POLYLOGARITHMIC WIDTH SUFFICES FOR GRADIENT DESCENT TO ACHIEVE ARBITRARILY SMALL TEST ERROR WITH SHALLOW RELU NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has revealed that overparameterized networks trained by gradient descent achieve arbitrarily low training error, and sometimes even low test error. The required width, however, is always polynomial in at least one of the sample size n, the (inverse) training error $1/\epsilon$, and the (inverse) failure probability $1/\delta$. This work shows that $\widetilde{O}(1/\epsilon)$ iterations of gradient descent on two-layer networks of any width exceeding $\mathrm{polylog}(n, 1/\epsilon, 1/\delta)$ and $\widetilde{\Omega}(1/\epsilon^2)$ training examples suffices to achieve a test error of ϵ. The analysis further relies upon a margin property of the limiting kernel, which is guaranteed positive, and can distinguish between true labels and random labels.

1 INTRODUCTION

Despite the extensive empirical success of deep networks, their optimization and generalization properties are still not well understood. Recently, the neural tangent kernel (NTK) has provided some insight into the problem. In the infinite-width limit, the NTK converges to a limiting kernel which stays constant during training; on the other hand, when the width is large enough, the function learned by gradient descent follows the NTK (Jacot et al., 2018). This motivates the study of overparameterized networks trained by gradient descent, using properties of the NTK. In fact, parameters related to NTK, such as the minimum eigenvalue of the limiting kernel, appear to affect optimization and generalization (Arora et al., 2019).

However, in addition to such NTK-dependent parameters, prior work also requires the width to depend polynomially on n, $1/\delta$ or $1/\epsilon$, where n denotes the size of the training set, δ denotes the failure probability, and ϵ denotes the target error. These large widths far exceed what is used empirically, constituting a significant gap between theory and practice.

Our contributions. In this paper, we narrow this gap by showing that a two-layer ReLU network with $\Omega(\ln(n/\delta)+\ln(1/\epsilon)^2)$ hidden units trained by gradient descent achieves classification error ϵ *on test data*, meaning both optimization and generalization occur. Unlike prior work, the width is fully polylogarithmic in n, $1/\delta$, and $1/\epsilon$; the width will additionally depend on the *separation margin* of the limiting kernel, a quantity which is guaranteed positive (assuming no inputs are duplicated with noisy labels), and can distinguish between true labels and random labels. The paper organization together with some details are described below.

Section 2 studies gradient descent on the training set. Using the ℓ_1 geometry inherent in classification tasks, we prove that with any width at least polylogarithmic and any constant step size no larger than 1, gradient descent achieves training error ϵ in $\widetilde{O}(1/\epsilon)$ iterations (cf. Theorem 2.2). As is common in the NTK literature (Chizat & Bach, 2019), we also show the parameters hardly change, which will be essential to our generalization analysis.

Section 3 gives a test error bound. Concretely, using the preceding gradient descent analysis, and standard Rademacher tools but exploiting how little the weights moved, we show that with $\widetilde{\Omega}(1/\epsilon^2)$ samples and $\widetilde{O}(1/\epsilon)$ iterations, gradient descent finds a solution with ϵ test error (cf. Theorem 3.2 and Corollary 3.3). (As discussed in Remark 3.4, $\widetilde{\Omega}(1/\epsilon)$ samples also suffices via a smoothness-based generalization bound, at the expense of large constant factors.)

Section 4 considers stochastic gradient descent (SGD) with access to a standard stochastic online oracle. We prove that with width at least polylogarithmic and sample complexity $\widetilde{O}(1/\epsilon)$, SGD achieves an arbitrarily small test error (cf. Theorem 4.1).

Section 5 discusses the separability condition, which is in general a positive number, but reflects the difficulty of the classification problem. Regarding random labels, we show that starting from a distribution with a good positive margin, but replacing the labels with random noise, the margin can degrade all the way down to $O(1/\sqrt{n})$, which (correctly) removes the possibility of generalization. In this way, our analysis can distinguish between true labels and random labels.

Section 6 concludes with some open problems.

1.1 Related work

There has been a large literature studying gradient descent on overparameterized networks via the NTK. The most closely related work is (Nitanda & Suzuki, 2019), which shows that a two-layer network trained by gradient descent with the logistic loss can achieve a small test error, under the same assumption that the neural tangent model with respect to the first layer can separate the data distribution. However, they analyze smooth activations, while we handle the ReLU. They require $\Omega(1/\epsilon^2)$ hidden units, $\widetilde{\Omega}(1/\epsilon^4)$ data samples, and $O(1/\epsilon^2)$ steps, while our result only needs polylogarithmic hidden units, $\widetilde{\Omega}(1/\epsilon^2)$ data samples, and $\widetilde{O}(1/\epsilon)$ steps.

Additionally on shallow networks, Du et al. (2018b) prove that on an overparameterized two-layer network, gradient descent can globally minimize the empirical risk with the squared loss. Their result requires $\Omega(n^6/\delta^3)$ hidden units. Oymak & Soltanolkotabi (2019) further reduces the required overparameterization, but it still has a $\mathrm{poly}(n)$ dependency. Based on the result in (Du et al., 2018b), Arora et al. (2019) further show that the two-layer network learned by gradient descent can achieve a small test error, assuming that on the data distribution the smallest eigenvalue of the limiting kernel is at least some positive constant. They also give a fine-grained characterization of the predictions made by gradient descent iterates; such a characterization makes use of a special property of the squared loss and cannot be applied to the logistic regression setting. Li & Liang (2018) show that stochastic gradient descent (SGD) with the cross entropy loss can learn a two-layer network with small test error, using $\mathrm{poly}(\ell, 1/\delta, 1/\epsilon)$ hidden units, where ℓ is at least the covering number of the unit sphere using balls whose radii are no larger than the smallest distance between two data points with different labels. In a high-dimensional space, ℓ could be very large. Allen-Zhu et al. (2018a) consider SGD on a two-layer network, and a variant of SGD on a three-layer network. The three-layer analysis further exhibits some properties not captured by the NTK. They assume a ground truth network with infinite-order smooth activations, and they require the width to depend polynomially on $1/\epsilon$ and some constants related to the smoothness of the activations of the ground truth network.

On deep networks, a variety of works have established low training error Allen-Zhu et al. (2018b); Du et al. (2018a); Zou et al. (2018); Zou & Gu (2019). Cao & Gu (2019a) assume that the neural tangent model with respect to the second layer of a two-layer network can separate the data distribution, and prove that gradient descent on a deep network can achieve ϵ test error with $\Omega(1/\epsilon^4)$ samples and $\Omega(1/\epsilon^{14})$ hidden units. Cao & Gu (2019b) consider SGD with an online oracle and give a general result. Under the same assumption as in (Cao & Gu, 2019a), their result requires $\Omega(1/\epsilon^{14})$ hidden units and sample complexity $\widetilde{O}(1/\epsilon^2)$. By contrast, with the same online oracle, our result only needs polylogarithmic hidden units and sample complexity $\widetilde{O}(1/\epsilon)$.

1.2 Notation

The dataset is denoted by $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$. For simplicity, we assume that $\|x_i\|_2 = 1$ for any $1 \le i \le n$, which is standard in the NTK literature.

The two-layer network has weight matrices $W \in \mathbb{R}^{m\times d}$ and $a \in \mathbb{R}^m$. We use the following parameterization, which is also used in (Du et al., 2018b; Arora et al., 2019).

$$f(x; W, a) := \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \sigma\left(\langle w_s, x\rangle\right),$$

with initialization

$$w_{s,0} \sim \mathcal{N}(0, I_d), \quad \text{and} \quad a_s \sim \text{unif}\left(\{-1, +1\}\right).$$

Note that in this paper, $w_{s,t}$ denotes the s-th row of W at step t. We fix a and only train W, as in (Li & Liang, 2018; Du et al., 2018b; Arora et al., 2019; Nitanda & Suzuki, 2019). We consider the ReLU activation $\sigma(z) := \max\{0, z\}$, though our analysis can be extended easily to Lipschitz continuous, positively homogeneous activations such as leaky ReLU.

We use the logistic (binary cross entropy) loss $\ell(z) := \ln\left(1 + \exp(-z)\right)$ and gradient descent. For any $1 \le i \le n$ and any W, let $f_i(W) := f(x_i; W, a)$. The empirical risk and its gradient are given by

$$\widehat{\mathcal{R}}(W) := \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i f_i(W)\right), \quad \text{and} \quad \nabla\widehat{\mathcal{R}}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell'\left(y_i f_i(W)\right) y_i \nabla f_i(W).$$

For any $t \ge 0$, the gradient descent step is given by $W_{t+1} := W_t - \eta_t \nabla\widehat{\mathcal{R}}(W_t)$. Also define

$$f_i^{(t)}(W) := \left\langle \nabla f_i(W_t), W\right\rangle, \quad \text{and} \quad \widehat{\mathcal{R}}^{(t)}(W) := \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i f_i^{(t)}(W)\right).$$

Note that $f_i^{(t)}(W_t) = f_i(W_t)$. This property generally holds due to homogeneity: for any W and any $1 \le s \le m$,

$$\frac{\partial f_i}{\partial w_s} = \frac{1}{\sqrt{m}} a_s \mathbb{1}\left[\langle w_s, x_i\rangle > 0\right] x_i, \quad \text{and} \quad \left\langle \frac{\partial f_i}{\partial w_s}, w_s \right\rangle = \frac{1}{\sqrt{m}} a_s \sigma\left(\langle w_s, x_i\rangle\right),$$

and thus $\left\langle \nabla f_i(W), W\right\rangle = f_i(W)$.

2 Empirical risk minimization

In this section, we consider a fixed training set and empirical risk minimization. We first state our assumption on the separability of the neural tangent model, and then give our main result and a proof sketch.

Here is some additional notation. Let $\mu_{\mathcal{N}}$ denote the Gaussian measure on $\mathbb{R}^d$, given by the Gaussian density with respect to the Lebesgue measure on $\mathbb{R}^d$. We consider the following Hilbert space

$$\mathcal{H} := \left\{ w : \mathbb{R}^d \to \mathbb{R}^d \,\middle|\, \int \|w(z)\|_2^2 \,\mathrm{d}\mu_{\mathcal{N}}(z) < \infty \right\}.$$

For each $1 \le i \le n$, define $\phi_i \in \mathcal{H}$ by

$$\phi_i(z) := x_i \mathbb{1}\left[\langle z, x_i\rangle > 0\right].$$

One can verify that $\|\phi_i\|_{\mathcal{H}}^2 = 1/2$, and thus ϕ_i is indeed in $\mathcal{H}$.

Assumption 2.1. There exists $\bar{v} \in \mathcal{H}$ and $\gamma > 0$, s.t. $\left\|\bar{v}(z)\right\|_2 \le 1$ for any $z \in \mathbb{R}^d$, and for any $1 \le i \le n$,

$$y_i \langle \bar{v}, \phi_i\rangle := y_i \int \left\langle \bar{v}(z), \phi_i(z)\right\rangle \mathrm{d}\mu_{\mathcal{N}}(z) \ge \gamma.$$

$\Diamond$

A more natural assumption is that the infinite-width limit of the NTK can separate the training set with a positive margin. This assumption actually implies Assumption 2.1 (cf. Proposition 5.1). Some other discussion on the separability assumption is also given in Section 5.

With Assumption 2.1, we state our main empirical risk result.

Theorem 2.2. *Under Assumption 2.1, given any risk target* $\epsilon \in (0,1)$ *and any* $\delta \in (0, 1/3)$*, let*

$$\lambda := \frac{\sqrt{2\ln(4n/\delta)} + \ln(4/\epsilon)}{\gamma/3}, \quad \text{and} \quad M := \max\left\{\frac{162\ln(n/\delta)}{\gamma^2}, 25\ln\left(\frac{2n}{\delta}\right), \frac{324\lambda^2}{\gamma^3}\right\}.$$

Then for any $m \geq M$ *and any constant step size* $\eta \leq 1$*, with probability* $1 - 3\delta$ *over the random initialization,*

$$\frac{1}{T}\sum_{t<T} \widehat{\mathcal{R}}(W_t) \leq \epsilon, \quad \text{where} \quad T := \lceil 2\lambda^2/\eta\epsilon \rceil.$$

Moreover for any $0 \leq t < T$*,*

$$\|W_t - W_0\|_F \leq \sqrt{6}\lambda, \quad \text{and} \quad \left\|w_{s,t} - w_{s,0}\right\|_2 \leq \frac{3\sqrt{6}\lambda}{\gamma\sqrt{m}} \text{ for any } 1 \leq s \leq m.$$

While the number of hidden units required by prior work all have a polynomial dependency on n, $1/\delta$ or $1/\epsilon$, Theorem 2.2 only requires $m = \Omega\left(\ln(n/\delta) + \ln(1/\epsilon)^2\right)$. In the rest of Section 2, we give a proof sketch of Theorem 2.2.

2.1 Properties at initialization

In this subsection, we give some nice properties of random initialization. The proofs are given in Appendix A.

Given an initialization (W_0, a), for any $1 \leq s \leq m$, define

$$\bar{u}_s := \frac{1}{\sqrt{m}} a_s \bar{v}(w_{s,0}), \tag{2.1}$$

where $\bar{v}$ is given by Assumption 2.1. Collect $\bar{u}_s$ into a matrix $\overline{U} \in \mathbb{R}^{m\times d}$. It holds that $\|\bar{u}_s\|_2 \leq 1/\sqrt{m}$, and $\|\overline{U}\|_F \leq 1$. Lemma 2.3 ensures that with high probability $\overline{U}$ has a positive margin at initialization.

Lemma 2.3. *Under Assumption 2.1, given any* $\delta \in (0,1)$ *and any* $\epsilon_1 \in (0,\gamma)$*, if* $m \geq \left(2\ln(n/\delta)\right)/\epsilon_1^2$*, then with probability* $1-\delta$*, it holds simultaneously for all* $1 \leq i \leq n$ *that*

$$y_i f_i^{(0)}\left(\overline{U}\right) = y_i \left\langle \nabla f_i(W_0), \overline{U}\right\rangle \geq \gamma - \sqrt{\frac{2\ln(n/\delta)}{m}} \geq \gamma - \epsilon_1.$$

For any W, any $\epsilon_2 > 0$, and any $1 \leq i \leq n$, define

$$\alpha_i(W, \epsilon_2) = \frac{1}{m}\sum_{s=1}^{m} \mathbb{1}\left[\left|\langle w_s, x_i\rangle\right| \leq \epsilon_2\right].$$

Lemma 2.4 controls $\alpha_i(W_0, \epsilon_2)$. It will help us show that $\overline{U}$ has a good margin during the training process.

Lemma 2.4. *Under the condition of Lemma 2.3, for any* $\epsilon_2 > 0$*, with probability* $1-\delta$*, it holds simultaneously for all* $1 \leq i \leq n$ *that*

$$\alpha_i\left(W_0, \epsilon_2\right) \leq \sqrt{\frac{2}{\pi}}\epsilon_2 + \sqrt{\frac{\ln(n/\delta)}{2m}} \leq \epsilon_2 + \frac{\epsilon_1}{2}.$$

Finally, Lemma 2.5 controls the output of the network at initialization.

Lemma 2.5. *Given any* $\delta \in (0,1)$*, if* $m \geq 25\ln(2n/\delta)$*, then with probability* $1-\delta$*, it holds simultaneously for all* $1 \leq i \leq n$ *that*

$$\left|f(x_i; W_0, a)\right| \leq \sqrt{2\ln\left(4n/\delta\right)}.$$

2.2 Convergence analysis of gradient descent

We analyze gradient descent in this subsection. First, define

$$\widehat{\mathcal{Q}}(W) := \frac{1}{n}\sum_{i=1}^{n} -\ell'\left(y_i f_i(W)\right).$$

We have the following observations.

- For any W and any $1 \le s \le m$, $\left\|\partial f_i/\partial w_s\right\|_2 \le 1/\sqrt{m}$, and thus $\left\|\nabla f_i(W)\right\|_F \le 1$. Therefore by the triangle inequality, $\left\|\nabla\widehat{\mathcal{R}}(W)\right\|_F \le \widehat{\mathcal{Q}}(W)$.
- If ℓ is 1-Lipschitz continuous, such as the logistic loss, then $\widehat{\mathcal{Q}}(W) \le 1$.
- If $-\ell' \le \ell$, such as the logistic loss, then $\widehat{\mathcal{Q}}(W) \le \widehat{\mathcal{R}}(W)$.

With the above observations, we give the following general result which does not require staying close to initialization. It plays an important role in the proof of our main result, and could be useful when analyzing neural networks beyond the NTK setting.

Lemma 2.6. *For any $t \ge 0$ and any $\overline{W}$, if $\eta_t \le 1$, then*

$$\eta_t\widehat{\mathcal{R}}(W_t) \le \left\|W_t - \overline{W}\right\|_F^2 - \left\|W_{t+1} - \overline{W}\right\|_F^2 + 2\eta_t\widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right).$$

Consequently, if we use a constant step size $\eta \le 1$ for $0 \le \tau < t$, then

$$\eta\left(\sum_{\tau<t}\widehat{\mathcal{R}}(W_\tau)\right) + \left\|W_t - \overline{W}\right\|_F^2 \le \left\|W_0 - \overline{W}\right\|_F^2 + 2\eta\left(\sum_{\tau<t}\widehat{\mathcal{R}}^{(\tau)}\left(\overline{W}\right)\right).$$

Proof. We have

$$\left\|W_{t+1} - \overline{W}\right\|_F^2 = \left\|W_t - \overline{W}\right\|_F^2 - 2\eta_t\left\langle\nabla\widehat{\mathcal{R}}(W_t), W_t - \overline{W}\right\rangle + \eta_t^2\left\|\nabla\widehat{\mathcal{R}}(W_t)\right\|_F^2. \tag{2.2}$$

The first order term of eq. (2.2) can be handled using the convexity of ℓ and homogeneity of ReLU:

$$\begin{aligned}\left\langle\nabla\widehat{\mathcal{R}}(W_t), W_t - \overline{W}\right\rangle &= \frac{1}{n}\sum_{i=1}^{n}\ell'\left(y_i f_i(W_t)\right) y_i\left\langle\nabla f_i(W_t), W_t - \overline{W}\right\rangle \\ &= \frac{1}{n}\sum_{i=1}^{n}\ell'\left(y_i f_i(W_t)\right)\left(y_i f_i(W_t) - y_i f_i^{(t)}\left(\overline{W}\right)\right) \\ &\ge \frac{1}{n}\sum_{i=1}^{n}\left(\ell\left(y_i f_i(W_t)\right) - \ell\left(y_i f_i^{(t)}\left(\overline{W}\right)\right)\right) = \widehat{\mathcal{R}}(W_t) - \widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right).\end{aligned} \tag{2.3}$$

The second-order term of eq. (2.2) can be bounded as follows

$$\eta_t^2\left\|\nabla\widehat{\mathcal{R}}(W_t)\right\|_F^2 \le \eta_t^2\widehat{\mathcal{Q}}(W_t)^2 \le \eta_t\widehat{\mathcal{Q}}(W_t) \le \eta_t\widehat{\mathcal{R}}(W_t), \tag{2.4}$$

because $\left\|\nabla\widehat{\mathcal{R}}(W_t)\right\|_F \le \widehat{\mathcal{Q}}(W_t)$, and $\eta_t, \widehat{\mathcal{Q}}(W_t) \le 1$, and $\widehat{\mathcal{Q}}(W_t) \le \widehat{\mathcal{R}}(W_t)$. Combining eqs. (2.2) to (2.4) gives

$$\eta_t\widehat{\mathcal{R}}(W_t) \le \left\|W_t - \overline{W}\right\|_F^2 - \left\|W_{t+1} - \overline{W}\right\|_F^2 + 2\eta_t\widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right).$$

Telescoping gives the other claim. □

Using Lemmas 2.3 to 2.6, we can prove Theorem 2.2. Below is a proof sketch; the full proof is given in Appendix A.

1. We first show that $\overline{U}$ defined in eq. (2.1) gives a positive margin at step t as long as the activation patterns do not change too much from the initialization.
2. We then show that such a phase lasts for a long time with a mild overparameterization by giving a strong control of $\|W_t - W_0\|_F$ via Lemma 2.6. Prior work only shows an $O(t)$ or $O(\sqrt{t})$ upper bound on $\|W_t - W_0\|_F$, which then requires the number of hidden units to be $\mathrm{poly}(1/\epsilon)$. By contrast, we are able to control $\|W_t - W_0\|_F$ by $O(\ln(t))$, which allows us to have a $\mathrm{polylog}(n, 1/\delta, 1/\epsilon)$ overparameterization.
3. Next we use Lemma 2.6 once again to get the empirical risk guarantee.
4. We also give an upper bound on $\left\|w_{s,t} - w_{s,0}\right\|_2$, or $\left\|W_t^\top - W_0^\top\right\|_{2,\infty}$. This will give us a Rademacher complexity bound in Section 3.

3 Generalization

To get a generalization bound, we naturally extend Assumption 2.1 to the following assumption, which is also made in (Nitanda & Suzuki, 2019) for smooth activations.

Assumption 3.1. There exists $\bar{v} \in \mathcal{H}$ and $\gamma > 0$, s.t. $\left\|\bar{v}(z)\right\|_2 \le 1$ for any $z \in \mathbb{R}^d$, and

$$y \int \langle \bar{v}(z), x \rangle \mathbb{1}\left[\langle z, x \rangle > 0\right] \mathrm{d}\mu_{\mathcal{N}}(z) \ge \gamma$$

for any (x, y) sampled from the data distribution $\mathcal{D}$. ◊

Here is our test error bound with Assumption 3.1.

Theorem 3.2. *Under Assumption 3.1, given any $\epsilon \in (0,1)$ and any $\delta \in (0, 1/4)$, let λ and M be given as in Theorem 2.2:*

$$\lambda := \frac{\sqrt{2\ln(4n/\delta)} + \ln(4/\epsilon)}{\gamma/3}, \quad \textit{and} \quad M := \max\left\{\frac{162\ln(n/\delta)}{\gamma^2}, 25\ln\left(\frac{2n}{\delta}\right), \frac{324\lambda^2}{\gamma^3}\right\}.$$

Then for any $m \ge M$ and any constant step size $\eta \le 1$, with probability $1 - 4\delta$ over the random initialization and data sampling,

$$P_{(x,y)\sim\mathcal{D}}\left(yf(x; W_k, a) \le 0\right) \le 2\epsilon + \frac{24\left(\sqrt{2\ln(4n/\delta)} + \ln(4/\epsilon)\right)}{\gamma^2\sqrt{n}} + 6\sqrt{\frac{\ln(2/\delta)}{2n}},$$

where k denotes the step with the minimum empirical risk before $\lceil 2\lambda^2/\eta\epsilon \rceil$.

Below is a direct corollary of Theorem 3.2.

Corollary 3.3. *Under Assumption 3.1, given any $\epsilon, \delta \in (0,1)$, using a constant step size and let*

$$n = \widetilde{\Omega}\left(\frac{1}{\gamma^4\epsilon^2}\right), \quad \textit{and} \quad m = \Omega\left(\frac{\ln(n/\delta) + \ln(1/\epsilon)^2}{\gamma^5}\right),$$

it holds with probability $1 - \delta$ that $P_{(x,y)\sim\mathcal{D}}\left(yf(x; W_k, a) \le 0\right) \le \epsilon$, where k denotes the step with the minimum empirical risk in the first $\widetilde{O}(1/\gamma^2\epsilon)$ steps.

To prove Theorem 3.2, we consider the sigmoid mapping $-\ell'(z) = e^{-z}/(1 + e^{-z})$, the empirical average $\widehat{\mathcal{Q}}(W_k)$, and the corresponding population average $\mathcal{Q}(W_k) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[-\ell'\left(yf(x; W_k, a)\right)\right]$. First of all, since $P_{(x,y)\sim\mathcal{D}}\left(yf(x; W_k, a) \le 0\right) \le 2\mathcal{Q}(W_k)$, it is enough to control $\mathcal{Q}(W_k)$. Next, as $\widehat{\mathcal{Q}}(W_k)$ is controlled by Theorem 2.2, it is enough to control the generalization error $\mathcal{Q}(W_k) - \widehat{\mathcal{Q}}(W_k)$. Moreover, since $-\ell'$ is supported on $[0, 1]$ and 1-Lipschitz, it is enough to bound the Rademacher complexity of the function space explored by gradient descent. Invoking the bound on $\left\|W_k^\top - W_0^\top\right\|_{2,\infty}$ finishes the proof. The proof details are given in Appendix B.

Remark 3.4. To get Theorem 3.2, we use the Lipschitz-based Rademacher complexity bound. One can also use the smoothness-based Rademacher complexity bound (Srebro et al., 2010, Theorem 1) and get a sample complexity $\widetilde{\Omega}(1/\gamma^4\epsilon)$. However, the bound will become complicated and some large constant will be introduced. It is an interesting open question to give a clean analysis based on smoothness. $\diamondsuit$

4 Stochastic gradient descent

There are some different formulations of SGD. In this section, we consider SGD with an online oracle. We randomly sample W_0 and a, and fix a during training. At step i, a data example (x_i, y_i) is sampled from the data distribution. We still let $f_i(W) := f(x_i; W, a)$, and perform the following update

$$W_{i+1} := W_i - \eta_i \ell'\left(y_i f_i(W_i)\right) y_i \nabla f_i(W_i).$$

Note that here i starts from 0.

Still with Assumption 3.1, we show the following result.

Theorem 4.1. *Under Assumption 3.1, given any $\epsilon, \delta \in (0,1)$, using a constant step size and $m = \Omega\left(\left(\ln(1/\delta)+\ln(1/\epsilon)^2\right)/\gamma^5\right)$, it holds with probability $1-\delta$ that*

$$\frac{1}{n}\sum_{i=1}^{n} P_{(x,y)\sim\mathcal{D}}\left(yf(x; W_i, a) \le 0\right) \le \epsilon, \quad \textit{for} \quad n = \widetilde{O}(1/\gamma^2\epsilon).$$

Below is a proof sketch; the details are given in Appendix C. For any W, define

$$\mathcal{R}_i(W) := \ell\left(y_i \left\langle \nabla f_i(W_i), W \right\rangle\right), \quad \text{and} \quad \mathcal{Q}_i(W) := -\ell'\left(y_i \left\langle \nabla f_i(W_i), W \right\rangle\right).$$

Due to homogeneity, $\mathcal{R}_i(W_i) = \ell\left(y_i f_i(W_i)\right)$ and $\mathcal{Q}_i(W_i) = -\ell'\left(y_i f_i(W_i)\right)$.

The first step is an extension of Lemma 2.6 to the SGD setting. The proofs are similar.

Lemma 4.2. *With a constant step size $\eta \le 1$, for any $\overline{W}$ and any $i \ge 0$,*

$$\eta\left(\sum_{t<i}\mathcal{R}_t(W_t)\right) + \left\|W_i - \overline{W}\right\|_F^2 \le \left\|W_0 - \overline{W}\right\|_F^2 + 2\eta\left(\sum_{t<i}\mathcal{R}_t\left(\overline{W}\right)\right).$$

With Lemma 4.2, we can also extend Theorem 2.2 to the SGD setting and get a bound on $\sum_{i<n}\mathcal{Q}_i(W_i)$, using a similar proof. To further get a bound on the cumulative population risk $\sum_{i<n}\mathcal{Q}(W_i)$, the key observation is that $\sum_{i<n}\left(\mathcal{Q}(W_i) - \mathcal{Q}_i(W_i)\right)$ is a martingale. Using a martingale Bernstein bound, we prove the following lemma; applying it finishes the proof of Theorem 4.1.

Lemma 4.3. *Given any $\delta \in (0,1)$, with probability $1-\delta$,*

$$\sum_{t<i}\mathcal{Q}(W_t) \le 4\sum_{t<i}\mathcal{Q}_t(W_t) + 4\ln\left(\frac{1}{\delta}\right).$$

5 On separability

Given a training set $\left\{(x_i, y_i)\right\}_{i=1}^n$, the linear kernel is defined as $K_0(x_i, x_j) := \left\langle x_i, x_j \right\rangle$. The maximum margin by a linear classifier is given by

$$\gamma_0 := \min_{q\in\Delta_n}\sqrt{(q\odot y)^\top K_0\,(q\odot y)}. \tag{5.1}$$

If the data is not linearly separable, $\gamma_0 = 0$.

In this paper we train the first layer of a two-layer network, and the kernel we consider is the NTK of the first layer:

$$\begin{aligned} K_1\left(x_i, x_j\right) &:= \mathbb{E}\left[\frac{\partial f(x_i; W_0, a)}{\partial W_0}, \frac{\partial f(x_j; W_0, a)}{\partial W_0}\right] \\ &= \left\langle x_i, x_j\right\rangle \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}\left[\mathbb{1}\left[\langle x_i, w\rangle > 0\right]\mathbb{1}\left[\left\langle x_j, w\right\rangle > 0\right]\right]. \end{aligned}$$

Similar to the definition of γ_0, the margin given by K_1 is defined as

$$\gamma_1 := \min_{q\in\Delta_n}\sqrt{(q\odot y)^\top K_1\,(q\odot y)}.$$

Regarding the relation between γ_1 and Assumption 2.1, we have the following result.

Proposition 5.1. *If $\gamma_1 > 0$, then there exists $\hat{v} \in \mathcal{H}$ s.t. $\|\hat{v}\|_{\mathcal{H}} = \gamma_1$, and $\left\|\hat{v}(z)\right\|_2 \le 1$ for any $z \in \mathbb{R}^d$, and $y_i\left\langle\hat{v}, \phi_i\right\rangle \ge \gamma_1^2$ for any $1 \le i \le n$.*

The proof is given in Appendix D, which uses the Fenchel duality theory. $\hat{v}$ given by Proposition 5.1 satisfies Assumption 2.1 with $\gamma = \gamma_1^2$, but there might exist some $\bar{v}$ with a much better γ, since the upper bound $\left\|\hat{v}(z)\right\|_2 \le 1$ in Proposition 5.1 might be very loose.

We can further ask how large γ_1 could be. (Oymak & Soltanolkotabi, 2019, Corollary I.2) shows that if for any two feature vectors x_i and x_j, we have $\|x_i - x_j\|_2 \ge \delta$ and $\|x_i + x_j\|_2 \ge \delta$ for some $\delta > 0$, then

$$\lambda_0 := \lambda_{\min}(K_1) \ge \frac{\delta}{100n^2}.$$

For arbitrary labels $y \in \{-1, +1\}^n$, since $\|q\odot y\|_2 \ge 1/\sqrt{n}$, we have the worst case bound $\gamma_1^2 \ge \delta/100n^3$. However, real world labels could give a much better γ_1. For example, a tighter lower bound on γ_1 is $\delta/100n_S^3$, where n_S denotes the number of support vectors, which might be much smaller than n.

On the other hand, given any training set $\left\{(x_i, y_i)\right\}_{i=1}^n$ which may have a large margin, if we replace y with random labels $\epsilon \sim \text{unif}\left(\{-1, +1\}^n\right)$, with high probability the margin becomes $O(1/\sqrt{n})$. To see this, let $\hat{q}$ denote the uniform probability vector $(1/n, \dots, 1/n)$. Note that

$$\begin{aligned} \mathbb{E}_{\epsilon\sim\text{unif}\left(\{-1,+1\}^n\right)}\left[(\hat{q}\odot\epsilon)^\top K_1\,(\hat{q}\odot\epsilon)\right] &= \mathbb{E}_{\epsilon\sim\text{unif}\left(\{-1,+1\}^n\right)}\left[\sum_{i,j=1}^n \frac{1}{n^2}\epsilon_i\epsilon_j K_1(x_i, x_j)\right] \\ &= \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}_{\epsilon\sim\text{unif}\left(\{-1,+1\}^n\right)}\left[\epsilon_i\epsilon_j K_1(x_i, x_j)\right] \\ &= \frac{1}{n^2}\sum_{i=1}^n K_1(x_i, x_i) = \frac{1}{2n}. \end{aligned}$$

Since $0 \le (\hat{q}\odot\epsilon)^\top K_1\,(\hat{q}\odot\epsilon) \le 1$ for any ϵ, by Hoeffding's inequality it holds with high probability that $(\hat{q}\odot\epsilon)^\top K_1\,(\hat{q}\odot\epsilon) = O(1/n)$, and thus the margin is $O(1/\sqrt{n})$.

6 Open problems

In this paper, we analyze gradient descent on a two-layer network in the NTK regime, where the weights stay close to the initialization. It is an interesting open question if gradient descent learns something beyond the NTK, after the iterates move far enough from the initial weights. It is also interesting to extend our analysis to other architectures, such as multi-layer networks, convolutional networks, and residual networks. Finally, in this paper we only discuss binary classification; it is interesting to see if it is possible to get similar results for other tasks, such as regression.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via overparameterization. *arXiv preprint arXiv:1811.03962*, 2018b.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, Nov 2002.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26, 2011.

Jonathan M. Borwein and Qiji J. Zhu. *Techniques of Variational Analysis, volume 20 of.* CMS Books in Mathematics, 2005.

Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning overparameterized deep relu networks. *arXiv preprint arXiv:1902.01384*, 2019a.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *arXiv preprint arXiv:1905.13210*, 2019b.

Lenaic Chizat and Francis Bach. A Note on Lazy Training in Supervised Differentiable Programming. `arXiv:1812.07956v2 [math.OC]`, 2019.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018a.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018b.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pp. 8157–8166, 2018.

Percy Liang. Stanford CS229T/STAT231: Statistical Learning Theory, Apr 2016. URL `https://web.stanford.edu/class/cs229t/notes.pdf`.

Atsushi Nitanda and Taiji Suzuki. Refined generalization analysis of gradient descent for overparameterized two-layer neural networks with smooth activations on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv preprint arXiv:1902.04674*, 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, pp. 2199–2207, 2010.

Martin J. Wainwright. UC Berkeley Statistics 210B, Lecture Notes: Basic tail and concentration bounds, Jan 2015. URL `https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf`.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. *arXiv preprint arXiv:1906.04688*, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

A OMITTED PROOFS FROM SECTION 2

Proof of Lemma 2.3. By Assumption 2.1, given any $1 \le i \le n$,

$$\mu := \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}\left[y_i\left\langle\bar{v}(w),x_i\right\rangle\mathbb{1}\left[\langle w,x_i\rangle>0\right]\right]\ge\gamma.$$

On the other hand,

$$y_i f_i^{(0)}\left(\overline{U}\right)=\frac{1}{m}\sum_{s=1}^{m}y_i\left\langle\bar{v}(w_{s,0}),x_i\right\rangle\mathbb{1}\left[\left\langle w_{s,0},x_i\right\rangle>0\right]$$

is the empirical mean of i.i.d. r.v.'s supported on $[-1,+1]$ with mean μ. Therefore by Hoeffding's inequality, with probability $1-\delta/n$,

$$y_i f_i^{(0)}\left(\overline{U}\right)-\gamma\ge y_i f_i^{(0)}\left(\overline{U}\right)-\mu\ge-\sqrt{\frac{2\ln(n/\delta)}{m}}.$$

Applying the union bound finishes the proof. □

Proof of Lemma 2.4. Given any fixed ϵ_2 and $1\le i\le n$,

$$\mathbb{E}\left[\alpha_i(W_0,\epsilon_2)\right]=P\left(\left|\langle w,x_i\rangle\right|\le\epsilon_2\right)\le\frac{2\epsilon_2}{\sqrt{2\pi}}=\sqrt{\frac{2}{\pi}}\epsilon_2,$$

because $\langle w,x_i\rangle$ is a standard Gaussian r.v. and the density of standard Gaussian has maximum $1/\sqrt{2\pi}$. Since $\alpha_i(W_0,\epsilon_2)$ is the empirical mean of Bernoulli r.v.'s, by Hoeffding's inequality, with probability $1-\delta/n$,

$$\alpha_i(W_0,\epsilon_2)\le\mathbb{E}\left[\alpha_i(W_0,\epsilon_2)\right]+\sqrt{\frac{\ln(n/\delta)}{2m}}\le\sqrt{\frac{2}{\pi}}\epsilon_2+\sqrt{\frac{\ln(n/\delta)}{2m}}.$$

Applying the union bound finishes the proof. □

To prove Lemma 2.5, we need the following technical result.

Lemma A.1. *Consider the random vector $X=(X_1,\ldots,X_m)$, where $X_i=\sigma(Z_i)$ for some $\sigma:\mathbb{R}\to\mathbb{R}$ that is 1-Lipschitz, and Z_i are i.i.d. standard Gaussian r.v.'s. Then the r.v. $\|X\|_2$ is 1-sub-Gaussian, and thus with probability $1-\delta$,*

$$\|X\|_2-\mathbb{E}\left[\|X\|_2\right]\le\sqrt{2\ln(1/\delta)}.$$

Proof. Given $a\in\mathbb{R}^m$, define

$$f(a)=\sqrt{\sum_{i=1}^{m}\sigma(a_i)^2}=\left\|\sigma(a)\right\|_2,$$

where $\sigma(a)$ is obtained by applying σ coordinate-wisely to a. For any $a,b\in\mathbb{R}^m$, by the triangle inequality, we have

$$\left|f(a)-f(b)\right|=\left|\left\|\sigma(a)\right\|_2-\left\|\sigma(b)\right\|_2\right|\le\left\|\sigma(a)-\sigma(b)\right\|_2=\sqrt{\sum_{i=1}^{m}\left(\sigma(a_i)-\sigma(b_i)\right)^2},$$

and by further using the 1-Lipschitz continuity of σ, we have

$$\left|f(a)-f(b)\right| \le \sqrt{\sum_{i=1}^{m}\left(\sigma(a_i)-\sigma(b_i)\right)^2} \le \sqrt{\sum_{i=1}^{m}(a_i-b_i)^2} = \|a-b\|_2 \,.$$

As a result, f is a 1-Lipschitz continuous function w.r.t. the ℓ_2 norm, and by Theorem 2.4 of (Wainwright, 2015), $f(X)$ is 1-sub-Gaussian. The remaining bound comes directly from Hoeffding's inequality. □

Proof of Lemma 2.5. Given $1 \le i \le n$, let $h_i = \sigma(W_0 x_i)/\sqrt{m}$. By Lemma A.1, $\|h_i\|_2$ is sub-Gaussian with variance proxy $1/m$. By Hoeffding's inequality, with probability $1 - \delta/2n$ over W_0,

$$\|h_i\|_2 - \mathbb{E}\left[\|h_i\|_2\right] \le \sqrt{\frac{2\ln(2n/\delta)}{m}} \le \sqrt{\frac{2\ln(2n/\delta)}{25\ln(2n/\delta)}} \le 1 - \frac{\sqrt{2}}{2}.$$

On the other hand, by Jensen's inequality,

$$\mathbb{E}\left[\|h_i\|_2\right] \le \sqrt{\mathbb{E}\left[\|h_i\|_2^2\right]} = \frac{\sqrt{2}}{2}.$$

As a result, with probability $1 - \delta/2n$, it holds that $\|h_i\|_2 \le 1$. By the union bound, with probability $1 - \delta/2$ over W_0, for all $1 \le i \le n$, we have $\|h_i\|_2 \le 1$.

For any W_0 such that the above event holds, and for any $1 \le i \le n$, the r.v. $\langle h_i, a\rangle$ is sub-Gaussian with variance proxy $\|h_i\|_2^2 \le 1$. By Hoeffding's inequality, with probability $1 - \delta/2n$ over a,

$$\left|\langle h_i, a\rangle\right| = \left|f(x_i; W_0, a)\right| \le \sqrt{2\ln\left(4n/\delta\right)}.$$

By the union bound, with probability $1 - \delta/2$ over a, for all $1 \le i \le n$, we have $\left|f(x_i; W_0, a)\right| \le \sqrt{2\ln\left(4n/\delta\right)}$.

The probability that the above events all happen is at least $(1 - \delta/2)(1 - \delta/2) \ge 1 - \delta$, over W_0 and a. □

Proof of Theorem 2.2. The condition on m ensures that Lemmas 2.3 to 2.5 hold with $\epsilon_1 = \gamma/9$ and $\epsilon_2 = \gamma/3$.

For any $1 \le i \le n$ and any step t, let $\xi_{i,t}$ denote the proportion of activation patterns for x_i that are different from step 0 to step t. Formally,

$$\xi_{i,t} := \frac{1}{m}\sum_{s=1}^{m}\mathbb{1}\left[\mathbb{1}\left[\langle w_{s,t}, x_i\rangle > 0\right] \ne \mathbb{1}\left[\langle w_{s,0}, x_i\rangle > 0\right]\right].$$

Let t_1 denote the first step such that there exists i with $\xi_{i,t_1} > 5\gamma/9$. By definition, for any $0 \le t < t_1$ and any $1 \le i \le n$, $\xi_{i,t} \le 5\gamma/9$.

The first claim we make is that $\overline{U}$ has a good margin before step t_1. For any $0 \le t < t_1$ and any $1 \le i \le n$,

$$\begin{aligned} y_i\left\langle \nabla f_i(W_t), \overline{U}\right\rangle &= y_i\left\langle \nabla f_i(W_0), \overline{U}\right\rangle + y_i\left\langle \nabla f_i(W_t) - \nabla f_i(W_0), \overline{U}\right\rangle \\ &\ge \gamma - \epsilon_1 + y_i\left\langle \nabla f_i(W_t) - \nabla f_i(W_0), \overline{U}\right\rangle. \end{aligned}$$

In addition,

$$\begin{aligned} y_i\left\langle \nabla f_i(W_t) - \nabla f_i(W_0), \overline{U}\right\rangle &= y_i\frac{1}{m}\sum_{i=1}^{m}\left(\mathbb{1}\left[\langle w_{s,t}, x_i\rangle > 0\right] - \mathbb{1}\left[\langle w_{s,0}, x_i\rangle\right]\right)\langle \bar{v}(w_{s,0}), x_i\rangle \\ &\ge -\frac{\xi_{i,t} m}{m}. \end{aligned}$$

Since $\epsilon_1 = \gamma/9$ and $\xi_{i,t} \leq 5\gamma/9$, we have

$$y_i \left\langle \nabla f_i(W_t), \overline{U} \right\rangle \geq \frac{\gamma}{3}. \tag{A.1}$$

Now we let

$$\overline{W} := W_0 + \lambda\overline{U}, \quad \text{where} \quad \lambda := \frac{\sqrt{2\ln(4n/\delta)} + \ln(4/\epsilon)}{\gamma/3}.$$

Due to Lemma 2.5, for any $0 \leq t < t_1$, $\widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right) \leq \epsilon/4$. Moreover since $\left\|\overline{U}\right\|_F \leq 1$, we have $\left\|\overline{W} - W_0\right\|_F \leq \lambda$.

Let $T := \lceil 2\lambda^2/\eta\epsilon \rceil$. The next claim we make is that $t_1 \geq T$. To see this, note that Lemma 2.6 ensures

$$\left\|W_{t_1} - \overline{W}\right\|_F^2 \leq \left\|W_0 - \overline{W}\right\|_F^2 + 2\eta\left(\sum_{t<t_1} \widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right)\right) \leq \lambda^2 + \frac{\epsilon}{2}\eta t_1.$$

Suppose $t_1 < T$, then we have $t_1 \leq 2\lambda^2/\eta\epsilon$ and $\left\|W_{t_1} - \overline{W}\right\|_F^2 \leq 2\lambda^2$, and thus

$$\|W_{t_1} - W_0\|_F \leq \left\|W_{t_1} - \overline{W}\right\|_F + \left\|\overline{W} - W_0\right\|_F \leq \sqrt{2}\lambda + \lambda \leq \sqrt{6}\lambda.$$

W.l.o.g., suppose at step t_1 hidden unit 1 has $\xi_{1,t_1} > 5\gamma/9$. By Lemma 2.4,

$$\alpha_1\left(W_0, \epsilon_2\right) \leq \epsilon_2 + \frac{\epsilon_1}{2} = \frac{7\gamma}{18}.$$

In addition to these $\alpha_1(W_0, \epsilon_2)m$ units, the other hidden units with different activation patterns from step t_1 to step 0 has proportion at most

$$\frac{\|W_{t_1} - W_0\|_F^2}{\epsilon_2^2}\frac{1}{m} \leq \frac{6\lambda^2}{\gamma^2/9}\frac{1}{m} \leq \frac{\gamma}{6},$$

since by the assumption $m \geq 324\lambda^2/\gamma^3$. As a result, we should have

$$\xi_{1,t_1} \leq \alpha_i(W_0, \epsilon_2) + \frac{\gamma}{6} \leq \frac{7\gamma}{18} + \frac{\gamma}{6} = \frac{5\gamma}{9},$$

which is a contradiction. Therefore $t_1 \geq T$.

Now we are ready to prove the claims of Theorem 2.2. First, for any $t < T \leq t_1$, the above argument works and implies that $\|W_t - W_0\|_F \leq \sqrt{6}\lambda$ and $\left\|W_t - \overline{W}\right\|_F \leq \sqrt{2}\lambda$.

Next we give the risk guarantee. Lemma 2.6 gives

$$\frac{1}{T}\sum_{t<T} \widehat{\mathcal{R}}(W_t) \leq \frac{\left\|W_0 - \overline{W}\right\|_F^2}{\eta T} + \frac{2}{T}\sum_{t<T} \widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Finally, we consider the bound on $\left\|w_{s,t} - w_{s,0}\right\|_2$ for any $1 \leq s \leq m$ and any $0 \leq t < T$. Note that

$$\begin{aligned}\sqrt{2}\lambda \geq \left\|W_t - \overline{W}\right\|_F \geq \left\langle W_t - \overline{W}, \overline{U}\right\rangle &= \left\langle W_t - W_0, \overline{U}\right\rangle - \left\langle \overline{W} - W_0, \overline{U}\right\rangle \\ &\geq \left\langle W_t - W_0, \overline{U}\right\rangle - \lambda.\end{aligned}$$

Moreover, due to eq. (A.1),

$$\begin{aligned}\left\langle W_t - W_0, \overline{U}\right\rangle = -\eta\sum_{\tau<t}\left\langle \nabla\widehat{\mathcal{R}}(W_\tau), \overline{U}\right\rangle &= \eta\sum_{\tau<t}\frac{1}{n}\sum_{i=1}^{n} -\ell'\left(y_i f_i(W_\tau)\right) y_i \left\langle \nabla f_i(W_\tau), \overline{U}\right\rangle \\ &\geq \eta\sum_{\tau<t}\widehat{\mathcal{Q}}(W_\tau)\frac{\gamma}{3}.\end{aligned}$$

As a result,

$$\eta \sum_{\tau<t} \widehat{\mathcal{Q}}(W_\tau) \leq \frac{3(\sqrt{2}+1)\lambda}{\gamma} \leq \frac{3\sqrt{6}\lambda}{\gamma}.$$

Furthermore, by the triangle inequality,

$$\begin{aligned} \left\|w_{s,t} - w_{s,0}\right\|_2 &\leq \eta \sum_{\tau<t} \left\| \frac{1}{n} \sum_{i=1}^{n} \ell' \left(y_i f_i(W_\tau)\right) y_i \frac{\partial f_i}{\partial w_{s,\tau}} \right\|_2 \\ &\leq \eta \sum_{\tau<t} \frac{1}{n} \sum_{i=1}^{n} \ell' \left(y_i f_i(W_\tau)\right) \left\| \frac{\partial f_i}{\partial w_{s,\tau}} \right\|_2 \\ &\leq \eta \sum_{\tau<t} \widehat{\mathcal{Q}}(W_\tau) \frac{1}{\sqrt{m}} \leq \frac{3\sqrt{6}\lambda}{\gamma\sqrt{m}}. \end{aligned}$$

□

B Omitted proofs from Section 3

The proof of Theorem 3.2 is based on Rademacher complexity. Given a sample $S = (z_1, \ldots, z_n)$ (where $z_i = (x_i, y_i)$) and a function class $\mathcal{H}$, the Rademacher complexity of $\mathcal{H}$ on S is defined as

$$\text{Rad}\left(\mathcal{H} \circ S\right) := \frac{1}{n} \mathbb{E}_{\epsilon \sim \{-1,+1\}^n} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \epsilon_i h(z_i) \right].$$

We will use the following general result.

Lemma B.1. *(Shalev-Shwartz & Ben-David, 2014, Theorem 26.5) If $h(z) \in [a, b]$, then with probability $1-\delta$,*

$$\sup_{h \in \mathcal{H}} \left(\mathbb{E}_{z \sim \mathcal{D}} \left[h(z)\right] - \frac{1}{n} \sum_{i=1}^{n} h(z_i) \right) \leq 2\text{Rad}\left(\mathcal{H} \circ S\right) + 3(b-a)\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

We also need the following contraction lemma. Consider a feature sample $X = (x_1, \ldots, x_n)$ and a function class $\mathcal{F}$ on X. For each $1 \leq i \leq n$, let $g_i : \mathbb{R} \to \mathbb{R}$ denote a K-Lipschitz function. Let $g \circ \mathcal{F}$ denote the class of functions which map x_i to $g_i(f(x_i))$ for some $f \in \mathcal{F}$.

Lemma B.2. *(Shalev-Shwartz & Ben-David, 2014, Lemma 26.9)* $\text{Rad}\left(g \circ \mathcal{F} \circ X\right) \leq K\text{Rad}\left(\mathcal{F} \circ X\right)$.

To prove Theorem 3.2, we need one more Rademacher complexity bound. Given a fixed initialization (W_0, a), consider the following classes:

$$\mathcal{W}_\rho := \left\{ W \in \mathbb{R}^{m \times d} \middle| \left\|w_s - w_{s,0}\right\|_2 \leq \rho \text{ for any } 1 \leq s \leq m \right\},$$

and

$$\mathcal{F}_\rho := \left\{ x \mapsto f(x; W, a) \middle| W \in \mathcal{W}_\rho \right\}.$$

Given a feature sample X, the following Lemma B.3 controls the Rademacher complexity of $\mathcal{F}_\rho \circ X$. A similar version was given in (Liang, 2016, Theorem 43), and the proof is similar to the proof of (Bartlett & Mendelson, 2002, Theorem 18) which also pushes the supreme through and handles each hidden units separately.

Lemma B.3. $\text{Rad}\left(\mathcal{F}_\rho \circ X\right) \leq \rho\sqrt{m/n}$.

Proof of Lemma B.3. We have

$$\begin{aligned}
\mathbb{E}_\epsilon\left[\sup_{W\in\mathcal{W}_\rho}\sum_{i=1}^n \epsilon_i f(x_i;W,a)\right] &= \mathbb{E}_\epsilon\left[\sup_{W\in\mathcal{W}_\rho}\sum_{i=1}^n \epsilon_i \sum_{s=1}^m \frac{1}{\sqrt{m}} a_s\sigma\left(\langle w_s, x_i\rangle\right)\right] \\
&= \mathbb{E}_\epsilon\left[\frac{1}{\sqrt{m}}\sup_{W\in\mathcal{W}_\rho}\sum_{s=1}^m\sum_{i=1}^n \epsilon_i a_s\sigma\left(\langle w_s, x_i\rangle\right)\right] \\
&= \mathbb{E}_\epsilon\left[\frac{1}{\sqrt{m}}\sum_{s=1}^m\left(\sup_{\left\|w_s-w_{s,0}\right\|_2\le\rho}\sum_{i=1}^n \epsilon_i a_s\sigma\left(\langle w_s, x_i\rangle\right)\right)\right] \\
&= \frac{1}{\sqrt{m}}\sum_{i=1}^m \mathbb{E}_\epsilon\left[\sup_{\left\|w_s-w_{s,0}\right\|_2\le\rho}\sum_{i=1}^n \epsilon_i a_s\sigma\left(\langle w_s, x_i\rangle\right)\right].
\end{aligned}$$

Note that for any $1 \le s \le m$, the mapping $z \mapsto a_s\sigma(z)$ is 1-Lipschitz, and thus Lemma B.2 gives

$$\begin{aligned}
\mathbb{E}_\epsilon\left[\sup_{W\in\mathcal{W}_\rho}\sum_{i=1}^n \epsilon_i f(x_i;W,a)\right] &\le \frac{1}{\sqrt{m}}\sum_{i=1}^m \mathbb{E}_\epsilon\left[\sup_{\left\|w_s-w_{s,0}\right\|_2\le\rho}\sum_{i=1}^n \epsilon_i a_s\sigma\left(\langle w_s, x_i\rangle\right)\right] \\
&\le \frac{1}{\sqrt{m}}\sum_{i=1}^m \mathbb{E}_\epsilon\left[\sup_{\left\|w_s-w_{s,0}\right\|_2\le\rho}\sum_{i=1}^n \epsilon_i \langle w_s, x_i\rangle\right].
\end{aligned}$$

Invoking the Rademacher complexity bound of linear classifiers (Shalev-Shwartz & Ben-David, 2014, Lemma 26.10) then gives

$$\mathrm{Rad}\left(\mathcal{F}_\rho\circ X\right) = \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{W\in\mathcal{W}_\rho}\sum_{i=1}^n \epsilon_i f(x_i;W,a)\right] \le \frac{\rho\sqrt{m}}{\sqrt{n}}.$$

□

Now we are ready to prove the main generalization result Theorem 3.2.

Proof. First fix an initialization (W_0, a), and consider $\mathcal{F}_\rho$. Let $\mathcal{H} := \left\{(x,y)\mapsto -\ell'\left(yf(x)\right)\middle| f\in\mathcal{F}_\rho\right\}$. Since for any $h\in\mathcal{H}$ and any z, $h(z)\in[0,1]$, Lemma B.1 ensures that with probability $1-\delta$ over the data sampling,

$$\sup_{h\in\mathcal{H}}\left(\mathbb{E}_{z\sim\mathcal{D}}\left[h(z)\right]-\frac{1}{n}\sum_{i=1}^n h(z_i)\right) = \sup_{W\in\mathcal{W}_\rho}\left(\mathcal{Q}(W)-\widehat{\mathcal{Q}}(W)\right) \le 2\mathrm{Rad}\left(\mathcal{H}\circ S\right)+3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Since for each $1\le i\le n$, the mapping $z\mapsto -\ell'(y_i z)$ is $(1/4)$-Lipschitz, Lemma B.2 further ensures that $\mathrm{Rad}\left(\mathcal{H}\circ S\right)\le \mathrm{Rad}\left(\mathcal{F}_\rho\circ X\right)/4$, and thus

$$\sup_{W\in\mathcal{W}_\rho}\left(\mathcal{Q}(W)-\widehat{\mathcal{Q}}(W)\right) \le \frac{\rho\sqrt{m}}{2\sqrt{n}}+3\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{B.1}$$

On the other hand, Theorem 2.2 ensures that under the condition of Theorem 3.2, for any fixed dataset, with probability $1-3\delta$ over the random initialization, we have

$$\widehat{\mathcal{Q}}(W_k)\le\widehat{\mathcal{R}}(W_k)\le\epsilon, \quad \text{and} \quad \left\|w_{s,k}-w_{s,0}\right\|_2\le\frac{3\sqrt{6}\lambda}{\gamma\sqrt{m}}.$$

As a result, invoking eq. (B.1) with $\rho = 3\sqrt{6}\lambda/(\gamma\sqrt{m})$, with probability $1-4\delta$ over the random initialization and data sampling,

$$\mathcal{Q}(W_k)\le\widehat{\mathcal{Q}}(W_k)+\frac{3\sqrt{6}\lambda}{2\gamma\sqrt{n}}+3\sqrt{\frac{\ln(2/\delta)}{2n}}\le\epsilon+\frac{12\left(\sqrt{2\ln(4n/\delta)}+\ln(4/\epsilon)\right)}{\gamma^2\sqrt{n}}+3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Invoking $P_{(x,y)\sim\mathcal{D}}\left(yf(x;W,a)\le 0\right)\le 2\mathcal{Q}(W)$ finishes the proof. □

C OMITTED PROOFS FROM SECTION 4

Proof of Lemma 4.2. Recall that $\left\|\nabla f_t(W_t)\right\|_F \le 1$, we have

$$\left\|W_{t+1}-\overline{W}\right\|_F^2 \le \left\|W_t-\overline{W}\right\|_F^2 - 2\eta\ell'\left(y_t f_t(W_t)\right) y_t \left\langle \nabla f_t(W_t), W_t-\overline{W}\right\rangle + \eta^2\left(\ell'\left(y_t f_t(W_t)\right)\right)^2. \tag{C.1}$$

Similar to the proof of Lemma 2.6, the first order term of eq. (C.1) can be handled using the convexity of ℓ and homogeneity of ReLU as follows

$$\ell'\left(y_t f_t(W_t)\right) y_t \left\langle \nabla f_t(W_t), W_t-\overline{W}\right\rangle \ge \mathcal{R}_t(W_t) - \mathcal{R}_t\left(\overline{W}\right), \tag{C.2}$$

and the second-order term of eq. (C.1) can be bounded as follows

$$\eta^2\left(\ell'\left(y_t f_t(W_t)\right)\right)^2 \le -\eta\ell'\left(y_t f_t(W_t)\right) \le \eta\ell\left(y_t f_t(W_t)\right) = \eta\mathcal{R}_t(W_t), \tag{C.3}$$

since $\eta, -\ell' \le 1$ and $-\ell' \le \ell$. Combining eqs. (C.1) to (C.3) gives

$$\eta\mathcal{R}_t(W_t) \le \left\|W_t-\overline{W}\right\|_F^2 - \left\|W_{t+1}-\overline{W}\right\|_F^2 + 2\eta\mathcal{R}_t\left(\overline{W}\right).$$

Telescoping gives the claim. □

With Lemma 4.2, we give the following result, which is an extension of Theorem 2.2 to the SGD setting.

Lemma C.1. *Under Assumption 3.1, given any $\epsilon \in (0,1)$, any $\delta \in (0,1/3)$, and any positive integer n_0, let*

$$\lambda := \frac{\sqrt{2\ln(4n_0/\delta)} + \ln(4/\epsilon)}{\gamma/3}, \quad \text{and} \quad M := \max\left\{\frac{162\ln(n_0/\delta)}{\gamma^2}, 25\ln\left(\frac{2n_0}{\delta}\right), \frac{324\lambda^2}{\gamma^3}\right\}.$$

For any $m \ge M$ and any constant step size $\eta \le 1$, if $n_0 \ge n := \lceil 2\lambda^2/\eta\epsilon \rceil$, then with probability $1-3\delta$,

$$\frac{1}{n}\sum_{i<n} \mathcal{Q}_i(W_i) \le \epsilon.$$

Proof. We first sample n_0 data examples $(x_0,y_0),\ldots,(x_{n_0-1},y_{n_0-1})$, and then feed (x_i,y_i) to SGD at step i. We only consider the first n_0 steps.

The proof is similar to the proof of Theorem 2.2. For $0 \le i,t < n_0$, let $\xi_{i,t}$ denote the proportion of activation patterns for x_i that are different from step 0 to step t. Let n_1 denote the first step before n_0 such that there exists some i with $\xi_{i,n_1} > 5\gamma/9$. If such a step does not exist, let $n_1 = n_0$.

Similar to the proof of Theorem 2.2, we can show that for any $0 \le i < n_1$,

$$y_i\left\langle \nabla f_i(W_i), \overline{U}\right\rangle \ge \frac{\gamma}{3}.$$

Consequently, still let $\overline{W} := W_0 + \lambda\overline{U}$, we have for any $0 \le i < n_1$, $\mathcal{R}_i\left(\overline{W}\right) \le \epsilon/4$.

Now consider $n := \lceil 2\lambda^2/\eta\epsilon \rceil$. Using Lemma 4.2, in the same way as the proof of Theorem 2.2, we can show that $n \le n_1$. Then invoking Lemma 4.2 again, we get

$$\frac{1}{n}\sum_{i<n}\mathcal{Q}_i(W_i) \le \frac{1}{n}\sum_{i<n}\mathcal{R}_i(W_i) \le \frac{\left\|W_0-\overline{W}\right\|_F^2}{\eta n} + \frac{2}{n}\sum_{i<n}\mathcal{R}_i\left(\overline{W}\right) \le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

□

Next we prove Lemma 4.3. We need the following martingale Bernstein bound.

Lemma C.2. *(Beygelzimer et al., 2011, Theorem 1) Let $(M_t, \mathcal{F}_t)_{t\geq 0}$ denote a martingale with $M_0 = 0$ and $\mathcal{F}_0$ be the trivial σ-algebra. Let $(\Delta_t)_{t\geq 1}$ denote the corresponding martingale difference sequence, and let*

$$V_t := \sum_{j=1}^{t} \mathbb{E}\left[\Delta_j^2 \middle| \mathcal{F}_{j-1}\right]$$

denote the sequence of conditional variance. If $\Delta_t \leq R$ a.s., then for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$M_t \leq \frac{V_t}{R}(e-2) + R\ln\left(\frac{1}{\delta}\right).$$

Proof of Lemma 4.3. For any $i \geq 0$, let z_i denote (x_i, y_i), and $z_{0,i}$ denote $(z_0, \ldots, z_i)$. Note that $\sum_{t<i}\big(\mathcal{Q}(W_t) - \mathcal{Q}_t(W_t)\big)$ is a martingale w.r.t. the filtration $\sigma(z_{0,i-1})$. The martingale difference sequence is given by $\mathcal{Q}(W_t) - \mathcal{Q}_t(W_t)$, which satisfies

$$\mathcal{Q}(W_t) - \mathcal{Q}_t(W_t) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[-\ell'\big(yf(x;W_t,a)\big)\right] + \ell'\big(y_t f(x_t;W_t,a)\big) \leq 1, \tag{C.4}$$

since $-1 \leq \ell' \leq 0$. Moreover, we have

$$\begin{aligned}
&\mathbb{E}\left[\big(\mathcal{Q}(W_t) - \mathcal{Q}_t(W_t)\big)^2 \middle| \sigma(z_{0,t-1})\right] \\
=&\mathcal{Q}(W_t)^2 - 2\mathcal{Q}(W_t)\mathbb{E}\left[\mathcal{Q}_t(W_t)\middle|\sigma(z_{0,t-1})\right] + \mathbb{E}\left[\mathcal{Q}_t(W_t)^2\middle|\sigma(z_{0,t-1})\right] \\
=&-\mathcal{Q}(W_t)^2 + \mathbb{E}\left[\mathcal{Q}_t(W_t)^2\middle|\sigma(z_{0,t-1})\right] \\
\leq&\mathbb{E}\left[\mathcal{Q}_t(W_t)^2\middle|\sigma(z_{0,t-1})\right] \\
\leq&\mathbb{E}\left[\mathcal{Q}_t(W_t)\middle|\sigma(z_{0,t-1})\right] \\
=&\mathcal{Q}(W_t).
\end{aligned} \tag{C.5}$$

Invoking Lemma C.2 with eqs. (C.4) and (C.5) gives that with probability $1-\delta$,

$$\sum_{t<i}\big(\mathcal{Q}(W_t) - \mathcal{Q}_t(W_t)\big) \leq (e-2)\sum_{t<i}\mathcal{Q}(W_t) + \ln\left(\frac{1}{\delta}\right).$$

Consequently,

$$\sum_{t<i}\mathcal{Q}(W_t) \leq 4\sum_{t<i}\mathcal{Q}_t(W_t) + 4\ln\left(\frac{1}{\delta}\right).$$

□

Finally, we prove Theorem 4.1.

Proof of Theorem 4.1. Suppose the condition of Lemma C.1 holds. Then we have for $n = \lceil 2\lambda^2/\eta\epsilon \rceil$, with probability $1-3\delta$,

$$\frac{1}{n}\sum_{i<n}\mathcal{Q}_i(W_i) \leq \epsilon.$$

Further invoking Lemma 4.3 gives that with probability $1-4\delta$,

$$\frac{1}{n}\sum_{i<n}\mathcal{Q}(W_i) \leq \frac{4}{n}\sum_{i<n}\mathcal{Q}_i(W_i) + \frac{4}{n}\ln\left(\frac{1}{\delta}\right) \leq 5\epsilon.$$

Since $P_{(x,y)\sim\mathcal{D}}\big(yf(x;W,a)\leq 0\big) \leq 2\mathcal{Q}(W)$, we get

$$\frac{1}{n}\sum_{i=1}^{n} P_{(x,y)\sim\mathcal{D}}\big(yf(x;W_i,a)\leq 0\big) \leq 10\epsilon.$$

For the condition of Lemma C.1 to hold, it is enough to let

$$n_0 = \Theta\left(\frac{\ln(1/\delta)}{\eta\gamma^2\epsilon^2}\right),$$

which gives

$$M = \Theta\left(\frac{\ln(1/\delta) + \ln(1/\epsilon)^2}{\gamma^5}\right) \quad \text{and} \quad n = \Theta\left(\frac{\ln(1/\delta) + \ln(1/\epsilon)^2}{\gamma^2\epsilon}\right).$$

□

D OMITTED PROOFS FROM SECTION 5

Proof of Proposition 5.1. Define $f : \mathcal{H} \to \mathbb{R}$ by

$$f(w) := \frac{1}{2}\int \|w(z)\|_2^2 \,\mathrm{d}\mu_{\mathcal{N}}(z) = \frac{1}{2}\|w\|_{\mathcal{H}}^2.$$

It holds that f is continuous, and f^* has the same form. Define $g : \mathbb{R}^n \to \mathbb{R}$ by

$$g(p) := \max_{1\le i\le n} p_i,$$

with conjugate

$$g^*(q) = \begin{cases} 0, & \text{if } q \in \Delta_n, \\ +\infty, & \text{o.w.} \end{cases}$$

Finally, define the linear mapping $A : \mathcal{H} \to \mathbb{R}^n$ by $(Aw)_i = y_i \langle w, \phi_i \rangle$.

Since f, f^*, g and g^* are lower semi-continuous, and $\mathbf{dom}\, g - A\mathbf{dom}\, f = \mathbb{R}^n$, and $\mathbf{dom}\, f^* - A^*\mathbf{dom}\, g^* = \mathcal{H}$, Theorem 4.4.3 of (Borwein & Zhu, 2005) ensures that

$$\inf_{w\in\mathcal{H}} \big(f(w) + g(Aw)\big) = \sup_{q\in\mathbb{R}^n} \big(-f^*(A^*q) - g^*(-q)\big).$$

with optimal primal-dual solutions $(\bar{w}, \bar{q})$. Moreover

$$\begin{aligned}
\inf_{w\in\mathcal{H}} \big(f(w) + g(Aw)\big) &= \inf_{w\in\mathcal{H}, u\in\mathbb{R}^n} \sup_{q\in\mathbb{R}^n} \big(f(w) + g(Aw+u) + \langle q, u\rangle\big) \\
&\ge \sup_{q\in\mathbb{R}^n} \inf_{w\in\mathcal{H}, u\in\mathbb{R}^n} \big(f(w) + g(Aw+u) + \langle q, u\rangle\big) \\
&= \sup_{q\in\mathbb{R}^n} \inf_{w\in\mathcal{H}, u\in\mathbb{R}^n} \Big(\big(f(w) - \langle A^*q, w\rangle\big) + \big(g(Aw+u) - \langle -q, Aw+u\rangle\big)\Big) \\
&= \sup_{q\in\mathbb{R}^n} \big(-f^*(A^*q) - g^*(-q)\big).
\end{aligned}$$

By strong duality, the inequality holds with equality. It follows that

$$\bar{w} = A^*\bar{q}, \quad \text{and} \quad \mathbf{supp}(-\bar{q}) \subset \arg\max_{1\le i\le n} (A\bar{w})_i.$$

Now let us look at the dual optimization problem. It is clear that

$$\sup_{q\in\mathbb{R}^n} \big(-f^*(A^*q) - g^*(-q)\big) = -\inf_{q\in\Delta_n} f^*(A^*q).$$

In addition, we have

$$\begin{aligned}
f^*(A^*q) &= \frac{1}{2}\int \left\|\sum_{i=1}^n q_i y_i \phi_i(z)\right\|_2^2 \mathrm{d}\mu_{\mathcal{N}}(z) \\
&= \frac{1}{2}\int \sum_{i,j=1}^n q_i q_j y_i y_j \big\langle \phi_i(z), \phi_j(z)\big\rangle \,\mathrm{d}\mu_{\mathcal{N}}(z) \\
&= \frac{1}{2}\sum_{i,j=1}^n q_i q_j y_i y_j \int \big\langle \phi_i(z), \phi_j(z)\big\rangle \,\mathrm{d}\mu_{\mathcal{N}}(z) \\
&= \frac{1}{2}\sum_{i,j=1}^n q_i q_j y_i y_j K_1(i,j) = \frac{1}{2}(q\odot y)^\top K_1(q\odot y),
\end{aligned}$$

and thus $f^*(A^*\bar{q}) = \gamma_1^2/2$. Since $\bar{w} = A^*\bar{q}$, we have that $\|\bar{w}\|_{\mathcal{H}} = \gamma_1$. In addition,

$$g(A\bar{w}) = -f^*\left(A^*\bar{q}\right) - f\left(\bar{w}\right) = -\gamma_1^2,$$

and thus $-\bar{w}$ has margin γ_1^2. Moreover, we have

$$\bar{w}(z) = \sum_{i=1}^{n} \bar{q}_i y_i \phi_i(z) = \sum_{i=1}^{n} \bar{q}_i y_i x_i \mathbb{1}\left[\langle z, x_i \rangle > 0\right],$$

and thus $\left\|\bar{w}(z)\right\|_2 \leq 1$. Therefore, $\hat{v} = -\bar{w}$ satisfies all requirements of Proposition 5.1. □