Benign overfitting of non-smooth neural networks beyond lazy training

Xingyu Xu
Tsinghua University

Yuantao Gu
Tsinghua University

Abstract

Benign overfitting refers to a recently discovered intriguing phenomenon that over-parameterized neural networks, in many cases, can fit the training data perfectly but still generalize well, surprisingly contrary to the traditional belief that overfitting is harmful for generalization. In spite of its surging popularity in recent years, little has been known in the theoretical aspect of benign overfitting of neural networks. In this work, we provide a theoretical analysis of benign overfitting for two-layer neural networks with possibly non-smooth activation function. Without resorting to the popular Neural Tangent Kernel (NTK) approximation, we prove that neural networks can be trained with gradient descent to classify binary-labeled training data perfectly (achieving zero training loss) even in presence of polluted labels, but still generalize well. Our result removes the smoothness assumption in previous literatures and goes beyond the NTK regime; this enables a better theoretical understanding of benign overfitting within a practically more meaningful setting, e.g. with (leaky-)ReLU activation function, small random initialization, and finite network width.

1 INTRODUCTION

In modern data science, neural networks have demonstrated its practical capacity to tackle many complicated tasks that were beyond the reach of classical machine learning methods. However, it remains mysterious why they can work so well (and, in opposition, why they fail in some cases) from a theoretical perspective. Classical theory like universal approximation theorem states that sufficiently large networks can fit any continuous function. One may translate it into a language that is more in spirit of this paper:

- With sufficient overparameterization, with an infinite number of noiseless training samples and with infinite computation power, we can train a neural network that generalizes well.

Unfortunately, this does not fully explain the success of neural networks: it is obvious that a linear interpolator will do the same job under the above idealized setting.

Thus the power of neural networks can be better exhibited only in a more practical setting: noisy and finite samples, with practical optimization algorithms. For example, the counterexample of linear interpolator given above does not generalize well if the training data is noisy. Similar issues occur for many classical learning methods. In fact, traditionally one often needs to cut down the number of parameters to alleviate performance deterioration caused by fitting noises. These had accumulated in the common beliefs that:

- Overparameterized model overfits on noisy data.
- Overfitting is harmful for generalization.

It seems contradictory to the above beliefs that neural networks, as an enormously overparameterized model, simultaneously show impressive generalization performance and strong fitting ability in many scenarios. This is the key observation in the important recent discovery: the *benign overfitting* phenomenon (Belkin et al., 2019). A brief summary of benign overfitting is:

- In many cases, overparameterized neural networks can be trained to zero training loss with simple algorithms like gradient descent, even in presence of noisy samples.
- But they still generalize well (in fact, often achieving state-of-the-art performance).

This indicates a new, unexplored statistical phenomenon contrary to classical beliefs, which can be helpful for understanding the success (and the failure) of neural networks. For this reason, benign overfitting has received surging attention in recent years, to name just a few, Bartlett et al. (2020), Wang et al. (2021), Mei and Montanari (2022), Hastie et al. (2022).

However, theoretical understanding of benign overfitting of neural networks is still limited. Previous literatures have restricted their scopes to the lazy training regime, also known as Neural Tangent Kernel (NTK) regime (Ja-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

cot et al., 2018; Arora et al., 2019b), which demands the weight never moves far from its initial value and often requires incredibly large network width, which does not align very well with the common usage of neural networks in practice. In addition, for technical reasons these results are often based on smooth activation functions, while in practice the non-smooth Rectified Linear Unit (ReLU) family has overwhelming popularity. It is thus important to go beyond these limitations and understand benign overfitting for non-smooth networks beyond lazy training regime.

In this paper we investigate the benign overfitting phenomenon for two-layer neural networks with possibly non-smooth activation functions. For well-separated binary classification problem, we prove that such networks can be trained to zero training loss with gradient descent, hence fitting all labels of the training data perfectly even in presence of adversary label pollution. Meanwhile, we prove they still attain minimax optimal generalization error:

- Non-smooth two-layer neural networks overfit benignly in binary classification for mixture model.

A notable feature of our results is that they are devoid of many restrictions commonly assumed in related literature. We allow non-smooth activations, including the popular ones in practice such as ReLU and leaky-ReLU. We allow constant network width and allow the weights to travel arbitrarily far from its initial values, transcending the lazy training paradigm commonly adopted in literature. As such, our results provide a better understanding of benign overfitting for two-layer networks under a more practical setting.

1.1 Related Works

A significant part of previous theoretical results on benign overfitting focused on the simplified setting of linear regression (Bartlett et al., 2020; Tsigler and Bartlett, 2020; Negrea et al., 2020; Chatterji et al., 2021; Hastie et al., 2022; Chinot et al., 2022). Recent works have moved on to more complicated settings like logistic regression (Montanari et al., 2019; Chatterji and Long, 2021; Wang et al., 2021), kernel-based estimators (Liang et al., 2020; Mei and Montanari, 2022). Considering the large and rapidly expanding volume of literatures in this area, the references we provide here are by no means comprehensive.

Though benign overfitting is initially motivated by attempts to understand neural networks, the theoretical aspect of benign overfitting for neural networks is much less cultivated. A few results were obtained in the lazy training regime (Allen-Zhu et al., 2019; Arora et al., 2019a), which roughly speaking is a linear (or quadratic) approximation of neural networks and, in spite of its simplicity, fails to account for several important aspects of practical usage of neural networks (Yang and Hu, 2021). Moving beyond the NTK regime, little is known. In the noiseless setting, Brutzkus et al. (2018) showed that two-layer neural networks with leaky ReLU activations can be trained to zero training loss by SGD and generalize well in low-dimensional regime. It is however not clear how their technique extends to noisy labels and high-dimensions.

Our result is most closely related to and inspired by Frei et al. (2022). In that paper, it was shown that two-layer neural networks with smooth, leaky activation functions overfit benignly for well-separated binary classification. Another related paper (Cao et al., 2022) considers a very different data model but also assumes a smoothified ReLU activation. Our result completely removes the smoothness and the leakiness assumptions. As we will see, removing these assumptions require novel ideas and substantial efforts:

- Non-smooth activation excludes the possibility to use Taylor approximations which are pervasive in Frei et al. (2022) and in other previous works, so a finer-grained analysis technique is necessary;

- Non-leaky activation may lead to severe problems known as dying ReLU in practice, making the discussion of abundance of active neurons crucial in the proof, both in training (Lemma 4.3) and in generalization (Lemma 4.7). Such results are novel and cannot be discovered under the setting of Frei et al. (2022). In establishing such results we also need to employ new tools such as random matrix theory and anti-concentration inequalities, and develop new arguments;

- The control of the empirical loss in Frei et al. (2022) relies on a proxy PL inequality which depends crucially on the smoothness of the activation function. We provide a finer analysis of the dynamics of the empirical loss that bypasses this problem. Our refined analysis also leads to a faster convergence rate compared with Frei et al. (2022);

- Moreover, in our general setting we will see that the local approach to control the generalization margin in Frei et al. (2022) fails to work (see Lemma 4.7) due to complicated statistical dependence stemming from lack of smoothness and leakiness. We will see that an almost entirely different "global" approach is required. Our new approach allows to utilize random matrix theory to completely avoid statistical dependence without resorting to any smoothness or leakiness assumption.

1.2 Basic Notations

We will use c, c', C, C', etc. to denote constants that may vary upon each occurrence. For a natural number n, we use $[n]$ to denote the set $\{1, \ldots, n\}$. The cardinality of a set A is denoted by $|A|$. The gradient $\partial\sigma$ of a possibly non-differentiable function σ is defined later by (3).

For a matrix $W = (w_1, w_2, \dots, w_m) \in \mathbb{R}^{p\times m}$ where $w_j \in \mathbb{R}^p$, denote its $(2,1)$-mixed norm by $\|W\|_{2,1}$, defined as $\|W\|_{2,1} := \sum_{j=1}^m \|w_j\|$.

2 PROBLEM FORMULATION

We consider a two-layer network $f_W(x)$ of width m, with activation function σ, defined by

$$f_W(x) := \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \sigma(\langle w_j, x\rangle) \tag{1}$$

where $x \in \mathbb{R}^p$ is the input data and $W = (w_1, \dots, w_m)$ with $w_j \in \mathbb{R}^p$, $j = 1, \dots, m$ being the weight vectors of the neurons. The weights of the second layer a_j's are initialized as i.i.d. Rademacher random variables[1] and fixed thereafter, following common conventions in related literature (Arora et al., 2019a; Frei et al., 2022; Cao et al., 2022).

Activation. The only restrictions we put on the activation function σ are $\sigma(0) = 0$ and the following:

$$0 \le \sigma(u) - \sigma(v) \le u - v, \qquad \forall u \ge v, \tag{2a}$$
$$\sigma(u) - \sigma(v) \ge \gamma(u - v), \qquad \forall u \ge v \ge 0. \tag{2b}$$

where $\gamma \in (0, 1]$ is some constant. We will regard γ as being far away from 0, thus $1/\gamma = O(1)$. In plain words, we assume the activation is increasing, Lipschitz continuous, and has non-vanishing gradients in the activated region $u > 0$. The prototype we have in mind is the (arguably most popular) ReLU family, including ReLU, leaky ReLU, Exponential Linear Unit (ELU), etc. With finer assumptions one may also include more complicated activations like Gaussian Error Linear Unit (GELU), but for simplicity we will content ourselves with the model above.

For purpose of gradient descent we need to compute the gradient of σ. In most cases σ will be differentiable almost everywhere, but actually we do not need to assume any differentiability here. Instead, we simply *define* the gradient $\partial\sigma(u)$ at u to be any number satisfying

$$\partial\sigma(u) \in \left[\varliminf_{v\to u} \frac{\sigma(v) - \sigma(u)}{v - u}, \varlimsup_{v\to u} \frac{\sigma(v) - \sigma(u)}{v - u}\right]. \tag{3}$$

By assumption we have $\partial\sigma(x) \in [\gamma, 1]$ for $x > 0$ and $\partial\sigma(x) \in [0, 1]$ for $x \le 0$.

Initialization. As mentioned above, we initialize the second layer a_j's of the neural network with Rademacher distribution. The hidden layer, on the other hand, is initialized by random Gaussians following standard practice, i.e., we draw the initial weights $W^{(0)} = (w_1^{(0)}, \dots, w_m^{(0)})$ from m i.i.d. samples from the rescaled Gaussian distribution $\mathcal{N}(0, \omega_{\mathsf{init}}^2 I_p)$. The parameter ω_{init} controls the magnitude of initialization and is usually set as a small number in practice. In accordance with this practice, we assume ω_{init} is sufficiently small throughout this paper, the quantitative meaning of which will be made clear later.

Data Model. Assume the unpolluted dataset consists of n labeled samples $(x_i, y_i^\star)_{i=1}^n$ which are i.i.d. samples drawn from some distribution $\mathsf{P}_\star$ on $\mathbb{R}^p \times \{-1, 1\}$. We can observe x_i's, but do not know what their true labels $y_i^\star$'s are. Instead, we can only access a polluted version y_i of $y_i^\star$. We assume that the pollution amounts to flip the labels of (at most) η-fraction of all the n data points but is otherwise arbitrary, allowing for adversary pollution. Formally speaking, we assume

$$\frac{1}{n}|\{i \in [n] : y_i \ne y_i^\star\}| \le \eta.$$

We further assume the distribution $\mathsf{P}_\star$ can be described by the following mixture model:

(i) Fix some $\mu \in \mathbb{R}^p$.

(ii) Generate $z \sim \mathsf{P}_z$ where P_z is a centered isotropic distribution[2] on $\mathbb{R}^p$ whose logarithmic Sobolev constant is bounded by some constant β (see Remark 1).

(iii) Generate $y^\star$ from a Rademacher distribution.

(iv) Set $x = y^\star\mu + z$. The final output is $(x, y^\star)$.

A few remarks on the generality of the model are in order.

Remark 1 (Log-Sobolev assumption). The assumption on the logarithmic Sobolev constant (Ledoux, 2001) of P_z is a purely technical one; it is used to derive Lipschitz concentration properties (cf. Ledoux (2001); please refer to the supplement material of this paper for details) for a simple proof of the generalization bound. It can be viewed as subgaussian assumption plus some geometric regularity assumption on the distribution. It is satisfied by a wide range of distributions adopted in previous literatures, e.g. Gaussian distribution, or strongly log-concave distributions. These subsume the models in Chatterji and Long (2021); Frei et al. (2022); Wang and Thrampoulidis (2022). Finally, we note that $\beta \gtrsim 1$ due to the centered and isotropic assumption (Ledoux, 2001).

Remark 2 (Centered and isotropic assumption). The centered and isotropic assumption is not restrictive since any distribution can be transformed to a centered and isotropic one by a simple affine transform.

Remark 3. Similar models appeared in Chatterji and Long (2021); Frei et al. (2022); Wang and Thrampoulidis (2022). Our model is a bit more general: we do not assume independence of the p features (i.e. dimensions) of z. Finding a

[1] A Rademacher random variable takes value -1 or 1 with equal probability $1/2$.

[2] This means $\mathsf{E}_{z\sim\mathsf{P}_z}[z] = 0$ and $\mathsf{E}_{z\sim\mathsf{P}_z}[zz^\top] = I$.

representation with independent features is usually nontrivial and requires considerable effort of feature engineering in practice, thus this generalization is meaningful.

Training. We train the neural network f_W by optimizing logistic loss with full-batch gradient descent. Denote by $\ell(u)$ the logistic loss function

$$\ell(u) := \log(1+\exp(-u)).$$

The empirical loss is defined by

$$\hat{L}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i f_W(x_i)). \tag{4}$$

From an iteration $W^{(t)} = (w_1^{(t)}, \dots, w_m^{(t)})$, the next iteration is computed by

$$W^{(t+1)} = W^{(t)} - \alpha \nabla_W \hat{L}(W^{(t)}),$$

where $\alpha > 0$ is the (constant) stepsize. The gradient should be understood in the sense of (3) for non-smooth σ.

Generalization. We characterize the generalization performance of a neural network f_W in terms of the expected classification error rate:

$$p_{\mathsf{err}}(W) := \mathsf{P}_{(x,y)\sim \mathsf{P}_\star}(y f_W(x) \le 0). \tag{5}$$

Assumption on Parameters. The key parameters in the above models include the input dimension p, the width m of the hidden layer, the initialization magnitude ω_{init}, the number of samples n, the separation μ of different classes, the stepsize α, and the fraction η of labels polluted. The lower bound γ of activated gradient and the upper bound β of logarithmic Sobolev constant are less crucial and can be regarded as absolute constants in this paper. We will fix a "failure probability" $\delta \in (0, 1/2)$, and make the following assumptions on parameters, with $C > 0$ some sufficiently large constant depending only on β, γ:

$$n \ge C\log(m/\delta), \tag{6a}$$
$$m \ge C\log(n/\delta), \tag{6b}$$
$$p \ge C(n\|\mu\|^2 + n^2\log(n/\delta)), \tag{6c}$$
$$\eta \le 1/C, \tag{6d}$$
$$\alpha \le 1/(Cp^2), \tag{6e}$$
$$\omega_{\mathsf{init}} \le \alpha/\sqrt{mp}. \tag{6f}$$

We need yet another assumption that may seem less obvious but is actually validated by information-theoretic limit:

$$\|\mu\| \ge C(p/n)^{1/4}\log^{1/4}(mp/n\delta). \tag{7}$$

The rationale for this assumption is that, by a result of Giraud and Verzelen (2019), the minimax generalization error is at least $c\exp(-c\min(\|\mu\|^2, n\|\mu\|^4/p))$. In our setting, according to (6c), we have $n\|\mu\|^4/p \le \|\mu\|^2$, thus the minimax generalization error is at least $c\exp(-cn\|\mu\|^4/p)$. This indicates that the assumption (7), up to logarithmic factors, is inevitable if we wish a trained classifier to generalize well (regardless of which classifier and which learning method we are using).

Remark 4 (Assumption on the corruption parameter η). We assumed $\eta \le 1/C$ for a sufficiently large constant C, i.e., the proportion of the corrupted labels is less than some small constant, which is unspecified in our paper yet assumed to be sufficiently small (say, $(1000\beta^2\gamma^2)^{-1}$). This is nearly the most general constraint from a theoretical perspective: η in general cannot exceed some small constant$\ll 1$ since with $\eta = 1/2$ one can clearly construct a dataset on which no algorithm is generalizable. Our assumption excludes these extreme cases but does not ask for much more.

3 MAIN RESULTS

Theorem 3.1. *Under the assumptions in Section 2, the following holds with probability at least $1-\delta$. For any number of iterations $t \ge C\hat{L}(W^{(0)})/(\alpha\|\mu\|^2\epsilon)$ where $\epsilon \in (0, 1/2)$ is arbitrary, the neural network overfits benignly, in the sense that*

1. *(Perfect fitting) The empirical loss is driven to arbitrarily small:*
$$\hat{L}(W^{(t)}) \le \epsilon, \tag{8}$$
and if $\epsilon < 1/(2n)$, all training labels (including polluted ones) are fitted perfectly:
$$y_i f_{W^{(t)}}(x_i) > 0,\ \forall 1 \le i \le n. \tag{9}$$
2. *(Generalization) The network $f_{W^{(t)}}$ generalizes well to new data: for $(x,y) \sim \mathsf{P}_\star$ we have (recall that p_{err} denotes the generalization error rate defined by (5))*
$$p_{\mathsf{err}}(W^{(t)}) \le \exp\left(-\frac{n\|\mu\|^4}{Cp}\right). \tag{10}$$

Since $\ell(u) \to 0$ only when $u \to \infty$, it is clear that (8) can hold only if the weights of the network can grow infinitely large as $\epsilon \to 0$ (hence $t \to \infty$). This indicates that our result allows (and requires) the weight to move arbitrarily far from its initial value, hence goes beyond the lazy training regime.

3.1 Comparison with Previous Works

It has been shown in Giraud and Verzelen (2019) that the minimax lower bound of generalization error in the noiseless version of our setting is $c\exp(-cn\|\mu\|^4/p)$, thus our result is minimax optimal in terms of sample efficiency.

Compared with Frei et al. (2022), we not only considered a more general setting but also obtained a better $O(1/t)$ convergence rate of $\hat{L}(W^{(t)})$ instead of the $O(1/\sqrt{t})$ rate there, due to some finer-grained analysis in our proof.

Note that in Frei et al. (2022) the activation function is assumed to be both leaky and smooth, i.e. σ is twice differentiable with $\sigma' \geq \gamma > 0$ and $|\sigma''| \leq H$ everywhere. Our work removes both assumptions: we make no assumption on second-order differentiability of σ and allow σ to be "deactivated" in the region $(-\infty, 0]$. In that region the gradient of σ can vanish or even not exist. Thus at least two new questions not involved in Frei et al. (2022) arise: the dying ReLU problem (the gradient can be vanishing so the training of NN gets stuck), and the failure of Taylor approximation.

From a big picture, the NTK approach can be regarded as a linearized regime, while the assumptions in Frei et al. (2022) can be regarded as some "semi-linear" regime. Though the activation function was not assumed to be actually linear there, it shared several crucial properties with a linear function, e.g. we have $\sigma(u)-\sigma(v) \asymp u-v$ for any u, v, the right hand side being a linear function. We even have $\sigma(u)-\sigma(v) \sim \sigma'(u)(u-v)$ uniformly for sufficiently close u, v since $|\sigma''| \leq H$. In contrast, our setting corresponds to a more "non-linear" regime, where the above approximation all fails and one has to confront some singular behaviors of NN due to non-linearity such as discontinuous gradients and dying ReLU.

4 ANALYSIS

In analyzing the evolution of the network during the training process we need to monitor several important aspects, which interweave with each other in a subtle way. This section provides a brief overview of these aspects. First we set up the basic technical background: we condition on some good events $\mathcal{E}$ throughout the whole proof to streamline the arguments. Then we analyze the dynamics of the state of neurons and provide a guarantee that sufficiently many neurons stay active. With this guarantee we are able to track the changes of empirical margins and losses, which are crucial to the proof of perfect fitting (8) and (9). The techniques established in these steps will be useful for the final part of our analysis: bounding the generalization error.

4.1 Good Event

It is convenient to condition on some good event $\mathcal{E}$, defined by the intersection of the following events:

(i) The sets $J^+ := \{j \in [m] : a_j = 1\}$ and $J^- := \{j \in [m] : a_j = -1\}$ have cardinality at least $m/3$:

$$\min(|J^+|, |J^-|) \geq m/3. \tag{11}$$

(ii) For all $i \in [n], j \in [m]$ we have

$$\Big|\{i \in [n] : y_i = a_j, \langle w_j^{(0)}, x_i\rangle > 0\}\Big| \geq \frac{n}{60\beta^2}, \tag{12}$$

$$\Big|\{j \in [m] : y_i = a_j, \langle w_j^{(0)}, x_i\rangle > 0\}\Big| \geq \frac{m}{60\beta^2}. \tag{13}$$

(iii) For any $\zeta = (\zeta_1, \ldots, \zeta_n) \in \mathbb{R}^n$ we have

$$\frac{3p}{4}\|\zeta\|^2 \leq \left\|\sum_{i=1}^n \zeta_i(x_i - y_i^\star \mu)\right\|^2 \leq \frac{5p}{4}\|\zeta\|^2. \tag{14}$$

(iv) For all $i \in [n]$ we have

$$|\langle \mu, x_i - y_i^\star \mu\rangle| \leq 16\beta\|\mu\|\sqrt{\log(n/\delta)}. \tag{15}$$

(v) For all $i, i' \in [n], i \neq i'$ we have

$$|\langle x_i - y_i^\star \mu, x_{i'} - y_{i'}^\star \mu\rangle| \leq 16\beta\sqrt{p\log(n/\delta)}. \tag{16}$$

(vi) For all $j \in [m]$ we have

$$\|w_j^{(0)}\| \leq 2\omega_{\mathsf{init}}\sqrt{p}. \tag{17}$$

(vii) For all $i \in [n]$ we have

$$|y_i f_{W^{(0)}}(x_i)| \leq 1. \tag{18}$$

The meaning of these events are technical; we will see later how some of them fit into our proof. For now it is useful to know that it is of no loss in utility by conditioning on $\mathcal{E}$.

Lemma 4.1. *The event $\mathcal{E}$ happens with probability at least $1-\delta$.*

Except for (12) and (13), the proof is a routine application of concentration inequalities and random matrix theory. The two exceptional equations are a bit more complicated and involve the anticoncentration property of the inner product $\langle w_j^{(0)}, x_i - y_i^\star \mu\rangle$. Details can be found in an extended version of this paper.

4.2 Undying ReLU

Now we encounter the first major technical challenge, known as dying ReLU problem when σ is the ReLU function, caused by the generality of our assumption on activation function: it is allowed to have zero gradient on negative input, hence the weight of neuron may not be updated anymore if all its inputs become negative. In such case the neuron is called *inactive*. The neural network can get trapped in the training process when too many neurons have turned inactive. We must show that this is unlikely to happen so that the neural network can be trained well.

Denote by $\mathcal{A}(t)$ the set of pairs (i, j) such that the neuron j is active for the sample x_i at the t-th iteration, i.e.

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

Denote its coordinate sections by $\mathcal{A}^i(t)$ and $\mathcal{A}_j(t)$ respectively, which are defined by

$$\mathcal{A}^i(t) := \{j \in [m] : (i,j) \in \mathcal{A}(t)\},$$
$$\mathcal{A}_j(t) := \{i \in [n] : (i,j) \in \mathcal{A}(t)\}.$$

In plain words, $\mathcal{A}^i(t)$ means the set of neurons that are active for input x_i at the t-th iteration, while $\mathcal{A}_j(t)$ means the set of training samples that activate the j-th neuron at the t-th iteration.

The importance of $\mathcal{A}(t)$ is apparent: we need to ensure the gradients are non-vanishing.

Proposition 4.1. *For any pair $(i,j) \in \mathcal{A}(t)$, we have $\partial\sigma(\langle w_j^{(t)}, x_i\rangle) \geq \gamma$.*

Proof. This follows from the definition of $\mathcal{A}(t)$ and our assumption (2b). □

Among the active pairs (i,j) it is important to distinguish the ones that have correct signs, i.e. $a_j = y_i$. Such a pair is not only active but also contributes positively to the margin $y_i f_{W^{(t)}}(x_i) = \sum_{j\in[m]} y_i a_j \sigma(\langle w_j^{(t)}, x_i\rangle)/\sqrt{m}$. For this reason we denote

$$\mathcal{T} := \{(i,j) \in [n] \times [m] : y_i = a_j\}.$$

and similar to the above we define

$$\mathcal{T}^i := \{j \in [m] : (i,j) \in \mathcal{T}\},$$
$$\mathcal{T}_j := \{i \in [n] : (i,j) \in \mathcal{T}\}.$$

First we show there are many correctly labeled active neurons upon initialization.

Lemma 4.2 (Initially active neurons)**.** *On the event $\mathcal{E}$ we have:*

$$|\mathcal{A}^i(0) \cap \mathcal{T}^i| \geq m/(60\beta^2), \qquad \forall i \in [n],$$
$$|\mathcal{A}_j(0) \cap \mathcal{T}_j| \geq n/(60\beta^2), \qquad \forall j \in [m].$$

Proof. This follows from (12) and (13). □

The following lemma, crucial to our analysis, asserts that correctly labeled active neurons are always active: they will never be deactivated in any iteration.

Lemma 4.3 (Active neurons stay active)**.** *On the event $\mathcal{E}$, the set $\mathcal{A}(t) \cap \mathcal{T}$ is monotonically increasing in t, i.e. for any $t \geq 0$,*

$$\mathcal{A}(t) \cap \mathcal{T} \subset \mathcal{A}(t+1) \cap \mathcal{T}.$$

In spite of its innocent looking, this lemma requires an rather indirect and involved proof. Due to space limits we will not discuss the detail here; it can be found in an extended version of this paper.

The above two lemmas together imply that we always have sufficiently many active neurons. In particular, for each neuron there are $\Omega(n)$ samples that activate it, and each sample activates $\Omega(m)$ neurons. These provide a quantitative guarantee that the network is sufficiently active, which will be important for later arguments.

4.3 Margins and Loss Decay

Next we evaluate the decay of training loss during training. Recalling (4), this is closely related to the margin $y_i f_{W^{(t)}}(x_i)$, which we inspect first. For convenience we denote

$$g(u) = -\ell'(u) = 1/(1+\exp(u)).$$

Lemma 4.4 (Growth of margin)**.** *On the event $\mathcal{E}$ the following holds. For all $i \in [n]$, we have, assuming $|\mathcal{A}^i(t) \cap \mathcal{A}^i(t+1)| \geq m/(60\beta^2)$, that*

$$\frac{\gamma^2\alpha p}{240\beta^2 n} \leq \frac{y_i f_{W^{(t+1)}}(x_i) - y_i f_{W^{(t)}}(x_i)}{g(y_i f_{W^{(t)}}(x_i))} \leq \frac{3\alpha p}{n}. \quad (19)$$

In particular, $y_i f_{W^{(t)}}(x_i) \geq y_i f_{W^{(0)}}(x_i) \geq -1$.

The assumption in the above lemma is automatically satisfied according to Lemma 4.3, which implies

$$\mathcal{A}^i(t) \cap \mathcal{A}^i(t+1) \supset (\mathcal{A}^i(t) \cap \mathcal{T}^i) \cap (\mathcal{A}^i(t+1) \cap \mathcal{T}^i)$$
$$\supset \mathcal{A}^i(0) \cap \mathcal{T}^i,$$

which has at least $m/(60\beta^2)$ elements in light of Lemma 4.2.

We return to discuss the implication of Lemma 4.4. It shows that $y_i f_{W^{(t)}}(x_i)$ is increasing in t and the growth of $y_i f_{W^{(t)}}(x_i)$ is proportional to the growth of $g(y_i f_{W^{(t)}}(x_i))$. Consequently, the decay of loss $\ell(y_i f_{W^{(t)}}(x_i))$ is roughly

$$\begin{aligned}&\ell(y_i f_{W^{(t)}}(x_i)) - \ell(y_i f_{W^{(t+1)}}(x_i))\\ \simeq\ &-\ell'(y_i f_{W^{(t)}}(x_i))(y_i f_{W^{(t+1)}}(x_i) - y_i f_{W^{(t)}}(x_i))\\ \simeq\ &\alpha p g(y_i f_{W^{(t)}}(x_i))^2/n\end{aligned}$$

given that the stepsize α is sufficiently small. With some efforts it is possible to make this computation rigorous:

Lemma 4.5 (Decay of empirical loss, preliminary form)**.** *On the event $\mathcal{E}$ the following holds. For all $i \in [n]$, assuming $|\mathcal{A}^i(t) \cap \mathcal{A}^i(t+1)| \geq m/(30\beta^2)$, the loss $y_i f_{W^{(t)}}(x_i)$ is decreasing in t. Moreover, with the same assumption we have*

$$\ell(y_i f_{W^{(t)}}(x_i)) - \ell(y_i f_{W^{(t+1)}}(x_i))$$
$$\asymp_{\beta,\gamma} \frac{\alpha p}{n} g(y_i f_{W^{(t)}}(x_i))^2. \quad (20)$$

Up to this point it is not clear why the above computations would lead to a deceasing training loss. There is in fact a

dilemma that entails a careful control of $g(y_i f_{W^{(t)}}(x_i))$. For the training loss to be small we need a large margin, which by (19) requires the sum of $g(y_i f_{W^{(t)}}(x_i))$ to be large; but the loss $\ell(y_i f_{W^{(t)}}(x_i))$ cannot be small if $g(y_i f_{W^{(t)}}(x_i))$ is too large:

Proposition 4.2. *For $u \geq -1$ we have*

$$g(u) \leq \ell(u) \leq 4g(u).$$

The solution to this dilemma is to look at the decay of $\ell(y_i f_{W^{(t)}}(x_i))$ directly with the help of (20). Combined with the above Proposition 4.2 (and with the fact that $y_i f_{W^{(t)}}(x_i) \geq -1$ by Lemma 4.4) we have

$$\begin{aligned} &\ell(y_i f_{W^{(t)}}(x_i)) - \ell(y_i f_{W^{(t+1)}}(x_i)) \\ \asymp_{\beta,\gamma}\ &\frac{\alpha p}{n} \ell(y_i f_{W^{(t)}}(x_i))^2. \end{aligned}$$

This allows to control the behavior of $\ell(y_i f_{W^{(t)}}(x_i))$ by elementary methods.

Proposition 4.3. *Let $(a_t)_{t\geq 0}$ be a sequence of non-negative numbers satisfying $a_{t+1} \leq a_t - \psi a_t^2$ for some constant $\psi > 0$ and for all $t \geq 0$. If $\psi a_0 \leq 1$, then*

$$a_t \leq 1/(\psi t),\ \forall t \geq 0.$$

Since $y_i f_{W^{(0)}}(x_i) \geq -1$ by (18), when α is sufficiently small one may confirm that $\alpha p \ell(y_i f_{W^{(0)}}(x_i))/n \ll 1$. Thus the above proposition can be applied to the sequence $\ell(y_i f_{W^{(t)}}(x_i))$ with $\psi = C'\alpha p/n$, where C' denotes the constant factor for the upper bound in (20). Together with the monotonicity of $\ell(y_i f_{W^{(t)}}(x_i))$ as described by Lemma 4.5 we obtain:

Lemma 4.6 (Decay of empirical loss, final form)**.** *On the event $\mathcal{E}$ the following holds. For all $i \in [n]$ we have*

$$\ell(y_i f_{W^{(t)}}(x_i)) \lesssim_{\beta,\gamma} \min\left(1, \frac{n}{\alpha p t}\right).$$

In particular, averaging over $i \in [n]$ and using the assumption $p \geq n\|\mu\|^2$, we have

$$\hat{L}(W^{(t)}) \lesssim_{\beta,\gamma} \min\left(1, \frac{1}{\alpha \|\mu\|^2 t}\right).$$

4.4 Generalization

Finally we analyze the generalization error of the trained neural network. First we show how to reduce the upper bound of generalization error to a lower bound of generalization margin by Lipschitz concentration of logarithmically-Sobolev bounded distributions.

Proposition 4.4. *If $y \in \{-1, 1\}$, the function $z \mapsto y f_W(y\mu + z)$ is $\frac{1}{\sqrt{m}}\|W\|_{2,1}$-Lipschitz.*

Proof. This follows from the assumption (2a) by

$$\begin{aligned} &|f_W(y\mu + z) - f_W(y\mu + z')| \\ \leq\ &\frac{1}{\sqrt{m}} \sum_{j=1}^m |\sigma(\langle w_j, y\mu + z\rangle) - \sigma(\langle w_j, y\mu + z'\rangle)| \\ \leq\ &\frac{1}{\sqrt{m}} \sum_{j=1}^m |\langle w_j, z - z'\rangle| \leq \frac{1}{\sqrt{m}} \sum_{j=1}^m \|w_j\| \|z - z'\|, \end{aligned}$$

and the definition that $\|W\|_{2,1} = \sum_{j=1}^m \|w_j\|$. □

Recall that if P_z is a probability distribution on $\mathbb{R}^p$ whose logarithmic Sobolev constant is bounded by β and $\phi : \mathbb{R}^p \to \mathbb{R}$ is L-Lipschitz, then for $z \sim \mathsf{P}_z$ we have, for any $t > 0$ that $\mathsf{P}(\phi(z) - \mathsf{E}\phi(z) \leq -t) \leq \exp(-t^2/\beta L^2)$; please refer to Ledoux (2001) for a systematic account.

Proposition 4.5 (Reduction to generalization margin)**.** *Conditioning on $\mathcal{E}$, for $(x, y) \sim \mathsf{P}_\star$ we have*

$$\mathsf{P}(y f_{W^{(t)}}(x) \leq 0) \leq \exp\left(-\frac{m(\mathsf{E} y f_{W^{(t)}}(x))^2}{\beta \|W^{(t)}\|_{2,1}^2}\right),$$

where the probability and the expectation are taken with respect to $(x, y) \sim \mathsf{P}_\star$.

Proof. This follows from applying the Lipschitz concentration property mentioned above to $\phi(z) = y f_{W^{(t)}}(y\mu + z)$ and $t = \mathsf{E} y f_{W^{(t)}}(x)$. □

Now we know that to show the generalization error rate is small, it would suffice to show the generalization margin is large compared with $\|W^{(t)}\|_{2,1}$. The latter is another major technical challenge which we discuss next.

4.4.1 Generalization Margin

For $(x, y) \sim \mathsf{P}_\star$, We would like to show that $\mathsf{E}(y f_{W^{(t)}})(x)$ is sufficiently large. Recall that $y f_{W^{(t)}}(x) = \sum_{j\in[m]} y a_j \sigma(\langle w_j^{(t)}, x\rangle)/\sqrt{m}$. One may envision at least three difficulties:

- We need enough active neurons with $\langle w_j^{(t)}, x\rangle > 0$; otherwise the network may have almost zero output.
- We need to show that if $y a_j = 1$, then $\langle w_j^{(t)}, x\rangle$ is not only positive but also large for many neurons.
- On the other hand, for neurons with $y a_j = -1$ we need to show that $\langle w_j^{(t)}, x\rangle$ cannot be too large, since otherwise the summand $y a_j \sigma(\langle w_j^{(t)}, x\rangle)$ will have a significant negative impact on the sum.

Obviously, to tackle these difficulties it is crucial to understand the behavior of $\langle w_j^{(t)}, x\rangle$. A feature of our approach is that, instead of controlling the local difference

$\langle w_j^{(t+1)}, x\rangle - \langle w_j^{(t)}, x\rangle$ which seems more intuitive, we directly analyze the global difference $\langle w_j^{(t)}, x\rangle - \langle w_j^{(0)}, x\rangle$. Compared with the local approach, the global approach will save a $\sqrt{n}$ factor which is crucial to attain the near-optimal dependence on $\|\mu\|$ given by (7). The following computation contrast these two approaches. Denote

$$\hat{G}(W) := \frac{1}{n}\sum_{i=1}^{n} g(y_i f_W(x_i)).$$

Lemma 4.7. *Assume $x = y\mu + z$ where $y \in \{-1, 1\}$ and $z \sim \mathsf{P}_z$ as in Section 2. Fix some $t > 0$. On the event $\mathcal{E}$, there exist constants $c, C' > 0$ depending only on β, γ such that the following holds for all $j \in [m]$ and all $\tau < t$ with probability at least $1 - (p/n)^{-10}$ with respect to z:*

$$\begin{aligned} &a_j y(\langle w_j^{(\tau+1)}, x\rangle - \langle w_j^{(\tau)}, x\rangle) \\ &\geq \frac{c\alpha}{\sqrt{m}}\left(\|\mu\|^2 - C'\max_{i\in[n]}|\langle x_i - y_i^\star\mu, z\rangle|\right)\hat{G}(W^{(\tau)}), \end{aligned} \tag{21}$$

meanwhile

$$\begin{aligned} &a_j y(\langle w_j^{(t)}, x\rangle - \langle w_j^{(0)}, x\rangle) \\ &\geq \frac{c\alpha}{\sqrt{m}}\left(\|\mu\|^2 - C'\sqrt{\frac{p\log(pm/n)}{n}}\right)\sum_{\tau=0}^{t-1}\hat{G}(W^{(\tau)}). \end{aligned} \tag{22}$$

Remark 5. The above bounds are tight: one may construct examples where the reverse inequalities, after replacing c, C' with another pair of constants, hold with a non-diminishing probability.

Conditioning on the event $\mathcal{E}$, it is possible to show that $\max_{i\in[n]}|\langle x_i - y_i^\star\mu, z\rangle| \asymp \sqrt{p\log(n)}$ with overwhelming probability with respect to z. Thus one may see that to control the behavior of $\langle w_j^{(t)}, x\rangle$, the local bound (21) requires $\|\mu\|^4 = \tilde{\Omega}(p)$, while the global bound (22) only requires $\|\mu\|^4 = \tilde{\Omega}(p/n)$. The latter is not only polynomially better but also matching the minimax lower bound $\|\mu\|^4 = \Omega(p/n)$ up to logarithmic factors.

Applying (22) and a few simple arguments to bound $\langle w_j^{(0)}, x\rangle$, we may prove under assumption (7) that

Corollary 4.1. *Assume $x = y\mu + z$ where $y \in \{-1, 1\}$ and $z \sim \mathsf{P}_z$ as in Section 2. Fix some $t > 0$. On the event $\mathcal{E}$, there exists constant $c > 0$ depending only on β, γ such that the following holds for all $j \in [m]$ with probability at least $1 - (p/n)^{-10}$:*

$$a_j y\langle w_j^{(t)}, x\rangle \geq \frac{c\alpha\|\mu\|^2}{2\sqrt{m}}\sum_{\tau=0}^{t-1}\hat{G}(W^{(\tau)}).$$

It can be seen that all the difficulties listed in the beginning have been resolved. From Corollary 4.1 we know that all neurons with $a_j y = 1$ are active for x, while all neurons with $a_j y = -1$ is deactivated. We may proceed as follows: for neurons with $a_j y = 1$ we have $y a_j \sigma(\langle w_j^{(t)} x,)\rangle \geq \gamma\langle w_j^{(t)}, x\rangle$, while for neurons with $a_j y = -1$ we have $y a_j \sigma(\langle w_j^{(t)} x,)\rangle \geq 0$ since $\langle w_j^{(t)}, x\rangle < 0$. Thus

$$\begin{aligned} y f_{W^{(t)}}(x) &= \frac{1}{\sqrt{m}}\sum_{j=1}^{m} y a_j \sigma(\langle w_j^{(t)}, x\rangle) \\ &\geq \frac{c\gamma\alpha\|\mu\|^2}{2m}|\{j\in[m] : a_j = y\}|\sum_{\tau=0}^{t-1}\hat{G}(W^{(\tau)}) \\ &\geq \frac{c\gamma\alpha\|\mu\|^2}{6}\sum_{\tau=0}^{t-1}\hat{G}(W^{(t)}), \end{aligned}$$

where the last inequality follows from (11). We have finished a major part of the proof the following lemma:

Lemma 4.8 (Large generalization margin)**.** *On the event $\mathcal{E}$, for some constant $c' > 0$ depending only on β, γ we have*

$$\mathsf{E}_{(x,y)\sim\mathsf{P}_\star}(y f_{W^{(t)}}(x)) \geq c'\gamma\alpha\|\mu\|^2\sum_{\tau=0}^{t-1}\hat{G}(W^{(\tau)}).$$

Sketch of proof. The above computation shows that $y f_{W^{(t)}}(x)$ is larger than a constant multiple of the right hand side with probability at least $1 - (p/n)^{-10}$. It remains to handle the "exceptional" case. In that case we use $|\langle w_j^{(t)}, x\rangle| \leq |\langle w_j^{(t)}, \mu\rangle| + \|w_j^{(t)}\|\|z\|$. It is relatively easy to control $|\langle w_j^{(t)}, \mu\rangle|$ and $\|w_j^{(t)}\|$ using the technique established in the proof of Lemma 4.7 (see also Lemma 4.9 below), while $\|z\|$ can be controlled using standard concentration inequality which implies that $\|z\|$ is "essentially" $O(\sqrt{p})$. □

4.4.2 Growth of Weights

Recalling Proposition 4.5, we still need to show that $\|W^{(t)}\|_{2,1}$ is relatively small. Fortunately, this is a much easier task so we simply state the corresponding result:

Lemma 4.9. *On the event $\mathcal{E}$, there exists some constant C' depending only on β, γ such that*

$$\|w_j^{(t+1)} - w_j^{(t)}\| \leq C'\alpha\sqrt{\frac{p}{nm}}\hat{G}(W^{(t)}).$$

Consequently, we have

$$\|W^{(t)}\|_{2,1} \leq \|W^{(0)}\|_{2,1} + C'\alpha\sqrt{\frac{mp}{n}}\sum_{\tau=0}^{t-1}\hat{G}(W^{(t)}).$$

Note that $\|W^{(0)}\|_{2,1} = \sum_j \|w_j^{(0)}\|$ can be well-controlled using (17).

We have by now collected all the necessary ingredients to prove our main result, which we will do immediately.

5 PROOF OF MAIN RESULTS

In this section we sketch the proof of Theorem 3.1.

Proof of (8). From Lemma 4.6 we have $\hat{L}(W^{(t)}) \leq C'/(\alpha\|\mu\|^2 t)$ for $t \geq C'/(\alpha\|\mu\|^2)$, where $C' > 0$ is some constant depending only on β, γ. From (18) we may infer that $\hat{L}(W^{(0)}) \geq \ell(1) \geq 1/4$, thus by taking $t \geq C\hat{L}(W^{(0)})/(\alpha\|\mu\|^2\epsilon)$ for sufficiently large C we readily obtain $\hat{L}(W^{(t)}) \leq \epsilon$, as desired. □

Proof of (9). Since $\hat{L}(W^{(t)}) \geq \frac{1}{n}\max_i \ell(y_i f_{W^{(t)}}(x_i))$, if there is some $i \in [n]$ such that $y_i f_{W^{(t)}} > 0$, we would have $\hat{L}(W^{(t)}) \geq \frac{1}{n}\ell(0) > \frac{1}{2n}$, contradicting (8) which says $\hat{L}(W^{(t)}) \leq \epsilon < 1/(2n)$. □

Proof of (10). Again by (18) we may infer $\hat{G}(W^{(0)}) \geq g(1) \geq 1/4$, thus from (17), (6f) and (6c) we know

$$\|W^{(0)}\|_{2,1} \leq 8\omega_{\mathsf{init}} m\sqrt{p}\hat{G}(W^{(0)}) \leq \alpha\sqrt{\frac{mp}{n}}\hat{G}(W^{(0)}).$$

Invoking Lemma 4.9, we may see that $\|W^{(t)}\|_{2,1} \leq 2C'\alpha\sqrt{mp/n}\sum_{\tau<t}\hat{G}(W^{(\tau)})$. This combined with Lemma 4.8 and Proposition 4.5 implies

$$\begin{aligned}\mathsf{P}(y f_{W^{(t)}}(x) \leq 0) &\leq \exp\left(-\frac{m\alpha^2\|\mu\|^4}{C\alpha^2(mp/n)}\right) \\ &= \exp\left(-\frac{n\|\mu\|^4}{Cp}\right),\end{aligned}$$

as desired. □

6 CONCLUSION

In this paper we proved with a fine-grained analysis of the network dynamics that non-smooth two-layer neural network overfits benignly in the binary classification problem, assuming the data comes from a well-separated mixture model. Our result allows the weight to travel infinitely far from the initial value, hence goes beyond the lazy training regime. This provides a better understanding of benign overfitting in a more practical setting.

Benign overfitting is an emerging area and many important problems remain open. Directly related to this paper are the following problems: (i) can we relax the overparameterization assumption (6c) to a more illuminating one, possibly of the form $mp \gg n$ (as mp is the total number of parameters in the network)? (ii) can we prove similar results beyond the mixture data model? (iii) how does the technique here helps to understand deeper networks? can we characterize the limits of benign overfitting capabilities of a two-layer network? We leave these directions for future research.

7 ACKNOWLEDGEMENTS

The authors are with Beijing National Research Center for Information Science and Technology (BNRist) and the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. This work was supported in part by the National Natural Science Foundation of China under Grant U2230201 and 61971266, Grant from the Guoqiang Institute, Tsinghua University, and in part by the Clinical Medicine Development Fund of Tsinghua University.

References

Z. Allen-Zhu, Y. Li, and Y. Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in Neural Information Processing Systems*, 2019.

S. Arora, S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, 2019a.

S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, 2019b.

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

M. Belkin, D. Hsu, S. Ma, and S. 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.

A. Brutzkus, A. Globerson, E. Malach, and S. Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. In *International Conference on Learning Representations*, 2018.

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

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

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

G. Chinot, M. Löffler, and S. van de Geer. On the robustness of minimum norm interpolators and regularized empirical risk minimizers. *The Annals of Statistics*, 50(4): 2306 – 2333, 2022.

S. Frei, N. S. Chatterji, and P. Bartlett. Benign overfitting without linearity: Neural network classifiers trained by

gradient descent for noisy linear data. In *Conference on Learning Theory*, 2022.

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

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

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

M. Ledoux. *The concentration of measure phenomenon*. American Mathematical Society, 2001.

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

S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

A. Montanari, F. Ruan, Y. Sohn, and J. Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

J. Negrea, G. K. Dziugaite, and D. Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, 2020.

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

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

K. Wang and C. Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization. *SIAM Journal on Mathematics of Data Science*, 4(1):260–284, 2022.

K. Wang, V. Muthukumar, and C. Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. In *Advances in Neural Information Processing Systems*, 2021.

G. Yang and E. J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, 2021.

A NOTATIONS AND BASIC COMPUTATIONS

As in the main text, we denote

$$g(u) := -\ell'(u) = 1/(1+\exp(u)), \quad g_i(W) := g(y_i f_W(x_i)), \quad \hat{G}(W) := \sum_{i=1}^{n} g_i(W).$$

The gradient $\nabla_W \hat{L}(W)$ can be computed explicitly:

$$\nabla_{w_j} \hat{L}(W) = -\frac{1}{n\sqrt{m}} \sum_{i=1}^{n} y_i a_j g(y_i f_W(x_i)) \partial\sigma(\langle w_j, x_i\rangle) x_i. \tag{23}$$

Define for an arbitrary weight matrix W that

$$\xi_j(x; W) := \frac{\sigma(\langle w_j - \alpha\nabla_{w_j}\hat{L}(W), x\rangle) - \sigma(\langle w_j, x\rangle)}{\langle -\alpha\nabla_{w_j}\hat{L}(W), x\rangle},$$
$$\lambda_i(x; W) := \frac{1}{m}\sum_{j=1}^{m} \xi_j(x; W)\partial\sigma(\langle w_j, x_i\rangle).$$

With $W^{(t+1)} = W^{(t)} - \alpha\nabla_W \hat{L}(W^{(t)})$, the update of $y f_{W^{(t)}}(x)$ for an arbitrary pair (x, y) is then:

$$\begin{aligned} y f_{W^{(t+1)}}(x) - y f_{W^{(t)}}(x) &= \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \xi_j(x; W^{(t)}) \langle -\alpha\nabla_{w_j}\hat{L}(W^{(t)}), yx\rangle \\ &= \frac{\alpha}{nm} \sum_{i=1}^{n}\sum_{j=1}^{m} a_j^2 g_i(W^{(t)}) \xi_j(x; W^{(t)}) \partial\sigma(\langle w_j, x_i\rangle)\langle y_i x_i, yx\rangle \\ &= \frac{\alpha}{n} \sum_{i=1}^{n} g_i(W^{(t)}) \lambda_i(x; W) \langle y_i x_i, yx\rangle \end{aligned} \tag{24}$$

where in the last line we used $a_j^2 = 1$.

B PROOF OF LEMMA 4.1

We prove Lemma 4.1 in the main text stating that the good event $\mathcal{E}$ happens with probability at least $1-\epsilon$. More concretely, we will show that each of Eqns. (11)–(18) in the main text hold with probability at least $1 - \delta/10$; the lemma then follows easily from a union bound. As we said, this is mostly a routine application of standard concentration inequalities, which we now introduce.

Lemma B.1 (Chernoff bound)**.** *If $X_1, \dots, X_n$ are i.i.d. random variables such that $|X_i| \le 1$ almost surely, with $\mu = \mathsf{E}X_i$ we have, for any $t > 0$, that*

$$\mathsf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > t\right) \le 2\exp(-nt^2/2). \tag{25}$$

The following results are concerned with the behavior of distributions with bounded logarithmic Sobolev constants. The most important one for us is the following Lipschitz concentration property, which is a standard corollary of Herbst's argument (Ledoux, 2001).

Lemma B.2 (Lipschitz concentration)**.** *If $\phi : \mathbb{R}^p \to \mathbb{R}$ is a Lipschitz function and P_z is a probability distribution on $\mathbb{R}^p$ with logarithmic Sobolev constant β, then for any $t > 0$ we have*

$$\mathsf{P}_z(\{z : \phi(z) - \mathsf{E}_{z\sim\mathsf{P}_z}\phi(z) > t\}) \le \exp(-t^2/\beta).$$

The same bound holds for the left tail $\mathsf{P}_z(\{z : \phi(z) - \mathsf{E}_{z\sim\mathsf{P}_z}\phi(z) < -t\})$.

As we have a collection of i.i.d. samples from P_z, it will be useful to investigate their joint distribution. A fundamental result in this vein is that the joint distribution also has bounded logarithmic Sobolev constant (Ledoux, 2001).

Lemma B.3 (Tensorization). *If P_z is a probability distribution on $\mathbb{R}^p$ with logarithmic Sobolev constant bounded by β, then its n-fold product P_z^n is a probability distribution on $\mathbb{R}^{p\times n}$ with logarithmic Sobolev constant bounded by β.*

The following corollaries, proved later in Section B.6, will be particularly helpful in this paper.

Proposition B.1. *Let P_z be a centered, isotropic distribution on $\mathbb{R}^p$ with logarithmic Sobolev constant bounded by β. Assume $z \sim \mathsf{P}_z$. Then $\langle v, z\rangle$ is centered $\frac{\|v\|}{\sqrt{2}}\beta$-subgaussian of variance $\|v\|^2$ for any $v \in \mathbb{R}^p$. Moreover, the following holds with probability at least $1-\exp(-u^2)$ for any $u > 4$:*

$$|\langle v, z\rangle| \le 4u\sqrt{\beta}\|v\|.$$

Proposition B.2. *Under the same assumption as in Proposition B.1, the following holds with probability at least $1-\exp(-pu^2)$ for any $u > 4$:*

$$\|z\| \le 4u\sqrt{\beta p}.$$

In particular, setting $\mathcal{E}_* = \{\|z\| \le 8u\sqrt{\beta p}\}$ with $u > 4$, we have

$$\mathsf{E}(\|z\|\mathbf{1}_{\mathcal{E}_*}) \le \sqrt{\beta}\exp(-pu^2).$$

Proposition B.3. *Let P_z be as in Proposition B.1 and let $z_1, z_2, \dots, z_n$ be i.i.d. samples from P_z. Then for any $u \in \left[4, \sqrt{p/16\beta n}\right]$, the following holds with probability at least $1-\exp(-nu^2)$. For any $\zeta_1, \zeta_2, \dots, \zeta_n \in \mathbb{R}$, we have*

$$(p - 16u\sqrt{\beta np})\sum_{i=1}^n \zeta_i^2 \le \left\|\sum_{i=1}^n \zeta_i z_i\right\|^2 \le (p + 16u\sqrt{\beta np})\sum_{i=1}^n \zeta_i^2$$

In particular, the following holds with probability at least $1-\exp(-cp/\beta)$, where $c > 0$ is some universal constant (say, $c = 1/1024$):

$$p/2 \le \min_{i\in[n]}\|z_i\|^2 \le \max_{i\in[n]}\|z_i\|^2 \le 2p.$$

B.1 Proof of Eqn. (11)

This follows from the Chernoff bound applied to $X_j := \mathbf{1}_{a_j=1}$ (hence $|J^+| = \sum_{j\in[m]} X_j$), which implies

$$\mathsf{P}\left(\left|\frac{|J^+|}{n} - \frac{1}{2}\right| > 1/6\right) \le 2\exp(-m/72) \le \delta/10,$$

by the assumption $m \ge C\log(1/\delta)$ (Eqn. (6b) in the main text). Note that when $\left|\frac{|J^+|}{n} - 1/2\right| \le 1/6$ we have $m/3 \le |J^+| \le 2m/3$, thus $|J^-| \ge m - |J^+| \ge m/3$, as desired.

B.2 Proof of Eqns. (12) and (13)

The proof relies crucially on the following anticoncentration property, which will be proved later in Section B.6.

Proposition B.4 (Anticoncentration of subgaussian random variables). *Assume X is some $\frac{1}{\sqrt{2}}\beta$-subgaussian random variable with $\mathsf{E}X = 0$ and $\mathsf{E}X^2 = 1$. Then*

$$\mathsf{P}\left(X > \frac{1}{200\beta^2}\right) \ge \frac{1}{20\beta^2}. \tag{26}$$

Return to the proof of Eqns. (12) and (13). By our data model we may write $x_i = y_i^\star\mu + z_i$ where $z_i \sim \mathsf{P}_z$. Note that $\langle w_j^{(0)}, x_i\rangle = y_i^\star\langle w_j^{(0)}, \mu\rangle + \langle w_j^{(0)}, z_i\rangle$. Recall that $w_j^{(0)}$'s are drawn i.i.d. from $\mathcal{N}(0, \omega_{\mathsf{init}}^2 I_p)$, we see that $\langle w_j^{(0)}, \mu\rangle$, $j \in [m]$ are i.i.d. centered Gaussian variables with variance $\omega_{\mathsf{init}}^2\|\mu\|^2$. By a well-known bound on the maximum of Gaussian variables (Vershynin, 2018) we have

$$\max_{j\in[m]}|\langle w_j^{(0)}, \mu\rangle| \le 4\omega_{\mathsf{init}}\|\mu\|\sqrt{\log(n/\delta)} \tag{27}$$

with probability at least $1-\delta/20$. Now we turn to inspect the term $\langle w_j^{(0)}, z_i\rangle$. Conditioning on z_i, this term becomes i.i.d. centered Gaussian variable with variance $\omega_{\mathsf{init}}^2\|z_i\|^2$ for $j\in[m]$. Each of such Gaussian random variables has a probability at least $1/5$ to exceed $\omega_{\mathsf{init}}\|z_i\|/10$. Moreover, since a_j is uniformly distributed on $\{-1,1\}$ and is independent of $w_j^{(0)}$, we have

$$\mathsf{P}\left(\langle w_j^{(0)}, z_i\rangle \ge \omega_{\mathsf{init}}\|z_i\|/10, a_j = y_i \,\middle|\, z_i, y_i\right) = \frac{1}{2}\mathsf{P}\left(\langle w_j^{(0)}, z_i\rangle \ge \omega_{\mathsf{init}}\|z_i\|/10 \,\middle|\, z_i\right) \ge 1/10.$$

Applying the Chernoff bound in a similar way as in Section B.1, we know that with probability at least $1-\exp(-cm)$, there exist a subset $J_i\subset[m]$ with $|J_i|\ge m/15$ such that $\langle w_j^{(0)}, z_i\rangle \ge \omega_{\mathsf{init}}\|z_i\|/10$ and $a_j = y_i^\star$ for all $j\in J_i$. Furthermore, conditioning on $\|z_i\|^2\ge p/2$, we have for all $j\in J_i$ that

$$\langle w_j^{(0)}, x_i\rangle \ge -4\omega_{\mathsf{init}}\|\mu\|\sqrt{\log(n/\delta)} + \omega_{\mathsf{init}}\|z_i\|/10 \ge \omega_{\mathsf{init}}(\sqrt{p}/20 - \|\mu\|\sqrt{\log(n/\delta)}) > 0,$$

where the last inequality follows from the assumption (6c) in the main text.

By Proposition B.3 we know that $\|z_i\|^2\ge p/2$ for all $i\in[n]$ with probability at least $1-\delta/60$. Combined with the above argument, we obtain that

$$\forall i\in[n],\quad \Big|\{j\in[m] : y_i = a_j, \langle w_j^{(0)}, x_i\rangle > 0\}\Big| \ge \frac{m}{15},$$

with probability at least $1-\delta/20-\delta/60-n\exp(-cm)$. When $m\ge C\log(n/\delta)$ as assumed in (6b), the probability is at least $1-\delta/10$. We have thereby proved that (13) holds with probability at least $1-\delta/10$.

The equation (12) follows from a similar argument. Conditioning on $w_j^{(0)}$, we infer from Proposition B.4 that $\langle w_j^{(0)}, z_i\rangle \ge \beta^{-2}\|w_j^{(0)}\|/200$ with probability at least $\beta^{-2}/20$ for each $i\in[n]$. Moreover, since $y_i^\star$ is uniformly distributed on $\{-1,1\}$ and is independent of z_i, we have

$$\mathsf{P}\left(\langle w_j^{(0)}, z_i\rangle \ge \beta^{-2}\|w_j^{(0)}\|/200, a_j = y_i^\star \,\middle|\, w_j^{(0)}, a_j\right) = \frac{1}{2}\mathsf{P}\left(\langle w_j^{(0)}, z_i\rangle \ge \beta^{-2}\|w_j^{(0)}\|/200 \,\middle|\, w_j^{(0)}\right) \ge \beta^{-2}/40.$$

The rest of the arguments is completely the same: conditioning on $\|w_j^{(0)}\|^2\ge \omega_{\mathsf{init}}^2 p/2$ which happens with probability at least $1-\delta/60$ we have

$$\forall j\in[m],\quad \Big|\{i\in[n] : y_i^\star = a_j, \langle w_j^{(0)}, x_i\rangle > 0\}\Big| \ge \frac{n}{50\beta^2},$$

which holds with probability at least $1-\delta/20-\delta/60-m\exp(-cn)$. When $n\ge C\log(m/\delta)$ as assumed in (6a), the probability is at least $1-\delta/10$. The desired equation (12) follows from the above result and the assumption that (recall that $C>0$ is a sufficiently large constant depending only on β,γ, thus we may choose $C>300\beta^2$)

$$|\{i\in[n] : y_i\ne y_i^\star\}| \le \eta n \le n/C \le \frac{n}{300\beta^2}.$$

B.3 Proof of Eqns. (14), (15), (16)

The equation (14) follows directly from Proposition B.3 and the assumption on p (Eqn. (6c)), which implies

$$\frac{3p}{4}\|\zeta\|^2 \le \left\|\sum_{i=1}^n \zeta_i(x_i - y_i^\star\mu)\right\|^2 \le \frac{5p}{4}\|\zeta\|^2$$

with probability at least $1-\exp(-16n)\ge 1-\delta/20$, given $n\ge C\log(1/\delta)$ (Eqn. (6a)).

The equation (15) follows from Proposition B.1 in the following way. By that proposition we know $|\langle\mu, x_i - y_i^\star\mu\rangle| \le 4u\sqrt{\beta}\|\mu\|$ with probability at least $1-\exp(-u^2)$ for each $i\in[n]$. Taking union bound, this inequality holds for all $i\in[n]$ with probability at least $1-n\exp(-u^2)$. Taking $u = 4\sqrt{\log(n/\delta)}$, we obtain the desired result.

The equation (16) follows from a combination of the above arguments. For convenience we denote $z_i = x_i - y_i^\star\mu$, hence $z_1,\dots,z_n$ are i.i.d. samples from P_z. First we fix some i and condition on z_i. By the argument used to prove (15) we know that $|\langle z_i, z_{i'}\rangle| \le 4u\sqrt{\beta}\|z_i\|$ with probability at least $1-\exp(-u^2)$ for any $i'\ne i$. We may then condition on (14) which implies $\|z_i\|^2\le 5p/4$ for all $i\in[n]$ with probability at least $1-\delta/20$. Then we have $|\langle z_i, z_{i'}\rangle| \le 8u\sqrt{\beta p}$ with probability at least $1-\exp(-u^2)$ conditioned on (14). Taking union bound over all $i,i'\in[n], i\ne i'$, we deduce that $\max_{i\ne i'}|\langle z_i, z_{i'}\rangle| \le 8u\sqrt{\beta p}$ with probability at least $1-n^2\exp(-u^2)$ conditioned on (14). Setting $u = 8\log(n/\delta)$, we have shown (16) holds with probability at least $1-\exp(-4\log(n/\delta)) - \delta/20 \ge 1-\delta/10$, as desired.

B.4 Proof of Eqn. (17)

It is well-known (or follows from Proposition B.2 with a slightly worse constant) that $\|w_j^{(0)}\| \le 2\omega_{\mathsf{init}}\sqrt{p}$ with probability at least $1-2\exp(-cp)$ if $w_j^{(0)} \sim \mathcal{N}(0, \omega_{\mathsf{init}}^2 I_p)$, cf. Vershynin (2018). Taking union bound, we obtain (17) with probability at least $1-2m\exp(-cp)$. By assumption (6c) and (6a) we know that $p \ge Cn \ge C\log(m/\delta)$, thus this probability is at least $1-\exp(-c'p) \ge 1-\delta/10$, as desired.

B.5 Proof of Eqn. (18)

Since $y_i \in \{-1,1\}$ it is clear that $|y_i f_{W^{(0)}}(x_i)| = |f_{W^{(0)}}(x_i)|$. Recall that

$$f_{W^{(0)}}(x_i) = \frac{1}{\sqrt{m}}\sum_{j=1}^m a_j \sigma(\langle w_j^{(0)}, x_i\rangle).$$

Since a_j's are i.i.d. Rademacher, they are 1-subgaussian; thus subgaussian concentration (which can be found in Vershynin (2018) or can be obtained as a corollary of Proposition B.1) implies with probability at least $1-\delta/30$ that

$$|f_{W^{(0)}}(x_i)| \le 2\sqrt{\frac{\log(1/\delta)}{m}}\left(\sum_{j=1}^m \sigma^2(\langle w_j^{(0)}, x_i\rangle)\right)^{1/2}.$$

Since $\sigma(0)=0$ and σ is Lipschitz, we know $\sigma^2(\langle w_j^{(0)}, x_i\rangle) \le |\langle w_j^{(0)}, x_i\rangle|^2$. Conditioning on x_i, $\langle w_j^{(0)}, x_i\rangle$ is a Gaussian random variable with variance $\omega_{\mathsf{init}}^2\|x_i\|^2$, hence is less than $2\sqrt{\log(mn/\delta)}\omega_{\mathsf{init}}\|x_i\|$ with probability at least $1-\delta/(30n)$ for all $j\in[m]$. Furthermore, condition on $\|x_i - y_i^\star\mu\| \le 2\sqrt{p}$ which holds for all $i\in[n]$ with probability at least $1-\delta/30$ as proved in Section B.3, we have $\|x_i\| \le \|\mu\| + \|x_i - y_i^\star\mu\| \le 3\sqrt{p}$ by assumption (6c), hence

$$|f_{W^{(0)}}(x_i)| \le 2\sqrt{\frac{\log(1/\delta)}{m}}(36m\log(mn/\delta)\omega_{\mathsf{init}}^2 p)^{1/2} = 12\sqrt{p\log(1/\delta)\log(mn/\delta)}\omega_{\mathsf{init}}. \tag{28}$$

By assumption (6f) we have $\omega_{\mathsf{init}} \le \alpha/\sqrt{mp}$, and by assumption (6e) we may further deduce $\omega_{\mathsf{init}} \le 1/(C\sqrt{mp^3})$. By our assumption (6b) on m and on p (Eqn. (6c)), it is clear that the right hand side of (28) is less than 1, as desired.

B.6 Proof of Auxiliary Propositions

Proof of Proposition B.1. Without loss of generality we may assume $\|v\|=1$. Then by centered isotropic assumption we have $\mathsf{E}|\langle v,z\rangle|^2 = \|v\|^2 = 1$. Note further that $\phi: z\mapsto\langle v,z\rangle$ is a Lipschitz function, the first part of the proposition the desired conclusion clearly follows from Lemma B.2. The second part then follows from the well-known subgaussian tail bound (cf. Vershynin (2018)). □

Proof of Proposition B.2. By the centered isotropic assumption we have $\mathsf{E}\|z\|^2 = p$. Note that $z\mapsto\|z\|$ is Lipschitz, the first part of the proposition follows from the same argument as in Proposition B.1. The second part of the proposition follows from the first part by integrating by parts:

$$\mathsf{E}(\|z\|\mathbf{1}_{\mathcal{E}_*}) = \int_{8u\sqrt{\beta p}}^\infty t\mathsf{P}(t<\|z\|\le t+dt) = \int_{8u\sqrt{\beta p}}^\infty \mathsf{P}(\|z\|>t)dt = 4\sqrt{\beta p}\int_{2u}^\infty \mathsf{P}(\|z\|>4v\sqrt{\beta p})dv,$$

where by the first part we know that the integral is no more than $\int_{2u}^\infty \exp(-pv^2)dv \le \exp(-2pu^2)/\sqrt{p}$ for $u>4$ by a well known Gaussian tail bound. The conclusion is immediate if we observe $4\exp(-pu^2) \le 4\exp(-16) < 1$. □

Proof of Proposition B.3. Let $A = [z_1,\dots,z_n] \in \mathbb{R}^{p\times n}$ be the matrix with columns z_i. Alternatively, one may view A as an (random) operator $\mathbb{R}^n \to \mathbb{R}^p$ defined by

$$(\zeta_1,\dots,\zeta_n) \mapsto \sum_{i=1}^n \zeta_i z_i.$$

The proposition is equivalent to saying that the singular values of A are bounded by

$$p - 16u\beta\sqrt{np} \le s_{\min}^2(A) \le s_{\max}^2(A) \le p + 16u\beta\sqrt{np} \tag{29}$$

with high probability.

By Lemma B.3, the matrix A follows a probability distribution in $\mathbb{R}^{p\times n}$ with logarithmic Sobolev constant bounded by β. Now fix some $v\in\mathbb{S}^n$. It is easy to verify $\mathsf{E}\|Av\|^2=p$. It follows immediately that $\mathsf{E}\|Av\|\le\sqrt{p}$. Furthermore, since the map $A\mapsto\|Av\|$ is Lipschitz, by Lemma B.2 we have

$$\begin{aligned}\mathsf{E}(\|Av\|-\mathsf{E}\|Av\|)^2&=\int_{t=0}^{\infty}t^2\mathsf{P}\left(t<\big|\|Av\|-\mathsf{E}\|Av\|\big|\le t+dt\right)\\&=\int_0^{\infty}2t\mathsf{P}\left(\big|\|Av\|-\mathsf{E}\|Av\|\big|>t\right)dt\\&\le\int_0^{\infty}4t\exp(-t^2/\beta)dt\\&\le 2\beta,\end{aligned}$$

where the second line follows from integration by parts. Using $\mathsf{E}(\|Av\|-\mathsf{E}\|Av\|)^2=\mathsf{E}\|Av\|^2-(\mathsf{E}\|Av\|)^2$ this yields $\mathsf{E}\|Av\|\ge\sqrt{p-2\beta}\ge\sqrt{p}-1$ given $p\ge C\beta$. Again by Lemma B.2 we obtain

$$\mathsf{P}\left(\big|\|Av\|-\sqrt{p}\big|>t+1\right)\le 2\exp(-t^2/\beta).$$

which implies

$$\mathsf{P}\left(\big|\|Av\|^2-p\big|>2\sqrt{p}(t+1)+(t+1)^2\right)\le 2\exp(-t^2/\beta). \tag{30}$$

Now, by a standard ϵ-net argument (taking $t=1.9\sqrt{\beta n}u$, cf. Vershynin (2018)), this implies

$$p-12u\sqrt{\beta np}-12\beta nu^2\le s_{\min}^2(A)\le s_{\max}^2(A)\le p+12u\sqrt{\beta np}+12\beta nu^2 \tag{31}$$

with probability at least $1-\exp(-nu^2)$, provided $u\ge 4$, $p\ge Cn$ and $n\ge C$. The desired conclusion (29) follows from the above inequality and the fact that $12\beta nu^2\le 4u\sqrt{\beta np}$ for $u\le\sqrt{p/16\beta n}$. □

Proof of Proposition B.4. Consider a truncated version of X defined by $\tilde{X}=X\mathbf{1}_{\beta^{-2}/200<|X|\le 3\beta}$. We have

$$\mathsf{E}\tilde{X}=\mathsf{E}\tilde{X}-\mathsf{E}X=\mathsf{E}X\mathbf{1}_{|X|\le\beta^{-2}/200}+\mathsf{E}X\mathbf{1}_{|X|>3\beta}.$$

It is trivial that $|\mathsf{E}X\mathbf{1}_{|X|\le\beta^{-2}/200}|\le\beta^{-2}/200$, while by subgaussian tail bound one may compute $|\mathsf{E}X\mathbf{1}_{|X|>3\beta}|\le\mathsf{E}|X|\mathbf{1}_{|X|>3\beta}\le 4\beta\exp(-9\beta)$. Since $\beta/\sqrt{2}\ge\mathsf{E}X^2=1$ we know $\beta\ge\sqrt{2}$ and hence $4\beta\exp(-9\beta)$ is much smaller than β^{-2}, say, $4\beta\exp(-9\beta)\le\beta^{-2}/500$. These imply

$$|\mathsf{E}\tilde{X}|\le\frac{1}{200\beta^2}+\frac{1}{500\beta^2}\le\frac{1}{100\beta^2}.$$

By a similar argument one may prove

$$|\mathsf{E}\tilde{X}^2-1|\le\frac{1}{200\beta^2}.$$

Let $\tilde{X}_+=\max(\tilde{X},0)$ and $\tilde{X}_-=\max(-\tilde{X},0)$, then $\tilde{X}_+,\tilde{X}_-$ are nonnegative and $\tilde{X}=\tilde{X}_+-\tilde{X}_-$, $\tilde{X}^2=\tilde{X}_+^2+\tilde{X}_-^2$. We thus have

$$|\mathsf{E}\tilde{X}_+-\mathsf{E}\tilde{X}_-|\le\frac{1}{100\beta^2}, \tag{32}$$

$$\mathsf{E}\tilde{X}_+^2+\mathsf{E}\tilde{X}_-^2\ge 1-\frac{1}{200\beta^2}. \tag{33}$$

Assume to the contrary that $\mathsf{P}(X>\beta^{-2}/200)<\beta^{-2}/20$, the following chain of reasoning would yield a contradiction, thereby proving $\mathsf{P}(X>\beta^{-2}/200)\ge\beta^{-2}/20$ as desired: since $|\tilde{X}|\le 3\beta$ and $\tilde{X}>0$ only if $X>\beta^{-2}/200$, we have

$$\mathsf{E}\tilde{X}_+\le 3\beta\mathsf{P}(X>\beta^{-2}/200)<3\beta^{-1}/20,$$

and

$$\mathsf{E}\tilde{X}_+^2\le 3\beta\mathsf{E}\tilde{X}_+<9/20.$$

However, from (32) we know

$$\mathsf{E}\tilde{X}_- \le \mathsf{E}\tilde{X}_+ + \frac{1}{100\beta^2} < \frac{3}{20\beta} + \frac{1}{100\beta} = \frac{4}{25\beta},$$

where we used $\beta \ge \sqrt{2}$ and hence $\beta^2 \ge \beta$. Therefore

$$\mathsf{E}\tilde{X}_-^2 \le 3\beta \mathsf{E}\tilde{X}_- < 12/25.$$

These together imply $\mathsf{E}\tilde{X}_+^2 + \mathsf{E}\tilde{X}_-^2 < 9/20 + 12/25 = 93/100 < 1 - \beta^{-2}/200$, contradicting (33). This completes the proof. □

C PROOF OF LEMMA 4.3 AND LEMMA 4.4

It turns out that Lemma 4.3 and Lemma 4.4 had better be proved in conjunction with each other and with another useful lemma (Lemma C.1 below). Before discussing the details, we prove a strengthened version of Proposition 4.1 in the main text which will be used later.

Proposition C.1. *If $(i,j) \in \mathcal{A}(t)$, then $\partial\sigma(\langle w_j^{(t)}, x_i\rangle) \ge \gamma$. Moreover, if $(i,j) \in \mathcal{A}(t) \cap \mathcal{A}(t+1)$, then $\xi_j(x_i; W^{(t)}) \ge \gamma$. In particular, we have*

$$\lambda_i(x_i; W^{(t)}) \ge \gamma^2 |\mathcal{A}^i(t) \cap \mathcal{A}^i(t+1)|/m.$$

Proof. The first assertion follows from the definition of $\mathcal{A}(t)$ which implies $\langle w_j^{(t)}, x_i\rangle > 0$, and the fact that $\partial\sigma(u) \ge \gamma$ when $u > 0$. The second assertion follows again from the definition, which implies $\langle w_j^{(t)}, x_i\rangle > 0$, $\langle w_j^{(t+1)}, x_i\rangle > 0$, thus by the assumption on σ we have

$$\xi_j(x_i; W^{(t)}) = \frac{\sigma(\langle w_j^{(t+1)}, x_i\rangle) - \sigma(\langle w_j^{(t)}, x_i\rangle)}{\langle w_j^{(t+1)}, x_i\rangle - \langle w_j^{(t)}, x_i\rangle} \ge \gamma.$$

The last assertion follows from

$$\begin{aligned}
\lambda_i(x_i; W^{(t)}) &= \frac{1}{m}\sum_{j=1}^m \xi_j(x_i; W^{(t)})\partial\sigma(\langle w_j^{(t+1)}, x_i\rangle) \\
&\ge \frac{1}{m} \sum_{j \in \mathcal{A}^i(t) \cap \mathcal{A}^i(t+1)} \xi_j(x_i; W^{(t+1)})\partial\sigma(\langle w_j^{(t)}, x_i\rangle) \\
&\ge \frac{\gamma^2}{m}|\mathcal{A}^i(t) \cap \mathcal{A}^i(t+1)|,
\end{aligned}$$

as desired. □

Return to the proof of Lemma 4.3 and Lemma 4.4. We will prove the in conjunction with the following lemma concerning the boundedness of the ratio $g_i(W^{(t)})/g_{i'}(W^{(t)})$ for $i, i' \in [n]$.

Lemma C.1. *On the event $\mathcal{E}$, for any $t \ge 0$ we have, with $C_r > 0$ some constant depending only on β, γ:*

$$\frac{\max_i g_i(W^{(t)})}{\min_i g_i(W^{(t)})} \le C_r. \tag{34}$$

Proof of Lemma 4.3, Lemma 4.4, and Lemma C.1. The proof is by a multiple induction on t. Denote by

$$\begin{aligned}
P(t):&\quad \mathcal{A}(\tau) \cap \mathcal{T} \subset \mathcal{A}(\tau+1) \cap \mathcal{T}, && \forall \tau \le t, \\
Q(t):&\quad \frac{\gamma^2\alpha p}{240\beta^2 n} g_i(W^{(\tau)}) \le y_i f_{W^{(\tau+1)}}(x_i) - y_i f_{W^{(\tau)}}(x_i) \le \frac{3\alpha p}{n} g_i(W^{(\tau)}), && \forall \tau \le t,\ \forall i \in [n], \\
R(t):&\quad \exp(y_{i'} f_{W^{(t)}}(x_{i'}) - y_i f_{W^{(t)}}(x_i)) \le C_r', && \forall i, i' \in [n].
\end{aligned}$$

the propositions that the conclusions of the three lemmas respectively hold at the t-th iteration (we will see soon how $R(t)$ implies Lemma C.1). We will show that $R(0)$ is true, and that for any $t \ge 0$ we have

$$R(t) \implies P(t), \quad P(t) \wedge R(t) \implies Q(t), \quad Q(t) \wedge R(t) \implies R(t+1).$$

It is evident that these together imply $P(t), Q(t), R(t)$ are true for all $t \geq 0$.

Before delving into the induction argument we first show how $R(t)$ implies Lemma C.1. In fact, observe that

$$\frac{1+\exp(z_1)}{1+\exp(z_2)} \leq 2(1+\exp(z_1-z_2)),\ \forall z_1, z_2 \in \mathbb{R}. \tag{35}$$

This is because the ratio is no more than $2/(1+\exp(z_2)) \leq 2$ when $z_1 \leq 0$, and is no more than $2\exp(z_1)/\exp(z_2) \leq 2\exp(z_1-z_2)$ when $z_1 > 0$. On the other hand, by similar arguments we have

$$\frac{1+\exp(z_1)}{1+\exp(z_2)} \geq \frac{1}{4}\exp(z_1-z_2),\ \forall z_1 \in \mathbb{R}, z_2 \geq -1, \tag{36}$$

which indicates $R(t)$ not only implies Lemma C.1, but is also equivalent to Lemma C.1 in some sense.

More precisely, since $g(z) = 1/(1+\exp(z))$, we have

$$\frac{g_i(W^{(t)})}{g_{i'}(W^{(t)})} \leq 2 + 2\exp(y_{i'} f_{W^{(t)}}(x_{i'}) - y_i f_{W^{(t)}}(x_i)), \tag{37}$$

which shows that were $R(t)$ true, we would have $g_i(W^{(t)})/g_{i'}(W^{(t)}) \leq 2 + 2C_r' \leq C_r$ as long as C_r is chosen to be larger than $2 + 2C_r$.

We now begin the induction argument.

Base case: $R(0)$**.** This follows from Eqn. (18) in the main text.

Induction Step I: $R(t) \implies P(t)$**.** For any $(i,j) \in \mathcal{A}(t) \cap \mathcal{T}$, we have $y_i a_j = 1$ and $\langle w_j^{(t)}, x_i\rangle > 0$ by definition. Thus

$$\begin{aligned}
&\langle w_j^{(t+1)}, x_i\rangle - \langle w_j^{(t)}, x_i\rangle \\
&= \frac{\alpha}{n\sqrt{m}} \sum_{k=1}^{n} y_k a_j g_k(W^{(t)}) \partial\sigma(\langle w_j^{(t)}, x_k\rangle)\langle x_k, x_i\rangle \\
&= \frac{\alpha}{n\sqrt{m}} g_i(W^{(t)}) \partial\sigma(\langle w_j^{(t)}, x_i\rangle)\|x_i\|^2 + \frac{\alpha}{n\sqrt{m}} \sum_{k\neq i} y_k a_j g_k(W^{(t)}) \partial\sigma(\langle w_j^{(t)}, x_k\rangle)\langle x_k, x_i\rangle \\
&\geq \frac{\alpha}{2n\sqrt{m}} g_i(W^{(t)})\gamma p - \frac{16\alpha}{\sqrt{m}} \hat{G}(W^{(t)}) \left(\|\mu^2\| + \sqrt{p\log(n/\delta)}\right) \\
&\geq \frac{\alpha\gamma p}{4C_r n\sqrt{m}} \hat{G}(W^{(t)}).
\end{aligned} \tag{38}$$

where in the second equality we used by definition that $y_i a_j = 1$, and in the penultimate inequality we used the following bounds: $\partial\sigma(\langle w_j^{(t)}, x_i\rangle) \geq \gamma$ since $\langle w_j^{(t)}, x_i\rangle > 0$; $\|x_i\|^2 \geq 5p/4 - \|\mu\|^2 \geq p/2$ by Eqns. (6c) and (14) in the main text; $|\langle x_i, x_{i_0}\rangle| \leq 16(\|\mu\|^2 + \sqrt{p\log(n/\delta)})$ by Eqns. (15) and (16) in the main text. The last inequality follows from $R(t)$, which implies $g_i(W^{(t)}) \geq C_r \hat{G}(W^{(t)})$, and the assumption (6c) that $p \geq Cn\|\mu\|^2 + Cn^2\log(n/\delta)$. An immediate consequence is that $\langle w_j^{(t+1)}, x_i\rangle > \langle w_j^{(t)}, x_i\rangle > 0$. This proves $(i,j) \in \mathcal{A}(t+1)$. Note that $(i,j) \in \mathcal{T}$ by definition, we have proved $(i,j) \in \mathcal{A}(t+1) \cap \mathcal{T}_j$. Since $(i,j) \in \mathcal{A}(t) \cap \mathcal{T}$ is arbitrary, this implies $\mathcal{A}(t) \cap \mathcal{T} \subset \mathcal{A}(t+1) \cap \mathcal{T}$, thereby proving $P(t)$ as desired.

Induction Step II: $P(t) \wedge R(t) \implies Q(t)$**.** It follows from (24) that

$$\begin{aligned}
y_i f_{W^{(t+1)}}(x_i) - y_i f_{W^{(t)}}(x_i) &= \frac{\alpha}{n} \sum_{k=1}^{n} g_k(W^{(t)}) \lambda_k(x_i; W^{(t)}) \langle y_k x_k, y_i x_i\rangle \\
&= \frac{\alpha}{n} \lambda_i(x_i; W^{(t)}) \|x_i\|^2 g_i(W^{(t)}) + \frac{\alpha}{n} \sum_{k\neq i} g_k(W^{(t+1)}) \lambda_k(x_i; W^{(t)}) \langle y_k x_k, y_i x_i\rangle
\end{aligned} \tag{39}$$

Recall that $\|x_i\|^2 \geq p/2$ as proved in Induction Step I, we have

$$\frac{\alpha}{n} g_i(W^{(t)}) \lambda_i(x_i; W^{(t)}) \|x_i\|^2 \geq \frac{\alpha p}{2n} \lambda_i(x_i; W^{(t)}) g_i(W^{(t)}).$$

But from the induction hypothesis $P(t)$ one may show $\lambda_i(x_i; W^{(t)}) \geq \gamma^2/(60\beta^2)$ since by Proposition C.1

$$\begin{aligned}\lambda_i(x_i; W^{(t)}) &\geq \frac{\gamma^2}{m}|\mathcal{A}^i(t) \cap \mathcal{A}^i(t+1)| \\ &\geq \frac{\gamma^2}{m}|\mathcal{A}^i(0) \cap \mathcal{T}^i| \geq \frac{\gamma^2}{60\beta^2},\end{aligned}$$

where the penultimate inequality follows from the induction hypothesis $P(t)$ since

$$\mathcal{A}^i(t) \cap \mathcal{A}^i(t+1) \supset (\mathcal{A}^i(t) \cap \mathcal{T}^i) \cap (\mathcal{A}^i(t+1) \cap \mathcal{T}^i) \supset \mathcal{A}^i(0) \cap \mathcal{T}^i,$$

and the last inequality follows from Lemma 4.2 in the main text. Consequently, we have

$$\frac{\alpha}{n} g_i(W^{(t)}) \lambda_i(x_i; W^{(t)}) \|x_i\|^2 \geq \frac{\gamma^2 \alpha p}{120\beta^2 n} g_i(W^{(t)}). \tag{40}$$

On the other hand, since $\lambda_k \leq 1$ and $\|x_i\|^2 \leq 2\|\mu\|^2 + 2\|x_i - y_i^\star \mu\|^2 \leq 2\|\mu\|^2 + 5p/4 \leq 2p$ we have

$$\frac{\alpha}{n} g_i(W^{(t)}) \lambda_i(x_i; W) \|x_i\|^2 \leq \frac{2\alpha p}{n} g_i(W^{(t)}). \tag{41}$$

Next we deal with the remainder term. From Eqns. (15) and (16) in the main text that we may show that $|\langle x_i, x_{i'} \rangle| \leq 16(\|\mu\|^2 + \sqrt{p \log(n/\delta)})$ for $i \neq i'$, thus

$$\frac{\alpha}{n} \left| \sum_{k \neq i}^{n} g_k(W^{(t)}) \lambda_k(x_i; W) \langle y_k x_k, y_i x_i \rangle \right| \leq 16\alpha \left(\|\mu\|^2 + \sqrt{p \log(n/\delta)} \right) \hat{G}(W^{(t)}). \tag{42}$$

By induction hypothesis $R(t)$ and (37) we have $\hat{G}(W^{(t)}) \leq C_r g_i(W^{(t)})$. Invoking the assumption (6c) that $p \geq Cn\|\mu\|^2 + Cn^2 \log(n/\delta)$ for sufficiently large C, we obtain

$$\frac{\alpha}{n} \left| \sum_{k \neq i}^{n} g_k(W^{(0)}) \lambda_k(x_i; W) \langle y_k x_k, y_i x_i \rangle \right| \leq \frac{\gamma^2 \alpha p}{240\beta^2 n} g_i(W^{(0)}).$$

The desired conclusion $Q(t)$ readily follows from summing up the above inequalities.

Induction step III: $Q(t) \wedge R(t) \implies R(t+1)$. We first show that bounding bounding $\exp(y_{i'} f_W(x_{i'}) - y_i f_W(x_i))$ is in a sense equivalent to bounding $g_i(W)/g_{i'}(W)$.

We return to bound $\exp(y_{i'} f_{W^{(t+1)}}(x_{i'}) - y_i f_{W^{(t+1)}}(x_i))$. By the induction hypothesis $Q(t)$ we have

$$\begin{aligned}&\exp(y_{i'} f_{W^{(t+1)}}(x_{i'}) - y_i f_{W^{(t+1)}}(x_i)) \\ &\leq \exp(y_{i'} f_{W^{(t)}}(x_{i'}) - y_i f_{W^{(t)}}(x_i)) \exp\left(\frac{\alpha p}{240\beta^2 n} (720\beta^2 g_{i'}(W^{(t)}) - \gamma^2 g_i(W^{(t)})) \right).\end{aligned}$$

We now see that the last factor is very small if $g_i(W^{(t)})/g_{i'}(W^{(t)})$ is large, hence shrinking the size of $g_i(W^{(t+1)})/g_{i'}(W^{(t+1)})$. This provides a negative feedback mechanism that ensures the ratio never grows too large. More precisely, we may distinguish two cases:

- If $g_i(W^{(t)})/g_{i'}(W^{(t)}) > 800\beta^2/\gamma^2$, then $720\beta^2 g_{i'}(W^{(t)}) - \gamma^2 g_i(W^{(t)}) < 0$ and hence

  $$\exp(y_{i'} f_{W^{(t+1)}}(x_{i'}) - y_i f_{W^{(t+1)}}(x_i)) \leq \exp(y_{i'} f_{W^{(t)}}(x_{i'}) - y_i f_{W^{(t)}}(x_i)) \leq C_r',$$

  where the last inequality follows from induction hypothesis $R(t-1)$.

- If $g_i(W^{(t)})/g_{i'}(W^{(t)}) \leq 800\beta^2/\gamma^2$, then by using $Q(t)$ again we have $y_i f_{W^{(t)}}(x_i) \geq -1$ and $y_{i'} f_{W^{(t)}}(x_{i'}) \geq -1$, thus (36) yields

  $$\exp(y_{i'} f_{W^{(t)}}(x_{i'}) - y_i f_{W^{(t)}}(x_i)) \leq \frac{4 g_i(W^{(t)})}{g_{i'}(W^{(t)})} \leq 3200\beta^2/\gamma^2.$$

On the other hand, since $g(z) \le 1$ we have $g_{i'}(W^{(t)}) \le 1$, thus

$$\exp\left(\frac{\alpha p}{240\beta^2 n}(720\beta^2 g_{i'}(W^{(t)}) - \gamma^2 g_i(W^{(t)}))\right) \le \exp\left(\frac{3\alpha p}{n}\right) \le 2,$$

where the last inequality follows from the assumption $\alpha \le 1/(Cp^2)$. To summarize, in this case we have

$$\exp(y_{i'} f_{W^{(t+1)}}(x_{i'}) - y_i f_{W^{(t+1)}}(x_i)) \le (3200\beta^2/\gamma^2) \cdot 2 = 6400\beta^2/\gamma^2 \le C_r', \tag{43}$$

as long as $C_r' \ge 6400\beta^2/\gamma^2$.

This completes the proof of $R(t+1)$. As a byproduct, from (37) we obtain

$$\frac{g_i(W^{(t)})}{g_{i'}(W^{(t)})} \le 2 + 2C_r' \le C_r, \tag{44}$$

by setting $C_r = 2 + 2C_r'$, which is what is desired by Lemma C.1. □

D PROOF OF LEMMA 4.5

By Lemma 4.4 and the monotonicity of g it is clear that $g_i(W^{(t)})$ is decreasing in t. This implies $g_i(W^{(t)}) \le g_i(W^{(0)}) \le 2/3$. Invoking Lemma 4.4 again, this in turn implies

$$y_i f_{W^{(t+1)}}(x_i) - y_i f_{W^{(t)}}(x_i) \le \frac{3\alpha p}{n} g_i(W^{(t)}) \le \frac{4\alpha p}{3n} \le 1,$$

since $\alpha \le 1/(Cp^2)$.

By Lemma 4.4 and mean-value theorem we have

$$\frac{\gamma^2 \alpha p}{240\beta^2 n} g(\theta_1) g_i(W^{(t)}) \le \ell(y_i f_{W^{(t)}}(x_i)) - \ell(y_i f_{W^{(t+1)}}(x_i)) \le \frac{3\alpha p}{n} g(\theta_2) g_i(W^{(t)}),$$

where $y_i f_{W^{(t)}}(x_i) \le \theta_1, \theta_2 \le y_i f_{W^{(t+1)}}(x_i) \le y_i f_{W^{(t)}}(x_i) + 1$. Since g is decreasing we have $g(\theta_1), g(\theta_2) \le g_i(W^{(t)})$. On the other hand, it is easy to verify that for $u \ge -1$ we have $1 \ge g(u+1)/g(u) \ge 1/10$, thus $g(\theta_1), g(\theta_2) \ge g_i(W^{(t)})/10$. Together these imply $g(\theta_1), g(\theta_2) \asymp g_i(W^{(t)})$ and hence

$$\ell(y_i f_{W^{(t)}}(x_i)) - \ell(y_i f_{W^{(t+1)}}(x_i)) \asymp_{\beta,\gamma} \frac{\alpha p}{n} g_i(W^{(t)})^2,$$

as desired.

E PROOF OF PROPOSITION 4.3

Let $b_t = \psi(t+1)a_t$. Then $b_0 \le 1$ by the assumption $\psi a_0 \le 1$. Moreover, $a_{t+1} \le a_t - \psi a_t^2$ is equivalent to

$$b_{t+1} \le \frac{t+2}{t+1} b_t - \frac{(t+2)}{(t+1)^2} b_t^2 = b_t + \frac{1}{t+1} b_t(1-b_t) - \frac{1}{(t+1)^2} b_t^2 \le b_t + \frac{1}{t+1} b_t(1-b_t).$$

Consequently, we have

$$1 - b_{t+1} \ge \left(1 - \frac{1}{t+1}\right)(1-b_t),$$

hence $1 - b_t \ge 0$ implies $1 - b_{t+1} \ge 0$ for all $t \ge 0$. Since $1 - b_0 \ge 0$, by induction we know that $1 - b_t \ge 0$ for all $t \ge 0$, i.e. $b_t \le 1$, therefore $a_t \le 1/(\psi(t+1)) \le 1/(\psi t)$.

F ANALYSIS OF GENERALIZATION

F.1 Proof of Lemma 4.7

Let us inspect the change of $\langle w_j^{(t)}, x\rangle$ for $(x, y) \sim \mathsf{P}_\star$. By (23) we have

$$\begin{aligned}\langle w_j^{(t+1)}, x\rangle - \langle w_j^{(t)}, x\rangle &= \frac{\alpha}{n\sqrt{m}} \sum_{i=1}^n y_i a_j g_i(W^{(t)}) \partial\sigma(\langle w_j^{(t)}, x_i\rangle)\langle y_i^\star \mu + z_i, y\mu + z\rangle \\ &= \frac{\alpha a_j y}{n\sqrt{m}} \sum_{i=1}^n y_i y g_i(W^{(t)}) \partial\sigma(\langle w_j^{(t)}, x_i\rangle)\langle y_i^\star \mu + z_i, y\mu + z\rangle,\end{aligned}$$

where we used $y^2 = 1$ to write $a_j = a_j y \cdot y$.

Denote $\delta_i = (y_i - y_i^\star)/2$, then $\delta_i \in \{-1, 1\}$ and $\sum_{i=1}^n |\delta_i| \le \eta n$. The above expression can be further written as

$$\begin{aligned}&a_j y(\langle w_j^{(t+1)}, x\rangle - \langle w_j^{(t)}, x\rangle) \\ &= \frac{\alpha}{n\sqrt{m}} \sum_{i=1}^n (y_i^\star + 2\delta_i) y g_i(W^{(t)}) \partial\sigma(\langle w_j^{(t)}, x_i\rangle)(y_i^\star y\|\mu\|^2 + y\langle \mu, z_i\rangle + y_i^\star\langle \mu, z\rangle + \langle z_i, z\rangle) \\ &= \underbrace{\frac{\alpha}{n\sqrt{m}} \sum_{i=1}^n g_i(W^{(t)}) \partial\sigma(\langle w_j^{(t)}, x_i\rangle)(\|\mu\|^2 + y_i^\star\langle \mu, z_i\rangle)}_{T_1(t)} + \underbrace{\frac{2\alpha}{n\sqrt{m}} \sum_{i=1}^n \delta_i g_i(W^{(t)})(y\|\mu\|^2 + y_i^\star y\langle \mu, z_i\rangle)}_{T_2(t)} \\ &\quad + \underbrace{\frac{\alpha}{n\sqrt{m}} \sum_{i=1}^n y_i y_i^\star y g_i(W^{(t)}) \partial\sigma(\langle w_j^{(t)}, x_i\rangle)\langle \mu, z\rangle}_{T_3(t)} + \underbrace{\frac{\alpha}{n\sqrt{m}} \left\langle \sum_{i=1}^n y_i y g_i(W^{(t)}) \partial\sigma(\langle w_j^{(t)}, x_i\rangle) z_i, z \right\rangle}_{T_4(t)}\end{aligned} \tag{45}$$

We control the four terms separately. For the first term we note that $|\langle \mu, z_i\rangle| \le \|\mu\|^2/2$ by Eqns. (15) and (7) in the main text, thus $\|\mu\|^2 + y_i^\star\langle \mu, z_i\rangle \ge \|\mu\|^2/2$, and all the summands is therefore nonnegative. Furthermore, by Lemma 4.3 we have $|\mathcal{A}_j(t)| \ge n/(60\beta^2)$, which together with Lemma C.1 implies $\sum_{i\in\mathcal{A}_j(t)} g_i(W^{(t)}) \ge C_r^{-1} n\hat{G}(W^{(t)})/(60\beta^2)$. Thus the first term

$$T_1(t) \ge \frac{\alpha}{120\beta^2 C_r\sqrt{m}} \gamma\|\mu\|^2 \hat{G}(W^{(t)}). \tag{46}$$

For the second term we use again $|\langle \mu, z_i\rangle| \le \|\mu\|^2/2$ and recall that $\sum_{i=1}^n |\delta_i| \le \eta n$, which together with Lemma C.1 yields

$$|T_2(t)| \le \frac{3\eta C_r \alpha}{\sqrt{m}} \|\mu\|^2 \hat{G}(W^{(t)}). \tag{47}$$

For the third term we simply bound

$$|T_3(t)| \le \frac{C_r\alpha}{\sqrt{m}} \hat{G}(W^{(t)})|\langle \mu, z\rangle|. \tag{48}$$

Define an event $\mathcal{E}_1'$ by

$$\mathcal{E}_1' := \left\{ |\langle \mu, z\rangle| \le 64\sqrt{\beta \log(p/n)}\|\mu\| \right\}.$$

By Proposition B.1 we have $\mathsf{P}(\mathcal{E}_1') \ge 1 - (p/n)^{-11}$. Conditioning on $\mathcal{E}_1'$ we have

$$|T_3(t)| \le \frac{64 C_r \alpha \|\mu\|}{\sqrt{m}} \sqrt{\beta \log(p/n)} \hat{G}(W^{(t)}). \tag{49}$$

By assumptions (6c) and (7) in the main text, the right hand sides of (47) and (49) can be absorbed by the right hand side of (46): by the assumption $\eta \le 1/C$ for sufficiently large we have $3\eta C_r \le 1/(960\beta^2 C_r)$; by the assumption (7) we have $\|\mu\| \ge C\sqrt{\log(p/n)}$ and hence $64 C_r \|\mu\| \sqrt{\beta \log(p/n)} \le \|\mu\|^2/(960\beta^2 C_r)$. These combined with (46), (47), (49) imply

$$T_1(t) + T_2(t) + T_3(t) \ge \frac{c\alpha}{\sqrt{m}} \|\mu\|^2 \hat{G}(W^{(t)}), \tag{50}$$

where $c > 0$ is some constant depending only on β, γ. In fact, one may take $c = \gamma/(160\beta^2 C_r)$.

It remains to control the last term $T_4(t)$. This is where the difference between the local approach and the global approach arises. We discuss these two approaches respectively.

The Local Approach. To control $\langle w_j^{(t)}, x\rangle$ with the local approach, we need to control all of $|\langle w_j^{(\tau+1)} - w_j^{(\tau)}, x\rangle|$ for $\tau < t$. The only efficient way to achieve this goal is to bound $\max_i |\langle z_i, z\rangle|$ so that

$$|T_4(\tau)| \le \frac{\alpha}{\sqrt{m}} \max_{i\in[n]} g_i(W^{(\tau)}) \max_{i\in[n]} |\langle z_i, z\rangle| \le \frac{C_r\alpha}{\sqrt{m}} \max_{i\in[n]} |\langle z_i, z\rangle| \hat{G}(W^{(\tau)})$$

holds uniformly in τ. The desired bound (Eqn. (21) in the main text) follows immediately from this and (50).

The Global Approach. To control the last term, it turns out useful to adopt a new viewpoint that one should look at the cumulative effect of the last term. More precisely, the cumulative effect of T_4 in the difference $\langle w_j^{(t)}, x\rangle - \langle w_j^{(0)}, x\rangle$ is $\sum_{\tau=0}^{t-1} T_4(\tau)$. One may invoke Eqn. (14) in the main text to see

$$\begin{aligned}\left\|\sum_{i=1}^{n}\sum_{\tau=0}^{t-1} y_i y g_i(W^{(\tau)}) \partial\sigma(\langle w_j^{(\tau)}, x_i\rangle) z_i\right\|^2 &\le 2p\sum_{i=1}^{n}\left(\sum_{\tau=0}^{T} g_i(W^{(\tau)})\partial\sigma(\langle w_j^{(\tau)}, x_i\rangle)\right)^2 \\ &\le 2C_r^2 pn\left(\sum_{\tau=0}^{t-1}\hat{G}(W^{(t)})\right)^2,\end{aligned}$$

thus

$$\left|\sum_{\tau=0}^{t-1} T_4(\tau)\right| \le \frac{2C_r\alpha\sqrt{p}}{\sqrt{nm}}\left(\sum_{\tau=0}^{t-1}\hat{G}(W^{(t)})\right)|\langle \phi_j^{(t)}, z\rangle| \tag{51}$$

where $\phi_j^{(t)} \in \mathbb{R}^p$ is some vector independent of z satisfying $\|\phi_j^{(t)}\| \le 1$.

Summing up, we have

$$\begin{aligned} a_j y(\langle w_j^{(t)}, x\rangle - \langle w_j^{(0)}, x\rangle) &= \sum_{\tau=0}^{t-1} T_1(\tau) + T_2(\tau) + T_3(\tau) + T_4(\tau) \\ &\ge \frac{c\alpha}{\sqrt{m}}\left(\|\mu\|^2 - C'\sqrt{\frac{p}{n}}|\langle \phi_j^{(t)}, z\rangle|\right)\sum_{\tau=0}^{t-1}\hat{G}(W^{(\tau)}). \end{aligned} \tag{52}$$

Define yet another event $\mathcal{E}_2'$ as

$$\mathcal{E}_2' = \left\{\max_{j\in[m]} |\langle \phi_j^{(t)}, z\rangle| \le 64\sqrt{\beta\log(pm/n)}\right\}.$$

It follows from Proposition B.1 again that $\mathsf{P}(\mathcal{E}_2') \ge 1 - (p/n)^{-11}$. On this event we have

$$\begin{aligned} a_j y(\langle w_j^{(t)}, x\rangle - \langle w_j^{(0)}, x\rangle) &= \sum_{\tau=0}^{t-1} T_1(\tau) + T_2(\tau) + T_3(\tau) + T_4(\tau) \\ &\ge \frac{c\alpha}{\sqrt{m}}\left(\|\mu\|^2 - C'\sqrt{\frac{p\log(pm/n)}{n}}\right)\sum_{\tau=0}^{t-1}\hat{G}(W^{(\tau)}). \end{aligned}$$

This proves the desired bound (Eqn. (22) in the main text) conditioning on $\mathcal{E}_1' \cap \mathcal{E}_2'$. By union bound this event occurs with probability at least $1 - 2(p/n)^{-11} \ge 1 - (p/n)^{-10}$, as desired.

F.2 Proof of Corollary 4.1

By Lemma 4.7 and its proof, we know that Eqn. (22) in the main text holds with probability at least $1 - 2(p/n)^{-11}$ (cf. the last paragraph in the proof of Lemma 4.7 in Section F.1). By assumption (7) in the main text we have $\|\mu\|^2 \ge$

$3C'\sqrt{p\log(pm/n)/n}$, hence

$$a_j y\langle w_j^{(t)}, x\rangle \geq a_j y\langle w_j^{(0)}, x\rangle + \frac{2c\alpha}{3\sqrt{m}}\|\mu\|^2 \sum_{\tau=0}^{t-1} \hat{G}(W^{(\tau)}).$$

To prove Corollary 4.1, it suffices to show

$$|\langle w_j^{(0)}, x\rangle| \leq \frac{c\alpha}{3\sqrt{m}}\|\mu\|^2 \hat{G}(W^{(0)}) \leq \frac{c\alpha}{3\sqrt{m}}\|\mu\|^2 \sum_{\tau=0}^{t-1} \hat{G}(W^{(\tau)}).$$

Recalling inequality (17), we invoke Proposition B.1 to see

$$|\langle w_j^{(0)}, x\rangle| \leq \|w_j^{(0)}\|\|\mu\| + 32\sqrt{\beta\log(p/n)}\|w_j^{(0)}\| \leq 2\|\mu\|\|w_j^{(0)}\| \leq 4\omega_{\mathsf{init}}\|\mu\|\sqrt{p},$$

with probability at least $1-(p/n)^{-11}$, where the penultimate inequality follows from the assumption (7). On the other hand, it follows from Eqn. (18) in the main text that $g_i(W^{(0)}) \geq g(1) \geq 1/4$, thus $\hat{G}(W^{(0)}) \geq 1/4$. Thus we may proceed to bound, recalling $\omega_{\mathsf{init}} \leq \alpha/\sqrt{mp}$ by (6f),

$$|\langle w_j^{(0)}, x\rangle| \leq 4\omega_{\mathsf{init}}\|\mu\|\sqrt{p} \leq 4\alpha\|\mu\|/\sqrt{m} \leq \frac{12\alpha}{\sqrt{m}}\|\mu\|\hat{G}(W^{(0)}) \leq \frac{c\alpha}{3\sqrt{m}}\|\mu\|^2\hat{G}(W^{(0)}),$$

where the last inequality follows from assumption (7) in the main text which implies $\|\mu\|^2 \geq 3c^{-1}\|\mu\|$. Overall, the probability that the above happens is at least $1-2(p/n)^{-11}-(p/n)^{-11} \geq 1-(p/n)^{-10}$. This completes the proof.

F.3 Proof of Lemma 4.8

By Lemma 4.7, we have a firm grasp of the the behavior of $\langle w_j^{(t)}, x\rangle$ on "good events". It remains to handle the "exceptional" bad events, described by the following lemma.

Lemma F.1. *For any $t>0$ and for any $x=y\mu+z$ with $y\in\{-1,1\}$, we have for some constant $C'>0$ depending only on β,γ that*

$$|\langle w_j^{(t)}, x\rangle| \leq \frac{C'\alpha}{\sqrt{m}}(\|\mu\|^2 + \|\mu\|\|z\| + \sqrt{p}\|z\|)\sum_{\tau=0}^{t-1}\hat{G}(W^{(\tau)}),\ \forall j\in[m]. \tag{53}$$

Proof. As in the proof of Lemma 4.7 in Section F.1, we write

$$\langle w_j^{(t)}, x\rangle = \langle w_j^{(0)}, x\rangle + \sum_{\tau=0}^{t-1}(\langle w_j^{(\tau+1)}, x\rangle - \langle w_j^{(\tau)}, x\rangle),$$

and decompose $\langle w_j^{(\tau+1)}, x\rangle - \langle w_j^{(\tau)}, x\rangle$ as in (45) and bound the term $T_1(\tau)$, $T_2(\tau)$ as in (46), (47). To bound $T_3(\tau)$ and $T_4(\tau)$ we again use (48) and (51), but we proceed by using the simplest bound instead of the probabilistic argument there: $|\langle\mu, z\rangle| \leq \|\mu\|\|z\|$ and $|\langle\phi_j^{(t)}, z\rangle| \leq \|z\|$, thus

$$|T_3(\tau)| \leq \frac{C_r\alpha}{\sqrt{m}}\hat{G}(W^{(t)})|\langle\mu, z\rangle| \leq \frac{C_r\alpha}{\sqrt{m}}\|\mu\|\|z\|\hat{G}(W^{(t)}),$$

and

$$\left|\sum_{\tau=0}^{t-1} T_4(\tau)\right| \leq \frac{2C_r\alpha}{\sqrt{nm}}\sqrt{p}\|z\|\left(\sum_{\tau=0}^{t-1}\hat{G}(W^{(t)})\right).$$

The conclusion then follows from the same argument as in the proof of Lemma 4.7, Section F.1. □

Return to the proof of Lemma 4.8. We rewrite the margin $yf_{W^{(t)}}(x)$ as

$$yf_{W^{(t)}}(x) = \frac{1}{\sqrt{m}}\sum_{j=1}^{m} a_j y\sigma(\langle w_j^{(t)}, x\rangle) = \frac{1}{\sqrt{m}}\sum_{j:a_j=y}\sigma(\langle w_j^{(t)}, x\rangle) - \frac{1}{\sqrt{m}}\sum_{j:a_j\neq y}\sigma(\langle w_j^{(t)}, x\rangle).$$

Handling good event. Denote by $\mathcal{E}_1$ the event that the conclusion of Corollary 4.1 holds, then $\mathsf{P}(\mathcal{E}_1) \geq 1 - (p/n)^{-10}$ by that corollary. Observe that on $\mathcal{E}_1$ we have $\langle w_j^{(t)}, x\rangle > 0$ if $a_j = y$ and $\langle w_j^{(t)}, x\rangle < 0$ if $a_j \neq y$ by the same corollary. Recalling our assumption on σ, it is obvious that $\sigma(u) \geq \gamma u$ if $u > 0$ and $\sigma(u) \leq 0$ if $u < 0$, thus on $\mathcal{E}_1$ we have

$$yf_{W^{(t)}}(x) \geq \frac{1}{\sqrt{m}} \sum_{j:a_j=y}^{m} \gamma \langle w_j^{(t)}, x\rangle \geq \frac{c\alpha |\{j : a_j = y\}|}{m} \|\mu\|^2 \sum_{\tau=0}^{t-1} \hat{G}(W^{(\tau)}) \geq \frac{c\alpha}{3} \|\mu\|^2 \sum_{\tau=0}^{t-1} \hat{G}(W^{(\tau)}),$$

where the last inequality follows from Eqn. (11) in the main text. These arguments imply

$$\mathsf{E}_{(x,y)\sim \mathsf{P}_\star}(yf_{W^{(t)}}(x)\mathbf{1}_{\mathcal{E}_1}) \geq \frac{c\alpha}{3} \|\mu\|^2 \mathsf{P}(\mathcal{E}_1) \sum_{\tau=0}^{t-1} \hat{G}(W^{(\tau)}) \geq \frac{c\alpha}{4} \|\mu\|^2 \sum_{\tau=0}^{t-1} \hat{G}(W^{(\tau)}), \tag{54}$$

where the last inequality follows from $\mathsf{P}(\mathcal{E}_1) \geq 1 - (p/n)^{-10} \geq 3/4$.

Handling exceptional event. We proceed to handle the exceptional case $\mathcal{E}_1^c$. To this end we apply Lemma F.1 and the idea that $\|z\|$ is essentially $O(\sqrt{p})$. More precisely, since $|\sigma(u)| \leq u$ by Lipschitz continuity of σ, we have

$$\begin{aligned}
\left|\mathsf{E}_{(x,y)\sim \mathsf{P}_\star}(yf_{W^{(t)}}(x)\mathbf{1}_{\mathcal{E}_1^c})\right| &\leq \mathsf{E}_{(x,y)\sim \mathsf{P}_\star} \frac{1}{\sqrt{m}} \sum_{j=1}^{m} |\langle w_j^{(t)}, x\rangle| \mathbf{1}_{\mathcal{E}_1^c} \\
&\leq C'\alpha \|\mu\|^2 \mathsf{P}(\mathcal{E}_1^c) \sum_{\tau=0}^{t-1} \hat{G}(W^{(\tau)}) + C'\alpha(\|\mu\| + \sqrt{p}) \mathsf{E}(\mathbf{1}_{\mathcal{E}_1^c} \|z\|) \sum_{\tau=0}^{t-1} \hat{G}(W^{(\tau)}),
\end{aligned}$$

We already know that $\mathsf{P}(\mathcal{E}_1^c) \leq (p/n)^{-10}$, which is sufficient to control the first term in the above expression. It remains to control the second term, which amounts to bounding $\mathsf{E}(\mathbf{1}_{\mathcal{E}_1^c}\|z\|)$. Denote

$$\mathcal{E}_2 = \left\{\|z\| \leq 40\sqrt{\beta p \log(p/n)}\right\},$$

then we have

$$\mathsf{E}(\mathbf{1}_{\mathcal{E}_1^c}\mathbf{1}_{\mathcal{E}_2}\|z\|) \leq 40\mathsf{E}(\mathbf{1}_{\mathcal{E}_1^c}\sqrt{p\log(p/n)}) \leq 40\sqrt{\beta p \log(p/n)}\mathsf{P}(\mathcal{E}_1^c) \leq (p/n)^{-8},$$

where in the last inequality we used the assumption (6c); while

$$\mathsf{E}(\mathbf{1}_{\mathcal{E}_1^c}\mathbf{1}_{\mathcal{E}_2^c}\|z\|) \leq \mathsf{E}(\mathbf{1}_{\mathcal{E}_2^c}\|z\|) \leq (p/n)^{-10},$$

where the last inequality follows from Proposition B.2. Summing up the above two inequalities we obtain

$$\left|\mathsf{E}_{(x,y)\sim \mathsf{P}}(yf_{W^{(t)}}(x)\mathbf{1}_{\mathcal{E}_1^c})\right| \leq C'\alpha(p/n)^{-8}(\|\mu\|^2 + \|\mu\| + \sqrt{p}) \sum_{\tau=0}^{t-1} \hat{G}(W^{(\tau)}) \leq \frac{c\alpha}{8} \|\mu\|^2 \sum_{\tau=0}^{t-1} \hat{G}(W^{(\tau)}) \tag{55}$$

where the last inequality follows from the assumptions (6c) and (7), assuming the constant C there is sufficiently large so that $C > 8C'/c$.

Putting things together. Combining (54) and (55) we obtain

$$\begin{aligned}
\mathsf{E}_{(x,y)\sim \mathsf{P}_\star}(yf_{W^{(t)}}(x)) &= \mathsf{E}_{(x,y)\sim \mathsf{P}_\star}(yf_{W^{(t)}}(x)\mathbf{1}_{\mathcal{E}_1}) + \mathsf{E}_{(x,y)\sim \mathsf{P}_\star}(yf_{W^{(t)}}(x)\mathbf{1}_{\mathcal{E}_1^c}) \\
&\geq \mathsf{E}_{(x,y)\sim \mathsf{P}_\star}(yf_{W^{(t)}}(x)\mathbf{1}_{\mathcal{E}_1}) - \left|\mathsf{E}_{(x,y)\sim \mathsf{P}_\star}(yf_{W^{(t)}}(x)\mathbf{1}_{\mathcal{E}_1^c})\right| \\
&\geq \frac{c\alpha}{8} \|\mu\|^2 \sum_{\tau=0}^{t-1} \hat{G}(W^{(\tau)}),
\end{aligned}$$

as desired. This completes the proof.

F.4 Proof of Lemma 4.9

By (23) we have

$$\|w_j^{(t+1)} - w_j^{(t)}\| = \alpha\|\nabla_{w_j}\hat{L}(W^{(t)})\| \le \frac{\alpha}{n\sqrt{m}}\sum_{i=1}^{n} g_i(W^{(t)})\|\mu\| + \frac{\alpha}{n\sqrt{m}}\left\|\sum_{i=1}^{n} y_i g_i(W^{(t)})\partial\sigma(\langle w_j^{(t)}, x_i\rangle)(x_i - y_i^\star\mu)\right\|.$$

But from Eqn. (14) in the main text we have

$$\left\|\sum_{i=1}^{n} y_i g_i(W^{(t)})\partial\sigma(\langle w_j^{(t)}, x_i\rangle)(x_i - y_i^\star\mu)\right\|^2 \le \frac{5p}{4}\sum_{i=1}^{n} y_i^2 g_i(W^{(t)})^2\partial\sigma(\langle w_j^{(t)}, x_i\rangle)^2 \le \frac{5C_r p}{4} n\hat{G}(W^{(t)})^2,$$

where in the last inequality we used Lemma C.1. Taking square roots and plugging this into the first equation, we obtain

$$\|w_j^{(t+1)} - w_j^{(t)}\| \le \frac{\alpha}{\sqrt{m}}\|\mu\|\hat{G}(W^{(t)}) + \alpha\sqrt{\frac{5C_r p}{4nm}}\hat{G}(W^{(t)}) \le C'\alpha\sqrt{\frac{p}{nm}}\hat{G}(W^{(t)}),$$

with $C' = 2\sqrt{C_r} + 1$, where we used the assumption (6c) in the main text to show $\|\mu\| \le \sqrt{p/n}$. This completes the proof.