On the Proof of Global Convergence of Gradient Descent for Deep ReLU Networks with Linear Widths

Quynh Nguyen [1]

Abstract

We give a simple proof for the global convergence of gradient descent in training deep ReLU networks with the standard square loss, and show some of its improvements over the state-of-the-art. In particular, while prior works require all the hidden layers to be wide with width at least $\Omega(N^8)$ (N being the number of training samples), we require a single wide layer of linear, quadratic or cubic width depending on the type of initialization. Unlike many recent proofs based on the Neural Tangent Kernel (NTK), our proof need *not* track the evolution of the entire NTK matrix, or more generally, any quantities related to the changes of activation patterns during training. Instead, we only need to track the evolution of the output at the last hidden layer, which can be done much more easily thanks to the Lipschitz property of ReLU. Some highlights of our setting: (i) all the layers are trained with standard gradient descent, (ii) the network has standard parameterization as opposed to the NTK one, and (iii) the network has a single wide layer as opposed to having all wide hidden layers as in most of NTK-related results.

1. Introduction

Understanding why gradient descent methods can often find a global optimum in minimizing the non-convex loss surface of neural nets is one of the key problems in deep learning theory. Recently, this problem has been approached from the perspective of the Neural Tangent Kernel (NTK) (Jacot et al., 2018). Consider an L-layer network with layer widths $\{n_0, \ldots, n_L\}$. Let $X \in \mathbb{R}^{N \times n_0}$ and $Y \in \mathbb{R}^{N \times n_L}$ be the

[1] MPI-MIS, Germany. Correspondence to: Quynh Nguyen <qnguyen@mis.mpg.de>.

training data, then the output at layer l is given by

$$F_l = \begin{cases} X & l = 0, \\ \sigma(F_{l-1} W_l) & l \in [L-1], \\ F_{L-1} W_L & l = L, \end{cases} \tag{1}$$

where $W_l \in \mathbb{R}^{n_{l-1} \times n_l}$, and $\sigma(x) = \max(0, x)$ applied componentwise. The training loss is defined as

$$\Phi(\theta) = \frac{1}{2} \|F_L(\theta) - Y\|_F^2,$$

where $\theta = (W_l)_{l=1}^L$. The NTK matrix is $K(\theta) = \left[\frac{\partial \text{vec}(F_L)}{\partial \text{vec}(\theta)}\right] \left[\frac{\partial \text{vec}(F_L)}{\partial \text{vec}(\theta)}\right]^T$. Our starting observation is that

$$\|\nabla \Phi(\theta)\|_2^2 \geq 2\lambda_{\min}(K(\theta)) \Phi(\theta).$$

This resembles the Polyak-Lojasiewicz (PL) inequality (Polyak, 1963) except that we do not have a constant on the RHS. Nevertheless, it suggests that if $\lambda_{\min}(K(\theta))$ is bounded away from zero, both at initialization and during training, then GD may still converge to a global optimum. To show this property at initialization, one can resort to standard concentration tools, which we discuss later. To show it holds throughout training, one popular way in the literature is to bound the traveling distance of GD from the initialization and the local Lipschitz constant of $K(\theta)$, and then invoke Weyl's eigenvalue inequality. As K involves the derivatives of σ at various activation neurons, it turns out that bounding the local Lipschitz constant of the NTK matrix can be done conveniently for sufficiently smooth activations (Du et al., 2019a). However, when σ' is not Lipschitz continuous (e.g. for ReLU), the entire analysis may become much more involved. In particular, most of the existing proofs for ReLU nets require to study various quantities related to the changes of the activation patterns during training, see e.g. (Allen-Zhu et al., 2019; Zou & Gu, 2019). In this paper, we discuss a simple alternative proof, where such analysis is not needed. In fact, starting from the following decomposition of the empirical NTK matrix

$$K = \sum_{l=1}^L \left[\frac{\partial \text{vec}(F_L)}{\partial \text{vec}(W_l)}\right] \left[\frac{\partial \text{vec}(F_L)}{\partial \text{vec}(W_l)}\right]^T \tag{2}$$

one can separate the sum into two terms: the first term involves $\sum_{l=1}^{L-1}(\cdot)$, and the second term corresponds to the derivative at W_L, which is given by $\mathbb{I}_{n_L} \otimes (F_{L-1}F_{L-1}^T)$. Since both terms are positive semidefinite, one obtains

$$\lambda_{\min}(K) \geq \lambda_{\min}\left(F_{L-1}F_{L-1}^T\right). \tag{3}$$

From here, in order to obtain a PL-like inequality, it suffices to bound the RHS of (3) at initialization and keep track of F_{L-1} during training. This can be done efficiently without studying the changes of activation patterns. Let us highlight that this is different from most of the prior works, where the output layer is fixed and only the hidden ones are optimized, in which case the PL analysis necessarily involves σ' from the hidden layers. Besides that, prior works also need to study other "smoothness" properties of the loss (local Lipschitz property of the gradient, or that of the Jacobian of the network), which again requires bounds on the various quantities related to the changes of the activation patterns at different layers, and thus make their proofs more involved.

In this work, we present an alternative proof framework, which does not require any analysis of the activation pattern changes. It consists of two key steps mentioned above: (a) bounding $\lambda_{\min}\left(F_{L-1}F_{L-1}^T\right)$ at initialization, and (b) bounding the changes of F_{L-1} during training. To show the convergence, we introduce a small trick to relate the loss of each iteration with that of the previous one. That means, an explicit analysis of the "smoothness" of the loss is not needed. Lastly, we show that our proof leads to improved bounds on layer widths in terms of the dependency on N.

Main Contributions. First, we provide some easy-to-check sufficient conditions on the initialization under which GD is guaranteed to converge to a global optimum. Then, we show that all these conditions are satisfied when the width of the last hidden layer exceeds the number of training samples (all the remaining layers can have constant widths). For LeCun's initialization, we show that these conditions are satisfied (and consequently, GD finds a global optimum) if the last hidden layer has quadratic width in case of two-layer networks, or cubic width in case of deep architectures. Although results with similar orders of over-parameterization have been recently obtained for deep nets with *smooth* activation functions (Huang & Yau, 2020; Nguyen & Mondelli, 2020), this is the first time, to the best of our knowledge, that such a result is proved for deep ReLU models.

2. Main Result

The GD updates are given by: $\theta_{k+1} = \theta_k - \eta\nabla\Phi(\theta_k)$. Define the shorthand $F_l^k = F_l(\theta_k)$. We omit the argument θ and write just F_l when it is clear from the context.

Let us first state some useful inequalities (see Lemma B.2 of (Nguyen & Mondelli, 2020)). There, the proof mainly uses the triangle inequality, and the Lip. property of ReLU.

Lemma 2.1 *For every $\theta = (W_p)_{p=1}^L$ and $l \in [L]$, it holds*

$$\|\nabla_{W_l}\Phi\|_F \leq \|X\|_F \prod_{\substack{p=1 \\ p\neq l}}^L \|W_p\|_2 \|F_L - Y\|_F. \tag{4}$$

Furthermore, let $\theta_a = (W_l^a)_{l=1}^L$, $\theta_b = (W_l^b)_{l=1}^L$, and $\max(\|W_l^a\|_2, \|W_l^b\|_2) \leq \bar{\lambda}_l$ for some $\bar{\lambda}_l \in \mathbb{R}$. Then, for every $l \in [L]$ we have

$$\|F_l(\theta_a) - F_l(\theta_b)\|_F$$
$$\leq \|X\|_F \left(\prod_{l=1}^l \bar{\lambda}_l\right) \sum_{p=1}^l \bar{\lambda}_p^{-1} \|W_p^a - W_p^b\|_2 \tag{5}$$

Our main theorem is the following.

Theorem 2.2 *Consider a deep ReLU network (1) where the width of the last hidden layer satisfies $n_{L-1} \geq N$. Let $(C_l)_{l=1}^L$ be any sequence of positive numbers. Define the following quantities:*

$$\alpha_0 = \sigma_{\min}\left(F_{L-1}^0\right), \quad \bar{\lambda}_l = \|W_l^0\|_2 + C_l, \quad \bar{\lambda}_{i\to j} = \prod_{l=i}^j \bar{\lambda}_l. \tag{6}$$

Assume that the following conditions are satisfied at the initialization:

$$\alpha_0^2 \geq 16 \|X\|_F \max_{l\in[L]} \frac{\bar{\lambda}_{1\to L}}{\bar{\lambda}_l C_l} \sqrt{2\Phi(\theta_0)} \tag{7}$$

$$\alpha_0^3 \geq 32 \|X\|_F^2 \bar{\lambda}_L \sum_{l=1}^{L-1} \frac{\bar{\lambda}_{1\to L-1}^2}{\bar{\lambda}_l^2} \sqrt{2\Phi(\theta_0)} \tag{8}$$

$$\alpha_0^2 \geq 16 \|X\|_F^2 \bar{\lambda}_L^2 \sum_{l=1}^{L-1} \frac{\bar{\lambda}_{1\to L-1}^2}{\bar{\lambda}_l^2} \tag{9}$$

Let the learning rate satisfy

$$\eta < \min\left(\frac{8}{\alpha_0^2}, \|X\|_F^{-2}\bar{\lambda}_{1\to L}^{-2}\left[\sum_{l=1}^{L-1}\bar{\lambda}_l^{-2}\right]\left[\sum_{l=1}^L\bar{\lambda}_l^{-2}\right]^{-2}\right) \tag{10}$$

Then the loss converges to a global minimum as

$$\Phi(\theta_k) \leq \left(1 - \eta\frac{\alpha_0^2}{8}\right)^k \Phi(\theta_0), \tag{11}$$

for every $k \geq 0$.

Proof: We show by induction for every $k \geq 0$ that

$$\begin{cases} \|W_l^r\|_2 \leq \bar{\lambda}_l, & l \in [L], r \in [0,k], \\ \sigma_{\min}\left(F_{L-1}^r\right) \geq \frac{1}{2}\alpha_0, & r \in [0,k], \\ \Phi(\theta_r) \leq \left(1 - \eta\frac{\alpha_0^2}{8}\right)^r \Phi(\theta_0), & r \in [0,k], \end{cases} \tag{12}$$

Clearly, (12) holds for $k = 0$. Assume that (12) holds up to iteration k. By the triangle inequality,

$$\left\|W_l^{k+1} - W_l^0\right\|_F \leq \sum_{s=0}^{k} \left\|W_l^{s+1} - W_l^s\right\|_F$$

$$= \eta \sum_{s=0}^{k} \left\|\nabla_{W_l}\Phi(\theta_s)\right\|_F$$

$$\leq \eta \sum_{s=0}^{k} \|X\|_F \prod_{\substack{p=1 \\ p\neq l}}^{L} \|W_p^s\|_2 \|F_L^s - Y\|_F$$

$$\leq \eta \|X\|_F \bar{\lambda}_{1\to L}\bar{\lambda}_l^{-1} \sum_{s=0}^{k} \left(1 - \eta\frac{\alpha_0^2}{8}\right)^{s/2} \|F_L^0 - Y\|_F,$$

where the 2nd inequality follows from (4), and the last one follows from induction assumption. Let $u := \sqrt{1 - \eta\alpha_0^2/8}$. The RHS of the previous expression is bounded as

$$\frac{8}{\alpha_0^2} \|X\|_F \bar{\lambda}_{1\to L}\bar{\lambda}_l^{-1}(1 - u^2)\frac{1 - u^{k+1}}{1 - u} \|F_L^0 - Y\|_F$$

$$\leq \frac{16}{\alpha_0^2} \|X\|_F \bar{\lambda}_{1\to L}\bar{\lambda}_l^{-1} \|F_L^0 - Y\|_F, \quad \text{since } u \in (0, 1)$$

$$\leq C_l, \quad \text{by (7).}$$

By Weyl's inequality, this implies

$$\left\|W_l^{k+1}\right\|_2 \leq \left\|W_l^0\right\|_2 + C_l = \bar{\lambda}_l \tag{13}$$

Next, we have

$$\left\|F_{L-1}^{k+1} - F_{L-1}^0\right\|_F$$

$$\leq \|X\|_F \bar{\lambda}_{1\to L-1} \sum_{l=1}^{L-1} \bar{\lambda}_l^{-1} \left\|W_l^{k+1} - W_l^0\right\|_2, \quad \text{by (5)}$$

$$\leq \frac{16}{\alpha_0^2} \|X\|_F^2 \bar{\lambda}_{1\to L-1}\bar{\lambda}_{1\to L} \sum_{l=1}^{L-1} \bar{\lambda}_l^{-2} \sqrt{2\Phi(\theta_0)}$$

$$\leq \frac{1}{2}\alpha_0, \quad \text{by (8).}$$

This implies that $\sigma_{\min}\left(F_{L-1}^{k+1}\right) \geq \frac{1}{2}\alpha_0$. So far, we have proved that the first two inequalities in (12) hold for $k + 1$. It remains to show that the third one also holds at $k+1$. Let us define the matrix $G = F_{L-1}^k W_L^{k+1}$. Then, one has

$$2\Phi(\theta_{k+1})$$

$$= 2\Phi(\theta_k) + \left\|F_L^{k+1} - F_L^k\right\|_F^2 + 2\operatorname{tr}(F_L^{k+1} - F_L^k)(F_L^k - Y)^T$$

$$= 2\Phi(\theta_k) + \left\|F_L^{k+1} - F_L^k\right\|_F^2 + 2\operatorname{tr}(F_L^{k+1} - G)(F_L^k - Y)^T$$

$$+ 2\operatorname{tr}(G - F_L^k)(F_L^k - Y)^T.$$

Let us bound each term individually. Using (4)-(5), we have

$$\left\|F_L^{k+1} - F_L^k\right\|_F \leq \|X\|_F \bar{\lambda}_{1\to L} \sum_{l=1}^{L} \bar{\lambda}_l^{-1} \left\|W_l^{k+1} - W_l^k\right\|_2$$

$$\leq \eta \|X\|_F^2 \bar{\lambda}_{1\to L}^2 \underbrace{\sum_{l=1}^{L} \bar{\lambda}_l^{-2}}_{Q_1} \left\|F_L^k - Y\right\|_F.$$

Furthermore, we have $F_L^{k+1} - G = (F_{L-1}^{k+1} - F_{L-1}^k)W_L^{k+1}$, and thus it holds

$$\operatorname{tr}(F_L^{k+1} - G)(F_L^k - Y)^T$$

$$\leq \left\|F_{L-1}^{k+1} - F_{L-1}^k\right\|_F \left\|W_L^{k+1}\right\|_2 \left\|F_L^k - Y\right\|_F$$

$$\leq \eta \|X\|_F^2 \bar{\lambda}_{1\to L-1}^2 \bar{\lambda}_L^2 \underbrace{\sum_{l=1}^{L-1} \bar{\lambda}_l^{-2}}_{Q_2} \left\|F_L^k - Y\right\|_F^2, \quad \text{by (5), (13)}$$

Lastly, by using the definition of G and the fact that $\nabla_{W_L}\Phi(\theta_k) = (F_{L-1}^k)^T(F_L^k - Y)$, we get

$$\operatorname{tr}(G - F_L^k)(F_L^k - Y)^T$$

$$= -\eta \operatorname{tr}((F_{L-1}^k)^T(F_L^k - Y)(F_L^k - Y)^T F_{L-1}^k)$$

$$\leq -\eta\frac{\alpha_0^2}{4} \left\|F_L^k - Y\right\|_F^2$$

where we used our assumption $n_{L-1} \geq N$ to obtain $\lambda_{\min}\left((F_{L-1}^k)^T(F_{L-1}^k)\right) = \sigma_{\min}\left(F_{L-1}^k\right)^2$, and the induction assumption implies that $\sigma_{\min}\left(F_{L-1}^k\right) \geq \frac{1}{2}\alpha_0$. Define Q_1, Q_2 as above. Putting all these bounds together, we get

$$\Phi(\theta_{k+1}) \leq \left[1 - \eta\frac{\alpha_0^2}{4} + \eta^2 Q_1^2 + \eta Q_2\right]\Phi(\theta_k)$$

$$\leq \left[1 - \eta\left[\frac{\alpha_0^2}{4} - 2Q_2\right]\right]\Phi(\theta_k) \qquad \text{By (10)}$$

$$\leq \left[1 - \eta\frac{\alpha_0^2}{8}\right]\Phi(\theta_k) \qquad \text{By (9)}$$

$\square$

We remark that Theorem 2.2 also holds for networks with biases. Furthermore, its statement is meaningful for $\alpha_0 > 0$, in which case there exists a θ such that F_{L-1} has full row rank, and thus the network can fit arbitrary labels.

Convergence of gradient descent for deep ReLU nets under the square loss has been studied in (Allen-Zhu et al., 2019; Zou & Gu, 2019), where the widths of all the hidden layers scales as order of $N^8\text{poly}(L)$. Here, Theorem 2.2 shows that a single wide layer of width N suffices, and in particular, all the remaining layers can have arbitrary constant widths. We remark that these results are not directly comparable at this point since it is not clear how likely the conditions (7)-(9) are satisfied when a common/practical initialization is considered. We address this question in the next section.

3. Applications of Theorem 2.2

This section discusses the satisfiability of the conditions (7)-(9) of Theorem 2.2 under different initialization schemes. To do so, it suffices to show a lower bound on α_0 (as given in (6)), and an upper bound on both $\bar{\lambda}_l$ and the initial loss $\sqrt{2\Phi(\theta_0)}$, and finally plugging these bounds into the above sufficient conditions to get a concrete requirement on layer widths. More details are given in the following sections.

3.1. Satisfiability of (7)-(9): Width N Suffices

First, let us show that all the conditions (7)-(9) of Theorem 2.2 are satisfied for a subset of points in parameter space. In particular, this does not require any additional requirements on the layer widths. Thus, if GD starts from one of these initial points, then it will converge to a global optimum provided that the width of the last hidden layer is at least N.

We apply Theorem 2.2 for $C_l = 1, l \in [L]$. Let $\theta_0 = (W_1^0, \ldots, W_{L-1}^0, W_L^0 = 0)$, where $(W_l^0)_{l=1}^{L-1}$ are chosen s.t. $\alpha_0 = \sigma_{\min}(F_{L-1}(\theta_0)) > 0$. Since $W_L^0 = 0$, we have $\sqrt{2\Phi(\theta_0)} = \|Y\|_F$. Let $\tilde{\theta}_0 = (\beta W_1^0, \ldots, \beta W_{L-1}^0, W_L = 0)$. Then, the condition (7) is satisfied at $\tilde{\theta}_0$ if

$$\beta^{2(L-1)}\alpha_0^2 \geq 16 \|X\|_F \|Y\|_F \prod_{l=1}^{L-1}(\beta \|W_l^0\|_2 + 1). \quad (14)$$

The LHS of the above inequality is a polynomial of degree $2(L-1)$ in β, whereas the RHS is a polynomial of degree $L-1$. Thus, once the parameters of θ_0 are fixed (as above), the inequality (14) will be satisfied for a sufficiently large value of β. Similarly, the condition (9) is satisfied at $\tilde{\theta}_0$ if

$$\beta^{2(L-1)}\alpha_0^2 \geq 16 \|X\|_F^2 \sum_{l=1}^{L-1} \prod_{\substack{p=1 \\ p \neq l}}^{L-1}(\beta \|W_p^0\|_2 + 1)^2. \quad (15)$$

The LHS of (15) has degree $2(L-1)$, whereas the RHS has degree $2(L-2)$. Thus, (15) is satisfied for large enough β. The same argument applies to (8). As a result, all the conditions of Theorem 2.2 are met at $\tilde{\theta}_0$ for large enough β.

3.2. LeCun's Initialization for Two-Layer ReLU Nets: Width N^2 Suffices

In this section, we show that all the conditions (7)-(9) of Theorem 2.2 are satisfied for two-layer networks under Le-Cun's initialization, provided that a stronger condition on the width is satisfied, namely $n_{L-1} = \Omega(N^2)$. The more general case of deep architecture is analyzed in Section 3.3.

For simplicity, we assume the following assumptions on the data (standard in the literature): *(i)* the samples are i.i.d. sub-gaussian vectors with $\|X_{i:}\|_2 = \sqrt{d}$ and $\|X_{i:}\|_{\psi_2} = \mathcal{O}(1)$, *(ii)* the ground-truth labels satisfy $\|Y_{i:}\|_2 = \mathcal{O}(1)$ for all

$i \in [N]$, and *(iii)* the number of outputs n_2 is a constant. For 2-layer nets, the conditions (7)-(9) become:

$$\alpha_0^2 \geq 16 \|X\|_F \max\left(\frac{\bar{\lambda}_2}{C_1}, \frac{\bar{\lambda}_1}{C_2}\right) \sqrt{2\Phi(\theta_0)} \quad (16)$$

$$\alpha_0^3 \geq 32 \|X\|_F^2 \bar{\lambda}_2 \sqrt{2\Phi(\theta_0)} \quad (17)$$

$$\alpha_0^2 \geq 16 \|X\|_F^2 \bar{\lambda}_2^2 \quad (18)$$

where $\alpha_0 = \sigma_{\min}(XW_1^0)$, $\bar{\lambda}_1 = \|W_1^0\|_2 + C_1$, $\bar{\lambda}_2 = \|W_2^0\|_2 + C_2$, $W_1^0 \in \mathbb{R}^{n_0 \times n_1}$ and $W_2^0 \in \mathbb{R}^{n_1 \times n_2}$. Pick $C_1 = C_2 = 1$. Consider LeCun's initialization:

$$(W_1^0)_{ij} \sim \mathcal{N}(0, 1/n_0), \ (W_2^0)_{ij} \sim \mathcal{N}(0, 1/n_1). \quad (19)$$

Then, by standard bounds on the operator norm of Gaussian matrices (see Theorem 2.12 of (Davidson & Szarek, 2001)), we have w.p. $\geq 1 - e^{-\Omega(n_1)}$ that

$$\bar{\lambda}_1 = \mathcal{O}\left(\frac{\sqrt{n_1}}{\sqrt{n_0}}\right), \quad \bar{\lambda}_2 = \mathcal{O}(1). \quad (20)$$

Using Matrix-Chernoff inequality, one can easily show that (see e.g. Lemma 5.2 of (Nguyen et al., 2021)) w.p. $\geq 1 - \delta$,

$$\alpha_0 \geq \sqrt{n_1 \lambda_*/4}, \quad (21)$$

as long as it holds $n_1 \geq \tilde{\Omega}(N/\lambda_*)$, where $\lambda_* = \lambda_{\min}\left(\mathbb{E}_{w \sim \mathcal{N}(0, n_0^{-1}\mathbb{I}_{n_0})}[\sigma(Xw)\sigma(Xw)^T]\right)$ and $\tilde{\Omega}$ hides logarithmic factors depending on δ. Similarly, by using a standard concentration argument, we have w.p. $\geq 1 - \delta$

$$\sqrt{2\Phi(\theta_0)} \leq \tilde{\mathcal{O}}(\sqrt{N}). \quad (22)$$

The proof of this can be found in Lemma C.1 of (Nguyen & Mondelli, 2020). Plugging all these bounds into (16)-(18), we obtain that all these conditions are satisfied for $n_1 = \Omega(N^2 \lambda_*^{-2})$. Lastly, our data assumptions stated above allow us to apply Theorem 3.3 of (Nguyen & Mondelli, 2020), which implies w.h.p. that $\lambda_* \geq \Omega(1)$ as long as $d \geq N^r$ for some fixed integer constant $r \geq 2$. Thus, all the conditions of Theorem 2.2 are satisfied for $n_1 = \Omega(N^2)$.

As a remark, quadratic bounds on the layer width for two-layer ReLU nets have been also claimed in (Song & Yang, 2020) (albeit for a different initialization). The main differences consist in the fact that (i) we train all layers of the network while (Song & Yang, 2020) train only the hidden layer, and (ii) our proof works for *arbitrary* depth, and (iii) it does not require to bound the number of sign flips of the activation neurons, nor the Lipschitz constant of the gradient. This last property makes our proof significant simpler.

3.3. LeCun's Initialization for Deep ReLU Nets: One Layer of Width N^3 Suffices

This section considers the same type of initialization as in Section 3.2, except two differences: i) we treat the case

of deep nets, and ii) the variance of the output weights at the initialization is chosen to be smaller. Some previous studies e.g. (Allen-Zhu et al., 2019; Zou & Gu, 2019; Chen et al., 2019) consider a regime in which the depth L may grow with N. In this section, we regard L as a constant and obtain improved bounds in terms of the dependency on N. More specifically, the best existing over-parameterization condition for the current setting (namely, deep ReLU nets + square loss) is given in (Zou & Gu, 2019), where every hidden layer is required to have $\Omega(N^8)$ neurons. The following corollary shows that all the conditions (7)-(9) of Theorem 2.2 are satisfied for $n_{L-1} = \Omega(N^3)$. In particular, all the remaining layers can have arbitrary *sublinear* widths.

Let us assume that the training samples are drawn i.i.d. from the following family of distributions, denoted as P_X.

Assumption 3.1 *P_X satisfies the following conditions:*

1. *For every Lipschitz continuous function f : $\mathbb{R}^{n_0} \to \mathbb{R}$, we have $\mathbb{P}(|f(x) - \mathbb{E}f(x)| > t) \le 2\exp\left(-ct^2/\|f\|_{\mathrm{Lip}}^2\right)$ for all $t > 0$, where $c > 0$ is some absolute constant.*

2. *$\mathbb{E}\|x\|_2 = \Theta(\sqrt{n_0})$, $\mathbb{E}\|x\|_2^2 = \Theta(n_0)$ and $\mathbb{E}\|x - \mathbb{E}x\|_2^2 = \Omega(n_0)$.*

Many distributions of interest satisfy Assumption 3.1, e.g. the standard Gaussian distribution and the uniform distribution on the sphere. In general, the first condition could cover all those distributions satisfying the log-Sobolev inequality with data-independent constant. The second condition is just about the scaling of the data. Here, we assume the data has norm of order $\sqrt{n_0}$, but it can also have other scalings.

The result of this section is stated below.

Corollary 3.2 *Consider a set of i.i.d. samples $\{X_{i:}\}_{i=1}^N$ from a data distribution which satisfies Assumption 3.1. Let the true labels satisfy $\|Y_{i:}\|_2 = \mathcal{O}(1)$ for all $i \in [N]$. Fix any even integer constant $r > 0$, and $\delta > 0$. Let the widths of the network satisfy the following conditions:*

$$n_l = \Theta(m), \quad \forall l \in [L-2], \tag{23}$$
$$n_{L-1} = \tilde{\Omega}\left(N^3 \max(m, n_0)^3\right) \tag{24}$$

where m is some variable which can depend on N, and $\tilde{\Omega}$ hides logarithmic terms in N and δ^{-1}. Let the weights of the network be initialized as

$$(W_l)_{ij} \sim \mathcal{N}\left(0, \frac{1}{n_{l-1}}\right), \forall l \in [L-1], \tag{25}$$
$$(W_L)_{ij} \sim \mathcal{N}\left(0, \frac{1}{n_{L-1}^{4/3}}\right). \tag{26}$$

Then, for a small enough learning rate, gradient descent converges to a global optimum w.p. at least

$$1 - \delta - N^2 \exp\left(-\Omega\left(\frac{\min(m, n_0)}{N^{2/(r-0.1)}\log(m)^{L-3}}\right)\right).$$

We remark that the probability of Corollary 3.2 can be made arbitrarily close to 1 as long as $\min(m, n_0)$ is not super-polynomially smaller than N. For instance, one can pick $m, n_0 = \tilde{\Theta}(N^{1/p})$ for a sufficiently large constant $p > 0$, in which case the main width condition of Corollary 3.2 becomes $n_{L-1} = \Omega(N^{3+\epsilon})$, for some small $\epsilon \in (0, 1)$.

In order to prove Corollary 3.2, we use the following bound on α_0 from (Nguyen et al., 2021). Below is a restatement of their result for the special case of our interest (the smallest singular value of the last hidden layer).

Lemma 3.3 *Let Assumption 3.1 hold. Let $(W_l)_{ij} \sim \mathcal{N}(0, \beta_l^2)$ for all $l \in [L-1], i \in [n_{l-1}], j \in [n_l]$. Fix any even integer $r > 0$, and $\delta > 0$. Suppose that*

$$n_{L-1} = \Omega\left(N\log(N)\log\left(\frac{N}{\delta}\right)\right), \tag{27}$$

$$\prod_{l=1}^{L-3}\log(n_l) = o\left(\min_{l\in[0,L-2]} n_l\right). \tag{28}$$

Then one has

$$\sigma_{\min}(F_{L-1})^2 = \Theta\left(n_0\prod_{l=1}^{L-1}n_l\beta_l^2\right) \tag{29}$$

w.p. $\ge 1 - \delta - N^2 \exp\left(-\Omega\left(\frac{\min_{l\in[0,L-2]} n_l}{N^{2/(r-0.1)}\prod_{l=1}^{L-3}\log(n_l)}\right)\right).$

We are now ready to prove the corollary.

Proof of Corollary 3.2. Let $a = \frac{4}{3}, b = -\frac{1}{6}, c = \frac{1}{2}$. By applying Theorem 2.2 for

$$C_l = \begin{cases} 1 & l \in [L-2], \\ n_{L-1}^c & l = L-1, \\ n_{L-1}^b & l = L, \end{cases}$$

it suffices to show that all the conditions (7),(8),(9) are satisfied under the stated assumptions. By Lemma 3.3, we have $\alpha_0^2 = \Omega(n_{L-1})$ with the same probability as stated in the corollary. By Theorem 4.4.5 of (Vershynin, 2018), we have w.p. $\ge 1 - e^{-\Omega(\min(n_0, m, n_{L-1}))}$ that

$$\bar{\lambda}_l = \begin{cases} \mathcal{O}\left(n_{L-1}^{(1-a)/2} + n_{L-1}^b\right) & l = L \\ \mathcal{O}\left(\frac{\sqrt{n_{L-1}}}{\sqrt{m}} + n_{L-1}^c\right) & l = L-1 \\ \mathcal{O}(1) & l \in [2, L-2] \\ \mathcal{O}\left(\frac{\max(\sqrt{m}, \sqrt{n_0})}{\sqrt{n_0}}\right) & l = 1 \end{cases}$$

Define the shorthand $D = \|X\|_F \sqrt{2\Phi(\theta_0)}$. Then, the condition (7) is satisfied for

$$
\begin{cases}
n_{L-1}^{1+b-\max(1/2,c)} \gtrsim D \max\left(1, \sqrt{\frac{m}{n_0}}\right), \\
n_{L-1}^{1+c-\max(b,(1-a)/2)} \gtrsim D \max\left(1, \sqrt{\frac{m}{n_0}}\right), \\
n_{L-1}^{1-\max(1/2,c)-\max(b,(1-a)/2)} \gtrsim D \max\left(1, \sqrt{\frac{m}{n_0}}\right).
\end{cases}
$$

Next, the condition (9) is satisfied for

$$
n_{L-1}^{1-2\max(1/2,c)-2\max(b,(1-a)/2)} \gtrsim \max\left(1, \frac{m}{n_0}\right) \|X\|_F^2 .
$$

Given the above condition, (8) is satisfied if it holds in addition that $\alpha_0 \gtrsim \bar{\lambda}_L^{-1} \sqrt{2\Phi(\theta_0)}$, or equivalently,

$$
n_{L-1}^{1/2+\max(b,(1-a)/2)} \gtrsim \sqrt{2\Phi(\theta_0)}.
$$

So far, all the conditions (7),(8),(9) are satisfied for

$$
\begin{cases}
n_{L-1}^{1/3} \gtrsim \|X\|_F^2 \max\left(1, \frac{m}{n_0}\right), \\
n_{L-1}^{1/3} \gtrsim \|X\|_F \sqrt{2\Phi(\theta_0)} \max\left(1, \frac{\sqrt{m}}{\sqrt{n_0}}\right), \\
n_{L-1}^{1/3} \gtrsim \sqrt{2\Phi(\theta_0)}.
\end{cases} \quad (30)
$$

By Lemma C.1 of (Nguyen & Mondelli, 2020), we have w.p. $\geq 1 - e^{-\Omega(n_0)}$ that $\|F_L(\theta_0)\|_F \lesssim \|X\|_F$. Moreover, Assumption 3.1 implies $\|x_i\| = \Theta(n_0)$ w.p. $\geq 1 - e^{-\Omega(n_0)}$. Taking the union bound, one obtains $\|X\|_F = \mathcal{O}(\sqrt{Nn_0})$ w.p. $\geq 1 - Ne^{-\Omega(n_0)}$. By our data assumption, it holds $\|Y\|_F = \mathcal{O}(\sqrt{N})$. These last two facts imply $\sqrt{2\Phi(\theta_0)} \leq \|F_L(\theta_0)\|_F + \|Y\|_F \leq \mathcal{O}(\sqrt{Nn_0})$. Combining all the bounds above, one conclude that all the conditions in (30) are satisfied for $n_{L-1} = \tilde{\Omega}\left(N^3 \max(m, n_0)^3\right)$.

4. Further Related Work

Some earlier works consider linearly separable data, see e.g. (Soudry et al., 2018; Brutzkus et al., 2018) and classification losses, and they show that (stochastic) gradient descent can find a global (max margin) solution which generalizes well. In contrast, the focus of this work is on the minimum level of over-parameterization required for general/worst-case data (Theorem 2.2 makes no assumption on the training data) that GD finds a global optimum for a regression loss.

(Du et al., 2019a) considers deep networks with the NTK parameterization (Jacot et al., 2018) and show that if all the hidden layers of the network have minimum width $\Omega(N^4)$, then GD converges to a global optimum. Their proof requires to track the evolution of the various Gram matrices defined recursively through all the layers. Recently, the width condition has been improved to $\Omega(N^3)$ by (Huang & Yau, 2020) for the same setting. (Nguyen & Mondelli, 2020) shows the global convergence of GD for deep pyramidal

nets with standard parameterization, where the width of the first hidden layer can scale linearly (or quadratically) in N.

For the setting of this paper, we are only aware of the results of (Allen-Zhu et al., 2019; Zou & Gu, 2019). One crucial difference which led to the proof of this work is that: in our setup, all the layers are optimized with standard GD, whereas in the prior works, the input and output layers are fixed, and only the inner layers are optimized. Interestingly, considering this slightly different setting allows us to reach a simple proof. In particular, to obtain a PL-like inequality, they used the part of the NTK corresponding to the gradient of the last hidden layer, whereas we used the gradient w.r.t the output layer. A few other works also study convergence of GD for ReLU networks, albeit for two-layer models, see e.g. (Arora et al., 2019; Du et al., 2019b; Oymak & Soltanolkotabi, 2020). Essentially, these results require that the number of hidden neurons scales at least as order of N^4.

Another closely related work is (Daniely, 2017). This paper shows that if the network is sufficiently large, the learning rate is small enough, the number of SGD steps is large enough, then SGD can learn any function in the conjugate kernel space associated to the output of the last hidden layer. Some crucial differences to our work are as follows. First, they consider online SGD, whereas we consider standard GD on a fixed data set. These two algorithms are different even if one sets the batch size to the size of the given training set, because in the former setting a fresh set of samples are drawn i.i.d. at every iteration, and thus the gradient at each iteration is not necessarily the same as the total gradient used by standard GD. Second, their result requires the size of the network to scale inversely with poly(ϵ), where ϵ is the desired error. Thus, in order for the error to get down arbitrarily close to 0, their network size needs to grow to infinity. In contrast, our work provides guarantees for network sizes independent of ϵ (i.e. only dependent on N). This difference also applies to several other works cited above. Third, they do not provide guarantees that the loss is monotonically decreasing over the iterations like ours. Fourth, their result shows that the loss after training is no worse than ϵ plus the minimum loss achievable by a function whose RKHS norm is bounded by a fixed variable M. However, it is not clear what is this minimum loss for different choices of M. This is especially important because their network size is also a polynomial function of M. In contrast, our work provides an end-to-end result where the loss is ensured to converge w.h.p. to 0 for a fixed size network. In terms of proof techniques, their proof is based on online SGD learning, while our proof is simpler and more transparent: it is based on the fact that if the smallest singular value of the last hidden layer is bounded away from zero during training, then GD linearly converges to a global optimum. Note that their proof as said is different, and it does not exploit this smallest singular value. Finally, let us highlight that this single idea is not

enough to prove the global convergence of the loss. To show the convergence, most of the prior works require to study the local Lipschitz property of the Jacobian (or gradient of the loss), which involves the study of various quantities related to the changes of activation patterns, see e.g. (Allen-Zhu et al., 2019; Zou & Gu, 2019; Oymak & Soltanolkotabi, 2020; Song & Yang, 2020; Du et al., 2019b; Arora et al., 2019). This has been a challenging task for deep ReLU nets since the derivative of ReLU is not Lipschitz continuous. In contrast, the proof of Theorem 2.2 does not require such analysis. The way it overcomes this issue is to introduce the transition matrix $G = F_{L-1}^k W_L^{k+1}$, and then decompose the loss at each iteration according to G. As shown, the remaining steps of the proof follow mostly from the triangle inequality. This makes the proof significantly simpler and more transparent. Furthermore, we have shown that this proof leads to improved over-parameterization condition in terms of the dependency on the number of training samples.

Acknowledgement

References

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

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

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

Chen, Z., Cao, Y., Zou, D., and Gu, Q. How much over-parameterization is sufficient to learn deep relu networks?, 2019. `arXiv:1911.12360`.

Daniely, A. Sgd learns the conjugate kernel class of the network. In *Neural Information Processing Systems (NeurIPS)*, 2017.

Davidson, K. R. and Szarek, S. J. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(140):317–366, 2001.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019a.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2019b.

Huang, J. and Yau, H.-T. Dynamics of deep neural networks and neural tangent hierarchy. In *International Conference on Machine Learning (ICML)*, 2020.

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

Nguyen, Q. and Mondelli, M. Global convergence of deep networks with one wide layer followed by pyramidal topology. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Nguyen, Q., Mondelli, M., and Montufar, G. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *International Conference on Machine Learning (ICML)*, 2021.

Oymak, S. and Soltanolkotabi, M. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Polyak, B. T. Gradient methods for minimizing functionals. *Zh. Vychisl. Mat. Mat. Fiz.*, 3(4), 1963.

Song, Z. and Yang, X. Quadratic suffices for over-parametrization via matrix chernoff bound, 2020. `arXiv:1906.03593`.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research (JMLR)*, 19:2822–2878, 2018.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018.

Zou, D. and Gu, Q. An improved analysis of training over-parameterized deep neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2019.