F}{\|W^{(0)}\|_F} \geq \frac{\gamma\|\mu\|}{C}.$$

The proof for Proposition 2 is provided in Appendix B.

4. Proof of Theorem 1

In this section we will assume that Assumptions (A1) through (A6) are in force for a large constant $C > 1$.

Theorem 1 consists of two claims: the first is that the test error of the trained neural network is close to the noise rate when $n\|\mu\|^4/p \gg 1$, and the second is that the empirical loss can simultaneously be made arbitrarily small despite the presence of noisy labels. Both of these claims will be established via a series of lemmas. All of these lemmas are proved in Appendix A.

The first claim will follow by establishing a lower bound for the *expected normalized margin* on clean points, $\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\tilde{y}f(x; W^{(t)})/\|W^{(t)}\|_F]$. We do so in the following lemma which leverages the fact that $\mathsf{P}_{\mathsf{clust}}$ is λ-strongly log-concave.

Lemma 3 *Suppose that* $\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\tilde{y}f(x;W)] \geq 0$*. Then there exists a universal constant* $c > 0$ *such that*

$$\mathbb{P}_{(x,y)\sim\mathsf{P}}\big(y \neq \mathrm{sgn}(f(x;W))\big) \leq \eta + 2\exp\left(-c\lambda\left(\frac{\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\tilde{y}f(x;W)]}{\|W\|_F}\right)^2\right).$$

Lemma 3 demonstrates that the generalization bound will follow by showing a lower bound on the normalized margin of the neural network on clean samples at a given time. To derive such a result, we first need to introduce a number of structural results about the samples and the neural network objective function. The first such result concerns the norm of the weights at initialization.

Lemma 4 *There is a universal constant* $C_0 > 1$ *such that with probability at least* $1-\delta$ *over the random initialization,*

$$\|W^{(0)}\|_F^2 \leq \frac{3}{2}\omega_{\mathrm{init}}^2 mp \quad \textit{and} \quad \|W^{(0)}\|_2 \leq C_0\omega_{\mathrm{init}}(\sqrt{m}+\sqrt{p}).$$

Our next structural result characterizes some properties of random samples from the distribution. It was proved in Chatterji and Long (2021a, Lemma 10) and is a consequence of Assumptions (A1) through (A4).

Lemma 5 *For all* $\kappa > 0$*, there is* $C_1 > 1$ *such that for all* $c' > 0$*, for all large enough* C*, with probability at least* $1-\delta$ *over* P^n*, the following hold:*

E.1 For all $k \in [n]$*,*

$$p/C_1 \leq \|x_k\|^2 \leq C_1 p.$$

E.2 For all $i \neq j \in [n]$*,*

$$|\langle x_i, x_j\rangle| \leq C_1(\|\mu\|^2 + \sqrt{p\log(n/\delta)}).$$

E.3 For all $k \in \mathcal{C}$*,*

$$|\langle \mu, y_k x_k\rangle - \|\mu\|^2| \leq \|\mu\|^2/2.$$

E.4 For all $k \in \mathcal{N}$*,*

$$|\langle \mu, y_k x_k\rangle - (-\|\mu\|^2)| \leq \|\mu\|^2/2.$$

E.5 The number of noisy samples satisfies $|\mathcal{N}|/n \leq \eta + c'$*.*

Definition 6 *If the events in Lemma 4 and Lemma 5 occur, let us say that we have a* good run*.*

Lemmas 4 and 5 show that a good run occurs with probability at least $1-2\delta$. In what follows, we will assume that a good run occurs.

We next introduce a number of structural lemmas concerning the neural network optimization objective. The first concerns the smoothness of the network in terms of the first layer weights.

Lemma 7 *For an H-smooth activation ϕ and any $W, V \in \mathbb{R}^{m\times p}$ and $x \in \mathbb{R}^p$,*

$$|f(x;W) - f(x;V) - \langle \nabla f(x;V), W - V\rangle| \le \frac{H\|x\|^2}{2\sqrt{m}}\|W - V\|_2^2.$$

In the next lemma, we provide a number of smoothness properties of the empirical loss.

Lemma 8 *For an H-smooth activation ϕ and any $W, V \in \mathbb{R}^{m\times p}$, on a good run it holds that*

$$\frac{1}{\sqrt{C_1 p}}\|\nabla \widehat{L}(W)\|_F \le \widehat{G}(W) \le \widehat{L}(W) \wedge 1,$$

where C_1 is the constant from Lemma 5. Additionally,

$$\|\nabla \widehat{L}(W) - \nabla \widehat{L}(V)\|_F \le C_1 p \left(1 + \frac{H}{\sqrt{m}}\right)\|W - V\|_2.$$

Our final structural result is the following lemma that characterizes the pairwise correlations of the gradients of the network at different samples.

Lemma 9 *Let $C_1 > 1$ be the constant from Lemma 5. For a γ-leaky, H-smooth activation ϕ, on a good run, we have the following.*

(a) For any $i, k \in [n]$, $i \neq k$, and any $W \in \mathbb{R}^{m\times d}$, we have

$$|\langle \nabla f(x_i;W), \nabla f(x_k;W)\rangle| \le C_1\left(\|\mu\|^2 + \sqrt{p\log(n/\delta)}\right).$$

(b) For any $i \in [n]$ and any $W \in \mathbb{R}^{m\times d}$, we have

$$\frac{\gamma^2 p}{C_1} \le \|\nabla f(x_i;W)\|_F^2 \le C_1 p.$$

In the regime where $\|\mu\|^2 = o(p)$, Lemma 9 shows that the gradients of the network at different samples are roughly orthogonal as the pairwise inner products of the gradients are much smaller than the norms of each gradient. This mimics the behavior of the samples x_i established in Lemma 5.

With these structural results in place, we can now begin to prove a lower bound for the normalized margin on test points. To do so, our first step is to characterize the change in the unnormalized margin $y[f(x;W^{(t+1)}) - f(x;W^{(t)})]$ from time t to time $t+1$ for an independent test sample (x, y).

Lemma 10 *Let $C_1 > 1$ be the constant from Lemma 1. For a γ-leaky, H-smooth activation ϕ, on a good run, we have for any $t \ge 0$ and $(x, y) \in \mathbb{R}^p \times \{\pm 1\}$, and for each $i = 1, \ldots, n$, there exist $\xi_i = \xi(W^{(t)}, x_i, x) \in [\gamma^2, 1]$, such that*

$$y[f(x;W^{(t+1)}) - f(x;W^{(t)})] \ge \frac{\alpha}{n}\sum_{i=1}^n g_i^{(t)}\left[\xi_i\langle y_i x_i, yx\rangle - \frac{HC_1^2 p^2 \alpha}{2\sqrt{m}}\right],$$

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

Consider what Lemma 10 tells us when $(x, \tilde{y}) \sim \tilde{\mathsf{P}}$ is a clean test example. The lemma suggests that if $\langle y_i x_i, \tilde{y} x\rangle$ is always bounded from below by a strictly positive constant, then the margin on the test sample $(x, \tilde{y})$ will increase. Unfortunately, the presence of noisy labels will cause some of the $\langle y_i x_i, \tilde{y} x\rangle$ terms appearing above to be negative, allowing for the possibility that the unnormalized margin decreases on a test sample $(x, \tilde{y})$. If the losses $g(y_i f(x_i; W^{(t)}))$ for (noisy) samples satisfying $\langle y_i x_i, \tilde{y} x\rangle < 0$ are particularly large relative to the losses $g(y_{i'} f(x_{i'}; W^{(t)}))$ for (clean) samples satisfying $\langle y_{i'} x_{i'}, \tilde{y} x\rangle > 0$, then indeed Lemma 10 may fail to guarantee an increase in the unnormalized margin. However, if one could show that the g losses are essentially 'balanced' across *all* samples, then provided the fraction of noisy labels is not too large, one could ignore the effect of the noisy labels which contribute negative terms to the sum $\sum_i g_i^{(t)} \langle y_i x_i, \tilde{y} x\rangle$, and eventually show that the lower bound given in Lemma 10 is strictly positive. This provides a motivation for our next lemma, which directly shows that the losses on all samples are relatively balanced throughout training. This is the key technical lemma for our proof, and extends the results of Chatterji and Long (2021a) from the logistic regression setting to the two-layer neural network setting.

Lemma 11 *For a γ-leaky, H-smooth activation ϕ, there is an absolute constant $C_r = 16C_1^2/\gamma^2$ such that on a good run, provided $C > 1$ is sufficiently large, we have for all $t \geq 0$,*

$$\max_{i,j \in [n]} \frac{g_i^{(t)}}{g_j^{(t)}} \leq C_r.$$

With this loss ratio bound, we first derive an upper bound on the norm of the iterates $W^{(t)}$, sharper than what we get by applying the triangle inequality along with the bound on the norm of the gradient of the loss provided by Lemma 8. This will improve our final guarantee for the normalized margin.

Lemma 12 *There is an absolute constant $C_2 > 1$ such that for $C > 1$ sufficiently large, on a good run we have that for all $t \geq 0$,*

$$\|W^{(t)}\|_F \leq \|W^{(0)}\|_F + C_2 \alpha \sqrt{\frac{p}{n}} \sum_{s=0}^{t-1} \widehat{G}(W^{(s)}).$$

With the loss ratio bound provided in Lemma 11 and the tightened gradient norm bound of Lemma 12 established, we can now derive a lower bound on the normalized margin. Note that this lower bound on the normalized margin in conjunction with Lemma 3 results in the test error bound for the main theorem.

Lemma 13 *For a γ-leaky, H-smooth activation ϕ, and for all $C > 1$ sufficiently large, on a good run, for any $t \geq 1$,*

$$\frac{\mathbb{E}_{(x,\tilde{y})\sim \tilde{\mathsf{P}}}[\tilde{y} f(x; W^{(t)})]}{\|W^{(t)}\|_F} \geq \frac{\gamma^2 \|\mu\|^2 \sqrt{n}}{8 \max(\sqrt{C_1}, C_2)\sqrt{p}},$$

where C_1 and C_2 are the constants from Lemma 5 and Lemma 12, respectively.

Since Lemma 13 provides a positive margin on clean points, we have by Lemma 3 a guarantee that the neural network achieves classification error on the noisy distribution close to the noise level. The only remaining part of Theorem 1 that remains to be proved is that the training loss can be driven to zero. This is a consequence of the following lemma, the proof of which also crucially relies upon the loss ratio bound of Lemma 11.

Lemma 14 *For a γ-leaky, H-smooth activation ϕ, provided $C > 1$ is sufficiently large, then on a good run we have for all $t \geq 0$,*

$$\|\nabla \widehat{L}(W^{(t)})\|_F \geq \frac{\gamma\|\mu\|}{4}\widehat{G}(W^{(t)}).$$

Moreover, any $T \in \mathbb{N}$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(y_i \neq \mathrm{sgn}(f(x_i; W^{(T-1)}))\big) \leq 2\widehat{G}(W^{(T-1)}) \leq 2\left(\frac{32\widehat{L}(W^{(0)})}{\gamma^2\|\mu\|^2\alpha T}\right)^{1/2}.$$

In particular, for $T \geq 128\widehat{L}(W^{(0)})/\left(\gamma^2\|\mu\|^2\alpha\varepsilon^2\right)$, we have $\widehat{G}(W^{(T-1)}) \leq \varepsilon/2$.

We now have all the results necessary to prove our main theorem.

Proof [Proof of Theorem 1] By Lemma 5 and Lemma 4, a 'good run' occurs with probability at least $1 - 2\delta$. Since a good run occurs, we can apply Lemma 13. Using this as well as Lemma 3, we have with probability at least $1 - 2\delta$,

$$\begin{aligned}\mathbb{P}_{(x,y)\sim \mathsf{P}}\big(y \neq \mathrm{sgn}(f(x; W))\big) &\leq \eta + 2\exp\left(-c\lambda\left(\frac{\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\tilde{y}f(x;W)]}{\|W\|_2}\right)^2\right)\\ &\leq \eta + 2\exp\left(-c\lambda\left(\frac{\gamma^4 n\|\mu\|^4}{8^2\max(C_1, C_2^2)p}\right)\right).\end{aligned}$$

By Lemma 14, since $T \geq 32\widehat{L}(W^{(0)})/\left(\gamma\|\mu\|\alpha\varepsilon^2\right)$, we have

$$\widehat{G}(W^{(T-1)}) \leq \varepsilon/2.$$

Since $\varepsilon < 1/(2n)$ and $g(z) = -\ell'(z) < 1/2$ if and only if $z > 0$, we know that $y_i f(x_i; W^{(T-1)}) > 0$ for every $i \in [n]$. We are working with the logistic loss, and hence we have $\frac{1}{2}\ell(y_i f(x_i; W^{(T-1)})) \leq g(y_i f(x_i; W^{(T-1)}))$ for every $i \in [n]$, which implies that

$$\widehat{L}(W^{(T-1)}) = \frac{1}{n}\sum_{i=1}^{n}\ell(y_i f(x_i; W^{(T-1)})) \leq \frac{2}{n}\sum_{i=1}^{n}-\ell'(y_i f(x_i; W^{(T-1)})) = 2\widehat{G}(W^{(T-1)}) \leq \varepsilon.$$

■

5. Discussion

We have shown that neural networks trained by gradient descent will interpolate noisy training data and still generalize close to the noise rate when the data comes from a mixture of well-separated sub-Gaussian distributions and the dimension of the data is larger than the sample size. Our results mimic those established by Chatterji and Long (2021a) for linear classifiers, but they hold for the much richer class of two-layer neural networks.

Our proof technique relies heavily upon the assumption that the number of samples is much less than the ambient dimension. This assumption allows for every pair of distinct samples to be roughly mutually orthogonal so that samples with noisy labels cannot have an outsized effect on the ability for the network to learn a positive margin on clean examples. Previous work has established a similar 'blessing of dimensionality' phenomenon: Belkin et al. (2018) showed that the gap between a particular simplicial interpolation rule and the Bayes error decreases exponentially fast as the ambient dimension increases, mimicking the behavior we show in Theorem 1. In the linear regression setting, it is known that for the minimum norm solution to generalize well it is necessary for the dimension of the data p to be much larger than n (Bartlett et al., 2020). It has also been shown that if the ambient dimension is one, local interpolation rules necessarily have suboptimal performance (Ji et al., 2021). Taken together, these results suggest that working in high dimensions makes it easier for benign overfitting to hold, but it is an interesting open question to understand the extent to which working in the $p \geq n$ regime is necessary for benign overfitting with neural networks. In particular, when can benign overfitting occur in neural networks that have enough parameters to fit the training points $(mp > n)$ but for which the number of samples is larger than the input dimension $(n > p)$?

In this work we considered a data distribution for which the optimal classifier is linear but analyzed a model and algorithm that are fundamentally nonlinear. A natural next step is to develop characterizations of benign overfitting for neural networks trained by gradient descent in settings where the optimal classifier is nonlinear. We believe some of the insights developed in this work may be useful in these settings: in particular, it appears that in the $p \gg n$ setting, the optimization dynamics of gradient descent can become simpler as can be seen by the 'loss ratio bound' provided in Lemma 11. On the other hand, we believe the generalization analysis will become significantly more difficult when the optimal classifier is nonlinear.

Acknowledgments

We thank Christophe Giraud for pointing us to references that show that the test error in Theorem 1 is minimax optimal, and we thank Bryn Elesedy for bringing to our attention an issue regarding the definition of the activation in an earlier version of this work. We are thankful for the anonymous reviewers' helpful feedback. 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-2031883 and #814639.

References

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 S 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.

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

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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

Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. In *International Conference on Learning Representations (ICLR)*, 2018.

Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning over-parameterized deep ReLU networks. In *the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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

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

Niladri S Chatterji and Philip M Long. Foolish crowds support benign overfitting. *arXiv preprint arXiv:2110.02914*, 2021b.

Niladri S Chatterji, Philip M Long, and Peter L Bartlett. The interplay between implicit bias and benign overfitting in two-layer linear networks. *arXiv preprint arXiv:2108.11489*, 2021.

Geoffrey Chinot, Matthias Löffler, and Sara van de Geer. On the robustness of minimum-norm interpolators. *arXiv preprint arXiv:2012.00807*, 2020.

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

Spencer Frei and Quanquan Gu. Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Spencer Frei, Yuan Cao, and Quanquan Gu. Algorithm-dependent generalization bounds for over-parameterized deep residual networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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.

Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k-means. *Mathematical Statistics and Learning*, 1(3):317–374, 2019.

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

Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 91–99, 2021.

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 and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory (COLT)*, 2019.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. In *International Conference on Learning Representations (ICLR)*, 2020.

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

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

Frederic Koehler, Lijia Zhou, Danica J Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds, and benign overfitting. *arXiv preprint arXiv:2106.09276*, 2021.

M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical surveys and monographs. American Mathematical Society, 2001. ISBN 9780821837924.

Yue Li and Yuting Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*, 2021.

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

Tengyuan Liang and Benjamin Recht. Interpolating classifiers make few mistakes. *arXiv preprint arXiv:2101.11815*, 2021.

Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and min-ℓ_1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*, 2020.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory (COLT)*, 2020.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.

Stanislav Minsker, Mohamed Ndaoud, and Yiqiu Shen. Minimax supervised clustering in the anisotropic gaussian mixture model: A new take on robust interpolation. *Preprint, arXiv:2111.07041*, 2021.

Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *Preprint, arXiv:2007.12826*, 2021.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, pages 7263–7272, 2020.

John Shawe-Taylor, Peter L Bartlett, Robert C Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, 44(5): 1926–1940, 1998.

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*, 65(2):742–769, 2019.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research (JMLR)*, 19(70):1–57, 2018.

A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression. *Preprint, arXiv:2009.14286*, 2020.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer, 1999.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint*, arXiv:1011.3027, 2010.

M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. ISBN 9781108498029. URL `https://books.google.com/books?id=8C8nuQEACAAJ`.

Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum l1-norm interpolation of noisy data. *arXiv preprint arXiv:2111.05987*, 2021a.

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

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

Ke Alexander Wang, Niladri S Chatterji, Saminul Haque, and Tatsunori Hashimoto. Is importance weighting incompatible with interpolating classifiers? *arXiv preprint arXiv:2112.12986*, 2021c.

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

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 overparameterized deep ReLU networks. *Machine Learning*, 2019.

Contents

Appendix A. Omitted Proofs from Section 4

In this section we provide a proof of all of the lemmas presented in Section 4. We remind the reader that throughout this section, we assume that Assumptions (A1) through (A6) are in force.

First in Section A.1 we prove the concentration results, Lemmas 3 and 4. Next, in Section A.2 we prove the structural results, Lemmas 7, 8 and 9. In Section A.3 we prove Lemma 10 that demonstrates that the margin on a test point increases with training. In Section A.4 we prove Lemma 11 that guarantees that the ratio of the surrogate losses remains bounded throughout training, while in Section A.5 we prove Lemma 12 that bounds the growth of the norm of the parameters. Next, in Section A.6 we prove Lemma 13 that provides a lower bound on the normalized margin on a test point. Finally, in Section A.7, we prove Lemma 14 that is useful in proving that the training error and loss converge to zero.

A.1. Concentration Inequalities

In this subsection we prove the concentration results Lemmas 3 and 4.

A.1.1. PROOF OF LEMMA 3

Let us restate Lemma 3.

Lemma 15 *Suppose that* $\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\tilde{y}f(x;W)] \geq 0$. *Then there exists a universal constant* $c > 0$ *such that*

$$\mathbb{P}_{(x,y)\sim\mathsf{P}}\big(y \neq \operatorname{sgn}(f(x;W))\big) \leq \eta + 2\exp\left(-c\lambda\left(\frac{\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\tilde{y}f(x;W)]}{\|W\|_F}\right)^2\right).$$

Proof Following the proof of Chatterji and Long (2021a, Lemma 9), we have

$$\begin{aligned}\mathbb{P}_{(x,y)\sim\mathsf{P}}(y \neq \operatorname{sgn}(f(x;W)) &= \mathbb{P}_{(x,y)\sim\mathsf{P}}(y\operatorname{sgn}(f(x;W)) < 0)\\ &\leq \eta + \mathbb{P}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}(\tilde{y}f(x;W) < 0).\end{aligned}$$

It therefore suffices to provide an upper bound for $\mathbb{P}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}(\tilde{y}f(x;W) < 0)$. Towards this end, we first note that f is a $\|W\|_2$-Lipschitz function of the input x: let $x, x' \in \mathbb{R}^p$, and consider

$$\begin{aligned}|f(x;W) - f(x';W)| &= \left|\sum_{j=1}^m a_j[\phi(\langle w_j, x\rangle) - \phi(\langle w_j, x'\rangle)]\right|\\ &\overset{(i)}{\leq} \sum_{j=1}^m |a_j||\langle w_j, x - x'\rangle|\\ &\overset{(ii)}{\leq} \sqrt{\sum_{j=1}^m a_j^2}\sqrt{\sum_{j=1}^m \langle w_j, x - x'\rangle^2}\\ &= \|W(x - x')\|\\ &\overset{(iii)}{\leq} \|W\|_2\|x - x'\|.\end{aligned}$$

Above, (i) uses that ϕ is 1-Lipschitz, and (ii) follows by the Cauchy–Schwarz inequality. Inequality (iii) is by the definition of the spectral norm. This shows that $f(\cdot;W)$ is $\|W\|_2$-Lipschitz.

Since $\mathsf{P}_{\mathsf{clust}}$ is λ-strongly log-concave, by Ledoux (2001, Theorem 2.7 and Proposition 1.10), since $\tilde{y}f(x;W)$ is $\|W\|_2$-Lipschitz, there is an absolute constant $c > 0$ such that for any $q \geq 1$, $\|\tilde{y}f(x;W) - \mathbb{E}[\tilde{y}f(x;W)]\|_{L^q} \leq c\|W\|_2\sqrt{q/\lambda}$. This behavior of the growth of L^q norms is equivalent to $\tilde{y}f(x;W) - \mathbb{E}[\tilde{y}f(x;W)]$ having sub-Gaussian norm $c'\|W\|_2/\sqrt{\lambda}$ for some absolute constant $c' > 0$, by Vershynin (2010, Proposition 2.5.2). Thus, there is an absolute constant $c'' > 0$ such that for any $t \geq 0$,

$$\mathbb{P}(|\tilde{y}f(x;W) - \mathbb{E}[\tilde{y}f(x;W)]| \geq t) \leq 2\exp\left(-c''\lambda\left(\frac{t}{\|W\|_2}\right)^2\right). \tag{2}$$

Since we have the equality,

$$\mathbb{P}\big(\tilde{y} \neq \operatorname{sgn}(f(x;W))\big) = \mathbb{P}(\tilde{y}f(x;W) - \mathbb{E}[\tilde{y}f(x;W)] < -\mathbb{E}[\tilde{y}f(x;W)])$$

the result follows by taking $t = \mathbb{E}[\tilde{y}f(x;W)] \geq 0$ in (2) and using $\|W\|_2 \leq \|W\|_F$. ■

A.1.2. PROOF OF LEMMA 4

Now let us restate and prove Lemma 4.

Lemma 16 *There is a universal constant $C_0 > 1$ such that with probability at least $1-\delta$ over the random initialization,*

$$\|W^{(0)}\|_F^2 \le \frac{3}{2}\omega_{\text{init}}^2 mp \quad \textit{and} \quad \|W^{(0)}\|_2 \le C_0\omega_{\text{init}}(\sqrt{m}+\sqrt{p}).$$

Proof Note that $\|W^{(0)}\|_F^2$ is a ω_{init}^2-multiple of a chi-squared random variable with mp degrees of freedom. By concentration of the χ^2 distribution (Wainwright, 2019, Example 2.11), for any $t \in (0,1)$,

$$\mathbb{P}\left(\left|\frac{1}{mp\omega_{\text{init}}^2}\|W^{(0)}\|_F^2 - 1\right| \ge t\right) \le 2\exp(-mpt^2/8).$$

In particular, if we choose $t = \sqrt{8\log(4/\delta)/md}$ and use Assumption (A2), we get that $t \le 1/2$ and so with probability at least $1-\delta/2$, we have

$$\|W^{(0)}\|_F^2 \le \frac{3}{2}mp\omega_{\text{init}}^2.$$

As for the spectral norm, since the entries of $W^{(0)}/\omega_{\text{init}}$ are i.i.d. standard normal random variables, by Vershynin (2010, Theorem 4.4.5) there exists a universal constant $c > 0$ such that for any $u \ge 0$, with probability at least $1 - 2\exp(-u^2)$, we have

$$\|W^{(0)}\|_2 \le c\omega_{\text{init}}(\sqrt{m}+\sqrt{p}+u).$$

In particular, taking $u = \sqrt{\log(4/\delta)}$ we have with probability at least $1-\delta/2$, $\|W^{(0)}\|_2 \le c\omega_{\text{init}}(\sqrt{m}+\sqrt{p}+\sqrt{\log(4/\delta)}$. Since $\sqrt{p} \ge \sqrt{\log(4/\delta)}$ by Assumption (A2), the proof is completed by a union bound over the claims on the spectral norm and the Frobenius norm. ■

A.2. Structural Results

As stated above in this section we prove Lemmas 7, 8 and 9.

A.2.1. PROOF OF LEMMA 7

We begin by restating and proving Lemma 7.

Lemma 17 *For an H-smooth activation ϕ and any $W, V \in \mathbb{R}^{m\times p}$ and $x \in \mathbb{R}^p$,*

$$|f(x;W) - f(x;V) - \langle \nabla f(x;V), W - V\rangle| \le \frac{H\|x\|^2}{2\sqrt{m}}\|W-V\|_2^2.$$

Proof Since ϕ is twice differentiable, ϕ' is continuous and so by Taylor's theorem, for each $j \in [m]$, there exist constants $t_j = t_j(w_j, v_j, x) \in \mathbb{R}$,

$$\phi(\langle w_j, x\rangle) - \phi(\langle v_j, x\rangle) = \phi'(\langle v_j, x\rangle)\cdot\langle w_j - v_j, x\rangle + \frac{\phi''(t_j)}{2}(\langle w_j - v_j, x\rangle)^2,$$

where t_j lies between $\langle w_j, x\rangle$ and $\langle v_j, x\rangle$. We therefore have

$$\begin{aligned} f(x;W) - f(x;V) &= \sum_{j=1}^m a_j[\phi(\langle w_j, x\rangle) - \phi(\langle v_j, x\rangle)] \\ &= \sum_{j=1}^m a_j \left[\phi'(\langle v_j, x\rangle)\cdot\langle w_j - v_j, x\rangle + \frac{\phi''(t_j)}{2}\langle w_j - v_j, x\rangle^2\right] \\ &= \langle \nabla f(x;V), W-V\rangle + \sum_{j=1}^m a_j \frac{\phi''(t_j)}{2}\langle w_j - v_j, x\rangle^2. \end{aligned}$$

The last equality follows since we can write

$$\nabla f(x;V) = D_x^V a x^\top, \quad \text{where} \quad D_x^V := \operatorname{diag}(\phi'(\langle v_j, x\rangle)), \tag{3}$$

and thus

$$\langle \nabla f(x;V), W-V\rangle = \operatorname{tr}(x a^\top D_x^V (W-V)) = a^\top D_x^V (W-V)x = \sum_j a_j \phi'(\langle v_j, x\rangle)\langle w_j - v_j, x\rangle.$$

For the final term, we have

$$\begin{aligned} \left| \sum_{j=1}^m a_j \frac{\phi''(\xi_j)}{2}\langle w_j - v_j, x\rangle^2 \right| &\le \sum_{j=1}^m |a_j| \frac{|\phi''(t_j)|}{2}\langle w_j - v_j, x\rangle^2 \\ &\le \frac{H}{2\sqrt{m}}\sum_{j=1}^m \langle w_j - v_j, x\rangle^2 \\ &= \frac{H}{2\sqrt{m}}\left\|(W-V)x\right\|_2^2 \\ &\le \frac{H}{2\sqrt{m}}\left\|W-V\right\|_2^2 \left\|x\right\|_2^2. \end{aligned}$$

■

A.2.2. PROOF OF LEMMA 8

Next we prove Lemma 8 that establishes that the loss is smooth.

Lemma 18 *For an H-smooth activation ϕ and any $W, V \in \mathbb{R}^{m\times p}$, on a good run it holds that*

$$\frac{1}{\sqrt{C_1 p}}\|\nabla \widehat{L}(W)\|_F \le \widehat{G}(W) \le \widehat{L}(W) \wedge 1,$$

where C_1 is the constant from Lemma 5. Additionally,

$$\|\nabla \widehat{L}(W) - \nabla \widehat{L}(V)\|_F \le C_1 p\left(1 + \frac{H}{\sqrt{m}}\right)\|W - V\|_2.$$

Proof Since a good run occurs, all the events in Lemma 5 hold. We thus have

$$\begin{aligned}
\left\|\nabla\widehat{L}(W)\right\|_F &= \left\|\frac{1}{n}\sum_{i=1}^{n} g(y_i f(x_i;W)) y_i \nabla f(x_i;W)\right\|_F \\
&\overset{(i)}{\le} \frac{1}{n}\sum_{i=1}^{n} g(y_i f(x_i;W)) \left\|\nabla f(x_i;W)\right\|_F \\
&\overset{(ii)}{\le} \frac{\sqrt{C_1 p}}{n}\sum_{i=1}^{n} g(y_i f(x_i;W)) = \sqrt{C_1 p}\,\widehat{G}(W) \\
&\overset{(iii)}{\le} \frac{\sqrt{C_1 p}}{n}\sum_{i=1}^{n} \min(\ell(y_i f(x_i;W)),1) \\
&\overset{(iv)}{\le} \sqrt{C_1 p}(\widehat{L}(W)\wedge 1).
\end{aligned}$$

In (i) we have used Jensen's inequality. In (ii) we have used that ϕ is 1-Lipschitz so that $\|\nabla f(x_i;W)\|_F^2 = \left\|D_i^W a x_i^\top\right\|_F^2 = \left\|D_i^W a\right\|_2^2 \|x_i\|_2^2 \le C_1 p$ by Event (E.1), where $D_i^W = D_{x_i}^W$ is defined in Equation (3). In (iii) we use that $0 \le g(z) \le 1 \wedge \ell(z)$. In (iv) we use Jensen's inequality since $z \mapsto \min\{z,1\}$ is a concave function.

Next we show that the loss has Lipschitz gradients. First, we have the decomposition

$$\begin{aligned}
\|\nabla\widehat{L}(W) - \nabla\widehat{L}(V)\|_F &= \left\|\frac{1}{n}\sum_{i=1}^{n} [g(y_i f(x_i;W)) y_i \nabla f(x_i;W) - g(y_i f(x_i;V)) y_i \nabla f(x_i;V)]\right\|_F \\
&\le \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(x_i;W)\|_F |g(y_i f(x_i;W)) - g(y_i f(x_i;V))| \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(x_i;W) - \nabla f(x_i;V)\|_F \\
&\overset{(i)}{\le} \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(x_i;W)\|_F |f(x_i;W) - f(x_i;V)| \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(x_i;W) - \nabla f(x_i;V)\|_F. \qquad (4)
\end{aligned}$$

In (i), we use that $g = -\ell'$ (the negative derivative of the logistic loss) is 1-Lipschitz. Therefore, to show that the loss has Lipschitz gradients, it suffices to show that both the network and the gradient of the network are Lipschitz with respect to the first layer weights. We first show that the network is Lipschitz with respect to the network parameters:

$$\begin{aligned}|f(x;W)-f(x;V)|^2 &= \left|\sum_{j=1}^m a_j[\phi(\langle w_j,x\rangle)-\phi(\langle v_j,x\rangle)]\right|^2\\ &\le \left(\sum_{j=1}^m a_j^2\right)\cdot\sum_{j=1}^m|\phi(\langle w_j,x\rangle)-\phi(\langle v_j,x\rangle)|^2\\ &\le \sum_{j=1}^m|\langle w_j,x\rangle-\langle u_j,x\rangle|^2\\ &= \|(W-V)x\|^2\\ &\le \|x\|^2\|W-V\|_2^2. \end{aligned}\tag{5}$$

As for the gradients of the network, again recalling the D_x^W notation from Equation (3), we have

$$\begin{aligned}\|\nabla f(x;W)-\nabla f(x;V)\|_F^2 &= \|(D_x^W-D_x^V)ax^T\|^2\\ &\le \|x\|^2\|(D_x^W-D_x^V)a\|^2\\ &= \|x\|^2\sum_{j=1}^m a_j^2[\phi'(\langle w_j,x\rangle)-\phi'(\langle v_j,x\rangle)]^2\\ &\le \|x\|^2\cdot\frac{H^2}{m}\sum_{j=1}^m|\langle w_j,x\rangle-\langle v_j,x\rangle|^2\\ &= H^2\|x\|^2\cdot\frac{1}{m}\|(W-V)x\|^2\\ &\le \frac{H^2}{m}\|x\|^4\|W-V\|_2^2. \end{aligned}\tag{6}$$

Continuing from (4), we have

$$\begin{aligned}\|\nabla\widehat{L}(W)-\nabla\widehat{L}(V)\|_F &\le \frac{1}{n}\sum_{i=1}^n\|\nabla f(x_i;W)\|_F|f(x_i;W)-f(x_i;V)|\\ &\quad+\frac{1}{n}\sum_{i=1}^n\|\nabla f(x_i;W)-\nabla f(x_i;V)\|_F\\ &\overset{(i)}{\le} \sqrt{C_1p}\cdot\frac{1}{n}\sum_{i=1}^n|f(x_i;W)-f(x_i;V)|+\frac{C_1Hp}{\sqrt{m}}\|W-V\|_2\\ &\overset{(ii)}{\le} C_1p\left(1+\frac{H}{\sqrt{m}}\right)\|W-V\|_2. \end{aligned}\tag{7}$$

In (i) we use that ϕ being 1-Lipschitz implies $\|\nabla f(x_i;W)\|_F=\|x_i\|\|D_i^Wa\|\le\sqrt{C_1p}$ for the first term, and (6) together with (E.1). In (ii), we use (5) and (E.1). ■

A.2.3. PROOF OF LEMMA 9

Finally, we prove Lemma 9 that bounds the correlation between the gradients.

Lemma 19 *Let $C_1 > 1$ be the constant from Lemma 5. For a γ-leaky, H-smooth activation ϕ, on a good run, we have the following.*

(a) For any $i, k \in [n]$, $i \neq k$, and any $W \in \mathbb{R}^{m \times d}$, we have

$$|\langle \nabla f(x_i; W), \nabla f(x_k; W) \rangle| \leq C_1 \left(\|\mu\|^2 + \sqrt{p \log(n/\delta)} \right).$$

(b) For any $i \in [n]$ and any $W \in \mathbb{R}^{m \times d}$, we have

$$\frac{\gamma^2 p}{C_1} \leq \|\nabla f(x_i; W)\|_F^2 \leq C_1 p.$$

Proof Recall the notation $D_i^W := \mathrm{diag}(\phi'(\langle w_j, x_i \rangle) \in \mathbb{R}^{m \times m}$. By definition,

$$\begin{aligned}
\langle \nabla f(x_i; W), \nabla f(x_k; W) \rangle &= \mathrm{tr}(x_i a^\top D_i^W D_k^W a x_k^\top) \\
&= \mathrm{tr}\left(x_i^\top x_k a^\top D_i^W D_k^W a \right) \\
&= \langle x_i, x_k \rangle a^\top D_i^W D_k^W a \\
&= \langle x_i, x_k \rangle \sum_{j=1}^m a_j^2 \phi'(\langle w_j, x_i \rangle) \phi'(\langle w_j, x_k \rangle) \\
&= \langle x_i, x_k \rangle \cdot \frac{1}{m} \sum_{j=1}^m \phi'(\langle w_j, x_i \rangle) \phi'(\langle w_j, x_k \rangle). \qquad (8)
\end{aligned}$$

Since a good run occurs, all the events in Lemma 5 hold. We can therefore bound,

$$|\langle \nabla f(x_i; W), \nabla f(x_k; W) \rangle| \overset{(i)}{\leq} |\langle x_i, x_k \rangle| \overset{(ii)}{\leq} C_1 \left(\|\mu\|^2 + \sqrt{p \log(n/\delta)} \right).$$

Inequality (i) uses that $|\phi'(z)| \leq 1$, while inequality (ii) uses Event (E.2) from Lemma 5. This completes the proof for part (a). For part (b), we continue from (8) to get

$$\|\nabla f(x_i; W)\|_F^2 = \|x_i\|^2 \cdot \frac{1}{m} \sum_{j=1}^m \phi'(\langle w_j, x_i \rangle)^2.$$

By the assumption on ϕ, we know $\phi'(z) \geq \gamma > 0$ for every $t \in \mathbb{R}$. Now we can use Lemma 5, which states that $p/C_1 \leq \|x_i\|^2 \leq C_1 p$ for all i. In particular, we have

$$\frac{p}{C_1} \cdot \gamma^2 \leq \|x_i\|^2 \cdot \frac{1}{m} \sum_{j=1}^m \phi'(\langle w_j, x_i \rangle)^2 = \|\nabla f(x_i; W)\|_F^2 \leq C_1 p.$$

■

A.3. Proof of Lemma 10

Let us restate and prove the lemma. Recall that (x, y) are independent test samples.

Lemma 20 *Let $C_1 > 1$ be the constant from Lemma 1. For a γ-leaky, H-smooth activation ϕ, on a good run, we have for any $t \geq 0$ and $(x, y) \in \mathbb{R}^p \times \{\pm 1\}$, and for each $i = 1, \ldots, n$, there exist $\xi_i = \xi(W^{(t)}, x_i, x) \in [\gamma^2, 1]$, such that*

$$y[f(x; W^{(t+1)}) - f(x; W^{(t)})] \geq \frac{\alpha}{n} \sum_{i=1}^n g_i^{(t)} \left[\xi_i \langle y_i x_i, yx \rangle - \frac{HC_1^2 p^2 \alpha}{2\sqrt{m}} \right],$$

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

Proof First, note that since a good run occurs, Lemma 7 implies

$$\left| f(x; W^{(t+1)}) - f(x; W^{(t)}) - \langle \nabla f(x; W^{(t)}), W^{(t+1)} - W^{(t)} \rangle \right| \leq \frac{HC_1 p}{2\sqrt{m}} \left\| W^{(t+1)} - W^{(t)} \right\|_2^2.$$

In particular, we have for $y \in \{\pm 1\}$,

$$y[f(x; W^{(t+1)}) - f(x; W^{(t)})] \geq y \left[\langle \nabla f(x; W^{(t)}), W^{(t+1)} - W^{(t)} \rangle \right] - \frac{HC_1 p}{2\sqrt{m}} \left\| W^{(t+1)} - W^{(t)} \right\|_2^2. \tag{9}$$

We can therefore calculate

$$\begin{aligned}
y[f(x; W^{(t+1)}) - f(x; W^{(t)})] &\overset{(i)}{\geq} y \left[\langle \nabla f(x; W^{(t)}), W^{(t+1)} - W^{(t)} \rangle \right] - \frac{HC_1 p}{2\sqrt{m}} \left\| W^{(t+1)} - W^{(t)} \right\|_2^2 \\
&= y \left[\frac{\alpha}{n} \sum_{i=1}^n g_i^{(t)} \langle y_i \nabla f(x; W^{(t)}), \nabla f(x_i; W^{(t)}) \rangle \right] \\
&\quad - \frac{HC_1 p \alpha^2}{2\sqrt{m}} \left\| \nabla \widehat{L}(W^{(t)}) \right\|_2^2 \\
&\overset{(ii)}{\geq} \left[\frac{\alpha}{n} \sum_{i=1}^n g_i^{(t)} \langle y \nabla f(x; W^{(t)}), y_i \nabla f(x_i; W^{(t)}) \rangle \right] \\
&\quad - \frac{HC_1^2 p^2 \alpha^2}{2\sqrt{m}} \widehat{G}(W^{(t)}) \\
&= \alpha \left[\frac{1}{n} \sum_{i=1}^n g_i^{(t)} \xi_i \langle y_i x_i, yx \rangle - \frac{HC_1^2 p^2 \alpha}{2\sqrt{m}} \widehat{G}(W^{(t)}) \right].
\end{aligned}$$

The inequality (i) follows by (9), while (ii) uses Lemma 8. The last equality follows by defining

$$\xi_i = \xi(W^{(t)}, x, x_i) := \frac{1}{m} \sum_{j=1}^m \phi'(\langle w_j^{(t)}, x_i \rangle) \cdot \phi'(\langle w_j^{(t)}, x \rangle),$$

and re-using the identity (8) and using the fact that $\phi'(z) \in [\gamma, 1]$ for all $z \in \mathbb{R}$. The result follows by recalling the definition $\widehat{G}(W^{(t)}) = \frac{1}{n} \sum_{i=1}^n g_i^{(t)}$. ■

A.4. Proof of Lemma 11

Let us first restate the lemma.

Lemma 21 *For a γ-leaky, H-smooth activation ϕ, there is an absolute constant $C_r = 16C_1^2/\gamma^2$ such that on a good run, provided $C > 1$ is sufficiently large, we have for all $t \geq 0$,*

$$\max_{i,j\in[n]} \frac{g_i^{(t)}}{g_j^{(t)}} \leq C_r.$$

Before proceeding with the proof of Lemma 11, we introduce the following fact which will be used in our proof.

Fact 22 *For any $z_1, z_2 \in \mathbb{R}$,*

$$\frac{g(z_1)}{g(z_2)} \leq \max\left(2, 2\frac{\exp(-z_1)}{\exp(-z_2)}\right).$$

Proof By definition, $g(z) = 1/(1+\exp(z))$. Note that g is strictly decreasing, non-negative, and bounded from above by one. Further, one has the inequalities

$$\frac{1}{2}\exp(-z) \leq g(z) \leq \exp(-z) \quad \forall z \geq 0.$$

We do a case-by-case analysis on the signs of the z_i.

- If $z_1 \leq 0$ and $z_2 \leq 0$, then since $g(z_1) \leq 1$ and $g(z_2) \geq 1/2$, it holds that $g(z_1)/g(z_2) \leq 2$.
- If $z_1, z_2 \geq 0$, then since $g(z_1) \leq \exp(-z_1)$ and $g(z_2) \geq {}^1\!/_2 \exp(-z_2)$, we have $g(z_1)/g(z_2) \leq 2\exp(-z_1)/\exp(-z_2)$.
- If $z_1 \geq 0$ and $z_2 \leq 0$, then $g(z_1)/g(z_2) \leq 2$.
- If $z_1 \leq 0$ and $z_2 \geq 0$, then $g(z_1)/g(z_2) \leq 2/\exp(-z_2) \leq 2\exp(-z_1)/\exp(-z_2)$.

This proves the upper bound of $g(z_1)/g(z_2)$. ■

We now proceed with the proof of the loss ratio bound.

Proof [Proof of Lemma 11] In order to show that the ratio of the $g(\cdot)$ losses is bounded, it suffices to show that the ratio of exponential losses $\exp(-(\cdot))$ is bounded, since by Fact 22,

$$\max_{i,j=1,\dots,n} \frac{g(y_i f(x_i; W^{(t)}))}{g(y_j f(x_j; W^{(t)}))} \leq \max\left(2, 2\cdot \max_{i,j=1,\dots,n} \frac{\exp(-y_i f(x_i; W^{(t)}))}{\exp(-y_j f(x_j; W^{(t)}))}\right). \tag{10}$$

Thus in the remainder of the proof we will show that the ratio of the exponential losses is bounded by an absolute constant. To see the claim at iteration 0, since ϕ is 1-Lipschitz and $\phi(0) = 0$, we have by Cauchy–Schwarz,

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

Since a good run occurs, all the events in Lemma 5 and Lemma 4 hold. In particular, we have $\|W^{(0)}\|_2 \leq C_0 \omega_{\text{init}}(\sqrt{m}+\sqrt{p})$ and $\|x_i\| \leq \sqrt{C_1 p}$ for all $i \in [n]$. We therefore have the bound,

$$2C_0 \omega_{\text{init}} \sqrt{C_1 p}(\sqrt{m}+\sqrt{p}) \overset{(i)}{\leq} \frac{2C_0\sqrt{C_1}\alpha\sqrt{p}(\sqrt{m}+\sqrt{p})}{\sqrt{mp}} \overset{(ii)}{\leq} \frac{2C_0\sqrt{C_1}}{Cp^2}\left(1+\sqrt{\frac{p}{m}}\right) \overset{(iii)}{\leq} 1,$$

where inequality (i) uses Assumption (A6), inequality (ii) uses Assumption (A5), and the final inequality (iii) follows by taking $C > 1$ large enough. We thus have for all $i \in [n]$,

$$|f(x_i; W^{(0)})| \leq \|W^{(0)}\|_2 \|x_i\| \leq 2C_0 \omega_{\text{init}} \sqrt{C_1 p}(\sqrt{m}+\sqrt{p}) \leq 1. \tag{11}$$

Thus,

$$\max_{i,j=1,\ldots,n} \frac{\exp(-y_i f(x_i; W^{(0)}))}{\exp(-y_j f(x_j; W^{(0)}))} \leq \exp(2). \tag{12}$$

We now claim by induction that for all $t \geq 0$,

$$\max_{i,j=1,\ldots,n} \frac{\exp(-y_i f(x_i; W^{(t)}))}{\exp(-y_j f(x_j; W^{(t)})} \leq \frac{8C_1^2}{\gamma^2}.$$

The base case $t = 0$ holds by (12) and since $C_1 > 1$. Assume now the result holds at time t and consider the case $t+1$. Without loss of generality, it suffices to prove that the ratio of the exponential loss for the first sample to the exponential loss for the second sample is bounded by $8C_1^2/\gamma^2$. To this end, let us denote

$$A_t := \frac{\exp(-y_1 f(x_1; W^{(t)}))}{\exp(-y_2 f(x_2; W^{(t)}))}.$$

Since the induction hypothesis holds at time t, A_t is at most $8C_1^2/\gamma^2$. We want to show $A_{t+1} \leq 8C_1^2/\gamma^2$. To do so, we calculate the exponential loss ratio between two samples at time $t+1$ in terms of the exponential loss ratio at time t. Recalling the notation $g_i^{(t)} := g(y_i f(x_i; W^{(t)}))$, we can calculate,

$$\begin{aligned}
A_{t+1} &= \frac{\exp(-y_1 f(x_1; W^{(t+1)}))}{\exp(-y_2 f(x_2; W^{(t+1)}))} \\
&= \frac{\exp\left(-y_1 f_1\left(W^{(t)} - \alpha \nabla \widehat{L}(W^{(t)})\right)\right)}{\exp\left(-y_2 f_2\left(W^{(t)} - \alpha \nabla \widehat{L}(W^{(t)})\right)\right)} \\
&\overset{(i)}{\le} \frac{\exp\left(-y_1 f\left(x_1; W^{(t)}\right) + y_1 \alpha \left\langle \nabla f(x_1; W^{(t)}), \nabla \widehat{L}(W^{(t)}) \right\rangle\right)}{\exp\left(-y_2 f\left(x_2; W^{(t)}\right) + y_2 \alpha \left\langle \nabla f(x_2; W^{(t)}), \nabla \widehat{L}(W^{(t)}) \right\rangle\right)} \exp\left(\frac{HC_1 p\alpha^2}{\sqrt{m}} \|\nabla \widehat{L}(W^{(t)})\|^2\right) \\
&\overset{(ii)}{=} A_t \cdot \frac{\exp\left(y_1 \alpha \left\langle \nabla f(x_1; W^{(t)}), \nabla \widehat{L}(W^{(t)}) \right\rangle\right)}{\exp\left(y_2 \alpha \left\langle \nabla f(x_2; W^{(t)}), \nabla \widehat{L}(W^{(t)}) \right\rangle\right)} \exp\left(\frac{HC_1 p\alpha^2}{\sqrt{m}} \|\nabla \widehat{L}(W^{(t)})\|^2\right) \\
&= A_t \cdot \frac{\exp\left(-\frac{\alpha}{n} \sum_{k=1}^n y_1 y_k g_k^{(t)} \langle \nabla f(x_1; W^{(t)}), \nabla f(x_k; W^{(t)}) \rangle\right)}{\exp\left(-\frac{\alpha}{n} \sum_{k=1}^n y_2 y_k g_k^{(t)} \langle \nabla f(x_2; W^{(t)}), \nabla f(x_k; W^{(t)}) \rangle\right)} \exp\left(\frac{HC_1 p\alpha^2}{\sqrt{m}} \|\nabla \widehat{L}(W^{(t)})\|^2\right) \\
&= A_t \cdot \exp\left(-\frac{\alpha}{n}\left(g_1^{(t)} \|\nabla f(x_1; W^{(t)})\|_F^2 - g_2^{(t)} \|\nabla f(x_2; W^{(t)})\|_F^2\right)\right) \\
&\qquad \times \frac{\exp\left(-\frac{\alpha}{n} \sum_{k>1} y_1 y_k g_k^{(t)} \langle \nabla f(x_1; W^{(t)}), \nabla f(x_k; W^{(t)}) \rangle\right)}{\exp\left(-\frac{\alpha}{n} \sum_{k\neq 2} y_2 y_k g_k^{(t)} \langle \nabla f(x_2; W^{(t)}), \nabla f(x_k; W^{(t)}) \rangle\rangle\right)} \\
&\qquad \times \exp\left(\frac{HC_1 p\alpha^2}{\sqrt{m}} \|\nabla \widehat{L}(W^{(t)})\|^2\right). \qquad (13)
\end{aligned}$$

Inequality (i) uses Lemma 7 and Event (E.1) which ensures that $\|x_i\|^2 \le C_1 p$, and (ii) uses that A_t is the ratio of the exponential losses. We now proceed to bound each of the three terms in the product separately. For the first term, by Part (b) of Lemma 9, we have for any $i \in [n]$,

$$\frac{\gamma^2 p}{C_1} \le \|\nabla f(x_i; W^{(t)})\|_F^2 \le C_1 p. \qquad (14)$$

Therefore, we have

$$\begin{aligned}
&\exp\left(-\frac{\alpha}{n}\left(g_1^{(t)} \|\nabla f(x_1; W^{(t)})\|_F^2 - g_2^{(t)} \|\nabla f(x_2; W^{(t)})\|_F^2\right)\right) \\
&\quad = \exp\left(-\frac{g_2^{(t)} \alpha}{n}\left(\frac{g_1^{(t)}}{g_2^{(t)}} \|\nabla f(x_1; W^{(t)})\|_F^2 - \|\nabla f(x_2; W^{(t)})\|_F^2\right)\right) \\
&\quad \overset{(i)}{\le} \exp\left(-\frac{g_2^{(t)} \alpha}{n}\left(\frac{g_1^{(t)}}{g_2^{(t)}} \cdot \frac{\gamma^2 p}{C_1} - C_1 p\right)\right) \\
&\quad = \exp\left(-\frac{g_2^{(t)} \alpha \gamma^2 p}{C_1 n}\left(\frac{g_1^{(t)}}{g_2^{(t)}} - \frac{C_1^2}{\gamma^2}\right)\right). \qquad (15)
\end{aligned}$$

Inequality (i) uses (14). This bounds for the first term in (13).

For the second term, we again use Lemma 9: we have for any $i \neq k$,

$$|\langle \nabla f(x_i; W), \nabla f(x_k; W)\rangle| \leq C_1 \left(\|\mu\|^2 + \sqrt{p \log(n/\delta)} \right). \tag{16}$$

This allows for us to bound,

$$\begin{aligned}
&\frac{\exp\left(-\frac{\alpha}{n}\sum_{k>1} y_1 y_k g_k^{(t)} \langle \nabla f(x_1; W^{(t)}), \nabla f(x_k; W^{(t)})\rangle\right)}{\exp\left(-\frac{\alpha}{n}\sum_{k\neq 2} y_2 y_k g_k^{(t)} \langle \nabla f(x_2; W^{(t)}), \nabla f(x_k; W^{(t)})\rangle\rangle\right)} \\
&\overset{(i)}{\leq} \exp\left(\frac{\alpha}{n}\sum_{k\neq 1} g_k^{(t)} |\langle \nabla f(x_1; W^{(t)}), \nabla f(x_k; W^{(t)})\rangle| + \frac{\alpha}{n}\sum_{k\neq 2} g_k^{(t)} |\langle \nabla f(x_2; W^{(t)}), \nabla f(x_k; W^{(t)})\rangle| \right) \\
&\overset{(ii)}{\leq} \exp\left(\frac{\alpha}{n}\sum_{k\neq 1} g_k^{(t)} \cdot C_1 \left(\|\mu\|^2 + \sqrt{p\log(n/\delta)}\right) + \frac{\alpha}{n}\sum_{k\neq 2} g_k^{(t)} \cdot C_1 \left(\|\mu\|^2 + \sqrt{p\log(n/\delta)}\right) \right) \\
&\overset{(iii)}{\leq} \exp\left(2\frac{\alpha}{n}\sum_{k=1}^{n} g_k^{(t)} \cdot C_1 \left(\|\mu\|^2 + \sqrt{p\log(n/\delta)}\right) \right) \\
&= \exp\left(2C_1 \alpha \left(\|\mu\|^2 + \sqrt{p\log(n/\delta)}\right) \widehat{G}(W^{(t)}) \right).
\end{aligned} \tag{17}$$

Inequality (i) uses the triangle inequality. Inequality (ii) uses that $g_k^{(t)} \geq 0$ for all $k \in [n]$ and eq. (16). Inequality (iii) again uses that $g_k^{(t)} \geq 0$.

Finally, for the third term of (13), we have

$$\exp\left(\frac{HC_1 p\alpha^2}{\sqrt{m}} \|\nabla \widehat{L}(W^{(t)})\|^2 \right) \overset{(i)}{\leq} \exp\left(\frac{HC_1^2 p^2 \alpha^2}{\sqrt{m}} \widehat{G}(W^{(t)}) \right) \overset{(ii)}{\leq} \exp\left(\alpha \sqrt{p} \widehat{G}(W^{(t)}) \right). \tag{18}$$

Inequality (i) uses Lemma 8, while (ii) uses that for $C > 1$ sufficiently large, by Assumption (A5) we have $HC_1^2 p^2 \alpha / \sqrt{m} \leq \sqrt{p}$. Putting (15), (17) and (18) into (13), we get

$$\begin{aligned}
A_{t+1} &\leq A_t \cdot \exp\left(-\frac{g_2^{(t)} \alpha \gamma^2 p}{C_1 n} \left(\frac{g_1^{(t)}}{g_2^{(t)}} - \frac{C_1^2}{\gamma^2} \right) \right) \\
&\qquad \times \exp\left(2C_1\alpha \left(\|\mu\|^2 + \sqrt{p\log(n/\delta)} \right) \widehat{G}(W^{(t)}) \right) \cdot \exp\left(\alpha\sqrt{p} \widehat{G}(W^{(t)}) \right) \\
&\leq A_t \cdot \exp\left(-\frac{g_2^{(t)} \alpha \gamma^2 p}{C_1 n} \left(\frac{g_1^{(t)}}{g_2^{(t)}} - \frac{C_1^2}{\gamma^2} \right) \right) \\
&\qquad \times \exp\left(2C_1\alpha \left(\|\mu\|^2 + 2\sqrt{p\log(n/\delta)} \right) \widehat{G}(W^{(t)}) \right).
\end{aligned} \tag{19}$$

We now consider two cases: in the first case, the ratio $g_1^{(t)}/g_2^{(t)}$ is relatively small, in this case we will show that the exponential loss ratio will not grow too much for small enough step-size α. In the second case, if the ratio $g_1^{(t)}/g_2^{(t)}$ is relatively large, then the first exponential term in (19) will dominate and cause the exponential loss ratio to contract.

Case 1 ($g_1^{(t)}/g_2^{(t)} \leq \frac{2C_1^2}{\gamma^2}$): Continuing from (19), we have

$$\begin{aligned}
A_{t+1} &\leq A_t \exp\left(-\frac{g_2^{(t)}\alpha\gamma^2 p}{C_1 n}\left(\frac{g_1^{(t)}}{g_2^{(t)}} - \frac{C_1^2}{\gamma^2}\right)\right)\exp\left(2C_1\alpha\left(\|\mu\|^2 + 2\sqrt{p\log(n/\delta)}\right)\widehat{G}(W^{(t)})\right) \\
&\overset{(i)}{\leq} A_t \exp\left(\frac{g_2^{(t)} C_1 \alpha p}{n}\right)\exp\left(2C_1\alpha\left(\|\mu\|^2 + 2\sqrt{p\log(n/\delta)}\right)\widehat{G}(W^{(t)})\right) \\
&\overset{(ii)}{\leq} A_t \exp\left(\frac{C_1\alpha p}{n}\right)\exp\left(2C_1\alpha\left(\|\mu\|^2 + 2\sqrt{p\log(n/\delta)}\right)\widehat{G}(W^{(t)})\right) \\
&\overset{(iii)}{\leq} 2\frac{g_1^{(t)}}{g_2^{(t)}}\exp\left(\frac{C_1\alpha p}{n}\right)\exp\left(2C_1\alpha\left(\|\mu\|^2 + 2\sqrt{p\log(n/\delta)}\right)\right) \\
&= 2\frac{g_1^{(t)}}{g_2^{(t)}}\exp\left(C_1\alpha\left(\frac{p}{n} + 2\|\mu\|^2 + 4\sqrt{p\log(n/\delta)}\right)\right) \\
&\overset{(iv)}{\leq} \frac{4C_1^2}{\gamma^2}\exp\left(C_1\alpha\left(\frac{p}{n} + 2\|\mu\|^2 + 4\sqrt{p\log(n/\delta)}\right)\right) \\
&\overset{(v)}{\leq} \frac{4C_1^2\exp(1/8)}{\gamma^2} \leq \frac{8C_1^2}{\gamma^2}.
\end{aligned}$$

In (i) we use that $g_i^{(t)} \geq 0$, while in (ii) we use that $|g(z)| \leq 1$. In (iii), we use Fact 22 and that $\widehat{G}(W) \leq 1$. In (iv), we use the Case 1 assumption that $g_1^{(t)}/g_2^{(t)} \leq 2C_1^2/\gamma^2$. Finally, in (v), we take $C > 1$ sufficiently large so that by the upper bound on the step-size given in Assumption (A5), we have,

$$C_1\alpha\left(\frac{p}{n} + 2\|\mu\|^2 + 4\sqrt{p\log(n/\delta)}\right) \leq \frac{1}{Hn} + \frac{6}{C_1 H} \leq \frac{1}{8},$$

where we have used Assumption (A2) and assumed without loss of generality that $H \geq 1$.

Case 2 ($g_1^{(t)}/g_2^{(t)} > \frac{2C_1^2}{\gamma^2}$): Again using the bound in (19), we have that

$$\begin{aligned}
&A_{t+1}\\
&\le A_t \cdot \exp\left(-\frac{g_2^{(t)}\alpha\gamma^2 p}{C_1 n}\left(\frac{g_1^{(t)}}{g_2^{(t)}} - \frac{C_1^2}{\gamma^2}\right)\right)\cdot \exp\left(2C_1\alpha\left(\|\mu\|^2 + 2\sqrt{p\log(n/\delta)}\right)\widehat{G}(W^{(t)})\right)\\
&= A_t \exp\left(-\frac{g_2^{(t)}\alpha\gamma^2 p}{C_1 n}\left(\frac{g_1^{(t)}}{g_2^{(t)}} - \frac{C_1^2}{\gamma^2}\right)\right)\\
&\qquad \times \exp\left(2C_1\alpha\left(\|\mu\|^2 + 2\sqrt{p\log(n/\delta)}\right)g_2^{(t)}\cdot\frac{1}{n}\sum_{i=1}^n \frac{-g_i^{(t)}}{g_2^{(t)}}\right)\\
&\overset{(i)}{\le} A_t \exp\left(-\frac{g_2^{(t)}\alpha\gamma^2 p}{C_1 n}\left(\frac{g_1^{(t)}}{g_2^{(t)}} - \frac{C_1^2}{\gamma^2}\right)\right)\\
&\qquad \times \exp\left(2g_2^{(t)}C_1\alpha\left(\|\mu\|^2 + 2\sqrt{p\log(n/\delta)}\right)\cdot\max\left\{2, \frac{16C_1^2}{\gamma^2}\right\}\right)\\
&\overset{(ii)}{=} A_t \exp\left(-g_2^{(t)}\alpha\left[\frac{\gamma^2 p}{C_1 n}\left(\frac{g_1^{(t)}}{g_2^{(t)}} - \frac{C_1^2}{\gamma^2}\right) - \frac{32C_1^3}{\gamma^2}\left(\|\mu\|^2 + 2\sqrt{p\log(n/\delta)}\right)\right]\right)\\
&\overset{(iii)}{\le} A_t \exp\left(-g_2^{(t)}\alpha\left[\frac{C_1 p}{n} - \frac{32C_1^3}{\gamma^2}\left(\|\mu\|^2 + 2\sqrt{p\log(n/\delta)}\right)\right]\right)\\
&\overset{(iv)}{\le} A_t \le \frac{8C_1^2}{\gamma^2}.
\end{aligned}$$

In (i) we use the induction hypothesis that $A_t \le 8C_1^2/\gamma^2$ together with Fact 22. Equality (ii) uses that $C_1 > 1$ and that $\gamma < 1$. In (iii), we use the Case 2 assumption that $g_1^{(t)}/g_2^{(t)} \ge 2C_1^2/\gamma^2$. Finally, in (iv), we use Assumption (A2) so that we have $p \ge Cn\|\mu\|^2 \ge \frac{128C_1^2}{\gamma^2}n\|\mu\|^2$ and that $p \ge Cn^2\log(n/\delta) \ge \left(\frac{128C_1^2}{\gamma^2}n\sqrt{\log(n/\delta)}\right)^2$ and also the fact that $g_2^{(t)} \ge 0$.

This completes the induction that for all times $t \ge 0$, the ratio of the exponential losses is at most $8C_1^2/\gamma^2$. Using (10) completes the proof. ■

A.5. Proof of Lemma 12

We remind the reader of the statement of Lemma 12.

Lemma 23 *There is an absolute constant $C_2 > 1$ such that for $C > 1$ sufficiently large, on a good run we have that for all $t \ge 0$,*

$$\|W^{(t)}\|_F \le \|W^{(0)}\|_F + C_2\alpha\sqrt{\frac{p}{n}}\sum_{s=0}^{t-1}\widehat{G}(W^{(s)}).$$

Proof By the triangle inequality we have that

$$\|W^{(t)}\|_F = \left\|W^{(0)} + \alpha\sum_{s=0}^{t-1}\nabla\widehat{L}(W^{(s)})\right\|_F \le \|W^{(0)}\|_F + \alpha\sum_{s=0}^{t-1}\|\nabla\widehat{L}(W^{(s)})\|_F. \tag{20}$$

Now observe that

$$\begin{aligned}
&\|\nabla \widehat{L}(W^{(s)})\|_F^2 \\
&\quad = \frac{1}{n^2}\left\|\sum_{i=1}^n g_i^{(s)} y_i \nabla f(x_i; W^{(s)})\right\|_F^2 \\
&\quad = \frac{1}{n^2}\left[\sum_{i=1}^n \left(g_i^{(s)}\right)^2 \left\|\nabla f(x_i; W^{(s)})\right\|_F^2 + \sum_{i\neq j\in[n]} g_i^{(s)} g_j^{(s)} y_i y_j \langle \nabla f(x_i; W^{(s)}), \nabla f(x_j; W^{(s)})\rangle\right] \\
&\quad \leq \frac{1}{n^2}\left[\sum_{i=1}^n \left(g_i^{(s)}\right)^2 \left\|\nabla f(x_i; W^{(s)})\right\|_F^2 + \sum_{i\neq j\in[n]} g_i^{(s)} g_j^{(s)} \left|\langle \nabla f(x_i; W^{(s)}), \nabla f(x_j; W^{(s)})\rangle\right|\right] \\
&\quad \overset{(i)}{\leq} \frac{C_1}{n^2}\left[\sum_{i=1}^n \left(g_i^{(s)}\right)^2 p + \sum_{i\neq j\in[n]} g_i^{(s)} g_j^{(s)} \left(\|\mu\|^2 + \sqrt{p\log(n/\delta)}\right)\right] \\
&\quad \leq \frac{C_1}{n^2}\cdot \max_{k\in[n]} g_k^{(s)} \left[\sum_{i=1}^n g_i^{(s)} p + n\sum_{i=1}^n g_i^{(s)} \left(\|\mu\|^2 + \sqrt{p\log(n/\delta)}\right)\right] \\
&\quad = \frac{C_1}{n^2}\left(p + n\|\mu\|^2 + n\sqrt{p\log(n/\delta)}\right)\cdot \max_{k\in[n]} g_k^{(s)} \left[\sum_{i=1}^n g_i^{(s)}\right],
\end{aligned}$$

where (i) follows by Lemma 9. Now note that since $p \geq Cn\|\mu\|^2$ and $p \geq Cn^2\log(n/\delta)$ by Assumption (A2), we have that,

$$\|\nabla \widehat{L}(W^{(s)})\|_F^2 \leq \frac{3C_1^2 p}{n}\left(\max_{k\in[n]} g_k^{(s)}\right)\widehat{G}(W^{(s)}).$$

Next note that by the loss ratio bound in Lemma 11 we have that

$$\max_{k\in[n]} g_k^{(s)} \leq \frac{C_r}{n}\sum_{i=1}^n g_i^{(s)} = C_r \widehat{G}(W^{(s)}).$$

Plugging this into the previous inequality yields

$$\|\nabla \widehat{L}(W^{(s)})\|_F^2 \leq \frac{3C_1^2 C_r p}{n}\left(\widehat{G}(W^{(s)})\right)^2.$$

Finally, taking square roots, defining $C_2 := \sqrt{3C_1^2 C_r}$ and applying this bound on the norm in Inequality (20) above we conclude that

$$\|W^{(t)}\|_F \leq \|W^{(0)}\|_F + C_2\alpha\sqrt{\frac{p}{n}}\sum_{s=0}^{t-1}\widehat{G}(W^{(s)}),$$

establishing our claim. ■

A.6. Proof of Lemma 13

Let us restate the lemma for the reader's convenience.

Lemma 24 *For a γ-leaky, H-smooth activation ϕ, and for all $C > 1$ sufficiently large, on a good run, for any $t \geq 1$,*

$$\frac{\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\tilde{y}f(x;W^{(t)})]}{\|W^{(t)}\|_F} \geq \frac{\gamma^2\|\mu\|^2\sqrt{n}}{8\max(\sqrt{C_1},C_2)\sqrt{p}},$$

where C_1 and C_2 are the constants from Lemma 5 and Lemma 12, respectively.

Proof Using the refined upper bound for the norm of the weights given in Lemma 12, we have that,

$$\|W^{(t)}\|_F \leq \|W^{(0)}\|_F + C_2\alpha\sqrt{\frac{p}{n}}\sum_{s=0}^{t-1}\widehat{G}(W^{(s)}). \tag{21}$$

To complete the proof, we want to put together the bound for the unnormalized margin on clean samples given by Lemma 10 with the upper bound on the norm given in (21). First, let us recall the definition of the quantity ξ_i first introduced in (the proof of) Lemma 10,

$$\xi_i = \xi(W^{(s)}, x_i, x) = \frac{1}{m}\sum_{j=1}^{m}\phi'(\langle w_j^{(s)}, x\rangle)\phi'(\langle w_j^{(s)}, x_i\rangle) \in [\gamma^2, 1].$$

Since $\xi_i \in [\gamma^2, 1]$ for all $i \in [n]$ and $s \in \{0, 1, \ldots\}$, and $\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\tilde{y}x] = \mu$, by (E.3) and (E.4), we have

$$\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\xi_i\langle y_i x_i, \tilde{y}x\rangle] \geq \begin{cases} \frac{\gamma^2}{2}\|\mu\|^2, & i \in \mathcal{C}, \\ -\frac{3}{2}\|\mu\|^2, & i \in \mathcal{N}. \end{cases} \tag{22}$$

This allows for us to derive a lower bound on the increment of the unnormalized margin, for any $s \geq 0$:

$$\begin{aligned}
&\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\tilde{y}(f(x;W^{(s+1)}) - f(x;W^{(s)}))] \\
&\quad\overset{(i)}{\geq} \frac{\alpha}{n}\sum_{i=1}^{n} g_i^{(s)}\left[\mathbb{E}[\xi_i\langle y_i x_i, \tilde{y}x\rangle] - \frac{HC_1^2p^2\alpha}{2\sqrt{m}}\right] \\
&\quad\overset{(ii)}{\geq} \alpha\left[\frac{1}{n}\sum_{i\in\mathcal{C}} g_i^{(s)}\cdot\frac{\gamma^2}{2}\|\mu\|^2 - \frac{1}{n}\sum_{i\in\mathcal{N}} g_i^{(s)}\cdot\frac{3}{2}\|\mu\|^2 - \frac{HC_1^2p^2\alpha}{2\sqrt{m}}\widehat{G}(W^{(s)})\right] \\
&\quad= \frac{\alpha\gamma^2\|\mu\|^2}{2}\left[\left(1 - \frac{HC_1^2p^2\alpha}{2\gamma^2\|\mu\|^2\sqrt{m}}\right)\widehat{G}(W^{(s)}) - \left(1 + \frac{3}{\gamma^2}\right)\cdot\frac{1}{n}\sum_{i\in\mathcal{N}} g_i^{(s)}\right] \\
&\quad\overset{(iii)}{\geq} \frac{\alpha\gamma^2\|\mu\|^2}{2}\left[\left(1 - 2C_r\eta\left(1 + \frac{3}{\gamma^2}\right) - \frac{HC_1^2p^2\alpha}{2\gamma^2\|\mu\|^2\sqrt{m}}\right)\widehat{G}(W^{(s)})\right] \\
&\quad\overset{(iv)}{\geq} \frac{\alpha\gamma^2\|\mu\|^2}{8}\widehat{G}(W^{(s)}).
\end{aligned} \tag{23}$$

Above, inequality (i) uses Lemma 10, while (ii) uses (22). Inequality (iii) uses the fact that the loss ratio bound given in Lemma 11 implies,

$$\sum_{i\in\mathcal{N}} g_i^{(t)} \le |\mathcal{N}|\cdot \max_i g_i^{(t)} = \frac{|\mathcal{N}|}{n}\sum_{k=1}^{n}\max_i g_i^{(t)} \le C_r\cdot|\mathcal{N}|\cdot\widehat{G}(W^{(t)}) \le 2C_r\eta n\widehat{G}(W^{(t)}). \tag{24}$$

The last inequality (iv) follows by Assumption (A4) so that the noise rate satisfies $\eta \le 1/C \le [8C_r(1+3\gamma^{-2})]^{-1}$, and since the assumption (A5) implies $HC_1^2p^2\alpha/(2\gamma^2\|\mu\|^2) \le 1/4$ for $C>1$ sufficiently large.

We provide one final auxiliary calculation before showing the lower bound on the normalized margin. By Equation (11) we have $|f(x_i;W^{(0)})| \le 1$ for all i. Using the following lower bound on the derivative of the logistic loss, $-\ell'(y_if(x_i;W)) \ge \frac{1}{2}\exp(-|f(x_i;W)|)$, we therefore have

$$\widehat{G}(W^{(0)}) \ge \frac{1}{2}\exp(-1) \ge \frac{1}{6}. \tag{25}$$

Using this along with Lemma 4, we have that

$$\|W^{(0)}\|_F \le 2\omega_{\text{init}}\sqrt{mp} \le 2\alpha \le \alpha\sqrt{C_1p/n}\widehat{G}(W^{(0)}), \tag{26}$$

where we have used the assumption (A6) that $\omega_{\text{init}}\sqrt{mp} \le \alpha$ and that Assumption (A2) implies p/n is larger than some fixed constant. With this in hand, we can calculate a lower bound on the normalized margin as follows. First, note that since $\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\tilde{y}f(x;W^{(0)})] = 0$, we can use (23) to get,

$$\begin{aligned}\frac{\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\tilde{y}f(x;W^{(t)})]}{\|W^{(t)}\|_F} &= \frac{\mathbb{E}[\tilde{y}f(x;W^{(0)})] + \sum_{s=0}^{t-1}\mathbb{E}[\tilde{y}[f(x;W^{(s+1)}) - f(x;W^{(s)})]}{\|W^{(t)}\|_F}\\ &\ge \frac{\alpha\gamma^2\|\mu\|^2\sum_{s=0}^{t-1}\widehat{G}(W^{(s)})}{4\|W^{(t)}\|_F}.\end{aligned} \tag{27}$$

Now consider two disjoint cases.

Case 1 ($\|W^{(t)}\|_F \le 2\|W^{(0)}\|_F$): In this case, by using (27) we have that,

$$\begin{aligned}\frac{\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\tilde{y}f(x;W^{(t)})]}{\|W^{(t)}\|_F} &\ge \frac{\alpha\gamma^2\|\mu\|^2\sum_{s=0}^{t-1}\widehat{G}(W^{(s)})}{8\|W^{(0)}\|_F}\\ &\overset{(i)}{\ge} \frac{\alpha\gamma^2\|\mu\|^2\sum_{s=0}^{t-1}\widehat{G}(W^{(s)})}{8\alpha\sqrt{C_1p/n}\widehat{G}(W^{(0)})}\\ &\overset{(ii)}{\ge} \frac{\gamma^2\|\mu\|^2\sqrt{n}}{8\sqrt{C_1p}}\end{aligned}$$

where (i) uses (21) and (ii) uses that $\sum_{s=0}^{t-1}G(W^{(s)}) \ge G(W^{(0)})$. This completes the proof in this case.

Case 2 ($\|W^{(t)}\|_F > 2\|W^{(0)}\|_F$): By (21), we have the chain of inequalities,

$$2\|W^{(0)}\|_F < \|W^{(t)}\|_F \leq \|W^{(0)}\|_F + C_2\alpha\sqrt{\frac{p}{n}}\sum_{s=0}^{t-1}\widehat{G}(W^{(s)}).$$

In particular, we have $C_2\alpha\sqrt{p/n}\sum_{s=0}^{t-1}\widehat{G}(W^{(s)}) > \|W^{(0)}\|_F$, and so substituting the preceding inequality into (27) we get,

$$\begin{aligned}\frac{\mathbb{E}_{(x,\tilde{y})\sim\tilde{\mathsf{P}}}[\tilde{y}f(x;W^{(t)})]}{\|W^{(t)}\|_F} &\geq \frac{\alpha\gamma^2\|\mu\|^2\sum_{s=0}^{t-1}\widehat{G}(W^{(s)})}{4\|W^{(0)}\|_F + 4C_2\alpha\sqrt{p/n}\sum_{s=0}^{t-1}\widehat{G}(W^{(s)})}\\ &\geq \frac{\alpha\gamma^2\|\mu\|^2\sum_{s=0}^{t-1}\widehat{G}(W^{(s)})}{8C_2\alpha\sqrt{p/n}\sum_{s=0}^{t-1}\widehat{G}(W^{(s)})}\\ &= \frac{\gamma^2\|\mu\|^2\sqrt{n}}{8C_2\sqrt{p}},\end{aligned}$$

completing the proof. ■

A.7. Proof of Lemma 14

Lemma 25 *For a γ-leaky, H-smooth activation ϕ, provided $C > 1$ is sufficiently large, then on a good run we have for all $t \geq 0$,*

$$\|\nabla\widehat{L}(W^{(t)})\|_F \geq \frac{\gamma\|\mu\|}{4}\widehat{G}(W^{(t)}).$$

Moreover, any $T \in \mathbb{N}$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(y_i \neq \mathrm{sgn}(f(x_i;W^{(T-1)}))\big) \leq 2\widehat{G}(W^{(T-1)}) \leq 2\left(\frac{32\widehat{L}(W^{(0)})}{\gamma^2\|\mu\|^2\alpha T}\right)^{1/2}.$$

In particular, for $T \geq 128\widehat{L}(W^{(0)})/\left(\gamma^2\|\mu\|^2\alpha\varepsilon^2\right)$, we have $\widehat{G}(W^{(T-1)}) \leq \varepsilon/2$.

Proof In order to show a lower bound for $\|\nabla\widehat{L}(W^{(t)})\|_F = \sup_{U:\|U\|_F=1}\langle -\nabla\widehat{L}(W^{(t)}), U\rangle$, it suffices to construct a matrix V with Frobenius norm at most one such that $\langle -\nabla\widehat{L}(W^{(t)}), V\rangle$ is bounded from below by a positive constant. To this end, let $V \in \mathbb{R}^{m\times p}$ be the matrix with rows

$$v_j = a_j\mu/\|\mu\|. \tag{28}$$

Then $\|V\|_F = 1$ (since $a_j = \pm 1/\sqrt{m}$), and we have for any $W \in \mathbb{R}^{m\times d}$,

$$\langle\nabla f(x_i;W), V\rangle = \sum_{j=1}^{m} a_j\phi'(\langle w_j, x\rangle)\langle v_j, x\rangle = \left\langle\frac{\mu}{\|\mu\|}, x\right\rangle\frac{1}{m}\sum_{i=1}^{m}\phi'(\langle w_j, x\rangle). \tag{29}$$

Now, by Events (E.3) and (E.4), we have that

$$\begin{cases} y_i\langle\mu, x_i\rangle \geq \frac{1}{2}\|\mu\|^2, & i \in \mathcal{C},\\ |\langle\mu, x_i\rangle| \leq \frac{3}{2}\|\mu\|^2, & i \in \mathcal{N}.\end{cases} \tag{30}$$

Since $\phi'(z) \geq \gamma > 0$ for all z, (29) implies we have the following lower bound for any $W \in \mathbb{R}^{m\times d}$,

$$y_i \langle \nabla f(x_i; W), V\rangle \geq \begin{cases} \frac{\gamma}{2}\|\mu\|, & i \in \mathcal{C}, \\ -\frac{3}{2}\|\mu\|, & i \in \mathcal{N}. \end{cases} \tag{31}$$

This allows for a lower bound on $\langle -\widehat{\nabla} L(W^{(s)}), V\rangle$, since

$$\begin{aligned} \langle -\widehat{\nabla} L(W^{(s)}), V\rangle &= \frac{1}{n}\sum_{i=1}^n g_i^{(s)} y_i \langle \nabla f(x_i; W^{(s)}), V\rangle \\ &\overset{(i)}{\geq} \frac{1}{n}\sum_{i\in\mathcal{C}} g_i^{(s)} \cdot \frac{\gamma}{2}\|\mu\| - \frac{1}{n}\sum_{i\in\mathcal{N}} g_i^{(s)} \cdot \frac{3}{2}\|\mu\| \\ &= \frac{\gamma\|\mu\|}{2}\left[\widehat{G}(W^{(s)}) - \left(1+\frac{3}{\gamma}\right)\frac{1}{n}\sum_{i\in\mathcal{N}} g_i^{(s)}\right] \\ &\overset{(ii)}{\geq} \frac{\gamma\|\mu\|}{2}\left[\widehat{G}(W^{(s)}) - \left(1+\frac{3}{\gamma}\right)\cdot 2C_r\eta\widehat{G}(W^{(s)})\right] \\ &\overset{(iii)}{\geq} \frac{\gamma\|\mu\|}{4}\widehat{G}(W^{(s)}). \end{aligned} \tag{32}$$

Inequality (i) uses (31), while (ii) uses the previously-established inequality (24). Finally, inequality (iii) above uses Assumption (A4) so that the noise rate satisfies $\eta \leq 1/C \leq [4C_r(1+3/\gamma)]^{-1}$. We can therefore derive the following lower bound on the norm of the gradient,

$$\text{for any } t \geq 0, \quad \|\nabla \widehat{L}(W^{(t)})\|_F \geq \langle \nabla \widehat{L}(W^{(t)}), -V\rangle \geq \frac{\gamma\,\|\mu\|\,\widehat{G}(W^{(t)})}{4}. \tag{33}$$

We notice that the inequality of the form $\|\widehat{\nabla} L(W)\| \geq c\widehat{G}(W)$ is a proxy PL inequality, where the proxy loss function is $\widehat{G}(W)$ (Frei and Gu, 2021). We can therefore mimic the smoothness-based proof of Frei and Gu (2021, Theorem 3.1) to show that $\widehat{G}(W^{(T-1)}) \leq \varepsilon$ for $T = \Omega(\varepsilon^{-2})$. By Lemma 8, the loss $\widehat{L}(W)$ has $C_1 p(1+H/\sqrt{m})$-Lipschitz gradients. In particular, we have

$$\widehat{L}(W^{(t+1)}) \leq \widehat{L}(W^{(t)}) - \alpha\|\nabla\widehat{L}(W^{(t)})\|_F^2 + C_1 p \max\left\{1, \frac{H}{\sqrt{m}}\right\}\alpha^2\|\nabla\widehat{L}(W^{(t)})\|_F^2. \tag{34}$$

In particular, since Assumption (A5) requires $\alpha \leq 1/\left(2\max\left\{1, \frac{H}{\sqrt{m}}\right\}C_1^2p^2\right)$, we have that

$$\|\nabla\widehat{L}(W^{(t)})\|_F^2 \leq \frac{2}{\alpha}\left[\widehat{L}(W^{(t+1)}) - \widehat{L}(W^{(t)})\right].$$

Telescoping the above sum and scaling both sides by $1/T$, we get for any $T \geq 1$,

$$\frac{\gamma^2\|\mu\|^2}{16}\frac{1}{T}\sum_{t=0}^{T-1}\widehat{G}(W^{(t)})^2 \overset{(i)}{\leq} \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla\widehat{L}(W^{(t)})\|_F^2 \leq \frac{2\widehat{L}(W^{(0)})}{\alpha T}. \tag{35}$$

Inequality (i) uses the proxy PL inequality (33). Finally, note that since $|\ell''| \leq 1$, an identical calculation to that of (7) shows that the loss $\widehat{G}(W)$ has $C_1 p(1+H/\sqrt{m})$-Lipschitz gradients, and since $\alpha \leq 1/(2\max(1, H/\sqrt{m})C_1 p^2)$, we therefore have

$$\widehat{G}(W^{(t+1)}) - \widehat{G}(W^{(t)}) \leq -\frac{\alpha}{2}\|\nabla\widehat{G}(W^{(t)})\|_F^2.$$

In particular, the loss $\widehat{G}(W^{(t)})$ is a decreasing function of t, and hence $\widehat{G}(W^{(t)})^2$ is a decreasing function of t. Therefore, by (35),

$$\widehat{G}(W^{(T-1)}) = \min_{t<T} \widehat{G}(W^{(t)}) \le \frac{1}{T}\sum_{t=0}^{T-1} \widehat{G}(W^{(t)}) \le \sqrt{\frac{32\widehat{L}(W^{(0)})}{\gamma^2\|\mu\|^2\alpha T}} \le \varepsilon/2, \tag{36}$$

where in the last inequality we use that $T \ge 128\widehat{L}(W^{(0)})/\left(\gamma^2\|\mu\|^2\alpha\varepsilon^2\right)$. The proof is completed by noting that $\mathbb{1}(z \le 0) \le -2\ell'(z)$. ■

Appendix B. Non-NTK results, Proof of Proposition 2

For the reader's convenience, we restate Proposition 2 here.

Proposition 26 *Under the settings of Theorem 1, we have for some absolute constant $C > 1$ with probability at least $1 - 2\delta$ over the random initialization and the draws of the samples,*

$$\frac{\|W^{(1)} - W^{(0)}\|_F}{\|W^{(0)}\|_F} \ge \frac{\gamma\|\mu\|}{C}.$$

Proof We construct a lower bound on $\|W^{(1)} - W^{(0)}\|_F$ using the variational formula for the norm, namely $\|W^{(1)} - W^{(0)}\|_F \ge \langle W^{(1)} - W^{(0)}, V\rangle$ for any matrix V with Frobenius norm at most 1. By definition,

$$\langle W^{(1)} - W^{(0)}, V\rangle = \alpha\langle -\nabla\widehat{L}(W^{(0)}), V\rangle.$$

By Lemmas 4 and 5, a good run occurs with probability at least $1 - 2\delta$. On a good run we can use the results in Lemma 14. In particular, with the choice of V given in eq. (28), we have,

$$\begin{aligned}\|W^{(1)} - W^{(0)}\|_F &\ge \langle W^{(1)} - W^{(0)}, V\rangle \\ &= \alpha\langle -\nabla\widehat{L}(W^{(0)}), V\rangle \\ &\overset{(i)}{\ge} \frac{\alpha\gamma\|\mu\|}{4}\widehat{G}(W^{(0)}) \\ &\overset{(ii)}{\ge} \frac{\alpha\gamma\|\mu\|}{24},\end{aligned}$$

where inequality (i) uses eq. (33) and the last inequality (ii) uses (25). Thus, by Lemma 4, we have

$$\frac{\|W^{(1)} - W^{(0)}\|_F}{\|W^{(0)}\|_F} \ge \frac{\alpha\gamma\|\mu\|}{48\omega_{\text{init}}\sqrt{mp}} \ge \frac{\gamma\|\mu\|}{48},$$

where the last inequality uses Assumption (A6). ■